Basics of Biostatistics

Basics of Biostatistics

September, 2015

Marija J. Norušis
marija@norusis.com

Preface

This book is a first attempt at non-intimidating teaching material for an introductory course in biostatistics for students in low and middle income countries with limited access to MOOCs and other material on the internet. The emphasis is on basic concepts that are needed to understand results sections of medical publications. Examples are taken from open source journals which students can access. The focus is on problems that confront many developing nations: malaria, HIV, low birth weight infants. The book includes output from statistical software packages but, at present, is not tied to any software product that needs to be downloaded.

I am grateful for any corrections, suggestions, or comments that readers may have. Future versions will be downloadable under a Creative Commons license from www.norusis.com.

I am grateful to Dr. Richard Heller for encouragement and comments and to students and tutors at People's Open Access Education Initiative: Peoples-uni, www.peoples-uni.org, for corrections and suggestions.

Marija J. Norušis
marija@norusis.com

Contents

1 Summarizing and Displaying Data

- What types of graphical displays are useful for summarizing data?
- What are levels of measurement?
- What summary statistics describe central tendency? Dispersion?
- What is a standard score?

Whether you're presenting the results of a complicated international drug trial or gathering data for your own clinic or hospital, you have to summarize your observations. You know you can't just list ages, blood pressures, or number of malaria nets for everyone in the study and leave it to the readers and reviewers to draw their own conclusions. You have to select appropriate charts, tables and statistical measures to convey the information that is important.

There's no single best way to summarize and display data. The best way depends on the properties of the data and of the statistical measures, on your audience and on your personal tastes in charts and graphs. Charts and graphs that display only one or two variables are usually not too difficult to understand, while those that attempt to portray complex relationships, such as a map that shows the distribution and severity of dengue in different age groups over time may require quite a bit of effort to decipher.

Always remember that the purpose of visual displays is to make it easier for the audience to understand the results. If you're making the charts, even if you have software which can produce incredibly complicated displays, don't put more information on a chart than the human mind can absorb without overheating. Some published charts are impossible to understand without the author sitting next to you. (And it's possible that the author has relegated chart making to someone else and is embarrassed to admit that he/she doesn't quite understand it either.)

When calculating summary measures make sure they are appropriate for your data. The widespread availability of statistical software makes it so easy with a couple of clicks of the mouse to calculate all kinds of statistics, whether they make sense or not. (In this chapter "statistic" refers to any number calculated from your data values.) Don't calculate an average religion or ethnicity!

Scales of Measurement

When you conduct a study you typically collect multiple pieces of information, called variables, for each person or case. For example you may record height, weight, gender, and admitting hospital for a group of patients. Each patient has a value for each variable. If Faramola is 60 cm tall, 60 cm. is the value of the height variable. If you are using software to analyze your data you'll have to enter the values of the variables for all of the cases into what's called a data file.

How you assign values or symbols to what you are measuring is called the **scale of measurement**. Variables can be measured in different ways. Height can have many values, while gender is restricted to two values. Heights can be ordered from smallest to largest, while admitting hospital is just a name and can't be ordered based on the name alone. Variables are often classified into four categories based on how they are measured:

- **Nominal variables** have values that cannot be ordered in any meaningful way. Country of birth, name of drug, and color of hair are all examples of nominal variables. In a data file cases are assigned the same symbol if they have the same value of the attribute. The symbol doesn't mean anything.

- **Ordinal variables** have values that can be arranged in some sensible way, such as worst to best, but the distance between the coded values doesn't mean anything. If you rank your health on a scale from 1 (poor) to 10 (excellent), the distance between adjacent ratings doesn't have a clear meaning. A person with a rating of 10 is not twice as healthy as a person with a rating of 5, nor is the change in health from 5 to 6 the same as the change in health from 1 to 2. All you can say is that the person with a rating of 10 considers himself healthier than the person with a rating of 5.

- **Ratio variables** have values that can be ordered and the actual distance between values is interpretable. The distance between 60 cm and 62 cm is the same as the distance between 40 cm and 42 cm. Ratio variables also have an absolute 0 so that you can compute meaningful ratios between values. For example, a 20 year old person is twice as old as a 10 year old person. A variable that has meaningful distances between values but does not have an absolute zero point is said to be measured on an **interval scale.** Temperature is the usual example of an interval value since a 40 degree day cannot be said to be twice as hot as a 20 degree day, (unless temperatures are measured on the Kelvin scale which does have an absolute zero).

Before you start analyzing data you must consider the scale of measurement for the variables of interest. A statistical analysis which is appropriate for a variable like weight may be totally inappropriate for a variable like region of the country. Often numerical codes are chosen to represent nominal variables in a data file. For example the value of 1 may correspond to Africa, 2 to South East Asia, 3 to Europe and so on. That does not change the level of measurement of the variable. It is still a nominal variable. This is important if you are analyzing data files with a computer since statistical software will often spit out answers without any regard to whether the result makes sense.

Variables can also be classified based on the number of values they can have. **Discrete** or **categorical** variables, such as stage of a disease, drug administered, or number of Insecticide Treated Nets (ITNs) in a household, have a limited number of distinct values. **Continuous** variables such as blood pressure and weight can have many possible values. Ratio variables can be either discrete (family size) or continuous (height, weight). Ordinal variables (job satisfaction, health status) are usually discrete.

Frequency Table

If a variable has a small number of possible values a straightforward summary method is simply to count the number of times each value occurs. This is called a **frequency table**. For example, consider Figure 1.1 which shows the number of insecticide-treated bed nets (ITNs) per household in a village in south Ethiopia as reported by Loha et al. (2013). The symbol **N** or *n* is usually used for sample sizes in tables or charts.

Figure 1.1: Number of Insecticide Treated Nets in Household

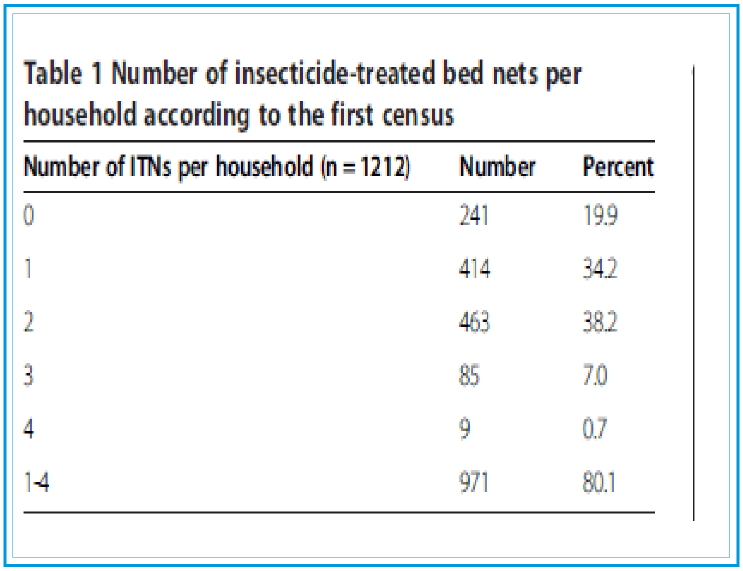


Table 1 Number of insecticide-treated bed nets per household according to the first census

Number of ITNs per household (n = 1212)	Number	Percent
0	241	19.9
1	414	34.2
2	463	38.2
3	85	7.0
4	9	0.7
1-4	971	80.1

From the table you see that 241 households did not have any ITNs. That's 19.9% of the 1212 households in the study (241/1212 times 100). A **percentage** tells you what part of the whole a particular piece is. The last row of the table tells you what percent of the households had any insecticide treated bed nets. That is, the percent that had 1, 2, 3 or 4 bed nets. A little over 80% of households had at least one ITN.

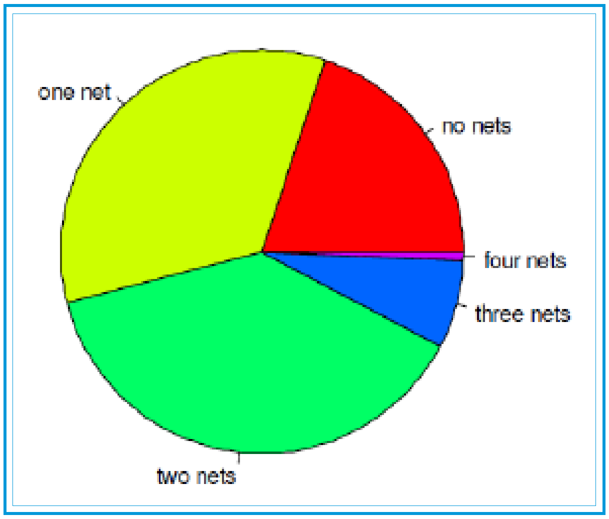
For variables where it makes sense to combine adjacent values, like age or income, you can make a frequency table where each row corresponds to a range of values. This is sometimes called a **grouped frequency table.** For example for age, you can set up categories like less than 5 years old, 5-11 , 12-18 years and older than 18, and count the number of cases in each category. Make sure that the categories don't overlap and that all possible values are included.

If you're the one actually analyzing the data instead of just reading someone else's results, a frequency table can reveal problems with your data that need to be corrected before you perform any additional analyses. For example if you find unlikely or impossible values in your table, such as 30 ITNs, you know that value should be checked and corrected. The person recording or entering the information probably made a mistake. If you don't correct errors in your data at the very beginning everything else you do may be incorrect. Check your ingredients before you start cooking!

Pie charts

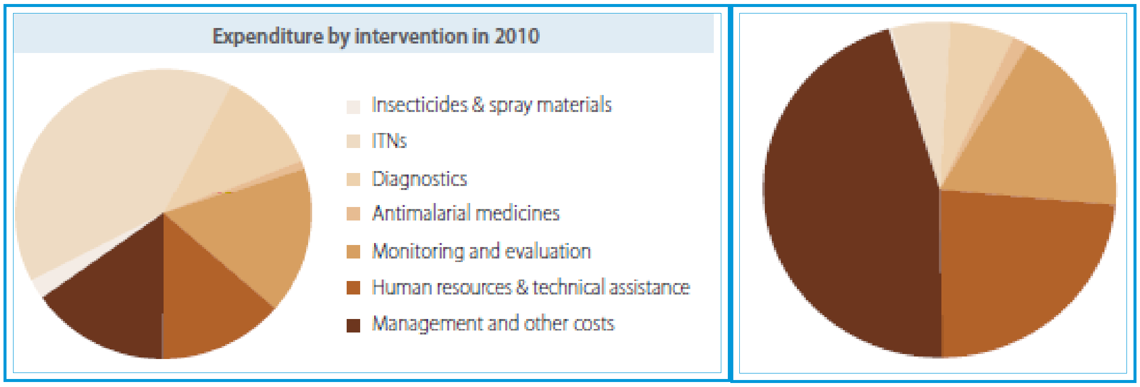
The information in a table is often easier to understand if you turn it into a visual display. Figure 1.2 is a **pie chart** for the ITN table. (A pie is a round cake.) There is a slice for each of the first five rows in the table. The size of the slice represents the counts or percentages for each value in the table. Many statisticians advise against the use of pie charts since humans aren't very good at visually comparing areas. That doesn't stop their use and you'll encounter pie charts in many publications and reports.

Figure 1.2: Pie Chart of Number of ITNs Owned by Household at First Census



Pie charts are not restricted to count data. They can be used whenever you subdivide a total into constituent parts. Figure 1.3 shows the total money spent on malaria control in India and Laos subdivided by type of activity (World Malaria Report, 2012).

Figure 1.3: Expenditures by Intervention India (left) and Laos

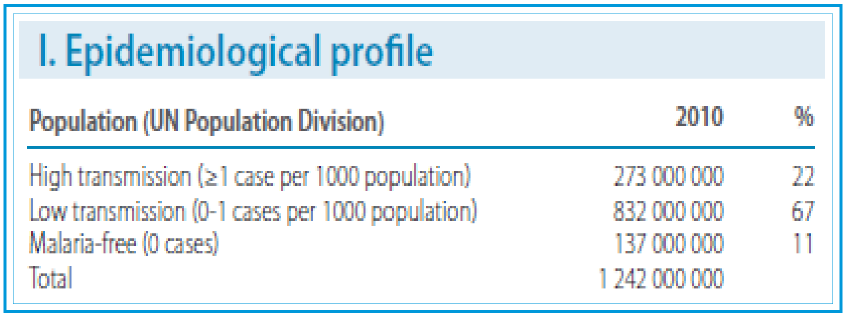


No doubt the designers of the report were concerned about the report being visually attractive so they selected shades of the same color to represent different interventions. Unfortunately that makes it difficult for the reader to distinguish between interventions represented by similar shades. You can tell, however, that India and Laos use their funds very differently. ITNs and diagnostics account for almost half of India's expenditures and roughly an eighth of Laos' expenditures. Laos reports that it spends almost half of its funds on management and other costs. Whenever data is self reported by individuals or agencies you have to be concerned about whether instructions and definitions are uniformly applied.

One important determinant of spending allocation is the incidence of malaria in a country. Look at the Epidemiological Profiles for India and Laos, shown in Figure 1.4 and Figure 1.5. A little more than a third of Laos' population (36%) lives in high transmission areas as compared to 22% for India. In

India 11% of the population is estimated to live in malaria free regions, compared to 41% for Laos. (The epidemiological profiles are grouped frequency tables where rows correspond to ranges of malaria incidence.)

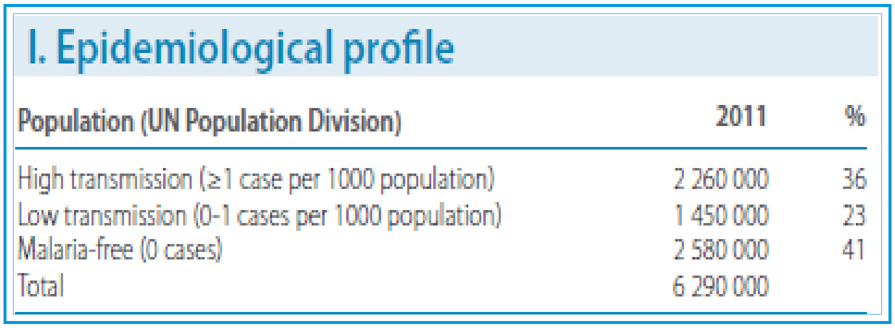
Figure 1.4: Epidemiological Profile for India



I. Epidemiological profile

Population (UN Population Division)	2010	%
High transmission (≥1 case per 1000 population)	273 000 000	22
Low transmission (0-1 cases per 1000 population)	832 000 000	67
Malaria-free (0 cases)	137 000 000	11
Total	1 242 000 000	

Figure 1.5: Epidemiological Profile for Laos



I. Epidemiological profile

Population (UN Population Division)	2011	%
High transmission (≥1 case per 1000 population)	2 260 000	36
Low transmission (0-1 cases per 1000 population)	1 450 000	23
Malaria-free (0 cases)	2 580 000	41
Total	6 290 000	

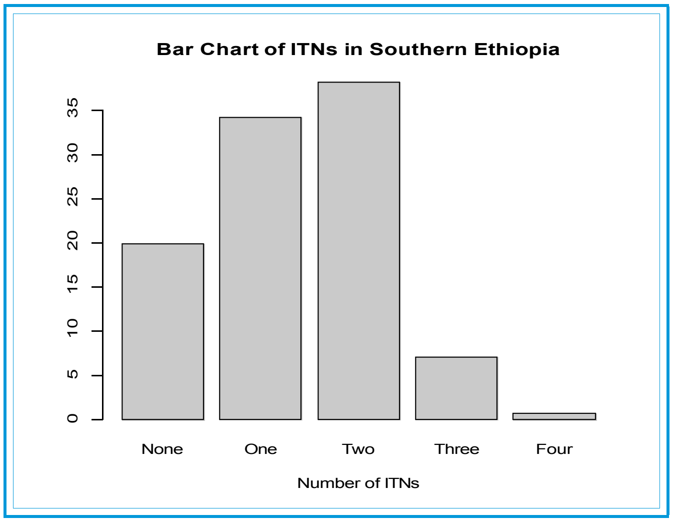
Exercise: Make Pie Charts of the epidemiological profiles for Laos and India . Combine the two epidemiological profiles into a single frequency table. Make sure to recalculate the percentages.

Bar charts

In a pie chart, the size of a slice depends on the number of cases or some other statistic, like the sum, for each category. In a bar chart the height of a bar depends on that statistic. Psychologists believe that humans process height better than area, so you'll often encounter bar charts when reading journals. Another advantage of bar charts is that it's easy to show multiple bars corresponding to various groupings of the data on the same chart. Bar charts come in many flavors as you will see.

In a bar chart you have as many bars as you have slices in a pie chart. Figure 1.6 is a bar chart of the data shown in Figure 1.1. The number of ITNs in a household is on the horizontal axis and the percentage of households is on the vertical axis. The pie chart and the bar chart are different visual displays of the same information.

Figure 1.6: Bar Chart of Number of Insecticide Treated Nets in a Household



Whenever you look at a chart, pay attention to the values on the axis which has the scale. That is, the axis that shows the percent or the count. If the scale does not start at 0 differences between bars are artificially magnified. You can make the difference between 20 and 22 appear very large if you start the scale axis at 19.

Histograms

If you have an ordered variable with many values you can display its distribution in a histogram. A **histogram** is like a bar chart except that adjacent values are combined into a single bar and distances on the horizontal axis are measured on a scale. The height of a bar depends on the number of cases that fall in that interval. The horizontal axis in a bar chart isn't ordered or scaled. For example, if you didn't have any cases with three nets, the bar for two nets would be next to the bar for four nets. You wouldn't see a space indicating that the three nets were never observed. In contrast, a histogram leaves blank space when there are no cases for a range of values.

Figure 1.7 shows a histogram of ages for a national survey of adults. Each bar represents a five year interval, instead of single years. If you didn't have any cases in a particular age group, say 40-45, unlike in a bar chart, you would see a gap corresponding to the 0 frequency. You see that the highest peak is for people in their fifties. Since this was a study of adults, the lowest acceptable age was 18. There was no upper limit on age.

A histogram shows detailed information about the distribution of a variable. You can tell if a distribution has a single peak or multiple peaks, whether there are gaps in the data, and whether there are **outliers**--values that are far removed from the rest. You can also tell whether the distribution is approximately **symmetric**, meaning that the two sides are mirror images of each other. If a distribution has one tail that extends farther from the center than the other the distribution is **skewed**. A distribution with a tail toward larger values is said to be skewed to the right, or have a positive skew; a distribution with a tail toward smaller values is skewed to the left, or negatively skewed. Income and age are variables that are often positively skewed. That's because there is a strict limit on how small the values can be but the upper limit is not restricted. Figure 1.8 shows what histograms look like for data with negative, positive and no skew.

Figure 1.7: Histogram of Age

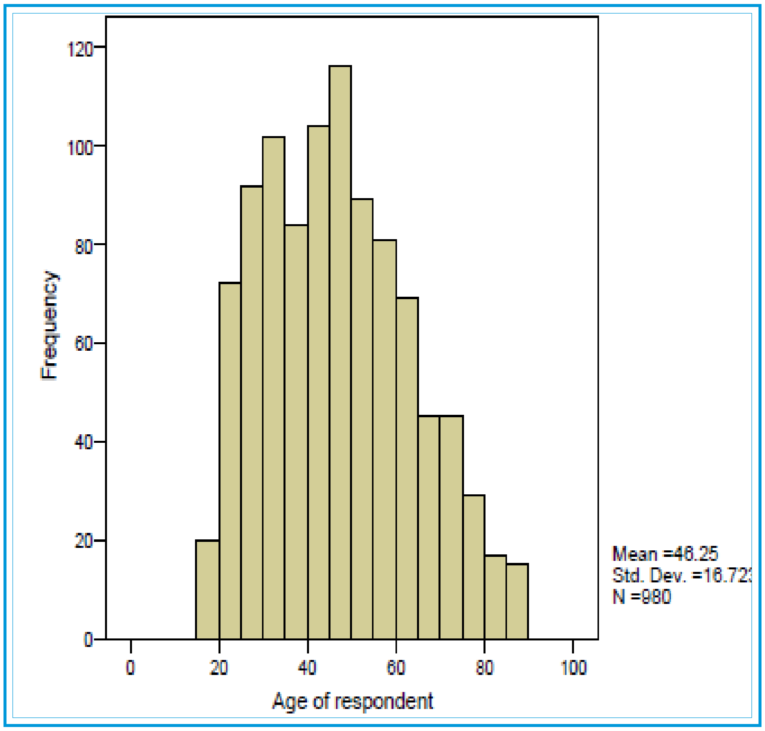
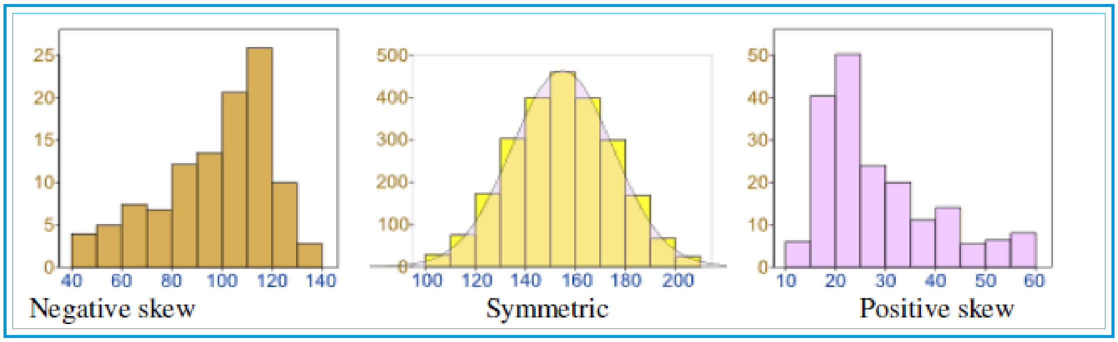


Figure 1.8: Distributions with Positive, Negative, and No Skewness



Distributions of variables are important for statistical testing of hypotheses and they are discussed in more detail in the chapter on Hypothesis Testing.

Grouped Bar Charts

Like pie charts, bar charts are used to display all kinds of summary measures besides counts. Figure 1.9 is a bar chart of ITN use in five countries over a three year period. For each year and country there are two bars, one for the proportion of the population with access to an ITN, the other for the proportion reporting sleeping under a net .

Figure 1.9: Access and Use of Insecticide Treated Nets

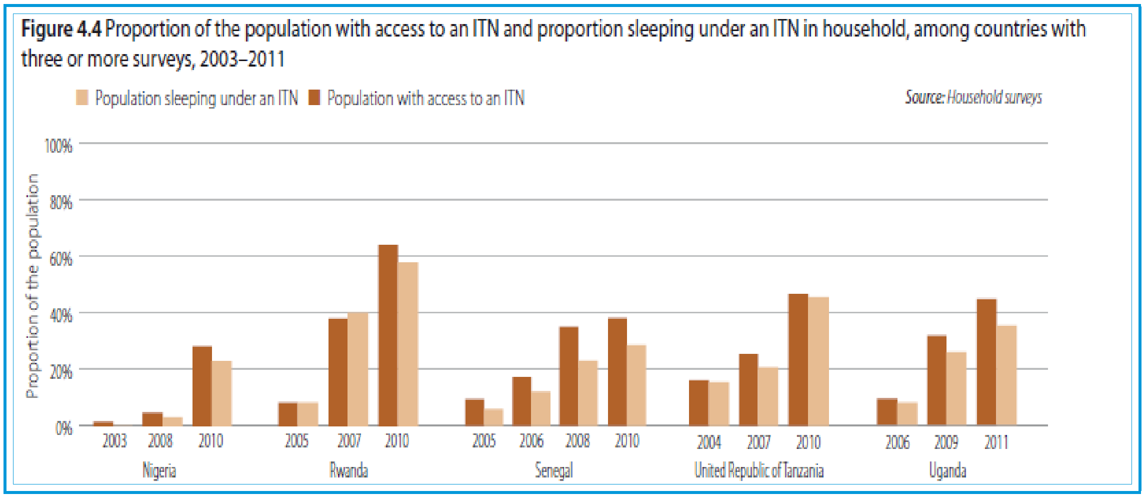


Figure 4.4 Proportion of the population with access to an ITN and proportion sleeping under an ITN in household, among countries with three or more surveys, 2003–2011

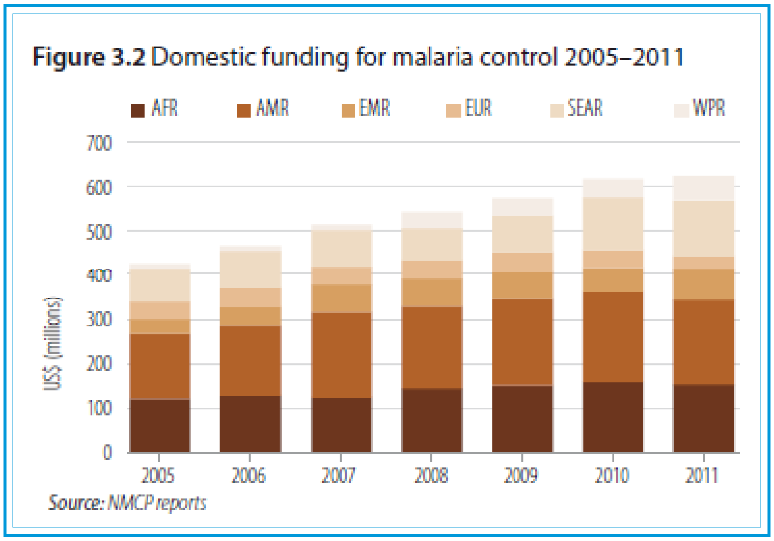
Source: Household surveys

You see that for each country the percent of the population with access to ITNs has increased with time as has the proportion sleeping under a net. Surveys show that for each year Rwanda has the greatest access and use.

Stacked Bar Charts of Sums

Instead of having bars next to each other, you can stack them one on top of another. As an example, consider Figure 1.10 from the World Malaria Report. Each bar represents total domestic funding for malaria control for one of the WHO regions in a year. You see that about 400 million US Dollars were spent for domestic funding of malaria control in 2005, increasing to over 600 million USD by 2011. The total spent is subdivided by region which is identified by the colors shown above the figure. Funding has increased or stayed the same in all regions except Europe, where it has decreased. In particular funding in South East Asia (SEAR) has increased since 2009.

Figure 1.10: Funding for Malaria Control



Cumulative Bar Charts

You know that the number of malaria cases and deaths vary by country. You can display the counts in a simple bar chart with each country corresponding to a bar. For a data set with 101 countries that's a lot of bars! A more informative display is to sort the countries by the number of deaths/cases and then to calculate the percentage of all malaria cases/ deaths that occur in that country and in all countries with a larger number of deaths. You cumulate the deaths across countries. That lets you identify the subset of countries that account for the largest proportion of malaria deaths.

Consider Figure 1.11. Nigeria has both the most malaria cases and the most deaths (25% of the cases and 30% of the deaths) so it is the first bar in both plots.

Figure 1.11: Cumulative Proportion of Malaria Cases and Deaths

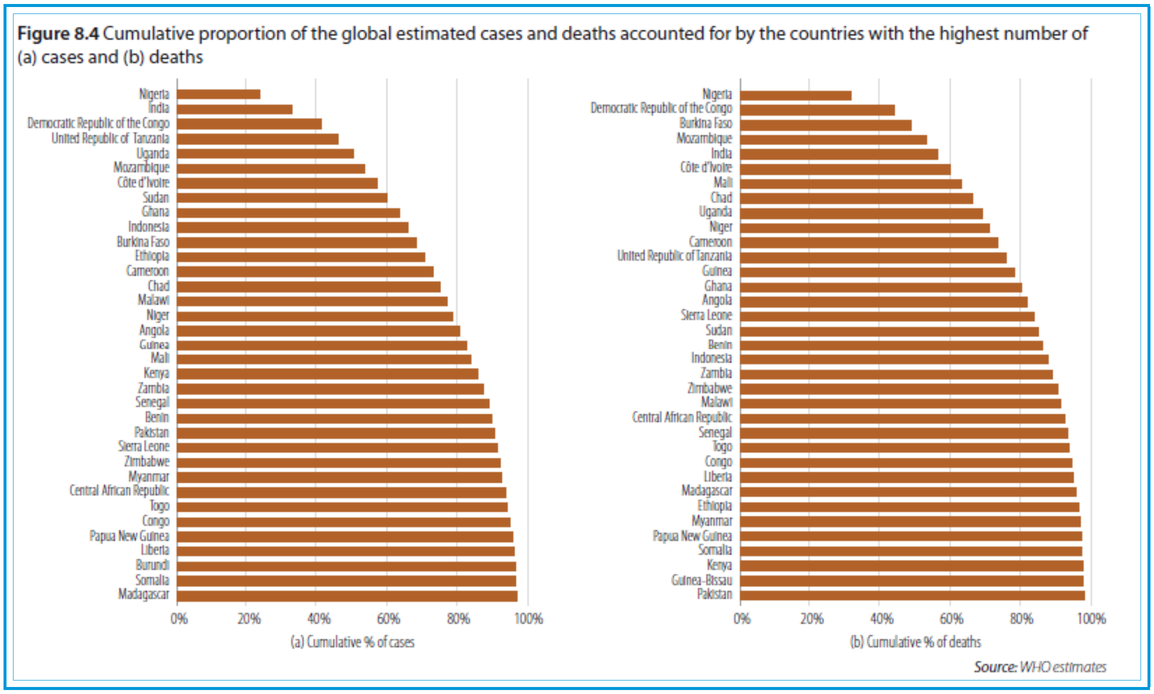


Figure 8.4 Cumulative proportion of the global estimated cases and deaths accounted for by the countries with the highest number of (a) cases and (b) deaths

Source: WHO estimates

In the case plot (on the left) the bar for India is the percent of total cases due to both Nigeria and India. The difference between the first two bars is the percent of cases for India. Each bar is calculated from the sum of that country's deaths and the deaths of all countries with larger values. You see that a small number of countries (with large populations) account for most of the cases and deaths. Eight countries are responsible for 60% of the cases of malaria; six countries are responsible for 60% of all malaria deaths.

Exercise: Which countries account for 50% of the malaria deaths? Which for 50% of the cases?

Line Charts

If a variable is ordered in some meaningful way, for example time, health status, or malaria incidence, an alternative to a series of bar charts is a line chart. For example, instead of showing ten bars that show the percentage of household using ITNs at 10 time points, you mark what would have been the height of the bar with a single point and then connect the points with a line. The horizontal scale doesn't have to have equal intervals but the categories should be ordered.

Figure 1.12 is a plot of ITN use by time for the 101 households in the malaria study. The horizontal axis is the week of study. For each week, instead of a bar, you place a point that shows the percentage of the households that use ITNs. Then you connect the points with a straight line. (It would take a lot of bars to show the same information!) If you have groups you can use different symbols for each group. Figure 1.12 uses different symbols for the group as a whole and for males and females. The intervals on the horizontal axis are equal, since they are weeks, but they don't have to be.

Figure 1.12: Insecticide treated Bed Net Use by Gender

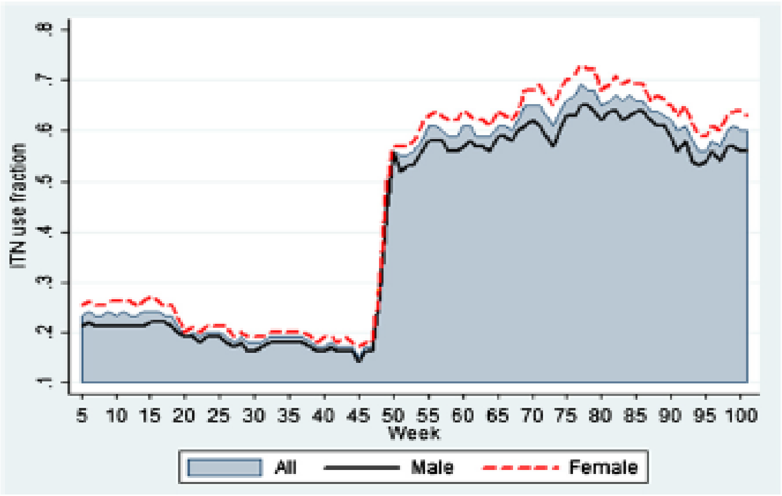


Figure 3 Insecticide-treated bed net use fraction by gender.

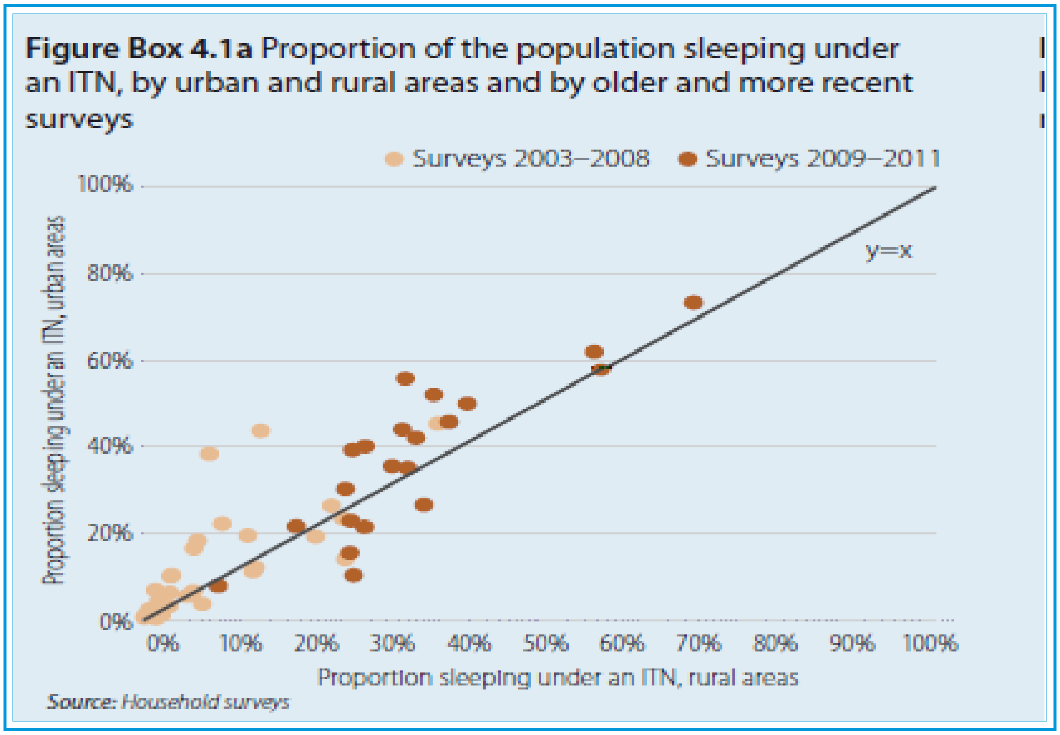
From the line chart in Figure 1.12 you see that something happened to cause a marked increase of ITN use. In fact, at week 48 ITNs were freely distributed to the participants. This caused the ITN use fraction to increase dramatically. The fraction was about 0.2 at week 47 and then skyrocketed to over 0.6 at week 48. Both before and after the intervention women were more likely to use nets than men. (Note that the vertical axis starts not at 0 but at 0.1. That makes the differences between men and women appear larger on the plot than if the axis started at 0.)

Scatterplots

In a line chart each location on the horizontal axis has one point, or several points if several groups are plotted on the same chart. For example, at week 10 you have a single value for the proportion of ITN use for males, another value for females and then one for both combined. You're summing or averaging the values of a variable for each location on the horizontal axis.

In a scatterplot you have two variables for each observations and you plot the values for each pair on the horizontal and vertical axis. Figure 1.13 is a scatterplot of the proportion of the population sleeping under ITNs for rural and urban areas as estimated from older and more recent surveys. For each survey you have two values: the proportion sleeping under ITNs for rural areas and the proportion sleeping under ITNs for urban areas. Those are the values that are shown in Figure 1.13.

Figure 1.13: Insecticide Treated Net Use In Urban and Rural Areas



Look at the rightmost point on the scatterplot. It corresponds to a survey with a value of 70% for rural areas and a little more than 70% for urban areas. The line that's drawn through the scatterplot is where points would fall if rural and urban areas had exactly the same proportions sleeping under ITNs. Surveys that found urban areas to have larger percentages of people sleeping under the net than rural areas are above the line. Surveys in which rural areas had larger percentages than urban areas are below the line. Regression lines which are used to summarize the data points on a scatterplot are discussed in great detail in the Chapter 6.

Descriptive Statistics

Frequency tables, bar charts, scatterplots and histograms are essential first steps in examining your data. From them you can see how often different values of a variable occur. Often, however, you need to summarize the information further, especially if a variable has many distinct values. You want to compute measures that describe the "typical" value as well as how much the data spread out around this value. You want to answer questions such as "what's the average number of ITNs in a household?" and "How much spread is there in the number of ITNs owned?" Sometimes there is no good answer to these questions. For example, if your data showed only households with no ITNs or 3 ITNs, trying to summarize the data by pooling the two groups would lead to misleading results.

You should never calculate summary descriptive statistics without first examining the distribution of the data values using tables and charts. Summary statistics can hide potentially serious problems with your data. Unless you examine all of the values you won't detect errors in data recording or entry, errors in setting up a data file or data values that may have been corrupted by your software.

Measures of Central Tendency

The arithmetic average, median and mode are the most frequently reported measures of "typical." The **mode** is the value that occurs most frequently. From Figure 1.1 you see that the largest number of households (463) had two ITNs, so that is the mode. It's not difficult to see why the mode alone is not a very good summary measure. If you know that the mode is two ITNs it's possible that almost all households have two nets. It's also possible that 605 households don't have any nets and 607 have two nets. The mode doesn't tell you anything about the distribution of the values. If you are reporting a mode make sure you also indicate the percent of cases in that category. You should say "The mode is 2, with 38.2% of households having two ITNs."

For a variable measured on a nominal scale, like type of insecticide used, or how the net was obtained, the mode is the only summary measure that you can compute. For variables that have values that can be ordered you can use this additional information for calculating a measure of central tendency.

Median

The median makes better use of the data than the mode. The **median** is the value that is in the middle when the data points are sorted from smallest to largest. For example, if the ages of five people are 28, 29, 30, 31 and 32, the median is 30 years, since it is the middle value. Half of the cases have values greater than the median and half have values less than the median. If you have an even number of cases there isn't a single number in the middle, instead there are two numbers. For example, if the values are 21, 28, 29, 30, 31 and 40, the middle values are 29 and 30. To calculate the median you add up the two middle values and divide by 2. In this case the median is (29+30)/2 = 29.5.

A shortcoming of the median as a summary measure is that it ignores much of the available information. The median for 10, 15, 17, 19, and 20 is the same as the median for 0, 0, 17, 100, 20000. The actual amount that values fall above or below the median has no effect on the median. This can be an advantage if you have one or more outlying values which are far removed from the rest since they will have no effect on the median.

Arithmetic Mean

The most commonly used measure of "typical" is the **arithmetic mean**, also known as the average. The mean uses the actual values of all of the cases. To compute the mean, add up the values of all of the cases and then divide by the number of cases.

The arithmetic mean of the five values 28, 29, 30, 98 and 190 is

$$Mean = \frac{(28+29+30+98+190)}{5} = 75$$

One of the disadvantages of the arithmetic is that cases with values that are very different from the rest can have a large effect on the mean, especially if the sample sizes are small. If you observe 5 households with 0, 1, 1, 2 and 20 ITNs, the arithmetic mean is 4.8. That's not a very good indicator of the typical number of nets in the households.

For a symmetric distribution the mean, median and mode are the same--the value that corresponds to the peak. If you have a distribution with several peaks, the sample mean may fall in between the peaks and not be at all a good measure of typical.

To prevent cases with outlying values from distorting the mean, you can calculate what's called a trimmed mean. A **trimmed mean** is calculated just like the usual arithmetic mean, except that a designated percentages of cases with the largest and smallest values are excluded from the calculation. This makes the trimmed mean less sensitive to outlying values. The 5% trimmed mean excludes 5% of largest values and the 5% of smallest values. It's based on the 90% of cases in the middle. The trimmed mean provides an alternative to the median when you have some data values that are far removed from the rest. (In your readings you may also encounter statistics called **M-estimators** which instead of discarding the largest and smallest sets of values give them less weight when calculating the mean.)

Exercise: In the table that follows why do you think the authors report the median instead of the mean for weeks of ITN use? How would you summarize the table?

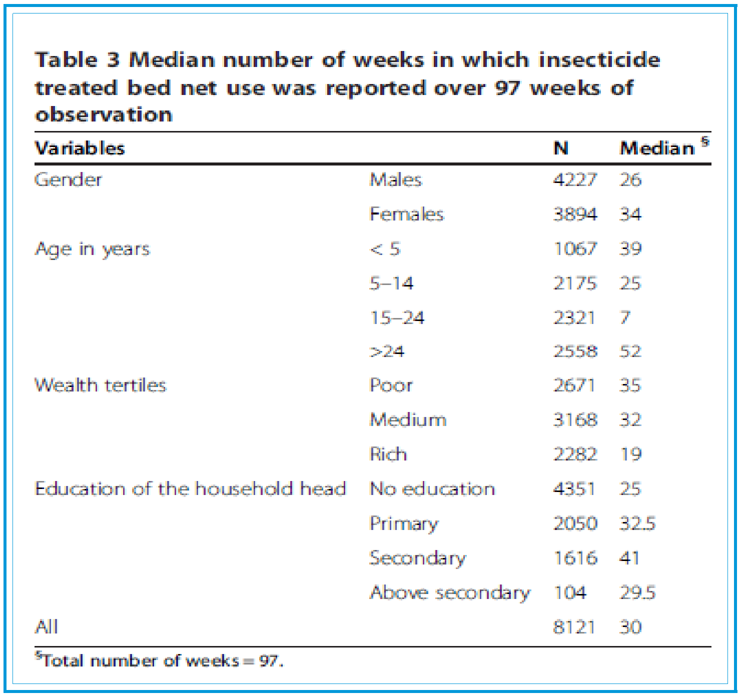


Table 3 Median number of weeks in which insecticide treated bed net use was reported over 97 weeks of observation

Variables		N	Median [§]
Gender	Males	4227	26
	Females	3894	34
Age in years	< 5	1067	39
	5–14	2175	25
	15–24	2321	7
	>24	2558	52
Wealth tertiles	Poor	2671	35
	Medium	3168	32
	Rich	2282	19
Education of the household head	No education	4351	25
	Primary	2050	32.5
	Secondary	1616	41
	Above secondary	104	29.5
All		8121	30

[§] Total number of weeks = 97.

Measures of Variability

Measures of central tendency don't tell you anything about how much the data values differ from each other. For example, the mean and median are both 50 for these two sets of ages: 50, 50, 50, 50, 50 and 10, 20, 50, 80, 90. However, the distribution of ages is very different between the two sets. Measures of variability quantify the spread of observations.

The **range** is the most straight forward measure of spread. It's the difference between the largest observed value (the **maximum**) and the smallest observed value (the **minimum**). A large value for the range tells you that the largest and smallest values differ substantially. It doesn't tell you anything about how much the values in between vary. For the first set of ages, the range is 0; for the second set the range is 90 – 10, or 80.

A better measure of variability is the **interquartile range** which is the range when cases with the 25% largest and 25% smallest values are excluded. It's the distance between the 75th and 25th percentile. Unlike the ordinary range, the interquartile range is not easily affected by extreme values.

Variance and Standard Deviation

The most commonly used measure of variability is the **variance**. It is based on the squared distances between the values of the individual cases and the mean. To calculate the squared distance between a value and the mean, just subtract the mean from the value and then square the difference. (One reason you must use the squared distance instead of the distance is that the sum of distances around the mean is always 0.) To get the variance, sum up the squared distances from the mean for all cases and divide the sum by the number of cases minus 1.

For example, to calculate the variance of the five numbers 28, 29, 30, 98, and 190, first find the mean. It is 75. The sample variance is then

$$s^2=\frac{((28-75)^2+(29-75)^2+(30-75)^2+(98-75)^2+(190-75)^2)}{(5-1)}=5026$$

If the variance is 0, all of the cases have the same value. The larger the variance, the more the values are spread out. To obtain a measure in the same units as the original data, you can take the square root of the variance and obtain what's known as the **standard deviation.** For this example the standard deviation is 70.9.

The Coefficient of Variation

The magnitude of the standard deviation depends on the units used to measure a variable. For example, the standard deviation for age measured in days is larger than the standard deviation of the same ages measured in years. (In fact, the standard deviation for age in days is 365.25 times the standard deviation for age in years.) Similarly, a variable such as yearly salary will usually have a larger standard deviation than a variable such as height in meters.

The **coefficient of variation** compares the standard deviation to the mean. It tells you what percent of the mean the standard deviation is. To compute the coefficient of variation, just divide the standard deviation by the mean and multiply by 100. (Take the absolute value of the mean if it is negative.)

$$\text{coefficient of variation} = \frac{\text{standard deviation}}{|\text{mean}|} \times 100$$

For example, if the age in days has a mean of 1500 days, and a standard deviation of 200 days, the coefficient of variation is

Coefficient of Variation=(200 days/1500 days) x 100 =13.3%

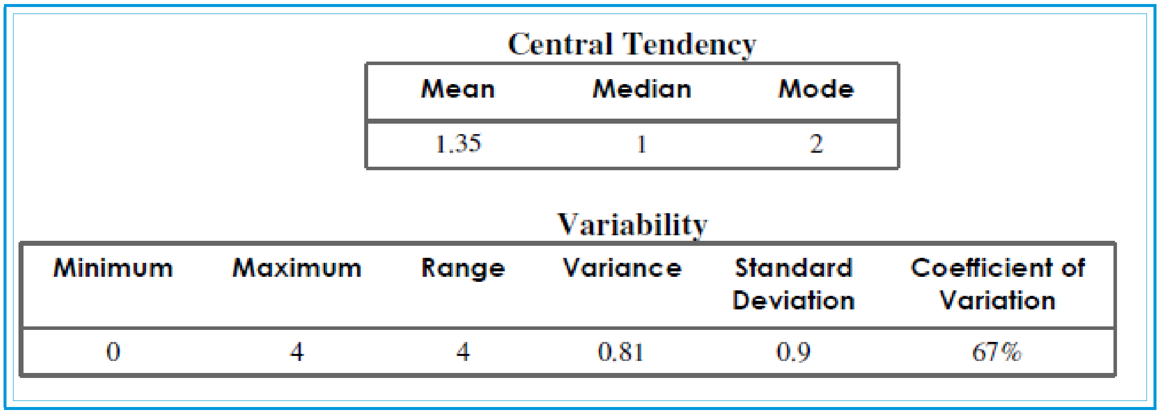
The coefficient of variation will be the same if age is recorded in days, months or years, provided it is recorded in the same detail. Since both the mean and the standard deviation are in the same units, the coefficient of variation is unit free.

The coefficient of variation allows you to compare the variability of different variables. For example, you can compare whether systolic blood pressures are more variable than diastolic blood pressures or whether the number of children is more variable than the number of ITNs.

Summarizing the Insecticide Treated Net Data

Figure 1.14 shows descriptive statistics for the number of insecticide treated nets in a household, as shown in Figure 1.1. The number of households in the study is 1212. When examining the results of any study always note how many observational units (people, animals, schools) were initially enrolled in the study and how many appear in the results. There are many possible reasons why observations are incomplete. Some are innocuous, like being unable to read a number on a form, but others can seriously discredit the results. If patients find a treatment unbearable and refuse to continue, eliminating them from the study will make the results impossible to interpret. In the malaria study if people who refuse to use ITNs didn't continue in the study, it would make ITN use appear to increase even if there is little change.

Figure 1.14: Summary Statistics of ITNs Owned for 1212 Households



Central Tendency

Mean	Median	Mode
1.35	1	2

Variability

Minimum	Maximum	Range	Variance	Standard Deviation	Coefficient of Variation
0	4	4	0.81	0.9	67%

The average number of nets is 1.35. Half of the households have one net or fewer. Thirty eight percent of households have two nets, the most frequently occurring number. The range is 4, the difference between the smallest value (0) and the largest value (4). The variance is 0.81, so the standard deviation is 0.90, its square root. The standard deviation is 67% of the mean, so that is the coefficient of variation.

Calculating the Summary Statistics

Since you have only 5 possible values for the number of insecticide treated nets, calculating the descriptive statistics is particularly easy. The sum of the number of nets is

sum= (241 x 0)+(414 x 1) + (463 x 2) + (85 x 3) + (9 x 4) = 1631

$$\text{mean} = \frac{1631}{1212} = 1.35$$

Similarly the variance is

$$s^2 = \frac{(241(0-1.35)^2+414(1-1.35)^2+463(2-1.35)^2+85(3-1.35)^2+9(4-1.35)^2)}{1211}$$

$$= 0.81$$

Exercise: You saw that 20% of all household do not have any ITNs. Above they are included in the computation of the statistics with a value of 0. It might be more informative to restrict the computation of the summary statistics to households that actually have nets and to report the number of households without nets separately. Recalculate the descriptive statistics for households that actually had ITNs. Compare the two sets of statistics. Which do you think is a better summary? In their results section the authors of the paper say that the average number of ITNs is 1.68. Do you recognize this number? Do you think the authors should have mentioned that the average does not include households that don't have any nets?

Percentiles

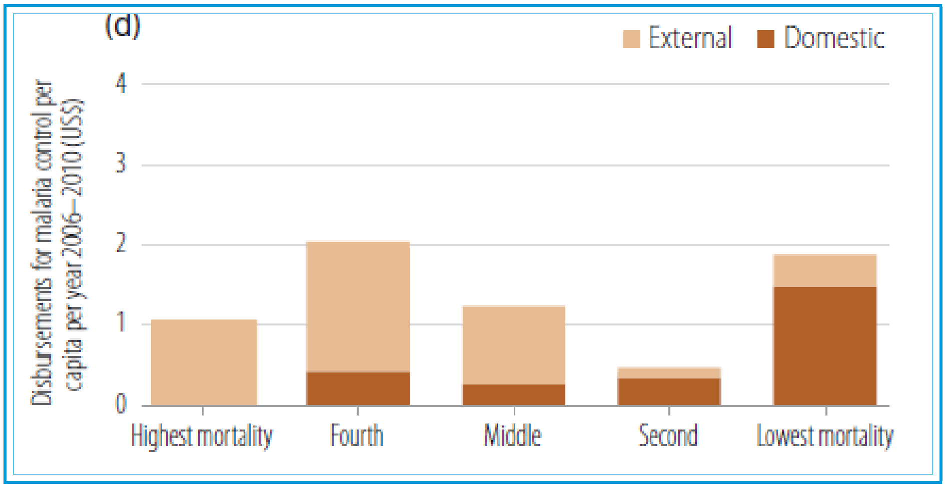
The median splits a sample into two equal parts based on the values of variable. You can compute values that split the sample in other ways as well. For example, you can find the value below which 25% of the observations fall, or the value below which 90 percent of the observations fall. Such values are called **percentiles**, since they tell you the percentage of cases with values below them. Twenty five percent of cases have values smaller than the 25th percentile, so 75% of the cases have values larger than the 25th percentile.

Many standardized tests give you percentile rankings as well as scores. That lets you determine where you stand in the pool of test-takers. A score at the 90th percentile is great, since 90% of people scored worse than you did; only 10% scored better. A percentile of 15 is nothing to brag about, since 85% of people scored better than you did. In contrast, when studying malaria cases and deaths, having low percentile values when compared to other countries is good since that means you have fewer cases and deaths.

Together, the 25th, 50th, and 75th percentiles are known as **quartiles**, since they split the sample into four groups with approximately equal numbers of cases. Values that divide the sample into three groups are called **tertiles**; into five groups, **quintiles** and into ten groups, **deciles.**

Figure 1.15 shows domestic and international per capita disbursements when countries are subdivided into five equal sized groups, quintiles, based on their observed malaria mortality. Countries with highest malaria mortality rates have the largest external disbursements. As mortality rates decrease domestic spending increases.

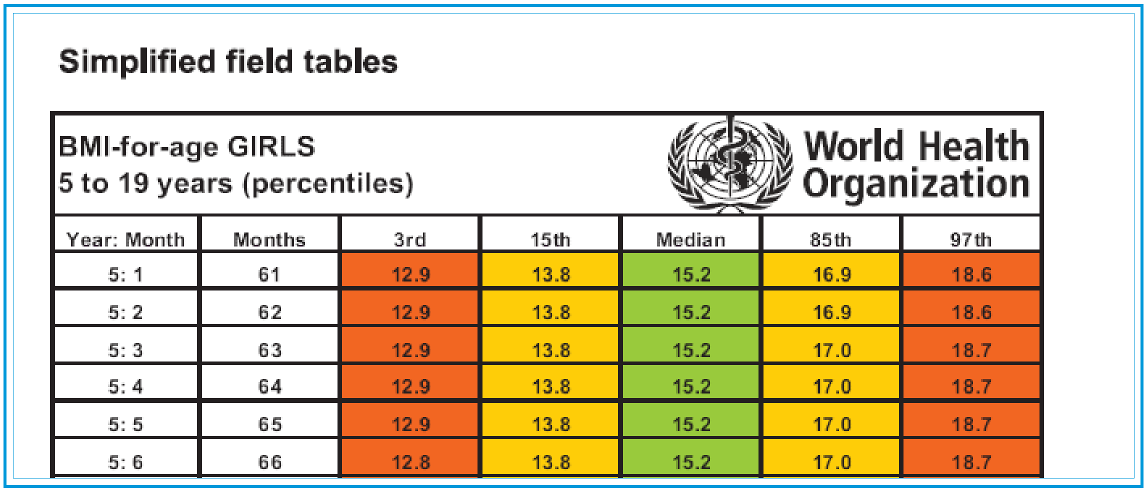
Figure 1.15: Spending by Source for Quintiles of Malaria Deaths



Exercise: Based on Figure 1.11, which countries fall into the top quarter of malaria cases? malaria deaths? Can you determine all of the countries that fall into the bottom quarter? (There are 101 countries total.)

Figure 1.16 shows World Health Organization estimated percentiles of body mass index for girls 5 to 19 years of age. (The body mass index is calculated as the weight in kilograms divided by the square of the height in centimeters.) Since values are shown by month of age, as you would expect the percentiles are constant over the six month interval shown. Half of all girls between 5 and 19 are estimated to have a BMI less than 15.2 and half greater than 15.2. Only 3 percent of girls have BMIs greater than 18.6 or less than 12.9. Having normative values is important for evaluating risk factors for individuals.

Figure 1.16: Percentiles of BMI for Girls 5-19 years of age



Simplified field tables

BMI-for-age GIRLS
5 to 19 years (percentiles)

World Health Organization

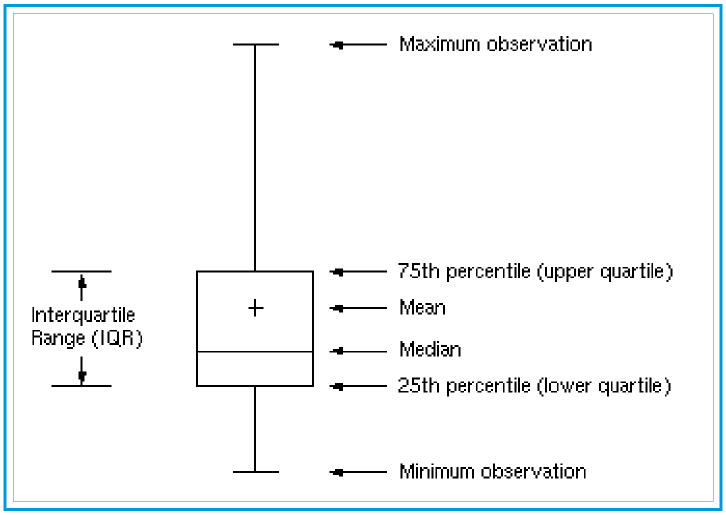
Year: Month	Months	3rd	15th	Median	85th	97th
5: 1	61	12.9	13.8	15.2	16.9	18.6
5: 2	62	12.9	13.8	15.2	16.9	18.6
5: 3	63	12.9	13.8	15.2	17.0	18.7
5: 4	64	12.9	13.8	15.2	17.0	18.7
5: 5	65	12.9	13.8	15.2	17.0	18.7
5: 6	66	12.8	13.8	15.2	17.0	18.7

Box Plots: Displaying Percentiles

When you use bar charts to compare means of groups, you ignore important information about the data. You can't tell anything about the distribution of data values by looking at just the means. If you present histograms for each group you suffer from the problem of too much information. It's difficult to compare a bunch of histograms.

A box plot is a display that conveys much more information than a bar chart but less than the individual histograms. It is particularly useful when you want to compare the distributions of several groups. A **box plot** simultaneously displays the median, the interquartile range, and the smallest and largest values for each group. Figure 1.17 is a basic annotated box plot. The length of the box indicates the variability of the observations. Long boxes have data values with more spread than short boxes. The lower boundary of each box is the 25th percentile, the upper boundary the 75th percentile. Half of the data values are within the lower and upper boundaries of the box. The lines extending from the ends of the box are called “whiskers” and the plot is often called a **box and whisker plot**. Usually in a box plot the whiskers extend to the smallest observed data values that aren't outliers, where an outlier is a point more than 1.5 box lengths from the edge of the box. (Sometimes box plots are displayed horizontally rather than vertically.)

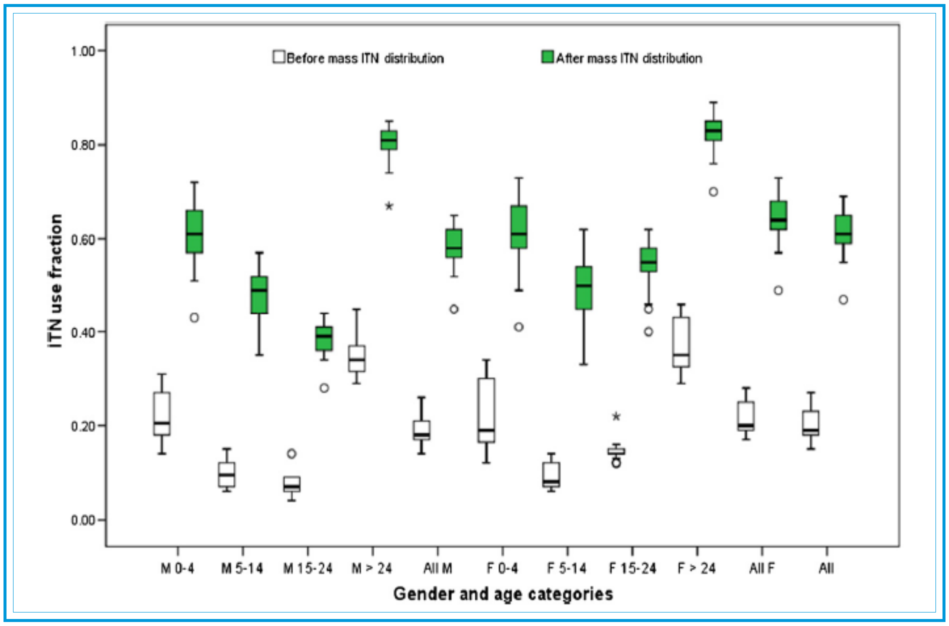
Figure 1.17: Schematic Box Plot


Maximum observation
Interquartile Range (IQR)
75th percentile (upper quartile)
Mean
Median
25th percentile (lower quartile)
Minimum observation

From a box plot you can tell about the shape of a distribution. If the median line is close to the center of the box, and the distribution of data values has a single peak and is continuous, the distribution of the variable is more or less symmetric. If the median is closer to the bottom of the box than to the top, the data are positively skewed. This means that it has a tail toward larger values. If the median is closer to the top of the box than to the bottom, the opposite is true: the distribution is negatively skewed. The length of the tail is shown by the whiskers.

Figure 1.18 is from the the same study as the frequency table. The author is describing the fraction of people who slept under an insecticide treated nets before and after free ITNs were distributed. There are separate boxes for age and sex groups.

Figure 1.18: Insecticide Treated Net Use Fraction by Gender and Age



The white boxes represent the fraction of people reporting ITN use the prior night before ITN distribution, the green, the fraction after the distribution. Surveys were conducted weekly so the boxes are based on the fraction reporting use each week. For every group net use increased after the distribution, so all of the green boxes are above their corresponding white boxes. Look at the last two boxes which represent all cases. The median ITN fraction before distribution was around 0.18. Half of the time fewer than 18% of all individuals slept under a net. After the net distribution, about 60% of the time people slept under a net. That's a big improvement.

The bottom of a box is at the 25th percentile, the top is at the 75th percentile. At the beginning of the study, seventy-five percent of the time between 17% and 22% of people reported sleeping under a net. After the net distribution, the interquartile range is from about 58% to 64%, indicating that half of the time between 58% and 64% of people slept under a net. The lengths of the before and after boxes are similar, because the interquartile range didn't change much.

The open circles and stars represent outlying cases. They are weeks in which net were used much more or much less than the others. The open circle weeks are not as far from the ends of the whiskers as are the star cases. Most boxes have lines (medians) which are not in the middle but closer to the bottom. That means that the distribution has a tail to the right. Frequent net use is reported more often than infrequent use.

The collected box plots are a good summary of the results. You see that for all age groups ITN use increased after the distribution. You can easily look for differences that may be attributable to age and sex.

Standard Scores

The mean often serves as a convenient reference point to which individual observations are compared. Whenever you receive an examination back, the first question you ask is, How does my performance compare with the rest of the class? An initially dismal-looking score of 65% may turn stellar if that's the highest grade. Similarly, a usually respectable score of 80 loses its appeal if it places you in the bottom quarter of the class. If the instructor just tells you the mean score for the class, you can only tell if your score is less than, equal to, or greater than the mean. You can't say how far it is from the average unless you also know the standard deviation.

For example, if the average score is 70 and the standard deviation is 5, a score of 80 is quite a bit better than average. It is two standard deviations above the mean. If the standard deviation is 15, the same score is not very remarkable. It is less than one standard deviation above the mean. You can determine the position of a case in the distribution of observed values by calculating what's known as a **standard score,** or ***z*- score.** (When variables have normal distributions the z -score is particularly useful since it allows you to position the case exactly in the distribution. That's a topic for chapter 3: The Normal Distribution.)

To calculate the standard score, first find the difference between the case's value and the mean and then divide this difference by the standard deviation:

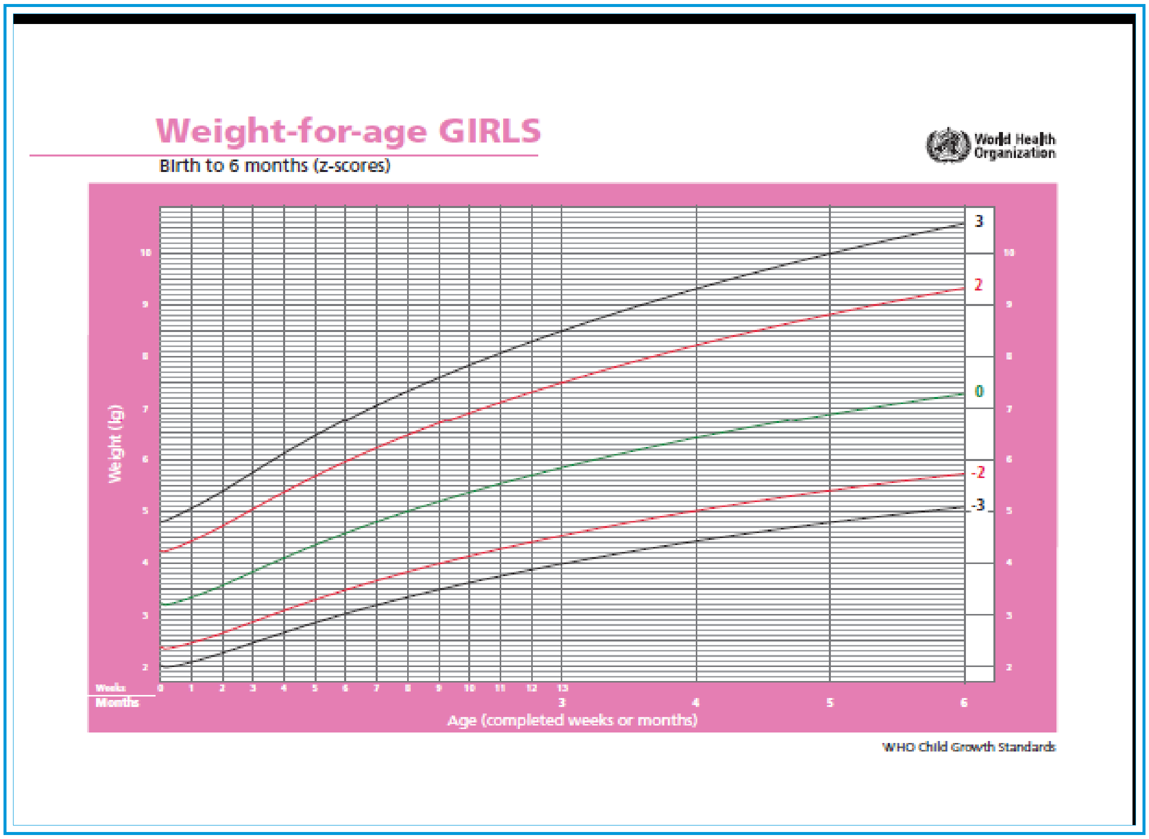
$$standard\ score = \frac{(value - mean)}{(standard\ deviation)}$$

A standard score tells you how many standard deviation units a case is above or below the mean. If a case's standard score is 0, the value for that case is equal to the mean. If the standard score is 1, the value for the case is one standard deviation above the mean. If the standard score is –1, the value for the case is one standard deviation below the mean. (For many types of distributions, including the normal distribution discussed in chapter 3, most of the observed values fall within plus or minus two standard deviations of the mean.) The mean of the standard scores for a variable is always 0, and their standard deviation is 1.

Standard scores allow you to compare relative values of several different variables for a case. For example, if a person has a standard score of 2 for income and a standard score of –1 for education, you know that the person has a larger income than most and somewhat fewer years of education. You can't meaningfully compare the original values, since the variables all have different units of measurement, different means, and different standard deviations.

Figure 1.19, distributed by the World Health Organization, plots z-scores for the weight of girls from birth to 6 months. The chart allows you to determine how a particular girl's weight compares to that of other girls of the same age. For each age the average weight and standard deviation of a large number of girls was obtained. Then for each age, the z-score for a particular weight is calculated by subtracting the mean weight for that age from the weight of interest and dividing by the standard deviation. (WHO actually uses the median instead of the mean for calculating the z-score. For a symmetric distribution that the mean and median are the same.)

Figure 1.19: Z- scores for Weight of Girls Birth to 6 Months



The green curve shows the average weight at each age. It corresponds to z-scores of 0. Girls whose weights are close to the green curve value for their age have average weights for their age. A girl whose weight is above the mean for her age will have a positive z-score. The heavier the girl the larger the z-score. If a girl weighs less than average she will have a negative z-score. The red curves show weights for z-scores of +2 and -2. You expect roughly 68% of girls to have z-scores between +1 and -1. Approximately 95% of girls to have z-scores between -2 and +2. Girls with z-scores larger than 2 in absolute value are quite atypical and may need to be evaluated.

Summary

In this chapter you worked with only the most basic of charts and summary statistics. But don't underestimate their importance. It doesn't matter how complicated or simple the paper you are reading or the data you are analyzing is, careful examination of the data values and of relationships between variables is essential. Complex statistical procedures and tests will never replace the insights you gain from basic analyses.

In the next chapter you'll learn how to draw conclusions about a population based on a sample. That's where the mysterious *p-value* you may have heard about makes an appearance.

Bibliography

World Malaria Report 2012

WHO Child Growth Standards

Loha et al: Freely Distributed Bed-net Use Among Chano Mille Residents, South Ethiopia: a Longitudinal Study. *Malaria Journal* 2013, 12:23.
Loha (2013)

2 Evaluating Results from Samples

- What is a population?
- What is a random sample?
- Is a sample a miniature of the population?
- What is the sampling distribution of a statistic?
- What factors determine how much sample means vary ?
- What's the standard error of the mean?
- What's an observed significance level?

In the previous chapter you used charts and descriptive statistics to summarize the observed data. Your only concern was how to describe the observations at hand. It's straightforward to say that at the beginning of a study people on average reported that they slept under insecticide treated nets (ITNs) 30% of the time and when free nets were distributed they said on average that they slept under ITNs 60% of the time. Unless there were errors recording or entering the data you can speak confidently about what you actually observed. But is that really enough? Do you want to draw conclusions only about the participants of your study, about the people you actually observed? That's unlikely. What you probably want to know is whether distributing free nets in malarial areas is an effective strategy for increasing ITN use for all people in malarial areas in your district, country, or region. You want to draw conclusions about more people than those that you actually observed. That's the topic of this chapter and the focus of much of the field of inferential statistics.

What is a Population?

In common usage the word population refers to all people, animals, plants, or anything else of interest, that inhabit a particular area. For example, there's the human population of Nigeria, or the population of wildebeest in the Maasai Mara. (Good luck counting them!) In statistics the term **population** refers to the totality of people or objects about which you want to draw conclusions. Before starting a study, you must carefully define your population. For example, if you want to study death rates from a particular disease in hospitals, your population might be the patients in a particular hospital, or all hospitals in the region, or perhaps all hospitals in a country or even the world.

Even when the population of interest is well defined you may not be able to study it. If you want to test the effectiveness of a new supplement for malnourished children, you would ideally like to draw conclusions about how well it works for all malnourished children. Unfortunately, the children whose parents agree to administer the supplement may differ in important ways from the population of all malnourished children. They may be sicker, poorer, or have parents who take better care of them. You can only study the population of children with willing parents and your conclusions are necessarily restricted to them.

What is a Sample?

The persons actually included in your study are called the **sample**. There are many different samples that can be selected from the same population. Think of all the ways you can select 100 school children from a town. You can include the first 100 from an alphabetical listing of student names. You can ask principals to select a certain number of students from their school. Or you can ask children to volunteer to participate. Common sense tells you that there are serious problems with all of these samples. An alphabetical list may results in too many students from a particular ethnic group with similar surnames. Principals may suggest students who they deem exemplary. Students who volunteer may be healthier than those who don't.

The term **statistic** is used to characterize results from a sample. The sample mean and variance are both examples of statistics. The term parameter is used to describe the characteristics of a population. For example, the percent of people using nets in your sample is a statistic, the percent of people using nets in the population is called a **parameter**. Parameters are usually designated with Greek symbols. For example the mean of a population is called called μ (mu), while the mean of a sample is called $\bar{X}$ (X bar). Similarly the standard deviation of the population is called σ (sigma), while the value for a sample is called s. Population values are usually not known but must be estimated from samples. If you knew the population values there would be little reason for you to perform experiments or conduct surveys.

Selecting a Sample

So what is a good sample? Obviously you want it to be selected from the entire population of interest. If you're studying the population of Tanzanian adults, you want your sample to include people of all ages from all areas of Tanzania. You want to make sure that you are not excluding any particular types of people and that all people have the same chance of being selected. (More sophisticated sampling plans allow for unequal probabilities of selection but these probabilities must be known so they can be included in the analysis of the data.)

What you shouldn't do is select people that you think are "typical" of the population. Such samples, called **judgment samples**, are fraught with problems, since the criteria for inclusion depend on someone's notion of who should be selected. There's no way to determine how well the sample really represents the population. Different people will select different samples from the same population based on their preconceived notion of what the population is really like.

Another bad sampling strategy is to select a **convenience sample**; a sample that relies on people who are convenient to include. Interviewers who stand in marketplaces waiting to question people are pursuing convenience samples. Any warm body that is willing to answer questions is fair game for the sample. Drawing statistically valid conclusions from a convenience sample is impossible, since the people included are not representative of any population except those willing to be interviewed at a particular location at a particular time. A **quota sample** attempts to improve a convenience sample by including a certain number of people in predesignated categories. An interviewer is told to get 40 “young” people, 50 “middle aged” people and 10 “old” people. That's still a bad sample since it depends on the whims of the interviewer.

Studies that rely on volunteers, people who offer to be part of your study, are easy to do. That’s why they’re so common. However, it is well known that volunteers can differ in important ways from people who don’t volunteer. People who volunteer their opinions or agree to participate in studies are often very different from those who don’t. They may be poorer or wealthier, healthier or sicker or better educated than people who don't. Don't rely on **volunteer samples**.

Surveys often pop up when you are using the internet. Your bank may ask whether you are satisfied with its services or a book seller may want to know what kind of books you read. Internet surveys combine the worst features of volunteer and convenience samples. You have no idea who is really answering the questions, why, or even how often. There's also the problem that internet availability depends both on demographic characteristics of the possible respondents and regional availability. It's not easy getting online in some places. Beware of drawing conclusions from poorly designed internet-based surveys.

Since all the above methods don't result in a “good” sample, it's natural to ask “How should I select a sample?” The answer may appear strange at first. The mechanism for selecting a sample should be chance. Not haphazard chance, but carefully planned chance. Sounds like a contradiction, but it’s not. Planned chance means that you have a method to ensure that each person in the population of interest has an equal probability of being selected for the sample, and that all possible samples are equally likely. Observations are selected independently so that one person's selection has no effect on the selection of any other person. This is called a **simple random sample**. For example, if you have a list of all student names in a district and want to select a simple random sample of 100 students, you can assign each student a random number which you obtain from a computer program that spits out random numbers. (In the past, statistics textbooks used to include lists of random digits!) Then you can sort the random numbers from smallest to largest, and select the 100 students with the largest or smallest random numbers. When samples are selected by chance, you actually know a lot about them because you can calculate mathematically the likelihood of various types of samples.

Statistical methods don't rely on magic. They rely on your following well-established principles of study design. If you don't, you can still generate all kinds of statistical tests, but the results will be meaningless. While it's possible to reanalyze good data if you've made a mistake in the analysis, a poorly designed study is doomed.

Even if you've carefully selected a random sample, you won't be able to draw correct conclusions about the population if you introduce bias into your survey or experiment by the way you ask questions, record answers, assign individuals to treatment groups, or evaluate responses to treatments. The design of surveys and medical experiments is an extensive topic about which much is written. Let's just take a quick look at some of the issues you should be aware of.

Surveys and experiments, including clinical trials, are the two most common sources of data. In a survey, you ask questions of people. In an experiment, you actually do something and then observe the response. Good design as well as selection of appropriate samples is essential for both of them. Unless you are doing research on mice, even experiments usually require you to ask questions of the study participants. You ask questions such as “Did you sleep under an insecticide treated net last night?” and record the answer, together with additional information about the respondent, such as their age, gender, place of residence, family size, or whatever else you think is important.

Asking Questions

Asking and answering questions is something you’ve done since your toddler days, so you might think it’s not a big deal, but it is critically important in all types of research. The way in which you ask questions affects the answers that you get. You don’t want to influence the response a person gives. Asking a respondent whether they agree with the statement that insecticide treated nets are the most important factor for the control of malaria may elicit different responses than asking “*What do you think is the most important factor for malaria control?”*. You must structure the question in such a way that it does not suggest an answer. The interviewer should always appear nonjudgmental, making sure the respondent doesn’t try to give answers designed to please the interviewer.

Eliciting even simple information requires thought. For example, “How old are you?” is an unimaginative question that you’ve been asked many times. You may recall that you haven’t always answered it truthfully. When you were 12 years old, you may have promoted yourself to the more glamorous teenage years. When you reach middle age, you may want to turn back the clock. Or someday, you simply may not remember the magic number anymore. Does that mean that you shouldn’t ask about age in a survey? Of course not. It means that you have to think about the best, most practical way to obtain reliable information about age. Asking to see a birth certificate is one way of obtaining accurate ages, but it’s obviously not a realistic tactic for most surveys. Asking people to give their birth date as well as their age is a better strategy. People are much less likely to manipulate their birth dates than their ages, since a birth date doesn’t change annually, and it’s work to calculate a birth year to match a desired age.

Although determining age has its problems, they are fairly minor compared to the difficulties encountered in getting information about sensitive topics such as drug use or criminal records. Even obtaining information about routine activities, such as time spent sleeping, poses challenges. Skilled researchers devote a lot of effort to designing the survey questionnaire. To minimize possible problems an actual questionnaire is often pretested on a random sample of people from the population to which a survey will be administered.

Exercise: In the Insecticide-Treated Net study, the experimenters weekly asked people whether they slept under a net the previous night. For each household they chose the same day of the week to ask the question. For example, for a particular household they would always conduct the interview on a Tuesday. Do you see any problems with this strategy?

Missing People and Missing Answers

You may have devised an excellent method for selecting a random sample of people to include in your study. It's unfortunate that not all people whom you want to include will agree to answer all of your questions or participate in your clinical trial. Your first response might be to ignore the uncooperative people and just replace them with more agreeable ones. If people who refuse to be part of your survey or experiment are different from people who participate (and you can't possibly be certain that they aren't), the results you draw based on cooperative people may not be true for the uncooperative people. That's potentially deadly for reaching conclusions about the entire population of interest.

If you encounter people who just refuse to answer certain questions, you can determine how people who answered a question differ from people who don't. For example, if you find that older people are less likely to answer questions about ITN use than younger people and that older people are less likely to use ITNs, you'll have to report results separately for the groups. What you should not do is ignore missing answers, since if they are not missing at random they can seriously affect your conclusions. (Answers are said to be missing at random if cases with missing values are otherwise no different from cases without missing values. It's as if someone went through a data file and randomly deleted some answers.)

Designing Experiments

Unlike a survey, an experiment involves actually doing something to people, animals, or objects. Instead of asking people whether they think that medicines that reduce cholesterol are effective for decreasing the risk of heart disease, you give them the medicine and observe the incidence of heart disease. Sometimes you study the same subject before and after an experimental treatment and determine if there has been a change. Or you might take several groups of people, do something different to each of the groups, and then compare the results.

Experimentation on people and animals poses ethical questions that deserve careful thought. Human experimentation requires informed consent from participants. Risks must be carefully and exhaustively detailed. Institutions have committees that regulate experiments involving humans or animals, and some governments, such as in the US and Europe, take strong measures when institutions do not comply with guidelines for human experimentation.

In experiments as well as surveys, the subjects must come from the population that you're interested in. (This is a lot easier to do for animals than people!) To properly assess the effect of different treatments, you must ensure that the groups receiving the treatments are as similar as possible. The best way to do this is to use chance to assign subjects to the different groups. This doesn't guarantee that the groups will have the same characteristics, but it does minimize the bias that is introduced when treatments are assigned to subjects by the investigator. (A study is termed biased if it favors one result over another. For example, if healthier patients receive the new treatment while sicker patients receive the old treatment, the study is biased in favor of the new treatment.)

Random Assignment to Treatments

Using chance to assign treatments to individuals does not mean that treatments are assigned haphazardly. As in surveys, chance requires a carefully selected systematic approach. If you want to study the effect of personal computer use on grades, you can't let classroom teachers decide which of their students receive personal computers and which do not. They might assign the computers only to the brighter students to reward them for past performance, or they might assign them only to the low-achieving students with the hope that the computers might help them. Any evaluation of the effect of the personal computers would be tainted by the differences between the students.

Random numbers are the preferred method for assigning experimental treatments to subjects. You can't make up your own table of numbers that you think are random; you have to use random numbers generated by computers. Computers don't have birthdays, license plates, children, or any other reason to prefer one number over another. Every number has the same chance of being selected. A simple method for assigning people to experimental conditions is to assign everyone a random number and then put people with even numbers in one group and those with odd numbers in the other. Or you can order the numbers from smallest to largest and assign people with numbers in the lower half to one group and numbers in the upper half to the other. You can use all sorts of systems based on random numbers to assign subjects to groups, even in very complicated experimental designs.

Unless you use a procedure that assigns your subjects to experimental conditions randomly, the results of your study may be difficult or impossible to interpret. Many assignment schemes that appear random to the inexperienced researcher turn out to have hidden flaws. For example, researchers at a hospital compared two treatments for a particular disease. Patients who were admitted on even-numbered days received one treatment, and those admitted on odd-numbered days received the other. Sounds random, but it failed. The number of patients admitted with the disease on even days gradually became larger than the number admitted on odd days. Why? Physicians figured out the scheme and admitted their patients on days when the procedure that they preferred was being used. This introduced bias, since physicians were making the decisions as to which patients received which treatment. That may have resulted in less sick patients being admitted on even days and sicker patients on odd-numbered days, or vice versa.

Exercise: In the 1954 clinical trial of the Salk polio vaccine in the United States, many different study designs were considered. Discuss the advantages and disadvantages of the following designs:

(a) Select a random sample of children and vaccinate them. Compare their polio rate to children in the USA.(b)Vaccinate children whose parents have volunteered them for the study and then compare the polio rate for vaccinated children with the rate for unvaccinated children in the same area.

(c) Vaccinate children in one city and compare their polio rate to children in another city.

Unbiased Evaluation

In experiments, just as in surveys, you must be careful not to let your prejudices influence the results. Some events, such as death, are not disputable. Others, such as "improvement," are not as clear. You must make sure that the endpoints that are being measured are well defined and unambiguous. Don't ask, "Are you better today?" Determine what constitutes better, and ask about the components, for example, freedom from pain, ability to sleep without interruption, performing activities of daily living, and so on.

The best way to make sure that measurements are obtained without bias is to make sure that neither the subject nor the evaluator is aware of the experimental procedure that a person has undergone. In medical studies, when a patient doesn't know which treatment he or she is receiving, the study is called **single blind**. If neither the researcher nor the patient knows, the study is termed **double blind**. People respond favorably to any treatment that they think will help them, even if it's a sugar pill. That is known as the **placebo effect**. If you think you've been given a "magic" pill to help you stay awake in class, you may be more alert than someone who hasn't been given the pill, even if the pill is totally ineffective. That's why it's important to make sure that all people are treated as similarly as possible. If one groups gets a pill, so should the other, even if it contains only sugar.

If you're evaluating a new treatment, you should make sure to include a group that doesn't receive the new treatment, known as a **control group**. For medical studies, the control group might receive the usual treatment for a condition. In studies of new instructional methods, the control group would receive standard instruction. Unless you have a control group that is being observed at the same time and under the same conditions as your experimental group, you will not be able to draw unbiased conclusions about the new method. A "new" treatment may have better survival rates not because it is better, but because patients are being diagnosed earlier today than they were in the past. Similarly, students may perform better in statistics classes today, not because the teachers are better, but because students today are more industrious.

Properties of Samples

Now that you've designed your study, selected a random sample, observed it without bias and objectively recorded all the necessary information, you're ready to reach conclusions about the population from which you selected your sample. On first thought, this might not seem very complicated. Why not assume that what's true for the sample is also true for the population? That would certainly be simple. But would it always be correct? Do you really believe that if 60% of your sample used nets after free distribution, that's exactly what you would see if you distributed nets in all malarial areas, or even in another sample in the same area? Common sense tells you that it's very unlikely that the results you see in a sample are identical to those you would obtain if you made measurements or inquiries of the entire population of interest. A sample is not a miniature of the population. If it was, one quick poll before an election would eliminate the need to even hold elections.

What is true instead is that different samples from the same population give different results, and it's highly unlikely especially for a continuous variable, such as age or weight, that any one sample will hit the population value exactly on the nose. To determine what you can realistically conclude about the population based on results from a sample, let's consider what results are possible when you select a sample from a population. Although we could use mathematical arguments to derive properties of samples, it's more appealing if you just start drawing samples from a population and see what you find.

Taking Samples from a Population

Let's start off with a population that has only two values: 1 if a person reported using a bed net the previous night; 0 if they did not. To construct a population in which half of the people report using a bed net the night before, label a million cards, half of them with "used net" and half of them with "didn't use net." Each card represents a person in your population. Then mix up the cards in a (very large) basket, and randomly select 10 cards, since we want to study the behavior of sample values based on 10 cases. Count the number of cards that say “used net.” If you find 6 cards out of the 10 you selected with “used net” written on them, your sample value for the average percent using a net the night before is 60%. Now you have the result of a single sample from your population. To see how sample results from the same population vary, you have to repeat the procedure over and over, selecting 10 “people” each time and then counting and recording the number of “used net” cards in each sample.

You can think of this activity as sending numerous survey takers into the same population. Each survey taker obtains a random sample of ten individuals, asks whether they slept under a net, and brings back the summarized results to you. Of course when you conduct a real survey that's not what happens. You get back results from the one survey that was funded. However, you know that the outcome of that survey is one of many different outcomes that could have been observed. If you asked everyone in the population if they slept under a bed net the previous night you'd almost certainly get a different result than that you obtained from your single survey. What concerns you is the question “How far off might my survey results be from the true population value?” That's why you're looking at the distribution of possible results.

In case you lack the patience to label a million cards and select 100 samples of 10 people each, when the probability of using a net is one half, I've done it for you. The results are shown in Figure 2.1.

Figure 2.1: Distribution of Results from a 100 samples of size 10 with p=0.5

Number Who Used ITN in Sample	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Total**
Number of Samples	0	0	1	11	23	26	23	11	4	1	0	100
Cumulative Percent	0	0	1	12	35	61	84	95	99	100	100	100

The first row of the figure corresponds to all possible results when you count the number of "use net" cards in a sample of 10 "people." You can find anywhere between 0 and 10 net users in a sample of 10 people. I took 100 samples of size 10 and counted how many net users were in each sample. That's shown in the row labeled *Number of Samples*. There were 26 samples with exactly 5 net users; 1 sample with 9 net users and 11 samples with 3 users. If you add up all of the values in the row labeled *Number of Samples* you'll get 100, the number of samples selected. The last row, *Cumulative Percent*, is the percent of all samples with that many nets or fewer. For example, the value for 4 nets is 35%. That tells you that out of my 100 samples, 35% of them had 4 or fewer net users.

From Figure 2.1 you clearly see that all samples didn't give you the same answer. Only 26% of the samples had exactly 50% net users, which we know is the population value. Some of the samples had values pretty far removed from 50%.

One hundred samples from a population aren't really enough to give us detailed information about the spread of sample values. Let's abandon our cards, and use the computer to draw 10,000 (!) samples from the same population. (This is called a computer simulation.) That should give us a very good idea of what the distribution of possible sample values looks like.

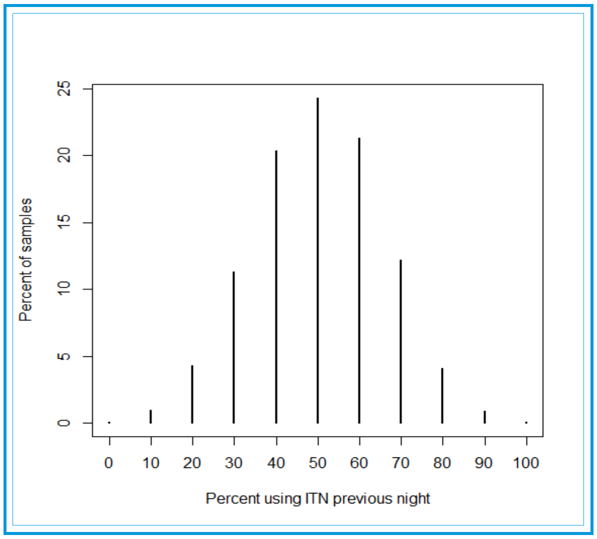
A Computer Model

Figure 2.2 shows the results, when you take ten thousand samples of ten people from a population in which the probability of using a net the previous night is fifty percent. The horizontal axis shows the percent of people in each sample who reported using nets the previous night, the vertical axis tells you the percent of samples with that number of nets. Look at 50% on the horizontal axis. That corresponds to samples of 10 people in which five use nets. Only 24% of all of the samples had net use fractions of exactly 50%. (If this seems a little surprising and you don't want to label a million cards, take a coin, designate one side “net” and flip it ten times, each time recording how many “nets” you observe. Keep selecting 10 “people” until you can't stand it anymore. Count how many times you found exactly five net users in each group of 10. Even though 50% is the probability of a someone using a net, in groups of 10 people you don't often find exactly 5 using nets.)

Fifty percent is, however, the most frequently occurring value. The distribution is symmetric so 50% is also the median and mean. The further away you move from 50% in either direction, the fewer samples you see. Most sample values cluster around the population value. You see that various outcomes are possible, but they are not all equally likely. For example, fractions of 0% and 100% are very unlikely.

From Figure 2.2 you can estimate how far off your sample results may be from the real population value when your sample size is 10 and the true population value is 50%. You see that it's not unlikely that your sample results are between 30% and 70% when the real value is 50%.

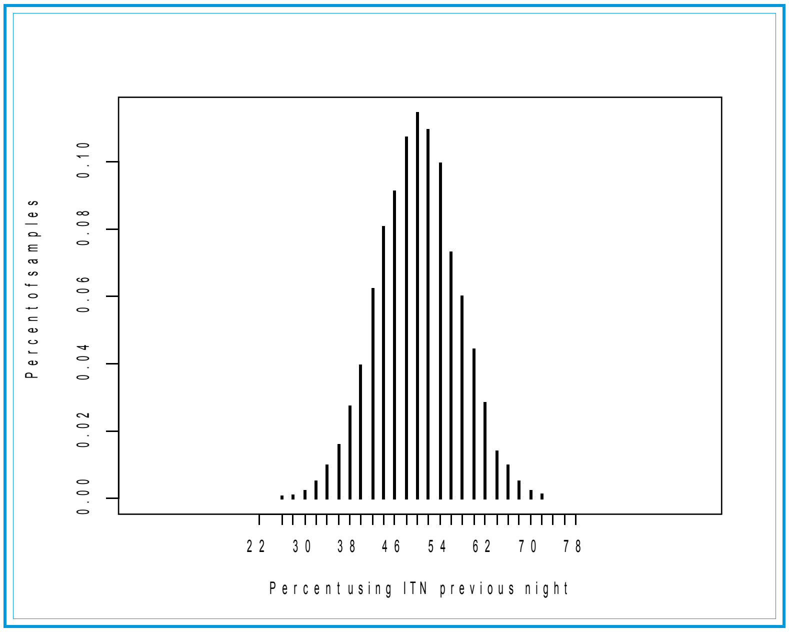
Figure 2.2: Samples of Ten Patients, Probability of Net Use is 50%



Consider what happens when you increase your sample size from 10 people to 50 people in a survey. The population value stays the same—50%. Figure 2.3 shows the distribution of sample results when the computer draws 10,000 samples from the same population as before, but with 50 people in each. The shape of the distributions in similar to that of samples of size 10.

Both distributions have the same mean and median (50%). What's different is that in Figure 2.3 the observed percent with nets has much less spread than it does in Figure 2.2. Sample means of thirty percent or seventy percent aren't unusual when you take samples of 10 people. They are quite unusual when you take samples of size 50. In Figure 2.3 you see very few samples that are far from 50% in either direction. About 95% of the samples means are between 46% and 64%. That's not surprising because it makes sense that large samples are less likely to produce unusual results than small samples.

Figure 2.3: Samples of 50 Patients, Probability of net use is 50%



Think of a coin analogy. If you flip a coin three times it's not unusual that it will come up on the same side each time. If you flip a coin 30 times, it's very unusual that it would come up the same side each time.

By this point you may be asking the age old question “Why should I care? I'm going to take a single sample from the population and draw conclusions about the population from that one sample.” You're correct. However, unless you consider the distribution of all possible sample values, you don't have a clue how close or how far from the true population value your sample value might be. If you understand the behavior of lots of samples from the same population, you can use that information to determine how close to the unknown population value your sample value may be. For example, you can calculate a range of values which you're reasonably confident will include the population value.

Sampling Distribution of a Statistic

The distribution of all possible sample values of a statistic (such as the percentage using a bed net the previous night), calculated from samples of a particular size from a population, is called the **sampling distribution** of the statistic. Figure 2.3 is an estimated sampling distribution of means of samples of size 50 from a distribution that has only two possible values: use a net and not use a net.

A sampling distribution has a mean and standard deviation, just as does the distribution from which you are taking the sample. Figure 2.4 contains summary statistics for the sampling distributions of means for the samples of size 1, 10, 50, 100, 1000. For each sample size, these are the values you get if you take the 10,000 sample means that the computer generated and find the average and standard deviation of the means.

Figure 2.4: Descriptive Statistics from 10,000 Samples of Different Sizes

Sample Size (N)	Mean	Median	Standard Deviation	Minimum	Maximum
1	50%	50%	50.00%	0.00%	100.0%
10	50%	50%	15.70%	0	100.0%
50	50%	50%	7.10%	24.00%	76.0%
100	50%	50%	5.09%	27.00%	69.0%
1000	50%	50%	1.57%	43.90%	56.6%

You see that the mean and median, 50%, are the same for the distribution of the individual values (samples of size 1) and for the means based on 10, 50, 100 and 1000 cases. It's always true that the mean of the sampling distribution of means is the same as the mean of the underlying population. You also see that the standard deviation of means based on 10 people is smaller than the standard deviation for the individual values but larger than the standard deviation based on samples from 50 people. Notice that the standard deviation for 10 cases is 10 times as large as the standard deviation for samples with 1000 cases.

Standard Error of the Mean

The standard deviation of the sample means has a special name. It is called the **standard error of the mean**. The standard error of the mean tells you how much the means from random samples of a particular size vary. The standard error of the mean depends on two factors: your sample size (N) and the standard deviation of the population (σ) from which you are taking the sample:

Standard Error of the Mean= $\sigma/\sqrt{(N)}$

The formula tells you that

- Sample means based on large samples vary less than sample means based on small samples from the same population.
- Sample means from populations with lots of variability vary more than sample means of the same size from populations with less variability

Let's use the formula for the standard error of the mean to calculate it for our example. The population standard deviation is 50%. That's shown in Figure 2.4 in the row labeled Samples of Size 1.

The standard of the mean for samples of 10 is

$$\text{Standard Error of the Mean=}\ \frac{50}{\sqrt{(10)}}\ \text{=15.8\%}$$

For samples of size 50 it is,

$$\text{Standard Error of the Mean=}\ \frac{50}{\sqrt{(50)}}\ \text{=7.07\%}$$

Compare these values to the estimated standard errors shown in Figure 2.4. The standard error of the mean for samples based on 10 observations is estimated as 15.70%. The standard error of the mean for samples based on 50 observations is estimated as 7.10%. The two sets of values are very close. The reason the standard errors differ a little is because in Figure 2.4 they estimated from a large number of samples, while the values based on the formula are mathematically derived exact values. The above formula explains why the standard error based on 10 cases is 10 times as large as the standard error for 1000 cases.

Put away the million labeled cards and unplug the computer. The formula for the standard error tells you what you need to know: how much sample means from the same population vary. You just have to supply the standard deviation and sample size. If you don't know the standard deviation in your population you can estimate it from your sample and use it in the formula. The standard error of the mean won't be exact then, but will be an estimate.

Exercise: In the previous example what would be the standard error of the mean if the sample size was increased to 200? decreased to 5?

The standard error of the mean depends not only on your sample size but also on the variability of the population from which you are sampling, the population standard deviation. If everyone in the population, or no one, uses nets, all samples from the population will give the same result. As the variability in the population increases so does the variability of the means of samples of a particular size. If you're studying the heights of two year old children you know they will have less variability, a smaller standard deviation, than the heights of adult men. If you sample 50 children you expect that the standard error of the mean will be smaller than if you sample 50 adults.

Exercise: Calculate the standard error of the mean for samples of size 10 and 50 when the population value for the percent using a net is 70%. Why are the values smaller than those you calculated for the same sample sizes when the population value was 50%?

Plotting Error Bars

In reading the medical literature you will often encounter figures with bracketed lines emanating from the end of bars or points. Usually the line is drawn so that it extends to one standard error above the end and one standard error below the end. Look at Figure 2.5 from (Vanden Eng et al., 2010). The white bar for Niger shows that about 65% of households own ITNs. The line that extends from the top of the bar goes to about 72%. The distance from 72% to 65% , 7%, is the standard error for the proportion of households owning a net. (Look carefully at the footnotes when you see error bars since sometimes confidence intervals are plotted instead of standard errors.)

Figure 2.5: Bar chart with Standard Errors: ITN Ownership and Use

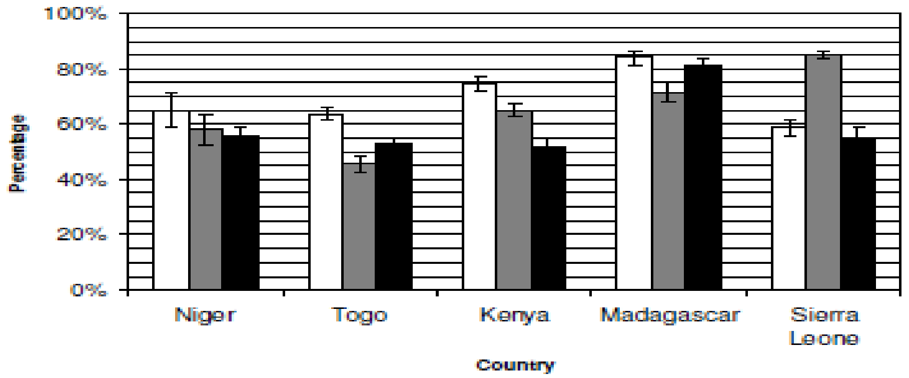


Figure 2 Indicators of ITN ownership and use: survey results following integrated campaigns targeting children <5 years of age in Niger, Togo, Kenya, Madagascar, and Sierra Leone. Bars represent: white: percentage of households that own an ITN (denominator: all households); dark: percentage of ITNs hanging (denominator: all ITNs); and black: percentage of all children less than 5 years of age sleeping under an ITN (denominator: all children less than 5 years of age). Note: household ownership and ITN use by children less than 5 years of age are standard RBM outcome indicators.

Knowing the standard error is important for drawing conclusions about the true rates in the population. The researchers did not include all households in their study instead they selected a sample from the population of households. The sample is one of many that could have been selected. If the standard error is small, the sample mean stands a better chance of being close to the population value than if the standard error is large.

The standard errors for the bars in Figure 2.5 are not all the same. There are two reasons for this: the numbers of households sampled differed for the countries, and the variability of the responses differed within the countries. The standard errors are smallest for the percentage of households owning nets in Madagascar and for the percentage of ITNs hanging in Sierra Leone. For both of these, the observed rates are close to 80%. If percentages are large (or small), you know that households are quite similar. There's not much variability. Most households in Madagascar own nets, while most households hang nets in Sierra Leone.

Large scale surveys must necessarily use more complicated sampling methods than selecting simple random samples from the population. For example, they may randomly select regions, then communities within a regions, and then households within a community. Analysis of data from complex surveys requires additional steps for calculating the actual standard errors.

Exercise: Write a short paragraph summarizing the results shown in Figure 5. Does there seem to be a relationship between the percent of households that own nets and the percent of these nets that are actually used? Do you think that when nets are difficult to obtain, those who own them are more likely to use them more?

Evaluating a Claim

You've demonstrated that sample means from random samples from the same population vary and that you can determine how much you expect them to vary for different sample sizes and different variability in the population. Let's see how you can actually use this information to evaluate a claim. In statistical terms what you're doing is testing a hypothesis.

A Distinguished Scientist claims that she has a better treatment for a disease of interest. Of 10 patients who received her new treatment, 70% were cured. Extensive literature on the topic indicates that, worldwide, only 50% of patients with this disease are cured. Based on the results of her experiment, can you tell if the physician has really made inroads into the treatment of this disease?

Are the Observed Results Unlikely?

To evaluate the scientist's claim, you have to ask yourself the question, are the results she observed (7 out of 10 cures) unlikely if the true population cure rate is 50%? You know that if half of all people with a disease are cured by the standard treatment, that doesn't mean that any time you treat 10 patients, exactly 5 will be cured.

Look at Figure 2.2 again. How often would you expect to see a "cure" rate of 70% or more in a sample of 10 patients, if the real cure rate in the population is 50%? Adding up the percentages shown on the vertical axis, you see that about 20% of the time you expect to see cure rates of 70% or more when the true population value is 50%. If the new treatment is no better than the standard, you would expect to see cure rates at least as large as those observed by the physician almost 1 out of 5 times you repeat the experiment. (In fact, it is possible to calculate mathematically that when the true cure rate is 50%, the probability of obtaining 7 or more cures in a sample of 10 is close to 17%.)

Of course, it's always possible that the new treatment is really less effective than the usual treatment. So if you want to test the hypothesis that the new treatment is not different from the standard treatment, you must evaluate the probability of results as extreme as the one observed in either direction—increasing or decreasing the cure rate. You can estimate from Figure 2.2 that the probability of 30% or fewer cures and the probability of 70% or more cures is about 40%. Based on these results, you have little reason to believe that the Distinguished Scientist is really onto something. Her results are certainly not incompatible with samples selected from a population in which the true cure rate is 50%.

The Effect of Sample Size

As you saw above, when the true cure rate is 50%, there's a good chance that you will observe from 3 to 7 "cured" patients in a sample of 10 by chance alone. In fact, most of the outcomes that can occur in a sample of 10 would not be considered unusual even if the new cure rate is quite a bit larger or smaller. Based on a sample of only 10 patients, it's very difficult to evaluate a new treatment. Another way of saying this is that a sample of size 10 has very little ability to identify true differences. It has little power. (In statistics, **power** means the probability of detecting a true difference when it exists.) Larger samples improve your chances of detecting a difference in the cure rates (if, in fact, there is one) because there is less variability in the possible outcomes.

You may wonder if you can ever tell from a sample of just 10 patients that a new treatment is better. The answer is yes. If you've observed 10 cures of a previously incurable disease, you have enough evidence to believe that the treatment is effective. Even a small number of successes is enough to give you pause.

What if the Distinguished Scientist tells you that she achieved a 70% cure rate based on 50 patients? Would you be more likely to believe that she's onto something? From Figure 2.3 you see that values greater than 70% or less than 30%, when the population value is 50%, are now noticeably less likely than they are for samples of 10 patients. Rates that were not particularly unusual when you had samples of 10 patients are quite unusual with 50 patients. Based on Figure 2.3, you see that your chance of finding a sample rate of 70% or more (or 30% or less) when the true rate is 50%, is very small. If you calculate the exact probability mathematically it is 0.66%. That means that fewer than 7 times in 1000 would a cure rate as extreme as the one observed (7-%) occur, in a sample of size 50, if the new treatment doesn’t differ from the standard treatment.

The usual rule of thumb used for “unusual” is a probability of 5% or less. That is, if results as extreme or more extreme than the ones you observe are expected to occur by chance alone in 5 (or fewer) samples out of 100, the results are considered unusual, or “statistically significant.” The probability of observing results as extreme as you observed when there is no effect is termed the **observed significance level** or the **p-value**.

Exercise: You saw that the observed result of 35 cures in a group of 50 patients was unusual if the true rate of cure is 0.50%. Do you believe that the physician has a better treatment? What questions would you ask her?

The Binomial Test

You can use a computer program to calculate a binomial test that compares an observed proportion or rate to a known standard or usual rate. To use the binomial test, your experiment or study must have only two possible outcomes, such as cured/not cured, pass/fail, buy/not buy, defective/not defective, and so on. All of the observations must be independent, and the probability of success must be the same for each member of the sample population. Observations are independent if one's subject's responses can't influence those of another. If students collaborate on an exam, their scores are not independent. If you cure the same patient with your treatment 10 times, the observations are not independent since they are coming from the same patient.

You specify the number of “successes” you observed in your sample and the value that you assume is true for the population. The program calculate the probability that you observe at least as many successes as you did, if the population value is correct. Try it on the Distinguished Scientist data.

Figure 2.6 contains output from a software package called Graph Pad. It's free and you do not have to download it. If you're at a computer click on the following link

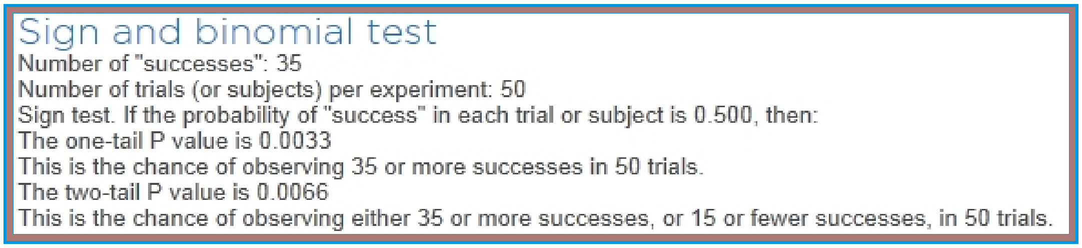
http://graphpad.com/quickcalcs/catMenu/

- Click on binomial and sign test and then click continue.
- Enter the number of “successes” you observed . In this example 35.
- Enter the number of trials (sample size). In this example 50.
- Enter the population value for the probability of success, which is 0.50.
- Click the Calculate Probabilities button.

The program will return the probability of observing at least the number of successes you observed. It will also calculate the probability of results as extreme as the ones you observed in either direction.

From Figure 2.6 you see that the probability of observing a result as extreme as 35 successes in 50 patients (trials) when the probability of success is 0.50, is 0.0033. The probability of observing a result as extreme in the other direction (15 or fewer successes) is also 0.0033. The probability of observing either 35 or more successes or 15 or fewer successes in 50 patients is 0.0066. It's pretty clear that it's unlikely you'll observe 35 cures in 50 patients if the probability of a cure is really 50%.

Figure 2.6: Graph Pad Test for the Binomial Distribution



Sign and binomial test

Number of "successes": 35
Number of trials (or subjects) per experiment: 50
Sign test. If the probability of "success" in each trial or subject is 0.500, then:
The one-tail P value is 0.0033
This is the chance of observing 35 or more successes in 50 trials.
The two-tail P value is 0.0066
This is the chance of observing either 35 or more successes, or 15 or fewer successes, in 50 trials.

Exercise: Using Graph Pad calculate the probability that in a sample of 50 people you observe 20 or fewer people using bed nets, when the population value is 60%. What's the probability that you observe 40 or more people using bed nets?

Summary

In this chapter, you saw that different samples from the same population give different results. The sampling distribution of a statistic tells you, for a particular sample size, the distribution of all possible sample values of the statistic. You use the sampling distribution of a statistic to determine how likely or unlikely various sample results are under particular circumstances.

The standard error of a statistic tells you how much results vary from sample to sample, taken from the same distribution. As the sample size increases, the variability of statistics calculated from the same population decreases. From the sampling distribution of a statistic you can calculate the probability of observing sample results that are at least as large as the one observed when there is no difference in the population.

In subsequent chapters, you'll learn more about testing hypotheses about a population, based on results observed in a sample. You'll also learn about the importance of the normal distribution in hypothesis testing.

Bibliography:

Vanden Eng et al. (2010) Assessing bed net use and non-use after long-lasting insecticidal net distribution: a simple framework to guide programmatic strategies. Malaria Journal 2010: 9:133. VandenEng (2010)

3 The Normal Distribution

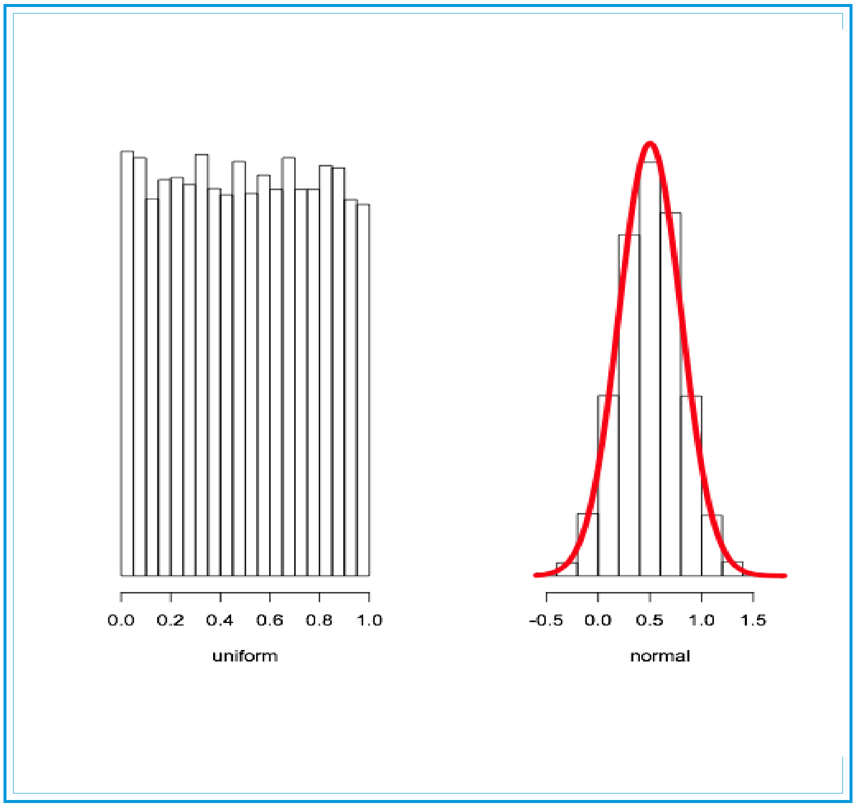
- What does a normal distribution look like?
- Why is it important in statistics?
- What is a standard normal distribution?
- What is a confidence interval?

Every variable and statistic has a probability distribution which tells you the likelihood of various values. The variable “sleeping under a net last night” has only two values: *yes* and *no* (if you exclude *can't remember*). The probability associated with answering *yes* completely determines the distribution. (The probability of *no* is just 1-probability of *yes*.) A variable like diastolic blood pressure has a lot of possible values and they are not all equally likely. You can't completely describe the distribution of diastolic blood pressure with just a mean and standard deviation.

Distributions can have the same mean and standard deviation and yet have very different shapes. Both of the histograms in Figure 3.1 come from distributions with a mean of 0.5 and a standard deviation of 0.3. The sample on the left comes from a uniform distribution, a distribution in which all values are equally likely. The second sample is from what is called a normal distribution.

The **normal distribution** is mathematically defined. There is a formula that exactly dictates what percent of the values fall into each interval based on the distance from the mean. The red curve in Figure 3.1 is the exact normal distribution superimposed on the histogram of the sample. Since samples aren't miniatures of the population, samples from a normal distribution are not exactly normal, especially if the sample size is small.

Figure 3.1: Two Distributions with Same Mean and Variance



The **normal distribution** is bell shaped. Most of the observations are bunched in the center. The mean, median and mode are exactly in the middle of the distribution, and are equal to one another. The farther you move from the center, in either direction, the fewer the number of observations. The distribution is symmetric. If you draw a line through the center and fold the distribution, the two sides are the same.

The normal distribution is very important in statistical analysis since many variables, such as height, weight, and blood pressure have distributions which are approximately normal. The normal distribution also serves as a reference point for characterizing data.

Figure 3.2 is a histogram of gestational age in weeks at which women enrolled in a malaria study delivered, as determined by ultrasound (Wylie et al. (2013)). Women with malaria deliver low birth weight infants so it is important to establish whether this is the result of prematurity or failure to grow, since the treatment differs for the two situations. The distribution of weeks of gestation looks approximately normal but is skewed, with more low gestational ages than you would expect if the data were a sample from a normal distribution.

Figure 3.2: Weeks Gestation at Delivery by Ultrasound

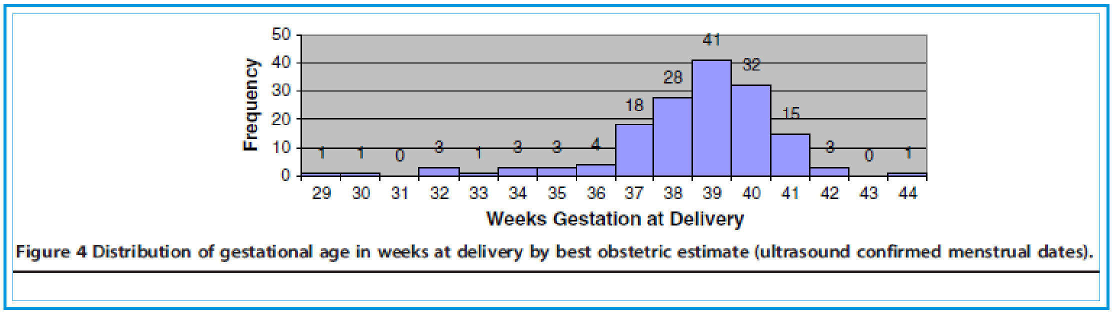
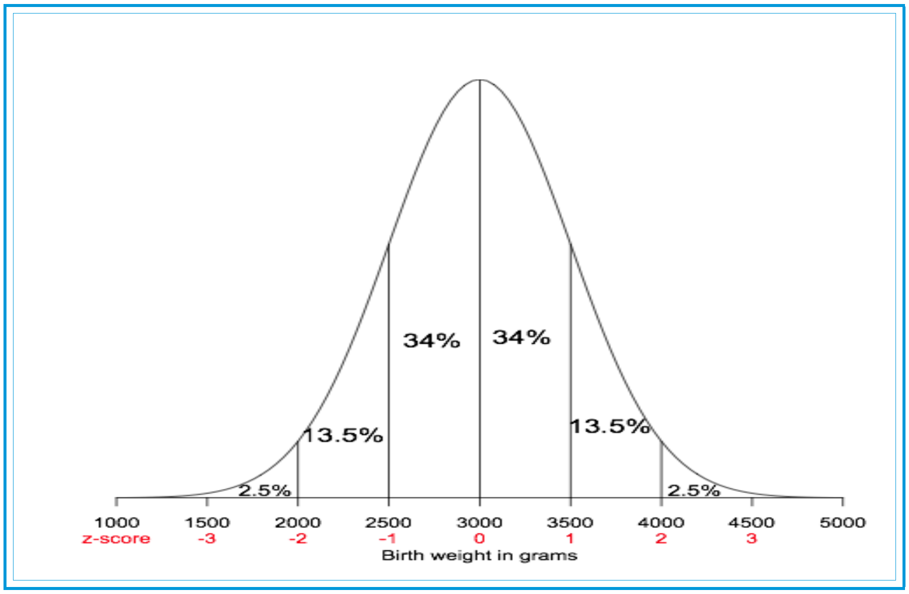


Figure 4 Distribution of gestational age in weeks at delivery by best obstetric estimate (ultrasound confirmed menstrual dates).

Finding Your Way Inside a Normal Distribution

If a variable has exactly a normal distribution and you know the population values for its mean and standard deviation, you know everything there is to know about the likelihood of various values. Consider Figure 3.3 which is a theoretical normal distribution of birth weights of full term newborns. The mean is 3000 grams, and the standard deviation is 500 grams.

Figure 3.3: Area Under Normal Curve



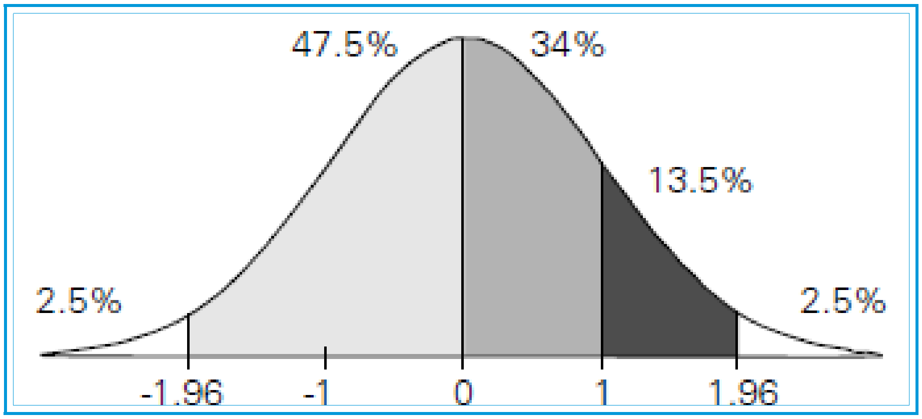
The percentages in the intervals tell you what percent of all newborns have weights within that interval. Thirty four percent of newborns weight between 2500 and 3000 grams and thirty four percent weigh between 3000 and 3500. These two percentages are equal because the normal distribution is symmetric and the two intervals are the same distance from the mean but on opposite sides. Since the standard deviation of the birth weights is 500, the value 2500 is one standard deviation below the mean and the value 3500 is one standard deviation above. In a normal distribution, 68% of all observations are within one standard deviation of the mean, and 95% are within two standard deviations. Only five percent of all observations are more than two standard deviations (1.96 standard deviations, to be more exact) from the mean.

The Standard Normal Distribution

Since a normal distribution can have any mean and standard deviation, the location of a case within the distribution is usually given by the number of standard deviations it is above or below the mean. This is known as a standard score or z-score and was discussed in more detail in Chapter 1. The standard score axis is labeled z-score in Figure 3.3.

A normal distribution in which all values are expressed as standard scores is called a **standard normal distribution**. It has a mean of 0 and a standard deviation of 1, as shown in Figure 3.4. Note that 2.5% of observations have standard scores greater than 1.96 and 2.5% less than -1.96.

Figure 3.4: Standard Normal Distribution

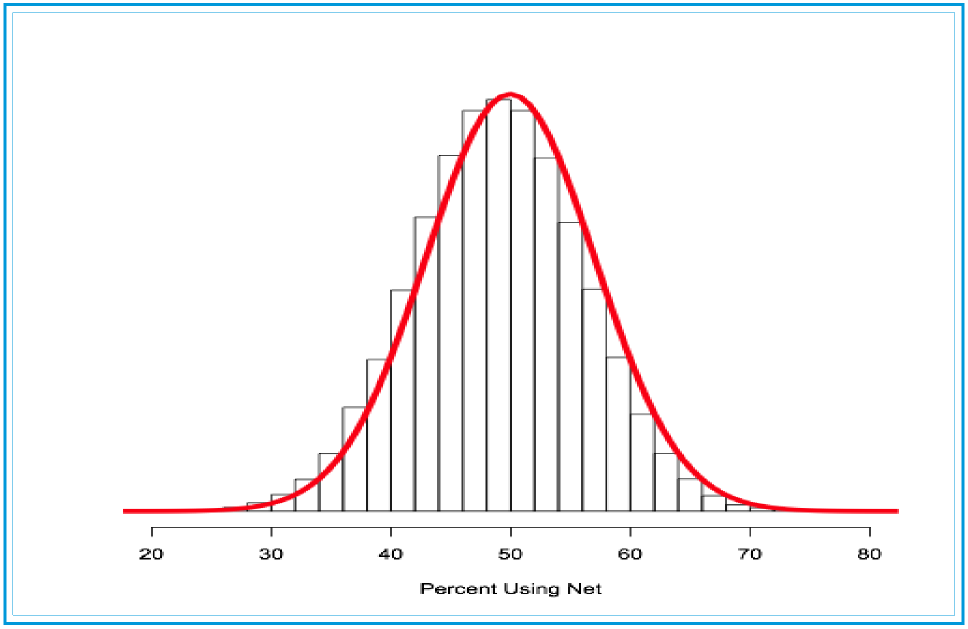


Exercise: Estimate the percent of babies with birth weights greater than 2500 g, Less than 3000 g, Greater than 4000 g, Greater than 3500 g, Between 3000 and 4000 g.

Distributions of Sample Means

In the previous chapter you saw that statistics calculated from random samples, such as sample means, also have a distribution. Some sample means are really close to the population value, some are farther away. Figure 3.5 Shows a sampling distribution of means when the population value for sleeping under an ITN is 50% and the sample size is 50. From the distribution you can tell how likely the various sample values for the mean are. There's a red curve drawn on the estimated sampling distribution. It is a normal distribution with the same mean (50%) and standard deviation (7.1%) as the distribution of all possible sample means. The normal distribution fits the observed distribution of sample means very well.

Figure 3.5: Distributions of Sample Means of 50 Cases from Binomial with p=0.5

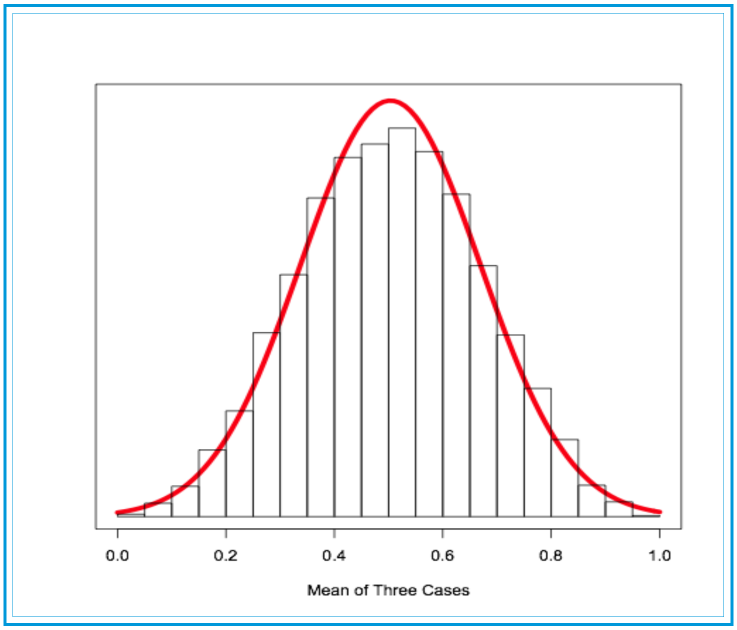


That's amazing! You start off with a variable that has only two values: *use a net*, or *don't use a net*. For a sample you calculate the percent using a net. When you look at the distribution of all possible sample values for the percent using a net, you see a normal distribution. Not only are many variables approximately normally distributed but, in many situations, so are their sample means. This remarkable finding is explained by the **Central Limit Theorem** which says that for samples of a sufficiently large size, the distribution of sample means is approximately normal. The original variable can have any kind of distribution. It doesn't have to be bell shaped at all.

How large a sample you need before the distribution of sample means is approximately normal depends on the distribution of the original values of a variable. For a variable that has a distribution not too far different from normal, sample means have a normal distribution, even if they're based on small sample sizes. If the distribution of the variable is very far from normal, larger samples sizes will be needed for the distribution of sample means to be normal. The important point is that the distribution of means always gets closer and closer to normal as the sample size gets larger and larger—regardless of what the distribution of the original variable looks like. That's why the normal distribution is very important in statistics. Not all variables are normally distributed, but for sufficiently large sample sizes, their sample means are normally distributed.

As an example, look back at Figure 3.1, which shows the uniform distribution in which all values are equally likely. Figure 3.6 shows what happens when you calculate sample means based on three observations from that distribution. Notice how the distribution is no longer flat but now has a the shape of the normal distribution. The distribution of individual values is flat, but thanks to the Central Limit Theorem the distribution of means is normal. The reason that statistician find it so exciting that sample means have a normal distribution is that the properties of the normal distribution can be used for calculating confidence intervals and testing hypotheses about population means.

Figure 3.6: Means of samples of three observations from uniform distribution

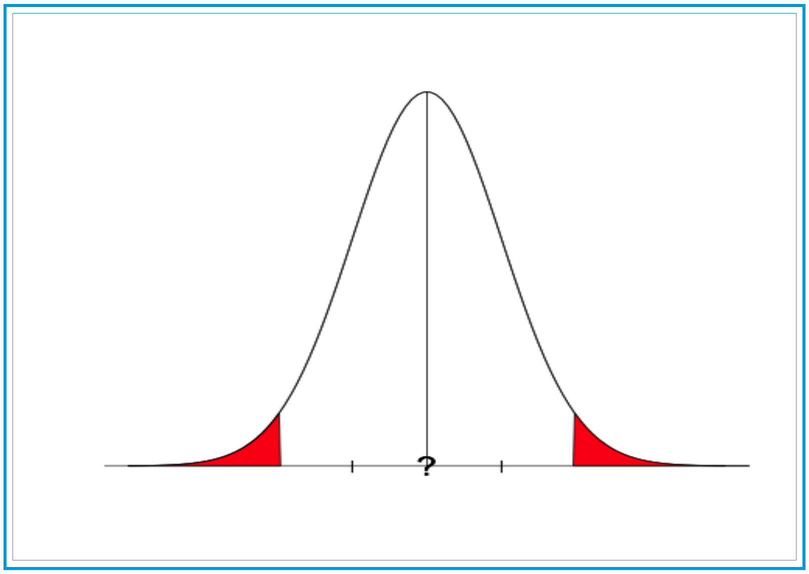


Confidence Intervals

When you conduct a survey or experiment you are interested in drawing conclusions about a population value based on results from your sample. You don't know the value of the mean for the entire population. If you did, you wouldn't be doing the survey or experiment. You do know that your sample mean is one of many possible means from the same population. It may be a lot bigger or smaller than the population value or it may be pretty close.

Look at Figure 3.7, a normal distribution of all possible sample means. The unknown population value is in the center, labeled "**?**". Since Figure 3.7 is a distribution of means, the standard deviation of the distribution is the standard error of the mean. From the properties of the normal distribution you know that 95% of sample means are within two standard errors of the population mean. The red region in the tails of Figure 3.7 is the area outside of two standard error units.

Figure 3.7: Sampling Distribution of Means



All you have to do is figure out where your sample mean is in this distribution. That's the problem. Since you don't know the population mean (the **?**) you can't place your sample mean. It might be a standard error above the population mean or may be half a standard error below the population mean. It can be anywhere in this distribution.

In Figure 3.8 the sample mean (Xbar) is one half standard error above the unknown population mean. In Figure 3.9 the sample mean is one standard error below. In Figure 3.10, the sample mean is two and a half standard deviations above the mean, in the red region of Figure 3.7.

Each of the sample means has a line drawn which extends two standard error units above it and two standard errors below it. Both of the first two intervals "trap" the unknown population value. It's only if your sample mean is in the unlucky red region of Figure 3.7 that an interval that extends two standard error units above and below the sample mean will not include the population value. That's shown in Figure 3.10. The interval around the sample mean does not trap the population value.

Figure 3.8: Sample Mean One Half Standard Deviation Above Population Mean

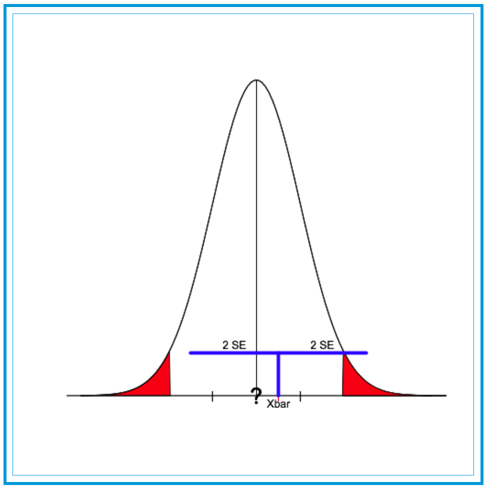


Figure 3.9: Sample Mean One Standard Deviation Below Population Mean

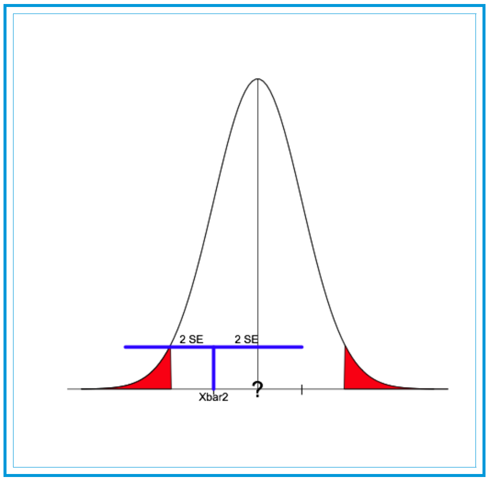
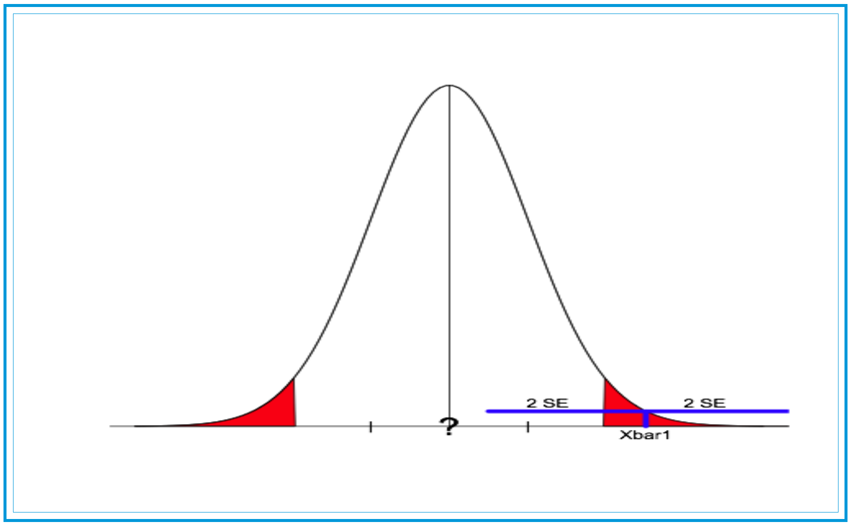


Figure 3.10: Sample Mean Two and a Half Standard Deviations Above Population Mean



Can you tell if your sample mean is in the dangerous red region? Of course not. But you do know that only five percent of the sample means fall into the red region, the area that corresponds to a standard score greater than 2 standard error units from the population mean. All you can do is calculate an interval around your sample mean and hope that your interval is one of the 95 out of a 100 that includes the population value. The interval around the sample mean that includes the population mean 95% of the times is called a 95% Confidence Interval for the Mean. (You can compute confidence intervals for other statistics besides an individual mean, for example you can compute a confidence interval for the difference of two means.)

Calculating a Confidence Interval

To compute a 95% confidence interval for the population mean:

- Calculate the standard error by dividing the standard deviation by the square root of the sample size
- Multiply the standard error by 2 (or 1.96 if you're fussy)
- Subtract twice the standard error from the observed sample mean to get the lower limit.
- Add twice the standard error to the observed sample mean to get the upper limit

For example, suppose you take a sample of 25 women with HIV and find the birth weights of their full term infants to be 2750 grams. If the standard deviation is 500 grams,

- The standard error of the mean is 100 grams $(\frac{500}{\sqrt{(25)}})$
- Twice the standard error is 200 grams
- The lower limit of the 95% confidence interval is 2550 grams

  (2750-2(100))
- The upper limit of the 95% confidence interval is 2950 grams

  (2750+2(100))

Does this 95% confidence interval trap the population mean birth weight for infants of HIV mothers? You don't know. The population value is a fixed number. It's what you get if you include the entire population of HIV babies in your sample. The population value either falls into your confidence interval or it doesn't. You don't know which is the case. All you know is that 95% of 95% confidence intervals include the population value. It's not strictly correct to say that the probability that a particular interval includes the population value is 95%. Instead you can say that you are 95% *confident* that the interval contains the unknown population mean. The following analogy may help. Before a baby is conceived the probability of a girl is roughly a half. Once the baby is conceived, it's either a boy or a girl. The same is true with a confidence interval once it's calculated. It either traps or does not trap the population value.

You can calculate confidence intervals that have higher than a 95% likelihood of trapping the unknown population. For example a 99% confidence interval includes the population value 99% of the time. The price you pay for a higher confidence level is that the interval becomes wider. You can make a confidence interval narrower, for a fixed level of confidence, by increasing the sample size since that will make the standard error smaller.

For small sample sizes, when the population standard deviation is not known, instead of multiplying by 2, a cutoff which is based on the normal distribution, you use values from what's called the *t*-distribution which is described in Chapter 5. The *t*-distribution depends on the sample size and takes into account the fact that, by using the sample standard deviation instead of a know population sample deviation, you're introducing additional uncertainty. The *t*-distribution looks like a normal distribution but it has more area in the tails. There's no need to get absorbed in minor differences in computing confidence intervals, since you'll no doubt use statistical software for the actual calculations. What matters is that you understand what you can conclude based on a confidence interval.

Exercise: This is an excerpt from the Results section of Vanden Eng (2010)

> The percent of households owning an ITN ranged from 58.6% (95% CI: 55.9-61.3) in Sierra Leone to 83.9% (95% CI: 81.3-86.5) in Madagascar. The percent of ITNs suspended over a sleeping space the previous night varied greatly among the countries. Although Sierra Leone had one of the lowest levels of ownership and use, it had the highest hanging percentage at 85.0% (95% CI: 83.5-86.5). Niger and Kenya had hanging percentages that lie between ownership and usage (58.0%, 95% CI: 52.6-63.5 and 65.1%, 95% CI: 62.8-67.4 respectively), whereas Togo (45.7%, 95% CI: 42.9-48.5) and Madagascar (71.5%, 95% CI: 67.9-75.1) had a smaller percentage of ITNs hanging than either ownership or use.

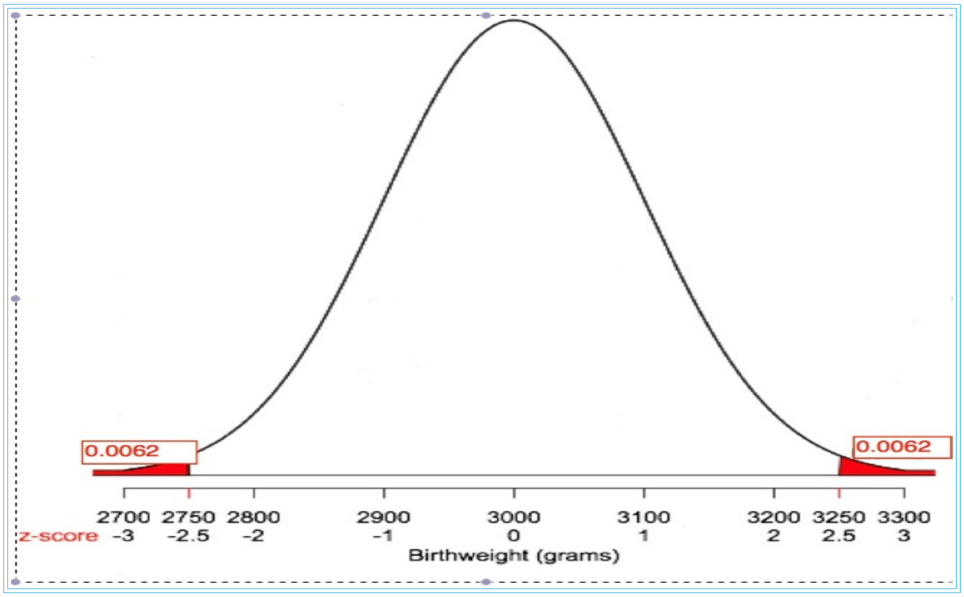
What is the 95% confidence interval for the percent of households owning an ITN in Sierra Leone? Does this confidence interval include the population value for percent of households owning an ITN in Sierra Leone? What would happen to the confidence interval if everything stayed the same but the sample size was increased? Decreased? Based on their confidence intervals, do you think that Sierra Leone are Madagascar are really different in the percent owning an ITN? Do their confidence intervals overlap?

Testing Hypotheses about Population Means

Since for sufficiently large sample sizes, the distribution of sample means is normal, you can use this information to test hypotheses about population means. Let's say you are interested in birth weights of infants born to HIV positive mothers. You want to know whether their average birth weight differs from the known population value. You know that in the population the average is 3000 grams and the standard deviation is 500 grams. Based on a sample of 25 newborns, you find that the average weight is 2750 grams. What can you conclude? Is the sample mean unusual if the true value is 3000 grams?

Look at Figure 3.11 which shows the distribution of sample means when the population average is 3000 grams and the standard error is 100 grams.You know that 95% of the sample means are between 2800 g. and 3200 g. (mean +/-2 SE), 68% within 2900 and 3100 grams (mean +/- 1 SE).

Figure 3.11: How Unlikely is a Sample Mean of 2750 grams?



Is your observed mean of 2750 grams unlikely if the population mean is 3000? To answer that question you need two steps:

- Calculate the z-score for a birth weight of 2750, so you can see where it falls within the distribution of sample means when the population mean is 3000 grams,

$$\text{z-score}=\frac{(\textit{sample mean}-\textit{population mean})}{(\textit{standard error of the mean})}=\frac{(2750-3000)}{100}=-2.5 \quad .$$

That's marked for you in Figure 3.11.

- Calculate the probability of observing a z-score of at least 2.5 in absolute value. From Figure 3.11 that's 0.0124, the solid red area in the normal curve. That's the **observed significance level**, often referred to as the *p*-value. (The next chapter talks more about why you are looking at differences in both directions.)

The observed significance level is smaller than the usual cutoff of 0.05 for "unusual", so you can conclude that it appears unlikely that babies born to women with HIV have the same average birth weight as the general population. Are you certain? Of course not. It's just an unusual result for a sample from a population with a mean of 3000 grams.

Previously, you found the 95% confidence interval for the population mean weight of birth weights of infants with HIV positive mothers to be from 2550 to 2950 grams. Notice that the value of 3000 grams is not included in the interval. That's because it's unlikely that the population value for the birth weight of the HIV babies is 3000. Any value not included in the 95% confidence interval is unlikely, using the 0.05 criterion for unlikely. In the next chapter you will learn in more detail about the mechanics and perils of hypothesis testing.

Summary

The normal distribution is important in statistics because many variables such as cholesterol, weight and blood pressure have distributions which are approximately normal and because for sufficiently large sample sizes the distribution of sample means is approximately normal. Although normal distributions can have different means and standard deviations, the proportional distribution of cases about the mean is the same. A standard normal distribution has mean of 0 and a standard deviation of 1. A confidence interval for a statistic is a range of values, based on a sample, that with a designated likelihood include the unknown population value.

Bibliography:

Wylie et al.: Gestational age assessment in malaria pregnancy cohorts: a prospective ultrasound demonstration project in Malawi. *Malaria Journal* 2013 12:183.Wylie(2013)

Vanden Eng et al. Assessing bed net use and non-use after long-lasting insecticidal net distribution: a simple framework to guide programmatic strategies. *Malaria Journal* 2010,9:133.VandenEng (2010)

4 Basics of Testing Hypotheses

- What is a null hypothesis? An alternative hypothesis?
- What are statistical tests?
- On what basis do you decide to reject the null hypothesis?
- What mistakes can you make when testing a hypothesis?
- Why is sample size important when testing hypotheses?

Read the abstract below (Ezeaka et al. (2009)). It's very similar to many published abstracts. The authors describe their observed results, draw conclusions and support them with mysterious words, symbols, and numbers, usually protected by parentheses. In this chapter you'll learn how to interpret what's hiding in the parentheses. In particular you'll focus on the p-values. The next chapter explains the cryptic letters and numbers that directly precede the **P=**.

> Numerous studies have reported that HIV-infected pregnant women are at increased risk of delivery of low birth weight (LBW) infants, of preterm deliveries and of intrauterine growth restriction. The objective of the study was to determine the effect of maternal HIV infection on the anthropometric characteristics of the babies at birth. A prospective study was carried out at the Lagos University Teaching Hospital, Nigeria. There were three times more LBW babies in the HIV-positive group than in the uninfected mothers (odds ratio = 3.47, 95% confidence interval = 1.69, 7.27; chi(2) = 12.99, **P = 0.0003**).The maternal weight (t = 15.85; **P = 0.0001**), maternal body mass index (BMI) (t = 15.07; **P = 0.0003**), birth weight of infants (t = 27.17; **P = 0.0001**) and birth length (t = 31.20; **P = 0.001)** were significantly less in HIV-positive mothers than in controls. In conclusion, poor maternal bodyweight and low BMI are significant contributors to LBW in HIV-infected women. Nutritional counselling, dietary intake and weight monitoring during pregnancy should be emphasized to improve pregnancy outcome in HIV-infected women.

Steps in Testing a Hypothesis

When you conduct a survey or perform an experiment, you're usually interested in answering questions about populations. The questions may be relatively straightforward, such as whether a new treatment is better than the standard or whether the ready availability of Insecticide Treated Nets increases their use. Or the questions may be considerably more complicated: what family and societal characteristics are predictors of whether a girl undergoes Female Genital Mutilation/Cutting (FGMC) or what factors predict neonatal mortality or morbidity for infants of HIV positive women? The Ezeaka abstract answers questions about anthropometric measurements of infants born to HIV-infected women and infants born to women who are HIV free.

Many different statistical tests are used to answer such questions, although the answers are usually not as definitive as you would like, since you want to draw conclusions about the population, not the sample. That introduces uncertainty. You've tested hypotheses in previous chapters. You considered whether the Distinguished Scientist had a better cure than the standard and whether newborns of HIV positive women come from a population with a weight of 3000 grams. In this chapter we'll discuss in more detail the steps involved in testing a hypothesis about the population.

Step 1: Specify the Null Hypothesis and the Alternative Hypothesis

All statistical tests require you to frame your question as a hypothesis. A **hypothesis** is just a factual statement that may or may not be true. The new treatment is not better than the standard is a hypothesis. ITN distribution increases the probability of using a net is also a hypothesis. So is maternal education decreases the likelihood of FGMC.

To use statistical tests you have to reformulate your question in terms of two related hypotheses: a null hypothesis and an alternative hypothesis. Together the null hypothesis and the alternative hypothesis cover all possible situations. The **null hypothesis** is sometimes called the “no difference” hypothesis. Usually it's what you are setting out to disprove. For example, if you've developed a new vaccine, you're interested in demonstrating that it's better than the standard. Your null hypothesis is that the new vaccine and the old vaccine are equally effective. Similarly the null hypothesis is that ITN distribution has no effect on ITN use and maternal education is not related to FGMC. The null hypothesis is stated in such a way that it can serve as a frame of reference for evaluating the observed results.The null hypothesis is assumed to be true until there's sufficient evidence to question it. In the previous chapter you tested the null hypothesis that infants born to mothers with HIV weigh the same as infants whose mothers don't have HIV.

The **alternative hypothesis** describes the state of affairs if the null hypothesis is false. The alternative hypothesis is that the new vaccine is different from the standard, that ITN distribution affects net use, that maternal education is related to FGMC, and that infants born to mothers with HIV don't weigh the same as infants whose mothers don't have HIV.

Exercise: For the Ezeaka abstract , identify all of the pairs of null and alternative hypotheses. For example, null hypothesis: low birth weight is equally likely for the HIV and control groups, alternative hypothesis: low birth weight is not equally likely for the two groups.

Step 2: Select the Appropriate Statistical Procedure

To test a hypothesis you have to select an appropriate statistical procedure which estimates whether your observed results are unusual if the null hypothesis is true. The statistical procedure depends both on the question you are posing and the characteristics of your data. There are numerous procedures for testing hypotheses about means, for quantifying the associations between variables and for building complex models that predict the values of a dependent variable from a set of independent variables.

A variable is termed **dependent** if its values are thought to depend on the values of other variables, called **predictor or independent** variables. For example, FGM/C may depend on family socioeconomic status, education and region. It's the dependent variable. Factors that are associated with it are the predictor variables. (Chapter 5 is an overview of the some of the most commonly used statistical procedures for testing hypotheses).

Don't be discouraged if you read papers that use statistical techniques that you're not familiar with. Remember that the goal of most statistical procedures is to test a hypothesis. The different procedures are after the same answer: *if the null hypothesis is true, are the observed sample results unlikely?* Each statistical procedure transforms the observed sample results in some way and produces what's called a **test statistic** which is used to calculate how unusual the observed results are if the null hypothesis is true.

In the previous chapter you calculated a *z*-score by subtracting the hypothesized population mean from the observed sample mean and dividing the result by the standard error. The test statistic is the *z*-score. The normal distribution is used to determine how unlikely the observed results are if the null hypothesis is true.

The actual mechanics of how a test statistic is computed don't matter, as long as the data meet the required assumptions for that procedure. Statistical software takes care of the transformations that are required and returns the value of the test statistic and the observed significance level. Remember that the **observed significance level** is the probability of observing results at least as extreme as those observed when the null hypothesis is true.

Statistical tests have unusual names. Some are named after the sampling distribution which is used to evaluate whether the value of the test statistic is unlikely if the null hypothesis is true. Others are named after the statisticians who introduced them. For example, if you don't know the value of the population standard deviation but estimate it from the sample, instead of calculating a *z*-score you calculate what's called a *t*-statistic and then use the *t*-distribution to calculate the observed significance level. (The *t*-distribution looks very much like the normal distribution but shifts the areas a little to make up for estimating the standard error. You'll frequently see *t*-values in papers. The chi-squared statistic is named after the distribution which is often used for testing hypotheses about count data.)

Exercise: In the previous abstract, identify the statistic used to test each of the hypotheses.

Step 3: Check Whether Your Data Meet the Required Assumptions for the Procedure

Every statistical test involves certain assumptions about the populations from which the data are sampled. Some statistical procedures go haywire when their assumptions are violated, while others remain fairly true to course. (A statistical procedure is termed **robust** if it can withstand violations of some of the underlying assumptions. That's a compliment for a statistical procedure.) Some assumptions, such as independence of observations and normality, are common to many procedures.

Observations are independent of each other if one observation isn't related in anyway to another. If you have an experimental design which takes multiple measurements of the same variable for an individual at different times or under different conditions (called a repeated measures design) or if your design involves clusters of individuals, for example, all members of a household, or students from the same school or patients from the same hospital, special statistical procedures are needed.

If your data fail to satisfy the assumptions required for a particular procedure, you can attempt to transform the data values to better comply with the necessary assumptions. For example, you might find that taking logs or square roots of the data will make the observations more normal. That's why sometimes you'll see results in papers given in logs or square roots, though it's better in many cases to report the results in the original units even when the data are transformed. You can also use **nonparametric tests** which make limited assumptions about the underlying distributions of the data.

Step 4: Assume that the Null Hypothesis is True

You must assume the null hypothesis is true and then determine if the observed sample results are unusual when the null hypothesis is true.

Step 5: Calculate the Observed Significance Level (the *p*-value)

If you're using statistical software to analyze your data, once you've selected a procedure, the software will give you the value of the test statistic and its observed significance level, the probability of observing results as extreme as the ones you've observed when the null hypothesis is true. This is the *p*-value that is reported in papers and presentations.

Exercise: For each of the hypotheses that you've identified, what is the observed significance level that is reported.

If you're using a statistical software package, the observed significance level may be displayed as only zeroes. That doesn't mean it's zero, just that the observed significance level is smaller than the number of decimals displayed. You can often click on a cell and see the observed significance level to more decimal places.

Step 6: Decide Whether to Reject the Null Hypothesis

If the observed significance level is small enough (usually less than 0.05), you can reject the null hypothesis. Traditionally, 0.05 is used as the cutoff for "unlikely" although there's nothing sacred about it. There's remarkably little difference between an observed significance level of 0.049 and 0.051. Both tell you that the observed results are unlikely if the null hypothesis is true. When publishing results give the actual observed significance level (p=0.039), not the cutoff value (p<0.05). You should always evaluate the actual strength of the evidence against the null hypothesis instead of relying on a magical number.

Sometimes, an alternative hypothesis specifies in advance the direction of the difference. For example, if you're administering nutritional supplements to infants you may be confident that they can only increase weight, or if you're studying ITN distribution you may know that their use can only decrease malaria incidence. If you can specify the direction of the difference *before* you analyze the data, you can adjust the observed significance level to take that into account. You can use what's called a **one-tailed test.** You reject the null hypothesis based on the probability for only the anticipated direction. You should use one-tailed tests with extreme caution because it's difficult to know the direction of the difference with certainty. In medicine, many treatments thought only to benefit patients have turned out to be harmful. Don't use one-tailed tests simply to make your observed results more unlikely.

Exercise: Which of the null hypotheses in the abstract do you reject? On what do you base your decision?

Step 7: Report Your Results Correctly

Because hypothesis testing depends on probabilities, and not certain knowledge, the conclusions you can draw are severely limited. Based on the observed significance level you can conclude only that the observed results are unlikely if the null hypothesis is true or that the observed results are not unlikely if the null hypothesis is true. You can either reject or not reject the null hypothesis. It's easy to get carried away and make statements that are unwarranted:

- **Don't overestimate the importance of tests of statistical significance.** Statistical tests and tests of significance provide useful information but their importance is often exaggerated. A *p*-value alone is not a good summary of the data. It tells you nothing about the magnitude of an effect or the associated variability. Graphical methods and confidence intervals are much more useful tools. Simple techniques often provide much more insight into the data. With large enough samples, small unimportant effects may be deemed “statistically significant”, while large and important effects may be missed if the study is poorly designed.

- **Don't equate statistical significance with practical significance.** When you reject the null the hypothesis, it is not necessarily true that the differences or associations you found are important. For large samples, even very small observed differences in means may be statistically significant. For example, if you find that a new treatment prolongs life by one week compared to the standard, it is of little practical importance. Always examine the actual observed differences and focus on those that are both statistically significant and practically meaningful.

- **Don't claim that you proved the null hypothesis is true.** You can never prove the null hypothesis is true. Think about it. Consider a null hypothesis that states that a coin is fair, that it has the same probability of landing on either side. If you flipped the coin 1000 times and it comes up 500 times on each side, have you shown that the coin is absolutely fair? Of course not. Your observed results are consistent with many different population values: 0.501 and 0.499; 0.5003 and 0.4997 to mention a few. Any value within the 95 percent confidence interval for the population probability cannot be excluded as a true value.

  Your failure to reject the null hypothesis may also mean that you haven't gathered enough evidence to reject it. Sometimes a legal analogy is made. The null hypothesis is compared to the presumption of innocence. Failure to find a defendant guilty doesn't prove innocence. All it says is that there was not enough evidence to establish guilt. If your sample size is small, you may fail to reject the null hypothesis even when population differences are large. That's why it's important, before you plan a study, to determine how big of a sample you need in order to detect what you consider to be an important difference.

- **Don't claim that the observed significance level is the probability that the null hypothesis is true.** The null hypothesis is either true or it is not. You don't know which.

- **Don't claim that you proved that the alternative hypothesis is true.** It's possible to get very unusual sample results when the null hypothesis is true. Just because your results are unusual doesn't mean the null hypothesis is indisputably false and the alternative hypothesis is true. The observed significance level tells you how often you would get sample results as extreme as the ones you observed if the null hypothesis is true. The observed significance level is never exactly 0.

To Err is Statistical

Whenever you test a hypothesis, you have two paths for making a mistake: you can reject the null hypothesis when it is true or not reject the null hypothesis when it is false. Figure 4.1 shows the possible outcomes and the creative names statisticians attach to them.

Figure 4.1: Outcomes of Testing a Null Hypothesis

	The null hypothesis is:	
Your action:	True	False
Reject	Type 1 error	You are correct
Not reject	You are correct	Type 2 error

You commit a **Type 1 error** whenever you reject a null that is true. You can think of it as finding an innocent person guilty. You commit a **Type 2 error** when you fail to reject a false null hypothesis. You set a guilty person free.You can decrease your Type 1 error rate by refusing to reject the null hypothesis unless the sample results are very, very unusual. Instead of using a cutoff of 0.05, you could use a smaller value, say 0.00001. The problem with this strategy is that you're increasing your chance of committing a Type 2 error. You're setting a very high standard for rejecting the null hypothesis when it is false. Whenever you test a hypothesis you must balance the likelihood of making one of the two possible mistakes. If you make one mistake less likely, you make the other more likely. You have to decide if it's better to set a guilty man free or imprison an innocent man.

Statistical Power

Failure to reject the null hypothesis when it is false is a serious issue for anyone designing a study. It's a terrible outcome if you have a treatment which is much better than the standard but you don't reject the null hypothesis that it is as good as the standard. You fail to find the real difference.

In statistics **power** refers to the probability of rejecting a false null hypothesis. The greater the power the more likely you are to reject the null hypothesis when it is false. Power depends on several factors:

- The size of the true difference in the population
- The variability within the groups
- The sample size
- The significance level at which you reject the null hypothesis

Let's look at an example. Your standard drug cures 70% of the population. Your new drug cures 71% of the population. That's a very small difference and it's highly unlikely that you'll be able to detect it since the two sampling distributions of means overlap so much. Do you care? Probably not. It's not an important difference. What if the new drug cures 90% of all cases? For the same sample size as before, your chances of detecting the difference improve, since the distribution of possible sample means don't overlap as much.

Figure 4.2 shows sampling distributions of the means when two population means are similar (30 and 31), Figure 4.3 when they are quite different (30 and 50) and the standard error is the same. When the means are similar the two distributions overlap a lot. When the means differ more, the two distributions are better separated. The larger the difference between two treatments the better your chance of finding the difference since there is less overlap between the distributions.

Figure 4.2: Similar population means

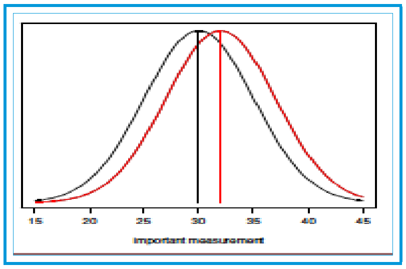
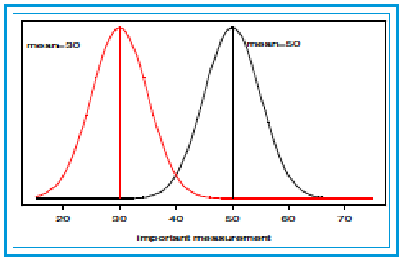


Figure 4.3: Different population means



The spread of sample means, the standard error of the mean, depends on two factors: the sample standard deviation and the sample size. You usually can't control how much the observations in a sample vary but you can control the sample size. The larger your sample size the better the separation between the two sampling distributions of the mean because the sample means cluster more tightly as the sample size increases.

Figure 4.4 shows two sampling distributions with means of 30 and 40 and a standard error of 5. The two distributions overlap a lot. If you decrease the standard error to 2 by increasing the sample size, you get the results shown in Figure 4.5. The difference between the population means is the same but the sampling distributions have much less overlap because the larger sample size results in a smaller standard error. You now stand a better chance of being able to reject the null hypothesis

Figure 4.4: Small sample size

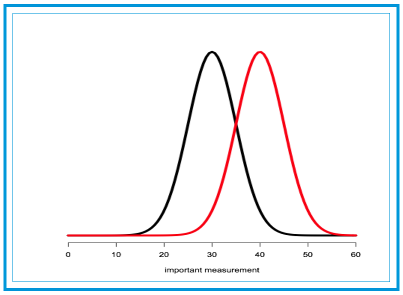
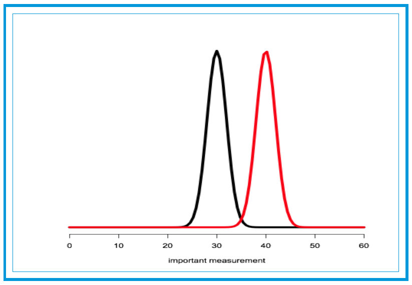


Figure 4.5: Larger sample size



Estimating Sample Size

Before you conduct any study you must determine how large a sample size you need to have a good chance of rejecting the null hypothesis if it is false. There's little point in evaluating differences between two treatments if the chances are slim that you'll be able to detect even big differences when they exist. Your ability to reject the null hypothesis depends on the actual difference between groups in the population, so you have to specify the smallest difference that you want to detect and the probability that you want for detecting that difference (the power).

If you're studying differences in infant birth weights between two groups, you probably don't care if you fail to detect a 10 gm difference in the two populations. The difference may be real but it's not important. You might decide that it's important to detect a difference in means of at least 300 grams. Ideally, you'd like to find a sample size that guarantees that you will always detect a difference of at least 300 grams, if it exists in the population. You've learned by now, no doubt, that certainty doesn't exist in the world of statistics. You must specify the probability that you want for detecting this difference. For example, you may want a 90% chance of detecting a difference of at least 300 grams. You must also indicate the observed significance level that you will use to reject the null hypothesis and an estimate of the standard deviation of the observations in each of the groups.

Figure 4.6 gives the sample size in each group required to find differences of varying sizes, when the standard deviations are 500 grams in each of the two groups and 5% is the significance level used to reject the null hypothesis. For each difference the sample size is shown for 80% power and 90% power.

Figure 4.6: Sample Size in Each Group to Detect Minimum Difference (alpha=0.05)

	Population Difference in Means Between the Two Groups (grams)			
Power	**100**	**200**	**300**	**500**
80%	393	99	44	16
90%	526	132	59	22

Look at the column labeled 100 grams. The first entry, 393, is the sample size required in each of two groups to have an 80% chance of detecting a difference of at least 100 grams if it exists in the population. If you want a 90% chance of detecting this difference, you'll need 526 infants in each group. You see that for a fixed power, the larger the difference, the smaller the sample size required to detect it. For a fixed difference, the greater the power the larger the sample size. This is not surprising. It makes sense that larger differences are easier to detect than smaller differences. Similarly, if you want more power you have to pay for it by increasing the sample size.

The sample size also depends on **alpha**, the significance level you use to reject the null hypothesis. If you use a large value for alpha it's easier to reject the null hypothesis and the required sample sizes are smaller. However, you are increasing the Type 1 error, the probability of rejecting the null hypothesis when it is true.

Whenever you fail to reject the null hypothesis, you have to worry about whether your sample size was large enough to detect important differences. That's why before you design a study it is important to determine an appropriate sample size for the magnitude of difference that you wish to detect.

Examples of Power Calculations in Studies

When you read papers that report negative results, those for which the null hypothesis is not rejected (p>0.05), remember that one possible explanation for the results is that the sample sizes were not big enough to find even large differences. Look for discussion of power when you are reading research results.

For example, in a study of the effects of rapid malaria diagnostic tests on treatment and outcome Msellem (2009) et al. explain:

> According to initial sample size calculations, a total of 850 participants were needed to identify an assumed reduction in malaria diagnosis of 50% with the addition of RDT to clinical diagnosis, at a 5% significance level and a power of 80%, without controlling for design effect of clustering within PHCUs.

In a study of the effects of impregnated bed sheets (shukas) on malaria prevalence Macintyre et al. (2003) state

> A minimum sample size of 440 (220 experimental and 220 control participants) was calculated using a confidence level of 95%, power of 80%, an estimate of 20% malaria parasitaemia in the control group, and a least extreme detectable difference of 10% among experimental group members. The estimated prevalence of malaria parasitaemia was based upon clinic reports that suggested a prevalence of 50% during the rainy season. Since this estimate was not based on blood samples we used a more conservative estimate of 20% to calculate the sample size.

Although the authors of the above paper were careful in estimating a sample size, they encountered one of the perils of sample size estimation. You have to have reasonably good information about the population before you can estimate a sample size. If your assumptions about the population are incorrect you may not have the power that you were aiming for. Macintyre et al. based their sample size estimates on a malaria prevalence rate of 20% in the untreated group. It turns out that the malaria prevalence rate in the control group was only 2%, a very big difference. You need much larger sample sizes to detect a reduction in prevalence between the control and experimental group when the control rate is 2% than when it is 20%.

The authors correctly conclude:

> Conclusions: These results suggest that permethrin-impregnated bedsheets may be protective against malaria prevention but *further studies with greater power are required to confirm this*.

Design Effect

Taking samples of people that are spread over different areas is expensive. Interviewers may have to travel long distances and be housed and supported in far flung locations. That's why it's attractive to sample clusters of people, such as those living in the same manyatta or village. The disadvantage of such a strategy is that people within the same cluster are more similar than people from different clusters. The observations don't vary as much as if they came from a simple random sample. You and your brother and sister are probably more similar in many respects than two unrelated people. Adding another person from the same cluster to a study doesn't give you as much information as adding another person from a different cluster.

The **design effect (deff)** is a measure of how much the sampling variability in a cluster sample differs from the sampling variability in a simple random sample. It's based on measuring the similarity between two randomly selected people within a cluster as compared to two randomly selected elements from different clusters. A deff of 2 tells you that you need twice as many cases from the clusters to get the same standard errors as you would if you took a random sample.

The design effect can vary for each question or measurement you take. For example, if you are sampling families, you don't get very much additional information if you ask each member how long the family has lived in their present housing. The responses should be highly correlated. However, measurements of their blood pressure will probably not be as highly correlated.

The design effect must be included in the computation of sample sizes and confidence intervals and tests of statistical significance. Many statistical software packages will make the appropriate adjustments when the design effect is given.

Sample Size Calculator

For studies based on simple random sampling, you can use a sample size calculator such as the one found at www.openepi.com. For complicated designs that involve several stages such as selecting cities within a region, then hospitals within the region, and then patients within a hospital, be sure to consults with a statistician for appropriate sample size estimates.

Exercise: Use the openepi software to calculate the sample size required to detect a difference of 400 grams, with a power of 90%. (Click Sample Size, then Mean Difference, then Enter New Data. Use 500 as the standard deviation in each of the groups. Click Calculate to see the results.)

Summary

Statistical hypothesis testing requires identification of a null hypothesis and an alternative hypothesis. The null hypothesis claims there is no difference in the population. The alternative hypothesis claims the opposite: there is a difference in the population. To evaluate a hypothesis you must determine the probability of observing results at least as extreme as the ones you observed, when the null hypothesis is true. This probability is called the observed significance level or the *p*-value. Statistical tests estimate this probability. If the observed significance is small, usually less than 0.05, you reject the null hypothesis. If you fail to meet the assumptions required for a statistical procedure the observed significance level may be incorrect.

When you test a hypothesis you can make two types of mistakes: reject the null hypothesis when it is true (type 1 error) or fail to reject the null hypothesis when it is false (type 2 error). Your ability to reject the null hypothesis when it is false depends on the size of the difference between the groups and the sample size in each group.

Bibliography:

Msellem MI, Mårtensson A, Rotllant G, Bhattarai A, Strömberg J, et al. Influence of Rapid Malaria Diagnostic Tests on Treatment and Health Outcome in Fever Patients, Zanzibar—A Crossover Validation Study. *PLoS Med* 2009*:* 6(4): e1000070. Msellem (2009)

Ezeaka VC, Iroha EO, Akinsulie AO, Temiye EO, Adetifa IM (2009). Anthropometric indices of infants born to HIV-1-infected mothers: a prospective cohort study in Lagos, Nigeria. *Int. J. Std. Aids*. 2009 Aug; 20(8):545-8. Ezeaka (2009)

Macintyre, K., Sosler, S., Letipila, F., Lochigan, M., Hassig, S., Omar, S. and Githure, A. A new tool for malaria prevention?: Results of a trial of permethrin-impregnated bedsheets (*shukas*) in an area of unstable transmission. *Int. J. Epidemiol.* 2003: 32 (1): 157-160 .
Macintyre (2003)

5 Statistical Procedures for Testing Hypotheses

- Why do you need statistical procedures?
- When do you use the t-test?
- How you test if two variables are independent?
- What is a relative risk ratio?
- What is an odds ratio?

Statistical procedures are used to estimate the observed significance level--the probability of obtaining sample results at least as extreme as those you've observed when the null hypothesis is true (the *p*-value). It would be great if there was one statistical procedure that could be used for all the different types of hypotheses but, unfortunately, that's not the case. Just as different diseases require different treatments, so different hypotheses require different statistical tests. For example, there are statistical procedures for testing the null hypothesis that two or more population means are equal; that two or more categorical variables are independent; that there is no linear relationship between a dependent variable and a set of independent variables. You already used the binomial procedure to test the null hypothesis that a cure rate is 70% and the one sample *z*-score procedure to test the null hypothesis that a sample of newborns of HIV positive mothers comes from a population with an average weight of 3000 grams.

The names of the statistical tests used, their values, and the observed significance levels are usually reported in scientific papers. For example, in the following abstract that summarizes differences between babies with HIV positive mothers and babies with HIV negative mothers (Ezeaka, 2009), the statistical tests used for determining the observed significance levels are identified as the *t*-test and the chi-square test:

> There were three times more LBW babies in the HIV-positive group than in the uninfected mothers (odds ratio = 3.47, 95% confidence interval = 1.69, 7.27; **chi(2) = 12.99**, P = 0.0003).The maternal weight (**t = 15.85**; P = 0.0001), maternal body mass index (BMI) (**t = 15.07**; P = 0.0003), birth weight of infants (**t = 27.17**; P = 0.0001) and birth length (**t = 31.20**; P = 0.001) were significantly less in HIV-positive mothers than in controls.

This chapter and the remaining chapters assume that you'll be using statistical software for calculating statistical tests so there's little emphasis on what numbers to square, divide or multiply to calculate the statistical tests or build the models. Instead the focus is on what tests are appropriate for a given problem and what the results mean. This makes the course more difficult since it's easier to multiply, divide and square than to think. We hope that in the long run you'll find this approach more valuable.

Some Warning on the Use of Significance Tests

Statistical tests and the *p*-values they produce are important part of data analysis but they should be used and interpreted carefully. Tests are not a substitute for good judgement on the part of the data analyst. Think before you start generating test results. Never perform a statistical test without graphically examining the data first. You can usually learn much more about your data and about the relationship between variables by examining plots and simple tables than by looking at *p*-values.

When reporting results, never report *p*-values alone, always include confidence intervals for the population values, since they provide information about the possible size of the true differences. You know that small, but practically unimportant, differences may have very small *p*-values if the samples are large. Large and important differences may not be statistically significant if the sample sizes are small and the tests have limited power. Journals are biased in favor of studies that report statistically significant differences so the 5% of studies that reject the null hypothesis when it's true stand a good chance of being published. Keep this in mind when you come upon studies that make claims counter to those published by other investigators or present astonishing results.

Assumption, Assumptions

Most statistical tests are based on the assumption that the data are random samples from one or more populations and that the observations are independent of each other. Since it can be difficult to randomly sample cases from the entire population of interest, in practice, selecting cases independently and observing them without bias is often the best that can be realistically achieved. (Imagine the difficulties involved in taking a random sample of all babies of HIV positive mothers in the world!) If your sampling strategy involves taking groups of observations from regions, hospitals or schools, you need special statistical procedures that incorporate the dependencies of the observations.

Statistical tests also require assumptions about the distributions of the variables in the population. Some tests require that data come from a particular distribution, such as the normal distribution, while other tests, called **nonparametric tests**, make less stringent assumptions. Before you use any statistical test, you should check whether your data meet the assumptions required for that procedure. You should also note whether the procedure is sensitive to the assumptions. That is, how much does the observed significance level that is calculated change if the assumptions are violated? There are special plots and tests that you can use to check whether the necessary assumptions appear to be violated.

Distributions of Test Statistics

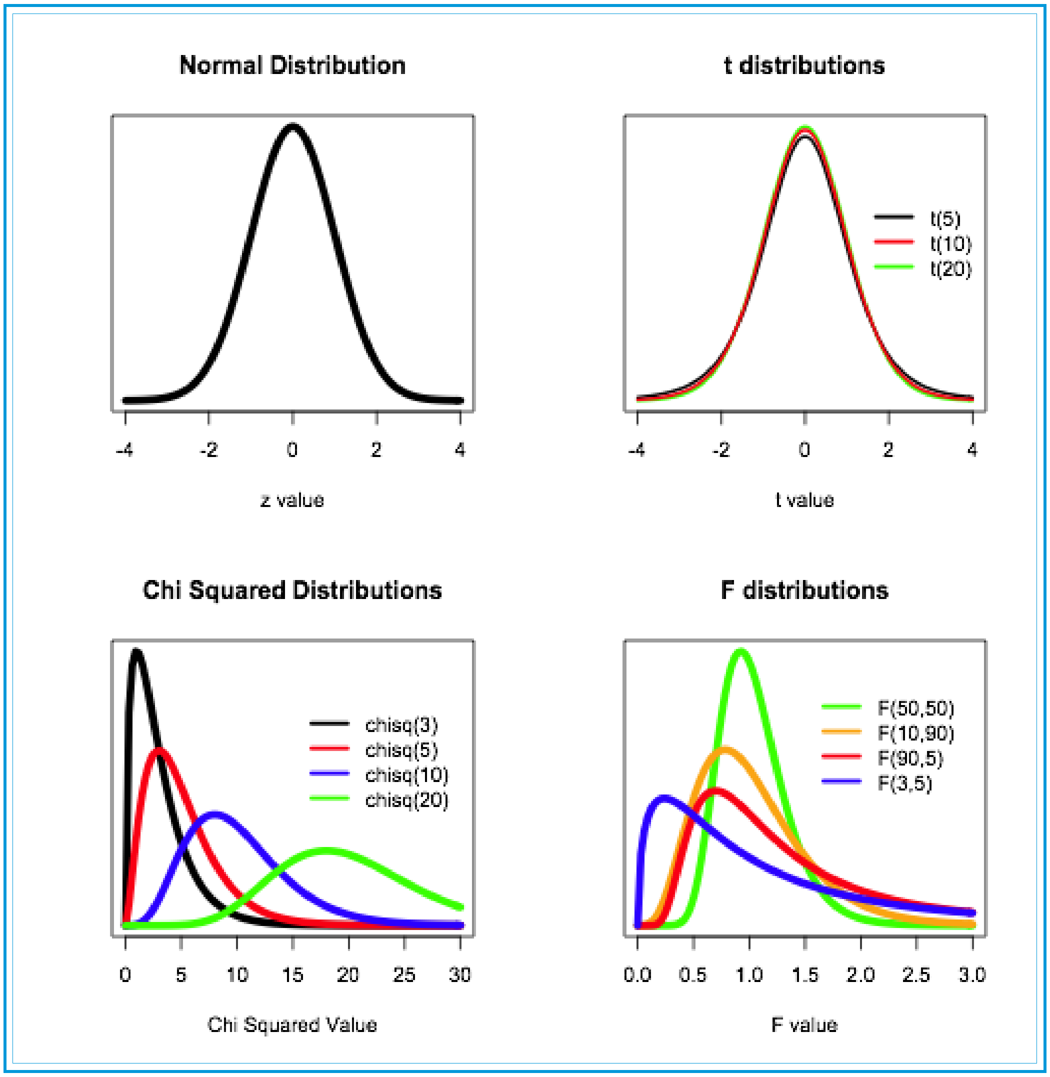
All statistical tests calculate a number based on your observed data (called the value of the test statistic). Different tests compute these numbers in different ways, but for all tests the value of the test statistic is used to calculate the probability of observing results as extreme as those you observed when the null hypothesis is true. Test statistics have mathematical distributions and sometimes tests take their names from these distributions.

For example, to test the null hypothesis that your sample comes from a population with a known mean and standard deviation, you use a *z*-score as the test statistic and the normal distribution as the reference distribution for calculating the *p*-value. You answer the question *How often would I get a z-score at least as large as the one I observed, in absolute value, if the null hypothesis is true?* Other common distributions that are used as reference distributions for test statistics to calculate *p*-values are the *t*-distribution, the chi-square distribution (χ^2) and the *F*-distribution. The statistical test determines which distribution is used.

Although you do not need to worry about the details of these distributions, you'll encounter the symbols *t*, *F*, and chi-square often when reading papers or analyzing data. Figure 5.1 shows you an idea of what these distributions look like. (It's good to have a mental image of what is being talked about, rather like photos of deceased frequently mentioned relatives.)

Some distributions change their shape based on what's known as their degrees of freedom. You can think of **degrees of freedom** as the number of independent pieces of information that go into calculating a test statistic. Degrees of freedom can be based on the size of the sample, or the number of groups involved in a comparison, or on the number of rows and columns in a table. Except for the normal distribution, you see several distributions of the same type on a single plot in Figure 5.1. The distributions have different degrees of freedom which are shown in the legend. The *F*-distribution has a pair of degrees of freedom, one for the numerator and one for the denominator of the test statistic.

Figure 5.1: Statistical Distributions Used in Testing Hypotheses



You see that the *t*-distribution looks very much like the normal distribution and for sample sizes greater than 30, it's indistinguishable from the normal distribution. For small sizes it has a little more area in the tails than a normal distribution. The *t*-distribution is used for testing hypotheses about means when the population standard deviations are not known. Estimating the standard deviation from the data introduces additional uncertainty.

Statistical Software

If you're enrolled in this course, you have access to a computer that you can use to run programs that calculate statistical tests. Some of the programs like SAS and IBM SPSS are very expensive, so you may not have them installed on the computer you are using. Don't worry. There is also open source (free) statistics software called **R** which has all of the capabilities of the expensive packages. With enough patience and skill you can program **R** to do almost anything, except cook your dinner. If analyzing data files is an important part of your job you may want to invest the time and effort required to learn **R** . A brief introduction to **R** by example is at www.statmethods.net.

If you just want to learn about statistical procedures or analyze small data sets, there is free web based software such as www.openepi.com and www.graphpad.com that calculate many of the commonly used statistical tests, though they may not have as many features as the more expensive software or **R** and cannot handle large data files. A good place to look for detailed and reliable information about statistical software packages and their output is www.ats.ucla.edu/stat.

Game Plan for Testing a Hypothesis Using Statistical Software

The steps you take to test a hypothesis using statistical software are straightforward:

- **Formulate a null and an alternative hypothesis about the population(s).** A computer can't do this for you. Make sure your hypothesis is appropriate for the data you have. Don't try to test hypotheses about average religions or average drugs administered. Computers don't know whether the numbers in your data file correspond to codes for religions or to the number of children. Software will happily calculate averages for nominal variables.

- **Select a statistical test for evaluating your hypothesis.** Keep in mind any assumptions a particular test requires. Figure 5.2 lists frequently encountered statistical hypotheses about means and proportions and the statistical tests used to test the hypotheses. The name of the test often used in statistical software packages is in the last column.

- **Identify the statistical procedure or commands in your software that compute the test that you want.** If there are several versions of a test, make sure you've selected the correct one. If there's a Help system or tutorial included in your software, use it. Most statistical software programs are organized by the type of test that is computed or model that is built.

- **Check to see if your data violate the assumptions required for the selected test.** Sometimes displays that help you evaluate the assumptions are part of the output for a particular procedure. If not, you have to look for them elsewhere, perhaps in other special purpose procedures.

- **Look for any warnings or errors the software displays when you run the procedure.** For example, the software may warn you if there are too many empty cells in a table or if counts are too small. Make appropriate modifications to your data, such as combining categories, or choose another statistical test which doesn't require the troubling assumptions.

- **Check that the sample sizes and other numbers that are shown in the output are correct**. For example, if you didn't enter all of the numbers correctly the number of cases in your sample will not match the sample size reported by the software.

- **In the output from the statistical procedure identify the test statistic, and the *p*-value, that correspond to your null hypothesis.** Check the labeling to see if it is for a one-tailed or two-tailed test. Don't choose the one tailed value because it is smaller. In most circumstances you should use the two tailed value. Note that many software packages report multiple test statistics within the same procedure. You may find a test that variances are equal when you're interested in testing equality of means. Don't confuse the results of additional tests with those of the null hypothesis you're testing. There's a reason that these additional tests are performed as part of your analysis. The Help system may tell you or you can enter the name of the unrecognized test into a search engine. If the output lists numerous tests for the same hypothesis, don't automatically select the one with the smallest *p*-value. If the observed significance level is identified as “asymptotic”, that means your sample sizes have to be reasonably large for the observed significance level to be accurate. If the *p*-value is labeled as “exact” that indicates that the observed significance level is calculated without the need for making assumptions about the distribution of the test statistic. Exact tests are useful for small data sets when the validity of the asymptotic results is in doubt. Do some research to see how the tests differ and the circumstances when one is preferred over another.

- **Reject the null hypothesis if the observed significance level is small (eg. p<0.05).**

- **Report the value of the test statistic, the observed significance level and the confidence intervals for population means, mean differences or ratios**. Include plots and descriptive statistics that summarize your findings and help the reader understand them.

Figure 5.2: Frequently Used Statistical Tests for Testing Hypotheses

Null Hypothesis	Test Statistic	Procedure Name
A sample comes from a population with a known mean and standard deviation	z-score	One sample normal test of means
A sample comes from a population with a known mean but the standard deviation is estimated	t-statistic	One sample t-test
Two independent population means are equal	t-statistic	Independent samples t-test
Two related (paired) population means are equal	t-statistic	Paired samples t-test
Two or more independent population means are equal	F-statistic	Oneway Analysis of Variance
Two or more independent population proportions are equal	χ^2 statistic	Chi squared test for binomial proportions
Rows and columns of a table are independent	χ^2 statistic	Crosstabulation

Testing That Two Population Means Are Equal

The *t*-test is probably the most frequently used statistical procedure. It's used to test the null hypothesis that two population means are equal. Another way of stating this is that the difference between the two true means is 0.

There are several different versions of the t-test for means:

- The **one sample t-test** is used to test the null hypothesis that a single sample comes from a population with a known mean. For example, you want to test whether the newborns in your sample come from a population in which the average birth weight is 3000 grams. The value 3000 grams is a known constant. You're not estimating it from a sample. You are, however, estimating the standard deviation from the sample.

- The **two independent sample t-test** is used to test the null hypothesis that two independent samples come from populations with the same mean. There are two slightly different versions depending on whether you assume that the variances are the same in the two populations. For example, in the HIV study, the two independent sample t-test is used to determine if babies of mothers with and without HIV have the same average maternal weights, maternal body mass indices, and infant weights and lengths in the population.

- The **paired samples (matched-cases) t-test** is used to test the null hypothesis that, in the population, two means are equal when you have measurements from pairs of people or objects that are similar in some important way. For example, you observed the same person before and after a treatment, or you obtain measurements from spouses or twins. The reason for pairing observations is to make the two groups being compared more similar by eliminating some of the differences between subjects. For example, you ask each person about their ITN use before and after information about their importance is distributed. Since the same person is asked the questions before and after you know the before and after responses come from people with identical values for variables like education or age that may affect the responses.

  If the variables on which the cases are matched aren't related to the variable you are studying, then analyzing the data with a paired test will actually make it harder for you to detect true differences than if you analyzed the data with a two independent samples t-test since the degrees of freedom are smaller for a paired design.

  If you observe the same subject under two conditions make sure that the effect of one treatment has worn off before the other is given. If you're giving the same task or test to people under different conditions, make sure that any improvement shown is not due to a learning effect—doing something better the second time because you've had practice.

Assumptions: If the samples sizes are small (say fewer than 30 cases) the distributions of the variables in the population should be approximately normal. If sample sizes are reasonably large, don't worry about normality.

How it's calculated: Find the difference between the two means. Calculate the standard error of the difference which is a measure of how much you expect sample means, based on the same number of cases from the same population to vary. Divide the difference between the two means by the standard error of the difference. The standard error of the difference is calculated differently depending on the type of t-test.

$$t\ (degrees\ of\ freedom) = \frac{(observed\ difference\ between\ means)}{(estimated\ standard\ error\ of\ the\ difference)}$$

Find the observed significance level-- the probability of a *t*-value in absolute value at least as large as the one you observed when the null hypothesis is true.

One Sample t-test Example

It's not unusual to wonder if a sample comes from a population with a known mean. For example, you may want to know if the average weight of newborns of HIV negative mothers in your district is 3000 grams, a national standard. The national standard is a known constant, you're not estimating it from your data.

Figure 5.3 is **R** output for one sample t-test of the null hypothesis that the sample of infants of HIV negative mothers comes from a population with an average weight of 3000 grams. The command used to generate the output in **R** is highlighted. (You have to enter all of the observed values into **R** and give them a name. The birth weights are stored in a variable called bwNoHIV.) In the output the alternative hypothesis is identified as “true mean is not equal to 3000”, since mu=3000 is specified in the **R** command as the true value you want to test against. You know that the null hypothesis is the opposite of the alternative hypothesis—so the null hypothesis is that the true mean *is* equal to 3000 grams.

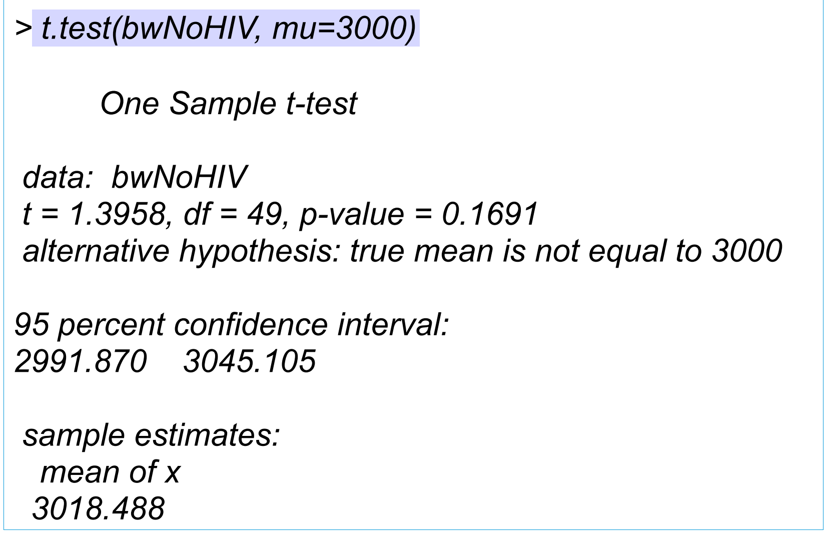


Figure 5.3: One Sample t-test output from R

```
> t.test(bwNoHIV, mu=3000)

        One Sample t-test

 data:  bwNoHIV
 t = 1.3958, df = 49, p-value = 0.1691
 alternative hypothesis: true mean is not equal to 3000

95 percent confidence interval:
2991.870    3045.105

 sample estimates:
   mean of x
  3018.488
```

From Figure 5.3, the value of the *t*-statistic, is 1.40. The degrees of freedom (df=49) used to calculate the *p*-value from the *t*-distribution are the number of cases (50) minus 1. The probability of observing a *t*-value as large as 1.40 in absolute value, when the null hypothesis is true is 0.1691. Since it's not less than the usual cutoff of 0.05, you cannot reject the null hypothesis that your infants come from a population with a mean of 3000 grams. Have you proved that the average true birthweight for the population from which your sample is selected is 3000 grams? Of course not! All you can say is that you don't have enough evidence to dismiss the null hypothesis.

The average observed weight is 3018.49 grams. The 95 percent confidence interval for the true average weight in the population is from 2991.87 grams to 3045.11 grams. Any value in this interval is a plausible value for the true population weight. The interval includes the value 3000 grams so it cannot be excluded as a plausible value.

Exercise: What displays would be useful for describing the results of this study? How would you change the **R** command if you wanted to test that the true average weight is 2900 grams?

Two Independent Samples t-test Example

In the one sample t-test example, you analyzed a single set of numbers—a sample of the birthweight of newborns in your district whose mothers didn't have HIV.

If you have two independent samples of infants, some born to HIV positive mothers and others to HIV negative mothers, and want to test whether the true average birth weights are equal, you use the two independent samples t-test. That's the test that was used to analyze the differences in anthropometric measures between the two groups in the study described at the beginning of this chapter.

Figure 5.4 is **R** output from the two independent samples *t*-test. (**R** output can be very sparse in comparison to some other software packages. You can, however, get all kinds of additional statistics and plots by asking for them.) The *t*-test is identified as the Welch Two Sample *t*-test. The Welch *t*-test doesn't require any assumptions about the equality of the variances in the two populations. There is also a version of the two sample t-test that assumes that the variances in the two populations are equal. If you're not sure which one to use, just go ahead and use the Welch test.

Figure 5.4: Two Independent Sample t-test output from R

```
> t.test(bwHIV,bwNoHIV)

        Welch Two Sample t-test

 data:  bwHIV and bwNoHIV
 t = -10.3448, df = 958.354, p-value < 2.2e-16

 alternative hypothesis: true difference in means is not equal to 0

 95 percent confidence interval:
    -84.52972 -57.57255

  sample estimates:
 mean of x mean of y
 2810.345  2881.396
```

The t-value is -10.34, the degrees of freedom are very large because there are 520 infants in the No HIV group and 450 cases in the HIV group. The *p*-value is very small so it is given in scientific notation. The value shown is 2.2 but the decimal must be moved 16 places to the left, making it a *very* small number (0.0000000000000022). The observed significance level is never 0, but software output sometimes shows values of 0, perhaps written as 0.0000 or the like. Sometimes you can click on a cell to see the value to more decimal places. If you're fortunate enough to obtain such a small observed significance level, don't submit the *p*-value to a journal with all of those zeroes. Just say it is less than the smallest number the software displays, say 0.00005 in this example. Based on the observed significance level you handily reject the null hypothesis.

The 95% confidence interval for the true difference in means is from -85 grams to -58 grams. Any values within this interval are plausible values for the true difference between the two groups. Infants of HIV positive mothers may weigh anywhere from 57 to 84 grams less than infants of mothers without HIV. You have to pay attention to which sample means is subtracted from which. In the t-test command the birth weights of the HIV infants is listed first, so the second mean (for HIV negative infants) is subtracted from the first. The negative sign tells you that infants of HIV positive women weigh less than infants of HIV negative women.

Paired Samples t-test Example

To study the effect of HIV on neonatal birth weight, Floridia et al. (2008) matched each of 600 infants born to HIV positive mothers to a control baby of the same gestational age and gender but with an HIV negative mother. This was done to minimize the effect of differences in gestational age and gender on birth weights. Their goal was to determine whether the true average birthweight differs between the HIV cases and controls.

You should use a paired t-test to analyze the results of this study. Although you have two samples, they are not independent, the infants are matched on the basis of their gestational age and gender. You know that birthweight is related to both gestational age and gender, so it is reasonable to create pairs on the basis of these variables.

Figure 5.5 is the output from the **R** procedure when observations are paired. The t-value is -3.08. The observed significance level is quite small, *p*-value = 0.002, so you can reject the null hypothesis that the true average weights are the same for the two groups. Infants of the HIV positive mothers weigh less. The observed average difference in weight between the HIV infants and controls is -93.62 grams. The 95% confidence interval is from -153 grams to -34 grams. The interval is wide but does not include 0. The true population difference may be anywhere between 34 and 153 grams.

Figure 5.5: R output for Paired Samples t-test

```
 >t.test(HIVweight, Controlweight,paired=TRUE)

          Paired t-test

    data:  HIVweight and Controlweight

   t = -3.0825, df = 599, p-value = 0.002147

    alternative hypothesis: true difference in means is not equal to
0

   95 percent confidence interval:

    -153.27497  -33.97374

     sample estimates:

    mean of the differences

           -93.62435
```

You can also analyze paired data by computing the differences between the values of the two variables for a case and then using a one sample t-test to test that the true difference is 0. The result are identical.

Analysis of Variance: Testing that Several Population Means Are Equal

The independent samples *t*-test is limited to testing hypotheses about two population means. Another statistical test, called **analysis of variance (ANOVA),** lets you test the null hypothesis that more than two population means are equal. For example, you have five different treatments and want to test whether the true final average diastolic blood pressure is the same for all treatments.

Instead of using a *t*-statistic, you calculate what's called an *F*-statistic and use the *F*-distribution to determine the observed significance level. (If you have two independent groups the results will be the same if you use a t-test that assumes equal variances in the two groups or the analysis of variance.)

Analysis of variance can also be used when cases are classified into groups based on several grouping variables. For example, subjects are classified by both gender and diagnosis. You can test the null hypotheses that the true average diastolic blood pressure is the same for all diagnoses, that the true average values are the same for males and females, and that for each diagnosis, males and females respond in the same way. That is, there is no interaction between the two variables.

In many situations multiple measurements are obtained from the same individual, say daily readings of blood pressure. Since multiple observations from the same individual are not independent, you need a modification of the usual analysis of variance called **Repeated Measures ANOVA**. Detailed discussion of advanced statistical techniques is beyond the scope of this course but it's important for you to recognize them in the literature and to understand the basic idea of what's being done.

Assumptions for ANOVA

Analysis of variance requires independent random samples from normal populations with equal variances. If the sample sizes in the groups are approximately equal, unequal variances don't matter much. There are other statistical tests for equality of means, such as the Welch and Brown-Forsythe tests, that don't require the assumption of equal variances. There also nonparametric tests that don't require the assumption of normality.

Calculating an ANOVA test

The *F*-test is calculated by finding the ratio of two estimates of variability: how much observations within a group vary, and how much the group means vary. If the null hypothesis that all population means are equal is true, these two estimates should be similar. If they're not, that's evidence against the null hypothesis that all population means are equal. If the value of the *F*-statistic is large, the population means vary more than you expect if the null hypothesis is true.

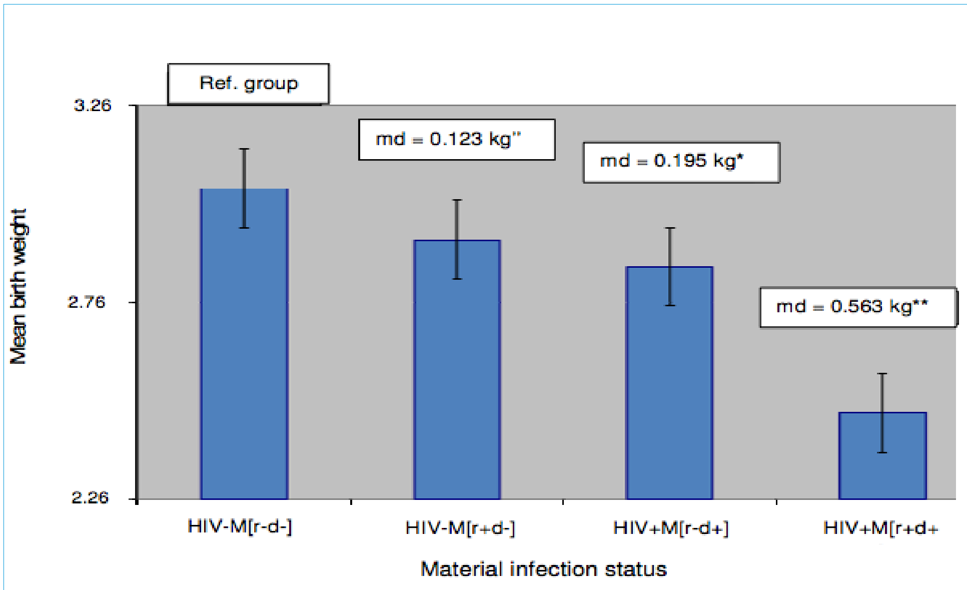
One Way Analysis of Variance Example

Laar (et al. 2010) studied 1154 pregnant women, 443 who were HIV positive and 711 who were HIV negative at recruitment. Malaria at recruitment (r) and at delivery (d) were also recorded. Four groups of women, based on HIV status and presence or absence of malaria, were of interest:

1. HIV-Mal[r-d-]: HIV - no malaria This is the reference group.
2. HIV-Mal[r+d-]: HIV - malaria at recruitment but not at delivery
3. HIV+Mal[r-d+]: HIV+ malaria only at delivery
4. HIV+Mal[r+d+]: HIV+ and malaria both at recruitment and at delivery

Analysis of variance (ANOVA) was used to test for differences in mean birth weight and gestation length across the four maternal infection categories. Figure 5.6 shows the average birth weights for the four groups. Note that the vertical axis doesn't start at 0, so the differences between the groups appear larger than they would if the scale started at 0. The lines at the end of the bars represent confidence intervals for the true mean for each group. The number in the box, labeled *md,* is the average difference in birthweight compared to the disease free reference group. The footnote shows an *F*-value of 11.24 with degrees of freedom of 3 and 722. The *p*-value is less than 0.001, so the null hypothesis that the true birth weight is the same in all groups is rejected.

Figure 5.6: Maternal Infection Status and Mean Birthweight



Overall test: F (3 722) = 11.24; p < 0.001; Error bars represent 95% Confidence interval for the mean.
Ref. group = Group compared with in the Post Hoc analysis (Hochberg GT2 test); ''p < 0.05; *p < 0.016; **p < 0.001

Figure 3. Association between maternal infection status and birth weight.

Multiple Comparison Procedures

When using ANOVA the null hypothesis that all population means are equal is rejected, the question of which groups are significantly different from each other remains. If you perform multiple *t*-tests between all possible pairs of means you encounter the problem that the more comparisons you make the more likely you are to call differences significant even when they are not. **Multiple comparison procedures**, sometimes called post hoc procedures, protect you from erroneously calling differences significant when they are not. There are numerous multiple comparison procedures, all with slightly different statistical properties and names. However, they all require larger differences between pairs of means before a difference is deemed significant than does the *t*-test. The result is that you are less likely to reject the null hypothesis when it is true but you are also less likely to find true differences.

In the footnote to Figure 5.6 the authors indicate that the Hochberg GT2 test is used to identify pairs of mean that are significantly different from the reference group. The observed significance level is also given in the footnote. The average difference between the disease free reference group and the HIV negative, malaria on recruitment only group is 0.123 kg. Based on the GT2 test the observed significance level for the difference is $p<0.05$, so you can reject the hypothesis that the true means are equal. Similarly there is a 0.195 kg difference between the reference group and the HIV positive and malaria only at delivery groups. The observed significance level is given as $p<0.016$.

Exercise: What's the difference in means between the Reference groups and the group that is positive for HIV and malaria at both recruitment and delivery? What is the observed significance level for the difference? What is the null hypothesis? Can you reject it?

Testing Hypotheses About Count Data

All of the statistical procedures described so far in this chapter are for testing hypotheses about population means. Now we'll take a look at statistical methods for testing hypotheses about counts.

Calculating Percentages

Well-planned tables and plots of counts and percentages are a critical component of data presentation. Figure 5.7 is a table, sometimes known as a **crosstabulation,** of the data presented by Laar et al. Instead of analyzing

actual birth weights in kilograms as you did previously, you're now looking at the counts of infants that meet the criteria for being low birth weight. HIV status forms the rows of the table, birth weight status forms the columns.
The numbers in parentheses are called **row percentages**. They tell you what percentage of the observations in each row fall into each cell in the row. For example, 22.4 percent (66/295) of HIV positive women delivered low birth weight infants and 14.2 percent (66/466) of HIV negative women delivered low birth weight infants. Overall 17.35% of the infants were low birth weight.

Figure 5.7: Two way table of Birth Weight and Maternal HIV Status

HIV Status	**Low Birth Weight Infant**		
	Yes (%)	**No (%)**	**Total (%)**
Positive	66 (22.4%)	229 (77.6%)	295 (100%)
Negative	66 (14.2%)	400 (85.8%)	466 (100%)
Total	132 (17.35)	629 (82.65%)	761(100%)

You can also calculate **column percentages** that express each cell count as the percent of its column total. For example, of the 132 low birth infants, 50% (66) were born to HIV positive women and 50% to HIV negative women. What does this percentage mean? Very little. You can't interpret the column percent without knowing how many HIV positive and HIV negative women are included in the study. If there are equal numbers of HIV positive and negative women in the study, you can conclude that the rate of low birth weight infants is the same in the two groups. However, if there are not equal numbers of women in the two groups, you can't draw any conclusions based only on these percentages. They cannot be interpreted correctly without knowing the proportion of HIV positive women in the study.

Whenever you display percentages in a table, make sure they are easy to interpret. If one of your variables can be thought of as predictor variable that has an effect on the values of the dependent variable, compute percentages so that they add up to 100% for each category of the predictor variable. In this example, low birth weight depends on maternal HIV status so it's the dependent variable. HIV status is the independent variable. You want to report the percent of low birth weight babies separately for each HIV status.

Testing Hypotheses about a Single Population Proportion

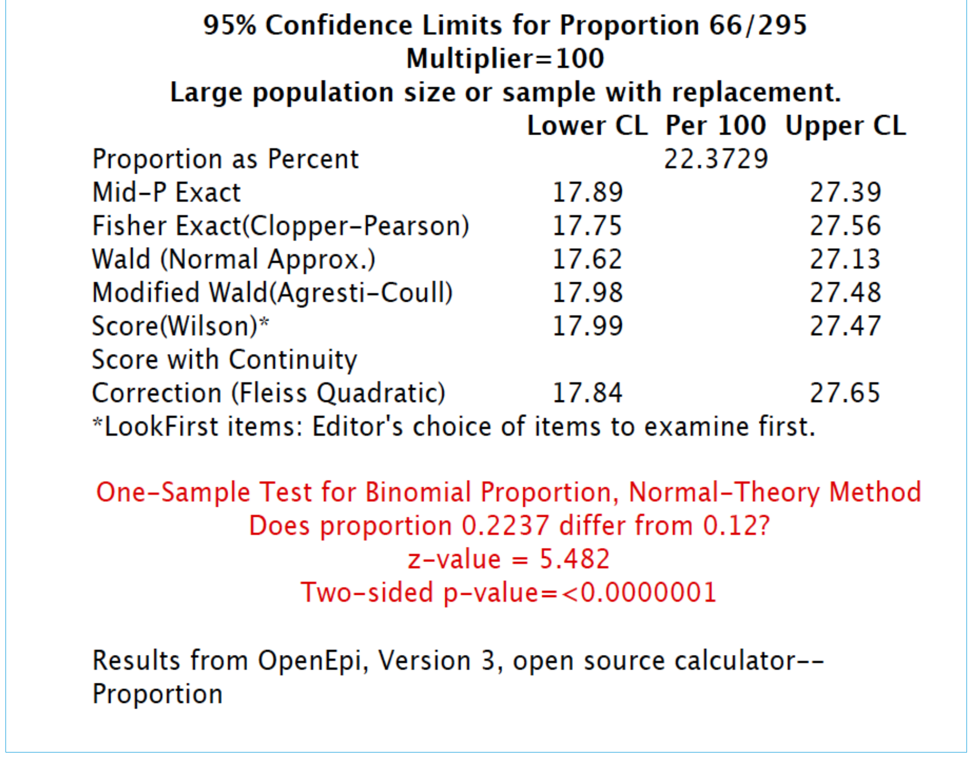
You see in Figure 5.7 that 22.4% of HIV positive women gave birth to LBW infants. If on the basis of large scale studies 12% of infants are known to be of LBW, you may want to test the null hypothesis that your sample of HIV women comes from that population. Another way of phrasing the hypothesis is that the HIV women in your sample come from a population with a 12% rate of LBW .

Figure 5.8 shows results from the OpenEpi calculator for the test that of a binomial proportion. (www.openepi.com/v37/Proportion/Proportion.htm) We are using OpenEpi since you can access it without having to install the program on your computer. You may prefer to download free and more complete software called EpiInfo, wwwn.cdc.gov/epiinfo, especially if you are analyzing data.

The first line of Figure 5.8 reports the observed proportion as 66/295. It's very important to check that you correctly entered numbers into a computer program and that the test results are based on them.

The results of the test of the null hypothesis that your sample comes from a population with a low birth weight rate of 12% are shown in red at the bottom. Since the sample size is large, the Normal-Theory Method is used to calculate the observed significance level. If you have a small sample the observed significance level is calculated exactly from the binomial distribution. You see that the observed significance level is very small and you can reject the null hypothesis that the sample comes from a population with a mean of 12%

Figure 5.8: One Sample Test of Binomial Proportion



```
              95% Confidence Limits for Proportion 66/295
                            Multiplier=100
            Large population size or sample with replacement.
                                      Lower CL  Per 100  Upper CL
Proportion as Percent                           22.3729
Mid-P Exact                             17.89               27.39
Fisher Exact(Clopper-Pearson)           17.75               27.56
Wald (Normal Approx.)                   17.62               27.13
Modified Wald(Agresti-Coull)            17.98               27.48
Score(Wilson)*                          17.99               27.47
Score with Continuity
Correction (Fleiss Quadratic)           17.84               27.65
*LookFirst items: Editor's choice of items to examine first.

One-Sample Test for Binomial Proportion, Normal-Theory Method
            Does proportion 0.2237 differ from 0.12?
                        z-value = 5.482
                Two-sided p-value=<0.0000001

Results from OpenEpi, Version 3, open source calculator--
Proportion
```

Many different methods for calculating the confidence interval for the true population proportion are also shown. Whenever you use OpenEpi and there are many ways to calculate test results or confidence intervals, one of the results has an asterisk indicating that the editors think it's the best method. Go with it. In this case the 95% confidence interval for the true proportion of LBW infants for mothers who are HIV positive is from 17.99% to 27.47%, based on the preferred Score(Wilson) method. The value of 12% is not included in the interval. Any value not in the 95 % interval can be excluded as a plausible value, using 5% as the cutoff for unlikely.

Testing that Two or More Population Proportions Are Equal

When you test the null hypothesis that two population means are equal, you compute the *t*-statistic and then, from the *t* distribution, calculate how unusual the observed *t*-value is if the null hypothesis is true. To test hypotheses about data that are counts, you compute what's called a chi-square statistic and compare its value to the chi-square distribution to see how unlikely the observed value is if the null hypothesis is true.

The most common null hypothesis about proportions, which are just means of binary variables where 0 and 1 are used for the possible outcomes, is that two or more population proportions are equal. You can use a chi-square test to test the null hypothesis that the true proportion of LBW infants is the same for HIV positive and HIV negative mothers. Another way of stating this is that you test whether maternal HIV status and infant birthweight are *independent.* Two variables are independent if knowing the value of one variable doesn't tell you anything about the value of the other variable.

Calculating the Chi-Square Test

The chi-square test is based on comparing observed and expected cell counts. The expected cell counts are the counts that you expect if the null hypothesis is true. Of course, even if the null hypothesis is true, the observed and expected values won't be identical, since the results you observe in a sample vary around the true population value. You want to determine whether the differences between the observed and expected counts are unusually large if the null hypothesis is true.

In Figure 5.7 you see that a total of 132 infants are of low birth weight. That's 17.35% of all infants. If the null hypothesis is true, 17.35% of infants born to HIV mothers and 17.35% of infants born to HIV negative mothers should be LBW. That means there should be 51.17 LBW infants for the 295 HIV positive mothers (295 times 0.17346), and 80.83 LBW infants for the 466 HIV negative mothers (466 times 0.17346).

Figure 5.9: Observed and Expected Values

	Low Birth Weight Infant		
HIV Status	**Yes** **Observed(Expected)**	**No** **Obs (Expected)**	**Total** **Obs (Expected)**
Positive	66 (E=51.17)	229 (E=243.83)	295 (E=295)
Negative	66 (E=80.83)	400 (E=385.17)	466 (E=466)
Total	132 (E=132)	629 (E=629)	761 (E=761)

The chi-square statistic is computed as

$$\chi^2 = \sum \left(\frac{(observed - expected)^2}{expected} \right)$$

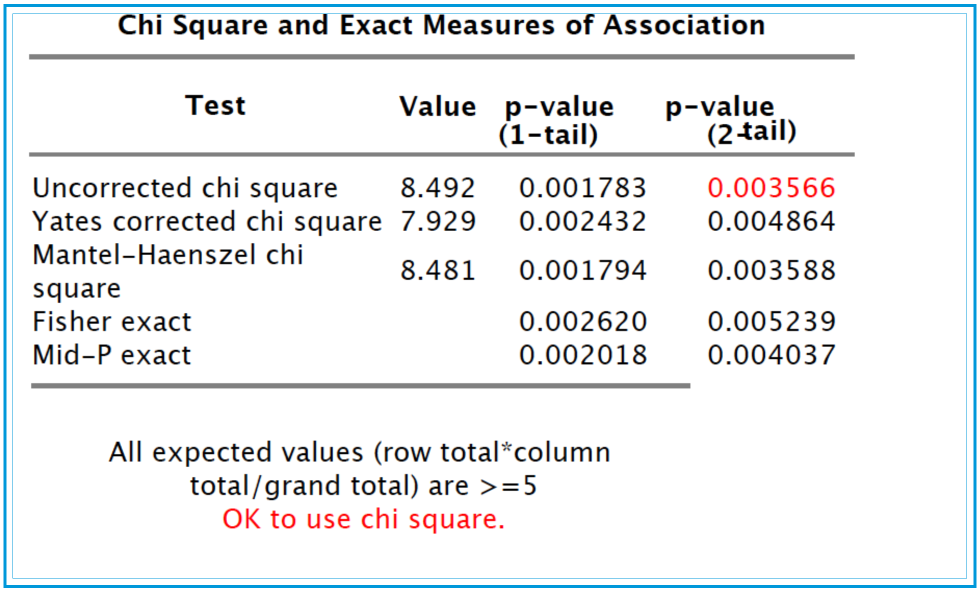
, where the sum is over all cells in a table, not including the total row and column. For each cell the expected value is the product of the number of cases in its row and its column, divided by the total sample size. The expected value for the first cell is (295 x 132)/761=51.17.

In this example,

$$\chi^2 = \frac{(66-51.17)^2}{51.17} + \frac{(229-243.83)^2}{243.83} + \frac{(66-80.83)^2}{80.83} + \frac{(400-385.17)^2}{385.17} = 8.492$$

For each cell in the table you can calculate its residual, the difference between the observed and expected counts. These residuals can be standardized in various ways to remove the effects of different sample sizes in each cell. By examining these residuals you can determine which cells have the largest discrepancies between the observed and the expected counts.

Figure 5.10: Chi-squareTest and Measures of Association



Chi Square and Exact Measures of Association

Test	Value	p-value (1-tail)	p-value (2-tail)
Uncorrected chi square	8.492	0.001783	0.003566
Yates corrected chi square	7.929	0.002432	0.004864
Mantel-Haenszel chi square	8.481	0.001794	0.003588
Fisher exact		0.002620	0.005239
Mid-P exact		0.002018	0.004037

All expected values (row total*column total/grand total) are >=5
OK to use chi square.

Figure 5.10 contains an assortment of tests from OpenEpi used to test the hypothesis that HIV status and LBW are independent, based on the data shown in Figure 5.7. Statisticians debate the merits of different tests, especially for small samples. The most straightforward approach is to use the uncorrected chi-square value of 8.492 and the corresponding two tailed *p*-value of 0.003566.

You can reject the null hypothesis that the true proportion of LBW infants is the same for mothers with and without HIV. Low birth weight and maternal HIV status do not appear to be independent. Below the table you use a note that tells you it's OK to use the chi-square test because the expected values in all cells of the table are greater than or equal to 5.

The value of the chi-square statistic doesn't tell you anything about the strength of the relationship between two variables since its value depends on the sample size. In the section Measuring Association you'll return to this table and calculate statistics which quantify the strength of the relationship between the two variables.

Assumptions for the Chi-Square Test

The chi-square test requires that all of the observation be independent. Each person contributes only one count to a table. For example, if you're comparing treatments you can't give two different drugs to the same person and then include them in the counts for both treatments. Another requirement for the chi-square tests is that most of the expected counts be greater than 5 and none less than 1. That's why in Figure 5.10 you see the note that all of the cells have expected values greater than 5 and it's OK to use the test. The software doesn't have any way of detecting if your observations are independent. You're responsible for that.

Exercise: Chi-square tests are often used to check whether two or more groups are comparable on a set of characteristics. For example you may want to see whether the people who refused to participate in a study are different in important ways to those who agreed to participate. Or to determine whether the people who drop out of a study are different from those who remain. Figure 5.11 is a comparison of women who remained in the Laar study all the way to delivery with women who were lost to follow-up. Why do you think such a comparison is important? The P value is the observed significance level from a chi squared test. What is the null hypothesis that is tested? For which variables do the two groups differ? Do you think any off these differences are potentially important? Can you think of any problems with the use of the chi square statistic for this purpose? (Hint: consider the effect of sample size. For large studies is it possible that even small differences between groups will have small *p*-values?)

Figure 5.11: Comparison of women who remained in study and those lost to followup

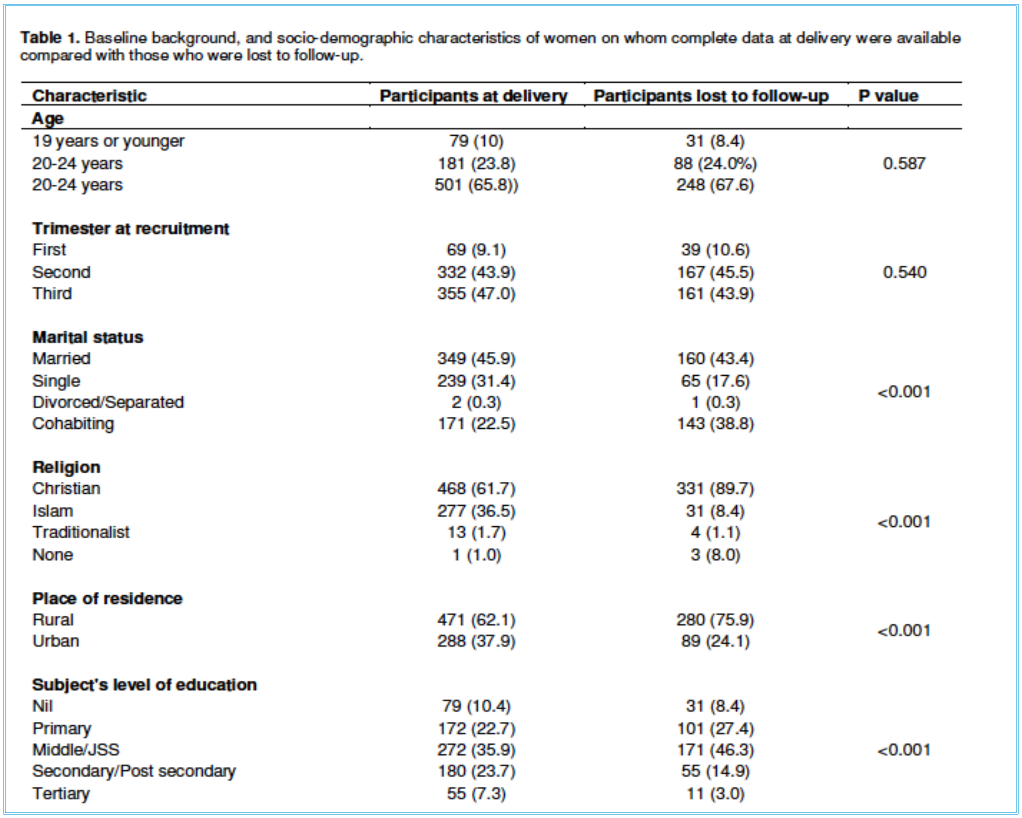


Table 1. Baseline background, and socio-demographic characteristics of women on whom complete data at delivery were available compared with those who were lost to follow-up.

Characteristic	Participants at delivery	Participants lost to follow-up	P value
Age			
19 years or younger	79 (10)	31 (8.4)	0.587
20-24 years	181 (23.8)	88 (24.0%)	
20-24 years	501 (65.8))	248 (67.6)	
Trimester at recruitment			
First	69 (9.1)	39 (10.6)	0.540
Second	332 (43.9)	167 (45.5)	
Third	355 (47.0)	161 (43.9)	
Marital status			
Married	349 (45.9)	160 (43.4)	<0.001
Single	239 (31.4)	65 (17.6)	
Divorced/Separated	2 (0.3)	1 (0.3)	
Cohabiting	171 (22.5)	143 (38.8)	
Religion			
Christian	468 (61.7)	331 (89.7)	<0.001
Islam	277 (36.5)	31 (8.4)	
Traditionalist	13 (1.7)	4 (1.1)	
None	1 (1.0)	3 (8.0)	
Place of residence			
Rural	471 (62.1)	280 (75.9)	<0.001
Urban	288 (37.9)	89 (24.1)	
Subject's level of education			
Nil	79 (10.4)	31 (8.4)	<0.001
Primary	172 (22.7)	101 (27.4)	
Middle/JSS	272 (35.9)	171 (46.3)	
Secondary/Post secondary	180 (23.7)	55 (14.9)	
Tertiary	55 (7.3)	11 (3.0)	

Testing for Independence in Larger Tables

You can use the chi-square statistic to test the null hypothesis that any number of rows and columns of a table are independent. The idea is the same as for a table with two rows: you compute expected counts based on the assumption of independence and then you compare the observed and expected counts.

In a paper studying the relationship between knowledge and health seeking behavior Karunamoothi and Kumera (2010), interviewed 228 Ethiopians about the measures they take to prevent malaria.

Figure 5.12 is a table from the paper showing the relationship between the educational level of the respondents and the primary type of preventive measures observed. To test whether the two variables are independent the authors used the chi-squared test. Their calculated chi-square value is 58.7, with 16 degrees of freedom and an observed significance level of p<0.001. (For a table the degrees of freedom for the chi-square statistic is the product of the number of rows minus 1 times the number of columns minus one. Since there are five rows and five columns the degrees of freedom are 16.) The authors reject the null hypothesis that education and type of preventive measure used are independent.

Figure 5.12: Preventive Measures and Educational Level

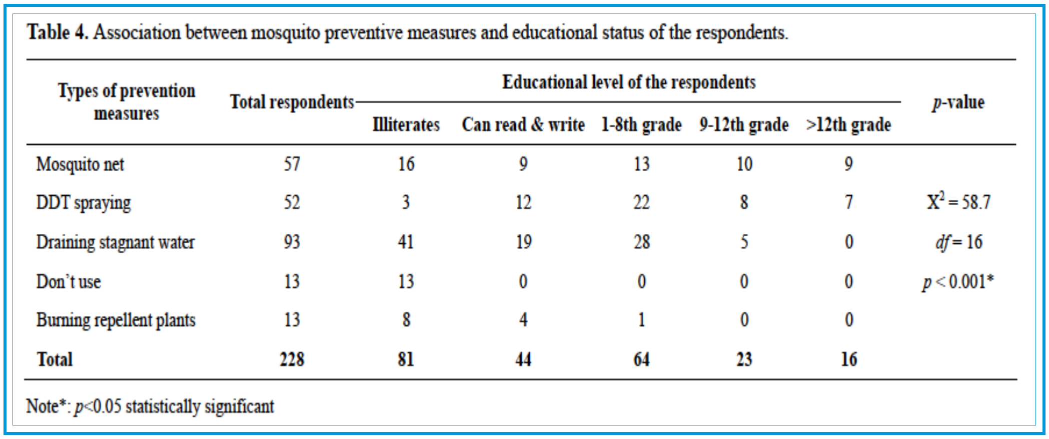


Table 4. Association between mosquito preventive measures and educational status of the respondents.

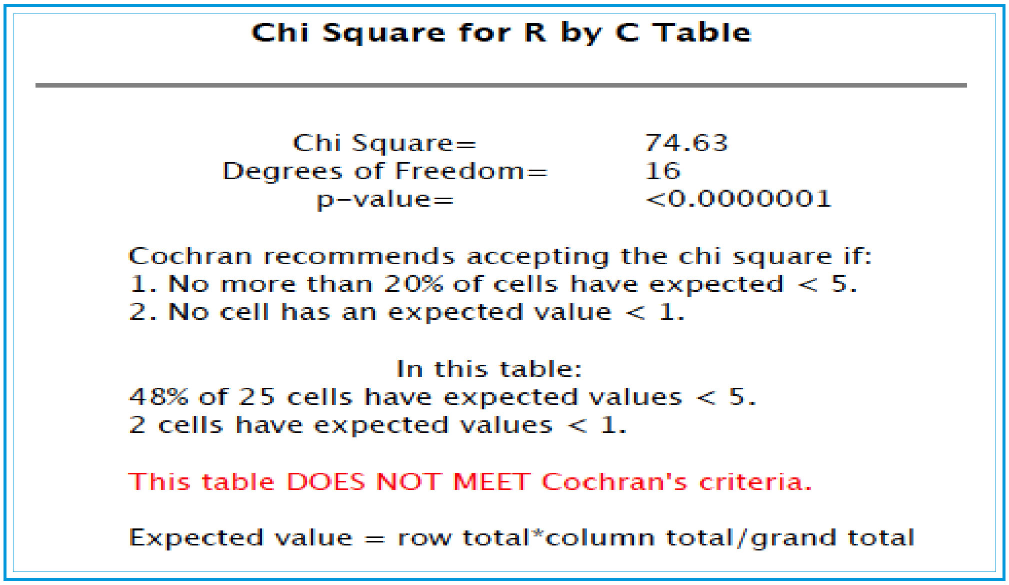
Types of prevention measures	Total respondents	Educational level of the respondents					*p*-value
		Illiterates	Can read & write	1-8th grade	9-12th grade	>12th grade	
Mosquito net	57	16	9	13	10	9	
DDT spraying	52	3	12	22	8	7	$X^2 = 58.7$
Draining stagnant water	93	41	19	28	5	0	$df = 16$
Don't use	13	13	0	0	0	0	$p < 0.001$*
Burning repellent plants	13	8	4	1	0	0	
Total	**228**	**81**	**44**	**64**	**23**	**16**	

Note*: $p<0.05$ statistically significant

Exercise: Compute the appropriate row or column percents for the table. Explain your choice. Calculate the expected number of cases in the cells that corresponds to illiterates who use mosquito nets. (Multiply the total number of cases who are illiterates by the total number of cases who use mosquito nets and then divide by the total sample size.) What is the difference between the observed count for the cell and the expected count?

Figure 5.13 shows the results from the OpenEpi software when the data from Figure 5.12 are entered into the calculator. The first item to notice is the warning that the table violates the assumptions required. Most of the expected cell count are not greater than 5 and 2 cells have expected values less than 1. The rule of thumb is that most cells should have expected values greater than 5 and no cells should have expected values less than one.

Figure 5.13: Chi square results from Open Epi software



```
              Chi Square for R by C Table

                Chi Square=              74.63
           Degrees of Freedom=           16
                 p-value=                <0.0000001

        Cochran recommends accepting the chi square if:
        1. No more than 20% of cells have expected < 5.
        2. No cell has an expected value < 1.

                       In this table:
        48% of 25 cells have expected values < 5.
        2 cells have expected values < 1.

        This table DOES NOT MEET Cochran's criteria.

        Expected value = row total*column total/grand total
```

If the assumptions are violated the observed significance level may be wrong. In particular small expected counts may cause the chi-square statistic to be too big. The next item to notice is the value of the chi-square statistic reported by Open Epi. It's quite different than that reported by the authors. Both Open Epi and the authors report the same degrees of freedom so the size of the table that's being analyzed is the same. Remember, just because it's published doesn't mean it's correct!

Exercise: If a respondent could chose more than one preventive measure for protecting themselves from malaria, could the data still be analyzed using a chi square test for independence? What would happen to the sample size in the table if the same person could appear several times? What do you suggest the authors do so the table can be analyzed using the chi-square test?

Measuring Association in 2 by 2 Tables

You used the chi-square statistic to test the null hypothesis that the proportion of low birth weight infants is the same for women with and without HIV. Based on the chi-square value you rejected the null hypothesis. Can you conclude anything about the strength of the relationship between HIV status and low birth weight infants based on the chi square value? Does a big chi -square value mean that the relationship is strong? The answer is no. The value of the chi-square statistic depends not only on the magnitude of the observed differences but also on the sample size in your table. If you multiply all cell entries in a table by 2, you double the value of the chi -square statistic. The relationship between the two variables remains the same but the value of the chi-square statistic doesn't.

In medicine and public health quantifying the strength of the relationship between a dichotomous predictor variable (risk factor) and an outcome, such as death, recurrence, or low birth weight, is very important. It's not enough to say HIV significantly increases the chances of a low birth weight baby. You want to answer the question *By how much?* Before you can answer that question you have to consider how the data were obtained.

Data often originate from one of these three different experimental designs:

- A **cohort** study in which you follow two groups (one with the risk factor and the other without) and record how often the event of interest occurs in each group. For example, you take samples of pregnant HIV positive and HIV negative women and follow the two groups until delivery. You then calculate the proportion of low birth weight infants in each group.
- A **cross-sectional** study in which you take a random sample of individuals and count how many fall into each of the four cells of a two way table. For example, you follow all pregnant women coming to your clinic and record whether they have HIV and whether they delivered a low birth weight infant.

- A **case-control** study in which you examine a group of individuals who have experienced the event of interest and a group who have not. For each of the individuals, you record whether the risk factor was present or not. For example, you take a sample of low birth weight infants and a sample of normal birth weight infants and record how many in each group had mothers with HIV.

Calculating the Relative Risk Ratio

In a cohort study or a cross-sectional study you can compute the incidence rate for the event of interest for cases with and without the risk factor. That's not possible in a case control study. Consider Figure 5.14 which is the same table you analyzed before. The Laars study was a cohort study which observed two groups of pregnant women, those with HIV and those without.

Figure 5.14: HIV Status and Low Birth Weight

	Low Birth Weight Infant (event)		
HIV Status (risk factor)	**Yes N (%)**	**No N (%)**	**Total N (%)**
Positive	66 (22.4%)	229 (77.6%)	295 (100%)
Negative	66 (14.2%)	400 (85.8%)	466 (100 %)
Total	132	629	761

The incidence of LBW infants in the HIV positive women group is 22.4%. The incidence in the HIV negative women is 14.2%. The **relative risk ratio** is the ratio of the two incidence rates. The relative risk ratio is

$$Relative\ Risk\ Ratio=\frac{(incidence\ of\ the\ event\ for\ the\ group\ with\ the\ risk\ factor)}{(incidence\ of\ the\ event\ for\ the\ group\ without\ the\ risk\ factor)}$$

$$RR=\frac{22.4}{14.2}=1.58$$

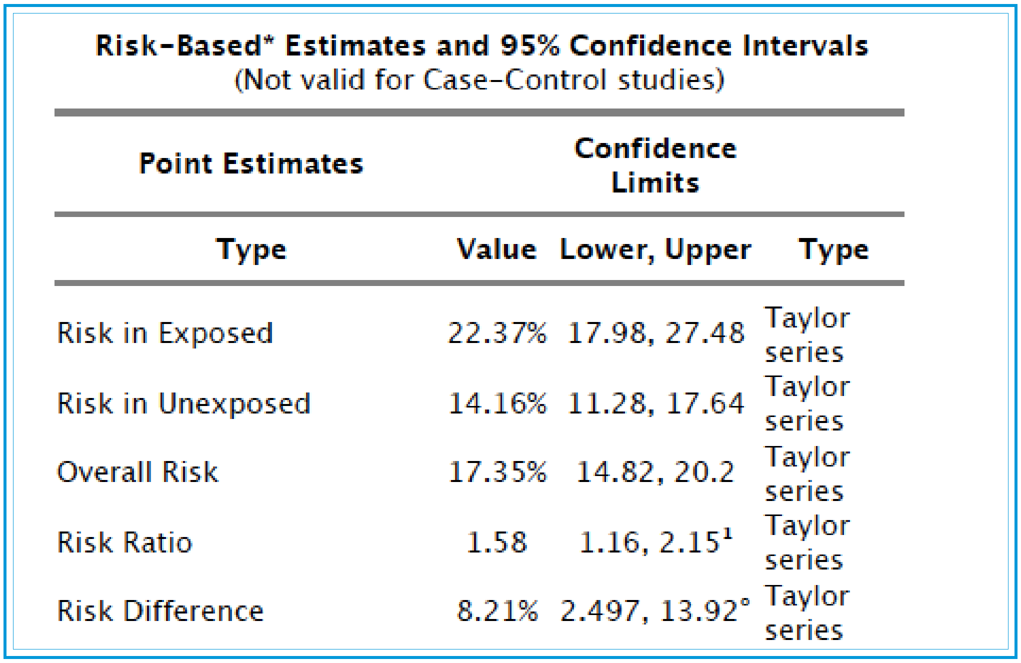
Women who are HIV positive are almost 1.6 times as likely to have a low birth weight infant as women without HIV. To put it another way, the risk of a low birth weight infant is 58% higher in women with HIV as compared to women without HIV.

Unlike the chi square statistic whose value depends on the sample size, the relative risk ratio is easy to interpret and can be compared across factors and different studies:

- If the relative risk ratio is 1, the risk factor and the outcome are independent
- If the relative risk ratio is greater than 1, people with the risk factor are more likely to experience the event than people without the risk factor.
- If the relative risk ratio is less than 1, people with the risk factor are less likely to experience the event than people without the risk factor.

Figure 5.15 is output from OpenEpi for risk based estimates, together with their confidence intervals. You see that the risk ratio is 1.58 with a 95% confidence interval of 1.16 to 2.15. Since the value of 1 is not included in the interval, you can reject the null hypothesis that the true value is 1 (no association). It appears that HIV positive women are more likely to have low birth weight infants than women without HIV.

Figure 5.15: Risk Based Estimates from Open Epi



Risk-Based* Estimates and 95% Confidence Intervals
(Not valid for Case-Control studies)

Point Estimates		**Confidence Limits**	
Type	**Value**	**Lower, Upper**	**Type**
Risk in Exposed	22.37%	17.98, 27.48	Taylor series
Risk in Unexposed	14.16%	11.28, 17.64	Taylor series
Overall Risk	17.35%	14.82, 20.2	Taylor series
Risk Ratio	1.58	1.16, 2.15[1]	Taylor series
Risk Difference	8.21%	2.497, 13.92°	Taylor series

The relative risk ratio doesn't tell you anything about the actual rates. The relative risk ratio is 2 regardless of whether the underlying rates are 40% and 20% or 2% and 1%. That's why it's useful to look at the actual difference between the rates. The difference in low birth weight rates between the two groups in this example, 8.21%, is labeled *Risk Difference* in Figure 5.15. The 95% confidence interval for the true difference is (2.5 to 13.9). The interval does not include 0, so you can reject the null hypothesis that the true difference in risk is 0. Testing that the risk difference is 0 is the same as testing that the relative risk ratio is 1.

Calculating the Odds Ratio

If your data are from a case-control study, you can't calculate the relative risk ratio because you can't calculate the probability that someone with and without the risk factor experiences the event. If Laars had taken a sample of low birth babies and a sample of normal birth weight babies and counted the number of HIV positive mothers in each group, he couldn't calculate the incidence rate of LBW since he decided how many cases of low birth weight to include in the study.

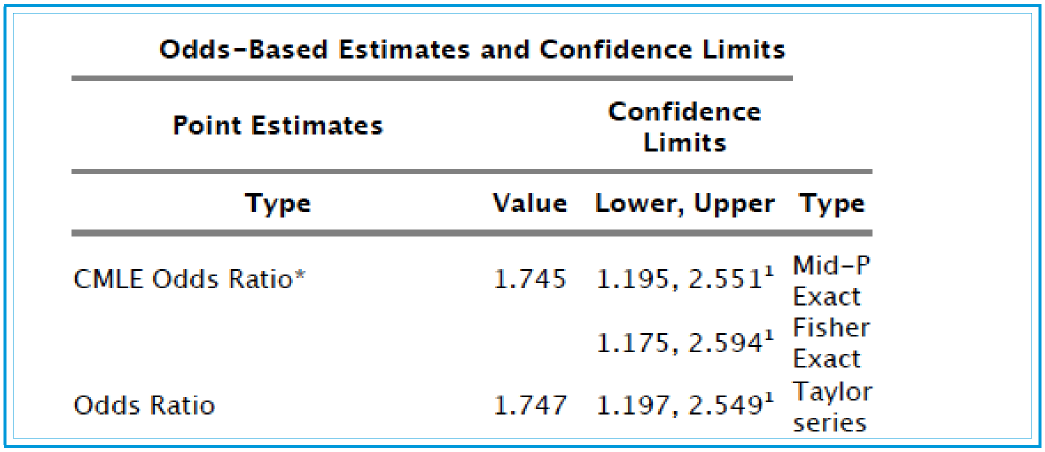
For a case-control study, the **odds ratio** is used to measure the relationship between the event and the risk factor. The odds ratio is the ratio of two odds: the odds that a case has the risk factor and the odds that the control has the risk factor. (The odds is the number of cases with the risk factor divided by the number of cases without the risk factor.) For this example:

- The odds that a LBW infant has a mother with HIV are 66/66= 1. If you choose a LBW infant, based on Figure 5.14, it's just as likely to have a mother with HIV as a mother without HIV.
- The odds that a normal weight infant has a mother with HIV are 229/400=0.573
- The odds ratio is 1/0.573=1.747

For a cohort or longitudinal study you can compute both an odds ratio and a relative risk ratio.

Several version of the odds ratio and its confidence limits, based on minor differences in statistical approaches, are shown in Figure 5.16. The 95% confidence interval for the odds ratio ranges from (1.197 to 2.549) and does not include the value of 1, so you can reject the null hypothesis that the true value is 1. There appears to be an association between HIV status and LBW. The odds of a LBW infant are 1.75 times greater for an HIV positive woman than the odds for an HIV negative woman. Sometimes Cochran's test is used to determine the observed significance level for the test that an odds ratio is 1. For a two by two table Cochran's test of conditional independence is just the uncorrected chi-square value in Figure 5.1.

Figure 5.16: Odds-Based Estimates from Open Epi



Odds-Based Estimates and Confidence Limits

Point Estimates		Confidence Limits	
Type	**Value**	**Lower, Upper**	**Type**
CMLE Odds Ratio*	1.745	1.195, 2.551[1]	Mid-P Exact
		1.175, 2.594[1]	Fisher Exact
Odds Ratio	1.747	1.197, 2.549[1]	Taylor series

If the event of interest occurs infrequently (less than 10% of the time) and if the total sample size is large, the odds ratio from a case control study can be used as an estimate of relative risk. However, odds ratios and relative risk ratios can differ a lot when the event is common. The odds ratio will be larger than the relative risk ratio if the ratio is larger than 1 and smaller if the ratio is less than 1. In this example, low birth weight is not a rare event so the odds ratio is larger than the relative risk ratio. As expected, both indicate that low birth weight is related to maternal HIV status. The odds ratio is more difficult to interpret than the relative risk ratio since it can't be expressed in terms of probabilities.

Exercise: In her abstract Ezeaka states:

> There were three times more LBW babies in the HIV-positive group than in the uninfected mothers (odds ratio = 3.47, 95% confidence interval = 1.69, 7.27; **chi(2) = 12.99**, P = 0.0003).

Can you think of reasons why her odds ratio for HIV is larger than that found by analyzing the Laars data? Remember that Laar did not exclude women who had malaria at recruitment. What groups would you analyze in the Laar data to estimate the effect of HIV status alone?

Exercise: Read the abstract of the Laar paper given below.

Preterm delivery and low birth weight among neonates born to HIV-positive and HIV-negative Ghanaian women
Laar A. K.[1]*, Ampofo W.[2], Tuakli J. M.[1], Norgbe G. K.[1] and Quakyi I. A.[3]

In sub-Saharan Africa, several hundreds of pregnancies are exposed to both malaria and HIV infections annually. Adverse perinatal outcomes as a result of these infections include preterm delivery (PTD), and low birth weight (LBW). These are not well characterized in Ghana. We determined whether malaria and HIV infections during pregnancy increase the risk of delivering a preterm or a LBW neonate. We enrolled 1,154 women at their first antenatal visit (443 HIV-positive and 711 HIV-negative), and prospectively collected data at delivery on 761 mother-infant pairs. Malaria parasitemia status, HIV status, hemoglobin concentration, and CD4+ cell count were determined using standard methods. We observed a significantly increased risk of LBW among HIV positive women with malaria at recruitment, odds ratio **(OR) = 4.4, 95% Confidence Interval [CI] (2.3 to 8.4),** at delivery, **OR = 2.5, 95% CI (1.1 to 3.7).** The risk among those who were dually-infected at recruitment and at delivery was more pronounced; **OR = 11.3; 95% CI (4.6 to 27.4)**. Dual infection was also associated with a 4-fold risk of delivering preterm; **OR = 3.96; 95% CI (1.8 to 8.5).** These findings demonstrate that neonates of HIV-positive women with multiple malaria infections are at particular risk of PTD and LBW in Ghana**.**

For each odd ratio given in the abstract, describe the two by two table it is based on. That is, what variable forms the rows of the table, what variable forms the columns, and what types of cases are included in the table. For which odds ratio can you reject the null hypothesis that the true value is 1? What do you base your decision on? Why are the confidence intervals widest for the dual infection cases? What type of study is this? Could the authors have calculated the relative risk ratio?

Summary

Statistical procedures are used to estimate observed significance levels for tests of the null hypothesis. The choice of procedure depends on the null hypothesis and on the characteristics of the data. It's important to check whether your data meet the assumptions required for a particular procedure. Otherwise the test results may not be correct. The *t*-test is used to test the null hypothesis that two population means are equal. The chi-square test is used to test that the rows and columns of a table are independent.

The relative risk ratio and the odds ratio quantify the association between a risk factor and an event. A ratio of 1 indicates that there isn't any association between the risk factor and the event. A ratio greater than 1 indicates that people with the risk factor are more likely to experience the event that people without the risk factor. A ratio less than 1 indicates that people with the risk factor are less likely to experience the event than people without the risk factor.

Bibliography:

Ezeaka, VC, Iroha EO, Akinsulie AO, Temiye EO, Adetifa IM. Anthropometric indices of infants born to HIV-1 infected mothers: a prospective cohort study in Lago, Nigeria. I*nt J STD AIDS*, 2009: 20(8):545-8. Ezeaka (2009)

Floridia M., Ravizza M.,Bucceri A., Lazier L., Vigano A., Alberico S., Guaraldi G. et al. Factors Influencing Gestational Age-Adjusted Birthweight in a National Series of 600 Newborns from Mothers with HIV. *HIV Clin Trials,* 2008 9(5):287-97. Floridia (2008)

Laar AK, Ampofo W., Tuakli JM, Norgbe GK and Quakyi IA, Preterm delivery and low birth weight among neonates born to HIV-positive and HIV-negative Ghanaian women. J*ournal of Public Health and Epidemiology,* 2010: Vol. 2(9), 224-237. Laar (2010)

Karunamoorthi, K. and Kumera, A. Knowledge and health seeking behavior for malaria among the local inhabitants in an endemic area of Ethiopia: implications for control. *Health,* 2010*:*2; 575-581. Karunamoorthi(2010)

6 Correlation and Linear Regression

- What is a correlation coefficient?
- What is linear regression?
- What does the slope tell you? The intercept?
- What is a residual?
- How can you tell how well a regression model fits the data?
- What are prediction intervals?

The word “correlation” is a vague term used in everyday conversation to describe some type of relationship between variables. You know that the amount you eat is correlated with the amount that you weigh; that how hard you work in school is correlated with how successful you'll be in life (well, maybe!), that infant gestational age is related to infant weight.

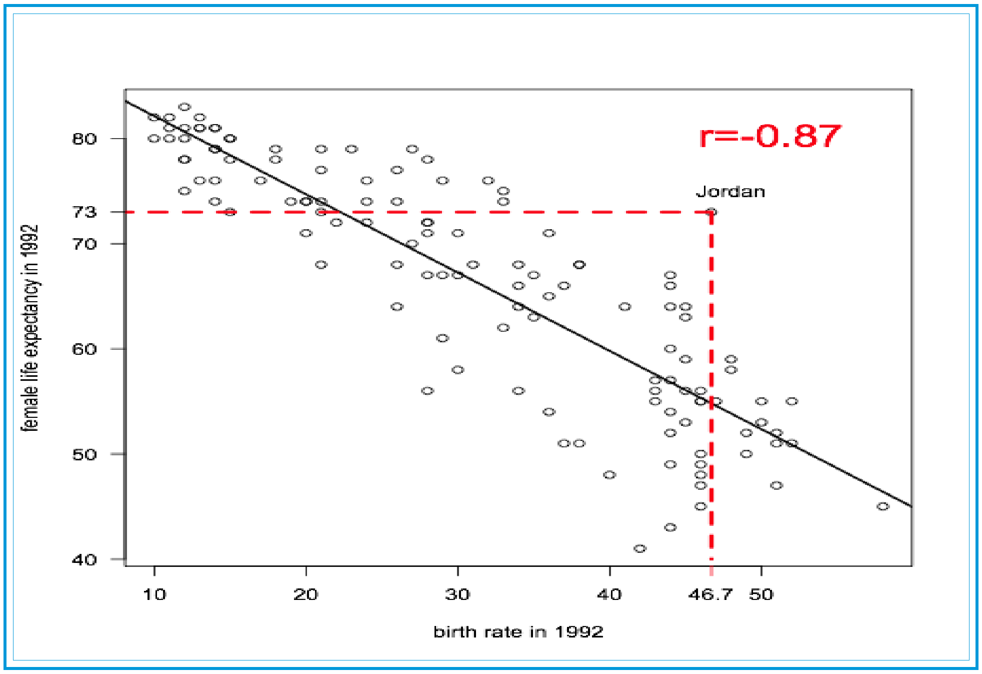
In statistics, correlation has a precise definition. It's a measure of the strength of the *linear* relationship between two variables. In this chapter you'll examine relationships between pairs of variables by plotting them and then, if the relationship is linear, you'll calculate the Pearson correlation coefficient as a summary measure. You'll also use linear regression to build a simple model to predict the values of a dependent variable based on the values of a single predictor variable. The next chapter introduces more complicated models.

Plotting Data

The first step in examining the relationship between two variables is to make a scatterplot, such as the one shown in Figure 6.1 for female life expectancy and birth rate in 122 countries. Each point in the plot represents a country. The point for *Jordan* is labeled and corresponds to a birth rate of 46.7 births per 1000 population (plotted on the horizontal or *x* axis) and a life expectancy of 73 years (plotted on the vertical or *y* axis).

You see that the points don't appear to be randomly scattered over the grid. Instead there is a pattern, a negative relationship between the two variables, since birth rate increases as female life expectancy decreases. If you had to describe the relationship between the two variables you might say it's linear, since points loosely scatter around a straight line.

Figure 6.1: Scatterplot of Female Life Expectancy and Birth Rate

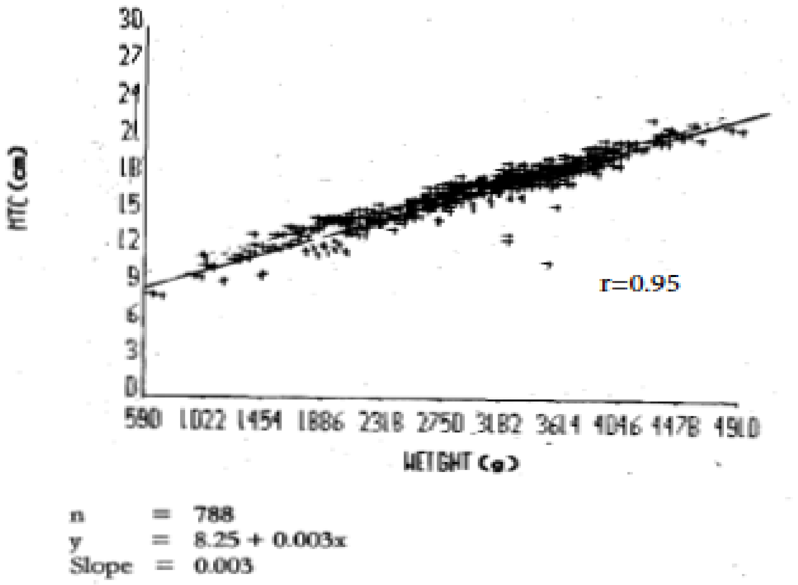


Predictors of Birth Weight

When variables are related to one another, it may be possible to use the values of one variable (the predictor or independent variable) to predict the values of another (the dependent variable). For example, in the absence of scales for weighing infants, easily available anthropometric measurements may serve as substitutes for identifying low birth weight infants who need further care; or maternal measurements during labor may predict birth weight and help to identify women who may need medical intervention during delivery. Usually the dependent variable is plotted on the vertical (y) axis and the independent variable is plotted on the horizontal (x) axis.

Ezeaka et al.(2003) examined the relationship between several infant measurements and birth weight. Figure 6.2 is a plot of infant maximum thigh circumference (MTC) and birth weight. The relationship between birth weight and MTC is said to be positive since as birth weight increases so does MTC. The points cluster closely around a straight line, suggesting that MTC may be a good predictor of birth weight.

Figure 6.2: Plot of Maximum Thigh Circumference and Birth Weight



Pay attention to the handful of points that are far removed from the rest. The individual birth weights and MTC values are not unusual and would not raise suspicion on a histogram. However, when the two values are considered together they're unusual. When analyzing data you should always check that outlying points are not the result of data entry errors. If the points are correct you should investigate whether the infants were in some way different from the rest, such as premature infants. Do not, however, remove the points from your analysis unless they are clearly wrong.

Buchmann et al. (2009) studied the relationship between symphysis- fundal height (SFH) and birthweight. The goal was to develop a simple formula to predict fetal weight at delivery. In Figure 6.3 you again see a positive relationship between SFH and birthweight, with the points clustering along a straight line. However, the points in Figure 6.3 don't cling as tightly to the line as the points do in Figure 6.2.

Figure 6.3: Symphysis-fundal Height and Birth Weight

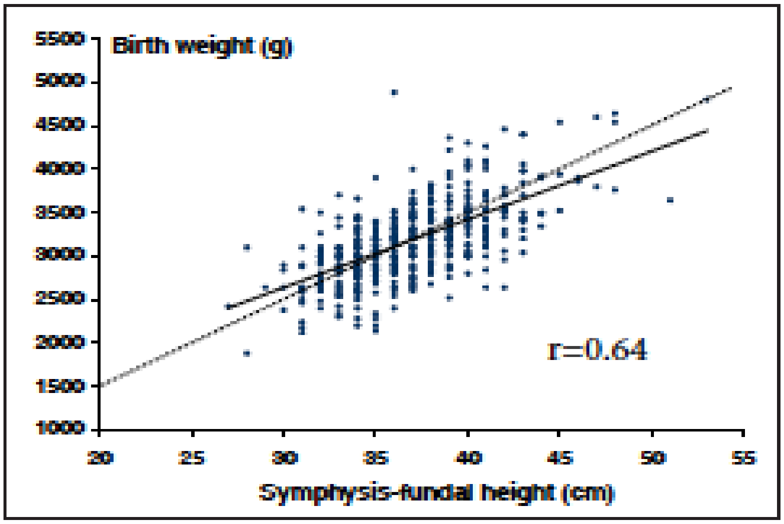


Fig. 1. Scatter plot for birth weight by symphysis-fundal measurement in the derivation study (N=504). The solid line represents the linear regression formula produced by the data (y=301+78x). The dotted line represents the simplified formula derived for clinical use (birth weight in g=100 ([SFH in cm]–5).

If you want to compare variables that may serve as predictors of infant birth weight you need to compute a measure which quantifies how strongly each of the predictor variables is related to birth weight.

The Pearson Correlation Coefficient

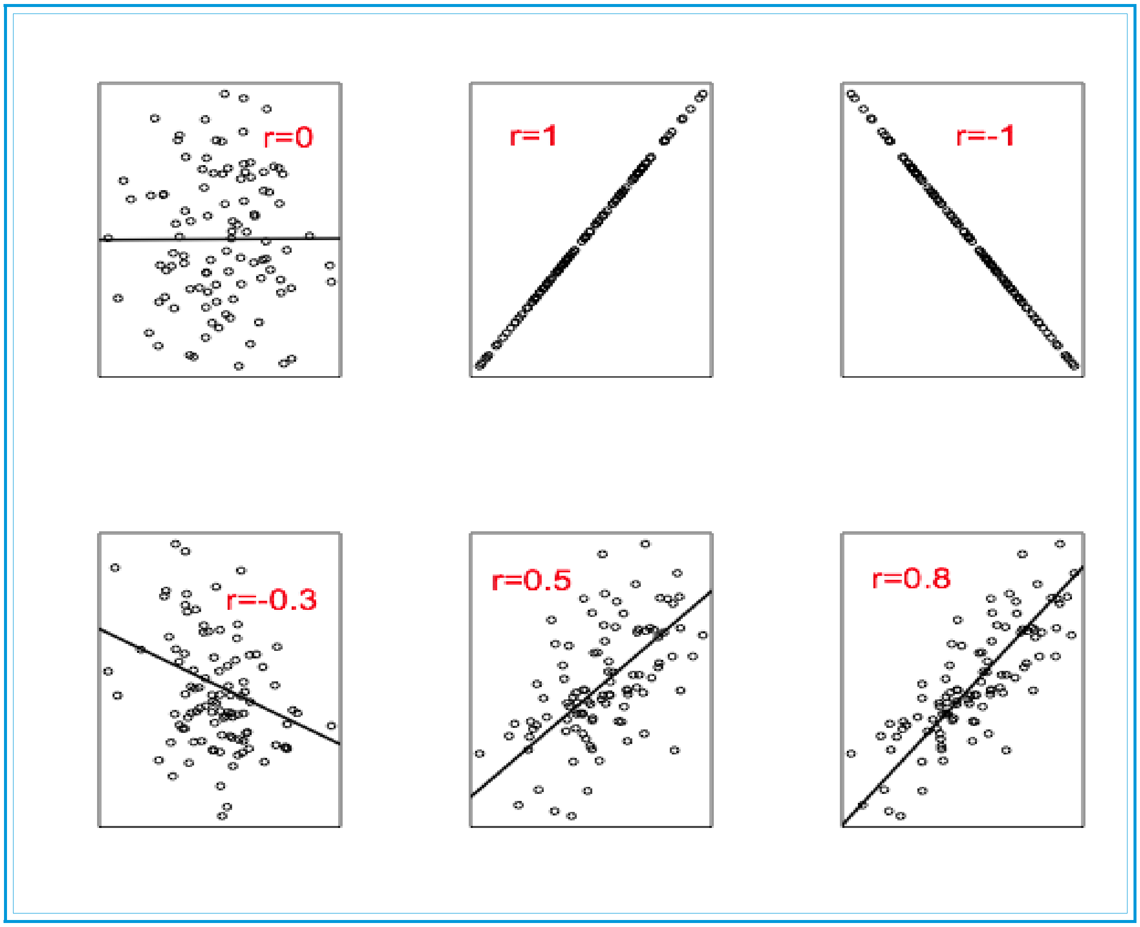
The **Pearson correlation coefficient** (usually denoted as *r)* is a statistic that measures the strength and direction of the *linear* relationship between two variables. The correlation coefficient does *not* tell you anything about the cause and effect relationship between the variables. Just because two variables are correlated that doesn't mean that one causes another.

The Pearson correlation coefficient:

- Ranges in value from -1 to +1. The absolute value of the correlation coefficient tells you how strongly the variables are linearly related. A value of either +1 or -1 means that you can perfectly predict the values of one variable from the values of the other. If the correlation coefficient is +1, all points fall on a line with values of both variables increasing together. If the correlation coefficient is -1, all points fall on a line but as values of one variable increase the values of the other variable decrease.
- Is 0 when there is no *linear* relationship between two variables.
- Is symmetric, since the correlation between birth weight and MTC is the same as the correlation between MTC and birth weight. When computing a correlation coefficient, neither variable is singled out as a dependent variable.
- Stays the same if you add a constant to all values or divide all the values by a constant
- When squared is the proportion of the variance of one variable that can be "explained" by the other.

Figure 6.4 shows plots, correlation coefficients and summary lines for correlations of different sizes. Remember that it is the absolute value of the correlation coefficient that tells you the strength of the linear relationship. Variables with a positive 0.5 correlation coefficient have as much of a linear relationship as variables with a negative 0.5 correlation coefficient.

Figure 6.4: Scatterplots with Correlation Coefficients

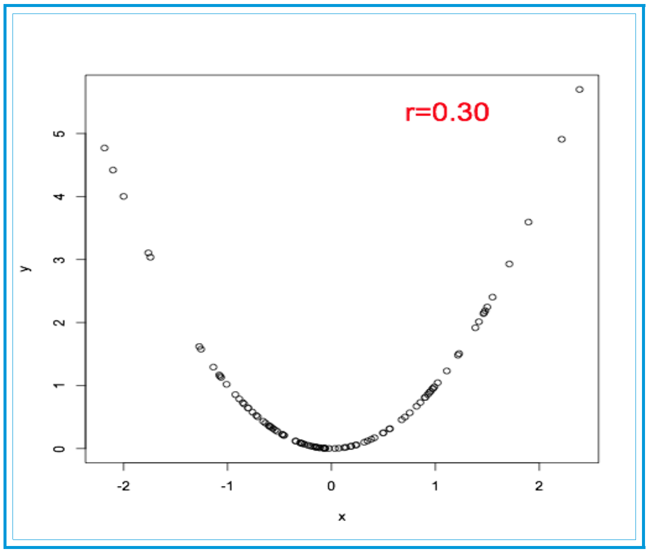


The correlation coefficient is also shown on the first three figures in the chapter. For Figure 6.1 the correlation coefficient (-0.87) is negative since as the birth rate increases female life expectancy decreases. For the other two figures the correlation coefficient is positive since as birth weight increases so do MTC and SFH. The largest correlation coefficient in absolute value (r=0.95) is for birthweight and MTC since the points cluster very tightly along the straight line. The correlation coefficient between birth weight and SFH is smaller (r=0.64).

Warnings About the Correlation Coefficient

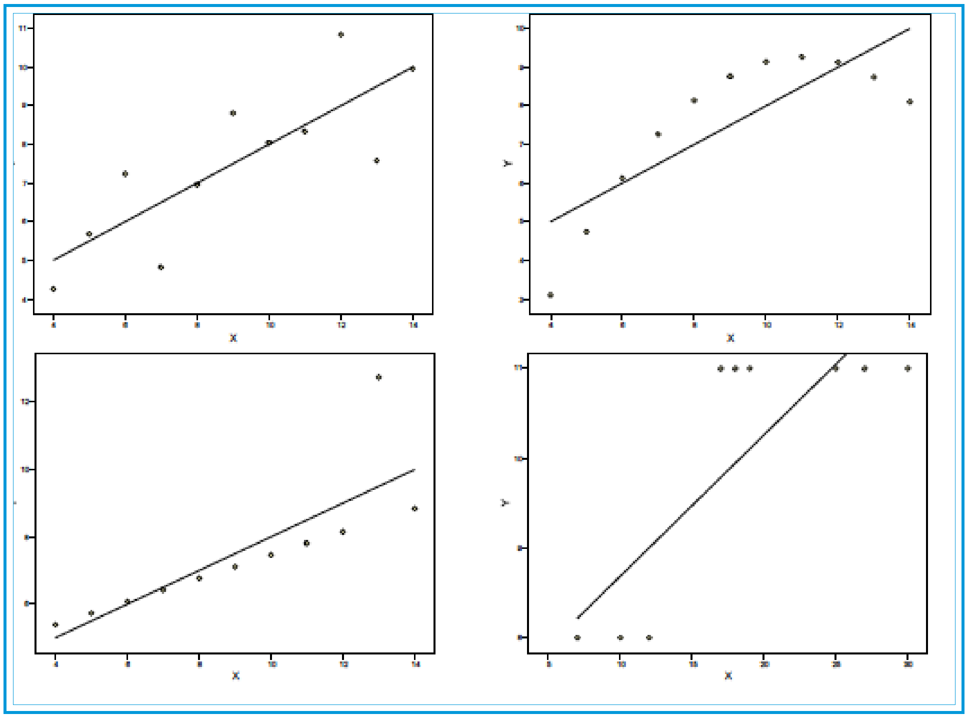
The correlation coefficient is one of the most popular statistics because of the mistaken belief that it completely describes the relationship between two variables with a single number between -1 and +1. Unfortunately that's not the case. Consider Figure 6.5 which shows a perfect nonlinear relationship between two variables. The correlation coefficient is small, only 0.30, because the correlation coefficient only measures the strength of the *linear* relationship. (Compare Figure 6.5 to the plot with the same correlation of 0.3 in Figure 6.4. The relationships between the two variables are very different.) A small value of the correlation coefficient may mean that two variables are not related in any way or that there is a relationship but it is not linear. That's why the first step should always be to plot the variables and see how they are related. It doesn't make sense to talk about how closely points cluster around a straight line if the line isn't a good summary.

Figure 6.5: Strong Nonlinear Relationship



Many different relationships can result in the same correlation coefficient. Figure 6.6 shows four different plots all with a correlation coefficient of 0.82. The correlation coefficient is an appropriate measure only for the first plot, since a straight line is a reasonable summary of the relationship between the two variables. In the second plot, the relationship between the two variables is not linear, so it doesn't make sense to describe how tightly the points cluster around a straight line. In the third plot, you see that the perfect relationship between two variables is distorted by a single point. It is very important to identify individual points that have a large effect on the correlation coefficient (or any other statistic). In the fourth plot there appear to be two groups of cases in which there is no linear relationship between the two variables. If you combine the two groups on the same plot, there appears to be a relationship. The punch line is clear: If you don't plot your data, you can't tell whether a correlation coefficient is an appropriate summary.

Figure 6.6: Scatterplots with r=0.82



The value of a correlation coefficient also depends on the range of values for which observations are taken. It's possible that even if there is a linear relationship between two variables you won't detect it if you consider only a small range of values of the variables. For example, height may be a poor predictor of weight if you restrict your range of heights to those over 1.8 meters.

Transforming to a Linear Relationship

Sometimes a relationship that appears not to be linear, such as the relationship between female life expectancy and GDP in US dollars as shown in Figure 6.7, can be made linear by transforming one or both of the variables. (When you transform a variable you change the scale on which it's measured. For example you take logs or square roots of the data values.)

Figure 6.8 is a plot of female life expectancy against the log of GDP. The relationship between the variables now appears to be linear. You transform variables because it's easier to work with a linear model rather than a more complicated function.

Figure 6.7: Plot of GDP and Female Life Expectancy

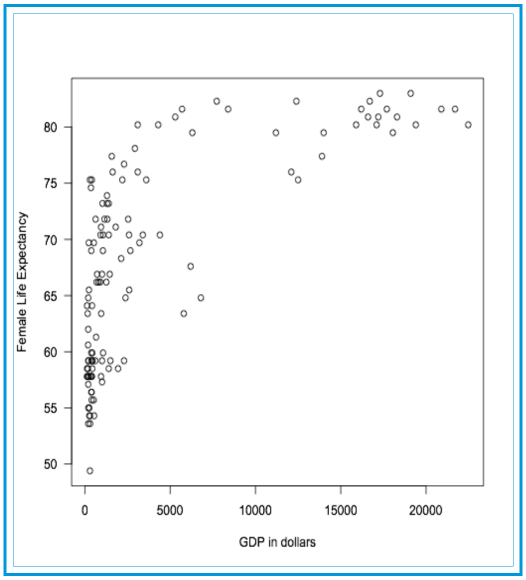
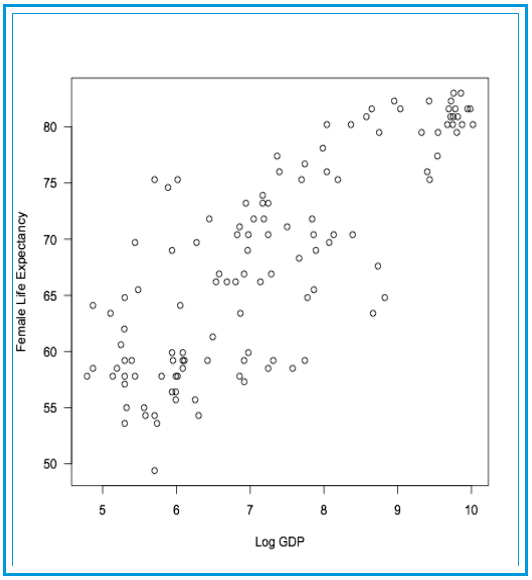


Figure 6.8: Plot of log GDP and Female Life Expectancy



Comparing Two Indexes

Often there are several different tests or measuring devices that purport to measure the same quantity. They may differ in accuracy, cost and perhaps even pain inflicted. Or there may be a "gold standard" that is known to be accurate. You may be faced with the task of comparing the results from the two alternative methods to determine if they can be used interchangeably. Even if the correlation coefficient is a perfect +1 that doesn't mean that the two methods are equivalent. The variables may have different means and standard deviations and have a perfect correlation. Two methods are equivalent if the correlation coefficient is 1 *and* the means and standard deviations are equal. (You can plot the difference between the two measurements against their sum to see how the difference varies over the range of values of the measurement.)

Testing Hypotheses About the Pearson Correlation Coefficient

You can use the correlation coefficient to just describe the observed relationship between two variables in your sample. However, it's more likely that you want to draw conclusions about the population based on the results you've observed in your sample. The symbol for the population correlation coefficient is ρ (rho). The correlation coefficient, like other statistics, has a sampling distribution since different samples from the same population result in different estimated correlation coefficients. To test hypotheses about the population your data must be an independent random sample from a population and, if the sample size is small, the two variables must jointly have a normal distribution.

Figure 6.9 from Ezeaka (2003) shows sample correlation coefficients and confidence intervals for the population correlation coefficient ρ between MTC and various anthropometric variables. You interpret the confidence interval for a correlation coefficients just as you do for a mean or any other statistic. The computation is different but the meaning of the interval is exactly the same for all statistics. For example, you are 95% confident that the true correlation between birth weight and maximum thigh circumference is between 0.94 and 0.96. If the 95% confidence interval does not include the value 0, you can reject the null hypothesis that the population value for the correlation coefficient is 0. The last column of Figure 6.9, labeled *P Value,* is the observed significance level for the test of the null hypothesis that the population value is 0. You see that there is a linear relationship between MTC and the other anthropometric measurements. You can reject the null hypothesis that the population coefficient is 0.

Figure 6.9: Correlations with Maximum Thigh Circumference

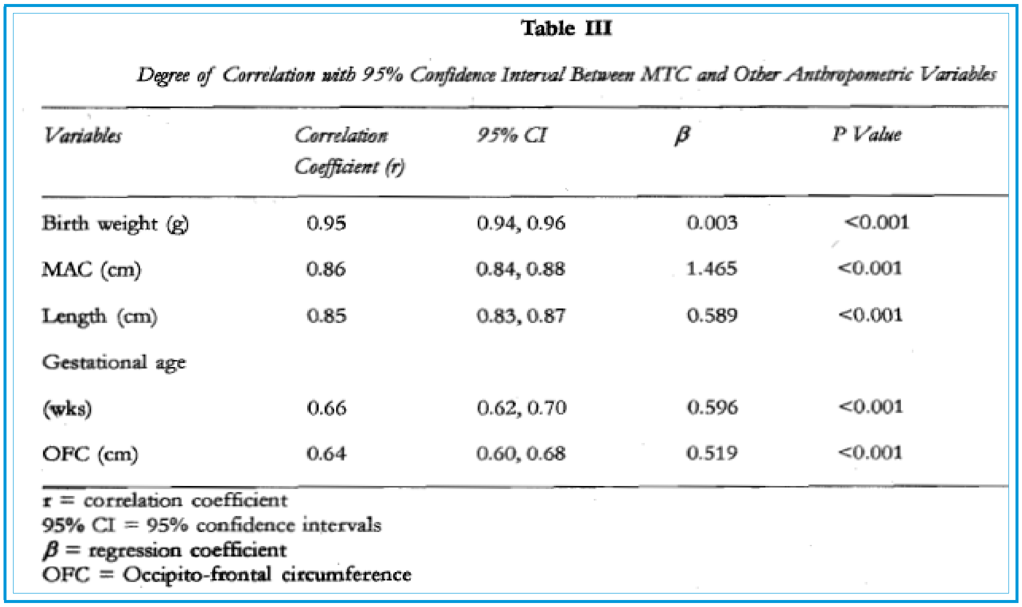


Table III

Degree of Correlation with 95% Confidence Interval Between MTC and Other Anthropometric Variables

Variables	*Correlation Coefficient (r)*	*95% CI*	*β*	*P Value*
Birth weight (g)	0.95	0.94, 0.96	0.003	<0.001
MAC (cm)	0.86	0.84, 0.88	1.465	<0.001
Length (cm)	0.85	0.83, 0.87	0.589	<0.001
Gestational age				
(wks)	0.66	0.62, 0.70	0.596	<0.001
OFC (cm)	0.64	0.60, 0.68	0.519	<0.001

r = correlation coefficient
95% CI = 95% confidence intervals
β = regression coefficient
OFC = Occipito-frontal circumference

Exercise: What explanation can you give as to why the confidence intervals for the correlation coefficients are so narrow? Read the abstract of the paper to see if your theory is correct.

Be sure to look at the magnitude of the correlation coefficients as well as the observed significance level. For large sample sizes, even small correlation coefficients will have small observed significance levels. *Statistically significant doesn't mean important or useful.*

When using statistical software it's tempting to compute correlation coefficients between many pairs of variables to see if something is statistically significant. Remember that even if the variables are not correlated in the population, you expect to find five significant sample coefficients in every 100, if you use a 5% cutoff for determining statistical significance. If you are examining many coefficients, you need a stricter cut off to protect you from rejecting too many times the null hypothesis when it is true.

Exercise: Elizabeth, et al. (2013) conducted a study in Uganda to identify predictors of infant birth weight. Figure 6.10 is a summary of the results:

Figure 6.10: Correlation of anthropometric measures with birth weight

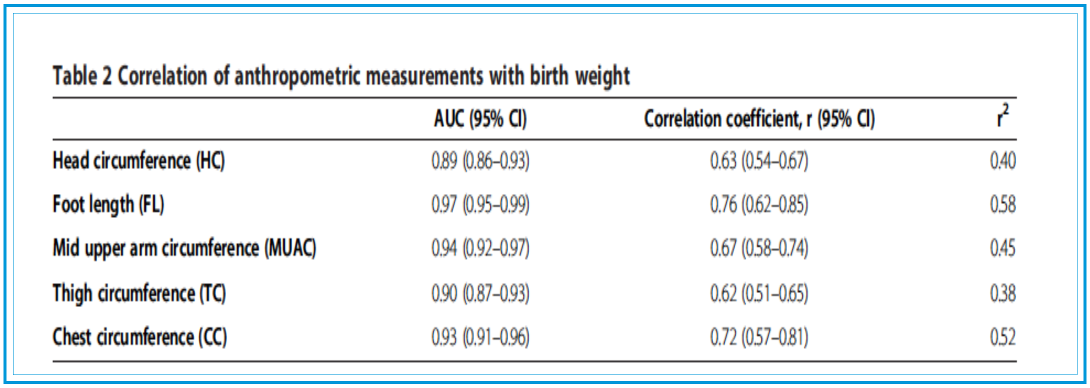


Table 2 Correlation of anthropometric measurements with birth weight

	AUC (95% CI)	Correlation coefficient, r (95% CI)	r^2
Head circumference (HC)	0.89 (0.86–0.93)	0.63 (0.54–0.67)	0.40
Foot length (FL)	0.97 (0.95–0.99)	0.76 (0.62–0.85)	0.58
Mid upper arm circumference (MUAC)	0.94 (0.92–0.97)	0.67 (0.58–0.74)	0.45
Thigh circumference (TC)	0.90 (0.87–0.93)	0.62 (0.51–0.65)	0.38
Chest circumference (CC)	0.93 (0.91–0.96)	0.72 (0.57–0.81)	0.52

Which variable, based on the correlation coefficient, appears to be the best predictor of birth weight? Compare the correlation coefficients with those from the Ezeaka study. What possible explanations can you think of for the differences between the two sets of correlation coefficients?

Correlation Coefficients Based on Ranks

For small sample sizes the Pearson correlation coefficient requires an assumption of normality. For ordinal data, or for interval data that do not satisfy the normality assumption, you can calculate a correlation coefficient that is based on ranks. For each variable you sort the data values from smallest to largest and then assign ranks, where 1 is given to the smallest number, 2 to the second smallest number, and so on. You replace the actual data values with their ranks and then compute the Pearson correlation coefficient which then changes its name to the **Spearman correlation coefficient.**

Like the Pearson correlation coefficient, the Spearman rank correlation ranges between -1 and +1, where -1 and +1 now indicate a perfect linear relationship between the *ranks* of the two variables. The interpretation is the same, except that the relationship is between ranks, not the actual values. To use the Spearman correlation coefficient, the actual values of the variables don't have to be linearly related but their ranks do. Correlation coefficients calculated from ranks aren't affected as much by outliers as are Pearson correlation coefficients.

Exercise: Elshibly et al. (2008) looked at maternal characteristics as predictors of birth weight in Sudanese infants. Their results are below:

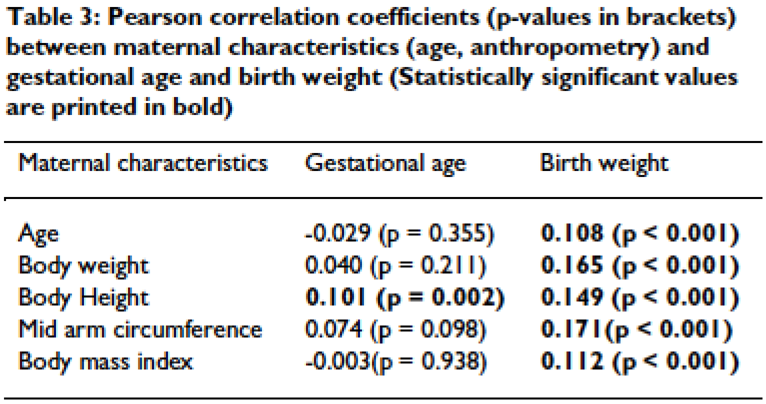


Table 3: Pearson correlation coefficients (p-values in brackets) between maternal characteristics (age, anthropometry) and gestational age and birth weight (Statistically significant values are printed in bold)

Maternal characteristics	Gestational age	Birth weight
Age	-0.029 (p = 0.355)	**0.108 (p < 0.001)**
Body weight	0.040 (p = 0.211)	**0.165 (p < 0.001)**
Body Height	**0.101 (p = 0.002)**	**0.149 (p < 0.001)**
Mid arm circumference	0.074 (p = 0.098)	**0.171(p < 0.001)**
Body mass index	-0.003(p = 0.938)	**0.112 (p < 0.001)**

Summarize their findings. Explain why the correlation coefficients are so much smaller than those reported by Ezeaka and by Elizabeth. Do you think any of these variables are good predictors of infant birth weight?

Linear Regression Model

Based on the Pearson correlation coefficient, you see that birth weight is linearly related to both maximum thigh circumference and to symphysis-fundal height (SFH). Can you use the correlation coefficient to actually predict birth weight from these measurements? Of course not. The correlation coefficient is an absolute number that just tells you about the strength of the linear association between two variables. There's no way to predict birth weight from MTC if all you know is that the correlation coefficient between the two variables is 0.95. To predict the values of a dependent variable from a predictor variable you must build a model that relates the two values. You need a model that returns birth weight when you have SFH.

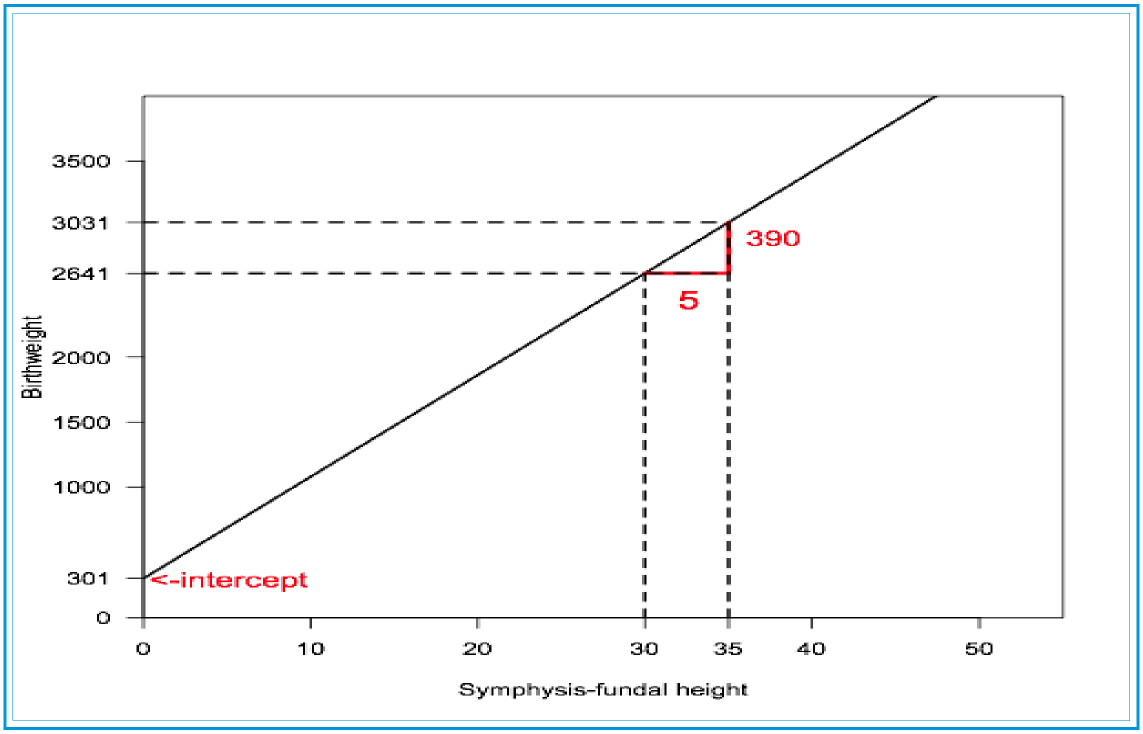
A Straight Line Model

Take a look again at Figure 6.3 again. Underneath the plot there's a cryptic statement that says *“The solid line represents the linear regression formula produced by the data (y=301 +78x).”* You may (or may not) recognize the equation as the formula for a straight line. The authors selected this line, based on the observed data, to predict the values of birth weight from SFH. For example, for an SFH of 30, the predicted birth weight is:

predicted birthweight in grams=301 + 78 x 30 = 2641 grams

Figure 6.11 is the line that relates birth weight to SFH, the regression line. The intercept is 301. That's the value for birth weight when SFH is 0. It's of no interest in this example since a height of 0 cannot realistically occur. The slope of 78 tells you that for a centimeter change in SFH there is a 78 gram change in birthweight. The diagram shows that for an SFH of 30 cm. the predicted birth weight is 2641 grams (301+78(30)). (The predicted value is the value on the line.) The predicted birth weight for an SFH of 35 cm. is 3031 grams. That's a difference of 390 grams in birthweight for a 5 centimeter change in SFH. For a one centimeter change in SFH the change is 78 grams (390/5), the slope.

Figure 6.11: The Regression Line Y=301+78X



If the slope is positive, as the values of one variable increase, so do the values of the other variable. If the slope is negative, the values of one variable increase, the values of the other variable decrease. If the slope is large, the line is steep, indicating that a small change in the predictor variable results in a large change in the dependent variable. If the slope is small, there is a more gradual increase or decrease. If the slope is 0, the value of the dependent variable doesn't change with values of the independent variable. It's the same as having a correlation coefficient of 0.

Unlike the Pearson correlation coefficient which is the same whether you correlate birth weight and SFH or SFH and birth weight, the values of the slope and intercept of a line depend on which variable is the dependent variable and which is the predictor variable. The values for the slope and intercept will be different if you're predicting birthweight from SFH or SFH from birth weight.

Exercise: What is the predicted birth weight based on an SFH of 33 cm.? Of 37 cm.?In Figure 6.9 the values for the slope are called β . What is the dependent variable for the equations the authors estimated? Can you write the equation to predict MTC from birthweight? What's the slope if you want to predict birthweight from MTC.

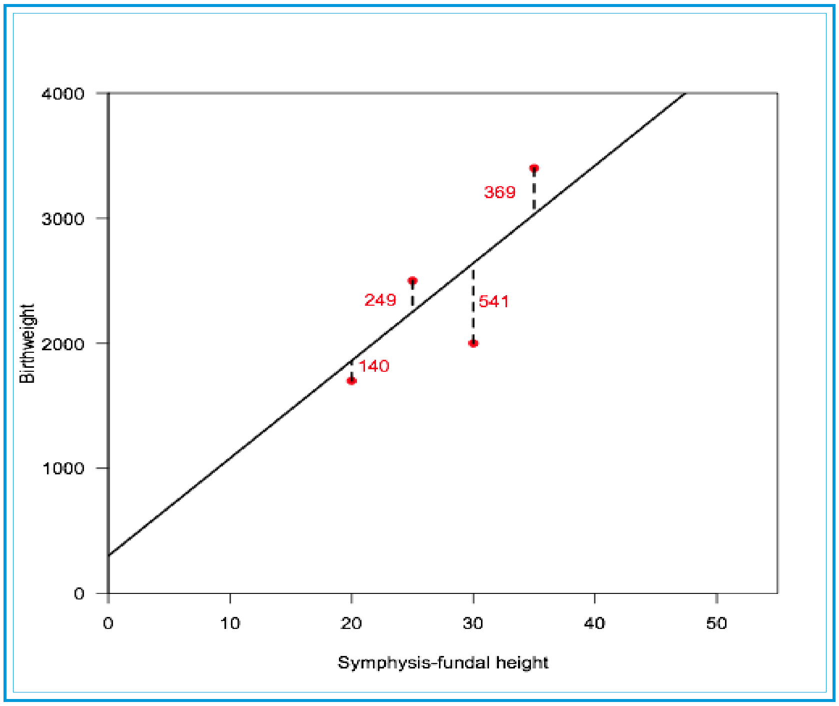
Least Squares Regression Line

The line in Figure 6.8 is the same line that you see in Figure 6.3, except that it's not surrounded by all of the data points. A lot of different lines can be drawn through the data points since the data don't fall perfectly on a line. You might wonder why the authors selected the line that they did.

The authors state that they are reporting the "*linear regression formula produced by the data.*" The line is properly called the **least squares regression line.** It is the line that has the smallest sum of squared distances from the observed data points to the values predicted by the line.

In Figure 6.12 you see the observed data points marked in red. For each point the dotted line is the distance between the observed weight and that predicted by the regression line. This is called the **residual**. The line that has the smallest sum of squared residuals is the least squares regression line. For these data the least squares regression line has an intercept of 301and a slope of 78. Those are the values reported by the authors. Any other line would have a larger sum of squared distances from the points to the line.

Figure 6.12: Residuals from Least Squares Line



Exercise: To make the regression line computations easier for medical workers, Buchmann et al. simplified the line to Weight=100 (SFH-5). What is the slope of the simplified line? The intercept?. Predict the birth weight for an SFH of 30 cm. based on the original line and the simplified line.

Exercise: You can very easily compute the least squares regression line using **R**. The output in Figure 6.13 is the regression line for Figure 6.1. Write the equation to predict female life expectancy from birth rate. What is the predicted female life expectancy for Jordan? (The birth rate for Jordan is 46.7per 1000 population and the female life expectancy is 73 years.) What is the residual, the difference between the observed and predicted life expectancy values?

Figure 6.13: R Output for the Regression Line that Predicts Female Life Expectancy from Birth Rate

```
> reg1=lm(mydata$lifeexpf~mydata$birthrat)
> reg1
Call:
lm(formula = mydata$lifeexpf ~ mydata$birthrat)
Coefficients:
   (Intercept)     mydata$birthrat
       89.5889         -0.7447
```

How Well Does the Regression Line Fit?

You know, that the least squares regression line is the line that has the smallest sum of squared distances from the points to the line. But that doesn't mean it fits the data well. Don't be tempted to use the size of the slope as a measure of how well the line fits the data. The value of the slope depends on the units in which the variables are measured. The slope is much larger when you predict birthweight in grams from maximum thigh circumference than when you predict birthweight in kilograms from maximum thigh circumference.

The absolute value of the Pearson correlation coefficient between the two variables tells you how well the line fits the data points. You see this in Figure 6.4, the larger the correlation coefficient in absolute magnitude, the closer the points are to the regression line. Although the value of the slope changes if you switch the dependent variable and the independent variable, the correlation coefficient stays the same. That means that you can predict birth weight from SFH as well as you can predict SFH from birth weight.

If you square the correlation coefficient, you obtain another useful measure. The square of the correlation coefficient, designated as R^2 , tells you what proportion of the variability in the dependent variable that is “explained” by the independent variable. You see R^2 reported in Figure 6.10. The R^2 for predicting birth weight from foot length is 0.58. More than half of the observed variability in birth weights can be attributed to differences in foot length.

All infants don't have the same birth weight. One of the explanations for this is that infants differ in size, some have long feet, some large maximum thigh circumferences and so on. If foot length was a perfect predictor of birth weight everyone with the same foot length would have exactly the same birth weight. All points would fall on the regression line and all of the residuals would be 0. Foot length would “explain” all of the observed variability in birth weights and R^2 would be 1. When all of the points don't fall exactly on the regression, you can still calculate how much of the observed variability in birth weights can be attributed to differences in foot lengths. That's what R^2 tells you. R^2 ranges from 0 to 1. Values close to 1 mean that most of the variability in the dependent variable is explained by the independent variable, values close to 0 mean that the independent variable is of little use in predicting the dependent variable when a linear model is used.

Some Warnings About Linear Regression

- Don't fit a straight line model to the data without first plotting the variables and making sure the relationship is linear over the entire range of values of the independent variable.
- Don't make predictions outside the range of the observed values of the independent variable that you use to build the model. For example, if the observed range for foot lengths is 6.0 to 9.3 cm., don't predict birth weights for infants whose foot lengths aren't within this range. You don't know if the relationship between the two variables is linear outside of your observed range.

- Don't expect that the model will fit another sample from the same population as well as it fits your data because the slope and intercept are the best values based on your data. Software packages produce adjusted R^2 values which modify the observed R^2 so it better reflects the relationship between the two variables in the population.

- Do beware of data points that have a large impact on the values of the slope and the intercept. For small datasets even a single point can make a big difference. Most software packages have special programs that help you identify these influential points.

Testing Hypotheses about the Linear Regression Model

A regression line, just like the correlation coefficient, can be used to just summarize observed data. Often you want to do more than that. You want to draw conclusions about the population from which the sample was selected. You want to calculate confidence intervals for the population slope and test the null hypothesis that the population value for the slope is 0.

To do this you have to make some assumptions about the relationship between the two variables in the population from which your sample was obtained:

- All the observations must be independent. For example, you can't include the same person more than once.

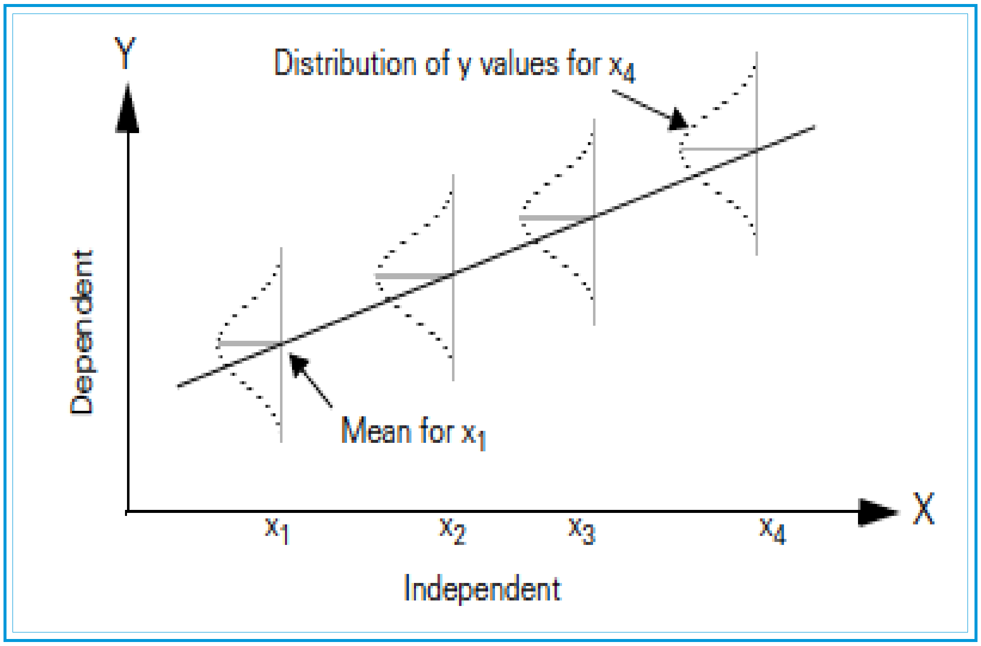
- For each value of the independent variable there is a normal distribution of values of the dependent variable. For example, for a foot length of 7 cm. there is a normal distribution of birth weights.

- All these distributions must have the same variance.

- The population means of the distributions must fall on a straight line.

Figure 6.14 is a diagram of the assumptions. For each value of the independent variable (*x1* to *x4*) there is a distribution of possible sample values for the dependent variable. In the population the distributions are normal, their means fall on a straight line, and the variances of the distributions are all the same.

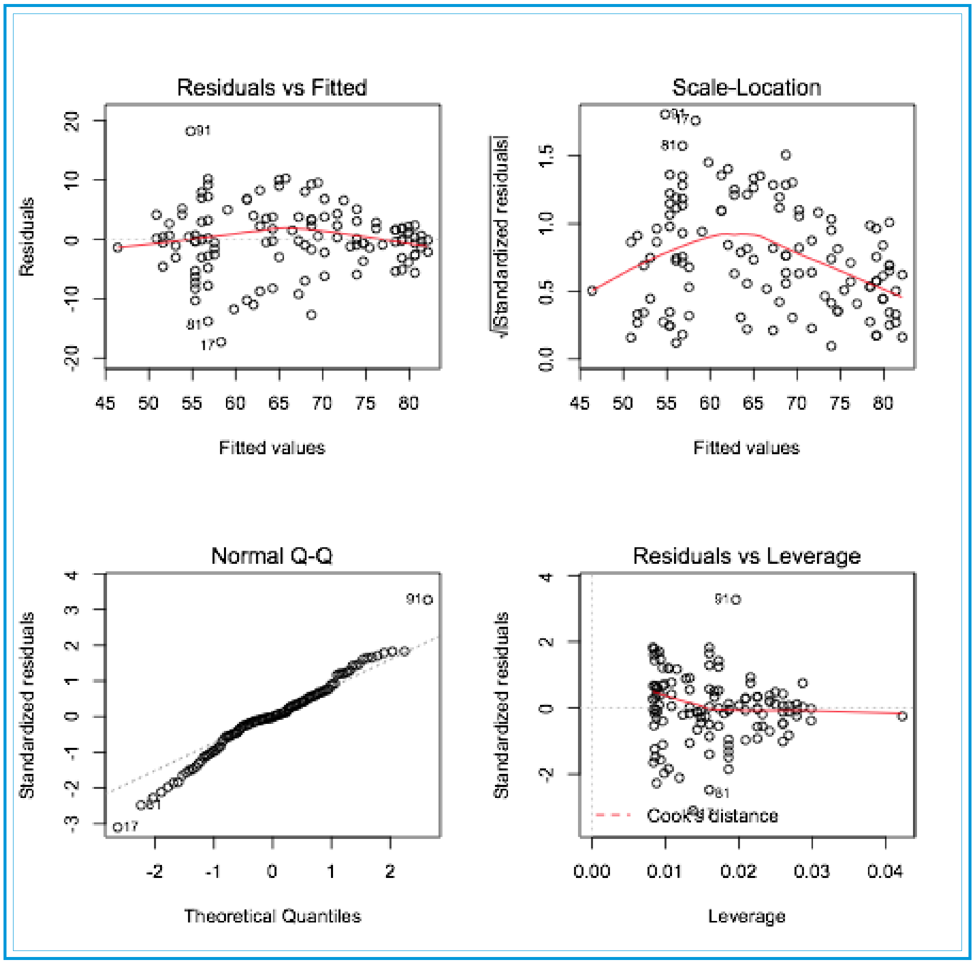
Figure 6.14: Linear Regression Assumptions



The statistical package that you are using will report the slope, the intercept and confidence intervals for the population values. You will also see a test of the null hypothesis that the population slope (β) is 0. That's the same test as the test for the hypothesis that the population correlation coefficient is 0.

The software package will also produce plots, such as those shown in Figure 6.15, that can be used to test the regression assumptions. A plot of residuals against predicted (fitted) values should not show any pattern. If it does it's possible that the linear model is incorrect. The second plot (standardized residuals against predicted values) should not show any pattern if the equal variance assumption is met. If you see increasing spread the variance may be increasing with the values of the independent variables. The Normal Q-Q plot is looking for violations of the normality assumption. If the residuals are normally distributed the points should fall on a straight line. The last plot identifies points that may have a large impact on the estimates of the regression coefficients.

Figure 6.15: Checking Linear Regression Assumptions with R



Confidence Intervals and Prediction Intervals

When you make a prediction using your regression line, you don't expect to be exactly correct. You know that if you took another sample of pregnant women from the same population you would get different values for the slope and the intercept of the regression line and therefore different predicted value. To make your prediction more useful you want to estimate the variability associated with it. If you know that for an SFH of 30 cm. the range of possible birth weights is from 2400 to 2600 grams, that's much more useful than if the plausible range of values is from 2000 to 4000 grams.

A model, such as linear regression, lets you make two different types of predictions: you can predict the average infant birthweight for all women who have an SFH of 30 cm. or you can predict the infant birthweight for a particular woman with an SFH of 30 cm. The predicted value is the same but the variability is different. You can predict the average value with less variability than you can predict the value for an individual case.

When estimating the variability of a prediction for all women with a particular SFH, you just have to worry that the slope and the intercept vary from sample to sample from the same population. For an individual value you have yet an additional worry: not all women with the same SFH have babies of the same weight. For each value of SFH there is a normal distribution of birth weights. For any value of the independent variable, the variability associated with the predicted value for an individual is always larger than the variability associated with predicting the mean.

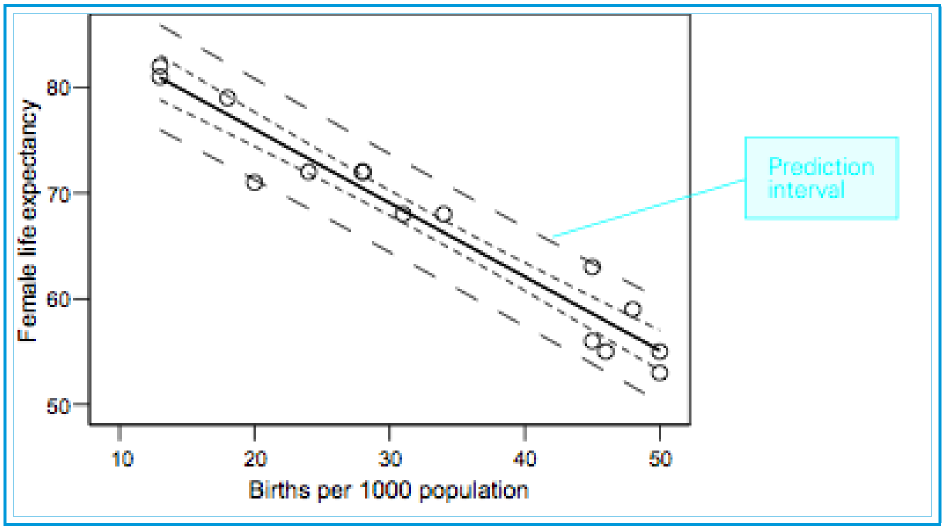
The variability for a prediction also depends on the actual value of the independent variable. Predictions are most stable for values of the independent variable close to the sample mean. That's because the regression line always passes through the point that corresponds to the mean of the dependent variable and the mean of the independent variable. Different samples from the same population don't change the predicted value as much for points close to the mean as they do for points farther away. As the distance from the mean increases, so does the variability associated with the prediction.

Prediction Intervals

Figure 6.16 illustrates what we've been talking about. You see fifteen data points and the least squares regression line based on the points. The two bands closest to the regression line are the 95% confidence intervals for the predicted mean value. The distance between the two bands is narrowest at the mean birth rate for all cases. Quite a few of the of the data points don't fall within these bands. That's because the interval is for predicting mean values, not values for individual cases.

The prediction intervals, shown by the two outermost bands, are much wider than the confidence intervals. The **prediction interval** is a range of values that you expect to include the actual value for a particular case with a designation likelihood. The prediction interval takes into account the two sources of variability for an individual prediction: sample regression lines vary about the population regression line and individual values vary about the mean. (Confidence intervals are for population values like the mean, slope, or intercept, not for values of individual cases. That's why the term prediction interval is used for individual cases.)

Figure 6.16: Prediction Intervals and Confidence Intervals



Summary

The Pearson correlation coefficient is a symmetric measure that ranges from -1 to +1 and its absolute value indicates the strength of the linear relationship between two variables. A linear regression model is used to predict the values of a dependent variable from a single predictor variable. A residual is the difference between the observed and predicted values for a case. The square of the correlation coefficient tells you what percent of the variability in the dependent variable is explained by the independent variable. Predicted values for the average of all cases with a particular value for an independent variable are less variable than predicted values for a particular case with the same value for the independent variable.

Bibliography

Elizabeth et al.: Determining an anthropometric surrogate measure for identifying low birth weight babies in Uganda: a hospital-based cross sectional study. *BMC Pediatrics* 2013: 13:54. Elizabeth(2013)

Buchmann et al.: A simple clinical formula for predicting fetal weight in labour at term—derivation and validation. *S Afr Med J:*2009: 99(6):457-460. Buchmann (2009)

Ezeaka et al.: Neonatal Maximum Thigh Circumference Tape: An Alternative Indicator of Low Birth Weight. *Nigerian Journal of Pediatrics* 2003; 30(2) : 60-66. *Ezeaka(2003)*

7 Statistical Models

- What is a statistical model?
- What is a multiple linear regression model?
- What is a logistic regression model?
- Why do survival data need special models?

In the previous chapter you used a simple statistical model to predict infant birth weight from single independent variables—maximum thigh circumference and symphysis fundal height. The model that describes the relationship between the dependent variable and the predictor was a straight line. The straight line has two **parameters**—the slope and the intercept—and you estimated the values of these from your data. (A parameter is the population value for a constant in a model.) Whenever you fit a model to data you have an equation that shows the relationship between the dependent variable and the predictor variables. The equation involves some unknown parameters that you estimate from your data values.

The straight line is one of many mathematical models that relate the value of a dependent variable to the values of one or more predictor variables. Some models, such as those in the life sciences or physical sciences, are based on known relationships such as that between the amount of a radioactive isotope remaining at a particular time and the decay parameter. If you don't have theory to guide, you want to select the simplest mathematical model that is consistent with the observed data.

Why Fit a Model?

There are different reasons for fitting statistical models. In the previous chapter your goal was prediction. You just wanted to estimate the birth weight from readily available infant measurements. As long as you have a good prediction you don't care what variable(s) it's based on.

Statistical models are also used as a tool for studying the relationship between a dependent variable and a set of independent variables. They help you organize and simplify the observed relationships so that you better understand the processes at work in the underlying population. The goal is not so much to make a prediction for an individual case as to explore the relationships between the variables. For example, you can build a model that relates maternal characteristics such as age, parity, prenatal care, nutrition, and gestational age to birth weight. Or you can estimate the likelihood that a person uses an ITN based on their age, education and gender. Your interest is not in predicting the actual birthweight for an infant or predicting whether a particular person use an ITN. Instead you want to identify the maternal, medical, and societal factors that are related to birth weight or to ITN use.

Another use of statistical models is to make the groups you are interested in comparable with respect to other predictor variables. For example, if you want to compare two treatments you want the groups receiving each treatment to be as similar as possible. It may not be practical to match the groups on a set of characteristics, so you can use a statistical model that adjusts for differences between groups based on a set of independent variables.

In the scientific literature you'll find many examples of interesting questions that are explored using statistical models. This chapter provides a brief overview of some of the commonly encountered models. You'll examine some published models and see the steps involved in building statistical models. (You'll have to study the models in more detail before you're ready to build your own.)

The Multiple Linear Regression Model

Amagloh et al. (2009) studied the relationship between infant birth weight and maternal characteristics in a sample of one hundred Ghanian women. Figure 7.1 is an estimated multiple linear regression equation for their data.

Figure 7.1: Multiple Regression Equation for Predicting Birthweight

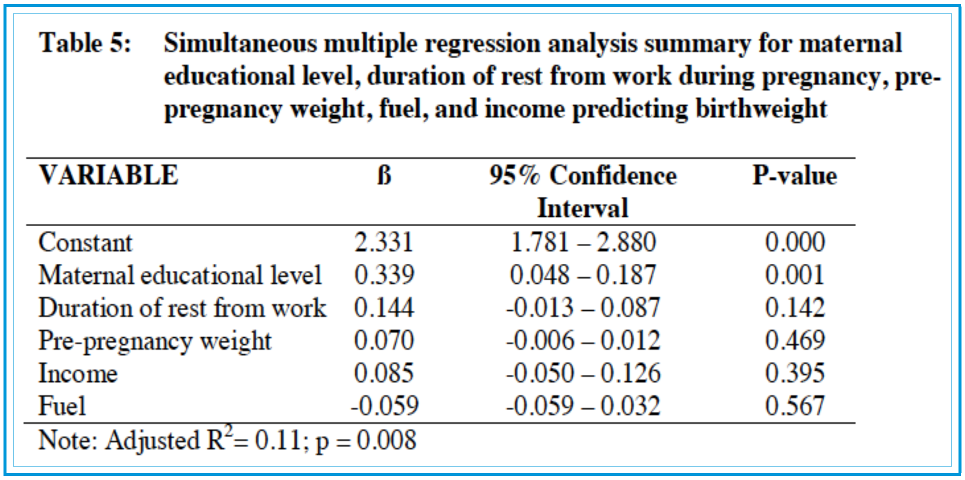


Table 5: Simultaneous multiple regression analysis summary for maternal educational level, duration of rest from work during pregnancy, pre-pregnancy weight, fuel, and income predicting birthweight

VARIABLE	ß	95% Confidence Interval	P-value
Constant	2.331	1.781 – 2.880	0.000
Maternal educational level	0.339	0.048 – 0.187	0.001
Duration of rest from work	0.144	-0.013 – 0.087	0.142
Pre-pregnancy weight	0.070	-0.006 – 0.012	0.469
Income	0.085	-0.050 – 0.126	0.395
Fuel	-0.059	-0.059 – 0.032	0.567

Note: Adjusted $R^2 = 0.11$; $p = 0.008$

The multiple linear regression model is an extension of the straight line regression model to allow more than one predictor variable. For the present study, the multiple regression equation is:

$$\text{predicted weight} = \mathbf{B_0} + \mathbf{B_1}(\text{maternal education}) + \mathbf{B_2}(\text{duration of rest})$$
$$+ \mathbf{B_3}(\text{pre-pregnancy weight}) + \mathbf{B_4}(\text{income}) + \mathbf{B_5}(\text{fuel})$$

Instead of a single slope, you now have a **partial regression coefficient (B)** for each predictor variable in the model, as well as the constant $\mathbf{B_{0.}}$ The method of least squares is used to estimate the coefficients. The coefficients are chosen so that the sum of the squared differences between the observed and predicted values (the residuals) of the dependent variable is as small as possible.

Exercise: Write the equation to predict weight using the coefficients in Figure 7.1. What do you need to know to calculate the predicted birth weight for a particular infant?

Read the paper by Amagloh. Can you determine what are possible values for the education variable? How about for fuel and income?

Testing Hypotheses About Population Values

In a multiple regression model, you can test hypotheses about the population coefficients from which the sample is selected. The sample partial regression coefficients (**B_1** through **B_5**) are then estimates of the unknown population coefficients β_1 through β_5 .

To test hypotheses about the population regression line when you have a single independent variable, your data must be a random sample from a population in which the following assumptions are met:

- The observations are independent
- The relationship between the two variables is linear
- For each value of the independent variable, there is a normal distribution of values of the dependent variable.
- Those distributions have the same variance.

You need only a slight modification of the assumptions for the multiple linear regression model. You must assume that the relationship between the dependent variable and the independent variables is linear and that for each *combination* of values of the independent variables, the distribution of the dependent variable is normal with a constant variance.

In Figure 7.1 you see the estimated partial regression coefficient for each variable in the model, the 95% confidence interval for the population value of the coefficient, and the observed significance level for the test that the true coefficient is 0. Only for education can you reject the null hypothesis that the true partial regression coefficient is 0. That's the only 95% confidence interval that doesn't include 0. The coefficient shows that for a one point increase in maternal educational level, however the authors define it, birthweight increases by 0.339 grams.

At the bottom of Figure 7.1 you see an adjusted R^2 of 0.11. That's the percent of the variability in birthweight that is accounted for by the predictor variables. It's called the *adjusted* R^2 because it adjusts the observed R^2 to better reflect how well the model would fit another sample from the same population. Adjusted R^2 also lets you compare models with different numbers of predictor variables because it adjusts for the increase expected in the sample R^2 when additional variables are included in a model, even if those variables aren't related to the dependent variable in the population.

Since the partial regression coefficients are calculated from your sample, the model fits your data better than it would another sample of cases from the same population. That's the case for all statistical models. **A model always fits the sample on which it is based better than it will fit another sample from the same population.**

R is the Pearson correlation coefficient between the observed and predicted values of the dependent variable. If the model predicts birth weight perfectly, R is 1. If there is no linear relationship between the observed and predicted values, R is 0.

The test that R^2 is 0 is the same as the test that all population partial regression coefficients are 0. That is,

$$\beta_1=\beta_2=\beta_3=...=\beta_k=0$$

That's the observed significance level shown with the adjusted R^2 value in Figure 7.1. Based on the small observed significance level (0.008) you can reject the null hypothesis that all partial regression coefficients are 0.

Interpreting the Partial Regression Coefficients

The partial regression coefficient for a variable tells you how much the value of the dependent variable changes when the value of that independent variable increases by 1 and *the values of the other independent variables stay the same*. A positive coefficient means that the predicted value of the dependent variable increases when the value of that independent variable increases. The positive coefficient for education indicates that as education increases so does infant birthweight, when the other independent variables stay the same. A negative coefficient tells you that the predicted value of the dependent variable decreases when the value of the independent variable increases. A negative coefficient for hours of exposure to smoke means that birthweight decreases as exposure to smoke increases.

A common mistake in regression analysis is to assume that a variable with a large coefficient is more important than a variable with a small coefficient. The size of the coefficient depends, among other things, on the units in which a variable is measured. If you multiply the values of a variable by 100, its coefficient will decrease by a factor of 100 while the coefficients for the other variables stay the same. You can't compare the coefficients for blood pressure in mm, income in dollars, and age in years. You can compute **standardized coefficients,** sometimes labeled Beta coefficients**,** which are the regression coefficients when the dependent variable and all independent variables are standardized to have a mean of 0 and a standard deviation of 1. They are shown in Figure 7.4

When you have one predictor variable in a regression equation, if you reject the null hypothesis that the population value for the slope is 0, you can conclude that there's a linear relationship between the dependent variable and the predictor. In statistical models with more than one predictor variable, the interpretation of the coefficients is not as straightforward. **The coefficient for a predictor variable depends not only on the relationship between the predictor and the dependent variable but also on its relationship to other predictor variables that are included in the model.**

Consider what happens if you try to predict female life expectancy in a country from four predictor variables: births per 1000 population, the logarithm of the number of doctors per 10,000 people, the log of the number of beds per 10,000 people, and the percent of the population living in an urban area. (Logarithms are taken of the number of doctors and the number of hospital beds to make the relationship between those variables and female life expectancy linear.)

Figure 7.2: Four Predictor Model for Female Life Expectancy

```
> model1=lm(formula=lifeexpf~birthrat+lnbeds+lndocs+urban)
> summary(model1)

Call:
lm(formula = lifeexpf ~ birthrat + lnbeds +  lndocs + newurban)

Coefficients:
                  Estimate      Std    Error       t value          Pr(>|t|)
(Intercept)        65.4584     3.66963           17.838          < 2e-16
birthrat          -0.31213     0.06214           -5.023            2.22e-06
lnbeds             1.48590     0.73291            2.027           0.0453
lndocs             3.09259     0.60726            5.093           1.67e-06
urban              0.03408     0.02982            1.143             0.2558

Residual standard error: 4.535 on 100 degrees of freedom
Multiple R-squared:  0.8412,       Adjusted R-squared:  0.8348
F-statistic: 132.4 on 4 and 100 DF,  p-value: < 2.2e-16
```

From Figure 7.2, based on the small observed significance levels, you can conclude that all of these variables except for the percent of the population living in an urban area are linearly related to female life expectancy. Now look at Figure 7.3, which is the same model without the log of the number of doctors. Note how all of the regression coefficients have changed. Percent urban, which was not statistically significant in the four-variable model, is now highly significant. That's because, **if the predictor variables are correlated, their coefficients in a model depend on the other variables in the model.** The percent of the population that is urban is highly correlated with the log of the number of doctors. That makes sense. The number of hospital beds, the number of doctors, and the percent of the population living in urban areas are all correlated. If you have hospital beds and number of doctors in an equation, the percent of urban population conveys little new information. Much of the information about it is already supplied by the other independent variables.

Figure 7.3: Three Predictor Model for Female Life Expectancy

```
> model1=lm(formula=lifeexpf~birthrat+lnbeds+newurban)
> summary(model1)
Call:
lm(formula = lifeexpf ~birthrat + lnbeds + urban)


Coefficients:
                   Estimate    Std. Error    t value    Pr(>|t|)
(Intercept)        69.65479    3.99301      17.444     < 2e-16
birthrat           -0.48269    0.05844      -8.260     5.91e-13
lnbeds              2.11207    0.80680       2.618     0.010209
urban               0.11188    0.02860       3.913     0.000166


Residual standard error: 5.064 on 101 degrees of freedom
Multiple R-squared:   0.8,    Adjusted R-squared:  0.794
F-statistic: 134.6 on 3 and 101 DF,  p-value: < 2.2e-16
```

Sometimes when two predictor variables are highly correlated with the dependent variable, you may find that neither of them has a statistically significant coefficient in the model, or it may be that they have signs which don't make sense. Remember that the coefficient of a variable always depends on the other variables in the model.

Exercise: Abu-Baker (2010) studied the relationship between secondhand smoke exposure and birth outcomes in Jordan. Figure 7.4 is the multiple regression model from her paper.

Figure 7.4: Model for Birth Weight and Second Hand Smoke

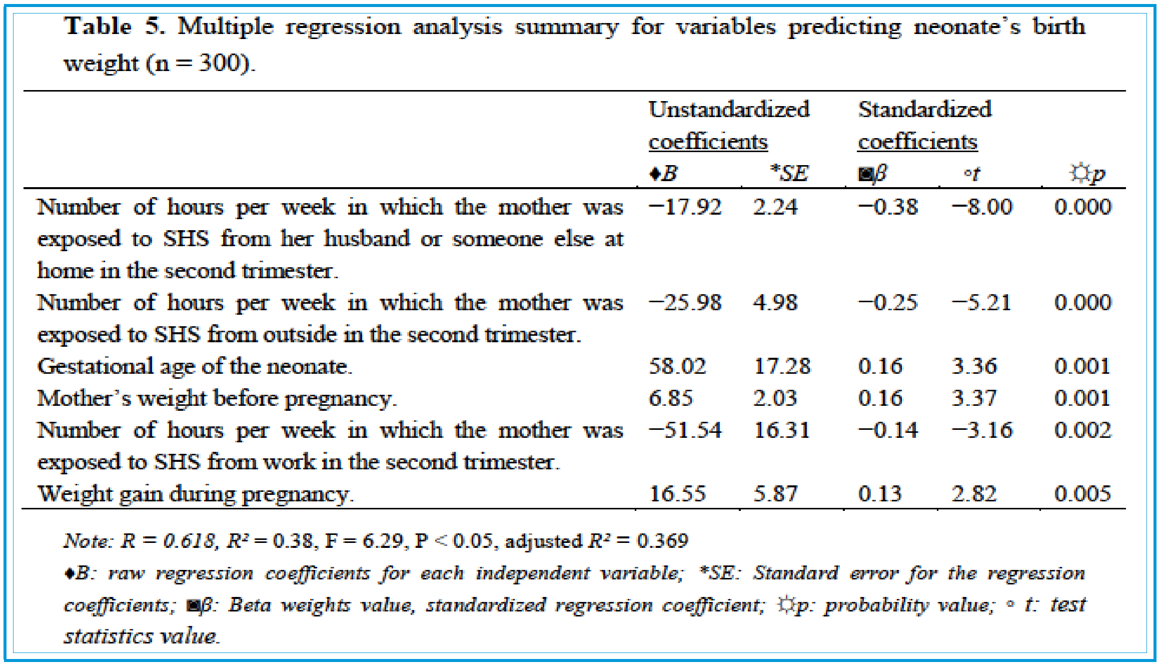


Table 5. Multiple regression analysis summary for variables predicting neonate's birth weight (n = 300).

	Unstandardized coefficients		Standardized coefficients		
	♦*B*	**SE*	◙*β*	°*t*	☼*p*
Number of hours per week in which the mother was exposed to SHS from her husband or someone else at home in the second trimester.	−17.92	2.24	−0.38	−8.00	0.000
Number of hours per week in which the mother was exposed to SHS from outside in the second trimester.	−25.98	4.98	−0.25	−5.21	0.000
Gestational age of the neonate.	58.02	17.28	0.16	3.36	0.001
Mother's weight before pregnancy.	6.85	2.03	0.16	3.37	0.001
Number of hours per week in which the mother was exposed to SHS from work in the second trimester.	−51.54	16.31	−0.14	−3.16	0.002
Weight gain during pregnancy.	16.55	5.87	0.13	2.82	0.005

Note: $R = 0.618$, $R^2 = 0.38$, F = 6.29, P < 0.05, adjusted $R^2 = 0.369$

*♦B: raw regression coefficients for each independent variable; *SE: Standard error for the regression coefficients; ◙β: Beta weights value, standardized regression coefficient; ☼p: probability value; ° t: test statistics value.*

Write the multiple regression equation. What percent of the observed variability in birth weights is explained by the model? Which variables are positively associated with birth weight? Which are negatively associated? For which variables can you reject the null hypothesis that the population value of the coefficient is 0? What do you base your conclusion on? Can you reject the null hypothesis that all the population coefficients are all 0? What do you base your conclusion on? Which predictor variables do you think are correlated?

Nominal Variables in Models

In a linear regression model, the coefficient for a predictor variable tells you the change in the dependent variable associated with a one-unit change in the predictor variable when the values of the other variables stay the same. That means that the independent variables must be measured on a scale which has equal distances between values. It makes sense to talk about the increase in birth weight for an additional year of maternal age. It doesn't make sense to talk about the increase in birthweight for a one unit change in region or religion or ethnic group.

In Figure 7.1 there's a coefficient for the education variable, and another coefficient for the fuel variable. That means that both education and fuel are assumed to be measured on a scale in which intervals are equal. If education is measured in actual years, a single coefficient may be appropriate. However, Amagloh assigned the values 1 to 5 for education, where 1=No education or up to primary level, 2=Junior High School, 3=Vocational training, 4=O'level/A'level/Senior High School, and 5=Tertiary. Since the partial regression coefficient predicts the same increase or decrease in predicted birth weight from one category to the next, that means that the difference between no education (code 1) and junior high school (code 2) is taken to be the same as the change from Senior High School (code 4) to Tertiary (code 5). That's probably not a realistic assumption.

The values for the fuel variable are equally troublesome. Since there is only one coefficient for *Fuel* it is treated as a continuous variable. Each value of *Fuel* is multiplied by the same coefficient. In fact, *Fuel* is a categorical variable and treating the five categories Firewood, Charcoal, LPG, Firewood and Charcoal, and Charcoal and LPG as equally-spaced is a poor strategy.

If you want to include nominal or ordinal variables in any kind of statistical model you must estimate a separate coefficient for each value of the variable. (One of the categories is selected as a reference category and the coefficients for the other categories indicate the difference from the reference category.) For example, you would have a coefficient for each of the education categories. For dichotomous variables that have only values of 0 and 1 (0=male;1=female) no modifications are necessary, since there is only one interval and the category coded with 0 is the reference category.

The Logistic Regression Model

If your dependent variable has only two values, such as lived/died; used an ITN/didn't use an ITN; low birth weight/not low birth weight, then you can't use a linear regression model, since a variable with only two values can't have a normal distribution with a constant variance. Instead you want to estimate the probability that an event occurs using a **binary logistic model**.

(Either of the possible outcomes can be considered “the event”, since the probability that one event occurs is just 1 minus the probability of the other. If the probability that you use an ITN is 0.60, then the probability that you don't use an ITN is 1-0.60, or 0.40.)

In logistic regression you estimate the probability that an event occur with the model

$$estimated\ probability(event)=\frac{1}{(1+e^{-Z})}$$

where Z is the linear combination

$$Z=B_0+B_1X_1+B_2X_2+...+B_pX_p$$

and X_1 to X_p are the values of the *p* independent variables and the B_i are the logistic regression coefficients.

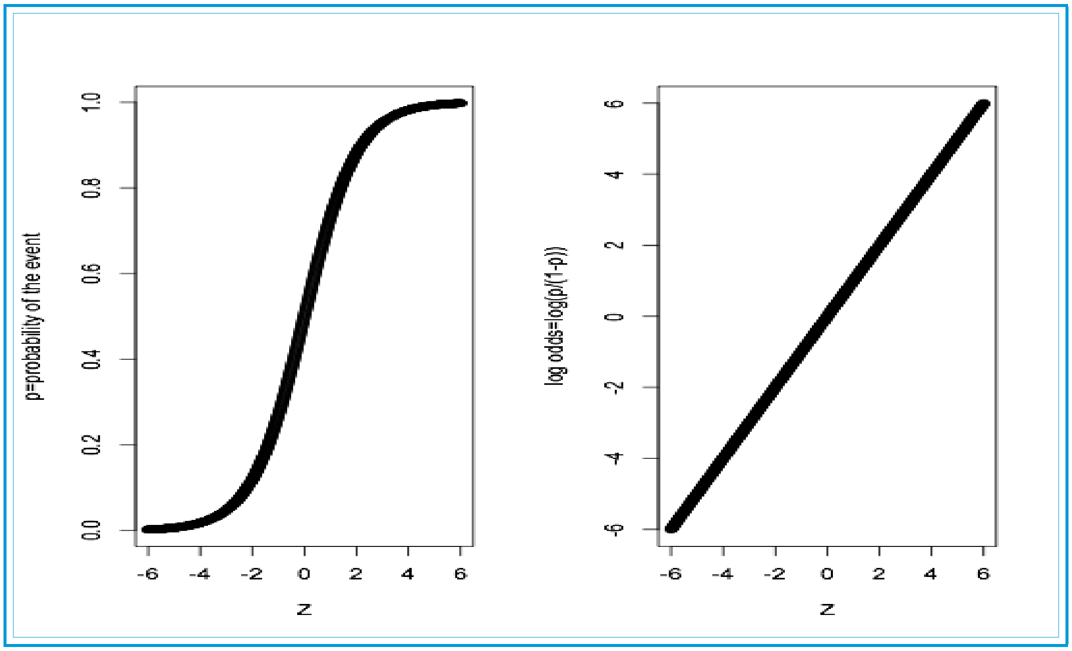
The first plot in Figure 7.5 shows what the logistic regression curve looks like when the sum of the predictors times their coefficients (Z) is on the horizontal axis and the probability of the event is on the vertical axis. The curve is S shaped. The relationship between the predictor variables and the probability is nonlinear.

The second plot in Figure 7.5 is the plot of

$$logit=\log\{\frac{(probability\ of\ event)}{(1-probability\ of\ event)}\}=\log\{\frac{(probability\ of\ event)}{(probability\ of\ no\ event)}\}$$

on the vertical axis and the linear combination of the predictor variables, Z, on the horizontal axis. The relationship between the two variables is now linear.

Figure 7.5: Logistic Regression Curve



The logistic regression equation can be written in terms of log of the odds as:

$$\log\{\frac{(probability(event))}{(probability(no\,event))}\} = B_0 + B_1 X_1 + B_2 X_2 + \ldots + B_p X_p$$

For example, if instead of predicting the actual birthweight you want to predict the probability of an infant being low birth weight, taken to be an observed birth weight less than 2500 grams, the logistic regression model is:

$$\log\{\frac{(probability(low\,birth\,weight))}{(probability(not\,low\,birth\,weight))}\} = B_0 + B_1(maternal\,education)$$

$$+B_2(duration\,of\,rest) + B_3(prepregnancy\,weight) + B_4(income) + B_5(fuel)$$

As before, the maternal education and fuel variables need to be equally spaced or dichotomous so that a single coefficient is appropriate.

The right hand side of the formula looks just like the multiple regression equation. The left hand side is the log of the odds ratio. The coefficients are estimated using an iterative method called **maximum likelihood** which results in coefficients that make the observed results most likely.

If your dependent variable is categorical and has more than two values, you can use an extension of the binary logistic regression model called multinomial regression. If the categories of the variable can be ordered in some way you can fit what's called an ordinal regression model.

Interpreting Binary Logistic Regression Coefficients

As in multiple linear regression, the value of a coefficient for a variable in a logistic regression model depends on the other variables in the logistic regression model. In multiple linear regression the regression coefficient is the estimated change in the dependent variable for a one-unit change in the independent variable when the other independent variables are held constant. The interpretation of the logistic regression is more complicated. The logistic regression coefficient for a variable is the change in the log odds for a one-unit change in the independent variable, when all other independent variables are held constant (and the variable is not included in any other terms in the model.) That's why the coefficient is sometimes called a **logit coefficient.**

Let's see what this means. An author whose identity could not be determined (Anonymous, 2012) analyzed data from Gambia to determine if there is an association between head of household exposure to malaria prevention messages and use of ITNs in children under age 5. Figure 7.6 shows the proportions of families with children under 5 who use ITNs subdivided by whether they were exposed to malaria prevention messages. You see that of those who heard or read the messages 61% use nets for their children, while 39% do not. Similarly of those who didn't hear or read messages 49% use nets and 51% do not.

Figure 7.6: Exposure to Malaria Prevention Messages and ITN Use

Message Received	Probability (Use ITN)	Probability (Not Use ITN)	Odds of ITN use	Log of Odds (base e)
Yes	0.61	0.39	0.61/0.39=1.54	0.43
No	0.49	0.51	0.46/0.51=0.95	-0.04

If you fit the logistic regression model:

$$\log\{\frac{(probability(use\ ITN))}{(probability(not\ use\ ITN))}\} = B_0 + B_1(message)$$

where *message* has two values: 1 if message is received, 0 if it is not, the coefficient for *message*, $\mathbf{B_1}$ is 0.47. That's the difference in the log of the odds of ITN use for the two groups (0.43- -0.04). If you calculate $e^{\mathbf{B1}}$, you get the odds ratio, a much more useful number. In this example, $e^{0.47}$ is 1.62. That's the ratio of the odds of ITN use in the group that has received the message (1.54) to the odds of ITN use in the group that hasn't received the message (0.95). An odds ratio of 1.62 indicates that the message-received group has a 62% increase in the odds of a child sleeping under an ITN compared to the odds for the group that did not receive the message.

In summary:

- If a coefficient is positive, the estimated odds ratio is greater than 1, which means that the odds of the event are increased.
- If a coefficient is negative, the estimated odds ratio is less than 1, which means that the odds of the even are decreased.
- If a coefficient is 0, the odds ratio is 1, which means the odds are unchanged.

Logistic Regression Example

Figure 7.7 shows an excerpt from the logistic regression model from Anonymous (2012). Since people who were exposed to malaria prevention messages may differ from those who aren't, the author wanted to look at the effect of exposure when other characteristics, such as age, region, and sex of head of household, are held constant. For each of the variables you see the logit coefficient (**B**), its standard error, and the odds ratios. Categorical variables are correctly treated in this model. Each category is a separate variable with its own coefficient.

Figure 7.7: Logistic Regression Parameter Estimates for ITN use

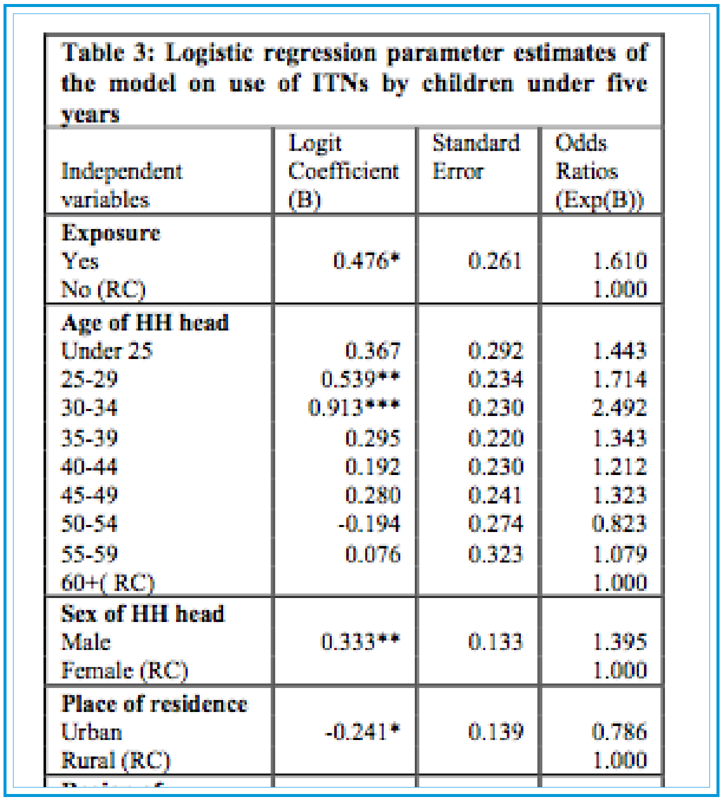


Table 3: Logistic regression parameter estimates of the model on use of ITNs by children under five years

Independent variables	Logit Coefficient (B)	Standard Error	Odds Ratios (Exp(B))
Exposure			
Yes	0.476*	0.261	1.610
No (RC)			1.000
Age of HH head			
Under 25	0.367	0.292	1.443
25-29	0.539**	0.234	1.714
30-34	0.913***	0.230	2.492
35-39	0.295	0.220	1.343
40-44	0.192	0.230	1.212
45-49	0.280	0.241	1.323
50-54	-0.194	0.274	0.823
55-59	0.076	0.323	1.079
60+(RC)			1.000
Sex of HH head			
Male	0.333**	0.133	1.395
Female (RC)			1.000
Place of residence			
Urban	-0.241*	0.139	0.786
Rural (RC)			1.000

***** p-value <0.01, ** p-value < 0.05, * p-value <0.10**

For each independent variables one of the categories of the variable is selected as the reference category (*RC* in the figure). For each variable, the odds for each category are compared to the odds for the reference category for that variable. That's why the odds ratio for the reference category is always 1. Look at the independent variable labeled *Exposure*. There are two groups, one of which was exposed to the malaria prevention message and the other which was not. The odds ratio for those exposed compared to those not exposed is 1.61. This is almost identical to the odds ratio you calculated based on the single predictor. That tells you that controlling for the other predictor variables had little effect on the coefficient for exposure.

You can test hypotheses and calculate confidence intervals for the population odds ratio. If the confidence interval for the odds ratio includes the value 1, you cannot reject the null hypothesis that the true odds ratio is 1. Testing whether the coefficient is 0 is equivalent to testing whether the odds ratio is 1.

Exercise: Zeleke et al. (2012) studied predictors of low birth weight babies in Ethiopia. Their logistic regression model is shown in Figure 7.8. What is the dependent variable? What are the predictor variables? For each give the reference category. Note that they do not even report the logistic regression coefficients instead reporting just the odds ratios.

Consider the variable Residence which has the values Urban and Rural. Based on the counts shown in the table, calculate the Crude OR for Rural compared to Urban. What is the 95% confidence interval for the OR? Does it include the value of 1? Can you reject the null hypothesis that place of residence is not related to LBW? Now look at the Adjusted OR. That's the OR estimated from the logistic regression model. What is the Adjusted OR value for Residence? Explain why the Crude OR and Adjusted OR values are different.

Summarize the results from Figure 7.8.

Figure 7.8: Factors Associated with LBW

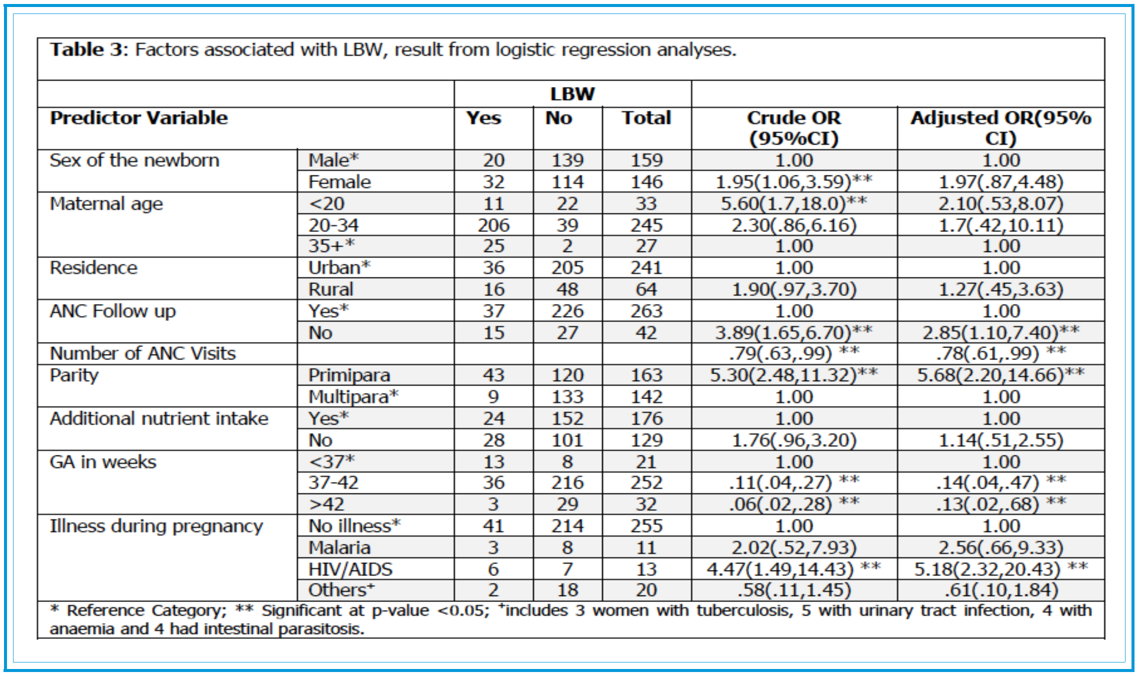


Table 3: Factors associated with LBW, result from logistic regression analyses.

		LBW				
Predictor Variable		**Yes**	**No**	**Total**	**Crude OR (95%CI)**	**Adjusted OR(95% CI)**
Sex of the newborn	Male*	20	139	159	1.00	1.00
	Female	32	114	146	1.95(1.06,3.59)**	1.97(.87,4.48)
Maternal age	<20	11	22	33	5.60(1.7,18.0)**	2.10(.53,8.07)
	20-34	206	39	245	2.30(.86,6.16)	1.7(.42,10.11)
	35+*	25	2	27	1.00	1.00
Residence	Urban*	36	205	241	1.00	1.00
	Rural	16	48	64	1.90(.97,3.70)	1.27(.45,3.63)
ANC Follow up	Yes*	37	226	263	1.00	1.00
	No	15	27	42	3.89(1.65,6.70)**	2.85(1.10,7.40)**
Number of ANC Visits					.79(.63,.99) **	.78(.61,.99) **
Parity	Primipara	43	120	163	5.30(2.48,11.32)**	5.68(2.20,14.66)**
	Multipara*	9	133	142	1.00	1.00
Additional nutrient intake	Yes*	24	152	176	1.00	1.00
	No	28	101	129	1.76(.96,3.20)	1.14(.51,2.55)
GA in weeks	<37*	13	8	21	1.00	1.00
	37-42	36	216	252	.11(.04,.27) **	.14(.04,.47) **
	>42	3	29	32	.06(.02,.28) **	.13(.02,.68) **
Illness during pregnancy	No illness*	41	214	255	1.00	1.00
	Malaria	3	8	11	2.02(.52,7.93)	2.56(.66,9.33)
	HIV/AIDS	6	7	13	4.47(1.49,14.43) **	5.18(2.32,20.43) **
	Others⁺	2	18	20	.58(.11,1.45)	.61(.10,1.84)

* Reference Category; ** Significant at p-value <0.05; ⁺includes 3 women with tuberculosis, 5 with urinary tract infection, 4 with anaemia and 4 had intestinal parasitosis.

Interactions Between Variables

Both the multiple linear regression models and the logistic regression models we've considered so far looked at the contributions of individual independent variables to predicting the dependent variable. You assumed that the effect of a variable is the same for all values of other variables in the model. For example, you assume that the effect of a laboratory value is the same over the entire range of ages or the same for men and women. For example, in Figure 1.1 there is one logistic regression coefficient for location (urban versus rural), and one logistic regression coefficient for the gender of the head of household. This implies that males in both rural and urban areas respond in the same way. The model does not allow for the interaction between the two variables. In statistics **interaction** means that two factors considered together have more (or less) of an effect than when they are considered individually. For example, the probability of rural-males using an

ITN is more (or less) than is predicted from the coefficient of rural and the coefficient of male. If you're studying the risk of lung cancer you may find that smoking and inhaling asbestos fibers both increase the risk of lung cancer when considered individually. However, people who smoke *and* inhale asbestos fibers are much more likely to develop lung cancer than you predict based on the individual smoking and asbestos coefficients. To capture the joint effect of two or more variables interaction terms must be included in a model. (For categorical variables interaction coefficients can be constructed in several ways.)

A linear regression model with an interaction term between the variables X_1 and X_2 is written as

$$\textbf{\textit{predicted dependent variable}} = \boldsymbol{B}_0 + \boldsymbol{B}_1 \boldsymbol{X}_1 + \boldsymbol{B}_2 \boldsymbol{X}_2 + \boldsymbol{B}_3 (\boldsymbol{X}_1 \boldsymbol{X}_2)$$

The coefficient $\mathbf{B_3}$ is for the interaction effect. If both of the variables are coded as 0 or 1 (absent or present), it tells you how much more the dependent variable changes when both variables are present than you would predict from their individual effects.

There are several different ways to code categorical variables and they result in coefficients which have different interpretations. The coefficient for each category may tell you how much the category effect differs from the average effect of all categories, or the category may be compared to a reference category. Different coding schemes result in different coefficients but not in different conclusions.

Bejon (2009) studied ITN use and children's age on malaria infection. The results of logistic regression models are in Figure 7.9. Two different models were examined: febrile vs asymptomatic, and any malaria vs uninfected. The interaction term between ITN use and age greater than 42 months (old) is shown as *ITN use*Old*. The odds ratio for the interaction term for the first model is 6.5, indicating that the odds of febrile malaria as compared to the odds of asymptomatic malaria are higher for older children who slept under an ITN as compared to all other children.

Figure 7.9: ITN Use Interaction with Age

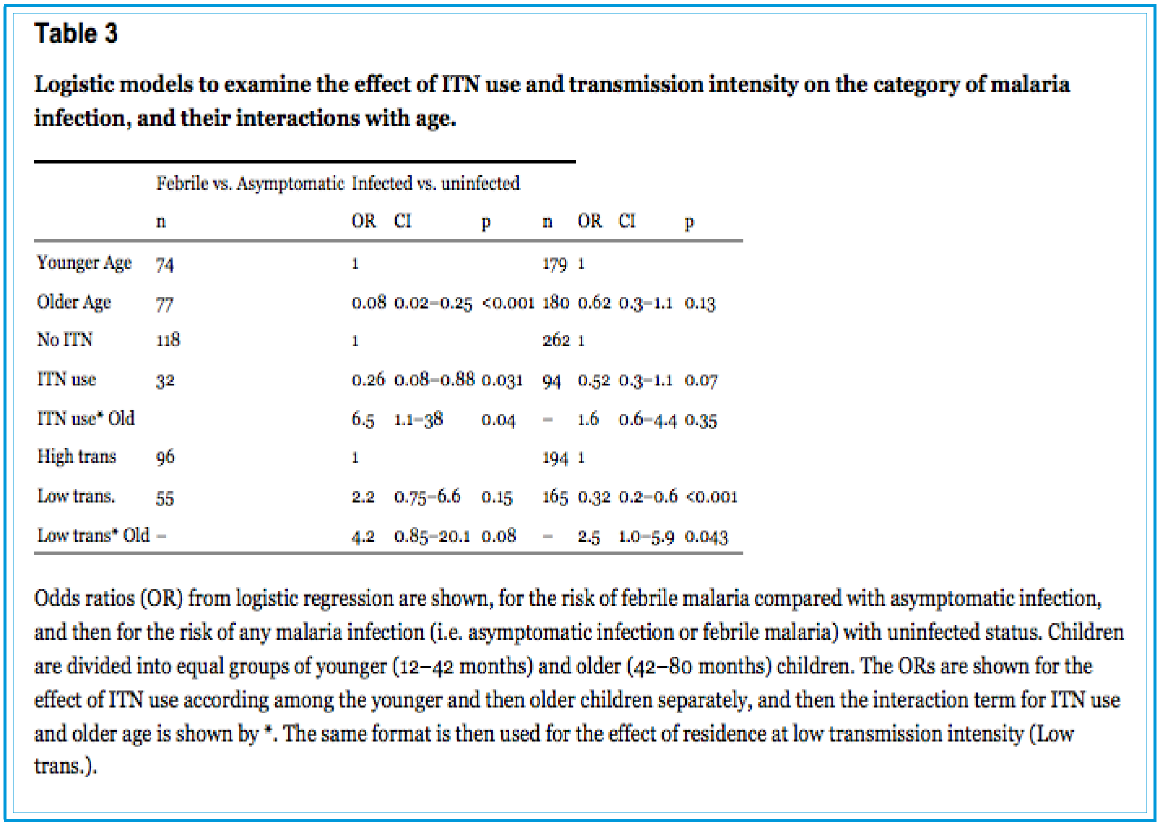


Table 3

Logistic models to examine the effect of ITN use and transmission intensity on the category of malaria infection, and their interactions with age.

	Febrile vs. Asymptomatic	Infected vs. uninfected						
	n	OR	CI	p	n	OR	CI	p
Younger Age	74	1			179	1		
Older Age	77	0.08	0.02–0.25	<0.001	180	0.62	0.3–1.1	0.13
No ITN	118	1			262	1		
ITN use	32	0.26	0.08–0.88	0.031	94	0.52	0.3–1.1	0.07
ITN use* Old		6.5	1.1–38	0.04	–	1.6	0.6–4.4	0.35
High trans	96	1			194	1		
Low trans.	55	2.2	0.75–6.6	0.15	165	0.32	0.2–0.6	<0.001
Low trans* Old	–	4.2	0.85–20.1	0.08	–	2.5	1.0–5.9	0.043

Odds ratios (OR) from logistic regression are shown, for the risk of febrile malaria compared with asymptomatic infection, and then for the risk of any malaria infection (i.e. asymptomatic infection or febrile malaria) with uninfected status. Children are divided into equal groups of younger (12–42 months) and older (42–80 months) children. The ORs are shown for the effect of ITN use according among the younger and then older children separately, and then the interaction term for ITN use and older age is shown by *. The same format is then used for the effect of residence at low transmission intensity (Low trans.).

Model Discrimination

Model Discrimination is the ability of a model to distinguish between cases who experience the event and those that don't, based on the estimated probability that the event occurs.

A perfect model always assigns higher probabilities to cases who experience the event than to cases who don't. The **c-statistic** is a measure of a model's ability to discriminate between the two sets of cases. It is the proportion of pairs of cases with different observed outcomes in which the model results in a higher probability for the case with the event than for the cases without the event. The *c*-statistic is equal to the area under the ROC (Receiver Operating Characteristic) curve, sometimes just called the Area Under the Curve (AUC). The *c*-statistic ranges in value from 0.5 to 1. A value of 0.5 indicates that the model is no better than flipping a coin for assigning cases to groups. A value of 1 means that the model always assigns higher probabilities to cases with the event than to cases without the event.

In the previous chapter you considered various easily obtainable infant measurements as predictors of birth weight. Elizabeth et al. (2013) evaluated five infant measures as predictors of birth weight. Besides reporting the correlation coefficients shown in Figure 7.10, they also report the AUC and its 95% confidence interval. To calculate the AUC each infant is classified into one of two groups based on birth weight (less than 2500 grams and greater than 2500 grams) and a logistic regression model is estimated using each of the variables alone. You see that foot length has the largest AUC, as well as the largest correlation coefficient.

Figure 7.10: Correlation of anthropometric measures with birth weight

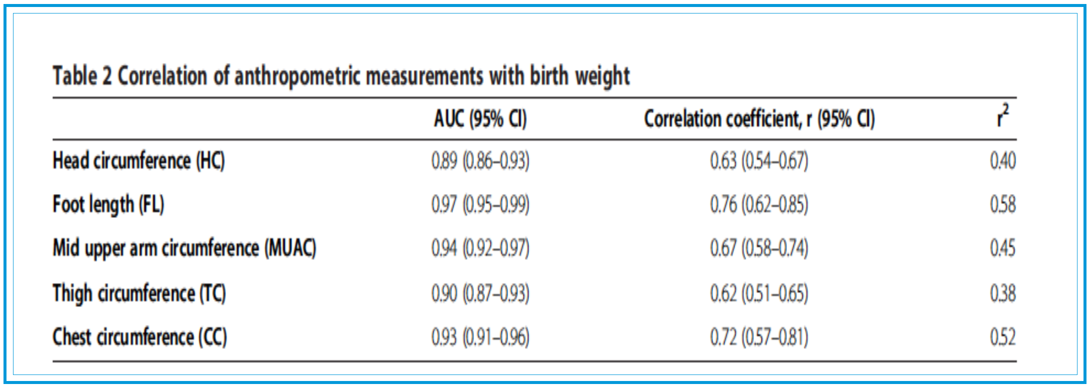


Table 2 Correlation of anthropometric measurements with birth weight

	AUC (95% CI)	Correlation coefficient, r (95% CI)	r^2
Head circumference (HC)	0.89 (0.86–0.93)	0.63 (0.54–0.67)	0.40
Foot length (FL)	0.97 (0.95–0.99)	0.76 (0.62–0.85)	0.58
Mid upper arm circumference (MUAC)	0.94 (0.92–0.97)	0.67 (0.58–0.74)	0.45
Thigh circumference (TC)	0.90 (0.87–0.93)	0.62 (0.51–0.65)	0.38
Chest circumference (CC)	0.93 (0.91–0.96)	0.72 (0.57–0.81)	0.52

Exercise: Read the Elizabeth (2013) paper and calculate the sensitivity and specificity values shown in their Table 3 from the counts. Calculate the statistics shown in Table 4 from the counts in Table 3.

Analyzing Survival Data

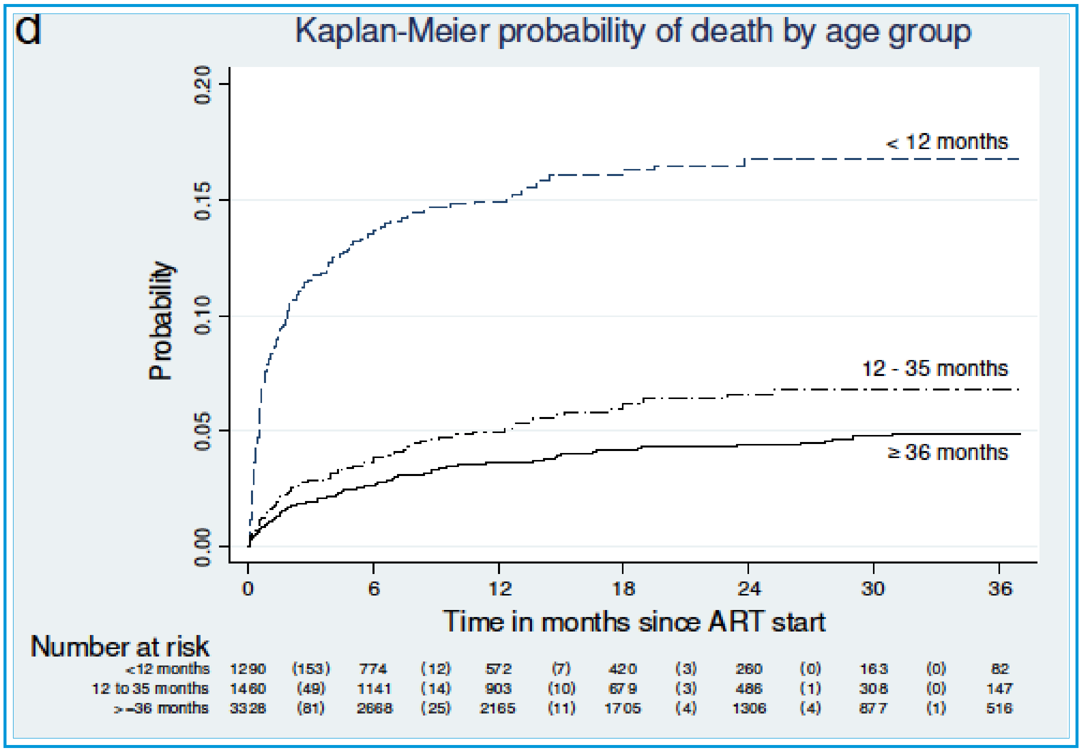
Often you are interested in the time interval between two events: diagnosis and death; surgery and recurrence of a disease; starting university and completion; release from prison and return. When you study time between two events you may face two problems: not everyone experiences the second event, and people are observed for different lengths of time. Not all patients experience recurrence of a disease and not all students finish their degrees. Cases who do not experience the final event are said to be **censored.** Similarly not everyone is diagnosed on the same day, or released from prison on the same day.

If you eliminate people who have not actually experienced the second event and calculate an average survival time or time to recurrence, your results will be flawed since data from people who are disease free or not back in person are ignored. Special statistical models, called survival time (or failure-time) methods, are used to analyze the time between two well defined events. Two frequently used methods for analyzing survival time data are Kaplan-Meier estimates and the Cox proportional hazards model.

Kaplan-Meier estimates are based on subdividing the time period after the initial event into smaller time intervals based on when events occur. Cases contribute survival information to intervals during which they have been observed. This makes best use of the available information. At each time point you obtain an estimate of the cumulative percent still surviving. You can estimate the percent dead by subtracting the percent surviving from 100%.

Consider Figure 7.11 which shows the percent of children who died at various time points after administration of antiretroviral treatment in South Africa (Davies et al., 2009). You see that 24 months after start of treatment, about 4% of children who started treatment when they were older than 36 months died. For children who started treatment when they were less than 12 months old, almost 17% are dead.

Figure 7.11: Kaplan Meier Survival Curve



Number at risk													
<12 months	1290	(153)	774	(12)	572	(7)	420	(3)	260	(0)	163	(0)	82
12 to 35 months	1460	(49)	1141	(14)	903	(10)	679	(3)	486	(1)	308	(0)	147
>=36 months	3328	(81)	2668	(25)	2165	(11)	1705	(4)	1306	(4)	877	(1)	516

Special methods are also available for modeling predictors of survival time. The **Cox proportional hazards model** is most often used. The **hazard function**, which tells you how likely a case is to experience an event given that the case has survived to that time, is expressed as a function of a linear combination of predictor variables.

The hazard ratio (HR) is the ratio of the rate at which patients in two groups experience an event. A hazard ratio of 1 means that the two groups experience the event at equal rates. A hazard ratio of 2 implies that, at any time, cases in one group are experiencing twice the rate of the event as the other group.

Figure 7.12 shows hazard ratios for the Davies study, both with and without adjustment (adjusted and crude HR) for the other predictor variables. As in logistic regression one of the categories is selected as the reference category. You see that children who started ART when they were less than 1 year old have a crude HR of 3.38 compared to children who started ART when they were older than 3 years. When adjusted for other variables in the model the HR decreases to 2.00.

Figure 7.12: Predictors of Mortality

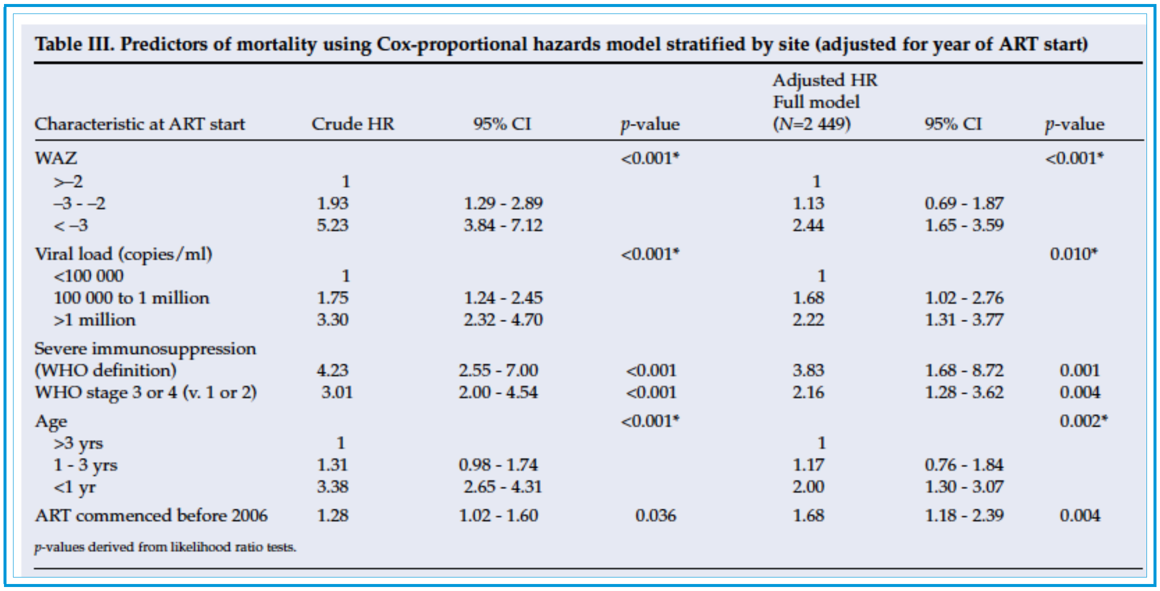


Table III. Predictors of mortality using Cox-proportional hazards model stratified by site (adjusted for year of ART start)

Characteristic at ART start	Crude HR	95% CI	*p*-value	Adjusted HR Full model (*N*=2 449)	95% CI	*p*-value
WAZ			<0.001*			<0.001*
>–2	1			1		
–3 - –2	1.93	1.29 - 2.89		1.13	0.69 - 1.87	
< –3	5.23	3.84 - 7.12		2.44	1.65 - 3.59	
Viral load (copies/ml)			<0.001*			0.010*
<100 000	1			1		
100 000 to 1 million	1.75	1.24 - 2.45		1.68	1.02 - 2.76	
>1 million	3.30	2.32 - 4.70		2.22	1.31 - 3.77	
Severe immunosuppression (WHO definition)	4.23	2.55 - 7.00	<0.001	3.83	1.68 - 8.72	0.001
WHO stage 3 or 4 (v. 1 or 2)	3.01	2.00 - 4.54	<0.001	2.16	1.28 - 3.62	0.004
Age			<0.001*			0.002*
>3 yrs	1			1		
1 - 3 yrs	1.31	0.98 - 1.74		1.17	0.76 - 1.84	
<1 yr	3.38	2.65 - 4.31		2.00	1.30 - 3.07	
ART commenced before 2006	1.28	1.02 - 1.60	0.036	1.68	1.18 - 2.39	0.004

p-values derived from likelihood ratio tests.

Exercise: Which variables in Figure 7.12 are associated with increased risk of death? How would you tell which variables are significantly associated with risk of death based only on the 95% confidence intervals.

Exercise: Read the Bejon (2009) paper and summarize the results for the Cox survival models. How do the results of these models compare to the logistic models? Are you convinced that there is an age by ITN interaction effect? Explain what that means.

Summary

This chapter focused on presenting the basics of the models you will most often encountered in the medical literature. If you are building the model yourself there are many other issues you must consider. For example: Do the data violate the assumptions needed for the particular model? How will you determine which predictor variables to include in a model? Is the relationship between each of these variables and the dependent variable linear? Do you need to transform any of the variables? Are interaction terms necessary? Are there points that have too much influence on the estimated coefficients? How well does the model fit?

Bibliography:

Abu-Baker et al. The Influence of Secondhand Smoke Exposure on Birth Outcomes in Jordan. *Int. J. Environ. Res. Public Health* 2010, *7*, 616-634. Abu-Baker (2010)

Amagloh et al. Evaluation of Some Maternal and Socio-Economic Factors Associated with Low Birthweight Among Women in the Upper East Region, Ghana.*African Journal of Food Agriculture Nutrition and Development:*2009(7). Amagloh (2009)

Elizabeth et al.: Determining an anthropometric surrogate measure for identifying low birth weight babies in Uganda: a hospital-based cross sectional study. *BMC Pediatrics* 2013: 13:54. Elizabeth(2013)

Zeleke et al. Incidence and correlates of low birth weight at a referral hospital in Northwest Ethiopia. *Pan African Medical Journal*.2012;12:4. Zelke(2012)

Davies et al., Outcomes of the South African National Antiretroviral Treatment Programme for children. *South African Medical Journal,* October 2009, Vol. 99, No. 10. Davies(2009)

Bejon P, Ogada E, Peshu, N, Marsh K. Interactions Between Age and ITN Use Determine the Risk of Febrile Malaria in Children. PloS One, 2009: 23 (4) Bejon (2009)

Index