Principles of Biostatistics

Class notes to accompany the textbook by Pagano and Gauvreau

Constantin Yiannoutsos., Ph.D.

CENTER FOR BIOSTATISTICS IN AIDS RESEARCH
HARVARD SCHOOL OF PUBLIC HEALTH

Contents

List of Figures

List of Tables

Chapter 1

What is statistics?

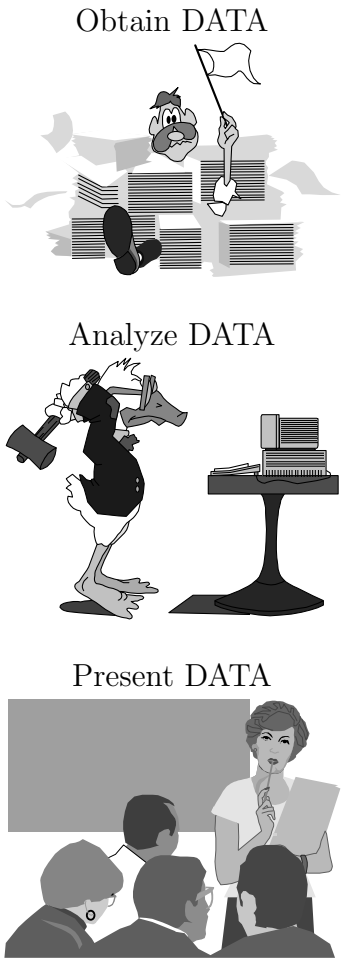


Figure 1.1: Functions of statisticians

What is statistics anyway?

- Statistics is the summary of information (data) in a meaningful fashion, and its appropriate presentation
- Statistics is the postulation of a plausible model explaining the mechanism that generates the data, with the ultimate goal to extrapolate and predict data under circumstances beyond the current experiment
- Bio-statistics is the segment of statistics that deals with data arising from biological processes or medical experiments

It all starts with DATA!

1.1 Types of Numerical DATA

As "data" we consider the result of an experiment. A rough classification is as follows:

- *Nominal data*
  Numbers or text representing unordered categories (e.g., 0=male, 1=female)
- *Ordinal data*
  Numbers or text representing categories where order counts (e.g., 1=fatal injury, 2=severe injury, 3=moderate injury, etc.
- *Discrete data*
  This is numerical data where both ordering and magnitude are important but only whole number values are possible (e.g., Numbers of deaths caused by heart disease (765,156 in 1988) versus suicide (40,368 in 1988, page 10 in text).
- *Continuous data*
  Numerical data where any conceivable value is, in theory, attainable (e.g., height, weight, etc.)

1.2 Ways to summarize data

1.2.1 Descriptive statistics

Measures of central tendency

The following are summaries of the "middle" (or most representative) part of the data

1. **Mean**
   The mean is the average of the measurements in the data. If the data are made up of n observations $x_1, x_2, ..., x_n$, the mean is given by the sum of the observations divided by their number, i.e.,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

where the notation $\sum_{i=1}^{n}$ means "sum of terms counted from 1 to n". For example if the data are $x_1 = 1$, $x_2 = 2$, $x_3 = 3$, then their average is $1/3(1+2+3) = 2$.

2. **Median**
   The median is the middle observation according to the observartions' rank in the data. In the previous example, the median is $m = 2$. The median is the observation with rank $(n+1)/2$ if n is odd, or the average of observations with rank $n/2$ and $(n+1)/2$ if n is even.

Note what would happen if $x_3 = 40$ in the above example. Then the mean is $\bar{x} = \frac{1}{3}(1+2+40) = 14.33$. However, the median is still $m = 2$. In general, the median is less sensitive than the mean to extremely large or small values in the data.

Remember: The mean follows the tail

Thus, when data are skewed to the left (there are a large number of small values), then the mean will be smaller than the median. Conversely, if the data are skewed to the right (there is a large number of high values), then the mean is larger than the median.

For example, the distribution of light bulb lifetimes (time until they burn out) is skewed to the right (i.e., most burn out quickly, but some can last longer). Next time you buy a light bulb, notice whether the mean or the median life of the light bulb is quoted on the package by the manufacturer. Which statistic would be most appropriate?

Measures of spread

The most common measures of spread (variability) of the data are the variance and the standard deviation.

1. **Variance** The variance is the average of the square deviations of the observations from the mean. The deviations are squared because we are only interested in the size of the deviation rather than the direction (larger or smaller than the mean). Note also that $\sum_{i=1}^{n}(x_i - \bar{x})$. Why? The variance is given by

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

where $x_1, x_2, ..., x_n$ are the data observations, and their mean. The variance of $x_1 = 1$,$x_2 = 2$,$x_n = 3$, is $\frac{1}{2}\left[(1-2)^2 + (2-2)^2 + (3-2)^2\right] = 1$.

The reason that we divide by n-1 instead of n has to do with the number of "information units" in the standard deviation. Try to convince yourself, that after estimating the sample mean, there are only $n-1$ independent (i.e., *a priori* unknown) observations in our data. Why? (hint. Use the fact that the sum of deviations from the sample mean is zero)

2. **Standard deviation**
The standard deviation is given by the square root of the variance. It is attractive, because it is expressed in the same units as the mean (instead of square units like the variance).

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

The standard deviation of $x1 = 1$,$x_2 = 2$,$x_n = 3$, is

$$s = \sqrt{\frac{1}{3-1}\left((1-2)^2 + (2-2)^2 + (3-2)^2\right)} = 1$$

1.2.2 Frequency tables and graphs

Tables

- Frequency tables (frequency distributions)
- Relative frequencies (percent of total)

Graphs

- Bar charts
- Frequency polygons
- Scatter plots
- Line graphs
- Box plots

1.3 Examples

1.3.1 Frequency tables

Table 1.1: Frequencies of serum cholesterol levels

Cholesterol level (mg/100 ml)	Frequency	Cumulative Frequency	Relative Frequency (%)	Cumulative Relative Frequency (%)
80-119	13	13	1.2	1.2
120-159	150	163	14.1	15.3
160-199	442	605	41.4	56.7
200-239	299	904	28.0	84.7
240-279	115	1019	10.8	95.5
280-319	34	1053	3.2	98.7
320-360	9	1062	0.8	99.5
360-399	5	1067	0.5	100.0
Total		1067		100.0

The choice of intervals in a frequency table is very important. Unfortunately, there are no established rules for determining them. Just make sure that a "cut-off" value is a beginning point of one of the intervals. In the table above, the value of 200 mg/100 ml of cholesterol is such a value.

1.3.2 Bar charts

Consider the following data set. This is an example of nominal data:

```
1 5 3 1 2 4 1 3 1 5
2 1 1 5 3 1 2 1 4 1
4 1 3 1 5 1 2 1 1 2
5 1 1 5 1 5 3 1 2 1
2 3 1 1 2 1 5 1 5 1
1 2 5 1 1 2 3 4 1 1
1 1 2 1 1 2 1 1 2 3
3 3 1 5 2 3 5 1 3 4
1 1 2 4 5 4 1 5 1 5
5 1 1 5 1 1 5 1 1 5
```

1. Motor vehicle, 2. Drowning, 3. House fire, 4. Homicide, 5. Other

Table 1.2: U.S.A. cigarette consumption, 1900-1990

Year	Number of Cigarettes
1900	54
1910	151
1920	665
1930	1485
1940	1976
1950	3522
1960	4171
1970	3985
1980	3851
1990	2828

The Stata output is as follows:

```
. tab accident

    acc_lab |      Freq.     Percent        Cum.
------------+-----------------------------------
   Motor Ve |         48       48.00       48.00
   Drowning |         14       14.00       62.00
   House Fi |         12       12.00       74.00
   Homicide |          7        7.00       81.00
      Other |         19       19.00      100.00
------------+-----------------------------------
      Total |        100      100.00

. label define acclab 1 "Motor vehicle" 2 "Drowning" 3 "House fire"
> 4 "Homicide" 5 "Other"
```

The following bar chart is produced by the STATA output below:

```
. label var cigs "Cigarette consumption"

. graph cigs, title(Cigarette consumption per capita US between 1900
> and 1990) xlab ylab
```

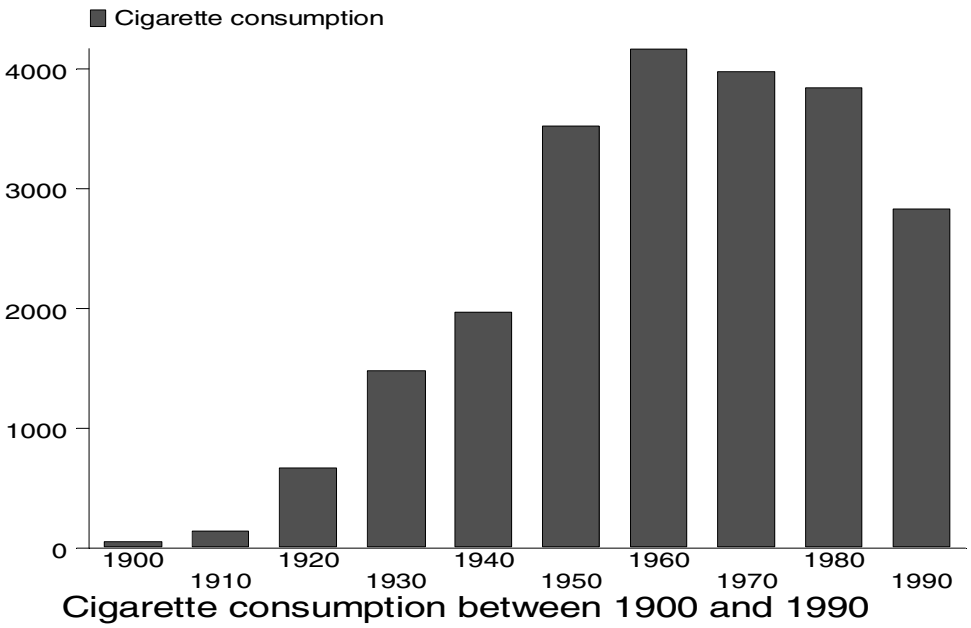


Figure 1.2: Example of a bar chart

1.3.3 The Box Plot

Follow these steps in order to produce a box plot:

1. Calculate the median m
2. Calculate the first and third quartile Q_1 and Q_3
3. Compute the inter-quartile range $IQR = Q_3 - Q_1$
4. Find the lower fence $LF = Q1 - 1.5 \times IQR$
5. Find the upper fence $UF = Q3 + 1.5 \times IQR$
6. Find the lower adjacent value LAV=smallest value in the data that is greater or equal to the lower fence
7. Find the upper adjacent value UAV=largest value in the data that is smaller or equal to the upper fence

```
. label val accident acclab

. graph accident, title(Reasons of death) xlab ylab
```

Figure 1.3: Frequency of deaths among children according to cause

8. Any value outside the LAV or UAV is called an outlier and should receive extra attention

Consider the following depression scale scores:

Table 1.3: The Koopmans (1981) data set of depression scores

2	5	6	8	8	9	9
10	11	11	11	13	13	14
14	14	14	14	14	15	15
16	16	16	16	16	16	16
16	17	17	17	18	18	18
19	19	19	19	19	19	19
19	20	20				

The Box Plot for Koopmans data is constructed as follows:

1. Calculate the median m Since the number of observations is 45 (odd number) the median is the $[(45+1)/2]$th i.e., the 23d observation. That is, $m = 16$

2. Calculate the first and third quartile Q_1 and Q_3 Split the data set into two equal parts (including the median in both of them), that is, split into the first and last 23

observations. Then Q_1 is the median of the first 23 observations (the 12th observation), and Q_3 is the median of the last 23 observations (the 34th observation). Thus, $Q_1 = 13$ and $Q_3 = 18$

3. Compute the inter-quartile range $IQR = Q3 - Q1$. $IQR = 18 - 13 = 5$
4. Find the lower fence $LF = Q_1 - 1.5 * IQR$. $LF = Q_1 - 1.5 * IQR = 13 - 1.5(5) = 5.5$
5. Find the upper fence $UF = Q_3 + 1.5 * IQR$. $UF = Q_3 + 1.5 * IQR = 18 + 1.5(5) = 25.5$
6. Find the lower adjacent value. LAV=smallest value in data > 5.5, $LAV = 6$
7. Find the upper adjacent value. UAV=largest value in data < 25.5, $UAV = 20$
8. Since 2 and 5 are lower than the LAV, these observations are outliers and must be investigated further

The Stata manipulation of the Koopmans data set is given below:

```
. graph depscore, box ylab title(Box Plot of Koopmans depression scores)
```

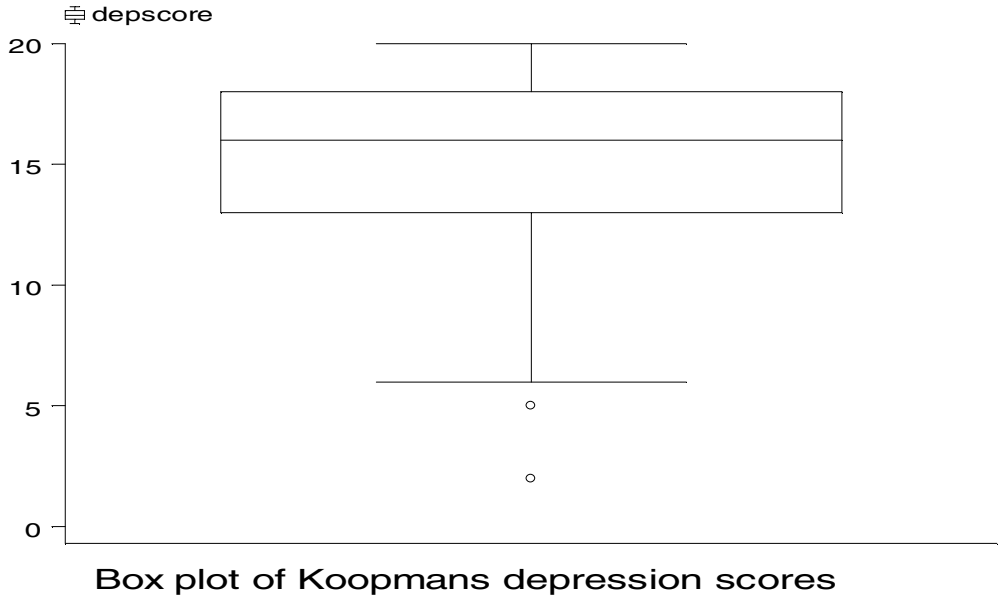


Figure 1.4: Box plot of Koopman's data

Chapter 2

Probability and Bayes Theorem

Ask Marilyn®

BY MARILYN VOS SAVANT

A particularly interesting and important question today is that of testing for drugs. Suppose it is assumed that about 5% of the general population uses drugs. You employ a test that is 95% accurate, which we'll say means that if the individual is a user, the test will be positive 95% of the time, and if the individual is a nonuser, the test will be negative 95% of the time. A person is selected at random and is given the test. It's positive. What does such a result suggest? Would you conclude that the individual is a drug user?

Consider what happens when you roll a die. Here are the possible outcomes of the single die roll. What if you rolled two dice? In that case you would get,

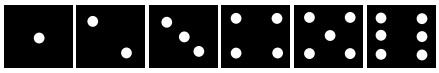


Figure 2.1: Possible outcomes of the roll of a single die

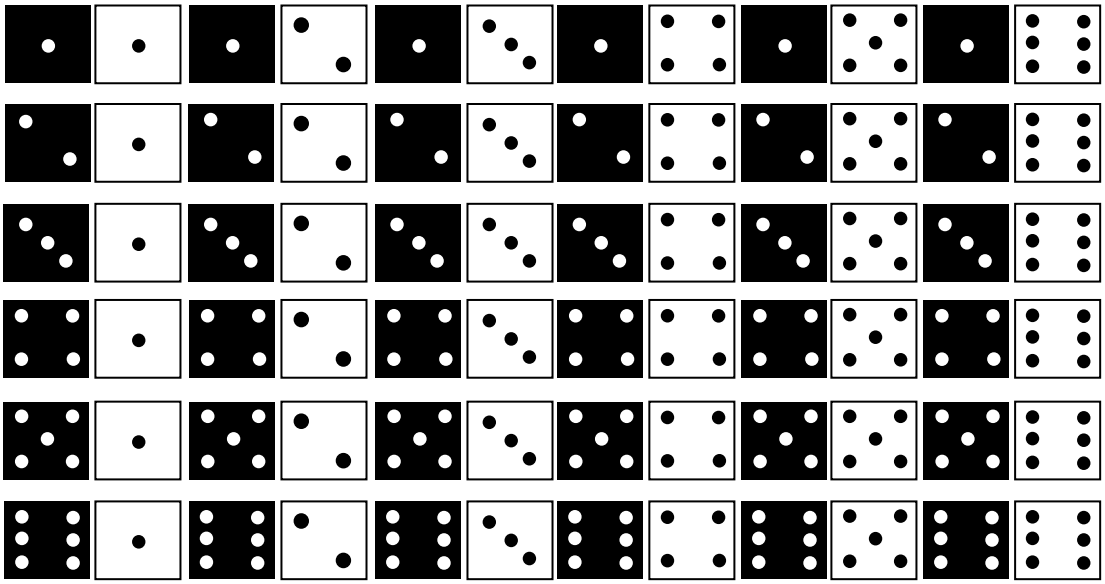


Figure 2.2: Possible outcomes of the roll of two dice

2.1 Events

The outcome of each die roll is called an *event*. Events are also coin flips, results of experiments, the weather and so on.

2.1.1 Special events

The set of all possible events is called the *sample space*. In the figures above, the sample space of a roll of a single die and two-dice was presented respectively.

An event that cannot happen is called a *null* event.

Two events that cannot both happen are called *mutually exclusive* events. For example event A="Male" and B="Pregnant" are two mutually exclusive events (as no males can be pregnant.

To envision events graphically, especially when all the outcomes are not easy to count, we use diagrams like the following one that shows two mutually exclusive events.

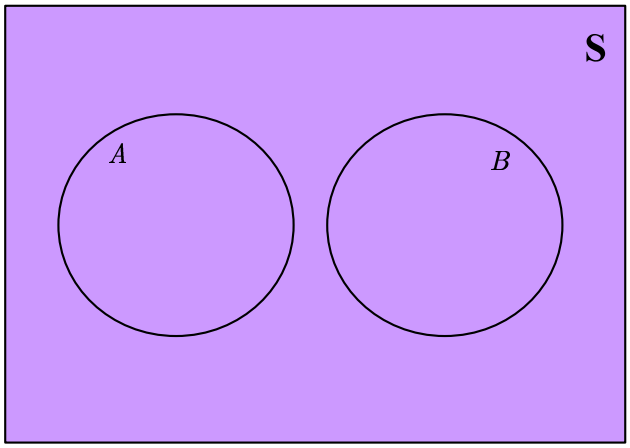


Figure 2.3: Mutually exclusive events

Note **S**, the sample space is context specific. In the previous example, S="Human", but in the example of the single die roll $S = \{1, 2, 3, 4, 5, 6\}$.

2.2 Operations on events

There are three main operations that one can perform on events:

1. Event intersection Consider the following figure: An *intersection* between events A and

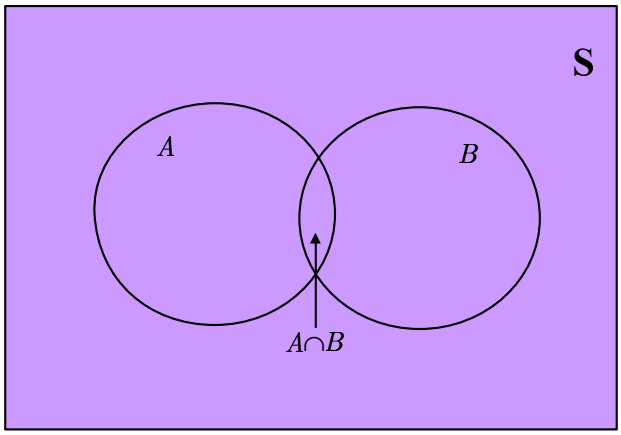


Figure 2.4: Intersection $A \cap B$

B are all the cases of overlap of the two events. For example if A="Face of die is odd" and B="Number is less than 3" then $A \cap B = \{1\}$. Note that if A and B are mutually exclusive, then their intersection is the null event (i.e., $A \cap B = \emptyset$).

2. Union of two events

The *union* of events A and B is comprised of all outcomes consistent to *either* A *or*

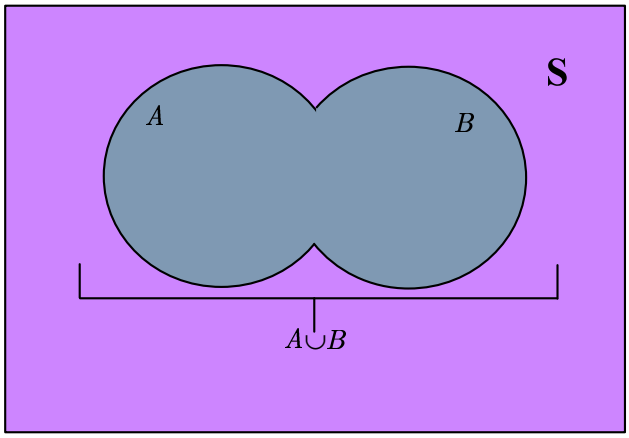


Figure 2.5: Union of events $A \cup B$

B or both. In the above example, the union of the two events A="Face of die is odd" and B="Number is less than 3" is $A \cup B = \{1, 2\}$ (Figure 2.5).

3. Complement of an event

The *complement of an event* A, denoted as A^c is comprised of all outcomes that are *not* compatible with A. Note that $A \cup A^c = S$ since all outcomes either will be contained in A or its complement (not A). This is seen by the following figure:

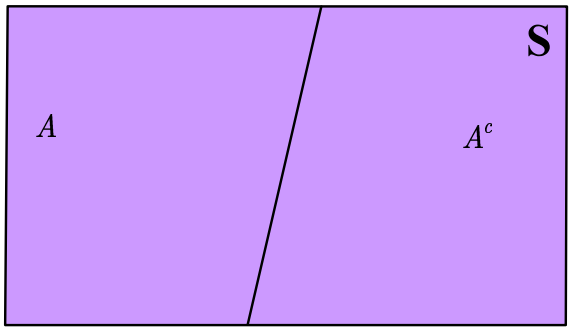


Figure 2.6: Complementary events

By the definition of the intersection, events A and A^c are mutually exclusive (i.e., $A \cap A^c = \emptyset$). This is because, there is no event that is consistent with both A and A^c (so that their intersection is the null event as shown above).

2.3 Probability

We define a measure of the likelihood of the occurrence of an event.

Definition: If an experiment is repeated n times under identical conditions, and event A occurs n_A times, then the probability that "event A occurs", denoted by P(A) is

$$P(A) = \frac{n_A}{n}$$

as n becomes large. According to this (the "frequentist") definition, probability is the "long-run frequency" of occurrence of the event.

For example, if A="Die comes up 1" then from Figure 1 we see that $n = 6$ and $n_A = 1$ so $P(A) = \frac{1}{6}$.

By the definition, $P(S) = 1$ and $P(\emptyset) = 0$ but in general $0 \leq P(A) \leq 1$.

2.3.1 Probabilities of special events

The following are probability calculations for some of the special or composite events that we discussed earlier:

1. Probability of the union of two events

   For any two events A, B it is always true

   $$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

   You can verify this visually from Figure 2.5, or by considering the fact that $A = (A \cap B^c) \cup (A \cap B)$ and $B = (B \cap A^c) \cup (A \cap B)$ so by taking the union we incorporate the intersection event $A \cap B$ twice in the calculations and thus need to remove it once. As a special case, when A and B are mutually exclusive events (since $A \cap B = \emptyset$) the above reduces to

   $$P(A \cup B) = P(A) + P(B)$$

   This is called the *additive rule*.

2. Probability of the complement of an event

   If A^c is the complementary event of A, then $P(A^c) = 1 - P(A)$.

   Proof:

   Since A^c="not A", either A occurs or A^c occurs. This means that $P(A \cup A^c) = 1$ (certainty). On the other hand not both events can occur at the same time. They are mutuallly exclusive. By the additive rule then $P(A \cup A^c) = P(A) + P(A^c) = 1$. It is easy to see then that $P(A^c) = 1 - P(A)$ (or equivalently, $P(A) = 1 - P(A^c)$).

2.4 Conditional Probability

The probability of an event B occurring conditional (or given) that event A occurs (or has occurred) is defined as:

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

and consequently, $P(A \cap B) = P(B|A)P(A)$.

Proof(Rozanov, 1977): By the definition of probability, $P(A) = \frac{n_A}{n}$ and $P(B) = \frac{n_b}{n}$. Now since A is given to occur, event B can only occur only among points that are compatible with the occurrence of A (i.e., here $n = n_A$). Also notice that given that A occurs, the occurrence of B means that both A and B will occur simultaneously (i.e., event $A \cap B$ will occur). By the definition of probability then,

$$P(B|A) = \frac{n_{A \cap B}}{n_A} = \frac{\frac{n_{A \cap B}}{n}}{\frac{n_A}{n}} = \frac{P(A \cap B)}{P(A)}$$

Consider the following events (Pagano & Gauvreau, 2000):

A= "A person in the U.S. is alive at age 60"

B= "A person in the U.S. will live to the age of 65"

Compute the probability of the event B | A="A 60 year-old person in the U.S. will live to the age of 65."

From life tables collected on the U.S. population, it is known that out of 100,000 individuals born, in 1988, 85,331 have reached 60 years of age while 79,123 have reached 65 years of age. Given the large n we can consider these proportions as reasonably accurate estimates of $P(A)$ and $P(B)$. That is,

$$P(A) = P(\text{"Lives to 60"}) \approx 0.85$$
$$P(B) = P(\text{"Lives to 65"}) \approx 0.79$$

Also, notice that $P(A \cap B) = P(\text{"Lives to 60" } \textit{and} \text{ "Lives to 65"}) = P(\text{"Lives to 65"}) = P(B) \approx 0.79$. Finally, $P(B|A) = \frac{P(A \cap B}{P(A)} = \frac{0.79}{0.85} \approx 0.93$. That is, a person has 79% chance of reaching 65 at birth, but a 60-year-old has 93% chance to reach the same age. The reason of course is that all situations where an individual would have died prior to having reached 60 years of age (i.e., the elements of S that are incompatible with A) have been excluded from the calculations (by the division with $P(A)$).

2.5 Independent events

Two events A and B are said to be *independent* if $P(B|A) = P(B)$. That is, knowledge that A has occurred does not affect the probability that B occurs. From the previous formula,

$$P(A \cap B) = P(A)P(B)$$

This is called the *multiplicative rule*.

For example, the event A="Heads" and B="Tails" as results of a coin toss are independent events. Having observed a head on the previous throw does not change the probability of tails in the current throw. That is, $P(B|A) = P(B) = 0.5$ and $P(A \cap B) = P(A) \times P(B) = 0.25$ (i.e., the sequence $\{H, T\}$ has probability 0.25).

2.6 Putting it all together

Consider the sum of two dice. The possible oucomes are

$$S = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$$

. A graphical presentation of S is presented in Table 2.1. What is the probability of the

Table 2.1: Sum of two dice

First	Second die					
die	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

event A="Sum of two dice is 7"?

To compute this probability, we must realize that each sum is a mutually exclusive event (since you cannot have 4+3 and 5+2 in the same toss), and thus, $P(A) = P[(1,6) \cup (2,5) \cup \ldots \cup (6,1)] = P(1,6) + P(2,5) + \ldots + P(6,1)$ by the additive rule. In addition, each die is rolled independently so for example, $P(1,6) = P(1 \cap 6) = P(1) \times P(6) = \left(\frac{1}{6}\right)\left(\frac{1}{6}\right) = \frac{1}{36}$, by the multiplicative rule. The same of course holds true for the other sums.

Thus,

$$P(\text{“Sum} = 7\text{”}) = \frac{1}{36} + \frac{1}{36} + \ldots + \frac{1}{36} = \frac{6}{36} = \frac{1}{6}$$

2.7 Diagnostic tests

Consider the following events:

- D = "Disease is present"
- D^c = "Disease is absent"
- T^+ = "Positive test result (test detects disease)"
- T^- = "Negative test result (test does not detect disease)"

In diagnostic-testing situations, the following "performance parameters" of the diagnostic procedure under consideration will be available:

- $P(T^+|D)$ = "Sensitivity (true positive rate) of the test"
- $P(T^+|D^c)$ = "Probability of a false positive test result"
- $P(T^-|D)$ = "Probability of a false negative test result"
- $P(T^-|D^c)$ = "Specificity (or true-negative rate) of the test"

In addition, in order to derive estimates of the PVP (i.e., the predictive value of a negative test $PVN = P(D^c|T^-)$) we will need an estimate of the overall probability of disease in the general population. This is called the *prevalence* of the disease $P(D)$.

Goal: Find $P(D|T^+)$ the predictive value of a positive test result (or PVP), that is, find the probability that a subject has the disease given a positive test.

2.7.1 X-ray screening for tuberculosis

In a large study 1820 individuals with or without tuberculosis (ascertained via an independent test) had an X-ray performed on them in order to ascertain the predictive ability of this examination (Pagano & Gauvreau, 2000). The situation is presented in Table 2.2. In addition, consider that the prevalence of the disease in the general population is $P(D) = 0.000093$ (i.e., 9.3 cases in 100,000).

From this table we can derive approximate estimates for the sensitivity and specificity of the X-ray as a diagnostic test. For example, $P(T^+|D) \approx \frac{22}{30} = 0.7333$. Notice that since D is *"given"*, the sample space is comprised only by the 30 positive cases in the first column. Similarly, $P(T^-|D^c) \approx \frac{1739}{1790} = 0.9715$.

Note You should not use as an estimate of prevalence the ratio $\frac{30}{1820} = 0.016$ from the table, as the 1820 subjects may not be representative of the general population. In fact the prevalence of TB in this hospital cohort is 1.6% or about 1,600 cases in 100,000. This is over 170 times higher than the prevalence of tuberculosis in the general population!

Table 2.2: X ray testing for the detection of tuberculosis

	Tuberculosis		
X-ray result	Yes	No	Total
Positive	22	51	73
Negative	18	1739	1747
Total	1790	30	1820

	Disease		
Test result	Yes	No	Total
Positive	$n_{D\cap T^+}$	$n_{D^c\cap T^+}$	n_{T^+}
Negative	$n_{D\cap T^-}$	$n_{D^c\cap T^-}$	n_{T^-}
Total	n_D	n_{D^c}	n

	Disease		
Test result	Yes	No	Total
Positive	$P(D\cap T^+)$	$P(D^c\cap T^+)$	$P(T^+)$
Negative	$P(D\cap T^-)$	$P(D^c\cap T^-)$	$P(T^-)$
Total	$P(D)$	$P(D^c)$	1

By similar calculations we obtain $P(T^-|D) =$ "false negative rate" $= \frac{8}{30} = 0.2667$ and $P(T^-|D) =$ "false positive rate"$= \frac{51}{1790} = 0.0285$.

Now let's calculate $P(D|T^+)$. By the definition of conditional probability,

$$P(D|T^+) = \frac{P(D \cap T^+)}{P(T^+)} = \frac{P(D|T^+)P(T^+)}{P(T^+)}$$

Since we do not know $P(T^+)$ let us consult the Figure 2.7. From the Figure it is seen that

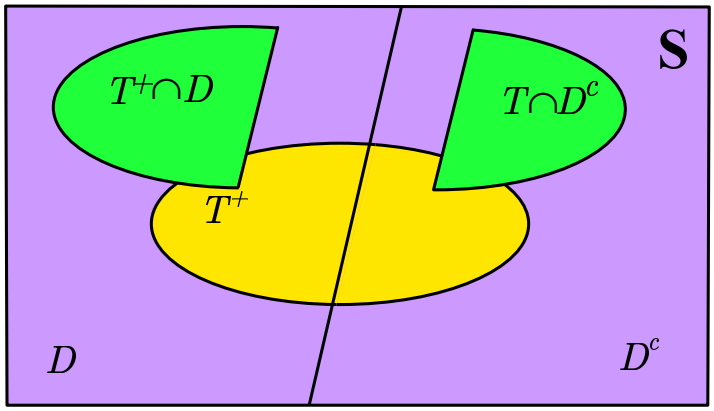


Figure 2.7: Components of T^+

$T^+ = (D \cap T^+) \cup (T^+ \cap D^c)$ so that

$$\begin{aligned} P(T^+) &= P\left[(D \cap T^+) \cup (T^+ \cap D^c)\right] \\ &= P(D \cap T^+) + P(T^+ \cap D^c) \\ &= P(T^+|D)P(D) + P(T^+|D^c)P(D^c) \end{aligned}$$

since $D \cap T^+$ and $T^+ \cap D^c$ are mutually exclusive events (using the additive rule). Then substituting above we have

$$\begin{aligned} P(D|T^+) &= \frac{P(D|T^+)P(D)}{P(T^+|D)P(D) + P(T^+|D^c)P(D^c)} \\ &= \frac{\text{sensitivity} \times \text{prevalence}}{\text{sensitivity} \times \text{prevalence} + \text{false positive} \times (1 - \text{prevalence})} \\ &= 0.00239 \end{aligned}$$

For every 100,000 positive x-rays, only 239 signal true cases of tuberculosis. This is called the "false positive paradox". Note also how we have incorporated the evidence from the positive X-ray in the calculation of the probability of tuberculosis.

Before the X-ray $P(D)$ = prior probability of disease = 0.000093. After the presence of

a positive test result we have $P(D|T^+)$ = posterior probability of disease (updated in the presence of evidence)= 0.00239. So, although the probability of tuberculosis is low, we have in fact reduced our degree of uncertainty 26-fold (0.00239/0.000093).

Now let's answer Marilyn's question. We have:

$$D = \text{“Drug user”} \qquad D^c = \text{“Not a drug user”}$$

$$T^+ = \text{“Positive drug test”} \qquad T^- = \text{“Negative test”}$$

We have also

$$P(\text{“Drug user”}) = P(D) = 0.05 \qquad \text{prevalence of drug use}$$

$$P(\text{“Positive test”} \mid \text{“Drug user”}) = P(T^+|D) \qquad \text{sensitivity of the test}$$

$$P(\text{“Negative test”} \mid \text{“Not a drug user”}) = P(T^-|D^c) \qquad \text{specificity of the test}$$

So finally,

$$\begin{aligned} P(\text{“Drug user”}|\text{“Positive test”}) &= \frac{P(D|T^+)P(D)}{P(T^+|D)P(D) + P(T^+|D^c)P(D^c)} \\ &= \frac{0.95 \times 0.05}{0.95 \times 0.05 + 0.05 \times 0.95} \\ &= 0.50 \end{aligned}$$

Why does this happen? To answer this consider a representative group from the general population as in Table 2.3. Approximately 48 ($\approx 50 \times 0.95$) out of the 50 drug users in this

Table 2.3: Expected number of drug users in 1,000 individuals randomly selected from the general population

	Drug use		
Drug test result	Yes	No	Total
Positive	48	48	96
Negative	3	903	906
Total	51	951	1002

group of 1,000 individuals, will test positive, but so will (by mistake; a false positive result) 48 ($\approx 0.05 \times 950$) of the 950 non drug users. Thus, only half of the 95 positive dru g test will have detected true cases of drug use (and thus $PVP \approx 50\%$). In general, when a disease (or, as in this case, drug use) is rare, even an accurate test will not easily reverse our initial (prior) low probability of its occurrence.

2.8 Bayes Theorem

If $A_1, A_2, \ldots, A_n$ are mutually exclusive events whose union is S (i.e., these events account for all possible outcomes or events without overlap), and suppose that the probabilities $P(B|A_i),\ P(A_i),\ i = 1 \ldots, n$ are known. Then, $P(A_i|B),\ i = 1, \ldots, n$ is given by

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B|A_1)P(A_1) + \ldots + P(B|A_i)P(A_i) + \ldots + P(B|A_n)P(A_n)}$$

It is easily seen that diagnostic testing is a special case of the Bayes Theorem. In the case of calculating the predictive value of a positive test (PVP), then $n = 2$ and $D \equiv A_1$, $D^c \equiv A_2$ and $T^+ \equiv B$. In the case of the PVN, then $T^- \equiv B$.

2.9 Bibliography

1. Principles of Biostatistics by M Pagano and K Gauvreau. Duxbury press

2. Probability Theory: A Consise Course by YA Rozanov, rev. English Edition. Edited and translated by RA Silverman. Dover publications Inc.

3. The Cartoon Guide to Statistics, by L Gonick and W Smith. Harper Collins

Chapter 3

Probability distributions

A random variable is a measurement whose observed values are the outcomes of a random experiment. In this sense, its values cannot be *a priori* determined. That is, we do not know what the values of the random variable are going to be before we collect the sample, run the experiment, etc.

The mechanism determining the probability or chance of observing each individual value of the random variable is called a probability distribution (as it literally distributes the probability among all the possible values of the random variables). Probability distributions are defined through frequency tables, graphs, or mathematical expressions.

There are two types of probability distributions corresponding to the two kinds of random variables:

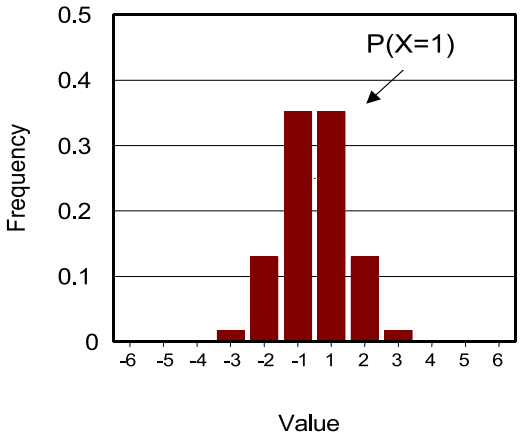
1. **Discrete probability distributions** (Figure 2A)
   These specify the chance of observing a small countable number of possible values (e.g., race, sex).

2. **Continuous probability distributions** (Figure 2B): These handle cases were all possible (real) numbers can be observed (e.g., height or weight). Note that large or infinite numbers of countable (i.e., discrete) values are usually handled by continuous distributions[1].

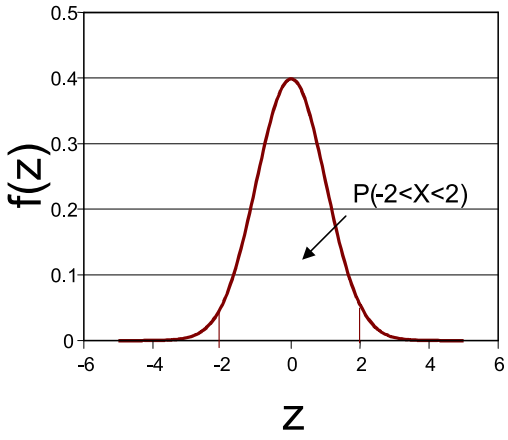
Note. Unlike the discrete case, in the case of a continuous distributions, the probability of observing any individual number is zero! Only probabilities of intervals have non-zero probability. Those probabilities are equal to the area between the x axis and the probability (density) curve.

[1] In fact one may argue that, given the finite precision with which measurements can be made, there are no truly continuous data!

Figure 3.1: Examples of probability distribution functions



Discrete probability distribution



Continuous probability distribution

3.1 Distributions of interest

3.1.1 The binomial distribution (discrete)

The binomial distribution describes how a number (n) of binary events (having only two possible outcomes) behave. These events, called Bernoulli trials have the following properties

1. Each event is identical to the others
2. All Bernoulli trial are mutually independent from all the others (i.e., information on the outcome of one does not affect the chances of any other)
3. There are two possible outcomes usually denoted as "success"=1 and "failure"=0
4. The probability of a "success" π is the same for all trials

The formula producing the probabilities of all possible arrangements of successes and failures is

$$P(X = j) = C_j^n \pi^j (1 - \pi)^{n-j}$$

where C_j^n is the number of ways of actually have j successes out of n trials. Actually, $C_j^n = \binom{n}{j} = \frac{n!}{j!(n-j)!}$. The notation $n! = n(n-1)...1$ is called "n factorial").

For example, if $n = 4$ and $j = 2$ then the possible ways to have two ones (successes) among four trials, is $4!/(2!2!) = 24/[(2)(2)] = 6$. Enumerating these we have: [1100], [1010], [1001], [0110], [0101], [0011].

Now if the probability of a success in each Bernoulli trial is $\pi = 0.5$ (say flipping a coin with "heads" considered as the "success") then the probability of two successes out of four trials is $P(X = 2) = (6)(0.5)^2(1 - 0.5)^{4-2} = (6)(0.25)(0.25) = 0.375$. In the coin-tossing experiment that would mean that there is about 38% probability to see two heads out of four tosses.

The mean of the binomial distribution $B(n, \pi)$ is

$$\mu = n\pi$$

This is intuitive. Consider the probability of "heads" $\pi = 0.5$ in a coin flip. Then if you toss the coin n times you would expect heads approximately half of the time. Less intuitive is the variance of the binomial distribution. This is given by

$$\sigma^2 = n\pi(1 - \pi)$$

The normal distribution (continuous)

The normal distribution $N(\mu_x, \sigma^2)$ is described by the following mathematical formula

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left[\frac{1}{2\sigma^2}(x - \mu_x)^2\right]$$

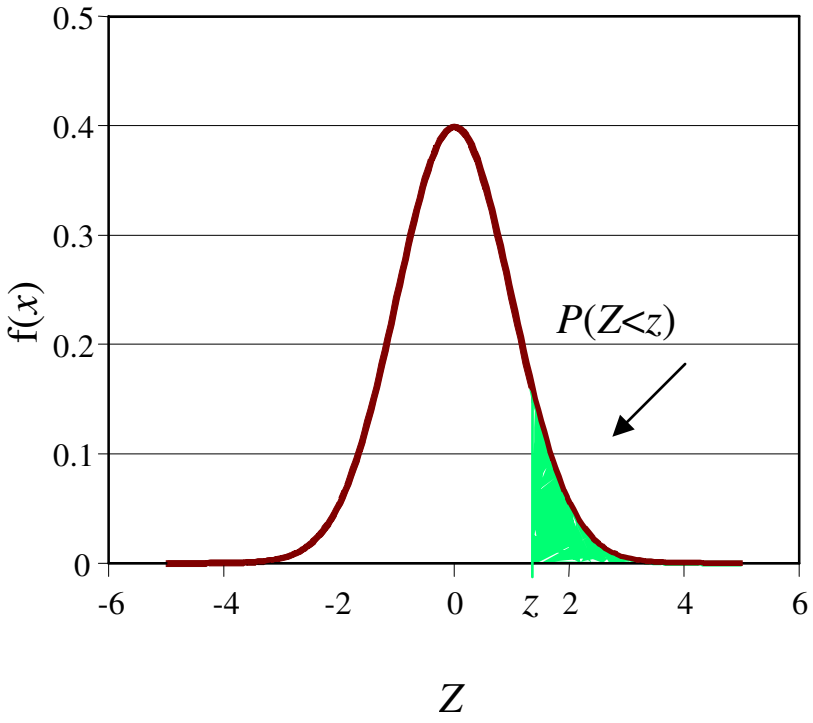
where μ_x and σ^2 are the population (parameters) mean and variance respectively. The function $f(x)$ is called a probability density function. It is symmetrical and centered around μ_x. Each probability is determined as the area between the density curve and the x axis (see Figure 2B).

The areas under the curve of the normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$ (the so-called "standard normal distribution") have been tabulated and are given in Table 3.1. This table presents probabilities in the tail of the standard normal distribution, i.e., $P(Z > z)$ for $z > 0.0$ (see Figure 3).

Table 3.1: Table A.l Areas in one tail of the standard normal curve

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.500	0.496	0.492	0.488	0.484	0.480	0.476	0.472	0.468	0.464
0.1	0.460	0.456	0.452	0.448	0.444	0.440	0.436	0.433	0.429	0.425
0.2	0.421	0.417	0.413	0.409	0.405	0.401	0.397	0.394	0.390	0.386
0.3	0.382	0.378	0.374	0.371	0.367	0.363	0.359	0.356	0.352	0.348
0.4	0.345	0.341	0.337	0.334	0.330	0.326	0.323	0.319	0.316	0.312
0.5	0.309	0.305	0.302	0.298	0.295	0.291	0.288	0.284	0.281	0.278
0.6	0.274	0.271	0.268	0.264	0.261	0.258	0.255	0.251	0.248	0.245
0.7	0.242	0.239	0.236	0.233	0.230	0.227	0.224	0.221	0.218	0.215
0.8	0.212	0.209	0.206	0.203	0.200	0.198	0.195	0.192	0.189	0.187
0.9	0.184	0.181	0.179	0.176	0.174	0.l7l	0.169	0.166	0.164	0.161
1.0	0.159	0.156	0.154	0.152	0.149	0.147	0.145	0.142	0.140	0.138
1.1	0.136	0.133	0.131	0.129	0.127	0.125	0.123	0.121	0.1l9	0.1l7
1.2	0.115	0.1l3	0.1ll	0.109	0.107	0.106	0.104	0.102	0.100	0.099
1.3	0.097	0.095	0.093	0.092	0.090	0.089	0.087	0.085	0.084	0.082
1.4	0.081	0.079	0.078	0.076	0.075	0.074	0.072	0.071	0.069	0.068
1.5	0.067	0.066	0.064	0.063	0.062	0.061	0.059	0.058	0.057	0.056
1.6	0.055	0.054	0.053	0.052	0.051	0.049	0.048	0.047	0.046	0.046
1.7	0.045	0.044	0.043	0.042	0.041	0.040	0.039	0.038	0.038	0.037
1.8	0.036	0.035	0.034	0.034	0.033	0.032	0.031	0.031	0.030	0.029
1.9	0.029	0.028	0.027	0.027	0.026	0.026	0.025	0.024	0.024	0.023
2.0	0.023	0.022	0.022	0.021	0.021	0.020	0.020	0.019	0.019	0.018
2.1	0.018	0.017	0.017	0.017	0.016	0.016	0.015	0.015	0.015	0.014
2.2	0.014	0.014	0.013	0.013	0.013	0.012	0.012	0.012	0.01l	0.01l
2.3	0.01l	0.010	0.010	0.010	0.010	0.009	0.009	0.009	0.009	0.008
2.4	0.008	0.008	0.008	0.008	0.007	0.007	0.007	0.007	0.007	0.006
2.5	0.006	0.006	0.006	0.006	0.006	0.005	0.005	0.005	0.005	0.005
2.6	0.005	0.005	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004
2.7	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003
2.8	0.003	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
2.9	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.001	0.001	0.001
3.0	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
3.1	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
3.2	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
3.3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3.4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Figure 3.2: Probabilities under the curve of the standard normal distribution



3.1.2 Reading the standard-normal table

To find a probability $P(Z > z)$ in the standard normal table that appears in the appendix of the textbook (or from the above Table) we search for z by proceeding down the left margin of the table going to a row that is just below z. Then we go across to a column that is as closest to z. The following figure helps clarify this for the case $P(Z > 0.16)$. This means

Figure 3.3: Probabilities under the curve of the standard normal distribution

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.500	0.496	0.492	0.488	0.484	0.480	0.476	0.472	0.468	0.464
0.1	0.460	0.456	0.452	0.448	0.444	0.440	0.436	0.433	0.429	0.425
0.2	0.421	0.417	0.413	0.409	0.405	0.401	0.397	0.394	0.390	0.386
0.3	0.382	0.378	0.374	0.371	0.367	0.363	0.359	0.356	0.352	0.348
0.4	0.345	0.341	0.337	0.334	0.330	0.326	0.323	0.319	0.316	0.312

that $P(Z > 0.16) = 0.436$. When reading a normal table, we take advantage of the following features of the normal distribution:

- The **symmetry** of the standard normal curve around zero (its mean). Thus, $P(Z \geq z) = P(Z \leq -z)$, where $z \geq 0$.

- The fact that (as in any distribution) the area under the curve is equal to 1. Thus, two complementary events, $P(Z \geq z) = 1 - P(Z \leq z)$.

We are usually faced with two problems:

1. Given a number $z \geq 0$[2] (say) find p such that the following is true:

   (a) $P(Z \geq z) = p$. To do this we read p directly from standard normal table

   (b) $P(Z \leq -z) = p$. In this case, we read $p_1 = P(Z \geq z)$ from the normal table, which by the symmetry of the normal distribution is equal to p

   (c) $P(Z \leq z) = p$. We read $p_1 = P(Z \geq z)$ from the normal table. Now $p = 1 - p_1$ since $P(Z \geq z)$ and $P(Z \leq z)$ are complementary events

   (d) $P(Z \geq -z) = p$. We Read $p_1 = P(Z \geq z)$ from the normal table and then $p = 1 - p_1$

   (e) Assuming that $z_1 \leq z_2$ we want to calculate $P(z_1 \leq Z \leq z_2) = p$. Since this is the area below z_2 i.e., $P(Z \leq z_2)$ with the "piece" $P(Z \leq z_1)$ removed, this is $P(z_1 \leq Z \leq z_2) = P(Z \leq z_2) - P(Z \leq z_1)$ (see above for the manner that these probabilities are calculated).

   In the special case $P(-z \leq Z \leq z) = 1 - 2P(Z > z)$

2. Given a probability p find z such that the following is true

   (a) $P(Z \geq z) = p$
   If $p \leq 0.5$. Then $z \geq 0$ and we look up p in the table. On the other hand, if $p \geq 0.5$ then $z \leq 0$ and we look up $p_1 = 1 - p$ in the table. z is the negative of the number located in the table

   (b) $P(Z \leq z) = p$
   If $p \leq 0.5$ then $z \leq 0$ and again we look up p in table. z is the negative of the number located there. On the other hand, if $p \geq 0.5$ then $z \geq 0$ and we look up $p_1 = 1 - p$ in the table.

   (c) $P(-z \leq Z \leq z) = p$. Look up $p_1 = (1-p)/2$ in the table. z is the closest number while $-z$ is its negative.

3.1.3 Examples

1. Find z such that $P(Z > z) = 0.025$. From above this can be looked-up directly in the standard normal table. We see that $z = 1.96$ is such that $P(Z > 1.96) = 0.025$.

2. Find z such that $P(Z < -z) = 0.05$. This is equal to $P(Z > z)$ which can be looked up on the table. We note that there are two numbers close to 0.05 but none fulfils the requirement exactly. We have $P(Z > 1.64) = 0.051$ while $P(Z > 1.65) = 0.049$. Interpolating between these two values we have $P(Z > 1.645) \approx 0.05$.

[2]Capital Z is the (normally distributed) random variable, while z is the values it assumes

3. Find z such that $P(-z < Z < z) = 0.95$. As above this probability is $1 - 2P(Z > z) = 0.95$ which means that $P(Z > z) = 0.025$ which means $z = 1.96$. That is, 95% of the area under the standard normal distribution is found between ± 1.96.

3.2 Standardization

A useful feature of the normal distribution is that if a variable X is distributed according to an arbitrary normal distribution $N(\mu, \sigma)$ then the variable $Z = \frac{X-\mu}{\sigma}$ is distributed as a standard normal distribution $N(0, 1)$ for which probabilities have been tabulated.

Intuitively this means that *for all normal distributions* the same amount of probability is concentrated under the normal distribution curve within the same number of standard deviations from the mean. Let's see how this works: In the case of the standard normal distribution, we know for example that 2.5% probability is concentrated above 1.96. That is, 2.5% probability is concentrated above 1.96 standard deviations above the mean (recall in the case of the standard normal, $\mu = 0$ and $\sigma = 1$). What we are saying is that *for any normal distribution* 2.5% probability is concentrated above $\mu + 1.96\sigma$, that is,

$$P\left(X > \mu + 1.96\sigma\right) = 0.025$$

Thus, any probability like $P(X > b)$ can be calculated by reference to the standard normal distribution, if one figures out how many standard deviations a is above the mean μ. This will be,

$$\begin{aligned} P\left(X > a\right) &= P\left(\frac{X-\mu}{\sigma} > \frac{a-\mu}{\sigma}\right) \\ &= P\left(Z > \frac{a-\mu}{\sigma}\right) \end{aligned}$$

where $Z = \frac{X-\mu}{\sigma}$ is the number of standard deviations above the mean. In other words, Z is distributed according to $N(0, 1)$. What the above says is that a is $z = \frac{a-\mu}{\sigma}$ standard deviations above zero. The probability associated with this event is of course easily obtained from the normal table in the textbook.

Other, more complex probabilities are obtained by simplifying the expression according to the methods that we discussed earlier. For example, recall the cholesterol level data, where the cholesterol level in the U.S. male population ages from 20-74 years was distributed according to the normal distribution $N(211, 46)$. What would be the probability that a randomly selected individual from this population has cholesterol level above $a = 220$? The answer is given if one thinks about how many standard deviations is a above the mean $\mu = 211$. That is,

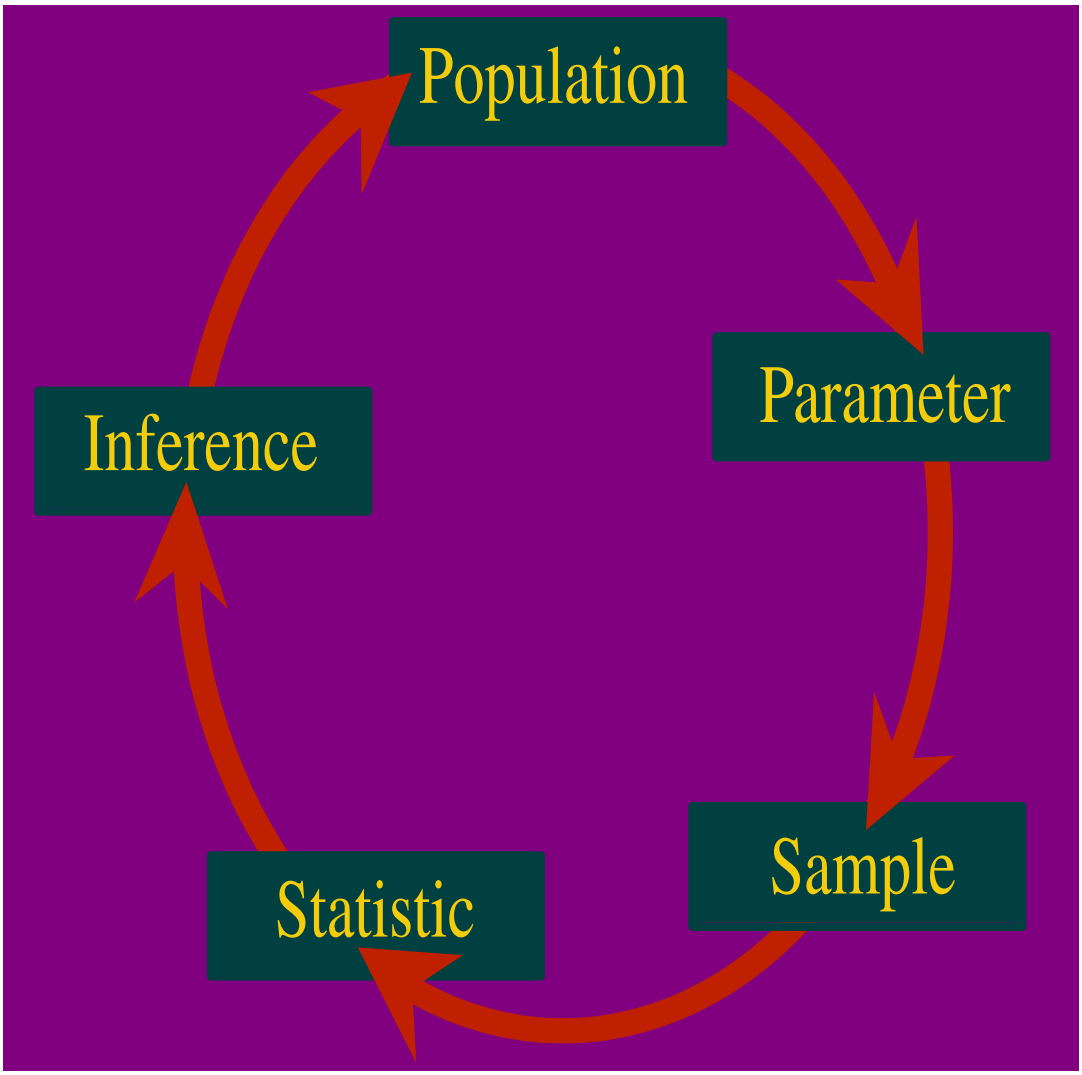
$$\begin{aligned} P\left(X > 220\right) &= P\left(\frac{X-\mu}{\sigma} > \frac{220-211}{46}\right) \\ &= P\left(Z > \frac{9}{46}\right) = P(Z > 0.196) \approx 0.42 \end{aligned}$$

That is, about 42% of the U.S. males ages 20-74 years-old have cholesterol above 220 mg/100 ml.

Chapter 4

Statistical inference

Figure 4.1: Progression of statistical analysis



Here we will make a parenthesis and introduce some basic statistical concepts.

1. **Population** is a set of measurements or items of interest, e.g., US males between the ages of 18-74, intravenous (IV) drug users, smokers, etc. A characteristic of the population is called a *parameter*

2. **Sample** is any subset from the population of interest A characteristic of the sample is called a *statistic*

That is, we are interested in a particular characteristic of the population (a parameter). To get an idea about the parameter, we select a (random) sample[1] and observe a related characteristic in the sample (a statistic). Then, based on assumption of the behavior of this statistic we make guesses about the related population parameter. This is called inference, since we infer something about the population. Statistical inference is performed in two ways: **Testing of hypotheses and estimation**

4.1 Sampling distributions

We mentioned that in order to make inferences about an unknown population parameter, such as the mean μ, you would need to take a random sample from the population and measure an appropriate quantity in the sample. In the case of the population mean, this quantity is the mean of the sample or *sample mean* $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Suppose now that one obtains repeated samples from the population and measures their means $\bar{x}_1, \bar{x}_2, ..., \bar{x}_n$. If you consider each sample mean as a single observation, from a random variable $\bar{X}$, then it will have a probability distribution. This is known as the *sampling distribution of means of samples of size n*.

4.2 The Central Limit Theorem

The central limit theorem is a powerful result allows the use of the normal distribution to make inferences. If the distribution of each observation in the population has mean μ and standard deviation σ *regardless of whether the distribution is normal or not*:

1. The distribution of the sample means $\bar{X}_n$ (from samples of size n taken from the population) has mean μ identical to that of the population

2. The standard deviation of this distribution is equal to $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$.

3. As n gets large, the shape of the sample distribution of the mean is approximately that of a normal distribution

[1] A random sample is one where every member of the population has equal chance of being selected

4.2.1 Cholesterol level in U.S. males 20-74 years old

The serum cholesterol levels for all 20-74 year-old US males has mean $\mu = 211$ mg/100 ml and the standard deviation is $\sigma = 46$ mg/100 ml. That is, each individual serum cholesterol level is distributed around $\mu = 211$ mg/100 ml, with variability expressed by the standard deviation σ.

Let's say that we take a sample size of size $n = 25$.

What if $\bar{x} = 217$ mg/100 ml?
$\bar{x} = 220$ mg/100 ml?
$\bar{x} = 230$ mg/100 ml?

If $\mu = 217$ mg/100 ml, then from the Central Limit theorem we have that

$$P(\bar{X} \geq 217) = P\left(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \geq \frac{217 - 211}{\frac{46}{\sqrt{25}}}\right) = P(Z \geq 0.65) = 0.258$$

Similarly $P(\bar{X} \geq 220) = P(Z \geq 0.98) = 0.164$ and, $P(\bar{X} \geq 230) = P(\bar{X} \geq 2.07) = 0.019$

Thus, less than 26% of the time will the means of the samples of size 25 will be above 217 mg/100 ml, about 16% of the time they will be above 220 mg/100 ml and less than 2% are the sample means expected to be larger than 230mg/100 ml.

To calculate the upper and lower cutoff points enclosing the middle 95% of the means of samples of size $n = 25$ drawn from this population we work as follows:
The cutoff points in the standard normal distribution are -1.96 and $+1.96$. We can translate this to a statement about serum cholesterol levels.

$$\begin{aligned} -1.96 \leq Z \leq 1.96 &\iff -1.96 \leq \frac{\bar{x}_{25} - 211}{\frac{46}{\sqrt{25}}} \leq 1.96 \\ &\iff 211 - 1.96(9.2) \leq \bar{x}_{25} \leq 211 - 1.96(9.2) \\ &\iff 1.93 \leq \bar{x}_{25} \leq 229.0 \end{aligned}$$

Approximately 95% of the sample means will fall between 193 and 229 mg/100 ml.

Note. This is a general result, i.e., 95% of *any* normal distribution is between $\mu \pm 1.96\sigma$. Here, $\sigma = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ and $\mu = \bar{x}$.

4.2.2 Level of glucose in the blood of diabetic patients

From previous studies it is known that in the general population the level of glucose in the blood has mean $\mu = 9.7$ mmol/L and standard deviation $\sigma = 2.0$ mmol/L.

In a group of $n = 64$ diabetic patients the average level of glucose is $\bar{X}_{64} = 13.6$ mmol/L. Assuming that diabetic patients do not have higher glucose levels in their blood compared to the rest of the population, what is the probability $P(\bar{X} > 13.6)$ if $\bar{X}_{\text{diabetic}} = \bar{X}_{\text{healthy}}$?

From the Central Limit Theorem, the distribution of $\bar{X}_{\text{diabetic}}$ is Normal with mean $\mu = 9.7$ mmol/L and standard deviation $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} =$. Thus,

$$\begin{aligned} P(\bar{X}_{64} > 13.6) &= P\left(\frac{\bar{X}_{64} - \mu}{\frac{\sigma}{\sqrt{n}}} > \frac{13.6 - 9.7}{\frac{2}{\sqrt{64}}}\right) \\ &= P\left(Z > \frac{3.9}{0.25}\right) = P(Z > 15.6) \end{aligned}$$

This is equivalent to asking what the probability is that a number is 15.6 standard deviations away from the mean. This of course is essentially zero!

Is this compatible with the hypothesis that diabetic patients have the same glucose levels as the rest of the population? Most people would say that this probability is "too small" or that the mean in the diabetics sample is "too far" from the hypothesized mean (of the healthy population), so that the hypothesis of equality of the diabetic and healthy means is suspect.

4.3 Hypothesis testing

In cases like the diabetic patients example statistics is hardly needed. However, in other cases answers are not as clear. We need a rigorous procedure to test statistical hypotheses. The steps involved in formally testing a statistical hypothesis are as follows:

1. State the null hypothesis H_o. Usually we will try to disprove it (i.e., "reject" it).

2. State the alternative hypothesis H_a.

3. Determine the α level of the test. This is the lowest level of probability resulting from assuming the null hypothesis is true, that you are willing to consider, before rejecting the null hypothesis (as having led you to a very unlikely event)

4. Specify the statistic T on which the test is based. In the cases that we are concerned with, this statistic is of the form

$$T = \frac{\hat{\theta} - \theta}{\text{s.e.}\hat{\theta}}$$

   where θ and $\hat{\theta}$ are the population parameter and sample statistic respectively, and s.e.(θ) the "standard error" is the standard deviation of the statistic $\hat{\theta}$.

5. Specify the decision rule for rejecting or not the null hypothesis. This must be based on the α level of the test and the test statistic T.

4.3.1 Hypothesis testing involving a single mean and known variance

Based on a random sample of size n we compute the sample mean $\bar{X}_n$. The testing of hypothesis in this case is carried out as follows:

1. The null hypothesis is
   (a) One-sided tests: $H_0 : \mu \geq \mu_0$ or $H_a : \mu \leq \mu_0$
   (b) Two-sided tests: $H_0 : \mu = \mu_0$
2. The alternative hypothesis is
   (a) One-sided tests: $H_a : \mu < \mu_0$ or $H_a : \mu > \mu_0$
   (b) Two-sided tests: $H_a : \mu \neq \mu_0$
3. Usually the α level will be 5% or 1% (the significance level of the test is $(1-\alpha)\%$, i.e., 95% or 99% respectively).
4. The test is based on the statistic
$$T = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$
   which is distributed according to the standard normal distribution.
5. Rejection rule: Reject the null hypothesis,
   (a) One-sided tests. If $T = \frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}} > z_{1-\alpha}$ or if $T < z_\alpha$ respectively
   (b) Two-sided tests. If $|T| > z_{1-\frac{\alpha}{2}}$

where $z_{1-\alpha}$ is the upper $(1-\alpha)\%$ tail and z_α is the lower tail of the standard normal distribution respectively.

Example: Level of glucose in the blood of diabetic patients (continued)

Based on a random sample of $n = 64$ diabetic patients with sample mean $\mu = 13.6$ mmol/L, we may ask the question:
Is it likely that the sample comes from a population with mean $\mu \leq 9.7$mmol/L (i.e., "no higher than" 9.7mmol/L) or from a population with a higher mean than that?

To test this question (hypothesis) we proceed as follows:

1. The null hypothesis is $H_o : \mu_{\text{diabetic}} \leq \mu_0 = \mu_{\text{healthy}} = 9.7$ mmol/L. This means that the diabetic population mean glucose level is at most that of the normal population if not lower
2. The alternative hypothesis is $H_a : \mu > \mu_0$ which means that the mean glucose level among diabetics is *higher than normal*

3. Let us choose $\alpha = 0.05$ (significance level is 95%)

4. The test statistic is $T = \frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{13.6-9.7}{\frac{2}{\sqrt{64}}} = 15.6$

5. Rejection rule (this is a one-sided test): Reject the null hypothesis if $T > 1.645 = z_{0.95}$

   Decision: Since $T = 15.6 > 1.645$ we reject H_o.

The data contradict the null hypothesis that diabetic patients have the same blood glucose level as healthy patients. On the contrary, the data suggest that diabetics have *significantly higher* glucose levels on average than individuals not suffering from diabetes.

4.4 Implications of each step in hypothesis testing

STEP 1. State the null hypothesis $H_0 : \mu = \mu_o$

By assuming that the mean is μ_o with known std. deviation σ (and sample size n) leads explicitly to the definition of the distribution of the sample mean $\bar{X}$. According to the Central Limit Theorem, this distribution will be normal with mean μ_o and std. deviation $\frac{\sigma}{\sqrt{n}}$. The sampling distribution of the mean dictates the probability of observing each sample

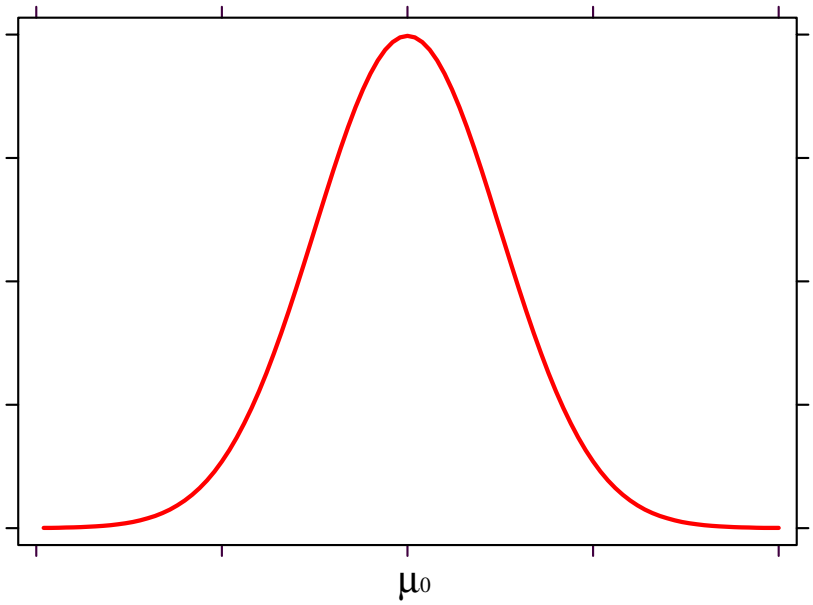


Figure 4.2: Sampling distribution of the mean under H_0

mean value.

STEP 2. State the alternative hypothesis

$H_a : \mu > \mu_0$ (other alternatives are possible)

STEP 3. Choose the α level of the test

Graphically, STEPS 2 and 3 are shown in Figure 4.3.

Steps 2 and 3 determine the location of the cutoff point(s) of the test. Step 2 implies that the

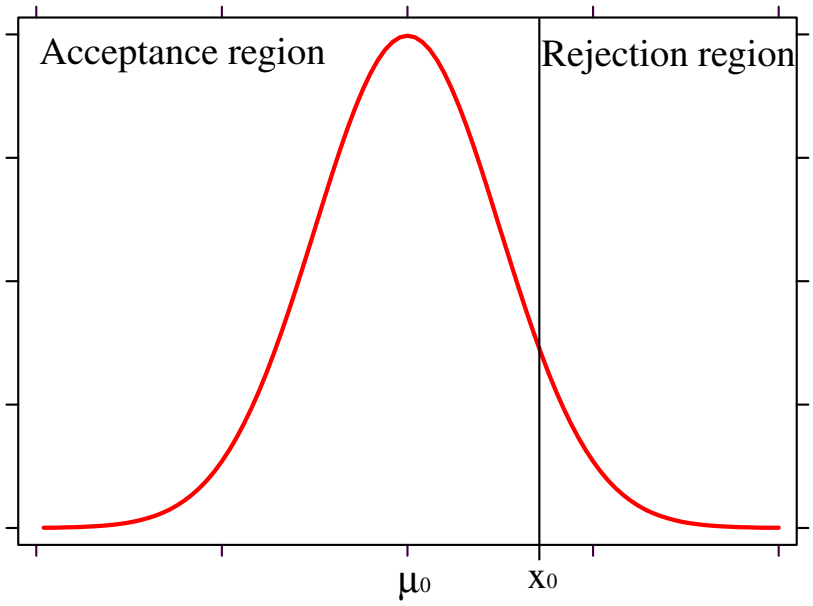


Figure 4.3: Impact of STEPS 2 and 3 on our assumptions

cutoff point $\bar{x_0}$ will be on the right tail of the sample mean distribution. Any observed value of above this point will raise suspicion about the veracity of the null hypothesis. Steps 2 and 3 have implications for the rejection rule. Step 3 determines how certain we want to be of our decision. A small alpha level indicates that we would be willing to reject the null hypothesis only for extremely unlikely values of $\bar{X}$. Larger alpha levels indicate a willingness to reject more easily. Compare this to a jury verdict. In the first case, we would want to be extra certain, while in the latter we would convict with weaker evidence. Calculation of the cutoff point $\bar{x_0}$ proceeds by translating the statement $P\left(\bar{X}_n > \bar{x_0}\right) = \alpha$, to a statement about Z (for which cutoff points have been tabulated). Since $P(Z > z_{1-\alpha}) = \alpha$, and $Z = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$, we have that $P\left(\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} > z_{1-\alpha}\right) = \alpha \equiv P\left(\bar{X}_n > \mu + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}\right) = \alpha$. This in turn immediately implies that $\bar{x}_0 = \mu + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$.

So, given α, we would go up to $z_{1-\alpha}$ std. deviations above the mean before rejecting the null hypothesis (in favor of the one-sided alternative $H_a : \mu > \mu_0$) at this α level.

If $H_a : \mu < \mu_0$ then the cutoff point will be $\bar{x}_0 = \mu - z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$. Thus, we reject H_0 for values of that are $z_{1-\alpha}$ std. deviations below the mean.

If the test is two-sided (alternative hypothesis of the form $H_a : \mu \neq \mu_0$) then the situation is as shown in Figure 4.4($\bar{x}_l$ and $\bar{x}_u$ are the lower and upper cutoff points respectively). Note now that the sum of the two tails is α, i.e. $P\left(\bar{X}_n < \bar{x}_l\right) + P\left(\bar{X}_n > \bar{x}_u\right) = \alpha$.

The point $\bar{x}_l$ is such that $P\left(\bar{X}_n < \bar{x}_l\right) = \frac{\alpha}{2}$ and the point $\bar{x}_u$ is such that $P\left(\bar{X}_n > \bar{x}_u\right) = \frac{\alpha}{2}$. Working in a similar manner as before, we see that since $P\left(Z > z_{1-\frac{\alpha}{2}}\right) = \frac{\alpha}{2}$, and

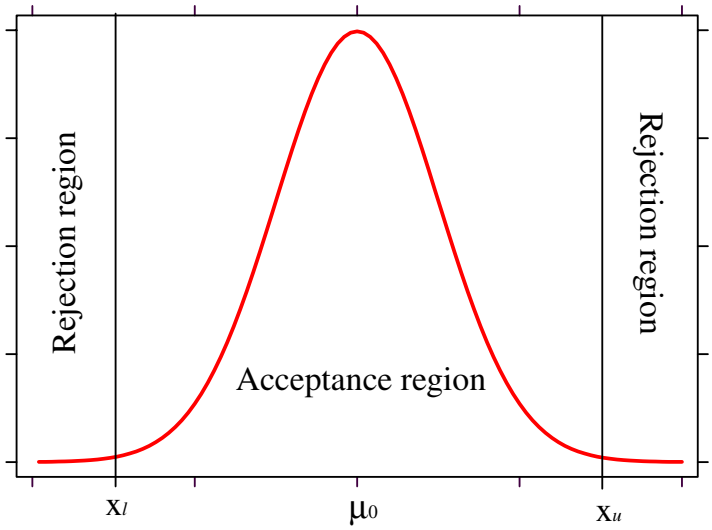


Figure 4.4: A two-sided alternative

$P\left(Z < -z_{1-\frac{\alpha}{2}}\right) = \frac{\alpha}{2}$, we have that $\bar{x}_l = \mu_0 - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$. Similarly, we have that $\bar{x}_u = \mu_0 + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$. This means that, given α, we would reject the null hypothesis if were $z_{1-\frac{\alpha}{2}}$ standard deviations above or below the mean.

4.4.1 Diabetes example

In the diabetes example, we assume that in the general population, the mean glucose level is $\mu_0 = 9.7$ mmol/L, with std. deviation $\sigma = 2.0$ mmol/L. If a sample of size $n = 64$ is selected the null hypothesis that diabetic persons have the same glucose level as healthy individuals $H_0 : \mu = \mu_0$ implies That is, $\bar{X}_{64}$ is distributed according to a normal distribution $N(9.7, 0.25)$ (recall that $\sigma_{\bar{x}=\frac{\sigma}{\sqrt{n}}}$ so, in this case it is $\sigma_{\bar{x}} = \frac{2.0}{\sqrt{64}} = \frac{2.0}{8} = 0.25$). The alternative hypothesis is $H_a : \mu > \mu_0$, along with an alpha level $\alpha = 0.05$ will produce a cutoff point $\bar{x}_0 = \mu_0 + z_{1-\alpha}\sigma_{\bar{x}} = 9.7 + 1.645(0.25) = 10.11$. The situation is shown in Figure 4.6.

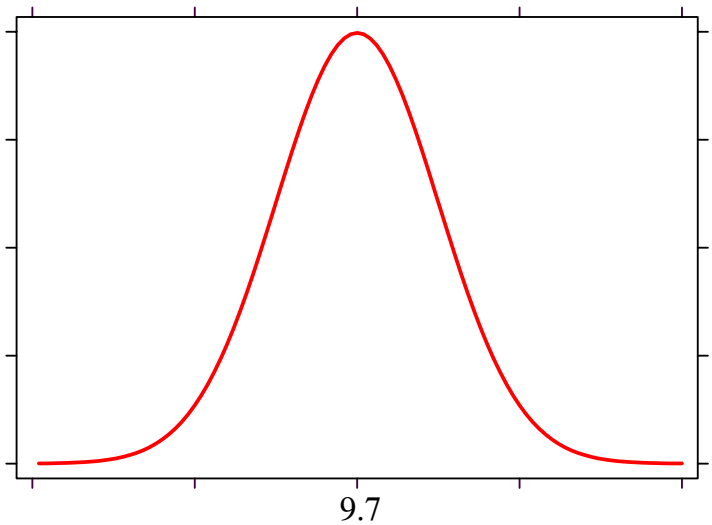


Figure 4.5: The distribution of the sample mean in the diabetes example

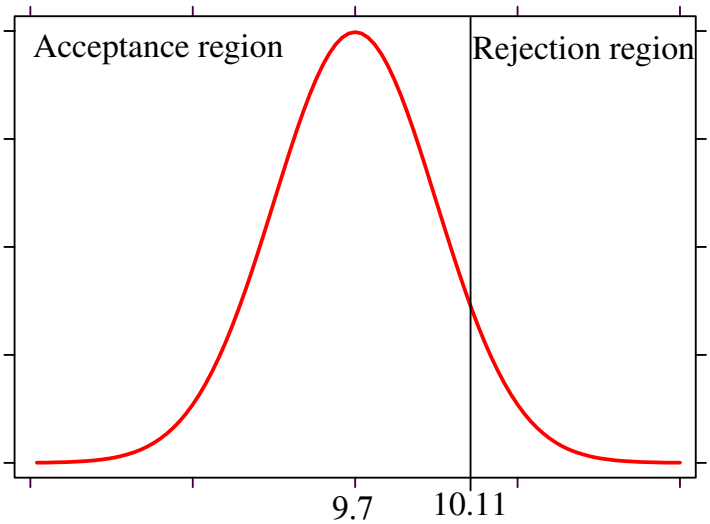


Figure 4.6: The diabetes example under a one-sided alternative and $\alpha = 0.05$

The cutoff point is $\bar{x}_0 = 10.11$ mmol/L. Since the observed value of was $\bar{x}_{64} = 13.6$ mmol/L, we reject H_0.

4.5 Hypothesis testing involving means and unknown variance

4.5.1 Concentration of benzene in a cigar

Suppose now that we want to determine whether the concentration of benzene in a brand of cigars is the same as that of cigarettes. Suppose further that we know that the mean concentration of benzene in cigarettes is $\mu = 81$g/g of tobacco, but are unsure of the variability of that measurement in cigars. Had we known , the test would be based on the statistic $Z = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$. Since we do not, we must estimate it, using the sample standard deviation $s = \sqrt{\frac{1}{n-1}\sum_{i=1} n(x_i - \bar{x}_n)^2}$. We can then "plug in" s into the previous test statistic and use the statistic t

$$\boxed{T = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}}$$

Note however, that we are given less information than we were, when the population standard deviation was known. Thus, T is not distributed according to a standard normal distribution. In fact we should expect T to be more variable than Z, and its distribution should reflect this.

The distribution of T is called the Student's t distribution (or t distribution).

The t distribution is symmetric, and centered around zero, it has "fatter" tails compared to the standard normal distribution and is defined by $n - 1$ "degrees of freedom" (where n is the sample size). Notice in Figure 6 how the t distribution approaches the standard normal distribution as the degrees of freedom increase. This is intuitively as expected, since when

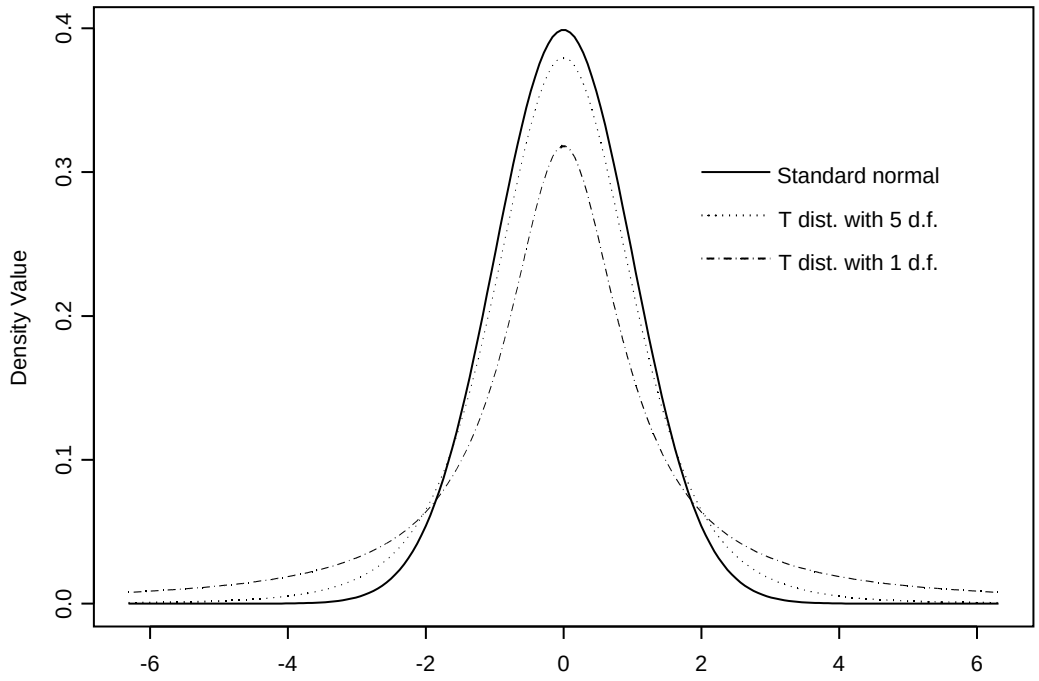


Figure 4.7: t distribution and standard normal distribution

we have a large sample size n, then the information increases (and thus the uncertainty introduced from having to estimate the standard deviation decreases). The degrees of freedom are essentially the number of independent pieces of information provided by the sample. Initially, every sample has n independent pieces of information (as many as the number of observations). However, after we calculate the sample mean, there are only $n-1$ independent pieces. Recall that $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$. Thus, if we know the first $n-1$ observations, we can compute the nth one (that would be $x_n = \bar{x} - \sum_{i=1}^{n-1}(x_i - \bar{x})$), and thus there are $n-1$ independent pieces of information. The test of hypotheses involving means with unknown variance proceeds as follows:

Based on a random sample of size n compute the sample mean $\bar{x}_n$

1. State the null hypothesis
   (a) *One-sided tests:* $H_0 : \mu \leq \mu_0$ or $H_0 : \mu \geq \mu_0$
   (b) *Two-sided tests:* $H_0 : \mu = \mu_0$
2. Set up the alternative hypothesis
   (a) *One-sided tests:* $H_a : \mu < \mu_0$ or $H_a : \mu > \mu_0$

(b) *Two-sided tests:* $H_a : \mu \neq \mu_0$

3. Choose the α level (the significance level of the test is $(1-\alpha)\%$).

4. The test statistic on which testing is based is $T = \frac{\bar{x}_n - \mu_0}{\frac{s}{\sqrt{n}}}$

5. Rejection rule: Rejection of the null hypothesis.

   (a) *One-sided tests:* Reject if $T > t_{n-1\,1-\alpha}$ or if $T < -t_{n-1;1-\alpha}$

   (b) *Two-sided tests:* Reject if $T > t_{n-1;1-\frac{\alpha}{2}}$ or if $T < -t_{n-1;1-\frac{\alpha}{2}}$

4.5.2 Concentration of benzene in cigars

A random sample of $n = 7$ cigars had mean benzene concentration $\bar{x}_7 = 151\mu$g/g and std. deviation $s = 9\mu$g/g. Is it possible that the benzene concentration is the same as that of the cigarettes (that have mean benzene concentration level $\mu = 81\mu$g/g)? To answer this question, we proceed as follows:

1. $H_0 : \mu = \mu_{\text{cigars}} = \mu_{\text{cigarettes}} = \mu_0 = 81\mu$g/g

2. $H_a : \mu \neq \mu_0$

3. The alpha level of the test is 5%

The question is:"What is the probability that the cigar population mean benzene concentration is $\mu = 81\mu$g/g?

Since $t = \frac{\bar{x}_7 - \mu_0}{\frac{s}{\sqrt{n}}}$ is distributed as a t distribution with $n - 1 = 6$ degrees of freedom, and $t = \frac{151-81}{\frac{9}{\sqrt{7}}} = 20.6$, the probability that a sample mean of 151 or higher would occur under the null hypothesis is less than 0.0001.

Since this is less than the alpha level of the test we reject the null hypothesis. Cigars have higher concentration of benzene than cigarettes.

4.5.3 Computer implementation

To carry out the above test of hypothesis by STATA we use the following command:

```
ttesti #obs #mean #sd #val
```

where `#obs` is the sample size, `#mean` is the sample mean, `#sd` is the sample standard deviation, and `#val` is the population mean under the null hypothesis.

Computer implementation of the benzene concentration example

```
. ttesti 7 151 9 81, level(95)

                                                         Number of obs =       7

------------------------------------------------------------------------------
Variable |      Mean    Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |       151     3.40168   44.3898   0.0000       142.6764    159.3236
------------------------------------------------------------------------------
Degrees of freedom: 6

                           Ho: mean(x) = 81

     Ha: mean < 81          Ha: mean ~= 81            Ha: mean > 81
        t =  20.5781             t =  20.5781            t =  20.5781
    P < t =   1.0000       P > |t| =   0.0000        P > t =   0.0000
```

Since we are performing a two-sided test, we concentrate in the middle part of the STATA output. Since `P > |t| = 0.0000`, which is much smaller than 0.05, we reject the null hypothesis.

4.6 Analyses involving two independent Samples

In the following table, the population parameters and sample statistics derived from two groups under comparison are listed.

	Group 1	Group 2
Population		
mean	μ_1	μ_2
std. deviation	σ_1	σ_2
Sample		
mean	$\bar{X}_1$	$\bar{X}_2$
std. deviation	s_1	s_2
sample size	n_1	n_2

4.6.1 Serum iron levels and cystic fibrosis

Consider the comparison of the serum iron levels of healthy children versus children suffering from cystic fibrosis (Pagano & Gauvreau, 2000). The (population) mean of the serum iron level among healthy children is μ_1 while the mean serum iron level among children suffering from cystic fibrosis is μ_2. Comparison of these unknown means is performed by taking two samples of size $n_1 = 9$ and $n_2 = 13$ children from the two populations.

In the case of two independent samples consider the following issues:

1. The two sets of measurements are independent (because each comes from a different group (e.g., healthy children, children suffering from cystic fibrosis).

2. In contrast to the one-sample case, we are simultaneously estimating two population means instead of one. Thus, there are now two sources of variability instead of one (one from each sample) instead of just one as was the case in the one-sample tests. As a result, the standard deviation is going to be roughly double (!) compared to the one-sample case.

When comparing two independent samples, the following assumptions must hold:

1. The two samples must be independent from each other
2. The individual measurements must be roughly normally distributed
3. The variances in the two populations must be roughly equal
4. If 1-3 are satisfied, inference will be based on the statistic

$$\boxed{T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}}$$

T is distributed according to a t distribution with $n_1 + n_2 - 2$ degrees of freedom.

In the above calculations $s_p^2 = \frac{n_1-1}{n_1+n_2-2} s_1^2 + \frac{n_2-1}{n_1+n_2-2} s_2^2 = \frac{\sum_{j=1}^{n}(x_{j1}-\bar{x}_1)^2 + \sum_{j=1}^{n}(x_{j2}-\bar{x}_2)^2}{n_1+n_2-2}$ where, $s_1^2 = \frac{1}{n_1-1}\sum_{j=1}^{n}(x_{j1} - \bar{x}_1)^2$ and $s_2^2 = \frac{1}{n_2-1}\sum_{j=1}^{n}(x_{j2} - \bar{x}_2)^2$ are the sample standard deviations in each sample respectively.

4.6.2 Testing of two independent samples (assuming equal variances)

STEP 1. Based on two random samples of size n_1 and n_2 observations compute the sample means $\bar{x}_1$ and $\bar{x}_2$, and the std. deviations s_1 and s_2.

STEP 2. Compute the pooled estimate of the population variance s_p^2 The pooled estimate of the standard deviation is $s_p = \sqrt{s_p^2}$.

Testing for difference of the means of two independent samples (assuming equal variances) proceeds as follows:

1. State the null hypothesis.
   (a) *One-sided tests:* $H_0 : \mu_1 \geq \mu_2$ or $H_0 : \mu_1 \leq \mu_2$
   (b) *Two-sided tests:* $H_0 : \mu_1 = \mu_2$
2. Set up the alternative hypothesis
   (a) *One-sided tests:* $H_a : \mu_1 < \mu_2$ or $H_a : \mu_1 > \mu_2$
   (b) *Two-sided tests:* $H_a : \mu_1 \neq \mu_2$
3. Choose the α level (the significance level of the test is $(1 - \alpha\%)$).

4. Rejection rule: Based on the observed value of T reject the null hypothesis, if

   (a) *One-sided tests:* If $T > t_{n_1+n_2-2;1-\alpha}$ or if $T < -t_{n_1+n_2-2;1-\alpha}$

   (b) *Two-sided tests:* If $|T| > t_{n_1+n_2-2;1-\frac{\alpha}{2}}$ (i.e., if $T < -t_{n_1+n_2-2;1-\frac{\alpha}{2}}$ or $T > t_{n_1+n_2-2;1-\frac{\alpha}{2}}$).

Example: Serum iron levels and cystic fibrosis (continued)

In this example we have $n_1 = 9$ and $n_2 = 13$ and $s_1 = 5.9\mu$mol/l, plus $\bar{x}_1 = 18.9\mu$mol/l and $\bar{x}_2 = 11.9\mu$mol/l and $s_2 = 6.3\mu$mol/l.

Based on two samples the pooled estimate of the population variance

$$\begin{aligned} s_p^2 &= \frac{n_1-1}{n_1+n_2-2}s_1^2 + \frac{n_2-1}{n_1+n_2-2}s_2^2 \\ &= \frac{(8)(5.9)^2+(12)(6.3)^2}{(9+13-2)} \\ &= 37.74 \end{aligned}$$

The estimate of the standard deviation is $s_p = \sqrt{37.74} = 6.14\mu$mol/l. The test of hypothesis is carried out as follows:

1. State the null hypothesis $H_0 : \mu 1 = \mu_2$
2. Set up the two-sided alternative hypothesis: $H_a : \mu_1 \neq \mu_2$
3. The α level is 5% (the significance level of the test is 95%)
4. The test statistic is $T = \frac{(18.9-11.9)}{6.14\sqrt{\frac{1}{9}+\frac{1}{13}}} = 2.63$
5. Rejection rule: Reject H_0, if $T > t_{20;0.975} = 2.086$ or if $T < -t_{20;0.975} = -2.086$. Since $T = 2.63 > 2.086$ we reject the null hypothesis.

That is, we are 95% sure that children suffering from cystic fibrosis have *significantly different* levels of iron in their serum compared to healthy children. It appears that these children have an iron deficiency. To carry out the above test of hypothesis by STATA we use the following command:

```
ttesti #obs1 #mean1 #sd1 #obs2 #mean2 sd2
```

where `#obs1` and `#obs2` are the sample sizes, `#mean1` and `#mean2` are the sample means, and `#sd1` and `#sd2` are the sample standard deviations for the two groups respectively.

Note. ttesti is the immediate version of the `ttest` command in STATA. We use the immediate versions of commands, when we do not have access to the raw data, but we do have access to the necessary summary statistics (like n, mean, standard deviation, etc.). If we had access to the raw data, say under variable names `X1` and `X2`, then the previous `ttest` command would be `ttest X1=X2` (and STATA would then proceed to calculate the means and standard deviations necessary). The computer output is as follows:

```
. ttesti 9 18.9 5.9 13 11.9 6.3
                                                    x: Number of obs =        9
                                                    y: Number of obs =       13
------------------------------------------------------------------------------
Variable |      Mean    Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |      18.9    1.966667   9.61017   0.0000       14.36486    23.43514
       y |      11.9    1.747306   6.81049   0.0000       8.092948    15.70705
---------+--------------------------------------------------------------------
    diff |         7    2.663838   2.62779   0.0161       1.443331    12.55667
------------------------------------------------------------------------------
Degrees of freedom: 20
                     Ho: mean(x) - mean(y) = diff = 0
Ha: diff < 0           Ha: diff ~= 0              Ha: diff > 0
        t =   2.6278             t =   2.6278             t =   2.6278
    P < t =   0.9919       P > |t| =   0.0161         P > t =   0.0081
```

The two-sided test corresponds to the middle alternative (`Ha:diff = 0`). The p-value (`P > |t| = 0.0161`) is less than the α level, so we reject H_0. Children with cystic fibrosis (group `y`) have different levels of iron in their blood from healthy children. The sample mean is less than that of the healthy children meaning that children with cystic fibrosis have lower blood iron levels.

4.6.3 Paired samples

Sixty-three adult males suffering from coronary artery disease were tested. The test involved challenging the subjects' cardiovascular system by riding a stationary bicycle until the onset of angina (chest pain). After resting, the subjects rode again until the repeat onset of angina and the percent of time of earlier onset of pain was recorded. On the first visit subjects were breathing clean air, while on a subsequent visit, the subject repeated the same series of tests, but CO was mixed in the air. The percent difference in the time to the onset of angina on the first series of tests (when breathing regular air) and the percent difference of time to onset of angina during the second series of tests (when breathing air mixed with CO) were compared.

In this study, each patient accounts for a pair of observations, so there are two issues to consider:

1. The two sets of measurements are not independent (because each pair is measured on the same patient) and each patient serves as his own "control". The advantage of this design is that we are able to account for individual (biological) patient variability. Someone that tends to experience angina faster on clean air will more likely experience angina faster when the air is mixed with CO. Similarly someone that experienced angina later when breathing clean air, will likely experience symptoms later when breathing CO as well.

2. It is not appropriate to think that we have $2n$ distinct (independent) data points (or units of information) available to us, since each data point on the same subject provides a great deal of information on the subsequent data points collected on the same subject.

4.6.4 Hypothesis testing of paired samples

In a random sample of size n paired observations, we compute the sample mean of the differences between the pairs of observations $d_i = x_{Ci} - x_{Ti}$, $i = 1, ..., n$ where "C" means control and "T" means treatment. We carry out the test like a usual single sample t test based on these differences that is,

1. State the null hypothesis
   (a) *One-sided tests:* $H_0 : \delta(= \mu_c - \mu_T) \geq 0$ or $H_0 : \delta \leq 0$
   (b) *Two-sided tests:* $H_0 : \delta(= \mu_c - \mu_T) = 0$
2. Set up the alternative hypothesis
   (a) *One-sided tests:* $H_a : \delta < 0$ $(\equiv \mu_C < \mu_T)$ or $H_a : \delta > 0$ $(\equiv \mu_C > \mu_T)$
   (b) *Two-sided tests:* $H_a : \delta \neq 0 (\equiv \mu_C \neq \mu_T)$
3. Choose the α level (the significance level of the test is $(1 - \alpha)\%$).
4. The test statistic is $T = \frac{\bar{d} - \delta}{\frac{s_d}{\sqrt{n}}} \sim t_{n-1}$, where $s_d = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (d_i - \bar{d})^2}$
5. Rejection rule: Reject the null hypothesis if,
   (a) *One-sided tests:* $T > t_{n-1;1-\alpha}$ or $T < -t_{n-1;1-\alpha}$
   (b) *Two-sided tests:* $|T| > t_{n-1;1-\frac{\alpha}{2}}$ (i.e., if $T > t_{n-1;1-\frac{\alpha}{2}}$ or $T < -t_{n-1;1-\frac{\alpha}{2}}$)

Example: CO study (continued)

The sample size is $n = 63$. The mean time to occurrence of angina was $\bar{x}_C = 3.35\%$ during the baseline (control) visit (when subjects were breathing clean air on both the stress and second measurement) and $\bar{x}_T = 9.63\%$faster when subjects were breathing air mixed with CO during the second ("treatment") visit. The difference between the two means is $\bar{d} = -6.63\%$ with standard deviation $s_d = 20.29\%$.

The null hypothesis is essentially asking the question "Is breathing CO associated with faster onset of angina" and is tested as follows:

1. The null hypothesis is $H_0 : \delta = \mu_C - \mu_T \geq 0$
2. The alternative hypothesis is $H_a : \delta < 0$ $(\equiv \mu_C < \mu_T)$
   That is, when breathing air mixed with CO angina occurs faster.
3. The α level is 5%
4. The test statistic is $T = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}} = -2.59\%$
5. Rejection rule: Reject H_0 if $T < -t_{62;0.95} = -1.673$. Since $T = -2.59 < -1.673 = t_{62;0.95}$ the null hypothesis is rejected.

Subjects when breathing air with CO experience angina faster than when breathing air without CO.

Computer implementation

To carry out the above test of hypothesis by STATA we use the one-sample t-test command as before, noting that our data are now comprised by differences of the paired observations and the mean under the null hypothesis is zero). The output is as follows:

```
. ttesti 63 -6.63 20.29 0
                                                        Number of obs =       63

------------------------------------------------------------------------------
Variable |      Mean    Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |     -6.63      2.5563  -2.59359   0.0118        1.52003    11.73997
------------------------------------------------------------------------------
Degrees of freedom: 62
                            Ho: mean(x) = 0
Ha: mean < 0           Ha: mean ~= 0              Ha: mean > 0
        t =  -2.5936             t =  -2.5936             t =  -2.5936
    P < t =  0.0059        P > |t| =   0.0118         P > t =   0.9941
```

Since `P<t=0.0059` is less than 0.05, we reject the null hypothesis. Subjects experience angina faster (by about 6.63%) when breathing air mixed with CO than when breathing clean air.

Chapter 5

Estimation

Hypothesis testing is one large part of what we call *statistical inference*, where by using a sample we *infer* (make statements) about the population that the sample came from. Another major part of statistical inference (and closely related to hypothesis testing) is estimation.

Estimation may be regarded as the opposite of hypothesis testing, in that we make a "guess" of the value (or range of values) of the unknown quantity. This is different from testing where a hypothesis about the value of this quantity must be made (what in hypothesis testing was the null hypothesis) and until shown otherwise, this hypothesized value is considered known. Nevertheless, estimation is closely related to testing, both conceptually (after all we still try to "guess" the true value of the unknown quantity) as well as in terms of mathematical implementation.

5.1 Confidence Intervals

There are two methods of estimating various population (patently unknown) quantities:

- Point estimation where a single guess about the value of the estimated quantity is proposed
- Confidence intervals where a whole range of values is considered

In what follows, we will concentrate on confidence intervals of the unknown population mean μ. Just as in hypothesis testing we will be concerned with two types of confidence intervals:

- One-sided confidence intervals
- Two-sided confidence intervals

We will also consider the case where the population standard deviation σ is known or unknown.

5.2 Estimation for the population mean (σ known)

Recall the distribution of the sample mean (a.k.a. sampling distribution of the mean). We recall that this distribution (especially for large sample sizes) is normal with mean μ and standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ where n is the size of the collected sample.

Then, the statistic

$$\boxed{Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}$$

is distributed according to the standard normal distribution. Because $P\left(-1.96 \leq Z \leq 1.96\right) = 0.95$ (i.e., Z is found between -1.96 and 1.96 about 95% of the time). Then,

$$P\left(-1.96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1.96\right) = 0.95$$

Working through the algebra (i.e., by multiplying all sides by $\sigma/\sqrt{n}$, then subtracting $\bar{X}$ and finally multiplying by -1.0, reversing the direction of the inequality) we get,

$$P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

This means that even though we do not know the exact value of μ, we expect it to be between $\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}$ and $\bar{x} + 1.96\frac{\sigma}{\sqrt{n}}$ 95% of the time. In this case, $\bar{x}$ is the point estimate of μ, while the interval $(\bar{X} - 1.96\sigma/\sqrt{n}, \bar{X} + 1.96\sigma/\sqrt{n})$ is the 95% confidence interval for μ.

5.2.1 Characteristics of confidence intervals

In the previous simple example we saw how a 95% two-sided confidence interval is constructed. If 95% is not an acceptably high confidence, we may elect to construct a 99% confidence interval. Similarly to the last case, this interval will be of the form

$$(\bar{X} - z_{0.995}\sigma/\sqrt{n}, \bar{X}z_{0.995}\sigma/\sqrt{n})$$

where $z_{0.995} = 2.58$, and consequently the 99% two-sided confidence interval of the population mean is

$$(\bar{X} - 2.58\sigma/\sqrt{n}, \bar{X} + 2.58\sigma/\sqrt{n})$$

All else being equal therefore, higher confidence (say 99% versus 95%) gets translated to a wider confidence interval. This is intuitive, since the more certain we want to be that the interval covers the unknown population mean, the more values (i.e., wider interval) we must allow this unknown quantity to take. In general, all things being equal:

- Larger variability (larger standard deviation) is associated with wider confidence intervals
- Larger sample size n results in narrower confidence intervals
- Higher confidence results in wider confidence intervals

In estimation we obviously want the narrowest confidence intervals for the highest confidence (i.e., wide confidence intervals are to be avoided).

5.2.2 Distribution of cholesterol levels

For all males in the United States who are hypertensive (have high systolic blood pressure) and smoke the distribution of cholesterol levels has an unknown mean μ and standard deviation $\sigma = 46$mg/100ml. If we draw a sample of size $n = 12$ subjects from this group of hypertensive smokers and compute their (sample) mean cholesterol level $\bar{x}_{12} = 217$mg/100ml, the 95% confidence interval based on information from this sample is

$$\left(217 - 1.96\frac{46}{\sqrt{12}}, 217 + 1.96\frac{46}{\sqrt{12}}\right) = (191,\ 243)$$

In other words we are 95% confident that the interval (191, 243) covers the unknown mean of the population of hypertensive smokers. Note that approximately 5% of the time the confidence interval that we compute will not cover the unknown population mean.

5.2.3 One-sided confidence intervals

Just as in the case of one-sided hypothesis testing, there are occasions where we are only interested in an upper or lower limit of the range of values that we will consider for the estimated quantity. In those cases we construct a *one-sided confidence interval*. In the case where only an upper limit is sought, we consider only the upper tail of the normal distribution. Conversely, when a lower limit is considered, we are concentrating in the lower tail of the normal distribution.

For example, an *upper* one-sided 95% confidence interval (when the population standard deviation is known is constructed as

$$\left(-\infty, \bar{X} + 1.645\frac{\sigma}{\sqrt{n}}\right)$$

while a *lower* one-sided 95% confidence interval is constructed as

$$\left(\bar{X} - 1.645\frac{\sigma}{\sqrt{n}}, +\infty\right)$$

Just as in the case of one-sided hypothesis testing, the advantage of using one-sided confidence intervals is obvious. Since we have to use z_α instead of $z_{\alpha/2}$, we need to consider lower cutoff point in the direction of interest. For example, if only high values are of interest, in the case of a 95% one-sided confidence interval we need to go only 1.645 standard deviations above the sample mean, instead of 1.96 standard deviations, as would be the case of the two-sided confidence interval.

5.2.4 Anemia and lead exposure

Suppose we select 74 children that have been exposed to high levels of lead, and we calculate their mean hemoglobin levels as $\bar{x}_{74} = 10.6$g/ml. Since there maybe some concern that exposure to lead is associated with lower levels of hemoglobin, we will probably be interested only in an upper limit of this value in the group of lead-exposed children. Based on this sample, and knowledge of the population standard deviation $\sigma = 0.85$g/ml, the 95% upper one-sided confidence interval is $10.6 + 1.645\frac{0.85}{74} \approx 10.8$. The unknown population mean hemoglobin level among lead-exposed children is at most 10.8g/ml.

5.3 Confidence intervals when σ is unknown

In most cases knowledge of the true variability of the measurement will not be available. In these cases, we proceed as before, substituting $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ but now basing our inference on the t distribution with $n-1$ degrees of freedom (where n is again the size of the sample).

1. Two-sided confidence intervals

$$\boxed{\left(\bar{X} - t_{n-1;1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + t_{n-1;1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right)}$$

2. One-sided confidence intervals

   - Upper one-sided confidence intervals

$$\boxed{\left(-\infty, \bar{X} + t_{n-1;1-\alpha}\frac{\sigma}{\sqrt{n}}\right)}$$

   - Lower one-sided confidence intervals

$$\boxed{\left(\bar{X} - t_{n-1;1-\alpha}\frac{\sigma}{\sqrt{n}}, +\infty\right)}$$

5.3.1 Antacids and plasma aluminumn level

In estimating the plasma aluminum level among infants that have taken antacids containing aluminum, a random sample of $n = 10$ infants was collected. The sample mean plasma aluminum level in this sample is $\bar{x}_{10} = 37.2\mu$g/l, while the sample standard deviation is $s = 7.13\mu$g/l. Since the mean and standard deviation of the plasma aluminum level in the population is unknown, a 95% two-sided confidence interval is based on the t distribution with $n - 1 = 9$ degrees of freedom as follows:

$$\left(\bar{X} - t_{n-1;1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + t_{n-1;1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = \left(37.2 - 2.262\frac{7.13}{\sqrt{10}}, 37.2 + 2.262\frac{7.13}{\sqrt{10}}\right) = (32.1, 42.3)$$

Compare the previous interval to the 95% confidence interval based on the normal distribution derived by pretending that the estimate of the standard deviation $s = 7.13\mu$g/l is the true population standard deviation. This interval is $(32.8, 41.6)$ and has length $8.8(= 41.6 - 32.8)\mu$g/l whereas the one based on the t distribution has length $10.2(= 42.3 - 32.1)\mu$g/l. This loss of accuracy (widening of the confidence interval) is the "penalty" we pay for the lack of knowledge of the true population standard deviation.

5.3.2 Computer implementation

Confidence intervals can be computed in STATA by using the command `ci`, or its immediate equivalent `cii`. The syntax is as follows:

ci [varlist] [weight] [if exp] [in range] [,level(#) by(varlist2) total]

cii #obs #mean #sd [, level(#)]

where is `ci` used when we have access to the complete data set, while `cii` is used when only the sample size (`#obs`), sample mean (`#mean`) and standard deviation (`#sd`) are known. In all cases, we can manipulate the alpha level of the confidence interval by using the option `level(#)`. For example, `level(95)` would calculate a 95% confidence interval (default), while `level(90)` would calculate a 90% confidence interval.

Example: Antacids and aluminum level (continued):

In the example above, a 95% (two-sided) confidence interval is as follows:

```
. cii 10 37.2 7.13

Variable |     Obs        Mean    Std. Err.       [95% Conf. Interval]
---------+--------------------------------------------------------------
         |      10        37.2    2.254704        32.09951    42.30049
```

This agrees with our hand calculations (i.e., that the 95% C.I. is $(32.1, 42.3)$).

Caution! STATA only produces two-sided confidence intervals. If you want to obtain one-sided confidence intervals for the aluminum example, you have to use the `level(#)` option as follows:

```
. cii 10 37.2 7.13, level(90)

Variable |     Obs        Mean    Std. Err.       [90% Conf. Interval]
---------+--------------------------------------------------------------
         |      10        37.2    2.254704        33.06687    41.33313
```

Thus, an upper 95% confidence interval would be $(-\infty, 41.3)$, while a lower 95% confidence interval would be $(33.1, +\infty)$.

5.4 Confidence intervals of a difference of two means

Recall that in the case of comparisons based of two groups we have:

	Group 1	Group 2
Population		
mean	μ_1	μ_2
std. deviation	σ_1	σ_2
Sample		
mean	$\bar{X}_1$	$\bar{X}_2$
std. deviation	s_1	s_2
sample size	n_1	n_2

In similar fashion as with two-sample tests of hypothesis, inference is based on the statistic,

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where s_p is the pooled estimate of the population standard deviation. In this case, t is distributed according to a t distribution with $n_1 + n_2 - 2$ degrees of freedom. Notice that $s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$, that is, the standard error of the difference of two means is the square root of the sum of the variances of each mean (recall that we have two sources of variability when we deal with two groups). This of course holds *only* when the two groups are independent!

The two-sided confidence interval for the difference of two means is

$$\left((\bar{x}_1 - \bar{x}_2) - t_{n_1+n_2-2;1-\alpha/2}s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}},\ (\bar{x}_1 - \bar{x}_2) + t_{n_1+n_2-2;1-\alpha/2}s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)$$

5.4.1 Serum iron levels and cystic fibrosis

In this example $\bar{x}_1 = 18\mu$mol/l is the sample mean iron level in a sample of $n_1 = 9$ healthy children with standard deviation $s_1 = 5.9\mu$mol/l, and the iron levels among $n_2 = 13$ children with cystic fibrosis where $\bar{x}_2 = 11.9\mu$mol/l and $s_2 = 6.3\mu$mol/l respectively.

The pooled estimate of the common standard deviation is $s_p = \sqrt{\frac{(9-1)(5.9)^2+(13-1)(6.3)^2}{9+13-2}} = 6.14\mu$mol/l. A two-sided confidence interval of the true difference in iron levels between healthy children and children with cystic fibrosis is then

$$\left((\bar{x}_1 - \bar{x}_2) - t_{n_1+n_2-2;1-\alpha/2}s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad , \quad (\bar{x}_1 - \bar{x}_2) + t_{n_1+n_2-2;1-\alpha/2}s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)$$

$$\Downarrow$$

$$\left((18.9 - 11.9) - (2.086)(6.14)\sqrt{\frac{1}{9} + \frac{1}{13}} \quad , \quad (18.9 - 11.9) - (2.086)(6.14)\sqrt{\frac{1}{9} + \frac{1}{13}}\right)$$

$$\Downarrow$$

$$(1.4 \quad , \quad 12.6)$$

How does this compare to the result of the hypothesis test (which as you may recall rejected the null hypothesis at the 5% level)?

5.5 Performing hypothesis testing using confidence intervals

To perform tests of hypotheses using confidence intervals we proceed as follows:

STEP 1. Formulate the null and alternative hypotheses as before

STEP 2. Choose the alpha level

STEP 3. Construct the $(1-\alpha)\%$ confidence interval as described above. Use a one-sided or two-sided confidence intervals depending on the test you want to carry out.

STEP 4. Rejection rule. Reject the null hypothesis (as described in STEP 1) if the confidence interval does not include the hypothesized value (in the null hypothesis).

In the example of the iron levels of children with cystic fibrosis versus healthy children, we carry out the test of no difference in the iron levels as follows:

STEP 1. H_o: $\mu_1 = \mu_2$ (or equivalently, $\mu_1 - \mu_2 = 0$)
H_a: $\mu_1 \neq \mu_2$ (or equivalently, $\mu_1 - \mu_2 \neq 0$)

STEP 2. The alpha level is 5%

STEP 3. The two-sided 95% confidence interval of the difference of the two means is $(1.4, 12.6)$

STEP 4. Since the hypothesized value of zero difference (equality of the two means) is not covered by this interval we reject the null hypothesis, in favor of the alternative. That is, children with cystic fibrosis do not have the same iron levels as healthy children (in fact they have lower levels).

5.5.1 Computer implementation

Confidence intervals for the difference between two means can be computed in STATA by using the command `ci` with the `by(varlist2)` option, if the data are given in two columns `(varlist1, varlist2)` where `varlist1` is the variable of interest and `varlist2` is a grouping variable.

```
ci [varlist] [weight] [if exp] [in range] [,level(#) by(varlist2) total]
```

If the data are not given as above, or access to the raw data is not possible, the `ttest` and `ttesti` commands must be used instead. The previous caution for calculating one-sided confidence intervals carries over to this case, that is, the option `level(#)` must be used. That is, to produce a 95% one-sided confidence interval we use the `ttest` or `ttesti` command with the option `level(90)`.

Example: Serum iron levels and cystic fibrosis (continued)

The computer output is as follows:

```
. ttesti 9 18.9 5.9 13 11.9 6.3

Two-sample t test with equal variances
------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |       9        18.9    1.966667         5.9    14.36486    23.43514
       y |      13        11.9    1.747306         6.3    8.092948    15.70705
---------+--------------------------------------------------------------------
combined |      22    14.76364    1.482474     6.95342    11.68066    17.84661
---------+--------------------------------------------------------------------
    diff |                 7    2.663838                1.443331    12.55667
------------------------------------------------------------------------------
Degrees of freedom: 20
                     Ho: mean(x) - mean(y) = diff = 0

     Ha: diff < 0                 Ha: diff ~= 0                 Ha: diff > 0
       t =   2.6278                  t =   2.6278                  t =   2.6278
   P < t =   0.9919          P > |t| =   0.0161          P > t =   0.0081
```

Thus, a two-sided 95% confidence interval for the difference in serum iron levels between healthy children and children suffering from cystic fibrosis is (1.4, 12.6) as we saw before.

5.5.2 One-sided tests

If a 95% one-sided confidence interval were required (corresponding to a one-sided hypothesis test), the computer solution would be as follows:

```
. ttesti 9 18.9 5.9 13 11.9 6.3, level(90)

Two-sample t test with equal variances
------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [90% Conf. Interval]
---------+--------------------------------------------------------------------
       x |       9        18.9    1.966667         5.9    15.24289    22.55711
       y |      13        11.9    1.747306         6.3    8.785799     15.0142
---------+--------------------------------------------------------------------
combined |      22    14.76364    1.482474     6.95342    12.21268    17.31459
---------+--------------------------------------------------------------------
    diff |                 7    2.663838                 2.40563    11.59437
------------------------------------------------------------------------------
Degrees of freedom: 20
                     Ho: mean(x) - mean(y) = diff = 0

     Ha: diff < 0                 Ha: diff ~= 0                 Ha: diff > 0
       t =   2.6278                  t =   2.6278                  t =   2.6278
   P < t =   0.9919          P > |t| =   0.0161          P > t =   0.0081
```

Thus the 95% lower one-sided confidence interval for the difference of the mean serum iron level is then $(2.4, +\infty)$

Chapter 6

Counts and Proportions

So far we were concerned with random variables that represented continuous measurements (weight, cholesterol level, etc.).

Now consider the case where there are consecutive experiments and where

1. There are two possible (mutually exclusive) outcomes, usually called a "success" and a "failure".

2. Each experiment is identical to all the others, and the probability of a "success" is p. Thus, the probability of a "failure" is $1 - p$.

Each experiment is called a *Bernoulli trial*. Such experiments include throwing the die and observing whether or not it comes up six, investigating the survival of a cancer patient, etc. For example, consider smoking status, and define $X = 1$ if the person is a smoker, and $X = 0$ if he or she is a non-smoker. If "success" is the event that a randomly selected individual is a smoker and from previous research it is known that about 29

$$P(X = 1) = p$$

$$P(X = 0) = 1 - p$$

This is an example of a Bernoulli trial (we select just one individual at random, each selection is carried out independently, and, each time the probability of that individual to be a "success" is constant)

6.1 The binomial distribution

If we selected two adults, and looked at their smoking status, we would have as possible outcomes,

- Neither is a smoker
- Only one is a smoker
- Both are smokers

If we define X as the number of smokers between these two individuals, then

- $X = 0$: Neither is a smoker
- $X = 1$: Only one is a smoker
- $X = 2$: Both are smokers

The probability distribution of X (recall that $S = \{0, 1, 2\}$ in this case) is

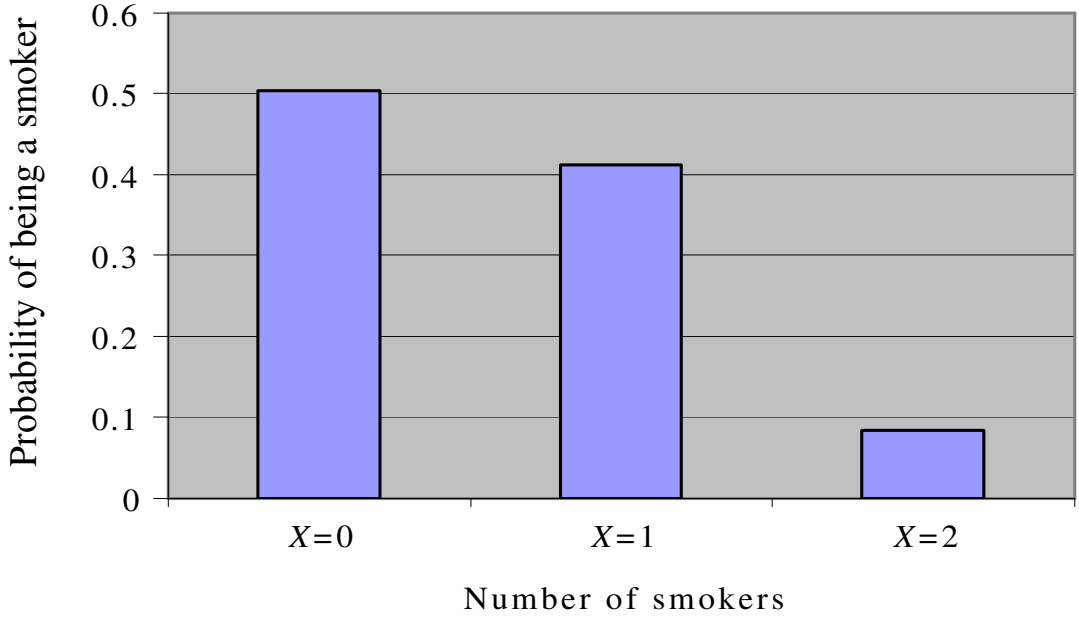
$$\begin{aligned} P(X = 0) &= (1 - p)^2 \\ &= (0.71)^2 = 0.5041 \end{aligned}$$

$$\begin{aligned} P(X = 1) &= P(\text{1st individual is a smoker } \textbf{OR} \text{ 2nd individual is a smoker}) \\ &= p(1 - p) + (1 - p)p \\ &= 2p(1 - p) = 0.4118 \end{aligned}$$

$$\begin{aligned} P(X = 2) &= p^2 \\ &= (0.29)^2 = 0.0841 \end{aligned}$$

Notice that $P(X = 0)+P(X = 1)+P(X = 2) = 0.5041+0.4118+0.0841 = 1.000$. The distribution of X is shown in the following figure. The bar chart above is a plot of the probability distribution

Figure 6.1: Distribution of the number of smokers among two individuals



of X, going over all the possible numbers that X can attain (in the previous example those were 0, 1, and $2 = n$). There is a special distribution that closely models the behavior of variables that "count" successes among n repeated Bernoulli experiments. This is called the *binomial distribution*.

In our treatment of the binomial distribution, we only need to know two basic parameters:

- The probability of "success" p

- The number of Bernoulli experiments n

One way of looking at p, the proportion of time that an experiment comes out as a "success" out of n repeated (Bernoulli) trials is as the mean of a sample of measurements that are zeros or ones. That is,

$$p = \frac{1}{n}\sum_{i=1}^{n} X_i,\ i = 1, ..., n$$

where X_i are zeros or ones.

Given the above parameters, the mean and standard deviation of X the count of successes out of n trials are: $\mu = np$ and $\sigma = \sqrt{np(1-p)}$ respectively.

6.2 Normal approximation to the binomial distribution

If n is sufficiently large, then the statistic

$$Z = \frac{x - np}{\sqrt{np(1-p}} \sim N(0, 1)$$

is approximately distributed as normal with mean 0 and standard deviation 1 (i.e., the standard normal distribution). A better approximation to the normal distribution is given by $Z = \frac{x-np+0.5}{\sqrt{np(1-p}}$ when $X < np$ and $Z = \frac{x-np-0.5}{\sqrt{np(1-p}}$ when $X > np$.

For example, suppose that we want to find the proportion of samples of size $n = 30$ in which at most six individuals smoke. With $p = 0.29$ and $n = 30$, $X = 6 < np = 8.7$. Thus, applying the continuity correction as shown above,

$$\begin{aligned} P(X \leq 6) &= P\left(Z \leq \frac{x - np + 0.5}{\sqrt{np(1-p}}\right) \\ &= P\left(Z \leq \frac{6 - (30)(0.29) + 0.5}{\sqrt{(30)(0.29)(0.71}}\right) \\ &= P(Z \leq -0.89) = 0.187 \end{aligned}$$

The exact binomial probability is 0.190, which is very close to the approximate value given above.

6.3 Sample distribution of a proportion

Our thinking in terms of estimation (including confidence intervals) and hypothesis testing does not change when dealing with proportions. Since the proportion of a success in the general population will not be known, we must estimate it. In general such an estimate will be derived by calculating the proportion of successes in a sample of n experiments by using the sample proportion $\hat{p} = \frac{x}{n}$ where x is the number of successes in the sample of size n. The sampling distribution of a proportion has mean $\mu = p$ and standard deviation $\sigma_{\hat{p}} = \sqrt{p(1-p)} = \frac{\sigma}{n}$ where $\sigma = \sqrt{np(1-p)}$ as above. We would expect that $\hat{p}$ might have similar properties as $\bar{X}$ and in fact, the statistic

$$Z = \frac{\hat{p} - p}{\sqrt{p(1-p)}} \sim N(0, 1)$$

is distributed according to the standard normal distribution. This approximation is particularly good when $np > 5$ and $n(1-p) > 5$. This means that the size of the sample is relative to the rarity of the event "success". The less rare such an event is (i.e., the higher p is) the smaller sample size would be required for the normal approximation to hold.

Example: Five-year survival among lung-cancer patients

Consider the five-year survival among patients under 40 years-old who have been diagnosed with lung cancer. The mean proportion of individuals surviving is $p = 0.10$ (implying that the standard deviation of the 5-year survival is $\sqrt{0.10(1-0.10)} = 0.30$).

If we select repeated samples of size $n = 50$ patients diagnosed with lung cancer, what fraction of the samples will have 20% or more survivors? That is, "what percent of the time $10(= 50(0.20))$ or more patients will be alive after 5 years"?

Since $np = (50)(0.1) = 5 \geq 5$ and $n(1-p) = (50)(0.9) = 45 > 5$ the normal approximation should be adequate. Then,

$$\begin{aligned} P(\hat{p} \geq 0.20) &= P\left(\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \geq \frac{0.20-p}{\sqrt{\frac{p(1-p)}{n}}}\right) \\ &= P\left(Z \geq \frac{(0.20)-(0.10)}{\sqrt{\frac{0.10(1-0.10)}{50}}}\right) = P(Z \geq 2.36) = 0.009 \end{aligned}$$

Only 0.9% of the time will the proportion of lung cancer patients surviving past five years be 20% or more.

6.3.1 Hypothesis testing involving proportions

In the previous example, we did not know the true proportion of 5-year survivors among individuals under 40 years of age that have been diagnosed with lung cancer. If it is known from previous studies that the five-year survival rate of lung cancer patients that are older than 40 years old is 8.2%, we might want to test whether the five-year survival among the younger lung-cancer patients is the same as that of the older ones. The steps for carrying out a test of hypotheses involving proportions proceeds in a manner identical to what we have seen previously.

The steps are as follows:

1. State the null hypothesis:
   - Two-sided hypothesis: $H_o : p = p_0$
   - One-sided hypothesis: $H_o : p \geq p_0$ or $H_o : p \leq p_0$
2. State the alternative hypothesis H_a
   - Two-sided hypothesis: $H_a : p \neq p_0$
   - One-sided hypothesis: $H_a : p < p_0$ or $H_a : p > p_0$
3. The alpha level of this test is α% (usually 5% or 1%)

4. The test statistic is $Z = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}$

5. We will reject the null hypothesis if:

   - Two-sided hypothesis: $P(|Z| > z_{1-\frac{\alpha}{2}})$, i.e., if $P(Z < z_{\frac{\alpha}{2}})$ or $P(Z > z_{1-\frac{\alpha}{2}})$
   - One-sided hypothesis: $P(Z < z_\alpha)$ or $P(Z > z_\alpha)$ respectively

   Equivalently, the rejection rule can be expressed as follows: Reject the null hypothesis if:

   - Two-sided hypothesis: $P(|Z| > z) < \alpha$, i.e., if $P(Z < -z) + P(Z > z) < \alpha$
   - One-sided hypothesis: $P(Z < -z) < \alpha$ or $P(Z > z) < \alpha$ respectively

   where z is the observed value of the statistic $\frac{\hat{p}-p_0}{\sqrt{\frac{p(1-p)}{n}}}$

From a sample of $n = 52$ patients under the age of 40 that have been diagnosed with lung cancer the proportion surviving after five years is $\hat{p} = 0.115$. Is this equal or not to the known 5-year survival of older patients? The test of hypothesis is constructed as follows:

1. $H_o : p = 0.082$

2. $H_a : p \neq 0.082$

3. The alpha level of this test is 0.01 (according to the textbook)

4. The test statistic is $Z = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}$

5. We will reject the null hypothesis if $P(|Z| > z) < \alpha$.

Since $P(|Z| > 0.87) = P(Z > 0.87) + P(Z < -0.87) = 0.192 + 0.192 = 0.384 > 0.05$, we do not reject the null hypothesis. That is, there is no evidence to indicate that the five-year survival of lung cancer patients who are younger than 40 years of age is different than that of the older patients.

6.3.2 Computer implementation

Use the command `bitest` or its immediate equivalent `bitesti`, or the command `prtest` or its immediate equivalent `prtesti`. The former performs an exact binomial test, while the latter performs the test based on the normal approximation. The syntax of both commands is as follows:

```
bitest varname = #p [if exp] [in range] [, detail]

bitesti #N #succ #p [, detail]
```

and

```
prtest varname = # [if exp] [in range] [, level(#)]

prtest varname = varname [if exp] [in range] [, level(#)]

prtesti #obs #p1 #p2 [, level(#)

prtesti #obs1 #p1 #obs2 #p2 [, level(#)]
```

Carrying out an exact binomial test in the problem above we have

```
. bitesti 52 6 0.082

N   Observed k   Expected k   Assumed p   Observed p
------------------------------------------------------------
    52          6        4.264       0.08200      0.11538

    Pr(k >= 6)           = 0.251946  (one-sided test)
    Pr(k <= 6)           = 0.868945  (one-sided test)
    Pr(k <= 1 or k >= 6) = 0.317935  (two-sided test)
```

We see that the p value associated with the two-sided test is 0.318 which is close to that calculated above. Using the normal approximation we have:

```
. prtesti 52 6 0.082,count

One-sample test of proportion                      x: Number of obs =       52

------------------------------------------------------------------------------
Variable |      Mean    Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |  .1153846    .0443047   2.60434   0.0092       .0285491    .2022202
------------------------------------------------------------------------------

                         Ho: proportion(x) = .082

      Ha: x < .082            Ha: x ~= .082            Ha: x > .082
       z =  0.877               z =  0.877               z =  0.877
   P < z =  0.8099        P > |z| =  0.3802         P > z =  0.1901
```

Note! The above output was produced with STATA version 7.0. To obtain the same output with STATA 6.0 or earlier you must omit the option `count` as follows:

```
. prtesti 52 6 0.082

One-sample test of proportion              x: Number of obs =       52

------------------------------------------------------------------------------
Variable |   Mean    Std. Err.        z  P>|z|   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |.1153846    .0443047  2.60434 0.0092   .0285491    .2022202
------------------------------------------------------------------------------

                     Ho: proportion(x) = .082

 Ha: x < .082            Ha: x ~= .082            Ha: x > .082
  z =  0.877               z =  0.877              z =  0.877
P < z = 0.8099         P > |z| = 0.3802         P > z = 0.1901
```

This closely matches our calculations. We do not reject the null hypothesis, since the p value associated with the two-sided test is $0.380 > \alpha$. Note that any differences with the hand calculations are due to round-off error.

6.4 Estimation

Similar to the testing of hypothesis involving proportions, we can construct confidence intervals where we can be fairly confident (at a pre-specified level) that the unknown true proportion lies. Again these intervals will be based on the statistic

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

where $\hat{p} = \frac{x}{n}$ and $\sqrt{\frac{p(1-p)}{n}}$ are the estimates of the proportion and its associated standard deviation respectively.

Two sided confidence intervals:

$$\boxed{\left(\hat{p} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}, \hat{p} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right)}$$

One-sided confidence intervals:

- Upper, one-sided interval:

$$\boxed{\left(0, \hat{p} + z_{1-\alpha}\sqrt{\frac{p(1-p)}{n}}\right)}$$

- Lower, one-sided interval :

$$\boxed{\left(\hat{p} - z_{1-\alpha}\sqrt{\frac{p(1-p)}{n}}, 1\right)}$$

In the previous example, if 6 out of 52 lung cancer patients under 40 years of age were alive after five years, and using the normal approximation (which is justified since $np = 52(0.115) = 5.98 > n$, and $52(1 - 0.115) = 46.02 > n$), an approximate 95% confidence interval for the true proportion p is given by

$$\left(\hat{p} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} \quad , \quad \hat{p} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right)$$

$$\Downarrow$$

$$\left(0.115 - 1.96\sqrt{\frac{0.115(1-0.115)}{52}} \quad , \quad 0.115 + 1.96\sqrt{\frac{0.115(1-0.115)}{52}}\right)$$

$$\Downarrow$$

$$(0.028 \quad , \quad 0.202)$$

In other words, we are 95% confident that the true five-year survival of lung-cancer patients under 40 years of age is between 2.8% and 20.2%. Note that this interval contains 8.2% (the five-year survival rate among lung cancer patients that are older than 40 years of age). Thus it is equivalent to the hypothesis testing test which did not reject the hypothesis that the five-year survival between lung cancer patients that are older than 40 years old versus younger subjects.

Computer implementation

To construct one- and two-sided confidence intermvals we use the `ci` command and its immediate equivalent `cii`. Their syntax is as follows:

```
ci varlist [weight] [if exp] [in range] [, level(#) binomial poisson exposure(varname)
by(varlist2) total ]
cii #obs #mean #sd [, level(#) ] (normal)
cii #obs #succ [, level(#) ] (binomial)
cii #exposure #events , poisson [ level(#) ] (Poisson)
```

In the previous example, using exact binomial confidence intervals we have

```
. cii 52 6
                                                    -- Binomial Exact --
Variable |    Obs       Mean    Std. Err.       [95% Conf. Interval]
---------+--------------------------------------------------------------
         |     52   .1153846     .0443047       .0435439    .2344114
```

which is close to our calculations that used the normal approximation. The normal approximation confidence interval can be constructed by the following command (note that $0.31902 = \sqrt{0.115(1-0.115)}$).

```
. cii 52 0.115 0.31902

Variable |     Obs        Mean    Std. Err.       [95% Conf. Interval]
---------+--------------------------------------------------------------
         |      52        .115    .0442401        .0261843    .2038157
```

6.4.1 Comparison between two proportions

We proceed when comparing two proportions just like a two-mean comparison. The sample proportion in the first and second groups are $\hat{p}_1 = \frac{x_1}{n_1}$ and $\hat{p}_2 = \frac{x_2}{n_2}$. Under assumptions of equality of the two population proportions, we may want to derive a pooled estimate of the sample proportion, using data from both groups, $\hat{p} = \frac{x_1+x_2}{n_1+n_2}$, that is divide the total number of successes in the two groups $(x_1 + x_2)$, by the total sample size $(n_1 + n_2)$. Using this pooled estimate, we can derive a pooled estimate of the standard deviation of the unknown proportion (assumed equal between the two groups as $s_p = \sqrt{\hat{p}(1+\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$. The hypothesis testing of comparisons between two proportions is based on the statistic

$$\boxed{z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}}$$

which is distribution according to a standard normal distribution. The test is carried out as follows:

1. Two-sided tests: $H_o : p_1 = p_2$ (or $p_1 - p_2 = 0$)
   - $H_a : p_1 \leq p_2$ (or $p_1 - p_2 \leq 0$)
   - $H_a : p_1 \geq p_2$ (or $p_1 - p_2 \geq 0$)
2. Two sided tests: $H_a : p_1 \neq p_2$ (or $p_1 - p_2 \neq 0$)
3. One sided tests:
   - $H_a : p_1 > p_2$ (or $p_1 - p_2 > 0$)
   - $H_a : p_1 < p_2$ (or $p_1 - p_2 < 0$)
4. Determine the alpha level of the test
5. Rejection rule:
   - Two-sided tests: Reject H_o if $P(|Z| > z) < \alpha$ (i.e., $P(Z > z) + P(Z < -z) < \alpha$).
   - One-sided tests: Reject H_o if $P(Z > z) < \alpha$ or reject H_o if $P(Z < -z) < \alpha$ respectively.

Example: Mortality of pediatric victims

In a study investigating morbidity and mortality among pediatric victims of motor vehicles accidents, information regarding the effectiveness of seat belts was collected. Two random samples were selected, one of size $n_1 = 123$ from a population of children that were wearing seat belts at the time of the accident, and another of size $n_2 = 290$ from a group of children that were not wearing seat belts at the time of the accident. In the first case, $x_1 = 3$ children died, while in the second $x_2 = 13$ died. Consequently, $\hat{p}_1 = 0.024$ and $\hat{p}_2 = 0.045$ and the task is to compare the two.

Carrying out the test of hypothesis as proposed earlier,

1. $H_o : p_1 = p_2$ (or $p_1 - p_2 = 0$)
2. $H_a : p_1 \neq p_2$ (or $p_1 - p_2 \neq 0$)
3. The alpha level of the test is 5%
4. The test statistic is

$$\begin{aligned} z &= \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \\ &= \frac{(0.024 - 0.045)}{\sqrt{0.039(1-0.039)\left(\frac{1}{123} + \frac{1}{290}\right)}} = -0.98 \end{aligned}$$

5. Rejection rule: Reject H_o if $P(|Z| > z) < \alpha$ (i.e., $P(Z > z) + P(Z < -z) < \alpha$).

This is, $P(Z > 0.98) + P(Z < -0.98) = 0.325 > \alpha$. Thus, there is no evidence that children not wearing seat belts are safer (die at different rates) than children wearing seat belts.

6.4.2 Confidence intervals of the difference between two proportions

Confidence intervals of the difference of two proportions are also based on the statistic

$$\boxed{z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}}$$

Note! Since we no longer need to assume that the two proportions are equal, the estimate of the standard deviation in the denominator is not a pooled estimate, but rather simply the sum of the std. deviations in each group. That is, the standard deviation estimate is

$$s_p = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

This an important deviation from hypothesis testing and may lead to inconsistency between decisions reached through usual hypothesis testing versus hypothesis testing performed using confidence intervals.

1. Two-sided confidence intervals:

$$\boxed{\left((\hat{p}_1-\hat{p}_2)-z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},(\hat{p}_1-\hat{p}_2)+z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right)}$$

2. One-sided confidence intervals:

   - Upper one-sided confidence interval:

$$\boxed{\left(-1,(\hat{p}_1-\hat{p}_2)+z_{\alpha}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right)}$$

   - Lower one-sided confidence interval:

$$\boxed{\left((\hat{p}_1-\hat{p}_2)-z_{\alpha}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},1\right)}$$

A two-sided 95% confidence interval for the true difference death rates among children wearing seat belts versus those that did not is given by

$$\begin{aligned}
&\left((\hat{p}_1-\hat{p}_2)-z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\quad,\quad(\hat{p}_1-\hat{p}_2)+z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right)\\
&\qquad\qquad\Downarrow\\
&((0.024-0.045)-1.96(0.018)\quad,\quad(0.024-0.045)+1.96(0.018))\\
&\qquad\qquad\Downarrow\\
&(-0.057\quad,\quad 0.015)
\end{aligned}$$

That is, the true difference between the two groups will be between 5.7% in favor of children wearing seat belts, to 1.5% in favor of children not wearing seat belts. In this regard, since the zero (hypothesized under the null hypothesis) difference is included in the confidence interval we do not reject the null hypothesis. There is no evidence to suggest a benefit of seat belts.

6.4.3 Computer implementation

To carry out the comparisons above we use the command `prtesti`. The syntax is as follows:

```
. prtesti 123 3 290 13, count

Two-sample test of proportion                     x: Number of obs =      123
                                                  y: Number of obs =      290

------------------------------------------------------------------------------
Variable |      Mean    Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |  .0243902    .0139089   1.75357   0.0795      -.0028707    .0516512
       y |  .0448276    .0121511   3.68919   0.0002       .0210119    .0686432
---------+--------------------------------------------------------------------
    diff | -.0204373    .0184691                         -.0566361    .0157614
         |  under Ho:   .0207648  -.984228   0.3250
------------------------------------------------------------------------------

               Ho: proportion(x) - proportion(y) = diff = 0

       Ha: diff < 0            Ha: diff ~= 0            Ha: diff > 0
        z = -0.984               z = -0.984               z = -0.984
      P < z = 0.1625          P > |z| = 0.3250         P > z = 0.8375
```

The option `count` specifies that integer counts, not proportions are being used. This is a STATA version 7.0 option. This option is not available in version 6.0. There, the same output would be produced as follows:

```
. prtesti 123 3 290 13

Two-sample test of proportion                     x: Number of obs =      123
                                                  y: Number of obs =      290

------------------------------------------------------------------------------
Variable |      Mean    Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |  .0243902    .0139089   1.75357   0.0795      -.0028707    .0516512
       y |  .0448276    .0121511   3.68919   0.0002       .0210119    .0686432
---------+--------------------------------------------------------------------
    diff | -.0204373    .0184691                         -.0566361    .0157614
         |  under Ho:   .0207648  -.984228   0.3250
------------------------------------------------------------------------------

               Ho: proportion(x) - proportion(y) = diff = 0

       Ha: diff < 0            Ha: diff ~= 0            Ha: diff > 0
        z = -0.984               z = -0.984               z = -0.984
      P < z = 0.1625          P > |z| = 0.3250         P > z = 0.8375
```

Chapter 7

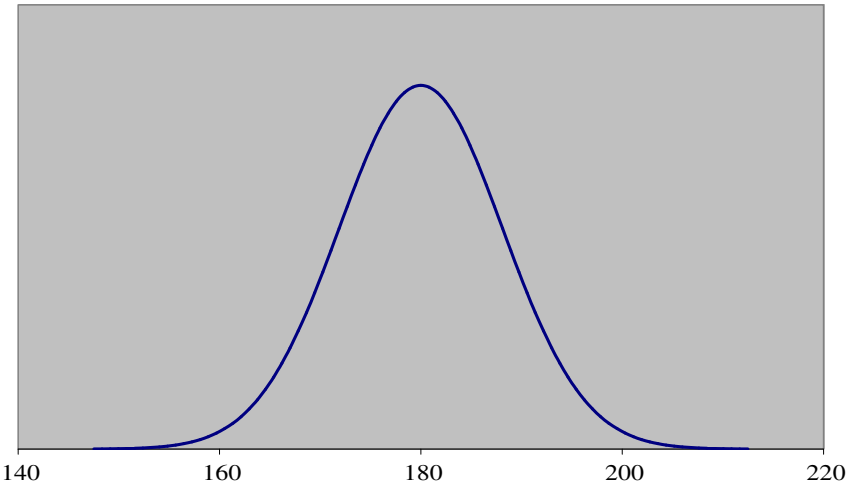
Power and sample-size calculations

Consider the following example. U.S. males ages 20-24 have mean cholesterol level $\mu = 180$ mg/ml. By comparison, we can assume that the mean cholesterol level in the overall population of males 20-74 is higher. Thus, we are carrying out the one-sided test of hypothesis:

$$H_o : \mu \leq 180 \text{ mg/100 ml}$$
$$H_a : \mu > 180 \text{ mg/100 ml}$$

If we collect a sample of $n = 25$ U.S. males ages 20-74 and measure their cholesterol level, we can use $\bar{X}_n$, the sample mean (mean cholesterol level in our sample) to make inferences about the (20-74 year-old male) population mean. The distribution of $\bar{X}_n$ is normal, with mean $\mu = 180$ mg/100 ml (by the assumption implicit in the null hypothesis), and standard deviation $\sigma = \frac{46}{\sqrt{25}}$ mg/ml (by the central limit theorem). The situation is pictured in the following figure:

Figure 7.1: Distribution of $\bar{X}_n \sim N(180, 9.2)$



7.1 Types of error

In the one-sided case above, when determining the α level of a hypothesis test, you also determine a cutoff point beyond which you will reject the null hypothesis. This point can be found from test

statistic $z = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$, and the fact that we will reject the null hypothesis if $Z \geq z_\alpha$. In the cholesterol level example, the cutoff cholesterol level corresponding to a 5% α level is found as follows:

$$\begin{aligned} Z &\geq z_\alpha \\ &\Leftrightarrow \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \geq z_\alpha \\ &\Leftrightarrow \frac{\bar{X}_n - 180}{\frac{46}{\sqrt{25}}} \geq 1.645 \\ &\Leftrightarrow \bar{X}_n \geq (1.645)\left(\frac{46}{25}\right) + 180 = 195.1 \end{aligned}$$

Thus, if the sample mean from a group of $n = 25$ males 20-74 years old is higher than 195.1 mg/dL,

Figure 7.2: Implication of setting the α level of a test

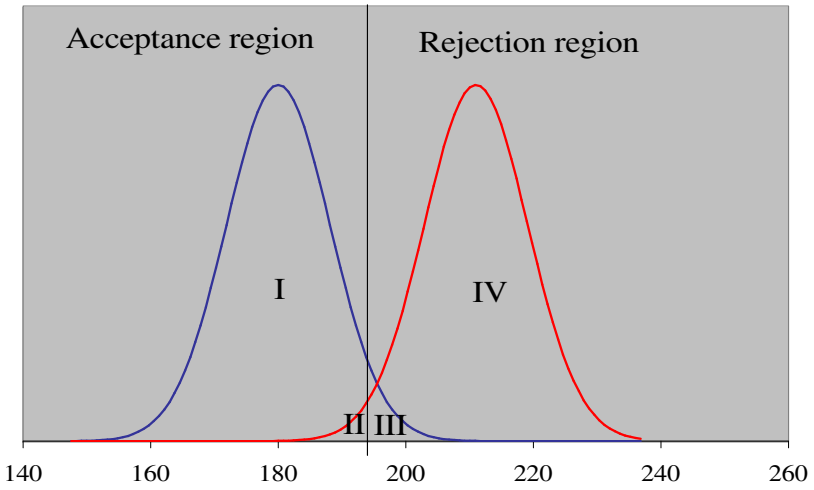
then we will reject the null hypothesis, and decide that the 20-74 population mean cholesterol level is higher than the 20-24 year old population. How often will such a sample mean be over 195.1 mg/dL even if the 20-74 year old males cholesterol level is the same as that of the 20-24 year olds?

This will happen α% of the time. That is, α% of the time we will be rejecting the null hypothesis even though it is true. This is called an *error of Type I.*

The alpha level of the test is the maximum allowed probability of type-I error

What if the "true" mean cholesterol of 20-74 year-olds is $\mu_1 = 211$ mg/dL? This situation is given in the following figure.

Figure 7.3: Normal distributions of $\bar{X}_n$ with means $\mu_o = 180$ mg/dL and $\mu_1 = 211$ mg/dL and identical std. deviations $\sigma = 9.2$mg/dL



There is a chance that even though the mean cholesterol level is truly $\mu_1 = 211$ mg/dL, that the sample mean will be to the left of the cutoff point (and in the acceptance region). In that case we would have to accept the null hypothesis (even though it would be false). That would be an *error of Type II*. The probability of a type-II error is symbolized by β.

What is this probability in this example? The probability of a type-II error is

$$\begin{aligned}
\beta &= P\left(\bar{X}_n \leq 195.1 | \mu = \mu_1 = 211\right) \\
&= P\left(\frac{\bar{X}_n - 211}{\frac{46}{\sqrt{25}}} \leq \frac{195.1 - 211}{\frac{46}{\sqrt{25}}}\right) \\
&= P\left(Z \leq -1.73\right) = 0.042
\end{aligned}$$

from the Appendix in the textbook.

There are four areas in the previous figure that are immediately identifiable:

- I. The distribution has mean $\mu = 180$ mg/dL and the null hypothesis is not rejected. In this case, the test made the correct decision.
- II. In this case, the distribution of cholesterol levels among 20-74 year-old males has a higher mean compared to that of 20-24 year-olds and the test has erroneously failed to reject the null hypothesis. This is an error of Type II.
- III. The mean cholesterol of the 20-74 year-olds is the same as that of 20-24 year-olds but the null hypothesis is rejected. This is an error of Type I.
- IV. The sample mean among 20-74 year-old individuals is truly higher than that of 20-24 year-olds and the test has correctly rejected the null hypothesis.

7.2 Power

The error associated with case II is called error of Type II. Just as we had defined the probability of a Type I error as α, and the probability of a Type II error as β, we have a special name for the probability associated with case IV, that is, the probability that the test will correctly reject the null hypothesis. This is called the *power* of the test. In other words, power is the chance that the test as defined will pick up true differences in the two populations.

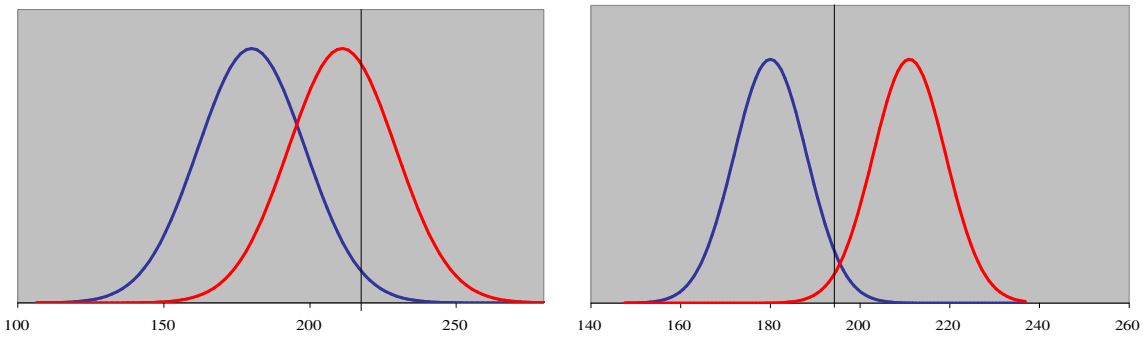
$$\boxed{\text{Power} = 1 - \beta}$$

In the example above, the power is $1 - 0.042 = 0.958$ (or 95.8%).

7.3 Sample size

In the beginning of a new experiment, we may ask ourselves how many subjects we should include. In other words, we want to determine the sample size of the study. All else being equal, the only way to increase the power of an experiment (i.e., increase its chance of detecting true differences) is by increasing the sample size. Consider the two cases: In the first case (left), the distributions

Figure 7.4: Separation of the null and alternative distributions with increasing sample size



are based on a sample size of $n = 5$ subjects versus $n = 25$ in the original situation (right).

To determine the sample size we need:

1. Null and alternative means
2. Standard deviation (or estimate of variability)
3. Alpha level of the test
4. Desired power

Items 1 and 2 can sometimes be substituted by "standardized differences" $\delta = \frac{\mu_1 - \mu_2}{\sigma}$ where σ is the assumed common standard deviation.

Example: Cholesterol example (continued):

For example, if $\alpha = 1\%$, the desired power is $1 - \beta = 95\%$, and the two means are $\mu_0 = 180$ mg/dL and $\mu_1 = 211$ mg/dL respectively, then the cutoff point is

$$\bar{x} = \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}} = 180 + (2.32)\left(\frac{46}{\sqrt{n}}\right)$$

Also, by virtue of the desired power,

$$\bar{x} = \mu_1 - z_\beta \frac{\sigma}{\sqrt{n}} = 211 - (1.645\left(\frac{46}{\sqrt{n}}\right)$$

So,

$$180 + (2.32)\left(\frac{46}{\sqrt{n}}\right) = 211 - (1.645\left(\frac{46}{\sqrt{n}}\right)$$

and thus,

$$n = \left[\frac{(2.32 - (-1.645))}{211 - 180}(46)\right]^2 = 34.6 \approx 35$$

In general,

$$n = \left[\frac{(z_\alpha + z_\beta)}{\mu_0 - \mu_1}\sigma\right]^2$$

To be assured of a 95% chance of detecting differences in cholesterol level between 20-74 and 20-24 year-old males (power) when carrying out the test at a 1% α level, we would need about 35 20-74 year-old males.

7.4 Computer implementation

Power and sample size calculations are performed in STATA via the command `sampsi`. The syntax is as follows (the underlined parts of the command are used to calculate the sample size and are omitted when calculating the power).

```
sampsi #1 #2 [,alpha(#) power(#) n1(#) n2(#) ratio(#) pre(#)
post(#)sd1(#)sd2(#) method(post|change|ancova|all) r0(#)
r1(#) r01(#) onesample onesided ]
```

In the two-sample case, when `n2` (and `ratio`) and/or `sd2` is omitted, they are assumed equal to `n1` and `sd1` respectively. You can use options `n1` and `ratio` (=`n2`/`n1`) to specify `n2`. The default is a two-sample comparison. In the one-sample case (population mean is known exactly) use option `onesample`. Options `pre(#)`, `post(#)`, `method(post|change|ancova|all)`, `r0(#)`, `r1(#)` refer to repeated-measures designs and are beyond the scope of this course.

7.4.1 Power calculations

In the cholesterol example, we will use the STATA command `sampsi` to compute the power of the study ($n = 25$, and $\sigma = 46$mg/ml) as follows:

```
. sampsi 180 211, alpha(.05) sd1(46) n1(25) onesample onesided

Estimated power for one-sample comparison of mean
  to hypothesized value

Test Ho: m =    180, where m is the mean in the population

Assumptions:
         alpha =   0.0500  (one-sided)
 alternative m =      211
            sd =       46
 sample size n =       25

Estimated power:
         power =   0.9577
```

The power is $0.9577 \approx 0.958$ as we saw earlier.

7.4.2 Sample size calculations

The sample size under $\alpha = 0.01$ and power 95%, is calculated as follows:

```
  . sampsi 180 211, alpha(.01) power(.95) sd1(46) onesample onesided

Estimated sample size for one-sample comparison of mean
  to hypothesized value

Test Ho: m =    180, where m is the mean in the population
Assumptions:

         alpha =   0.0100  (one-sided)
         power =   0.9500
 alternative m =      211
            sd =       46
Estimated required sample size:
             n =       35
```

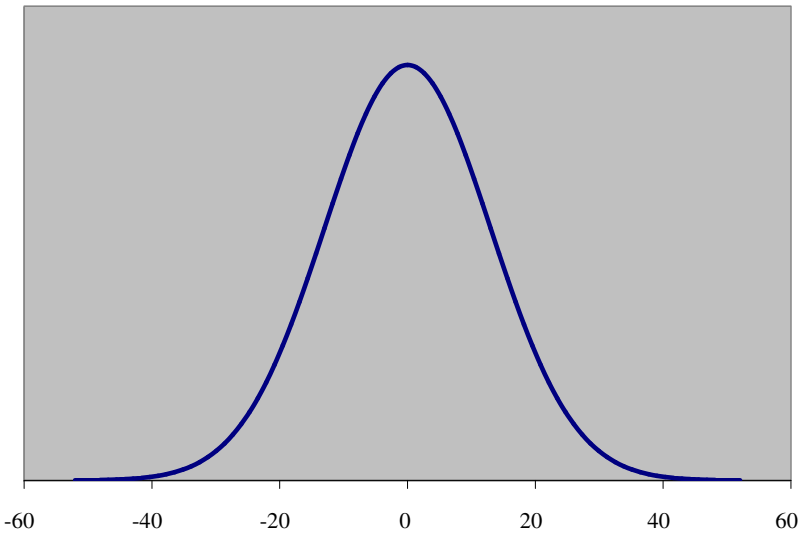
We use options `onesample` and `onesided` since we are working with a single sample ($(\mu)_0 = 180$ mg/ml is the population mean) and carry out a one-sided test.

7.5 The two (independent) sample case

In the two-sample case, recall that the null hypothesis is (usually) $H_o : \mu_1 = \mu_2$. This is equivalent to $H_o : \delta = \mu_1 - \mu 2 = 0$. Power and sample-size calculations are based on the distribution of the sample difference of the two means $\bar{d} = \bar{X}_1 - \bar{X}_2$ under some *a priori* assumptions.

The distribution of the sample difference of two means, assuming two equal-size $n_1 = n_2 = n$

Figure 7.5: Distribution of $\bar{d} \sim N(0, 13.01)$



(say) independent samples and known and equal variances ($\sigma_1 = \sigma_2 = s$) is $\bar{d} \sim N(\delta, \sigma_\delta)$, where $\sigma_\delta = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = \sigma\sqrt{\frac{2}{n}}$ if $n_1 = n_2 = n$.
Example: Cholesterol example (continued)

If the mean cholesterol among 20-24 year-olds were also unknown, while the standard deviation in both groups were $\sigma = 46$ mg/ml, two samples would be collected. One from a group of 20-24 year-olds and one from a group of 20-74 year-olds. Under the null $H_o : \mu_1 = \mu_2$ $(= d = 0)$ the distribution of $\bar{d} \sim N(0, \sigma_{\bar{d}})$ with $\sigma_d = \sigma\sqrt{\frac{2}{n}} = 13.01$ mg/dL, as depicted in Figure 7.5. If we wanted to carry out the cholesterol study in a manner identical to the previous discussion (with the exception of course that neither population mean is assumed known *a priori*) the procedure would be as follows:

1. H_o: $d = 0$
2. H_a: $d = 31$ mg/ml (corresponding to the situation where $\mu_1 = 180$ mg/ml and $\mu_2 = 211$ mg/ml)
3. α=0.01
4. Power=$1 - \beta = 0.95$

To calculate the sample size for each group (n) we can use the previous one-sample formula, with the appropriate estimate of the variance of course. That is, each group will be comprised of individuals from each population,

$$n\prime = \left[\frac{(z_\alpha + z_\beta)}{\delta_\alpha}\sigma_d\right]^2 = \left[\frac{(z_\alpha + z_\beta)}{\delta_\alpha}\sigma\sqrt{2}\right]^2 = 2\left[\frac{(z_\alpha + z_\beta)}{\delta_\alpha}\sigma\right]^2 = 2n$$

where n is the size of the identically defined one-sample case. That is, the sample size in the two-sample case will be roughly double that of the one-sample case. In this case, $n\prime = 2\left[\frac{(2.32+1.645)}{31}(46)\right]^2 =$

69.23 The required sample size is at least 70 subjects per group (double the $n = 35$ subjects required in the identical one-sample study).

Note! The total required sample is 140 subjects, or *four times that of the single-sample study.* This is the penalty of ignorance of both means versus just one out of the two means.

7.6 Computer implementation

The computer implementation is given below.

```
  . sampsi 180 211, alpha(.01) power(.95) sd1(46) onesided
Estimated sample size for two-sample comparison of means

Test Ho: m1 = m2, where m1 is the mean in population 1
                    and m2 is the mean in population 2
Assumptions:

         alpha =   0.0100  (one-sided)
         power =   0.9500
            m1 =      211
            m2 =      180
           sd1 =       46
           sd2 =       46
         n2/n1 =     1.00

Estimated required sample sizes:
            n1 =       70
            n2 =       70
```

7.7 Power calculations when testing a single proportion

Power and sample size calculations can be carried out when the comparisons involve proportions. We discuss the one-sample case, as the two-sample case is a bit more complicated. When determining power or sample size for the comparison of a proportion (against a known population value), we make use of the normal approximation of the binomial distribution. That is, we use the fact that, at least for large n,

$$Z = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

where p is the unknown population proportion, and $\hat{p} = \frac{x}{n}$ is the ratio of successes over the total number of trials (experiments).

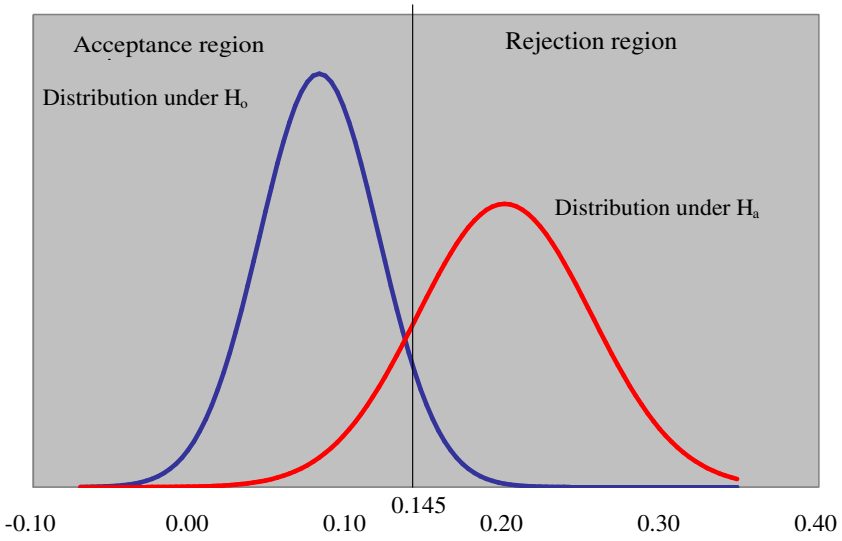
Example: Five-year survival rate of lung-cancer patients

Suppose that we are testing the hypothesis that the 5-year survival of lung-cancer patients under 40 is 8.2%, equal to 5-year survival rates of lung-cancer patients over 40 years-old. That is, we test H_o : $p_o \leq 0.082$ versus $H_a : p_0 > 0.082$. If the 5-year survival among younger patients is as high as 20% (i.e., $p_a = 0.200$), and the study sample size is $n = 52$, then $\sigma_{\hat{p}} = \sqrt{\frac{p_0(1-p_0}{n}} = \sqrt{\frac{0.082(1-0.082}{52}}] = 0.038$ under the null hypothesis, and $\sigma_{\hat{p}} = \sqrt{\frac{p_a(1-p_a}{n}} = \sqrt{\frac{0.200(1-0.200}{52}}] = 0.055$, under the alternative. The null hypothesis will be rejected if $Z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} > z_\alpha = 1.645$, that is, if

$$\hat{p} > p_0 + z_\alpha\sqrt{\frac{p_0(1-p_0)}{n}} = 0.082 + 1.645\sqrt{\frac{0.082(1-0.082)}{52}} \approx 0.145$$

This situation is depicted in the following figure: **Example:** Five-year survival of lung-cancer

Figure 7.6: Distribution of $\hat{p}$ under the null hypothesis $\hat{p} \sim N(0.082, 0.038)$ (blue) and alternative hypothesis $\hat{p} \sim N(0.200, 0.055)$ (red)



patients (continued):

First of all,

$$\begin{aligned} z_\beta &= \frac{(p_a - p_0) - z_\alpha\sqrt{\frac{p_0(1-p_0)}{n}}}{\sqrt{\frac{p_a(1-p_a)}{n}}} \\ &= \frac{(0.200 - 0.082)\, 1.645\sqrt{\frac{0.082(1-0.082)}{52}}}{\sqrt{\frac{0.200(1-0.200)}{52}}} \approx 1.00 \end{aligned}$$

Thus, the probability of a type-II error is $\beta = P(Z > z_\beta) = 0.159$ and thus, the power of a test for

a single proportion based on $n = 52$ subjects is $1 - \beta = 0.841$ or about 84%.

If the power is 95% and the alpha level of the test is 1%, then the required sample size is

$$\begin{aligned} n &= \left[\frac{z_\beta \sqrt{\frac{p_a(1-p_a)}{n}} + z_\alpha \sqrt{\frac{p_0(1-p_0)}{n}}}{(p_a - p_0)} \right]^2 \\ &= \left[\frac{1.645\sqrt{\frac{0.200(1-0.200)}{52}} + 2.32\sqrt{\frac{0.082(1-0.082)}{52}}}{(0.200 - 0.082)} \right]^2 = 120.4 \end{aligned}$$

That is, about 121 lung cancer patients under 40 years old will be necessary to be followed, and their 5-year survival status determined in order to ensure power of 95% when carrying out the test at the 1% alpha level.

7.7.1 Computer implementation of power calculations for propotions

The power of a study testing $H_o : p = 0.082$ (known) versus H_a : p=0.200, involving $n = 52$ subjects, with a statistical test performed at the 5% alpha level, is computed with STATA as follows:

```
. sampsi .082 .200, n(52) alpha(.05) onesample onesided

Estimated power for one-sample comparison of proportion
  to hypothesized value

Test Ho: p = 0.0820, where p is the proportion in the population

Assumptions:

         alpha =   0.0500  (one-sided)
 alternative p =   0.2000
 sample size n =       52

Estimated power:

power =   0.8411
```

Notice that omission of estimates for the standard deviation (`sd1` and/or `sd2`) produced power calculations for proportions.

7.7.2 Computer implementation for sample-size calculations for proportions

In the case where the same hypotheses as above are tested, assuming a power of 95% and alpha level of 1%, the required sample size will be computed as follows:

```
. sampsi .082 .200, alpha(.01) power(0.95) onesample onesided

Estimated sample size for one-sample comparison of proportion
  to hypothesized value

Test Ho: p = 0.0820, where p is the proportion in the population

Assumptions:

         alpha =   0.0100  (one-sided)
         power =   0.9500
 alternative p =   0.2000

Estimated required sample size:

n =      121
```

Thus, $n = 121$ subjects will be necessary to be involved in the study and followed for 5-year survival.

Chapter 8

Contingency tables

Consider the following table:

	Wearing Helmet		
Head Injury	Yes	No	Total
Yes	17	218	235
No	130	428	558
Total	147	646	793

If we want to test whether the proportion of unprotected cyclists that have serious head injuries is higher than that of protected cyclists, we can carry out a general test of hypothesis involving the two proportions $p_1 = 17/147 = .115$, and $p_2 = 218/646 = .337$.

The normal variable associated with the difference between p_1 (protected cyclists having head injuries), and p_2 (unprotected cyclists with head injuries) is $z = -5.323$, the null hypothesis is rejected at the 95% significance level.

Now suppose that you wanted to determine whether there is any *association* between wearing helmets and frequency of brain injuries based on the same data. Then you must perform a different test, called the chi-square test because it is based on the χ^2 distribution, which we will cover momentarily. This test is set up as follows:

1. H_o: Suffering a head injury is not associated with wearing a helmet
2. H_a: There is an association between wearing a helmet and suffering a head injury
3. Specify the alpha level of the test
4. Rejection rule (two-sided only):
   Reject H_o if the chi-square statistic is too large (see discussion below)

Consider the following output corresponding to that test:

```
. prtesti 147 .115 646 .337

Two-sample test of proportion                 x: Number of obs =      147
                                              y: Number of obs =      646
------------------------------------------------------------------------------
Variable |      Mean    Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |      .115    .0263125   4.37055   0.0000       .0634285    .1665715
       y |      .337    .0185975   18.1207   0.0000       .3005495    .3734505
---------+--------------------------------------------------------------------
    diff |     -.222    .0417089    5.3226   0.0000       -.303748    -.140252
------------------------------------------------------------------------------
               Ho: proportion(x) - proportion(y) = diff = 0

       Ha: diff < 0                 Ha: diff ~= 0                 Ha: diff > 0
        z = -5.323                   z = -5.323                   z = -5.323
      P < z = 0.0000           P > |z| = 0.0000           P > z = 1.0000
```

Now let us consider the implication of the null hypothesis:

If the distinction between the two groups (helmet wearers and non-helmet wearers) is an artificial one, then the head-injury rate is better estimated by $\hat{p} = \frac{235}{793} = 0.2963$, the overall incidence rate of head injuries.

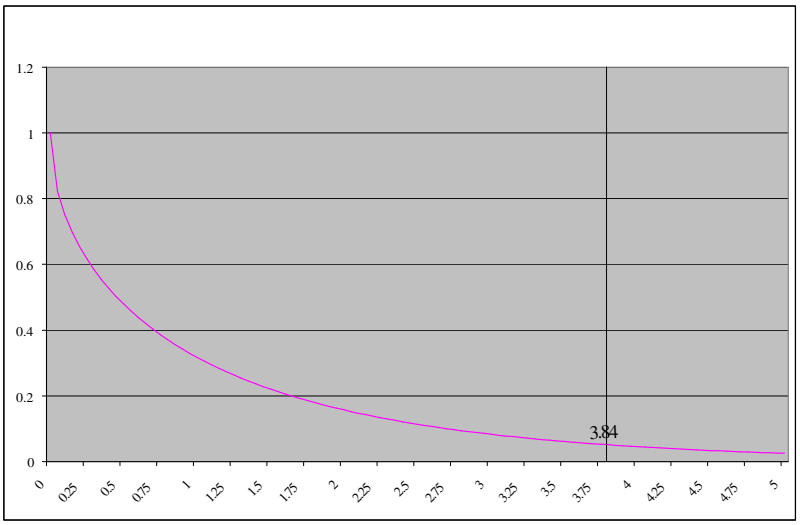
The expected number of injured protected cyclists is $(0.2963)(147) = 43.6$ on average (versus the observed 17). Similarly, the number of injured unprotected cyclists should be $(0.2963)(646) = 191.4$ (versus the observed 218). The expected number of uninjured helmeted cyclists is $(1 - 0.2963)(147) = 103.4$ (versus the observed 130), and the expected number of unprotected uninjured cyclists is $(1 - 0.2963)(646) = 454.6$ (versus the observed 428). Our test should be based on quantifying whether deviations from these two expected numbers are serious enough to warrant rejection of the null hypothesis.

In general, the chi square test looks like this:

$$\chi^2 = \sum_{i=1}^{rc} \frac{(O_i - E_i)^2}{E_i}$$

E_i is the expected number of head injuries, O_i is the observed number, r is the number of rows in the table, and c is the number of columns. Then, χ^2 is distributed according to the chi square distribution with $df = (r-1)(c-1)$ degrees of freedom. Critical percentiles of the chi-square distribution can be found in the appendix of your textbook. The chi-square distribution with one degree of freedom is shown below:

Figure 8.1: Chi-square distribution with one degree of freedom



Returning to the above example, the chi square test is:

$$\begin{aligned}
\chi^2 &= \sum_{i=1}^{rc} \frac{(|O_i - E_i| - 0.5)^2}{E_i} \\
&= \frac{(|17 - 43.6| - 0.5)^2}{43.6} + \frac{(|130 - 103.4| - 0.5)^2}{103.4} + \frac{(|217 - 191.4| - 0.5)^2}{191.4} + \frac{(|428 - 454.6| - 0.5)^2}{454.6} \\
&= 15.62 + 6.59 + 3.56 + 1.50 = 27.27
\end{aligned}$$

It is clear that large deviations of the observed counts from the expected ones will lead to large chi-square statistics. Thus, large values of χ^2 contradict the null hypothesis. The cutoff point of the chi-square distribution is determined by the number of degrees of freedom and the alpha level of the test. In the case of the previous example, the number of degrees of freedom is $(r-1)\times(c-1) = (2-1)\times(2-1) = 1$. For $\alpha = 0.05$ the point to the right of which lies 5% of the chi-square distribution with one degree of freedom is 3.84.

The chi-square test in the previous example is implemented as follows:

1. H_o: Suffering a head injury is not associated with wearing a helmet
2. H_a: There is an association between wearing a helmet and suffering a head injury
3. $\alpha = 0.05$
4. Rejection rule (two-sided only):
   Reject H_o if the chi-square statistic is higher than $\chi^2(1)_{0.05} = 3.84$

Comparing the observed value of the statistic to 3.84 we reject the null hypothesis as 27.27 is much higher than 3.84. In fact the p value of the test is give easily by STATA as follows:

```
. display chiprob(1,27.27)
1.769e-07
```

In other words the probability under the chi-square distribution with one degree of freedom to the right of 27.27 is 0.0000001769 or 1.769 in ten million! This is the chance of the null hypothesis (of no association between wearing a helmet and suffering a head injury) is correct!

Since this probability is smaller than $\alpha = 0.05$ this is an alternative justification for rejecting the null hypothesis.

8.0.3 Computer implementation

Carrying out the test using STATA is as follows:

```
. tabi 17 218\ 130 428, chi

           |          col
       row |         1          2 |     Total
-----------+----------------------+----------
         1 |        17        218 |       235
         2 |       130        428 |       558
-----------+----------------------+----------
     Total |       147        646 |       793

          Pearson chi2(1) =  28.2555   Pr = 0.000
```

Columns 1 and 2 and rows 1 and 2 correspond to "Yes" and "No" (wore helmet?) and (sustained head injury?) respectively. The p-value of the test is $0.0000 < 0.05$, so we reject the null hypothesis. There is an association between wearing a helmet and serious head injury. STATA calculates the Pearson χ^2 a bit differently (the value of the STATA-generated statistic is 28.2555 instead of 27.27, which is slightly different from our hand-calculations).

8.1 The odds ratio

The chi square test of association answers only the question of association. It does not comment on the nature or direction of the association. For a further investigation of this hypothesis one must rely on a different test.

We define as the odds of having the disease if exposed is

$$\frac{P(\text{disease}|\text{exposed})}{[1 - P(\text{disease}|\text{exposed})]}$$

The odds of having the disease if unexposed is

$$\frac{P(\text{disease}|\text{unexposed})}{[1 - P(\text{disease}|\text{unexposed})]}$$

The odds ratio (OR) is defined as:

$$\text{OR} = \frac{\frac{P(\text{disease}|\text{exposed})}{[1-P(\text{disease}|\text{exposed})]}}{\frac{P(\text{disease}|\text{unexposed})}{[1-P(\text{disease}|\text{unexposed})]}}$$

Consider the following 2×2 table: An estimate of the odds ratio is

	Exposed	Unexposed	Total
Disease	a	b	$a+b$
No disease	c	d	$c+d$
Total	$a+c$	$b+d$	n

$$\widehat{\text{OR}} = \frac{\frac{a/(a+c)}{c/(a+c)}}{\frac{b/(b+d)}{d/(b+d)}} = \frac{a/c}{b/d} = \frac{ad}{bc}$$

To construct statistical tests of hypotheses involving the odds ratio we must determine its distribution. The OR itself is not distributed normally; but its logarithm is.

In fact, the statistic

$$Z = \frac{\ln\left(\frac{ad}{bc}\right)}{\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}}$$

is approximately distributed according to the standard normal distribution. Tests and confidence intervals are derived as usual.

8.1.1 Testing the hypothesis of no association (using the odds ratio)

The statistical test constructed around the odds ratio is as follows:

1. H_o: There is no association between exposure and disease
   (a) One-sided alternatives
      - H_o: OR ≥ 1
      - H_o: OR ≤ 1
   (b) Two-sided alternative
   H_o: OR $= 1$
2. H_a: There is an association between exposure and disease
   (a) One-sided alternatives
      - H_a: OR < 1
      - H_o: OR > 1
   (b) Two-sided alternative
   H_a: OR $\neq 1$
3. Specify the alpha level of the test

4. The test statistic is $Z = \frac{\ln\left(\frac{ad}{bc}\right)}{\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}} \sim N(0,1)$

5. Rejection rule:

   (a) One-sided alternatives

   - Reject the null hypothesis if $Z > z_{1-\alpha}$
   - Reject the null hypothesis if $Z < z_{1-\alpha}$

   (b) Two-sided alternative
   Reject the null hypothesis if $Z > z_{1-\frac{\alpha}{2}}$ or if $Z < -z_{1-\frac{\alpha}{2}}$

8.1.2 Confidence intervals

A $(1-\alpha)\%$ two-sided confidence interval of the *log-odds* ratio is given by

$$\left(\ln(\widehat{\text{OR}}) - z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}},\ \ln(\widehat{\text{OR}}) + z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right)$$

Thus, the $(1-\alpha)\%$ confidence interval of the true odds ratio is given by

$$\boxed{\left(\exp\left\{\ln(\widehat{\text{OR}}) - z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right\},\ \exp\left\{\ln(\widehat{\text{OR}}) + z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right\}\right)}$$

where $\ln(x)$ is the *natural logarithm* (or logarithm base e) of x, i.e., $e^{\ln(x)} = x$; $\exp(x)$ is the same as e^x. Finally, $e \approx 2.718$.

This confidence interval can also be used to perform a hypothesis test by inspecting whether it covers 1 (the OR hypothesized value under the null hypothesis. **Example:** Consider the following data on use of EFM (Electronic Fetal Monitoring) and frequency of Caesarean birth deliveries. The table is as follows: To test the test of hypothesis of no association between EFM and Caesarean

Caesarean delivery	EFM exposure		Total
	Exposed	Unexposed	
Yes	358	229	587
No	2,492	2,745	5,237
Total	2,850	2,974	5,824

births, we base our inference on the statistic

$$Z = \frac{\ln\left(\frac{ad}{bc}\right)}{\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}} = \frac{\ln\left(\frac{358\times 2745}{229\times 2492}\right)}{\sqrt{\frac{1}{358}+\frac{1}{229}+\frac{1}{2492}+\frac{1}{2745}}} = \frac{\ln(1.72)}{0.089} = 6.107$$

Since $Z = 6.107 > 1.96 = z_{0.975}$, we reject the null hypothesis (in favor of the two-sided alternative).

These data support a rather strong (positive) association between EFM and Caesarean births.

On the other hand, the 95% confidence interval is given by

$$(\exp\{\ln(1.72) - (1.96)(0.089)\},\ \exp\{\ln(1.72) + (1.96)(0.089)\}) = \left(e^{0.368},\ e^{0.716}\right) = (1.44,\ 2.052)$$

Notice that 1 is not contained in the above confidence interval. This is consistent to the result of the test of hypothesis, which rejected the null hypothesis of no association between EFM exposure and risk of caesarean sections. The estimated odds ratio among women monitored via EFM, is from 44% higher to over double that of women that were not monitored by EFM.

8.1.3 Computer implementation

This example is handled by the `cci` (case-control; immediate version):

```
. cci 358 229 2492 2745, cornfield
                                                        Proportion
                 |   Exposed   Unexposed  |     Total     Exposed
-----------------+------------------------+----------------------
           Cases |       358          229  |       587      0.6099
        Controls |      2492         2745  |      5237      0.4758
-----------------+------------------------+----------------------
           Total |      2850         2974  |      5824      0.4894
                 |                         |
                 |      Point estimate     |  [95% Conf. Interval]
                 |-------------------------+----------------------
      Odds ratio |          1.722035       |  1.446551    2.049976 (Cornfield)
 Attr. frac. ex. |          .4192916       |  .3087003    .5121894 (Cornfield)
 Attr. frac. pop |          .2557178       |
                 +-------------------------------------------------
                             chi2(1) =    37.95  Pr>chi2 = 0.0000
```

The odds ratio is 1.72 with a 95% confidence interval (1.447, 2.050). Thus, the null hypothesis of no association is rejected as both limits of the confidence interval are above 1.0.

8.2 Combining 2×2 contingency tables

Consider the following data involving the association between myocardial infarction (MI) and coffee consumption among smokers and non-smokers:

Smokers

Non-smokers

The question that naturally arises is whether we should combine the information in those two tables and use all the available data in order to ascertain the effect of coffee on the risk of Myocardial Infarction (MI). However, if the association between coffee and MI is different in the group of smokers compared to the group of non-smokers, then such an analysis would be inappropriate. In

Myocardial infarction	Coffee consumption		Total
	Yes	No	
Yes	1,011	81	1,092
No	390	77	467
Total	1,401	158	1,559

Myocardial infarction	Coffee consumption		Total
	Yes	No	
Yes	383	66	449
No	365	123	488
Total	748	189	937

general we have g tables ($i = 1, ..., g$) that are constructed as follows ($g = 2$ in the previous example):

Disease	Exposure		Total
	Yes	No	
Yes	a_i	b_i	N_{1i}
No	c_i	d_i	N_{2i}
Total	M_{1i}	M_{2i}	T_i

We employ the following strategy:

1. Analyze the two tables separately. Based on the individual estimates of the odds ratios,
2. Test the hypothesis that the odds ratios in the two subgroups are sufficiently close to each other (they are homogeneous).
   (a) If the assumption of homogeneity is not rejected then perform an overall (combined) analysis.
   (b) If the homogeneity assumption is rejected, then perform separate analyses (the association of the two factors is different in each subgroup).

The test of homogeneity is set-up as follows:

1. H_o: $\text{OR}_1 = \text{OR}_2$ (the two odds ratios do not have to be 1, just equal)
2. H_a: $\text{OR}_1 \neq \text{OR}_2$ (only two-sided alternatives are possible with the chi square test)
3. The test statistic is
$$X^2 = \sum_{i=1}^{g} w_i (y_i - Y)^2 \sim \chi^2(1)$$
4. Rejection rule. Reject the null hypothesis (conclude that the two subgroups are not homogeneous if $X^2 > \chi^2(1)_{1-\alpha}$.

In the discussion above X^2 has an approximate chi-square distribution with 1 degree of freedom. Here $y_i = \ln(\widehat{\text{OR}}_\text{i}) = \ln\left(\frac{a_i c_i}{b_i d_i}\right)$, $w_i = \frac{1}{\left[\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}\right]}$, $i = 1, ..., g$ [1] and $Y = \frac{\sum_{i=1}^{g} w_i y_i}{\sum_{i=1}^{g} w_i}$. That is, we combine the individual odds ratios, producing a weighted average, weighing each of them inversely proportional to the square of their variance to downweight odds ratios with high variability. High variability means low information. The statistic is the combined (weighted) distances of the individual log-odds ratios from the weighted mean. If the deviations are large (lack of homogeneity) the null hypothesis will be rejected and the conclusion will be that the association between the two primary factors is different (heterogeneous) from one subgroup (as defined by the third factor) to another.

In the previous example, among smokers, $\widehat{\text{OR}}_1 = \frac{(1011)(77)}{(390)(81)} = 2.46$. Thus, $y_1 = \ln(\widehat{\text{OR}}_1) = \ln(2.46) = 0.900$. Among non-smokers, $\widehat{\text{OR}}_2 = \frac{(383)(123)}{(365)(66)} = 1.96$. Thus, $y_2 = \ln(\widehat{\text{OR}}_2) = \ln(1.96) = 0.673$. The weights are $w_1 = \frac{1}{\left[\frac{1}{1011}+\frac{1}{390}+\frac{1}{81}+\frac{1}{77}\right]} = 34.62$ and $w_2 = \frac{1}{\left[\frac{1}{383}+\frac{1}{365}+\frac{1}{66}+\frac{1}{123}\right]} = 34.93$. The common odds ratio is $Y = 0.786$.

From the expression of the test for homogeneity that we just described, we have

$$\begin{aligned} X^2 &= \sum_{i=1}^{2} w_1 (y_i - Y)^2 \\ &= (34.62)(0.900 - 0.786)^2 + (34.93)(0.673 - 0.786)^2 = 0.896 \end{aligned}$$

Since 0.896 is not larger than any usual critical value (as seen in the Appendix), we do not reject the null hypothesis. There is no evidence that the association between coffee consumption and occurrence of MI differs between smokers and non-smokers. It is appropriate to proceed with the overall analysis.

8.2.1 Confidence intervals of the overall odds ratio

The confidence intervals of the overall ratio are constructed similarly to the one-sample case. The only difference is the estimate of the overall log odds ratio, and its associated standard error. In general a $(1-\alpha)\%$ confidence interval for the combined odds ratio $Y = \ln(\text{OR})$ based on the standard normal distribution is constructed as follows:

$$\left(Y - z_{1-\frac{\alpha}{2}}\text{s.e.}(Y),\ Y + z_{1-\frac{\alpha}{2}}\text{s.e.}(Y)\right)$$

The combined log odds ratio is $Y = \frac{\sum_{i=1}^{g} w_i y_i}{\sum_{i=1}^{g} w_i}$ and $\text{s.e.}(Y) = \frac{1}{\sum_{i=1}^{g} w_i}$, where the w_i are defined as before. Since $Y = \ln(\text{OR})$, the $(1-\alpha)\%$ confidence interval of the common odds ratio is

$$\boxed{\left(\exp\left\{Y - z_{1-\frac{\alpha}{2}}\text{s.e.}(Y)\right\},\ \exp\left\{Y + z_{1-\frac{\alpha}{2}}\text{s.e.}(Y)\right\}\right)}$$

In the previous example, a 95% confidence interval is

$$\left(e^{0.786-(1.96)(0.120)},\ e^{0.786+(1.96)(0.120)}\right) = \left(e^{0.551},\ e^{1.021}\right) = (1.73,\ 2.78)$$

[1] If any of the cell counts is zero then $w_i = \frac{1}{\left[\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}\right]}$. Notice that $w_i = [\text{s.e.}(\text{OR})]^2$.

Thus, at the 95% level of significance, coffee drinkers have from 73% higher risk for developing MI, to almost triple the risk, compared to non-coffee drinkers. Since this interval does not contain 1, this implies that we should reject the null hypothesis of no (overall) association between coffee consumption and MI adjusted for smoking status.

8.2.2 The Mantel-Haenszel test for association

The Mantel-Haenszel is based on the chi square distribution and the simple idea that if there is no association between "exposure" and "disease", then the number of exposed individuals a_i contracting the disease should not be too different from $m_i = \frac{M_i N_i}{T_i}$. To see why this makes sense, recall that under independence (i.e., no association), the probability $P(A \cap B) = P(A)P(B)$. If A is the event"Subject has the disease", and B "Subject is exposed" then $P(A \cap B) = \frac{a_i}{T_i}$, and $P(A) = \frac{M_i}{T_i}$ while $P(B) = \frac{N_i}{T_i}$, $i = 1, ..., g$.

Thus, under the assumption of independence (no association), $P(A \cap B) = \frac{a_i}{T_i} = \frac{M_i}{T_i}\frac{N_i}{T_i} = P(A)P(B)$ and finally, $a_i = \frac{M_i N_i}{T_i}$. A less obvious estimate of the variance of ai is $\sigma_i = \sqrt{\frac{M_{1i} M_{2i} N_{1i} N_{2i}}{T_i^2 (T_i - 1)}}$.

The Mantel Haenszel test is constructed as follows:

1. H_o: $\mathrm{OR}_1 = 1$
2. H_a: $\mathrm{OR}_1 \neq 1$ (only two-sided alternatives can be accommodated by the chi-square test)
3. The Mantel-Haenszel test statistic is $X^2_{MH} = \frac{\left[\sum_{i=1}^g a_i - \sum_{i=1}^g m_i\right]^2}{\sum_{i=1}^g \sigma_i^2} \sim \chi^2(1)$
4. Rejection rule: Reject H_o if $X^2 > \chi^2(1)_{1-\alpha}$.

where $m_i = \frac{M_i T_i}{T_i}$, $i = 1, ..., g$ are the expected counts of diseased exposed individuals.

In the previous example, $a_1 = 1011$, $m_1 = 981.3$, $\sigma_1^2 = 29.81$, $a_2 = 383$, $m_2 = 358.4$, $\sigma_2^2 = 37.69$.

Thus,

$$\begin{aligned} X^2_{MH} &= \frac{\left[\sum_{i=1}^g a_i - \sum_{i=1}^g m_i\right]^2}{\sum_{i=1}^g \sigma_i^2} \\ &= \frac{\left[(1011 + 863) - (981.3 + 358.4)\right]^2}{29.81 + 37.69} = 43.68 \end{aligned}$$

Since 43.68 is much larger than 3.84 the 5% tail of the chi-square distribution with 1 degree of freedom we reject the null hypothesis. Coffee consumption has a significant positive association with the risk of M.I. across smokers and non-smokers.

8.2.3 Computer implementation

To run the Mantel-Haenszel test we must have each individual subject's data comprised of their coffee drinking (e.g., 0=Yes, 1=No) and similarly whether they have suffered an MI and whether they are a smoker or not. In the absence of the raw data (2496 lines in total) we will have the counts from the tables. Unlike the one-sample case however, there is no immediate command that performs the M-H test directly. First create a dataset containing the frequencies from the table as follows:

```
. input smoke MI coffee count
            smoke         MI     coffee       count
  1. 0 0 0 1011
  2. 0 0 1 81
  3. 0 1 0 390
  4. 0 1 1 77
  5. 1 0 0 383
  6. 1 0 1 66
  7. 1 1 0 365
  8. 1 1 1 123
  9. end

. label define yesno 1 No 0 Yes
. label val smoke yesno
. label val MI yesno
. label val coffee yesno
```

To check whether we have recreated the tables correctly we do the following:

```
. sort smoke

. by smoke: tab MI coffee [freq=count]

-> smoke=   Smoker
           | coffee
        MI |       Yes         No |     Total
-----------+----------------------+----------
       Yes |      1011         81 |      1092
        No |       390         77 |       467
-----------+----------------------+----------
     Total |      1401        158 |      1559


-> smoke= Non-smok
           | coffee
        MI |       Yes         No |     Total
-----------+----------------------+----------
       Yes |       383         66 |       464
        No |       365        123 |       488
-----------+----------------------+----------
     Total |       748        189 |       952
```

Then we carry out the M-H test but remembering that each line of data is not a single line, but it represents as many subjects as the number in the variable `count`.

This is done as follows:

```
. cc MI coffee [freq=count], by(smoke) cornfield

           smoke |       OR      [95% Conf. Interval]    M-H Weight
-----------------+-------------------------------------------------
          Smoker |   2.464292     1.767987   3.434895      20.26299 (Cornfield)
        Non-smok |   1.955542     1.404741   2.722139      25.70971 (Cornfield)
-----------------+-------------------------------------------------
           Crude |   2.512051     1.995313   3.162595               (Cornfield)
    M-H combined |   2.179779     1.721225   2.760499
-----------------+-------------------------------------------------
Test for heterogeneity (M-H)     chi2(1) =     0.933  Pr>chi2 = 0.3342

                  Test that combined OR = 1:
                              Mantel-Haenszel chi2(1) =     43.58
                                              Pr>chi2 =    0.0000
```

Following the earlier strategy, the analysis can be performed from the previous output

1. Analyze the two tables separately.
   The odds ratio among smokers is 2.464292, and among non-smokers is 1.955542.

2. Test of the homogeneity of the association between coffee consumption and MI
   The test of homogeneity ("test for heterogeneity" in STATA) has a p-value $0.3342 > 0.05$. We do not reject the hypothesis of homogeneity in the two groups. A combined analysis can be carried out over both smokers and non-smokers

3. Since the assumption of homogeneity was not rejected we perform an overall (combined) analysis. From this analysis, the hypothesis of no association between coffee consumption and myocardial infarction is rejected at the 95% alpha level (since the M-H p-value $0.0000 < 0.05$).

   By inspection of the combined Mantel-Haenszel estimate of the odds-ratio (2.179779) we see that the risk of coffee drinkers (adjusting for smoking status) is over twice as high as that of non-coffee drinkers.

Chapter 9

Analysis of Variance

Patients from 3 centers, Johns Hopkins, Rancho Los Amigos, and St. Louis, were involved in a clinical trial. As a part of their baseline evaluations, the patients' pulmonary performance was assessed. A good marker of this is the Forced Expiratory Volume in 1 second FEV_1. The data are presented in Table 12.1 of the textbook and the STATA output below.

It was important to the investigators to ascertain whether the patients from the 3 centers had on average similar pulmonary function before the beginning of the trial.

The DATA set

```
. list
          FEV1              center
  1.      3.23      Johns Hopkins
  2.      3.47      Johns Hopkins
  3.      1.86      Johns Hopkins
  4.      2.47      Johns Hopkins
  5.      3.01      Johns Hopkins
   .       .
   .       .
   .       .
 10.      3.36      Johns Hopkins
 11.      2.61      Johns Hopkins
 12.      2.91      Johns Hopkins
   .       .
   .       .
   .       .
 57.      2.85           St Louis
 58.      2.43           St Louis
 59.       3.2           St Louis
 60.      3.53           St Louis
```

Johns Hopkins: (center==1)
Rancho Los Amigo (center==2)
St Louis (center==3)

STATA Summary Statistics

```
. sort center

. by center: summarize FEV1

-> center=Johns Ho

Variable |     Obs        Mean   Std. Dev.       Min        Max
---------+--------------------------------------------------------
    FEV1 |      21     2.62619   .4961701       1.69       3.47

-> center=Rancho L

Variable |     Obs        Mean   Std. Dev.       Min        Max
---------+--------------------------------------------------------
    FEV1 |      16      3.0325   .5232399       1.71       3.86

-> center=St. Loui

Variable |     Obs        Mean   Std. Dev.       Min        Max
---------+--------------------------------------------------------
    FEV1 |      23    2.878696   .4977157       1.98       4.06
```

To address the investigators' concerns we must compare the average pulmonary function of the patients at the 3 sites. Since the population mean and standard deviation of the pulmonary function at each site is not known, we must estimate them from the data.

In general, when k such groups are involved we have the following:

	Group 1	**Group 2**	...	**Group k**
Population				
Mean	μ_1	μ_2	...	μ_k
Std. deviation	σ_1	σ_2	...	σ_k
Sample				
Mean	$\bar{x}_1$	$\bar{x}_2$	...	$\bar{x}_k$
Std. deviation	s_1	s_2	...	s_k
Sample size	n_1	n_2	...	n_k

We must use the sample information, in order to perform inference (hypothesis testing, confidence intervals, etc.) on the population parameters.

A statistical test addressing this question is constructed as follows:

1. $H_o : \mu_1 = \mu_2 = \ldots = \mu_k$
2. H_a: At least one pair is not equal
3. The level of significance is $(1-\alpha)\%$

Question: Can we use the two-sample t test to perform these comparisons?
Answer: The two-sample t test cannot directly address this hypothesis, because these comparisons involve more than 2 groups. However, we can use the two-sample t test to perform every possible pair-wise comparison among the k groups. In the case of $k = 3$, $g = 3$ pair-wise comparisons must be performed (i.e., Group 1 vs Group 2, Group 1 vs. Group 3, and Group 2 vs. Group 3).

9.1 t test for equality of k group means

A test of the overall hypothesis of equality among the k means, based on two-sample t tests of all g pair-wise group comparisons, is constructed as follows:

1. $H_o : \mu_1 = \mu_2 = \ldots = \mu_k$ versus H_a : At least one pair is not equal
2. Perform g two-sample t tests (g is the number of all possible pair-wise comparisons):
   (a) $H_{o,l} : \mu_i = \mu_j$ i, j are two of the k groups, and $l = 1, \ldots, g$.
   (b) $H_{a,l} : \mu_i \neq \mu_j$
   (c) The significance level is $(1-\alpha^*)\%$ (we'll discuss later what this level means)
   (d) The test statistic is
   $T = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{s^2_{p,ij}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}}$, where $s^2_{p,ij}$ is the pooled estimate of the variance within groups i and j given by $s^2_{p,ij} = \frac{(n_i-1)s_i^2+(n_j-1)s_j^2}{n_i+n_j-2}$.
   (e) **Pair-wise test rejection rule:** Reject $H_{o,l}$ if $T > t_{n_i+n_j-2;\alpha^*/2}$, or if $T < -t_{n_i+n_j-2;\alpha^*/2}$.
3. **Rejection rule (of the overall test):** Reject H_o if any of the g pair-wise tests rejects its null hypothesis $H_{o,l}$. Otherwise, do not reject H_o.

There are several issues of note:

i. Although $g = 3$ if $k = 3$, and thus it is relatively small, if $k = 10$, then $g = 45$ (check this!). Thus, g tends to increase rapidly with increasing k.

ii. If the level of significance of the overall test is $(1-\alpha)\%$ and that of each of the g sub-tests is also $(1-\alpha)\%$, then the level of significance of the overall test is lower than $(1-\alpha)\%$.

Example: Consider the case where $\alpha = 0.05$ (then the significance level is 95%) and $k = 3$ (then $g = 3$). Then if event A="The overall test correctly rejects H_o", and A_l="Pair-wise test l correctly rejects $H_{o,l}$", then $P(A) = P(A_1 \cap A_2 \cap \ldots \cap A_g) = P(A_1)P(A_2)\ldots P(A_g) = (1-\alpha)^g$ assuming independence among sub-tests. If $\alpha = 0.05$, $P(A) = (1-\alpha)^g = (0.95)^3 = 0.857 <$

0.95. Consequently, the probability of a Type-I error is $1 - 0.857 = 0.143$ instead of only 0.05 (under the assumption of independence). Even if the individual pair-wise tests are not independent however, the significance level of the overall test can be much smaller than anticipated. Thus, the two-sample t test is not totally satisfactory.

Pulmonary function example (continued):

($l = 1$) **Johns Hopkins versus Rancho Los Amigos**

```
. ttest FEV1 if center==1 | center==2, by(center)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
Johns Ho |      21     2.62619    .1082732    .4961701    2.400337    2.852044
Rancho L |      16      3.0325      .13081    .5232399    2.753685    3.311315
---------+--------------------------------------------------------------------
combined |      37    2.801892    .0889105    .5408216    2.621573    2.982211
---------+--------------------------------------------------------------------
    diff |           -.4063096    .1685585               -.7485014   -.0641177
------------------------------------------------------------------------------
Degrees of freedom: 35
               Ho: mean(Johns Ho) - mean(Rancho L) = diff = 0

     Ha: diff < 0                 Ha: diff ~= 0                 Ha: diff > 0
       t =  -2.4105                 t =  -2.4105                 t =  -2.4105
   P < t =   0.0107           P > |t| =   0.0213           P > t = 0.9893
```

($l = 2$) **Johns Hopkins versus St. Louis**

```
. ttest FEV1 if center==1 | center==3, by(center)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
Johns Ho |      21     2.62619    .1082732    .4961701    2.400337    2.852044
St Louis |      23    2.878696    .1037809    .4977157    2.663467    3.093924
---------+--------------------------------------------------------------------
combined |      44    2.758182    .0765034    .5074664    2.603898    2.912466
---------+--------------------------------------------------------------------
    diff |           -.2525052    .1500002               -.5552179    .0502075
------------------------------------------------------------------------------
Degrees of freedom: 42
               Ho: mean(Johns Ho) - mean(St Louis) = diff = 0

     Ha: diff < 0                 Ha: diff ~= 0                 Ha: diff > 0
       t =  -1.6834                 t =  -1.6834                 t =  -1.6834
   P < t =   0.0499           P > |t| =   0.0997           P > t =   0.9501
```

($l = 3$) **Ranch Los Amigos versus St. Louis**

```
. ttest FEV1 if center==2 | center==3, by(center)

Two-sample t test with equal variances

------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
Rancho L |      16      3.0325      .13081    .5232399    2.753685    3.311315
St Louis |      23    2.878696    .1037809    .4977157    2.663467    3.093924
---------+--------------------------------------------------------------------
combined |      39    2.941795    .0812345    .5073091    2.777344    3.106245
---------+--------------------------------------------------------------------
    diff |            .1538044    .1654468               -.1814227    .4890314
------------------------------------------------------------------------------
Degrees of freedom: 37
               Ho: mean(Rancho L) - mean(St Louis) = diff = 0

     Ha: diff < 0                 Ha: diff ~= 0                 Ha: diff > 0
       t =   0.9296                  t =   0.9296                  t =   0.9296
   P < t =   0.8207          P > |t| =   0.3586          P > t =   0.1793
```

Conclusions

1. We consider the two-sided alternative in each case, i.e., $H_{a,l:\mu_i \neq \mu_j}$, $l = 1, 2, 3$.

2. The only difference that was found to be significant at an (unadjusted) 5% alpha level, was the comparison of Rancho Los Amigos and Johns Hopkins (two-sided p value=0.0213).

3. Since $g = 3$ three possible pair-wise comparisons were possible. Each pair-wise test p value should be less than $\alpha^* = \alpha/3 \approx 0.017$ for the mean difference to be statistically significant if the overall test were carried out with a 5% α-level. In the case where the overall test were carried out at the 10% α-level, then the p value of the pairwise comparison would have to be less than $\alpha^* = \alpha/3 \approx 0.033$.

4. Given that the pair-wise comparison between Rancho Los Amigos and Johns Hopkins had a p value of 0.0213, given the Bonferroni adjustment, the comparison would be statistically significant only if the overall test had been carried out at the 90% level of significance (as $0.0213 < 0.033$) but not at the 5% level of significance (as $0.0213 > 0.017$).

9.2 Analysis of Variance

A general procedure that directly addresses the above null hypothesis is called Analysis of Variance (ANOVA). This technique focuses in the analysis of the sources of variation among

the observations.

The following assumptions accompany the Analysis of Variance:

a. The k groups are independent from each other

b. The k group variances are equal, i.e., $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$

c. The populations are approximately normal

9.2.1 Sources of Variation

The total variability in Y is

$$S_y = \frac{1}{N-1} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_{..}\right)^2$$

where $N = \sum_{i=1}^{k} n_i$, $\bar{Y}_{i.} = \frac{1}{n_i} \sum j = 1 Y_{ij}$ and $\bar{Y}_{..} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}$ is the total sample size, the mean of each group i, and the overall mean of all the observations. This variability can be divided into two parts as follows:

$$\begin{aligned} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_{..}\right)^2 = & \quad \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left[\left(Y_{ij} - \bar{Y}_{i.}\right) + \left(Y_{i.} - \bar{Y}_{..}\right)\right]^2 \\ = & \underbrace{\sum_{i-1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_{i.}\right)^2}_{\substack{\text{Variability due} \\ \text{to} \\ \text{differences} \\ \textit{within} \text{ groups}}} + \underbrace{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{i.} - \bar{Y}_{..}\right)^2}_{\substack{\text{Variability due} \\ \text{to} \\ \text{differences} \\ \textit{between} \text{ groups}}} = \quad SS_w + SS_b \end{aligned}$$

It can be shown that these are two sources of variation can be written as follows:

a. Variability *within* each of the k groups, which to a certain extent is inherent. This is estimated by

$$\boxed{s_w^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2 + \ldots + (n_k - 1)\, s_k^2}{n_1 + n_2 + \ldots + n_k - k}}$$

where $s_i^2 = \frac{1}{(n_i - 1)} \sum_{j=1}^{n} \left(Y_{ij} - \bar{Y}_{i.}\right)^2$ is the sample variance in group i. Notice that this is a *pooled* estimate of the variance for k groups.

b. Variability between the groups is estimated by

$$s_b^2 = \frac{n_1\left(\bar{Y}_{1\cdot} - \bar{Y}_{\cdot\cdot}\right)^2 + n_2\left(\bar{Y}_{2\cdot} - \bar{Y}_{\cdot\cdot}\right)^2 + \ldots + n_k\left(\bar{Y}_{k\cdot} - \bar{Y}_{\cdot\cdot}\right)^2}{k-1}$$

It increases as individual groups have increasingly different means.

This is akin to the expression of the one-sample variance $s^2\frac{1}{(n-1)}\sum_{i=1}^{n}(x_i-\bar{x})^2$ with the group sample means playing the role of the individual observations. The Analysis of Variance procedure is based on the following simple idea: Under the null hypothesis of equality of the k group means, the division of the population into k groups is artificial (i.e., there is only one homogeneous group). Thus, any variability *between* or *within* the groups comes from the inherent variability of the individual observations in the population. Thus, both s_b^2 and s_w^2 are estimating the same quantity (i.e., the population variance) and should be close in value (this non-technical argument can be duplicated mathematically).

9.2.2 The F Test of Equality of k Means

Since s_b^2 and s_w^2 are close in value (under the hypothesis of equality of the k group means), their *ratio* should be close to 1.

On the other hand, consider what happens when the k group means are *not* equal. Then s_b^2 increases (as the squared deviations from the overall mean increase) and the ratio of the *between* to *within* variability ratio becomes significantly larger than 1. It can be shown that the ratio of s_b^2 over s_w^2 has an F distribution with $k-1$ (the number of groups minus 1) and $n-k$ degrees of freedom (the remaining degrees of freedom from $k-1$ to the total $n-1$). The criterion of what is a "large" (statistically significant) deviation from 1.0 is determined by comparing the ratio to the tail of $F_{k-1,n-k}$, i.e., an F distribution with $k-1$ *numerator* degrees of freedom associated with the *between* groups variability, and $n-k$ *denominator* degrees of freedom associated with the *within* group variability. Critical values of the F distribution can be found in Appendix A.5 of the textbook).

The test of hypothesis of equality of the k population means is constructed as follows:

1. $H_o : \mu_1 = \mu_2 = \ldots = \mu_k$
2. H_a : At least two means are not equal
3. Tests are carried out at the $(1-\alpha)\%$ level of significance
4. The test statistic is $F = \frac{s_b^2}{s_w^2} = \frac{MS_b}{MS_w}$
5. **Rejection rule:** Reject H_o, if $F > F_{k-1,n-k;\alpha}$

Table 9.1: The Analysis of Variance (ANOVA) Table

Source of variability	Sums of squares (SS)	Df	Mean squares (MS)	F
Between groups	$SS_b=(k-1)s_b^2$	$k-1$	$MS_b=s_b^2$	$F=\frac{s_b^2}{s_w^2}$
Within groups (error)	$SS_w=(n-k)s_w^2$	$n-k$	$MS_w=s_w^2$	
Total	$(k-1)s_b^2+(n-k)s_w^2$	$n-1$	$s_b^2+s_w^2$	

9.2.3 Computer implementation

```
. oneway FEV1 center, tabulate
           | Summary of Forced Expiratory Volume 1 sec
    Center |        Mean   Std. Dev.       Freq.
-----------+------------------------------------
   Johns Ho |   2.6261905   .49617009          21
   Rancho L |      3.0325    .5232399          16
   St. Loui |   2.8786957   .49771572          23
-----------+------------------------------------
     Total |   2.8313333   .52178139          60
                      Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      1.58283745      2   .791418723      3.12     0.0520
 Within groups      14.4802556     57   .254039573
------------------------------------------------------------------------
    Total           16.0630931     59   .272255815
Bartlett's test for equal variances:  chi2(2) =   0.0583  Prob>chi2 = 0.971
```

Several pieces of information can be extracted from this table:

1. $s_b^2 = .791418723$, $s_w^2 = .254039573$. The between groups variability contains $k-1=2$ degrees of freedom, while the within groups variability contains $n-k=57$ degrees of freedom.

2. The ratio of the between- to within-groups variability is $F = 3.12$. This value is between $2.79 = F_{2,57;0.10} < F < F_{2,57;0.05} = 4.00$. Thus we would reject the null hypothesis at the 90% level, but would not reject it at the 95% level. There seems to be a slight difference (at least at the 90% level of significance) between patients in the three clinical centers.

3. STATA lists the Bartlett's test for equality of the individual population variances. This is a chi-square test, with the usual rejection rule (i.e. reject the hypothesis of equality

of variances if the p-value listed is lower than a pre-specified α-level). From the output above we are reassured that the hypothesis of equal group variances holds.

9.2.4 Remarks

1. The ANOVA procedure is a very powerful and direct way to compare an arbitrary number of k groups for any value of k. If $k = 2$, then the F-test is equivalent to a two-sample t test. In fact, the value of the F statistic in this case is equal to the square of the two-sample T statistic.

2. The MS_w (within Mean Squares) should be preferred in general as an estimate of the population variance to MS_b, because the latter tends to be inflated when the null hypothesis is rejected (i.e., apart from variability inherent in the individual observations in the population, it contains variability attributable to the differences between the groups. MS_w however, should still be preferred even when the F test does not reject the null hypothesis.

9.3 *Post-hoc* Tests

Even though the analysis of variance technique addresses the question of equality of the group means, there are many instances where we would like to know which subgroups are different. Recall that it suffices for only two groups to be different for the F test to reject the null hypothesis. Normally these tests will be carried out *after* the null hypothesis has been rejected. This is why they are called *post-hoc* tests.

Most of these tests perform all pair-wise comparisons between the k group means, adjusting the p value for the fact that multiple comparisons are carried out.

Note! It is not meaningful to carry out such tests if the overall hypothesis has not been rejected. In some borderline cases, one or more pair-wise comparison may appear statistically significant, even though the overall hypothesis has not been rejected by the F test.

9.3.1 The Bonferroni test

The Bonferroni *post-hoc* test proceeds as follows:

1. Carry out all pair-wise tests l such that $H_{o,l} : \mu_i = \mu_j$, i, j are two of the k groups, and $l = 1, \ldots, g$.

2. $H_{a,l} : \mu_i \neq \mu_j$

3. The significance level is $(1 - \alpha^*)\%$, where $\alpha^* = \alpha/g$ approximately.

4. The test statistic is $T = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{s_w^2\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}$ where s_w^2 is the variability *within* groups taken from the ANOVA table.

Note! We take advantage of information derived from all k groups in calculating the pooled estimate of the variance.

5. **Rejection rule:** Reject $H_{o,l}$ if $T > t_{n-k;\alpha^*/2}$ or if $T < -t_{n-k;\alpha^*/2}$ (notice the degrees of freedom).

The option bon in the oneway command of STATA produces the following output:

```
    Comparison of Forced Expiratory Volume 1 second by Clinical center
                                (Bonferroni)

Row Mean-|
Col Mean |   Johns Ho   Rancho L
---------|----------------------
Rancho L |     .40631
         |      0.055
         |
St. Loui |    .252505   -.153804
         |      0.307      1.000
```

The $g = 3$ possible pair-wise comparisons between the 3 clinical centers are listed in the STATA output. The first entry is the difference between the two group sample means. For example the difference between the mean pulmonary function among patients at Rancho Los Amigos ("column mean") and Johns Hopkins ("row mean") is $\bar{x}_1 - \bar{x}_2 = 3.0325 - 2.6261905 = 0.40631$. After adjusting for multiple comparisons, the p-value of the test (second entry) is 0.055. Thus, we would reject the null hypothesis that patients at Johns Hopkins and Rancho Los Amigos have the same pulmonary function levels as measured by FEV_1, at the 0.10 level but not at the 0.05 level (since the p-value of the test is $0.05 < 0.055 < 0.10$).

Note that STATA simplifies carrying out the Bonferroni multiple test procedure by printing out an **adjusted** p-value. This means that you should compare it to the α-level of the pair-wise test and not to $\alpha^*\alpha/g$. In that regard, STATA makes it unnecessary to have think in terms of α^*, and we can thus consistently carry out all tests at the usual level of significance.

Chapter 10

Correlation

Consider the diphtheria, pertussis, and tetanus (DPT) immunization rates, presented on page 398 of your textbook. Now consider the following question:

> Is there any *association* between the proportion of newborns immunized and the level of infant mortality?

Notice the inadequacy of chi-square-based tests in order to address this question. The data are continuous and even in a small data set as the one considered here, the problem is beyond the scope of any test for categorical data (as continuous data have too many "categories" for such test to be appropriate).

Example: DPT Immunization and Infant Mortality Consider the following two-way scatter plot of the under-5 mortality rate on the y axis and the DPT levels (percent of the population immunized) on the x axis (under five mortality rate data set).
By simple inspection of the graph it is clear that as the proportion of infants immunized against DPT *increases*, the infant mortality rate *decreases*.

Now consider:

X: Percent of infants immunized against DPT

Y: Infant mortality (number of infants under 5 dying per 1,000 live births)

A measure of this association is the *Pearson correlation coefficient* ρ, the average of the product of the standardized (normalized) deviates from the mean of each population. It is estimated by

$$\begin{aligned} r &= \frac{1}{(n-1)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) \\ &= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^{n} (x_i - \bar{x})^2 \right] \left[\sum_{i=1}^{n} (y_i - \bar{y})^2 \right]}} \end{aligned}$$

where s_x and s_y are the standard deviations of x and y respectively.

```
. label var under5 Mortality rate per 1000 live births
. label var immunize Percent immunized
. graph under5 immunize, xlab ylab
```

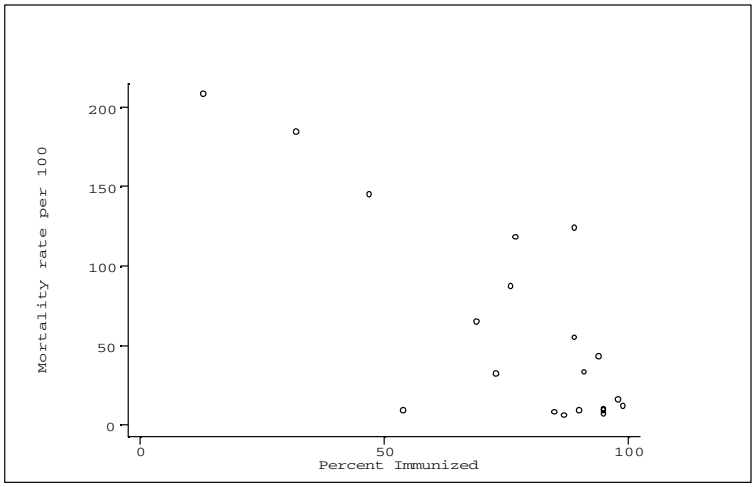


Figure 10.1: Scatter plot of DPT immunization and under-5 mortality rate

10.1 Characteristics of the Correlation Coefficient

The correlation coefficient can take values from -1 to +1. **Positive** values of ρ (close to +1) imply a *proportional* relationship between x and y. **Negative** values of ρ (close to -1) imply an *inversely proportional* relationship between x and y.

If $|\rho|$ is close to 1, this implies a *functional* (perfect) relationship between x and y, meaning that if we know one of them, it is like knowing the other exactly. Independent variables are **uncorrelated**.

Note!. The correlation coefficient is not a measure of whether the relationship between X and Y is linear.

Consider Figures 2 and 4. The correlation coefficient is zero in the former case (Figure 2) and greater than zero in the latter case (Figure 4). However, in neither case is the relationship between X and Y linear. Considering the data from table 17.1 (DPT data set), we have the following:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{20}\sum_{i=1}^{20} x_i = 77.4\%, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{20}\sum_{i=1}^{20} y_i = 59.0 \text{ per 1,000 live births}$$

$$\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} (x_i - 77.4)(y_i - 59.0) = -22706$$

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{20} (x_i - 77.4)^2 = 10630.8 \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{20} (y_i - 59.0)^2 = 77498$$

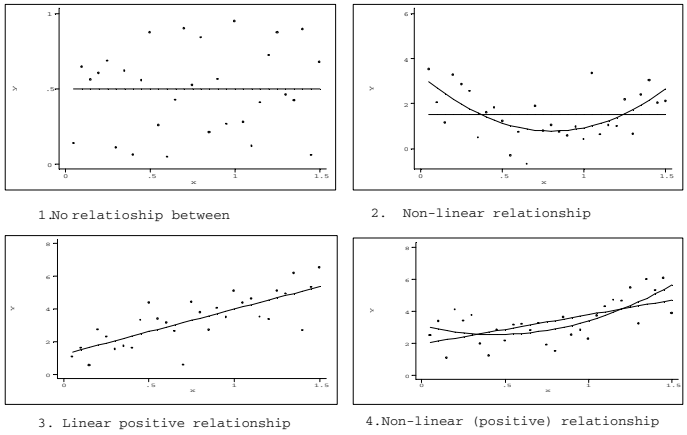


Figure 10.2: Examples of relationships between two measures

The correlation coefficient is

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{-22706}{\sqrt{(10630.8)(77498)}} = -0.79$$

This implies that there is a fairly substantial negative association between immunization levels for DPT and infant mortality.

10.2 Hypothesis Testing for $\rho = 0$

The test of the hypothesis of zero correlation is constructed as follows:

1. $H_o : \rho = 0$
2. (a) $H_a : \rho > 0$
   (b) $H_a : \rho < 0$
   (c) $H_a : \rho \neq 0$
3. The level of significance is $(1 - \alpha)\%$
4. The test is based on the statistic $T = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = r\sqrt{\frac{n-2}{1-r^2}} \sim t_{n-2}$
5. **Rejection rule**:
   (a) Reject H_o if $t > t_{n-2;1-\alpha}$
   (b) Reject H_o if $t < t_{n-2;\alpha}$
   (c) Reject H_o if $t > t_{n-2;1-\alpha/2}$ or $t < t_{n-2;\alpha/2}$

In the previous example, if α is 5% (significance level 95%), since

$$t = r\sqrt{\frac{n-2}{1-r^2}} = -0.79\sqrt{\frac{20-2}{1-(-0.79)^2}} = -5.47$$

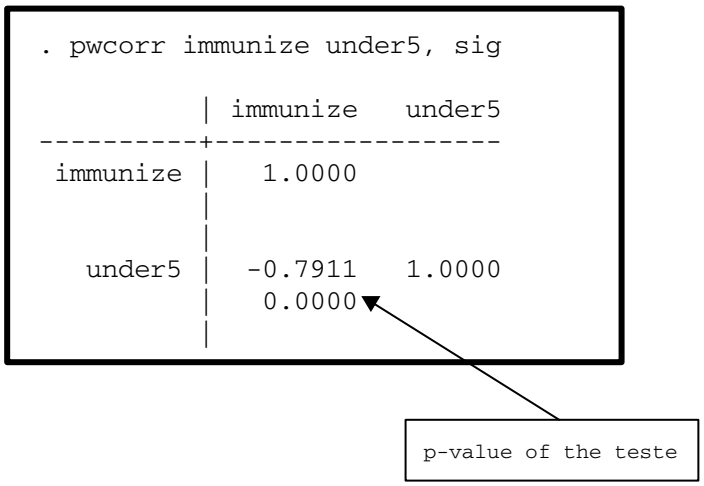
Since $-5.47 << t_{18;0.025}$ we reject the null hypothesis at the 95% level of significance. There is a statistically significant *negative* correlation between immunization levels and infant mortality. This means that as immunization levels rise, infant mortality decreases.

Note! We cannot estimate how much infant mortality would decrease if a country were to increase its immunization levels by say 10%.

10.2.1 Computer Implementation

We use the STATA command pwcorr with the option sig to obtain the Pearson correlation coefficient and carry out the test of hypothesis that was described above. The output is as follows:

Figure 10.3: Scatter plot of DPT immunization and under-5 mortality rate



Chapter 11

Simple Linear Regression

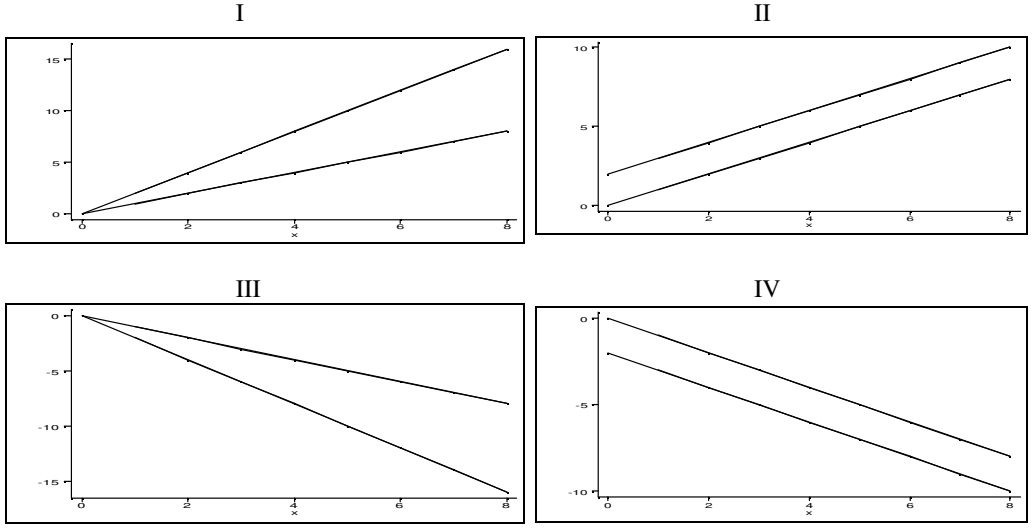
We all know what a straight line is. Along with the simple way of drawing a line (e.g., by using a ruler), there is a mathematical way to draw a line. This involves specifying the relationship between two **coordinates** x (measured on the horizontal or x axis) and y (measured on the vertical or y axis). We can use a line in attempting to describe (model) the relationship between x_i and y_i.

The equation relating the x_i to the y_i is $y = \alpha + \beta x + \epsilon$. The linear part of the relationship between x and y is:

$$\boxed{\mu_{y|x} = \alpha + \beta x}$$

α is called the **intercept** of the line (because if $x_i = 0$ the line "intercepts" the y axis at α), and β is called the **slope** of the line. The additional term ϵ, is an error term that accounts for random variability from what is expected from a linear relationship.

Figure 11.1: Examples of linear relationships



I. Both lines have the same intercept.

II. Both lines have the same slope (they are **parallel**) but different intercept.

III. Both lines have the same intercept but different **negative** slopes.

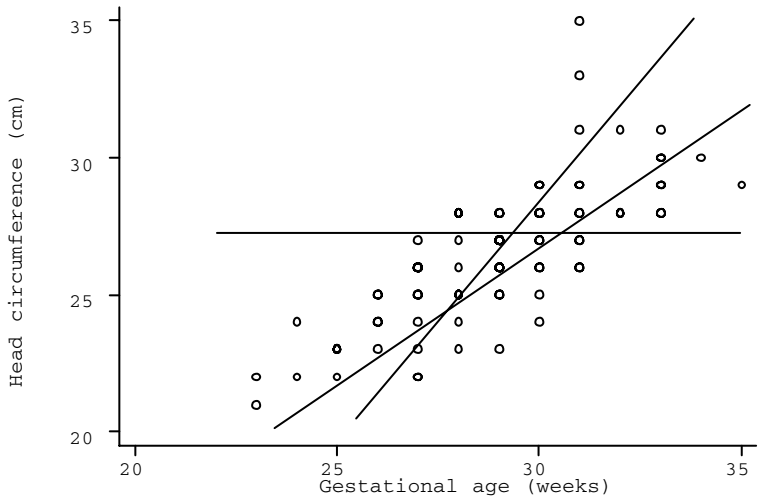
IV. Both lines have the same (negative) slope but different intercepts.

The appeal of a linear relationship is the *constant slope*. This means that for a fixed increase Δx in x, there will be a fixed change $\Delta y (= \beta \Delta x)$. This is going to be a fixed *increase* if the slope is positive, or a fixed *decrease* if the slope is negative, regardless of the value of x. This is in contrast to a *non-linear* relationship, such a *quadratic* or *polynomial*, where for some values of x, y will be increasing, and for some other values y will be decreasing (or vice versa).

Example: Head circumference and gestational age

Consider the problem described in section 18.4 of the textbook. The question that is asked is whether there is a relationship between the head circumference (in inches) and the gestational age (in weeks) of newborn infants. A scatter plot of the data is as follows: Even though it

Figure 11.2: Possible linear relationships between gestational age and head circumference



seems that head circumference is increasing with increasing gestational age, the relationship is not a perfect line. If we want to draw a line through the plotted observations that we think best describes the trends in our data we may be confronted with many candidate lines.

11.1 Determining the Best Regression Line

The regression line (whatever it is) will not pass through all data points Y_i. Thus, in most cases, for each point X_i the line will produce an estimated point $\widehat{Y}_i = \alpha + \beta x_i$ and most probably, $\widehat{Y}_i \neq Y_i$. For each choice of α and β (note that each pair α and β completely defines the line) we get a new line. The "best-fitting line" according to the least-squares method is the one that *minimizes* the sum of square deviations $\sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i\right)^2 = \sum_{i=1}^{n} \left[Y_i - \left(\widehat{\alpha} + \widehat{\beta} x_i\right)\right]^2$.

11.1.1 The least-squares line

The *least-squares estimates* of α and β that determine the least-squares line are

$$\boxed{\widehat{\beta} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \qquad \boxed{\widehat{\alpha} = \bar{y} - \widehat{\beta}\bar{x}}$$

11.1.2 Explaining Variability

The total variability in the data is given by

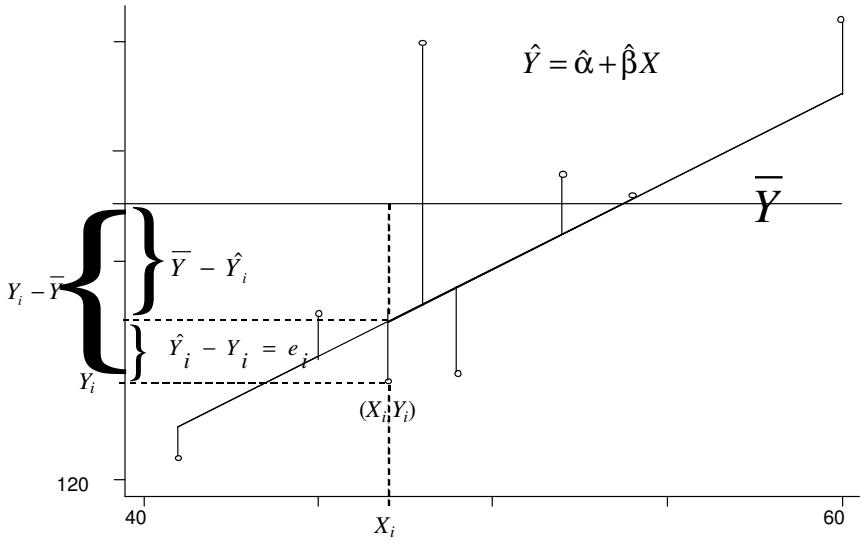
$$\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 = \sum_{i=1}^{n}\left[\left(Y_i - \widehat{Y}_i\right) + \left(\widehat{Y}_i - \bar{Y}\right)\right]^2$$

It turns out that

$$\boxed{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 = \underbrace{\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2}_{\substack{\text{unexplained}\\ \text{variability}}} + \underbrace{\sum_{i=1}^{n}\left(\widehat{Y}_i - \bar{Y}\right)^2}_{\substack{\text{explained}\\ \text{variability}}}}$$

as we can see by inspection of the following Figure.

Figure 11.3: Explaining variability with regression



There are two parts to the total variability in the data. One part, $SSR = \sum_{i=1}^{n}\left(\widehat{Y}_i - \bar{Y}\right)^2$, that is explained by the linear association of x and y, and the other, $SSE = \sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2$, that is left unexplained, because the regression model cannot further explain why there are still distances between the estimated points and the data (this is called *error sum of squares*).

11.1.3 Degrees of Freedom

The total variability in the data is given by $\underbrace{\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2}_{SSY}=\underbrace{\sum_{i=1}^{n}\left(Y_i-\widehat{Y}_i\right)^2}_{SSE}+\underbrace{\sum_{i=1}^{n}\left(\widehat{Y}_i-\bar{Y}\right)^2}_{SSR}$.

1. The total sum of squares $SSY=\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2$ is made up of n terms of the form $\left(Y_i-\bar{Y}\right)^2$. Once however the mean $\bar{Y}$ has been estimated, only $n-1$ terms are needed to compute SSY. The n^{th} term is known since $SSY=\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)=0$, for all $Y_i, i=1,\ldots,n-1$. Thus, the *degrees of freedom*[1] of SSY are $n-1$.

2. On the other hand, it can be shown that the sum of squares due to regression, SSR, is computed from a single function involving β and has thus only one degree of freedom associated with it (which is "expended" in the estimation of β).

3. SSE has the remaining $n-2$ degrees of freedom.

11.1.4 Assumptions of the linear regression model

1. The y values are distributed according to a normal distribution with mean $\mu_{y|x}$ and variance $\sigma^2_{y|x}$ which is unknown.

2. The relationship between x and y is described by the formula $\mu_{y|x}=\alpha+\beta x$.

3. The y are *independent*.

4. For *every* value x the standard deviation of the outcomes y is constant (and equal to $\sigma^2_{y|x}$). This concept is called *homoscedacity*.

11.2 Inference in Regression

In these models, y is our target (or *dependent* variable, the outcome that we cannot control but want to explain) and x is the explanatory (or *independent* variable, a factor that is, to a certain extend within our control). In our example, head circumference is our dependent variable, while gestational age is the independent variable.

Within each regression the primary interest is the assessment of the existence of the linear relationship between x and y. If such an association exists, then x provides information about y.

Inference on the existence of the linear association is accomplished via tests of hypotheses, and confidence intervals. Both of these center around the estimate of the slope $\widehat{\beta}$, since it is clear, that if the slope is zero, then changing x will have no impact on y (thus there is no association between x and y).

[1] You can think of the degrees of freedom as *unique* pieces of information

Source of variability	Sums of squares (SS)	df	Mean squares (MS)	F	Reject H_o if
Regression	SSR	1	$MSR = \frac{SSR}{1}$	$F = \frac{MSR}{MSE}$	$F > F_{1,n-2;1-\alpha}$
Residual (error)	SSE	$n-2$	$MSE = \frac{SSE}{n-2}$		
Total	SSY	$n-1$			

11.2.1 The F test of overall linear association

The test of hypothesis of no linear association is defined in a similar manner as in the ANOVA:

1. H_o: No linear association between x and y.
2. H_a: Linear association exists between x and y.
3. The test is carried out at the $(1-\alpha)\%$ level of significance.
4. The test statistic is $F = \frac{MSR}{MSE}$, where the numerator is part of the variability that can be explained through the regression model and the denominator is the unaccounted for variability or error.
5. **Rejection rule**: Reject H_o, if $F > F_{1,n-2;1-\alpha}$. This will happen if F is far from unity (just like in the ANOVA case).

The F test of linear association tests whether a line (other than the horizontal one going through the sample mean of the Y's) is useful in explaining some of the variability of the data. The test is based on the observation that, under the null hypothesis, $MSR \approx \sigma^2_{y|x}$, $MSE \approx \sigma^2_{y|x}$. If the population regression slope $\beta \approx 0$, that is, if the regression does not add anything new to our understanding of the data (i.e., does not explain a substantial part of the total variability), the two mean square errors MSR and MSE are estimating a common quantity (the population variance $\sigma^2_{y|x}$). Thus the ratio should be close to 1 if the hypothesis of no linear association between X and Y is present. On the other hand, if a linear relationship exists, (β is far from zero) then $SSR > SSE$ and the ratio will deviate significantly from 1.

11.2.2 Hypothesis testing for zero slope

The test of hypothesis of no linear association is defined as follows:

1. H_o: No linear association between x and y: $\beta = 0$.
2. H_a: A linear association exists between x and y:

   (a) $\beta \neq 0$ (two-sided test)
   (b) $\beta > 0$ } (one-sided tests)
   (c) $\beta < 0$ }

3. Tests are carried out at the $(1-\alpha)\%$ level of significance

4. The test statistic is $T = \frac{\widehat{\beta}}{\text{s.e.}(\widehat{\beta})}$ distributed as a t distribution with $n-2$ degrees of freedom

5. **Rejection rule**: Reject H_o, in favor of the three alternatives respectively, if

   (a) $t < t_{n-2;\alpha/2}$, or $t > t_{n-2;(1-\alpha/2)}$

   (b) $t > t_{n-2;(1-\alpha)}$

   (c) $t < t_{n-2;\alpha}$

11.2.3 Confidence Intervals for α and β

Confidence intervals of β_1 are constructed as usual, and are based on the standard error of $\widehat{\beta}$, the estimator, and the t statistic discussed above.

A $(1-\alpha)\%$ confidence interval is as follows:

$$\boxed{\left[\widehat{\beta} - t_{n-2;(1-\alpha/2)}\text{s.e.}\left(\widehat{\beta}\right),\ \widehat{\beta} + t_{n-2;(1-\alpha/2)}\text{s.e.}\left(\widehat{\beta}\right)\right]}$$

with $\text{s.e.}\left(\widehat{\beta}\right) = \frac{s_{y|x}}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$, where $s_{y|x}$ is the estimate of $\sigma_{y|x}$ and $s_{y|x} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i-\widehat{y}_i\right)^2}{n-2}} = \sqrt{MSE}$. One-sided confidence intervals are constructed in a similar manner.

In some occasions, tests involving the intercept are carried out. Both hypothesis tests and confidence intervals are based on the variance $\text{s.e.}(\widehat{\alpha}) = s_{y|x}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$. The statistic $T = \frac{\widehat{\alpha}}{\text{s.e.}(\widehat{\alpha})} \sim t_{n-2}$.

A $(1-\alpha)\%$ two-sided confidence interval is as follows:

$$\boxed{\left[\widehat{\alpha} - t_{n-2;(1-\alpha/2)}\text{s.e.}\left(\widehat{\alpha}\right),\ \widehat{\alpha} + t_{n-2;(1-\alpha/2)}\text{s.e.}\left(\widehat{\alpha}\right)\right]}$$

One-sided confidence intervals are constructed in a similar manner.

11.2.4 Computer Implementation

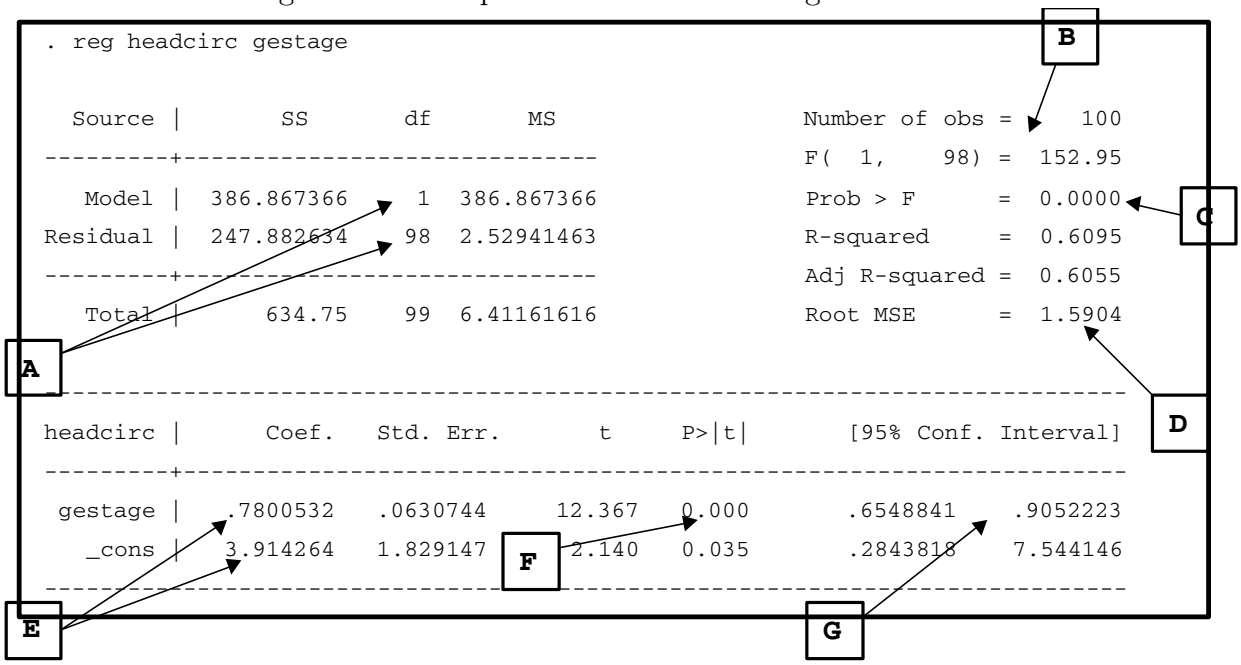
We use the *low birth weight infants* data

The following parts of the computer output are of interest:

A. **Degrees of freedom.** There is one degree of freedom associated with the model, and $n-2 = 98$ degrees of freedom comprising the residual.

B. **F test.** This is the *overall* test of the hypothesis of no linear association. Note that the numerator degrees of freedom for this test are the model degrees of freedom, while the denominator degrees of freedom for this test are the residual (error) degrees of freedom. This test is identical to the F test in the multi-mean comparison case, i.e., it measures deviations from unity.

Figure 11.4: Output of the low birth weight data



C. **Rejection rule of the F test.** Since the p-value is $0.000 << 0.05 = \alpha$, we will *reject* the null hypothesis. There appears to be strong evidence against for a linear association between gestational age and head circumference of the newborn.

D. **Root MSe.** This is the square root of the mean square error, and as mentioned before, can be used as an estimate of $\sigma_{y|x} = 1.5904$.

E. gestage is the estimate of the slope, $\widehat{\beta} = 0.7800532$. This means that for each additional week of gestational age, the head circumference increases by an average 0.78 inches. _cons is the estimate of the intercept $\widehat{\alpha} = 3.914264$. It means that at gestational age zero, the head circumference is approximately 3.91 inches (this is of course not true). Note that the fact that the model fails at gestage=0 does not mean that it is not useful, or that the linear association is not valid. Normally we would be interested in a finite range of values within which the linear relationship would be both useful and valid. This is one of these cases.

F. **p-value** of the t test described above. Since $0.000 << 0.05$, we must reject the null hypothesis. There is strong evidence of a *proportional* (positive) linear relationship between gestational age, and head circumference of a newborn. Notice also that the value of the t statistic 12.367 squared is equal to 154.95, the value of the F statistic. The F test of overall linear association and the t test of zero slope are *equivalent* in simple linear regression. This is not the case in multiple regression.

G. **The confidence interval for β.** The 95% confidence interval is the default. In the previous example, the 95% confidence interval for the population slope is $[0.6548841, 0.9052223]$. Since this interval excludes 0, we must reject the null hypothesis, and conclude that

there is a strong *positive* linear relationship between gestational age and head circumference of a newborn.