# Government 1000 Lecture Notes[*]

Jasjeet S. Sekhon

Center for Basic Research in the Social Sciences
Department of Government
Harvard University

http://jsekhon.fas.harvard.edu/
jasjeet_sekhon@harvard.edu

# Introduction

We are concerned here with statistical reasoning and measurement. In particular, we shall study the most used non-trivial statistical technique in the world: least squares. It is used in most every scientific field imaginable, from Astronomy to Zoology.

The intellectual history of least squares is a glorious one. Many of the greatest minds of the 18th and 19th centuries contributed to its creation: De Moivre, several Bernoullis, Gauss, Laplace, Quetelet, Galton, Pearson, and Yule.

Many of these great minds studied the mathematics of simple games of chance—i.e., dice and card games. It is a mystery to some why these great minds chose to study such trivial diversions.

The mystery is cleared up when one realizes that all of these scholars thought of themselves to be involved in the discovery of a method of statistical calculus that would do for social studies what Leibniz's and Newton's calculus did for physics.

It quickly came apparent this is would be most difficult because of a variety of issues. One of these issues is the difficulty involved with making valid inferences when we cannot rely on two of the standard methods of imposing structure on our data:

1. Experimentation

2. Rigorous mathematical theories such as Newtonian physics or the two theories of relativity.

It was hoped that statistical inference through the use of multiple regression, and other such methods, is able to provide to the social scientist what experiments and rigorous mathematical theories provide, respectively, to the micro-biologist and astronomer.

In essence, we are after the ability to control for multiple factors so that the scientist may observe the causal (or true) relationship serious problems between $x$ and $y$.

Least-squares also helped solve the problem of how to combine observations.

# Mill’s Methods of Inductive Inference

John Stuart Mill (in his *A System of Logic*) devised a set of five methods (or canons) by means of which to analyze and interpret our observations for the purpose of drawing conclusions about the causal relationships they exhibit. These methods have been used by generations of social science researchers.

**Method of Agreement:** “If two or more instances of the phenomenon under investigation have only one circumstance in common, the circumstance in which alone all the instances agree is the cause (or effect) of the given phenomenon.”

**Method of Difference:** “If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance in common save one, that one occurring only in the former; the circumstance in which alone the two instances differ is the effect, or the cause, or an indispensable part of the cause, of the phenomenon.”

These methods have been used by a vast number of researchers, including such famous ones as Durkheim and Weber. All such work as serious problems.

Here are some examples:

- The Protestant Ethic
- Deficits and interest rates
- Gun control
- The list goes on, and on....

Mill himself realized many of these problems:

> "Nothing can be more ludicrous than the sort of parodies on experimental reasoning which one is accustomed to meet with, not in popular discussion only, but in grave treatises, when the affairs of nations are the theme. "How," it is asked, "can an institution be bad, when the country has prospered under it?" "How can such or such causes have contributed to the prosperity of one country, when another has prospered without them?" Whoever makes use of an argument of this kind, not intending to deceive, should be sent back to learn the elements of some one of the more easy physical sciences."

# What is the Solution?

There is no single solution. But somethings certainly help:

- good research design
- statistics

A key and tricky concept in statistical inference is conditional probability.

Let's look at an example Mill himself brought up:

> "In England, westerly winds blow during about twice as great a portion of the year as easterly. If, therefore, it rains only twice as often with a westerly as with an easterly wind, we have no reason to infer that any law of nature is concerned in the coincidence. If it rains more than twice as often, we may be sure that some law is concerned; either there is some cause in nature which, in this climate, tends to produce both rain and a westerly wind, or a westerly wind has itself some tendency to produce rain."

$$
\begin{aligned}
\mathsf{H}: &\mathsf{P}(\text{rain}|\text{westerly wind, } \Omega) > \\
&\mathsf{P}(\text{rain}|\textbf{not} \text{ westerly wind, } \Omega),
\end{aligned}
$$

where $\Omega$ is a set of background conditions we consider necessary for a valid comparison.

# But Conditional Probability is Tricky

This example was made famous by Monty Hall (Let's Make a Deal). Let us assume that Monty Hall presents to you three envelopes. One of the envelopes contains a $100 bill the other two are empty. Monty tells you that he put the money into an envelope by random (using a discrete uniform distribution $\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$). You are asked to pick one envelope. You pick envelop $A$. Then, Monty tells you that he will open one of the other envelopes—one which does not contain any money. Monty opens envelope $C$. Monty then allows you the option of switching from the envelope you have chosen $(A)$ to the remaining unopened envelope $(B)$. Assume that Monty has been telling you the truth.

To be clear, let us assume the following:

1. Monty Hall would never open envelope you have chosen—i.e., $A$.
2. Monty would never open the envelope containing the money.
3. If the money is in envelope $A$, Monty will choose to open envelope $\mathrm{B}$ or $\mathrm{C}$ with equal probability.

Should you switch? Does it matter?

Answering questions like this is a lot easer if we know how to manipulate the formal rules of probability.

Probability also helps us make valid empirical inferences, both descriptive and causal.

# Basic Set Theory

Define an "experiment" as an expression in the world that produces our data. Let

$$\begin{aligned}\Omega &= \text{the Sample Space of all possible outcomes,}\\ \omega &= \text{any one of the possible outcomes which makeup } \Omega,\end{aligned}$$

Let $\in$ denote the "contained in" operator. Therefore, by definition, $\omega \in \Omega$.

Let $\{\}$ denote elements of a set. For example, if we are interesting in the outcomes of a six sided die $\Omega = \{1, 2, 3, 4, 5, 6\}$.

Let us define $\phi$ as the null set. It contains no outcomes. By definition, for any set $A$, $\phi \in A$, and $\phi \in \Omega$.

An event is defined as a subset of the sample space. For example, we may define:

$$A = \{1, 2, 3\}, \qquad (1)$$

$$B = \{4, 5, 6\}. \qquad (2)$$

It follows that $A \in \Omega$ and $B \in \Omega$, but $A \notin B$.

# Unions

If $A$ and $B$ are events, $A \cup B$ is the set containing all events in $A$ and all events in $B$.

For example, $A = \{1, 2\}$, $B = \{3, 2\}$, $A \cup B = \{1, 2, 3\}$.

The union operator has the cumulative and the associative property.

The cumulative property is used to obtain the following:

$$A \cup B = B \cup A \tag{3}$$
$$A \cup A = A \tag{4}$$
$$A \cup \phi = A \tag{5}$$
$$A \cup \Omega = \Omega. \tag{6}$$

It also follows that if $A \in B \; then \; A \cup B = B$.

The associative property is illustrated by the following example :
$A \cup B \cup C = (A \cup B) \cup C = A \cup (B \cup C)$.

# Intersections

$C = AB = A \cap B$. This implies that C contains only those events which occur *both* in $A$ and $B$.

Like the union operator, the intersection operator obeys the cumulative and the associative property.

The cumulative property is used to obtain the following:

$$A \cap B = B \cap A \tag{7}$$

$$A \cap A = A \tag{8}$$

$$A \cap \phi = \phi \tag{9}$$

$$A \cap \Omega = A. \tag{10}$$

It also follows that if $A \in B$, **then** $A \cap B = A$.

The associative property is illustrated by the following example:
$A \cap B \cap C = (A \cap B) \cap C = A \cap (B \cap C)$.

# Complements

$A^c$ or $\bar{A}$ denotes the set of all of the points in $\Omega$ not in $A$.

Properties:

$$(A^c)^c = A \tag{11}$$

$$\phi^c = \Omega \tag{12}$$

$$\Omega^c = \phi \tag{13}$$

$$A \cup A^c = \Omega \tag{14}$$

$$A \cap A^c = \phi \tag{15}$$

# Disjoint Events

Disjoint events are events which are mutually exclusive.

In other words, $A$ and $B$ are disjoint sets if they have no common sample point $(\omega)$.

$A$ and $B$ are disjoint if and only if (iff) $A \cap B = \phi$.

# Some Laws

## Distributive Laws

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C) \tag{16}$$

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C) \tag{17}$$

These are called the distributed laws because of the analogy with $a \times (b + c) = (a \times b) + (a \times c)$.

## De Morgan's Laws, named after Augustus De Morgan

$$(A \cup B)^c = A^c \cap B^c \tag{18}$$

$$(A \cap B)^c = A^c \cup B^c \tag{19}$$

# Basic Definitions of Probability

Probability is a function often denoted by $P(\cdot)$ or $\Pr(\cdot)$.

A probability space is defined by three axioms:

1. for any event A, $P(A) \geq 0$.

2. $P(\Omega) = 1$.

3. if $A$ and $B$ are disjoint sets, then $P(A \cup B) = P(A) + P(B)$.

There are several consequences from the forgoing:

1. $\mathrm{P}(\phi) = 0$.

2. $\mathrm{P}(A^c) = 1 - \mathrm{P}(A)$.
   **Proof:**
   By definition, $A$ and $A^c$ are disjoint and $A \cup A^c = \Omega$. It follows from axiom 3 that $\mathrm{P}(\Omega) = \mathrm{P}(A) + \mathrm{P}(A^c)$. Since $\mathrm{P}(\Omega) = 1$ (by axiom 2), we obtain $\mathrm{P}(A^c) = 1 - \mathrm{P}(A)$.

3. $0 \leq \mathrm{P}(A) \leq 1$. This follows from axioms 1, 2 and the fact that $A \cup A^c = \Omega$.

4. if $A \in \mathrm{B}$, $\mathrm{P}(A) \leq \mathrm{P}(\mathrm{B})$.

5. for any sets $A$ and $\mathrm{B}$, $\mathrm{P}(A \cup \mathrm{B}) = \mathrm{P}(A) + \mathrm{P}(\mathrm{B}) - \mathrm{P}(A \cap \mathrm{B})$.
   And $\mathrm{P}(A \cap \mathrm{B}) = \mathrm{P}(A) + \mathrm{P}(\mathrm{B}) - \mathrm{P}(A \cup \mathrm{B})$.

# Conditional Probability

Let $A$ be an event such that $P(A) > 0$. The conditional probability of an event $B$ given $A$ occurs, denoted $P(B|A)$, is

$$P(B|A) = \frac{P(A \cap B)}{P(A)}. \tag{20}$$

Observe that if $P(A) = 0$ the conditional probability given $A$ is undefined.

We often write:

$$P(A \cap B) = P(B|A)P(A). \tag{21}$$

It also follows that

$$P(A \cap B) = P(A|B)P(B). \tag{22}$$

The probability measure $P_A(B) = P(B|A)$ is called the conditional distribution given $A$.

It is easily seen that our definition of the conditional distribution given $A$ is sensible. Whatever prior information that led us to use $P$ is not disturbed. We simply reassign the probabilities to exclude outcomes not in $A$ since we know $A$ occurs. Thus the probability of $B$ occurring should now be proportional to $P(A \cap B)$.

Hence $P_A(B) = cP(A \cap B)$ for all events $B$, and as $P_A(A) = 1$ we see that $c = P(A)^{-1}$.

The relative frequency interpretation yields the same formula. Suppose we perform the experiment a large number of times, say $n$, and observe $A$ $N(A)$ times and $A \cap B$ $N(A \cap B)$ times. The proportion of times $B$ occurs in the $N(A)$ experiments when $A$ occurs is $\frac{N(A \cap B)}{N(A)}$ and this should be approximately $P_A(B)$. But this is:

$$\frac{N(A \cap B)/n}{N(A)/n} \approx \frac{P(A \cap B)}{P(A)} \tag{23}$$

# Bayes' Rule

Bayes Theorem is commonly ascribed to the Reverent Thomas Bayes (1701-1761) who left one hundred pounds in his will to Richard Price "now I suppose Preacher at Newington Green." Price discovered two unpublished essays among Bayes' papers which he forwarded to the Royal Society. This work made little impact, however, until it was independently discovered a few years later by the great French mathematician Laplace. English mathematicians then quickly rediscovered Bayes' work.

The following is called Bayes Rule:

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} && (24) \\ &= \frac{P(B|A)P(A)}{P(B)} && (25) \end{aligned}$$

# Simple Card Example

**Card Example:** A card is selected from a deck of cards and found to be a spade. What is the probability that it is a face card?

**Solution:** Let $A$ be the event that the selected card is a spade and $B$ the event the selected card is a face card.

We seek:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \tag{26}$$

$$= \frac{\frac{3}{52}}{\frac{13}{52}} \tag{27}$$

$$= \frac{3}{13} \tag{28}$$

# Monty Hall Problem: Solution

This example was made famous by Monty Hall (Let's Make a Deal). Let us assume that Monty Hall presents to you three envelopes. One of the envelopes contains a $100 bill the other two are empty. Monty tells you that he put the money into an envelope by random (using a discrete uniform distribution $\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$). You are asked to pick one envelope. You pick envelop $A$. Then, Monty tells you that he will open one of the other envelopes—one which does not contain any money. Monty opens envelope $C$. Monty then allows you the option of switching from the envelope you have chosen ($A$) to the remaining unopened envelope ($B$). Assume that Monty has been telling you the truth.

To be clear, let us assume the following:

1. Monty Hall would never open envelope you have chosen—i.e., $A$.

2. Monty would never open the envelope containing the money.

3. If the money is in envelope $A$, Monty will choose to open envelope $B$ or $C$ with equal probability.

Should you switch? Does it matter?

Let there be three envelopes: $\Omega \in \{A, B, C\}$. One of these envelopes contains money following a discrete uniform distribution. It follows that:

$$P(B = \$100 \cup C = \$100) = \frac{2}{3} \tag{29}$$

$$P(C = \$100 \cap C = \$0) = 0 \tag{30}$$

Without loss of generality, let us assume that envelope C was revealed to be the one which is empty. Should you switch? *Let* $C = \$0$ *denote the event that "envelope C will be revealed to be empty."*

We are interested in two conditional probabilities: $P(A = \$100|C = \$0)$ and $P(B = \$100|C = \$0)$. We want to know the probability that $A$ contains the money given that C was revealed to be empty and the probability that B contains the money given that C was revealed to be empty.

The value of both of these probabilities can be obtaining using Bayes Rule. But let us first determine $P(C = \$0)$ because we will require this value to apply Bayes' Rule.

$$\begin{aligned}
P(C = \$0) = {} & P(C = \$0|A = \$100)P(A = \$100) \qquad (31)\\
& + P(C = \$0|B = \$100)P(B = \$100)\\
& + P(C = \$0|C = \$100)P(C = \$100)\\
= {} & \frac{1}{2}\frac{1}{3} + 1\frac{1}{3} + 0\frac{1}{3} \qquad (32)\\
= {} & \frac{1}{6} + \frac{1}{3} + 0 \qquad (33)\\
= {} & \frac{1}{6} + \frac{2}{6} \qquad (34)\\
= {} & \frac{1}{2} \qquad (35)
\end{aligned}$$

Let us now calculate $P(A = \$100|C = \$0)$:

$$P(A = \$100|C = \$0) = \frac{P(A = \$100 \cap C = \$0)}{P(C = \$0)} \tag{36}$$

$$\text{By Bayes' Rule :}$$

$$= \frac{P(C = \$0|A = \$100)P(A = \$100)}{P(C = \$0)} \tag{37}$$

$$= \frac{\frac{1}{2}\frac{1}{3}}{\frac{1}{2}} \tag{38}$$

$$= \frac{1}{3} \tag{39}$$

Therefore, the probability that envelope $A$ contains the money remains unchanged.

Let us now calculate the probability that envelope $\mathrm{B}$ contains the money after envelope $\mathrm{C}$ has been revealed not to contain any money.

$$\mathrm{P}(\mathrm{B} = \$100 | \mathrm{C} = \$0) = \frac{\mathrm{P}(\mathrm{B} = \$100 \cap \mathrm{C} = \$0)}{\mathrm{P}(\mathrm{C} = \$0)} \tag{40}$$

$$\text{By Bayes' Rule}:$$

$$= \frac{\mathrm{P}(\mathrm{C} = \$0 | \mathrm{B} = \$100) \mathrm{P}(\mathrm{B} = \$100)}{\mathrm{P}(\mathrm{C} = \$0)} \tag{41}$$

$$= \frac{1\frac{1}{3}}{\frac{1}{2}} \tag{42}$$

$$= \frac{1}{3}\frac{2}{1} \tag{43}$$

$$= \frac{2}{3} \tag{44}$$

Therefore, the probability that envelope $A$, our original choice, contains the money is, as before, $\frac{1}{3}$ while the probability that envelope $\mathrm{B}$ now contains the money is $\frac{2}{3}$. Therefore, we should switch!

# Monty Hall, Version 2

What if we changed assumption 3 from:

> If the money is in envelope $A$, Monty will choose to open envelope $\mathrm{B}$ or $\mathrm{C}$ with equal probability.

To:

> If the money is in envelope $\mathrm{A}$, Monty will choose to open envelope $\mathrm{B}$ with probability $\frac{3}{4}$ and envelope $\mathrm{C}$ with probability $\frac{1}{4}$.

Let's work through the probabilities again. As before, *let* $\mathrm{C} = \$0$ *denote that event that "envelope C was revealed to be empty."* As before, let $A = \$100$ denote the event that envelope $A$ contains \$100. And, as before, let $\mathrm{B} = \$100$ denote the event that envelope $\mathrm{B}$ contains \$100.

We are interested in two conditional probabilities: $\mathrm{P}(A = \$100|\mathrm{C} = \$0)$ and $\mathrm{P}(\mathrm{B} = \$100|\mathrm{C} = \$0)$. We want to know the probability that $A$ contains the money given that $\mathrm{C}$ was revealed to be empty and the probability that $\mathrm{B}$ contains the money given that $\mathrm{C}$ was revealed to be empty.

As before, we know that we are going to need to use Bayes Rule. So let's calculate $P(C = \$0)$ first.

$$\begin{aligned}
P(C = \$0) &= P(C = \$0|A = \$100)P(A = \$100) \\
&\quad + P(C = \$0|B = \$100)P(B = \$100) \\
&\quad + P(C = \$0|C = \$100)P(C = \$100) && (45)\\
&= \frac{1}{4}\frac{1}{3} + 1\frac{1}{3} + 0\frac{1}{3} && (46)\\
&= \frac{1}{12} + \frac{1}{3} + 0 && (47)\\
&= \frac{1}{12} + \frac{4}{12} && (48)\\
&= \frac{5}{12} && (49)
\end{aligned}$$

$$P(A = \$100 | C = \$0) = \frac{P(A = \$100 \cap C = \$0)}{P(C = \$0)} \tag{50}$$

$$\text{By Bayes' Rule :}$$

$$= \frac{P(C = \$0 | A = \$100) P(A = \$100)}{P(C = \$0)} \tag{51}$$

$$= \frac{\frac{1}{4}\frac{1}{3}}{\frac{5}{12}} \tag{52}$$

$$= \frac{1}{5} \tag{53}$$

$$P(B = \$100 | C = \$0) = \frac{P(B = \$100 \cap C = \$0)}{P(C = \$0)} \tag{54}$$

$$\text{By Bayes' Rule :}$$

$$= \frac{P(C = \$0 | B = \$100) P(B = \$100)}{P(C = \$0)} \tag{55}$$

$$= \frac{1\frac{1}{3}}{\frac{5}{12}} \tag{56}$$

$$= \frac{1}{3}\frac{12}{5} \tag{57}$$

$$= \frac{4}{5} \tag{58}$$

You should clearly switch!

# Before Monty Opens Anything

Say you choose envelope $A$, but you do not *yet* know which envelope Monty will open. Before you know this information, what is your probability of winning if you switch versus staying with $A$? Obviously, you can only switch after Monty has opened an envelope. Recall that assumption assumption 3 has been changed from:

> If the money is in envelope $A$, Monty will choose to open envelope $\mathrm{B}$ or $\mathrm{C}$ with equal probability.

To:

> If the money is in envelope A, Monty will choose to open envelope $\mathrm{B}$ with probability $\frac{3}{4}$ and envelope C with probability $\frac{1}{4}$.

The two relevant probabilities are $\mathrm{P}(A = 1|\mathrm{B} = 0 \cup \mathrm{C} = 0)$ (i.e., the probability of $A$ containing the money given that either $\mathrm{B}$ or $\mathrm{C}$ will be revealed to be empty and $\mathrm{P}(switch = 1)$ (i.e., the probability of the switched to envelope containing the money given that either $\mathrm{B}$ or $\mathrm{C}$ have been revealed to be empty).

It is straightforward to obtain both of these.

$$P(A = 1|B = 0 \cup C = 0) = \frac{P\left(A = 1 \cap (B = 0 \cup C = 0)\right)}{P(B = 0 \cup C = 0)} \tag{59}$$

$$= \frac{P((B = 0 \cup C = 0)|A = 1)P(A = 1)}{P(B = 0 \cup C = 0)} \tag{60}$$

$$= \frac{1\frac{1}{3}}{1} \tag{61}$$

$$= \frac{1}{3} \tag{62}$$

$$P(switch = 1) = P(C = 1|B = 0)P(B = 0) + P(B = 1|C = 0)P(C = 0) \tag{63}$$

We have already calculated $P(B = 1|C = 0)$ (see Equations 54—58), but we do not yet know $P(C = 1|B = 0)$. In order to obtain this we must calculate $B = 0$.

$$\begin{aligned} P(B = 0) &= P(B = 0|A = 1)P(A = 1) + P(B = 0|B = 1)P(B = 1) \\ &\quad + P(B = 0|C = 1)P(C = 1) \\ &= \frac{3}{4}\frac{1}{3} + 0\frac{1}{3} + 1\frac{1}{3} \\ &= \frac{3}{12} + \frac{1}{3} \\ &= \frac{7}{12} \end{aligned} \tag{64–67}$$

We must still obtain $P(C = 1|B = 0)$.
pause

$$P(C = 1|B = 0) = \frac{P(C = 1 \cap B = 0)}{B = 0} \tag{68}$$

$$= \frac{P(B = 0|C = 1)P(C = 1)}{P(B = 0)} \tag{69}$$

$$= \frac{1\frac{1}{3}}{\frac{7}{12}} \tag{70}$$

$$= \frac{12}{21} \tag{71}$$

$$= \frac{4}{7} \tag{72}$$

We may now return to our story and obtain $P(switch = 1)$.

$$
\begin{aligned}
P(switch = 1) &= P(C = 1|B = 0)P(B = 0) + P(B = 1|C = 0)P(C = 0) && (73)\\
&= \frac{4}{7}\frac{7}{12} + \frac{4}{5}\frac{5}{12} && (74)\\
&= \frac{28}{84} + \frac{70}{60} && (75)\\
&= \frac{1}{3} + \frac{1}{3} && (76)\\
&= \frac{2}{3} && (77)
\end{aligned}
$$

We should clearly *switch*!

For completeness we shall calculate $P(A = 1|B = 0)$:

$$\begin{aligned}
P(A = 1|B = 0) &= \frac{P(A = 1 \cap B = 0)}{P(B = 0)} && (78)\\
&= \frac{P(B = 0|A = 1))P(A = 1)}{P(B = 0)} && (79)\\
&= \frac{\frac{3}{4}\frac{1}{3}}{\frac{7}{12}} && (80)\\
&= \frac{\frac{3}{12}}{\frac{7}{12}} && (81)\\
&= \frac{3}{7} && (82)
\end{aligned}$$

# The Independence Of Monty Hall

What if Monty didn't always open another envelope?

Say he only opened an envelope when you had originally picked the correct envelope?

Clearly, the above discussion would be wrong, but why?

The answer is the importance of independence.

# Exit Polls

In the California gubernatorial election in 1982, several TV stations predicted, on the basis of questioning people when they existed the polling place, that Tom Bradley, then mayor of Los Angeles, would win the election beating the only other candidate George Deukmejian.

When the voters were counted, however, he lost by a considerable margin. What happened?

To give our explanation we need some notation and some numbers.

Suppose we choose a person at random, let $B =$ "The person votes for Bradley" and suppose that $P(B) = 0.45$.

Since there were only two candidates, this makes the probability of voting for Deukmejian $P(B^c) = 0.55$.

Let $A =$ "The voter stops and answers a question about how she voted" and suppose that $P(A|B) = 0.4$, $P(A|B^c) = 0.3$. That is, 40% of Bradley voters will respond compared to 30% of the Deukmejian voters.

We are interested in computing $P(B|A)$: the fraction of voters in our sample that voted for Bradley.

By the definition of conditional probability,

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

To evaluate the numerator we use the multiplication rule,

$$P(B \cap A) = P(B)P(A|B) = 0.45 \times 0.4 = 0.18. \tag{83}$$

Similarly,

$$P(B^c \cap A) = P(B^c)P(A|B^c) = 0.55 \times 0.3 = 0.165 \tag{84}$$

Since $P(A) = P(B \cap A) + P(B^c \cap A)$,

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{0.18}{0.18 + 0.165} = 0.5217 \tag{85}$$

# Independence

Intuitively, two events $A$ and $B$ are independent if the occurrence of $A$ has no influence on the probability of occurrence of $B$.

The formal definition is: $A$ and $B$ are independent if $P(A \cap B) = P(A)P(B)$.

$A$ and $B$ are considered independent if the occurrence of one in no way depends on the occurrence of the other. In other words:

$$P(A|B) = P(A) \tag{86}$$

$$P(B|A) = P(B). \tag{87}$$

Bayes Rule is always true. That is, regardless of independence

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

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

Two classic examples of independent events. In each case it should be clear that the intuitive definition is satisfied, so we will only check the conditions of the formal one.

Example 1 Flip two coins. $A =$ "The first coin shows Heads," $B =$ "The second coin shows Heads." $P(A) = \frac{1}{2}$, $P(B) = \frac{1}{2}$, $P(A \cap B) = \frac{1}{4}$.

Example 2 Roll two dice. $A =$ "The first die shows 4," $B =$ "The second die shows 2." $P(A) = \frac{1}{6}$, $P(B) = \frac{1}{6}$, $P(A \cap B) = \frac{1}{36}$.

An example of events that are not independent:

Example 3 Roll two dice. $A =$ "The sum of the two dice is 9", $B =$ "The first die is 2." $A = (6,3),(5,4),(4,5),(3,6)$ so $P(A) = \frac{4}{36}$. $P(B) = \frac{1}{6}$, but $P(A \cap B) = 0$ since $(2,7)$ is impossible.

In general, if $A$ and $B$ are disjoint events that have positive probability, they are not independent since $P(A)P(B) > 0$ (by definition), but $P(A \cap B) = 0$ (because they are disjoint).

A finite sequence $A_1, A_2, \cdots, A_n$ or an infinite sequence $A_1, A_2, \cdots$ of events is said to be disjoint if $A_i \cap A_j = \phi$, for all $i \neq j$.

Recall that if $A$ and $B$ are disjoint,

$$P(A \cup B) = P(A) + P(B). \tag{88}$$

Also recall that, it is always true that:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \tag{89}$$

Independence is about the probability of the occurrence of events and disjoint is a description of the sample space. It may be useful to think of "independence" as a property of random variables and "disjoint" as a property of events.

Example: three events Let:

$A =$ "Alan and Barney have the same birthday"

$B =$ "Barney and Carl have the same birthday,"

$C =$ "Carl and Alan have the same birthday."

Let us assume that leap years do not exist and that all birthdays are equally likely. Since there are 365 ways two boys can have the same birthday out of $365^2$ possibilities, $P(A) = P(B) = P(C) = \frac{1}{365}$.

Likewise, there are 365 ways all three boys can have the same birthday out of $365^3$ possibilities, so

$$P(A \cap B) = \frac{1}{365^2} = P(A)P(B), \tag{90}$$

that is, $A$ and $B$ are independent. Similarity, $B$ and $C$, are independent and $C$ and $A$ are independent, so $A$, $B$, and $C$ are pairwise independent.

The three events $A$, $B$, and $C$ are not independent, however, since $A \cap B = A \cap B \cap C$ and hence

$$P(A \cap B \cap C) = \frac{1}{365^2} \neq \left(\frac{1}{365}\right)^3 = P(A)P(B)P(C) \tag{91}$$

The moral of this example is that in order to determine if several events are independent, you pair-wise comparisons are not enough.

# Independence and Experimental and Nonexperimental Research

In a successful randomized experiment, the treatment (T) is independent of all baseline variables (Z).

This implies that: $P(T|Z) = P(T)$

Note that Z need not be independent of the outcome of interest. If Y is the outcome of interest: $P(Y|Z) \neq P(Y)$.

In a non-randomized experiment, we cannot, other than in very rare circumstances, obtain this independence condition—i.e., in general $P(T|Z) \neq P(T)$.

For example: There is no association between the probability of an incumbent House candidate winning and the amount of money the candidate spends on the reelection bid. Why?

Answer: $P(\text{spending money}|Z) \neq P(\text{spending money})$, where Z is a long list of baseline variables such as House voting record, general quality of the candidate, the constituency service the candidate performs. The upshot is that there is very little agreement on the effect of money.

How to obtain conditional independence is a central concern of only this class but of all empirical research.

The most used “solution” in the social sciences is (ordinary least squares) regression. This is a particular way of modeling the conditional mean: $P(Y|X)$, where $X$ are some variables we wish to condition on.

In the campaign and money literature, these variables are usually measures of the quality of the candidates, their voting record, constituency characteristics and other such political factors.

A necessary (but not sufficient) condition of such an approach to work is that conditional independence holds: $P(Y|Z, X) = P(Y|X)$, where $X$ are observed variables and $Z$ are unobserved variables.

Thus, regression is a model of the conditional mean. There is a close (but all too imperfect) relationship between the often used regression model and the Neyman-Rubin-Holland Causal Model: both are based on conditional means.

Foreshadowing Note: When a regression model is not consistent with the Neyman-Rubin-Holland Causal Model, it cannot be interpreted as offering direct causal estimates. But it may still offer useful information as in the “Peasants or Bankers? The American Electorate and the U.S. Economy” article we will be discussing later in the course.

In order to understand both regression and the Neyman-Rubin-Holland Causal Model we need to understand not only conditional probability but also distributions.

For example: in the “Quality Meets Quantity” article we are interesting in comparing the following two conditional probabilities:

$$\mathrm{P}(\mathrm{revolution}|\mathrm{foreign\ threat},\ Z), \tag{92}$$

$$\mathrm{P}(\mathrm{revolution}|\mathrm{no\ foreign\ threat},\ Z), \tag{93}$$

where Z is the set of background conditions we consider necessary for valid comparisons (such as village autonomy and dominant classes who are economically independent). In the article Z is denoted by $\Omega$ but that may cause confusion given our definition of $\Omega$.

Recall that Equation 92 $= \frac{1}{8}$ and Equation 93 $= \frac{2}{69}$. Note We have rounded Equation 93 from the number reported in the article.

We wish to know if Equation 92 is SIGNIFICANTLY larger than 93. In other words, we need to rule out that Equation 92 is larger than Equation 93 just by chance.

When comparing any two probabilities (conditional or otherwise) we are interesting in the question of significance.

In order to answer this equation we need to make some claims about how revolutions and foreign threat are distributed. That is, we need to consider both revolutions and foreign threat to be random variables.

The use of many purely algebraic concepts such as the mean crucially depend on the distribution which is assumed. For example, the use of the mean did not become widespread in society until the normal distribution was discovered and until it became generally believed. This did not occur until the late 19th century. Without an implied distribution, the mean may be a completely uninformative concept. This will become clear in a lecture or two.

# Random Variables I

The concept of a random (or chance) variable is almost as old as probability theory itself, although a precise definition was not formulated until the 20th Century.

There are several different types of variables.
The most basic is discrete vs. continuous.

There are more fine grained distinctions.

1. Nominal The nominal scale is the least powerful. It only maps the attributes of the object into a name. This mapping is simply a classification of entities. The only relationship is whether the measure of two attributes are the same or different. If our concept is worth anything at all, we should be able to come up with a nominal measurement of it.

2. Ordinal This scale ranks the entities according to some criterion. An ordinal scale is more powerful than a nominal scale because it orders the entities. The ordering might be “greater than”’, “better than” or “more complex”.

3. Interval This scale orders values but there is also a notion of “relative distance” between two values. The difference between 6 and 10 degrees is larger than the difference between 6 and 8 degrees.

4. Ratio If there exists a meaningful zero value and the ratio between two measures is meaningful, a ratio scale can be used. A task that takes four days to complete is twice as long as a task that takes two days to complete.

Here are some examples of random variables:

1. the height of a person selected at random from a certain population (real valued)
2. the probability that Al Gore will be the Democratic nominee in 2004 (real valued, bounded)
3. the number of people that reside in Cambridge that will vote in the 2008 Presidential election (discrete, count)
4. the number of wars that the United States will fight in the next ten years (discrete, count)
5. the amount of money the United States will spend on foreign aid in 2010 (discrete/real valued)
6. the number of people who will die of AIDS in 2010 (discrete, count)
7. the amount of rainfall in Los Angeles on a given day of the year (real valued, bounded on bottom)
8. the number of flaws in a square yard of a certain material (discrete, count)
9. and to use an illustration form classical probability, the number of heads obtained in tossing a coin 100 times (discrete, count).

There are some other random variables which are more nonintuitive:

1. the number of wars that the United States fought in the last 10 years (discrete, count).

2. the amount of money the United States spent on foreign aid in 1987 (discrete/real valued).

3. the number of people who died of AIDS in 1990 (discrete, count).

# Random Variables II

Most people do not have a correct notion of “randomness”.

Random variables actually follow certain (mathematical) rules.

The most important distribution in statistics is the normal distribution. It is continuous, unimodal, symmetric and unbounded. Another important distribution is the binomial, which we have already seen—e.g., “the number of heads in several tosses of a coin.”

The normal distribution has two parameters: the mean ($\mu$) and variance ($\sigma^2$). The standard normal distribution is assumed to have a mean of zero and a variance of one. The concepts of mean and variance obviously are not limited to just this distribution.

We shall discuss probability densities and then examine the concepts of mean, variance, covariance and correlation without explicit reference to any distribution.

# Probability Density Functions (pdfs)

Let $X$ be a random variable with distribution $F$. And let $f(x)$ be the probability density function at $x$.

1. Sum over all possible values of $x$ is 1.0
   (a) For discrete: $\Sigma_{\text{all Possible } x} f(x) = 1$
   (b) For continuous (normal): $\int_{-\infty}^{\infty} f(x)dx = 1$

2. The function $f(x) \geq 0$ for each possible $x$.

Probabilities can be computed from densities:

1. For discrete: $P(x) = f(x)$

2. For continuous: $P(x) = 0$ (why?)

3. For both: $P(a \leq X \leq b) = \int_a^b f(x)dx$.

Let's look at some examples.

# Simple Discrete Example

Let Y $= 1$ denote a vote for the Republican Party and Y $= 0$ denote a vote for the Democratic Party. $\Omega = \{0, 1\}$. A valid probability distribution for Y is:

- P(Y=1) = 0.6
- P(Y=0) = 0.4

Another valid distribution for Y could be:

- P(Y=1) = 0.9
- P(Y=0) = 0.1

A generalized form for such distributions is called the Bernoulli Distribution.

# Bernoulli Distribution

This distribution corresponds to our previous voting example or to the flip of one possibly unfair coin.

- First Principles about the process that generates $Y_i$:

  1. $Y_i$ as 2 mutually exclusive outcomes
  2. The 2 outcomes are exhaustive

Mathematical expression for the pdf

- $P(Y_i = 1|\pi) = \pi$
  $P(Y_i = 0|\pi) = 1 - \pi$

- The parameter $\pi$ can be interpreted as a probability:

  ⋆ $P(Y = 1) = \pi$
  ⋆ $P(Y = 0) = 1 - \pi$

- The common summary of the Bernoulli distribution is:

$$P(Y = y|\pi) = \pi^{y}(1-\pi)^{1-y} \tag{94}$$

where $y \in \{0, 1\}$.

Using Equation 94 we can find out the probability of obtaining $y = 1$ or $y = 0$ given the parameter $\pi$ which can be interpret to be the probability of obtaining $y = 1$.

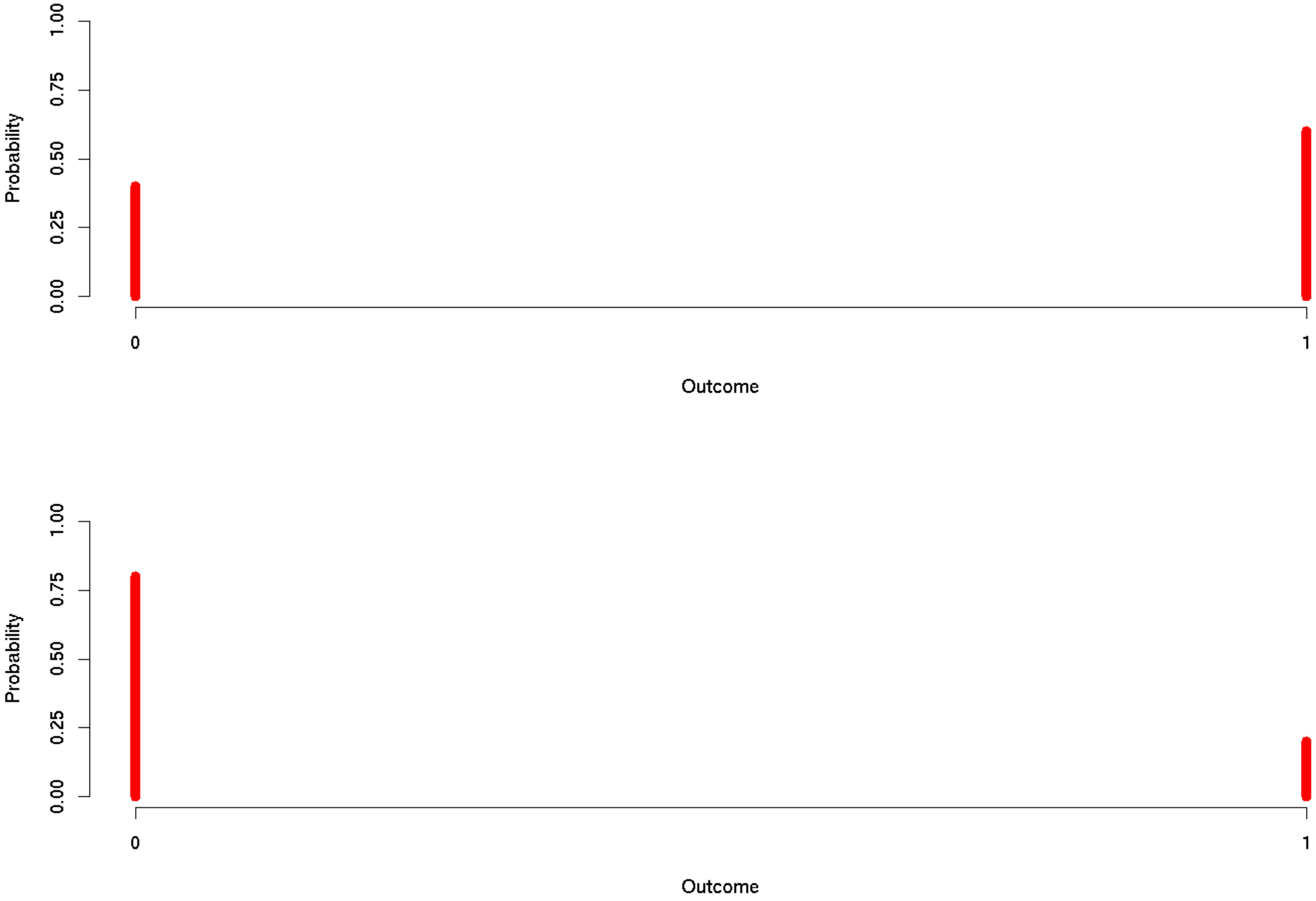
For example, if $\pi = .4$ our probability of obtaining $y = 1$ is:

$$P(Y = 1|\pi = .4) \quad = .4^{1}(1 - .4)^{1-1} \quad = 0.4 \tag{95}$$

And our probability of obtaining $y = 0$ is:

$$P(Y = 0|\pi = .4) \quad = .4^{0}(1 - .4)^{1-0} \quad = 0.6 \tag{96}$$

# Graphical Summary of Two Bernoulli Distributions



# Binomial Distribution

This is the most common discrete distribution. It results when there are many independent Bernoulli trials with the same $\pi$. The number of trails is denoted by $n$ and the number of successes (i.e., $y = 1$) by $s$. $n$ could be the number of voters and $s$ the number of votes for the Republican candidate.

To get from the Bernoulli to the Binomial we need to assume:

- There are $n$ trails—e.g., $n$ voters or $n$ tosses of a coin.
- The probability of a success or failure is defined by the Bernoulli distribution.
- The trails are statistically independent.

The binomial distribution depends on binomial coefficients which were first discovered by Jia Xian in the 11th century. Pascal and Newton independently discovered them.

The binomial distribution is covered in Wannacott and Wonnacott chapter 4 and Freedman et al. chapter 15. The Freedman discussion of distributions is generally better.

If the previous three conditions hold, then $S$ is called a binomial variable. The binomial PDF which gives the probability of exactly $s$ successes in $n$ trails when each trail has probability $\pi$ of a success is:

$$P(s) = \binom{n}{s}\pi^s(1-\pi)^{n-s} \tag{97}$$

where $\binom{n}{s}$ is the binomial coefficient.

The binomial coefficient $\binom{n}{s}$ is the number of combinations of $n$ things taken $s$ at a time. It is defined as:

$$\binom{n}{s} = \frac{n!}{s!(n-s)!} \tag{98}$$

where, in turn, the factorial $n!$ is defined by:

$$n! = n(n-1)(n-2)\cdots 1 \tag{99}$$

For example: How many different ways are there to get 1 revolution (i.e., one success) out of 8 chances (i.e., trials)?

Answer:

$$\frac{n!}{s!(n-s)!} = \frac{8!}{1!7!} = 8 \tag{100}$$

For example: How many different ways are there to get 2 revolutions out of 69 chances (i.e., trials)?

Answer:

$$\frac{n!}{s!(n-s)!} = \frac{69!}{2!67!} = 2346 \tag{101}$$

Question: What's the probability of observing 2 revolutions out of 69 chances if we use the binomial distribution with $\pi = \dfrac{2}{69}$?

Answer:

$$P(2) \quad = \quad \binom{69}{2}\left[\frac{2}{69}\right]^2 (1-\frac{2}{69})^{69-2} = 0.2746708 \tag{102}$$

Question: What's the probability of observing $\mathrm{round}(\frac{1}{8} * 69) = 9$ revolutions out of 69 chances if we use the binomial distribution with $\pi = \frac{2}{69}$?

Answer:

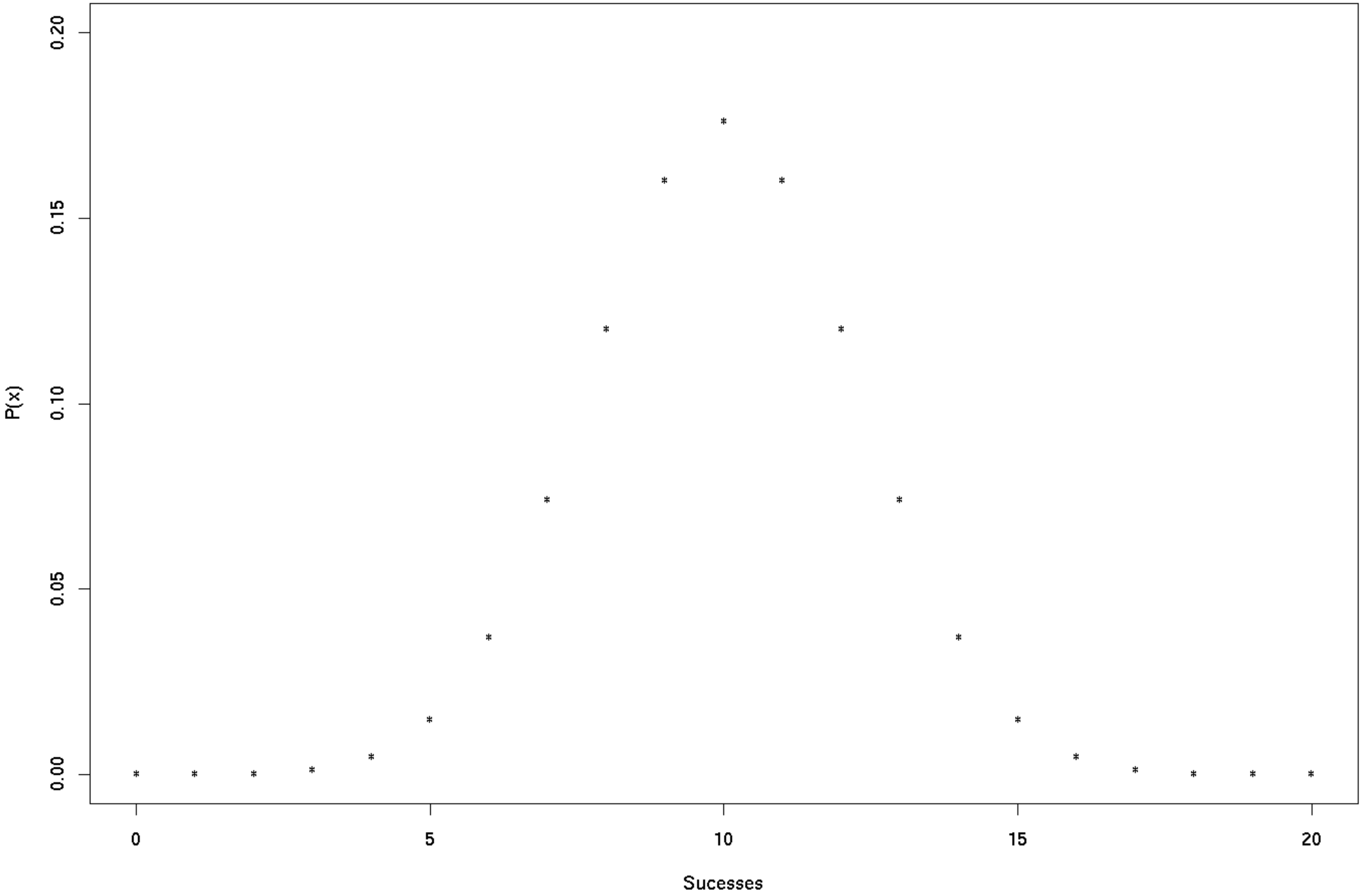
$$P(9) \quad = \quad \binom{69}{9} \left[\frac{2}{69}\right]^{9} (1 - \frac{2}{69})^{69-9} = 0.0001401329 \tag{103}$$

Question: What's the probability of observing 2 revolutions out of 69 chances if we use the binomial distribution with $\pi = \frac{1}{8}$?

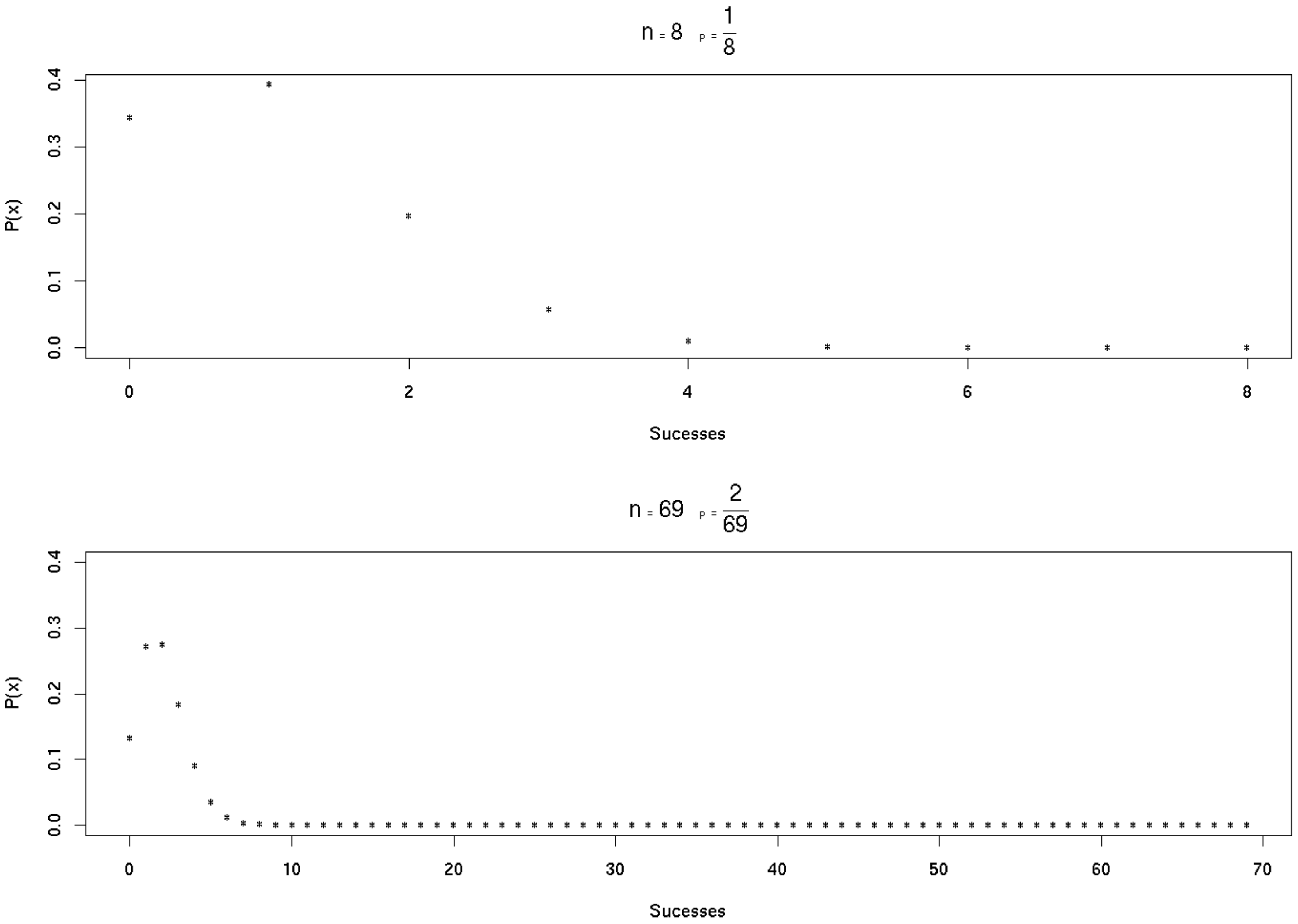
Answer:

$$P(2) \quad = \quad \binom{69}{2} \left[\frac{1}{8}\right]^{2} (1 - \frac{1}{8})^{69-2} = 0.004771859 \tag{104}$$

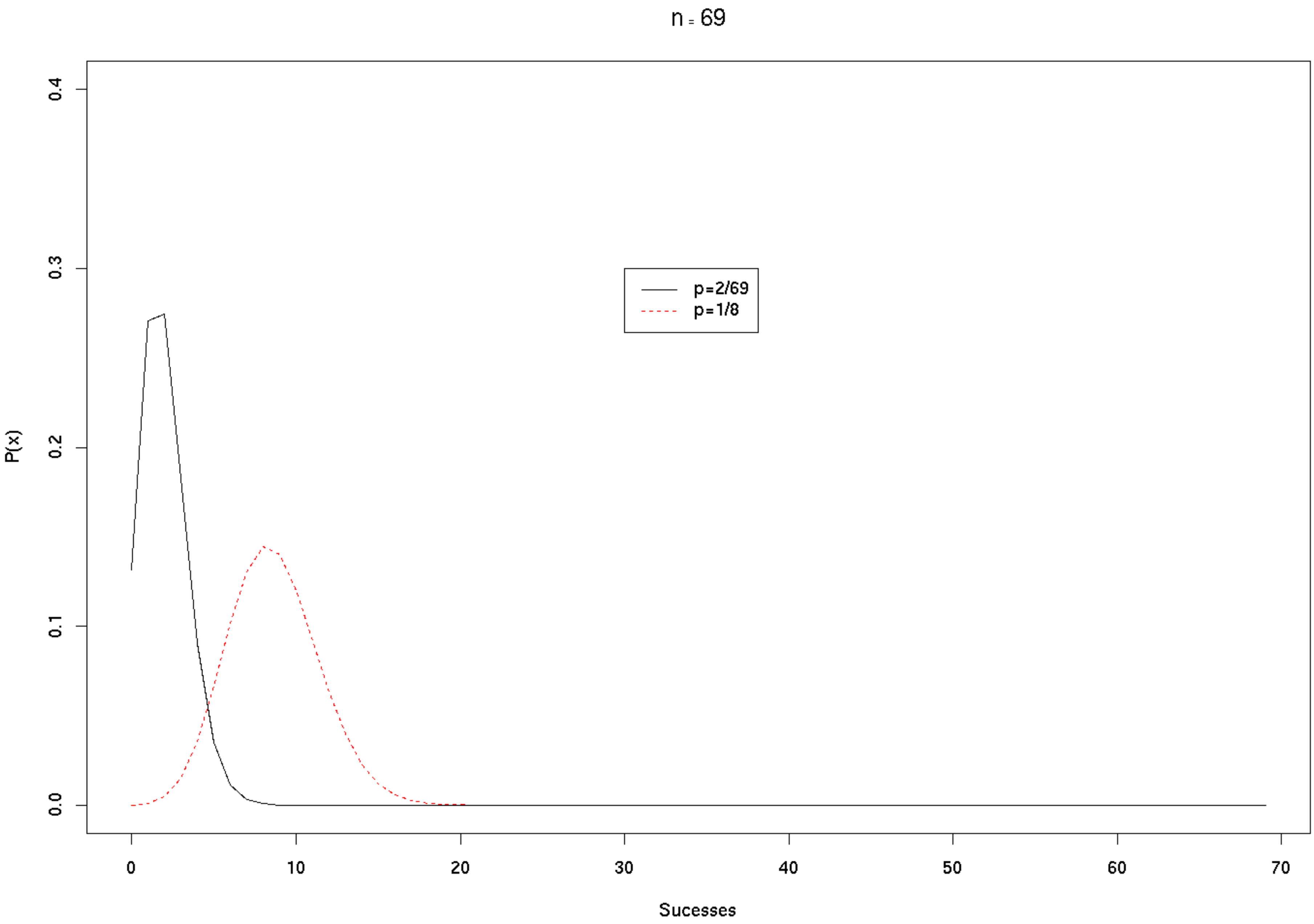
# Binomial PDF with $pi = 0.5$ and $n = 20$



# Binomial PDFs of Data From “Quality Meets Quantity”)



# Binomial PDFs of Data From “Quality Meets Quantity”
## RESCALED



# Binomial: More Details

There are two special cases of the binomial coefficient which are not covered by Equation 97:

- $\binom{n}{n} = 1$
- $\binom{n}{0} = 1$

In **R**:

- `gamma(n+1)` $=$ n!
- `choose(n,s)` $= \binom{n}{s}$
- `rbinom()` produces pseudo-random draws from the binomial distribution.
- `dbinom()` is the probability density function for the binomial distribution.

A lecture on how to use these **R** functions will be presented in section.

# Normal Distribution

- This is the most commonly used distribution in statistics.

- First discovered around 1720 by Abraham de Moivre, while he was developing the mathematics of chance.

- Around 1870, the Belgian mathematician and social scientist Adolph Quetelet had the idea of using the curve as an ideal to which data can be compared.

- With the work of Quetelet and others (such as Francis Galton) the normal distribution became an ideal defended by data and data defended by the ideal.

- “Everybody believes in the [normal approximation], the experimenters because they think it is a mathematical theorem, the mathematicians because they think it is an experimental fact” G. Lippmann (French Physicist, 1845-1921).

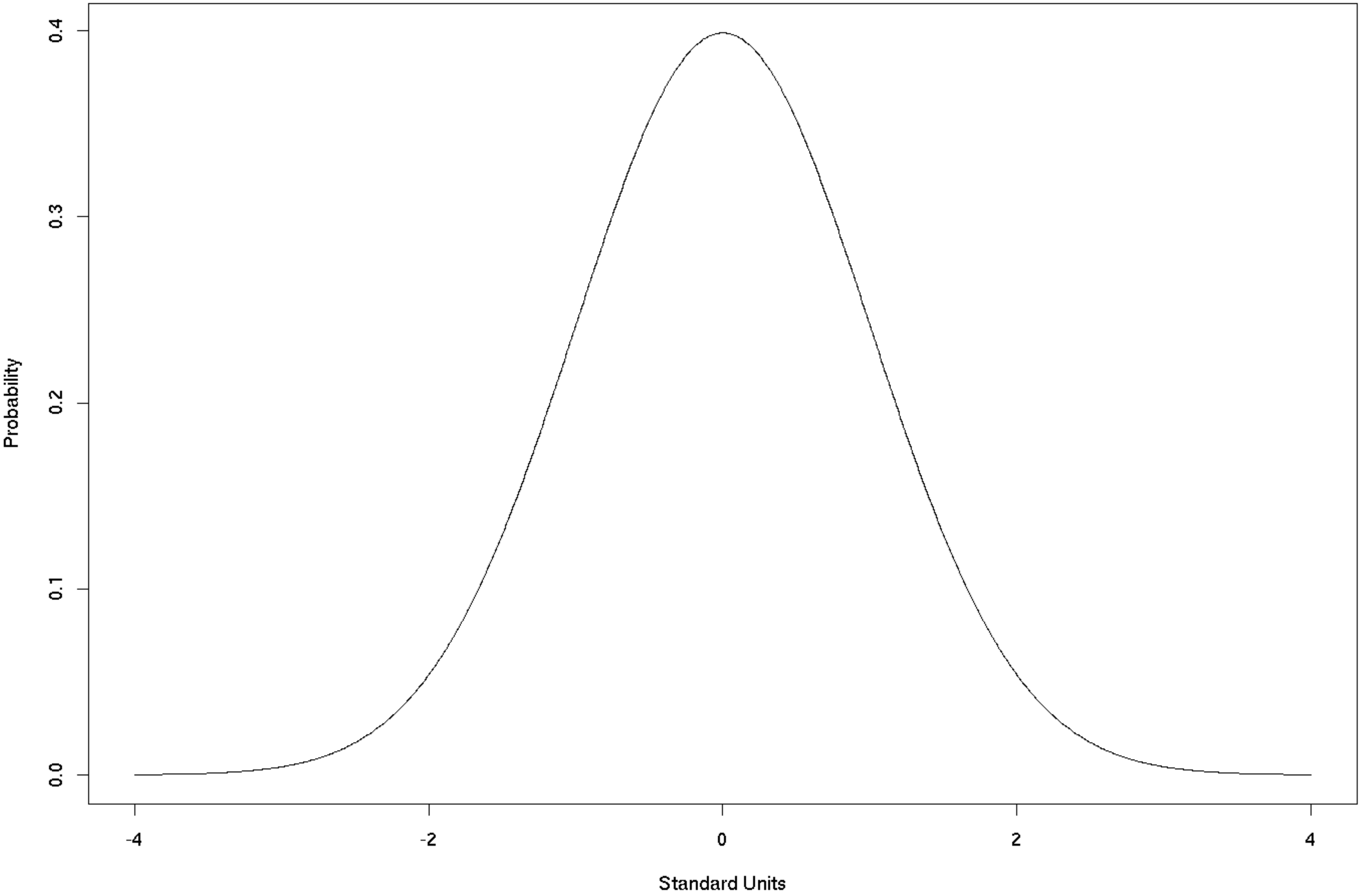
- The univariate normal density:

$$N(y_i|\mu, \sigma^2) = (2\pi\sigma^2)^{-1/2}\text{exp}\left(\frac{-(y_i - \mu)^2}{2\sigma^2}\right) \tag{105}$$

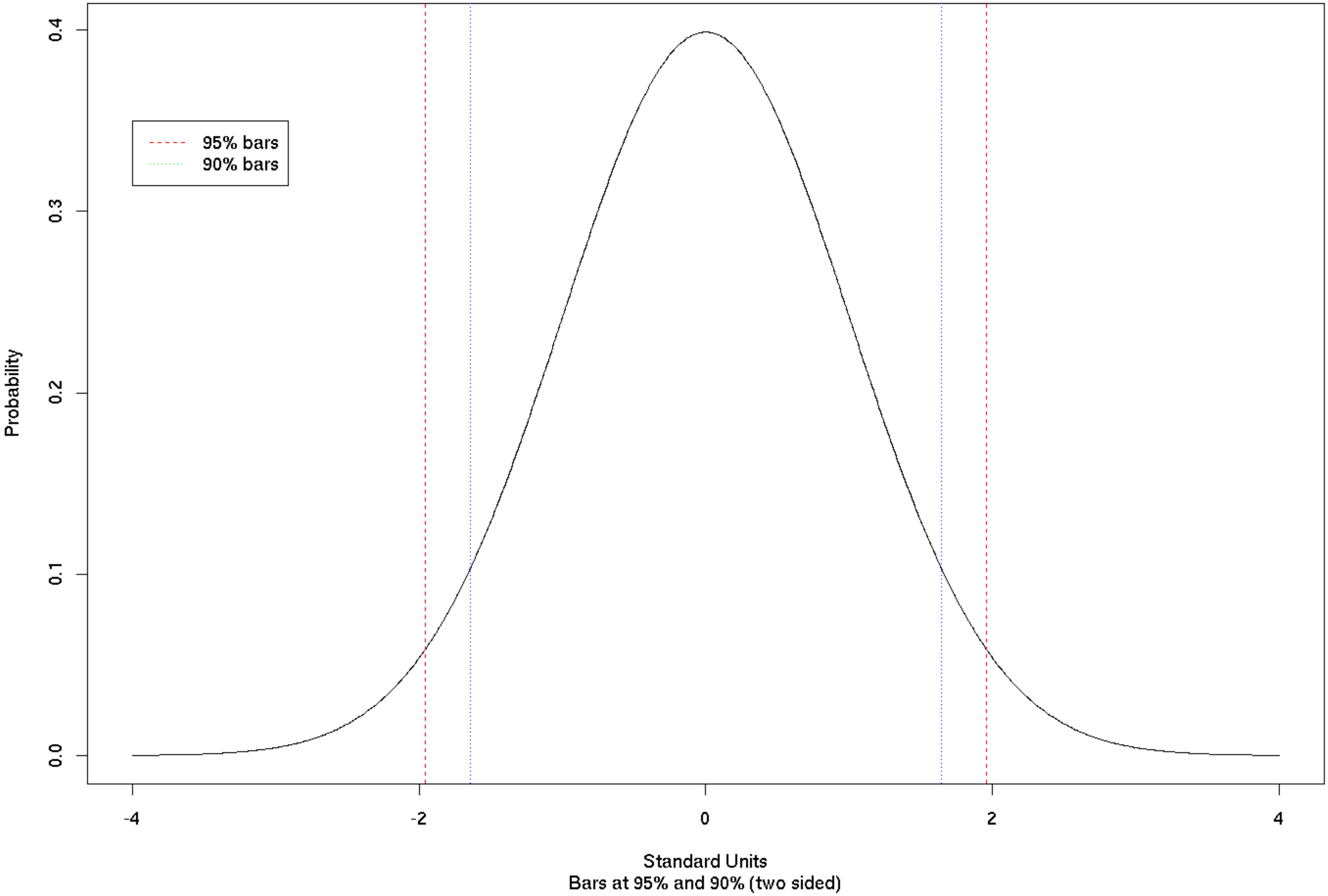
- The standardized univariate normal density:

$$N(y_i|0, 1) = (2\pi)^{-1/2}\text{exp}\left(\frac{-y_i^2}{2}\right) \tag{106}$$

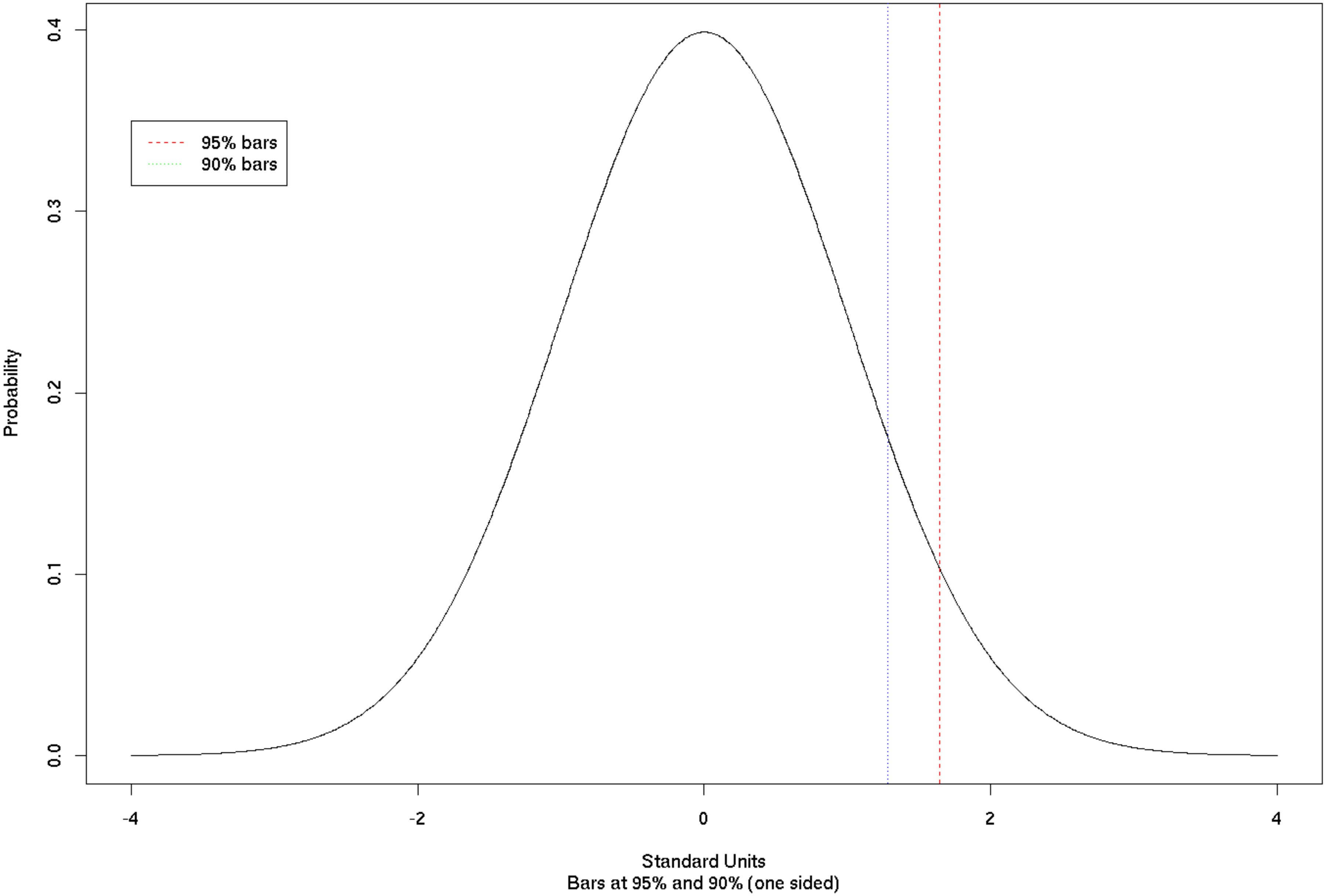
# Standard Normal PDF



# Standard Normal PDF, two-sided CI



# Standard Normal PDF, one-sided CI



# Cumulative Distribution Function (CDF)

The Cumulative Distribution Function is defined as
$F(y) \equiv P(Y \leq y) = \int_{-\infty}^{y} f(z)dz.$

This is the cumulative probability density up to $y$ and it is denoted $F(y)$.

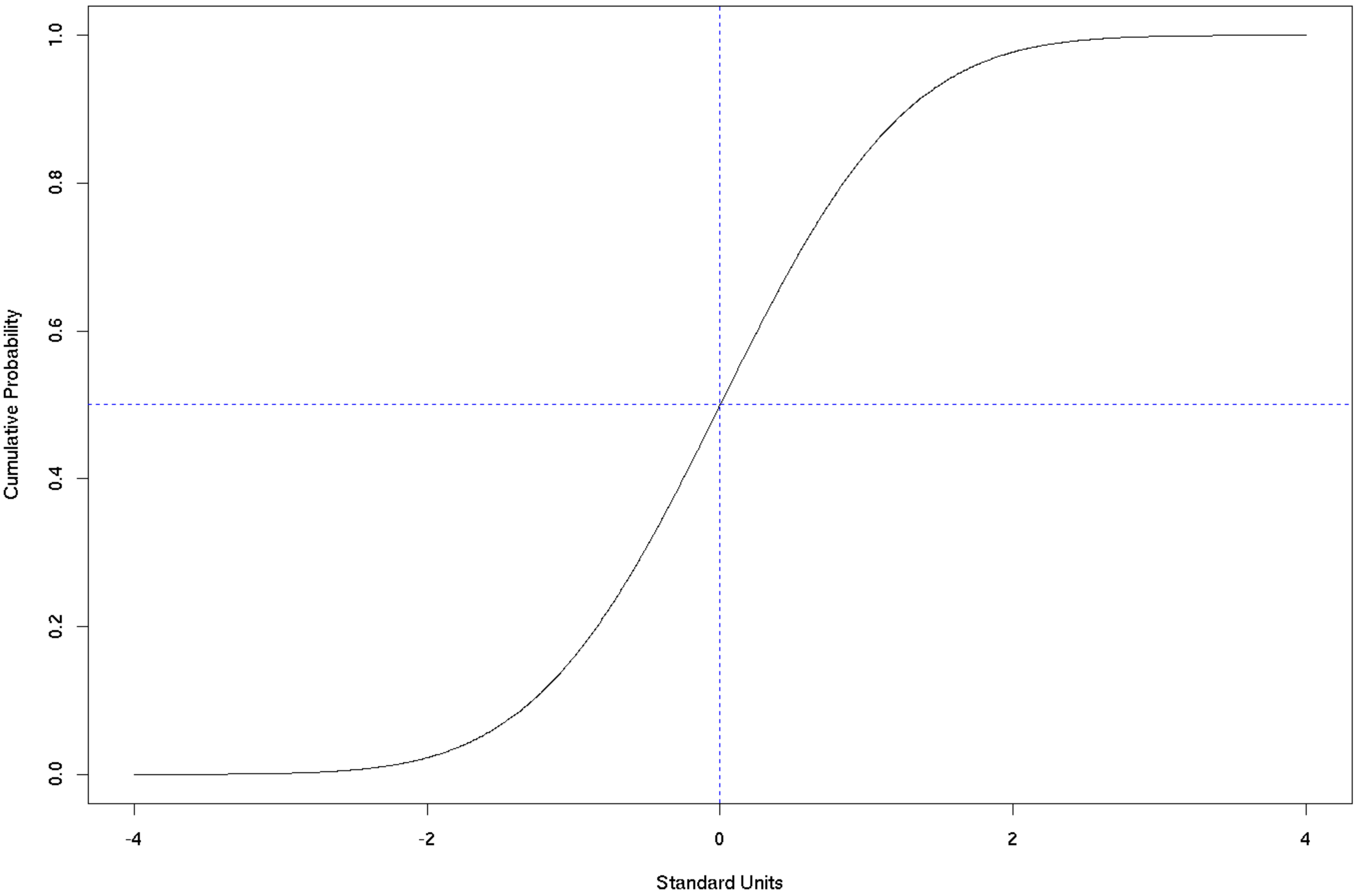
It has some properties:

- $F(-\infty) = 0$
- $F(\infty) = 1$
- $\partial F(y)/\partial y = f(y)$

Hence, there is a mapping from the probability distribution function (PDF) and the cumulative distribution function (CDF):

$$
\begin{aligned}
P(a \leq Y \leq b) &= \int_{a}^{b} f(y)dy && (107)\\
&= \int_{-\infty}^{b} f(y)dy - \int_{-\infty}^{a} f(y)dy && (108)\\
&= F(b) - F(a) && (109)
\end{aligned}
$$

# Cumulative Standard Normal PDF (from the bottom)



# Normal Distribution R

R has various functions associated with the normal, they will be discussed in section:

- `rnorm()`: generates pseudo-random draws from a normal distribution
- `dnorm()`: the probability distribution function for a normal distribution
- `pnorm()`: the cumulative distribution function
- `qnorm()`: the quantile distribution function. You tell this function the probability you want and it returns the quantile. This is the reverse of what `pnorm` does.

For examples see `http://jsekhon.fas.harvard.edu/gov1000/normal1.R`

# Mean and Median

Estimating the mean ought to be familiar to everyone. Let $\mu$ denote the mean of $n$ realizations of the random variable X: $x_1, x_2, \cdots, x_n$

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i. \tag{110}$$

$\hat{\mu}$ is our estimate of $\mu$.

The median should also be familiar. It is the .5 quantile.

Whether the mean or median is a better measure depends on the underlying distribution of the variable of interest. Social scientists, and lay people, usually (often implicitly) assume the normal distribution. Therefore, they generally use the mean.

# Rules of Summation; Variance and Covariance

Here are eight rules of summation. In the course of describing them we will also define variance, covariance and correlation.

Rule 1

The summation of a constant $k$ times a variable is equal to the constant times the summation of that variable:

$$\sum_{i=1}^{n} kX_i = k \sum_{i=1}^{n} X_i \tag{111}$$

Rule 2

The summation of the sum of observations on two variables is equal to the sum of their summations

$$\sum_{i=1}^{n} (X_i + Y_i) = \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} Y_i \tag{112}$$

Rule 3

The summation of a constant over $n$ observations equals the product of the constant and $n$.

$$\sum_{i=1}^{n} k = k \times n \tag{113}$$

These three rules can be used to to derive some other rules.

## Rule 4

The summation of the deviations of observations on X about its mean is zero.

$$\sum_{i=1}^{n}(X_i - \bar{X}) = 0 \tag{114}$$

Proof

$$x_i = X_i - \bar{X} \tag{115}$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{116}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}) \tag{117}$$

$$= \frac{\sum_{i=1}^{n} X_i}{n} - \bar{X} \tag{118}$$

$$= \bar{X} - \bar{X} \tag{119}$$

$$= 0 \tag{120}$$

We define the variance of X to be

$$var(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{121}$$

The covariance between two random variables (X, Y) is defined to be:

$$cov(X, Y) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) \tag{122}$$

And the correlation is:

$$cor(X, Y) = \frac{cov(X, Y)}{\sqrt{(var(X)}\sqrt{(var(Y)}} \tag{123}$$

Please see page 167 in Wonnacott and Wonnacott and chapters 8 and 9 in Freedman et al. for more details.

## Rule 5

The covariance between X and Y is equal to the mean of the products of observations on X and Y minus the product of their means:

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \bar{X}\bar{Y} \tag{124}$$

Proof:

$$\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})=\frac{1}{n}\sum_{i=1}^{n}X_iY_i-\frac{1}{n}\sum_{i=1}^{n}\bar{X}Y_i-\frac{1}{n}\sum_{i=1}^{n}X_i\bar{Y}+\frac{1}{n}\sum_{i=1}^{n}\bar{X}\bar{Y} \tag{125}$$

Using Rule 1, we obtain:

$$\mathrm{cov}(X,Y)=\frac{1}{n}\sum_{i=1}^{n}X_iY_i-\frac{1}{n}\bar{X}\sum_{i=1}^{n}Y_i-\frac{1}{n}\bar{Y}\sum_{i=1}^{n}X_i+\frac{1}{n}\sum_{i=1}^{n}\bar{X}\bar{Y} \tag{126}$$

Recalling the definition of the mean of X and the mean of Y we obtain:

$$=\frac{1}{n}\sum_{i=1}^{n}X_iY_i-\bar{X}\bar{Y}-\bar{Y}\bar{X}+\frac{1}{n}\sum_{i=1}^{n}\bar{X}\bar{Y} \tag{127}$$

$$=\frac{1}{n}\sum_{i=1}^{n}X_iY_i-2\bar{X}\bar{Y}+\bar{X}\bar{Y} \tag{128}$$

$$=\frac{1}{n}\sum_{i=1}^{n}X_iY_i-\bar{X}\bar{Y} \tag{129}$$

## Rule 6

The variance of X is equal to the mean of the squares of observations on X minus its mean squared. Rule 6 follows from Rule 5 since it applies to the case in which X and X are the two variables (instead of X and Y).

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \bar{X}^2 \tag{130}$$

■

It is interesting to note that when X and Y have a mean of zero, the definitions of covariance and variance become:

$$cov(x, y) = \frac{1}{n}\sum_{i=1}^{n}x_i y_i \tag{131}$$

$$var(x) = \frac{1}{n}\sum_{i=1}^{n}x_i^2 \tag{132}$$

In certain situations it will be necessary to use summations which apply to two random variables, called double summations. Specifically, let $X_{ij}$ be a random variable which takes on $N$ values for each outcome of $i$ and $j$. There will, of course, be $N^2$ total outcomes. New we define the double summation of these $N^2$ outcomes as:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} X_{ij} = \sum_{i=1}^{n} (X_{i1} + X_{i2} + \cdots + X_{in}) \tag{133}$$

$$= (X_{1\,1} + X_{1\,2} + \cdots + X_{1\,n}) + \cdots + (X_{n\,1} + X_{n\,2} + \cdots + X_{n\,n}) \tag{134}$$

We here list two rules of double-summation.

## Rule 7

$$\sum_{i=1}^{n}\sum_{j=1}^{n} X_i Y_j = \left(\sum_{i=1}^{n} X_i\right)\left(\sum_{j=1}^{n} Y_j\right) \tag{135}$$

Note that the double summation in Rule 7 is very different from the single summation $\sum_{i=1}^{n} X_i Y_i$, which contains $n$ rather than $n^2$ terms.

## Rule 8

$$\sum_{i=1}^{n}\sum_{j=1}^{n} (X_{ij} + Y_{ij}) = \sum_{i=1}^{n}\sum_{j=1}^{n} X_{ij} + \sum_{i=1}^{n}\sum_{j=1}^{n} Y_{ij} \tag{136}$$

# Mathematical Expectation

In order to study random variables and their probability distributions, it is useful to define the concept of mathematical expectation of a random variable and of the functions of a random variables.

Consider the experiment of rolling a single die. Let X be the value that shows on the die. The probability distribution of X is

| x | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| f(x) | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{1}{6}$ |

What would the average value of X be if the experiment were repeated an infinite number of times? Intuitively, you would expect $X = 1$ on $\frac{1}{6}$ of the throws, $X = 2$ on $\frac{1}{6}$ of the throws, and so on. So, on average, the value of X is

$$1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = 3.5 \tag{137}$$

That is, 3.5 is the average value of X that occurs in infinitely many trials of the experiment. This average is the *expected value* of the random variable X, despite the fact that X cannot actually take the value 3.5.

# Variance

The variance of a random variable provides a measure of the spread, or dispersion, around the mean. It is denoted $\sigma^2$, and (in the discrete case) it is defined as

$$var(X) = \sigma^2 = \sum_{i=1}^{n} p_i \left[X_i - E(X)\right]^2 \tag{138}$$

Thus, the variance is a weighted average of the squares of the deviations of outcomes on X from its expected value, with the corresponding probabilities of each outcome occurring serving as weights.

The variance is in itself an expectation since,

$$var(X) = E\left[X - E(X)\right]^2 \tag{139}$$

The positive square root of the variance is called the standard deviation and is denoted by $\sigma$.

# Properties of the Expectations Operator

Result 1

$$E(aX + b) = aE(X) + b, \tag{140}$$

where $X$ is a random variable and $a$ and $b$ are constants.

## Result 2

$$E[(aX)^2] = a^2 E(X^2) \tag{141}$$

Note that $E(X^2) \neq [E(X)]^2$.

For example, let $X = 1$ where a coin appears heads and $X = 0$ when it appears tails. Then for a fair coin, $p_1 = \frac{1}{2}$ and $p_0 = \frac{1}{2}$, so that

$$E(X^2) = \frac{1}{2}(1^2) + \frac{1}{2}(0^2) = \frac{1}{2}1 = \frac{1}{2}, \tag{142}$$

However,

$$E(X) = \frac{1}{2}(1) + \frac{1}{2}(0) = \frac{1}{2} \tag{143}$$

and

$$[E(X)]^2 = \frac{1}{4} \tag{144}$$

## Result 3

$$var(aX+b) = a^2 var(X) \tag{145}$$

PROOF

By definition

$$var(aX+b) = E\left[(aX+b) - E(aX+b)\right]^2 \tag{146}$$

By Result 1 we have that $E(aX+b) = aE(X)+b$, therefore

$$\begin{aligned}
var(aX+b) &= E\left[aX - E(aX)\right]^2 && (147)\\
&= E\left[aX - aE(X)\right]^2 && (148)\\
&= E\left[a(X - E(X))\right]^2 && (149)\\
&= a^2 E\left[X - E(X)\right]^2 && (150)\\
&= a^2 var(X) && (151)
\end{aligned}$$

Now, we can use the expectations operator to prove some results concerning the covariance between two random variables.

## Result 4

If $X$ and $Y$ are random variables, then

$$E(X + Y) = E(X) + E(Y) \tag{152}$$

## Result 5

$$var(X+Y) = var(X) + var(Y) + 2cov(X, Y) \tag{153}$$

PROOF

$$\begin{aligned}
var(X+Y) &= E[(X+Y) - E(X+Y)]^2 && (154)\\
&\text{by Result 4}\\
&= E[(X+Y) - E(X) - E(Y)]^2 && (155)\\
&= E[(X - E(X)) + (Y - E(Y))]^2 && (156)\\
&= E[X - E(X)]^2 + E[Y - E(Y)]^2 + 2E[(X - E(X))(Y - E(Y))] && (157)\\
&= var(X) + var(Y) + 2\,cov(X, Y) && (158)
\end{aligned}$$

Result 6

If X and Y are independent, then $E(XY) = E(X)E(Y)$.

Result 7

If X and Y are independent, then $cov(X, Y) = 0$.

PROOF

$$\begin{aligned} cov(X, Y) &= E[X - E(X)][Y - E(Y)] \qquad (159) \\ &= E[XY - E(X)Y - XE(Y) + E(X)E(Y)] \qquad (160) \\ &= E(XY) - E(X)E(Y) \qquad (161) \end{aligned}$$

by Result 6

$$= 0. \qquad (162)$$

## Result 8

$$var(\bar{X}) = \frac{\sigma^2}{n}, \tag{163}$$

where $\bar{X}$ is the sample mean of a random variable with mean $\mu$ and variance $\sigma^2$.

PROOF:

$$var(\bar{X}) = var(\frac{1}{n}\sum_{i=1}^{n} X_i) \tag{164}$$

by Result 3

$$= \left(\frac{1}{n}\right)^2 var(\sum_{i=1}^{n} X_i) \tag{165}$$

by Results 5 and 7

and the assumption that all $X_i$ are independent

$$= \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} \sigma^2$$

$$= \left(\frac{1}{n}\right)^2 n\sigma^2 \tag{166}$$

$$= \frac{\sigma^2}{n}. \tag{167}$$

Result 8 shows that the variance of the estimator of the mean $\bar{X}$ falls as the sample size increases. Thus, with more and more information, we get more and more accuracy in our estimates of the mean $\mu$. What happens to this variance as we get infinite data?

## Result 9

$$\sigma^2 = E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] \tag{168}$$

## Result 10

Note that

$$E(X^2) = \sigma^2 + \mu^2 \tag{169}$$

$$= Var(X) + [E(X)]^2 \tag{170}$$

# Sampling

Simple random sampling, or random sampling without replacement, is the sampling design in which $n$ distinct units are selected from the $N$ units in the population in such a way that every possible combination of $n$ units is equally likely to be the sample selected.

The sample may be obtained through $n$ selections in which at each step every unit of the population not already selected has equal chance of selection.

Equivalently, one may make a sequence of independent selections from the whole population, each unit having equal probability of selection at each step, discarding repeat selections and continuing until $n$ distinct units are obtained.

With simple random sampling, the probability that the $i$th unit of the population is included in the sample is $\pi_i = \frac{n}{N}$, so that the inclusion probability is the same for each unit.

Designs other than simple random sampling may give each unit equal probability of being included in the sample, but only with simple random sampling does each possible sample of $n$ units have the same probability.

# Estimating The Population Mean

With simple random sampling (with replacement), the sample mean $\bar{y}$ is an unbiased estimator of the population mean $\mu$. The population mean $\mu$ is the average of the $y$-values in the whole population:

$$\mu = \frac{1}{N}(y_1 + y_2 + \cdots + y_N) \tag{171}$$

$$= \frac{1}{N}\sum_{i=1}^{N} y_i \tag{172}$$

The sample mean $\bar{y}$ is the average of the $y$-values in the sample:

$$\bar{y} = \frac{1}{n}(y_1 + y_2 + \cdots + y_n) \tag{173}$$

$$= \frac{1}{n}\sum_{i=1}^{n} y_i \tag{174}$$

$\bar{y}$ is an unbiased estimate of $\mu$.

# Sampling: Some Underlying Ideas

The estimator $\bar{y}$ is a random variable, the outcome of which depends on which sample is selected. With any given sample, the value of $\bar{y}$ may be either higher or lower than the population mean $\mu$. But the expected value of $\bar{y}$, then over all possible samples, equals $\mu$. The estimator $\bar{y}$ is said to be an $unbiased$ estimator of the population quantity $\mu$.

In fact, $\bar{y}$ is not only biased, but also *design-unbiased*.

It is called *design-unbiased* because the unbiasedness of the sample mean for the population mean with simple random sampling does not depend on any assumptions about the population itself. This is true because the probability with respect to which the expectation is evaluated arises from the probabilities, due to the design, of selecting different samples.

# Estimating the Sample Variance

With simple random sampling, the sample variance $\hat{\sigma}^2$, often also denoted by $s^2$, is an unbiased estimator of the finite population variance $\sigma^2$. The finite population variance is defined as

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu)^2 \tag{175}$$

The sample variance is defined as

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{176}$$

The variance of the estimator $\bar{y}$ with simple random sampling is

$$var(\bar{y}) = \left(\frac{N-n}{N-1}\right)\frac{\sigma^2}{n} \tag{177}$$

An unbiased estimator of this variance is

$$\widehat{var}(\bar{y}) = \left(\frac{N-n}{N-1}\right)\frac{s^2}{n} \tag{178}$$

Recall that the square root of the variance of the estimator is its standard error.

The quantity $\frac{N-n}{N-1}$ is termed the finite population correction factor.

If the population is large relative to the sample size, so that the sampling fraction $\frac{n}{N}$ is small, the finite population correction factor will be close to one, and the variance of the sample mean $\bar{y}$ will be approximately equal to $\frac{\sigma^2}{n}$.

Omitting the finite population correction factor in estimating the variance of $\bar{y}$ in such a situation will tend to give a slight overestimate of the true variance.

When sampling small populations, however, the finite population correction factor may have an appreciable effect in reducing the variance of the estimator, and it is important to include it in the estimate of that variance.

Note that as sample size $n$ approaches the population size $N$ in simple random sampling, the finite population correction factor approaches zero, so that the variance of the estimator $\bar{y}$ approaches zero.

# Estimating The Population Total

To estimate the population total $\tau$, where

$$\tau = \sum_{i=1}^{N} y_i \tag{179}$$

$$= N\mu \tag{180}$$

the population mean is multiplied by $N$. And unbiased estimator of the population total is

$$\hat{\tau} = N\bar{y} \tag{181}$$

$$= \frac{N}{n}\sum_{i=1}^{n} y_i \tag{182}$$

Since the estimator $\hat{\tau}$ is $N$ times the estimator $\bar{y}$, the variance of $\hat{\tau}$ is $N^2$ times the variance of $\bar{y}$. Thus,

$$var(\hat{\tau}) = N^2 var(\bar{y}) = N(N-n)\frac{\sigma^2}{n} \tag{183}$$

An unbiased estimator of this variance is

$$\widehat{var}(\hat{\tau}) = N^2 \widehat{var}(\bar{y}) = N(N-n)\frac{s^2}{n} \tag{184}$$

# Estimating Proportions

When the population variable of interest may take on only the values zero and one, the population total $\tau$ is the number of units in the population with the attribute, and the population mean $\mu$ is the proportion of units in the population with the attribute.

To estimate a population proportion using simple random sampling, the usual methods associated with estimating a population mean could be used.

However, several special features are worth noting:

1. The formulae simplify considerably with attribute (i.e., zero, one) data.;
2. exact confidence intervals are possible;
3. a sample size sufficient for a desired absolute precision may be chosen without any information about population parameters. This is possible because the population parameters are bounded.

# Estimating a Population Proportion

Writing $p$ for the proportion in the population with the attribute,

$$p = \frac{1}{N}\sum_{i=1}^{N} y_i = \mu \tag{185}$$

The finite population variance is

$$\sigma^2 = \frac{\sum_{i=1}^{N}(y_i - p)^2}{N-1} \tag{186}$$

$$= \frac{\sum_{i=1}^{N} Y_i^2 - Np^2}{N-1} \tag{187}$$

$$= \frac{Np - Np^2}{N-1} \tag{188}$$

$$= \frac{N}{N-1}p(1-p) \tag{189}$$

Letting $\hat{p}$ denote the proportion in the sample with the attribute,

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y} \tag{190}$$

The sample variance is

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} \tag{191}$$

$$= \frac{\sum_{i=1}^{n} y_i^2 - n\hat{p}^2}{n-1} \tag{192}$$

$$= \frac{n}{n-1}\hat{p}(1-\hat{p}) \tag{193}$$

Thus, the relevant statistics can be computed from the sample proportion alone.

Since the sample proportion is the sample mean of a simple random sample, it is unbiased for the population proportion, and has variance

$$var(\hat{p}) \;=\; \left(\frac{N-n}{N-1}\right)\frac{p(1-p)}{n} \tag{194}$$

An unbiased estimator of this variance is

$$\widehat{var}(\hat{p}) \;=\; \left(\frac{N-n}{N-1}\right)\frac{\hat{p}(1-\hat{p})}{n} \tag{195}$$

# Confidence Intervals For A Proportion

An approximate confidence interval for $p$ based on a normal distribution is given by

$$\hat{p} \pm t\sqrt{\widehat{var}(\hat{p})}, \tag{196}$$

where $t$ is the upper $\alpha/2$ point of the $t$-distribution with $n-1$ degrees of freedom—in this class we just use the normal approximation.

The normal approximation on which this interval is based improves the larger the sample size and the closer $p$ is to 0.5.

Confidence limits may also be obtained based on the exact hypergeometric distribution of the number of units in the sample with the attribute. I will not discuss the exact method.

# Sample Size for Estimating a Proportion

To obtain an estimator $\hat{p}$ having probability at least $1-\alpha$ of being no farther than $d$ from the population proportion, the sample size formula based on the normal approximation gives

$$n = \frac{Np(1-p)}{(N-1)\frac{d^2}{z^2} + p(1-p)}, \tag{197}$$

where $z$ is the upper $\alpha/2$ point of the normal distribution. When the finite population correction can be ignored, the formula reduces to

$$n_0 = \frac{z^2 p(1-p)}{d^2} \tag{198}$$

Note that the formulae depend on the unknown population proportion $p$.

If no estimate of $p$ is available prior to the survey, a "worst case" value of $p = .5$ can be used in determining sample size.

The quantity $p(1-p)$, and hence the value of $n$ required by the formula, assumes its maximum value when $p$ is one-half.

## Examples

Assuming the worst case and no finite sample correction, to be 95% certain that we are within 4% we will need a sample size of:

$$\frac{z^2p(1-p)}{d^2} = \tag{199}$$

$$= (1.96)^2 * .5 * (1-.5)/(.04^2) \tag{200}$$

$$= 600.25 \tag{201}$$

Is this correct?

$$t\sqrt{\widehat{var}(\hat{p})} = CI \tag{202}$$

$$1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \tag{203}$$

$$1.96\sqrt{\frac{.5(1-.5)}{600.25}} = 0.04 \tag{204}$$

Assuming the worst case and no finite sample correction, to be 95% certain that we are within 1% we will need a sample size of:

$$
\begin{aligned}
\frac{z^2 p(1-p)}{d^2} &= && (205)\\
&= (1.96)^2 * .5 * (1-.5)/(.01^2) && (206)\\
&= 9604 && (207)
\end{aligned}
$$

Is this correct?

$$
\begin{aligned}
t\sqrt{\widehat{var}(\hat{p})} &= CI && (208)\\
1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} &= && (209)\\
1.96\sqrt{\frac{.5(1-.5)}{9604}} &= 0.01 && (210)
\end{aligned}
$$

Assuming the worst case and no finite sample correction, to be 95% certain that we are within 0.27% (the margin of victory in the 2000 national Presidential vote totals) we will need a sample size of:

$$\frac{z^2 p(1-p)}{d^2} = \tag{211}$$

$$= (1.96)^2 * .5 * (1 - .5)/(.0027^2) \tag{212}$$

$$= 131742.1 \tag{213}$$

Is this correct?

$$t\sqrt{\widehat{var}(\hat{p})} = CI \tag{214}$$

$$1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \tag{215}$$

$$1.96\sqrt{\frac{.5(1-.5)}{131742.1}} = 0.0027 \tag{216}$$

What if we are willing to be only 90% certain?

$$\frac{z^2 p(1-p)}{d^2} = \tag{217}$$

$$= (1.644)^2 * .5 * (1 - .5)/(.0027^2) \tag{218}$$

$$= 92686.42 \tag{219}$$

Is this correct?

$$t\sqrt{\widehat{var}(\hat{p})} = CI \tag{220}$$

$$1.644\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \tag{221}$$

$$1.644\sqrt{\frac{.5(1-.5)}{92686.42}} = 0.0027 \tag{222}$$

# Regression, It's Finally Here!!

We now start our discussion of regression estimation. This is the most commonly used non-trivial statistical apparatus in the social sciences.

It is hoped that statistical inference through the use of multiple regression, and other such methods, is able to provide to the social scientist what experiments and rigorous mathematical theories provide, respectively, to the micro-biologist and astronomer.

We are after the ability to take multiple factors into account so that we may infer the causal (or true) relationship between $x$ and $y$.

The textbooks do a good job of discussing the details of the methods involved. But they lose focus on the big picture.

# Presidential Approval

Our running example in this section will be the U.S. Presidential Approval series.

This series is based on the following question:

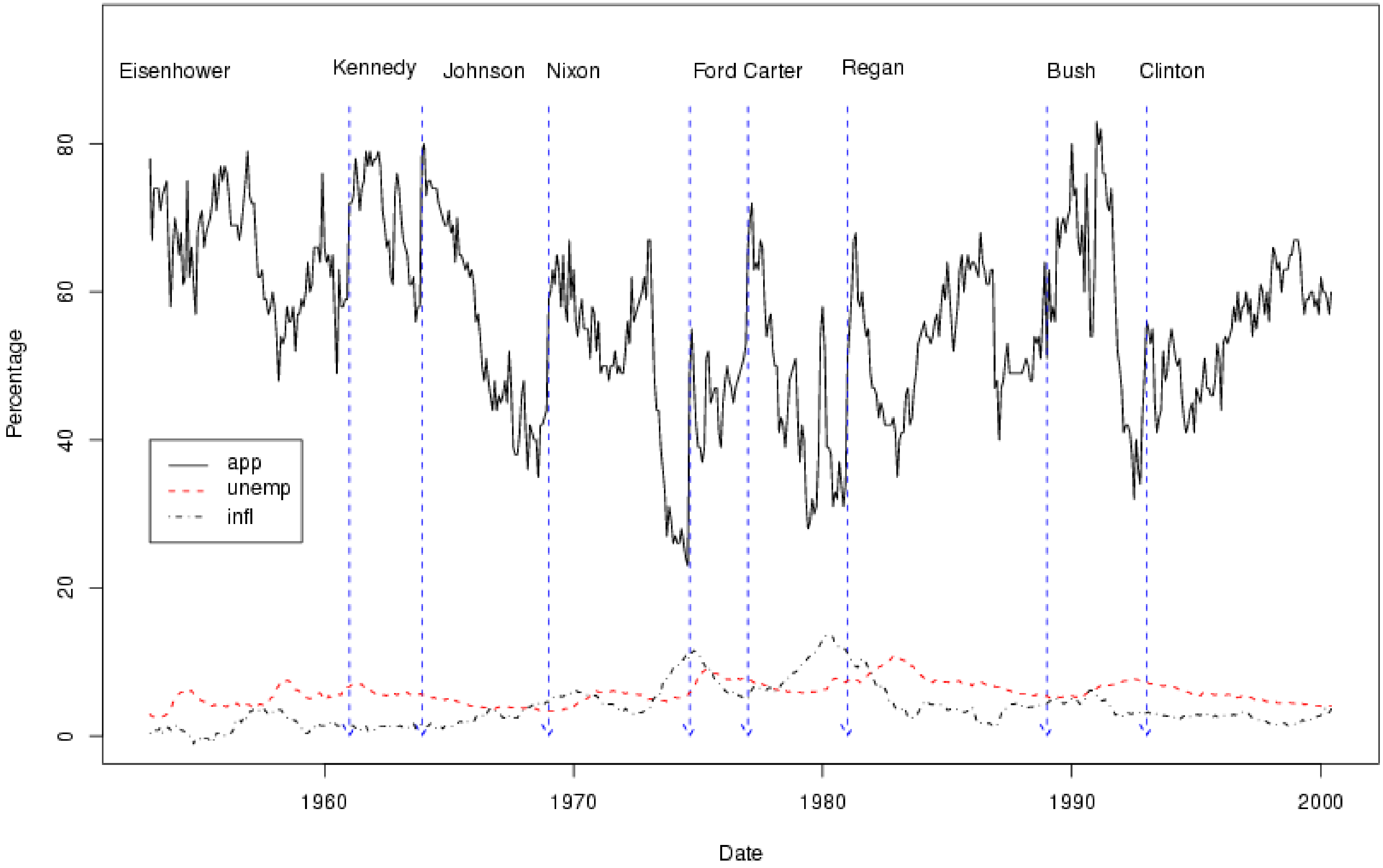
> ”Do you approve or disapprove of the way [Bush] is handling his job as president?”

[Bush] is obviously replaced by who the president happens to be. The question has three valid responses: approve, disapprove, and no opinion.

First, we are interesting in how this series is related to unemployment and inflation.

Figure 1 plots all three series.

# Figure 1: U.S. Presidential Approval, Unemployment and Inflation



Here are some basic summary statistics concerning these three series:

- mean(approval) = 56.0579
- mean(unemployment rate) = 5.809825
- mean(inflation rate) = 3.899265

- cov(approval,unemployment) = -5.318496
- cor(approval,unemployment) = -0.2905604

- cov(approval,inflation) = -19.4367
- cor(approval,inflation) = -0.5580672

- cov(unemployment,inflation) = 1.476466
- cor(unemployment,inflation) = 0.3326103

Questions:
What has a stronger impact on approval, unemployment or inflation?
How much does approval change if unemployment increases by 1%?
Are voters forward looking or retrospective?

# Data Generating Process (DGP)

A fundamental methodology of modern statistics is to assume that observed data are generated by some stochastic process—i.e., some probability distribution. Given this we can define a DGP.

Suppose observed data are realizations of a stochastic process of the following variety:

$$Y = f\left(X_1, X_2, \cdots X_k, \epsilon\right), \tag{223}$$

let $Z$ denote a set which contains all of the variables in the DGP.

The probability law P gives a complete description of the stochastic process. If P where known, we would be able to know every aspect of our DGP, such as the conditional means, conditional variance, etc.

P is determined by the nature of the world, it is not known. The problem of estimation and inference arises precisely because it is unknown.

If we observe a realization of the sequence $Z$, then we can infer some knowledge of $P$ from this realization. In practice, observation of the entire sequence is impossible. Instead, we have a realization $z^n = (z_1, z_2, \cdots, z_n)$ of a finite history. We call $z^n$ a sample of size $n$. We usually hope that this sample is *random*.

We will learn $P$ using the information available $(z^n)$. Note it is impossible to learn $P$ precisely because of the limited (i.e., finite) amount of information. This is related to the Law of Small Numbers.

We can, however, learn $P$ arbitrarily well as the sample size $n$ goes to $\infty$.

# Models

We are interested in the relationship between $Y$ and $X$—i.e., in explaining the behavior of $Y$ using $X$. A function of $X$, $f(X)$, is used to approximate $Y$. This function is called a model or a predictor for $Y$.

In practice, a linear function is most often used:

$$f(X) = \alpha + \beta X, \tag{224}$$

where $\alpha$ is the intercept and $\beta$ is the slope.

Obviously, any function $f(X)$ may be incorrect.

It is also important to note that while we observe $Y$ and $X$, the parameters $\alpha$ and $\beta$ are not directly observed. We must estimate $\alpha$ and $\beta$. But how do we do this?

This is where the concept of a loss function comes in.

# Loss Function

How well the model $f(X)$ will explain $Y$ is described by a what is called a "loss function." In general, there exists a discrepancy between $f(X)$ and $Y$. When $f(X) \neq Y$, a "loss" will occur. A function which tells us how big this "loss" will be, is called a loss function.

A loss function $l(Y, f(X))$ is a real-valued function that describes how well the model $f(X)$ can explain $Y$.

For example,

$$l(Y, f(X)) = \sum_i (Y_i - f(X_i))^p, \tag{225}$$

where $0 \leq p \leq \infty$, is a loss function.

These least square predictor is the loss function where $p = 2$. This is an arbitrary choice. But it is a choice with some nice properties.

A perfectly good loss function which has some nice properties (some of them better than least-squares) is:

$$l(Y, f(X)) = \mathrm{median}\ (Y_i - f(X_i))^2. \tag{226}$$

This is the "least-median of squares" loss function.

A another alternative which is sometimes discussed is:

$$l(Y, f(X)) = \sum_i abs(Y_i - f(X_i)). \tag{227}$$

This function doesn't really have much to offer relative to the previous two.

The expected loss is defined as $E[l(Y, f(X))]$. When $l(Y, f(X)) = (Y - f\,(X))^2$, the expected loss is mean square error (MSE).

We wish to minimized the loss function. The estimator which minimizes the MSE is called the least squares estimator.

# Regression Models

The concept of a regression function is related to material from conditional probability which we have already covered. In particular, the conditional mean $E(Y|X)$ is called the "regression function" of $Y$ on $X$.

A regression model consisting of $Y$, $X$ and $\epsilon$ satisfies the following property:

$$Y = E(Y|X) + \epsilon, \tag{228}$$

where the disturbance has the property $E(\epsilon|X) = 0$.

**Proof:**

$$Y = E(Y|X) + \epsilon \tag{229}$$

Then

$$\epsilon = Y - E(Y|X). \tag{230}$$

and

$$\begin{aligned} E(\epsilon|X) &= E\left[(Y - E(Y|X)|X\right] && (231)\\ &= E(Y|X) - E\left[E(Y|X)|X\right] && (232)\\ &= E(Y|X) - E(Y|X) = 0 && (233) \end{aligned}$$

Remarks:

1. The regression function $E(Y|X)$ is used to predict $Y$ from knowledge of $X$.

2. The term $\epsilon$ is called the “regression disturbance.” The fact $E(\epsilon|X) = 0$ implies that $\epsilon$ contains no systematic information of $X$ in predicted $Y$. In other words, all information of $X$ that is useful to predict $Y$ has been summarized by $E(Y|X)$.

# Properties

Under a set of assumptions, OLS is unbiased and efficient.

Under these assumptions, OLS is BLUE: the Best Linear Unbiased Estimator.

Assumptions A1—A4 taken together are sufficient for unbiasedness, and assumptions A1—A5 taken together are sufficient to prove efficiency.

# Classical Assumptions

A1. $Y_t = \sum_k X_{kt}\beta_k + \epsilon_t,$ $t = 1, 2, 3, \cdots, n \quad k = 1, \cdots, K,$ where $t$ indexes the observations and $k$ the variables. This is a very strong assumption.

A2. All of the $X$ variables are nonstochastic. They are fixed and there is no disturbance associated with them. Therefore, every $X$ can be moved outside of an expectation. This assumption is easy to weaken.

A3. There is no deterministic linear relationship between any of the $X$ variables. More precisely, the $k \times k$ matrix $\sum X_t X_t'$ is non-singular for every $n > k$.

A4. $E[\epsilon_t] = 0$ for every $t$, $t = 1, 2, 3, \cdots, n$. Since every $X$ is assumed to be nonstochastic (A2), (A4) implies that $E[X_t\epsilon_t] = 0$. (A4) *always* holds if there is an intercept and if (A1)–(A3) hold.

A5. The variance of the random error, $\epsilon$ is equal to a constant, $\sigma^2$, for all values of every $X$ (i.e., $var[\epsilon_t] = \sigma^2$), and $\epsilon$ is normally distributed. This assumption implies that The errors associated with any two observations are independent and identically distributed. This assumption can be significantly weakened, but the assumption of normality plays a key role.

# Simple Regression, Approval Example

Simple regression is a way to obtain the total effect of one variable on another. For example, if we estimate the following simple regression model:

$$\text{Approval}_t = \beta_0 + \beta_1 \text{Unemployment}_t + \epsilon_t, \tag{234}$$

we observe that $\beta_0 = 69.303$ and $\beta_1 = -2.280$. We may then interpret $\beta_1$ to be the total effect of one percent of unemployment on approval—it is the slope of unemployment's effect on approval.

What is the relationship between our estimate of $\widehat{\beta}_1$ and cov(approval,unemployment) and cor(approval,unemployment)?

Recall that The covariance between two random variables (X, Y) is defined to be:

$$\mathrm{cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right) \tag{235}$$

$$\mathrm{cor}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sqrt{(var(X)}\sqrt{(var(Y)}} \tag{236}$$

Note that:

- mean(approval) = 56.0579
- mean(unemployment rate) = 5.809825

- var(approval) = 143.6153
- var(unemployment rate) = 2.332944

- cov(approval,unemployment) = -5.318496
- cor(approval,unemployment) = -0.2905604

- $\dfrac{\mathrm{cov}(\mathrm{approval}, \mathrm{unemployment})}{\mathrm{var}(\mathrm{unemployment})}$ = -2.280

  And this is the same as our slope estimate: $\beta_1 = -2.280$!

# Derivation of Simple Least-Squares Parameter Estimates

In this section we explore how we obtain our estimates of $\alpha$ and $\beta$. This section requires some knowledge of calculus. It is **not** essential to understand this section to understand subsequent sections **nor** will you be tested on this material.

Chapter 11 of Wonnacott and Wonnacott (appendix 11-2). Also see Pindyck and Rubinfeld page 16ff. Alternatively, see judge, Hill et al. Part 2 or $many$ other books.

Our goal is to minimize $\sum_i^n (Y_i - \hat{Y}_i)^2$, where $\hat{Y}_i = \alpha + \beta X_i$ is the fitted value of $Y_i$ corresponding to a particular observation $X_i$.

We minimize the expression by taking the partial derivatives with respect to $\alpha$ and $\beta$, setting each equal to 0, and solving the resulting pair of simultaneous equations:

$$\nabla_\alpha \sum_i^n (Y_i - \alpha - \beta X_i)^2 = -2 \sum_i^n (Y_i - \alpha - \beta X_i) \tag{237}$$

$$\nabla_\beta \sum_i^n (Y_i - \alpha - \beta X_i)^2 = -2 \sum_i^n X_i (Y_i - \alpha - \beta X_i) \tag{238}$$

Equating these two derivatives to zero and dividing by $-2$, we obtain:

$$\sum_i^n (Y_i - \hat{\alpha} - \hat{\beta} X_i) = 0 \tag{239}$$

$$\sum_i^n X_i (Y_i - \hat{\alpha} - \hat{\beta} X_i) = 0 \tag{240}$$

$$\tag{241}$$

We may now rewrite these two equations to obtain what are called the **normal equations**:

$$\sum_i^n Y_i = \hat{\alpha} n + \hat{\beta} \sum_i^n X_i \tag{242}$$

$$\sum_i^n X_i Y_i = \hat{\alpha} \sum_i^n X_i + \hat{\beta} \sum_i^n X_i^2 \tag{243}$$

We can solve for $\hat{\alpha}$ and $\hat{\beta}$ simultaneously by multiplying Equation 242 by $\sum_i^n X_i$ and multiplying Equation 243 by $n$:

$$\sum_i^n X_i \sum_i^n Y_i = \hat{\alpha} n \sum_i^n X_i + \hat{\beta} (\sum_i^n X_i)^2 \tag{244}$$

$$n \sum_i^n X_i Y_i = \hat{\alpha} n \sum_i^n X_i + \hat{\beta} n \sum_i^n X_i^2 \tag{245}$$

Subtracting Equation 244 from Equation 245, we obtain

$$n \sum X_i Y_i - \sum_i^n X_i Y_i = \hat{\beta} \left[ n \sum_i^n X_i^2 - (\sum_i^n X_i)^2 \right] \tag{246}$$

It follows that:

$$\hat{\beta} = \frac{n \sum_i^n X_i Y_i - \sum_i^n X_i \sum_i^n Y_i}{n \sum_i^n X_i^2 - (\sum_i^n X_i)^2} \tag{247}$$

You are responsible for knowing that $\hat{\beta}$ is. You should also know that it results from minimizing the least squares loss function. You don't need to know the exact derivation.

$$\hat{\beta} = \frac{n\sum_i^n X_i Y_i - \sum_i^n X_i \sum_i^n Y_i}{n\sum_i^n X_i^2 - (\sum_i^n X_i)^2} \tag{248}$$

Note that $n\sum_i^n X_i^2 - (\sum_i^n X_i)^2 = n\sum_i^n (X_i - \bar{X})^2$. One may use this note which comes from the rules of summation (remember those?) to greatly simplify the equation which defines $\beta$ if the mean of X and the mean of Y is zero:

$$\begin{aligned} \hat{\beta} &= \frac{n\sum_i^n X_i Y_i}{n\sum_i^n X_i^2} \qquad (249) \\ &= \frac{cov(XY)}{var(X)} \qquad (250) \end{aligned}$$

Note the differences between the simple measure of covariance and the regression coefficient. There is a lot of intuition to be found examining the expression about, particularly Equation 250.

Given our solution of $\hat{\beta}$, we may obtain our solution for $\hat{\alpha}$ from Equation 242

$$\hat{\alpha} = \frac{\sum_{i}^{n} Y_i}{n} - \hat{\beta} \frac{\sum_{i}^{n} X_i}{n} \tag{251}$$

# The Differences Between Simple and Multiple Regression

As you read this discussion you may want to jump ahead and take a look at Figure 2 in these notes, which shows the direct and indirect relationships between inflation, unemployment and presidential approval.

One way to think about the difference between simple and multiple regression is to consider the difference between direct effects and indirect effects and the related concept of total effects.

The direct effect of, say, unemployment on presidential approval is the effect that unemployment has on approval if everything else is held constant.

The indirect effect of unemployment on approval is the direct effect that unemployment has on other variables times the direct effect these other variables have on approval.

Therefore, the indirect effect of unemployment on approval is the effect that unemployment has on approval when unemployment moves variables which themselves have a direct effect on approval.

The total effect of unemployment on approval is the direct effect unemployment has on approval plus the indirect effect it has on approval—i.e., the total effect is a sum of the direct and indirect effects.

Simple regression is a way to obtain the total effect of one variable on another.

It is often assumed that multiple regression is a way to obtain the direct effect of one variable on another.

But the use of multiple regression is much more complicated. The research issues we spoke about at the beginning of the term are very important to consider and are often overlooked.

## Derivation of Multiple Least-Squares Parameter Estimates

Without matrix algebra the derivation of multiple regression is rather tedious. In order to simply matters but to still communicate a sense of what is going on, we restrict our selves to a three parameter model:

Our goal is to minimize $\sum_i^n (Y_i - \hat{Y}_i)^2$, where $\hat{Y}_i = \hat{\alpha} + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i}$.

We can do this by calculating the partial derivaties with respect to the three unknown parameters $\alpha$, $\beta_1$, $\beta_2$, equating each to 0 and solving.

To further simplify the algebra we deviate the observed variables by their means. These mean deviated variables are denoted, as before, by $y_i$, $x_{1i}$ and $x_{2i}$.

Therefore,

$$\text{ESS} = \sum_i^n (y_i - \alpha + \beta_1 X_{1i} + \beta_2 X_{2i}) \tag{252}$$

Then,

$$\frac{\partial ESS}{\partial \beta_1} = \hat{\beta}_1 \sum_i^n x_{1i}^2 + \hat{\beta}_2 \sum_i^n x_{1i}x_{2i} - \sum_i^n x_{1i}y_i \tag{253}$$

$$\frac{\partial ESS}{\partial \beta_2} = \hat{\beta}_1 \sum_i^n x_{1i}x_{2i} + \hat{\beta}_2 \sum_i^n x_{2i}^2 - \sum_i^n x_{2i}y_i \tag{254}$$

These can be rewritten as:

$$\sum_i^n x_{1i}y_i = \hat{\beta}_1 \sum_i^n x_{1i}^2 + \hat{\beta}_2 \sum_i^n x_{1i}x_{2i} \tag{255}$$

$$\sum_i^n x_{2i}y_i = \hat{\beta}_1 \sum_i^n x_{1i}x_{2i} + \hat{\beta}_2 \sum_i^n x_{2i}^2 \tag{256}$$

To solve, we multiply Equation 255 by $\sum_i^n x_{2i}^2$ and Equation 256 by $\sum_i^n x_{1i}x_{2i}$ and substract the latter from the former.

Then,

$$\sum_i^n x_{1i}y_i \sum_i^n x_{x2i}^2 - \sum_i^n x_{2i}y_i \sum_i^n x_{1i}x_{2i} \quad = \quad \hat{\beta}_1 \left[ \sum_i^n x_{1i}^2 x_{2i}^2 - (\sum_i^n x_{1i}x_{2i})^2 \right]$$

Thus,

$$\hat{\beta}_1 \quad = \quad \frac{(\sum_i^n x_{1i}y_i)(\sum_i^n x_{x2i}^2) - (\sum_i^n x_{2i}y_i)(\sum_i^n x_{1i}x_{2i})}{(\sum_i^n x_{1i}^2)(\sum_i^n x_{2i}^2) - (\sum_i^n x_{1i}x_{2i})^2} \tag{257}$$

And

$$\hat{\beta}_2 \quad = \quad \frac{(\sum_i^n x_{2i}y_i)(\sum_i^n x_{x1i}^2) - (\sum_i^n x_{1i}y_i)(\sum_i^n x_{1i}x_{2i})}{(\sum_i^n x_{1i}^2)(\sum_i^n x_{2i}^2) - (\sum_i^n x_{1i}x_{2i})^2} \tag{258}$$

If we do the same for $\alpha$ we find that:

$$\hat{\alpha} \quad = \quad \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 X_2 \tag{259}$$

The equations for the estimates of $\widehat{\beta}_1$ and $\widehat{\beta}_2$ can be rewritten as:

$$\widehat{\beta}_1 = \frac{cov(X_{1i}, Y_i)\ var(X_{2i}) - cov(X_{2i}, Y_i)\ cov(X_{1i}, X_{2i})}{var(X_{1i})\ var(X_{2i}) - [cov(X_{1i}, X_{2i})]^2} \tag{260}$$

And,

$$\widehat{\beta}_2 = \frac{cov(X_{2i}, Y_i)\ var(X_{1i}) - cov(X_{1i}, Y_i)\ cov(X_{1i}, X_{2i})}{var(X_{1i})\ var(X_{2i}) - [cov(X_{1i}, X_{2i})]^2} \tag{261}$$

For an example using **R** code see `http://jsekhon.fas.harvard.edu/gov1000/mr1.R` and its output file `http://jsekhon.fas.harvard.edu/gov1000/mr1.Rout`.

# Multiple Regression, Approval Example

$$\text{Approval}_t = \alpha_0 + \alpha_1 \text{inflation}_t + \alpha_2 \text{Unemployment}_t + \epsilon_t, \qquad (262)$$

we observe that $\alpha_0 = 69.7785$, $\alpha_1 = -2.1394$ and $\alpha_2 = -0.9258$. It is clear that the slope associated with unemployment has *greatly* changed from $-2.280$ in the simple regression to $-0.9258$. In other words, in the simple regression model a 1 unit increase in the unemployment level decreases presidential approval by 2.28 units, but in the multiple regression model a 1 unit increase in the unemployment level decreases approval by only .9258 units. What's going on?

The multiple regression is giving us the effect of unemployment on approval *holding inflation constant*. In other words, the indirect effect of unemployment on approval which works through inflation is not taken into consideration. However, in the simple regression model the direct and indirect effect of unemployment—i.e., the total effect—is estimated.

Recall that the multiple regression model gives the direct effect, and the simple regression model the total effect which is equal to the direct effect plus the indirect effect.

We know the direct effect of unemployment on approval is the $\alpha_2$ coefficient in equation 262, $\alpha_2 = -0.9258$. The indirect effect is equal to the direct effect of inflation on approval (which is $\alpha_1 = -2.1394$) **times** the effect of unemployment on inflation, which we have not calculated.

Given that, aside from the intercept $\alpha_0$, we are only considering two independent variables in our multiple regression model (equation 262), the direct effect of unemployment on inflation can be found by estimating the following simple regression:

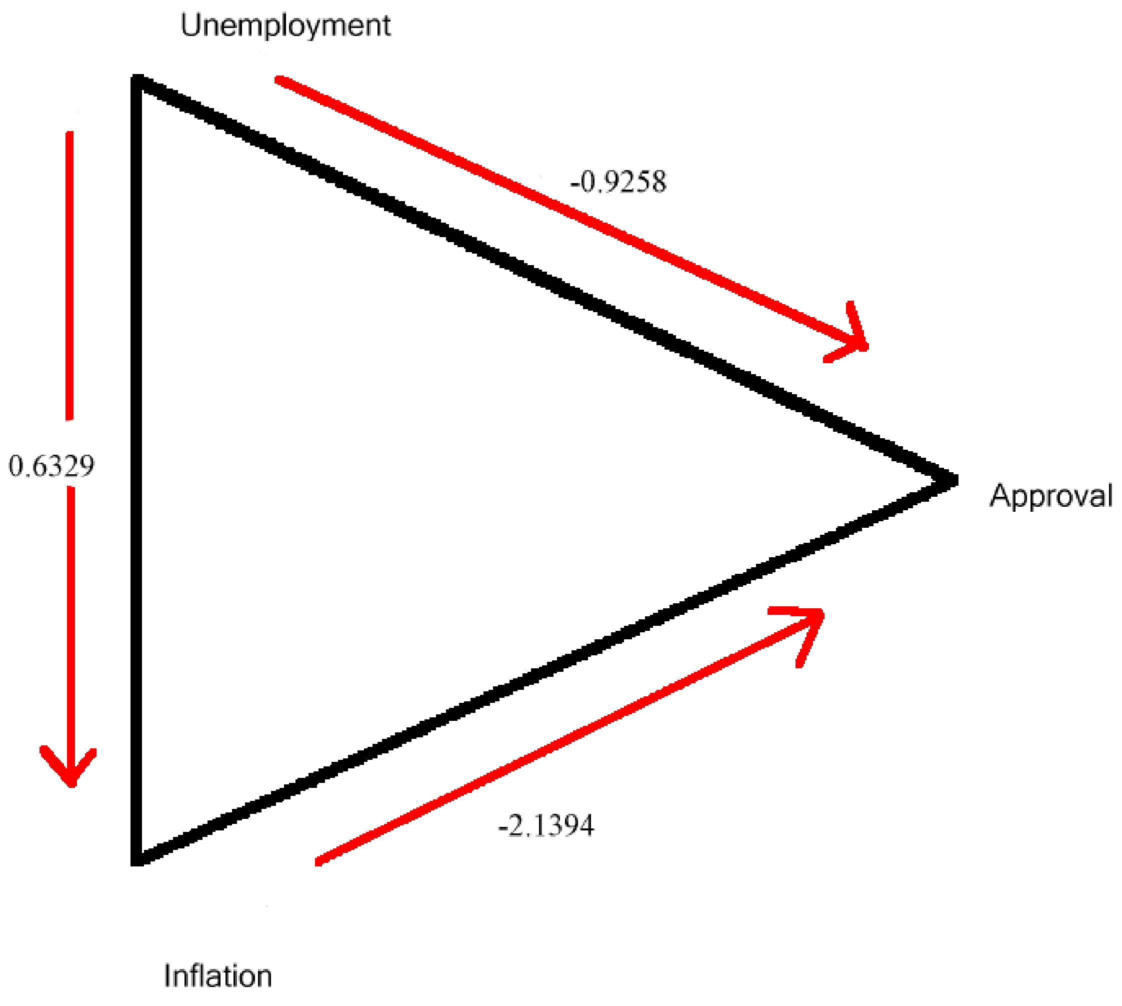
$$\text{Inflation}_t = \gamma_0 + \gamma_1 \text{Unemployment}_t + \epsilon_t. \tag{263}$$

If we estimate this model we find that the intercept, $\gamma_0$, equals $0.2224$ and the slope of the effect of unemployment on inflation, $\gamma_1$, equals 0.6329.

Therefore, the total effect of unemployment on approval must equal $-0.9258 + (0.6329 * -2.1394) = -2.280$. And this is exactly what we found when we estimated Equation 234. Small differences can arise between these two numbers because of degrees of freedom differences (remember that issue for our sampling discussion and estimating variances?).

Figure 2, on the next page, shows the direct and indirect relationships between inflation, unemployment and presidential approval.

Figure 2: The Relationship Between Unemployment and Presidential Approval Given Inflation



Direct and Indirect Effects of Unemployment on Approval

# Caution about the Direct/Indirect Distinction

The direct and indirect effects discussed here are accounting identities. They should not be taken to imply causal associations. They cannot be interpreted as causal without a set of assumptions being met.

This accounting occurs because of the geometry of least squares. The least squares estimator has mathematical proprieties which are just true and statistical properties which require assumptions.

An important mathematical property is that of orthogonal projection.

Another important mathematical property is that the mean residual will be zero when an intercept is included. But this isn't the same as unbiasedness. Unbiasedness is a statistical property which requires some additional assumptions.

The central statistical assumption is the correct specification assumption previously mentioned: $E(\epsilon|X) = 0$.

## What does the correct specification assumption imply?

It implies that $\epsilon$, the residual, contains no systematic information of $X$ in predicted $Y$. In other words, all information of $X$ that is useful to predict $Y$ has been summarized by $\mathrm{E}(Y|X)$.

# Hypothesis Testing

We frequently wish to test hypotheses about the regression models we estimate. Such hypotheses normally take the form of equality restrictions on some of the parameters. They are usually of the following form:

1. testing whether a single parameter takes on a certain value: $\beta_2 = 1$ or more commonly $\beta_2 = 0$.

2. whether two parameters are related in a specific way: $\beta_3 = 2\beta_4$

3. whether a nonlinear restriction holds: $\dfrac{\beta_1}{\beta_3} = \dfrac{\beta_2}{\beta_4}$

The hypothesis that the restriction or set of restriction to be tested does in fact hold is called the null hypothesis and is usually denoted $H_0$.

The model in which the restrictions do not hold is usually called the alternative hypothesis, or sometimes the maintained hypothesis, and is usually denoted $H_1$.

The terminology “maintained hypothesis” reflects the fact that in a statistical test only the null hypothesis $H_0$ is under test. Rejecting $H_0$ does not in any way oblige us to accept $H_1$, since it is not $H_1$ that we are testing.

Hypothesis tests usually involve the use of a test statistic.

A test statistic, such as T, is a random variable of which the probability distribution is known under the null hypothesis.

We then see how likely the observed value of T is to have occurred, according to that probability distribution.

If T is a number that could easily have occurred by chance, then we have no evidence against the null hypothesis $H_0$.

However, if it is a number that would occur by chance only rarely, we do have evidence against the null.

The size of a test is the probability that the test statistic will reject the null hypothesis when it is true—i.e., $P(H_0^R|H_0^T)$. The size of a test is also called its significance level.

We perform tests in the hope that they will reject the null hypothesis when it is false. Accordingly, the power of a test is of great interest. The power of a test statistic $T$ is the probability that $T$ will reject the null hypothesis when the latter is not true—i.e., $P(H_0^R|H_0^F)$.

The power of a consistent test increases with the sample size. As $n \to \infty$, power goes to 1. But the size of a test does not change when $n$ increases.

There are two types of errors:

1. Type I rejecting a null hypothesis when it is true—$P(H_0^R|H_0^T)$.
2. Type II accepting the null when it is false—$P(H_0^A|H_0^F)$

# How To Conduct These Tests

Please see section 12-2 in Wonnacott and Wonnacott.

# Key Assumptions of the Linear Model

There are three key assumptions that we need. All of them can be weakened, but that involves advanced material.

- 1. The relationship between Y and X is linear:

$$Y_i = \alpha + \beta X_i + \epsilon_i. \tag{264}$$

- 2. The X's are nonstochastic variables whose values are fixed.

- ⋆ a. The error term has zero expected value and constant variance for all observations; that is, $E(\epsilon_i) = 0$ and $E(\epsilon_i^2) = \sigma^2$.
  - ⋆ b. The random variables $\epsilon_i$ are statistically independent. Thus, $E(\epsilon_i \epsilon_j) = 0$, for all $i \neq j$.
  - ⋆ c. The error term is normally distributed. Thus $\epsilon \sim N(0, \sigma^2)$.

Not all of these assumptions are required for all actions involved with the linear model.

# Properties of Beta

Recall our simple regression model:

$$Y_i \quad = \alpha + \beta X_i + \epsilon_i. \tag{265}$$

Where we assume that $\epsilon \sim N(0, \sigma^2)$. This implies that $y$ is a random variable.

Note that our estimate for $\beta$ is defined as follows:

$$\hat{\beta} = \frac{n \sum_i^n X_i Y_i - \sum_i^n X_i \sum_i^n Y_i}{n \sum_i^n X_i^2 - (\sum_i^n X_i)^2} \tag{266}$$

Recall that this can be simplified considerably if the mean of X and the mean of Y is zero:

$$\hat{\beta} = \frac{n \sum_i^n X_i Y_i}{n \sum_i^n X_i^2} \tag{267}$$

Our definition of $\widehat{\beta}$ implies that $\widehat{\beta}$ is random because it depends on Y which is itself random.

Let $y_i = Y_i - \bar{Y}$ and $x_i = X_i - \bar{X}$.

$\widehat{\beta}$ has some properties under a set of assumptions we discuss later.

The first result is that:

$$E(\widehat{\beta}) = \beta, \tag{268}$$

so $\widehat{\beta}$ is an unbiased estimator of $\beta$.

The second result is that

$$Var(\widehat{\beta}) = \frac{\sigma^2}{\sum x_i^2}, \tag{269}$$

so that the variance of $\widehat{\beta}$ depends solely on the error variance ($\sigma^2$), the variance of the $X$'s, and the number of observations.

The mean and variance of the estimator of the intercept term are:

$$\begin{aligned} E(\hat{\alpha}) &= \alpha && (270) \\ Var(\hat{\alpha}) &= \sigma^2 \frac{\sum X_i^2}{n \sum (X_i - \bar{X})^2} && (271) \end{aligned}$$

Finally, the covariance between $\hat{\alpha}$ and $\hat{\beta}$ is given by:

$$Cov(\hat{\alpha}, \hat{\beta}) = \frac{-\bar{X}\sigma^2}{\sum x_i^2} \qquad (272)$$

# Estimating Sigma

In the previous equations the population variance, $\sigma^2$, appears. We need to estimate this.

We use the following sample estimate of the true variance $\sigma^2$.

Let $e_i = Y_i - \hat{Y}_i$, then

$$\hat{\sigma}^2 = \frac{\sum \hat{e}_i^2}{n-k}, \tag{273}$$

where $k$ is the number of parameters.

Note that another name for $\hat{\sigma}^2$ is simply $s^2$.

# Sampling Variance of Beta

With information about the means and variances of the least-squares estimators and their covariances, we are ready to discuss statistical testing of the linear model.

First, note that since $\widehat{\beta}$ is a weighted average of the $Y_i$'s, and that the $Y_i$'s are normally distributed, the estimator $\widehat{\beta}$ will be normally distributed.

This follows because a linear combination of independent normally distributed variables with also be normally distributed.

Even if the $Y$'s are not normally distributed, the distribution of $\widehat{\beta}$ can be shown to be asymptotically normal (under reasonable conditions) by appeal to the central-limit theorem—see Achen.

Roughly speaking the central-limit theorem states that the distribution of the sample mean of an independently distributed variable will tend toward normality as the sample size gets infinitely large.

To sum up:

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum x_i^2}\right) \tag{274}$$

$$\hat{\alpha} \sim N\left(\alpha, \sigma^2 \frac{\sum X_i^2}{n \sum x_i^2}\right) \tag{275}$$

$$\mathrm{cov}(\hat{\alpha}, \hat{\beta}) = -\frac{\bar{X}\sigma^2}{\sum x_i^2} \tag{276}$$

Notice that the variance of $\hat{\beta}$ varies directly with the variance of $\epsilon$.

Thus, *ceteris paribus*, we obtain more precise estimates of the slope when the variance of the error term is small.

On the other hand, the variance of $\hat{\beta}$ varies inversely with $\sum x_i^2$.

Thus, the larger the variance of $X_i$, the better you are likely to do in estimating $\beta$.

# Back to the Approval Example

Recall the Presidential Approval Figure. Let us consider five Presidential Approval Models.

Model 1:

$$\text{Approval}_t = \alpha + \beta_1 \text{Unemp}_t + \epsilon_t$$

Model 2:

$$\text{Approval}_t = \alpha + \beta_1 \text{Infl}_t + \epsilon_t$$

Model 3:

$$\text{Approval}_t = \alpha + \beta_1 \text{Unemp}_t + \beta_2 \text{Infl}_t + \epsilon_t$$

Model 4:

$$\text{Approval}_t = \alpha + \beta_1 \text{Unemp}_t + \beta_2 \text{Infl}_t + \beta_3 \text{Approval}_{t-1} + \epsilon_t$$

Model 5:

$$\begin{aligned}\text{Approval}_t = {} & \alpha + \beta_1 \text{Unemp}_t + \beta_2 \text{Infl}_t + \beta_3 \text{Approval}_{t-1} + \beta_4 \text{Eisenhower} \\ & + \beta_5 \text{Kennedy} + \beta_6 \text{Johnson} + \beta_7 \text{Nixon} \\ & + \beta_8 \text{Ford} + \beta_9 \text{Carter} + \beta_{10} \text{Regan} + \beta_{11} \text{Bush} + \epsilon_t\end{aligned}$$

# Dummy Variables

- Wonnacott and Wonnacott give examples of two and three category dummy variables in Chapter 14. We give an example of a variable with 9 categories: the nine administrations in our dataset.

- Variables used in regression models are usually continuous. But sometimes they are weaker measures.

- Dummy variables (which are also called indicator variables) take on the value of either 0 or 1. For example, the Clinton indicator variable would take on the value of 0 for those observations when Clinton was not in office and 1 when he was.

- Recall our previous discussion of random variables—see Random Variables I

- Dummy variables are usually used with nominal or ordinal variables.

- Recall that the nominal scale is the least powerful. It only maps the attributes of the object into a name. This mapping is simply a classification of entities. The only relationship is whether the measure of two attributes are the same or different.

- Note that even if quantification of a concept is generally thought to be impossible, if we can't even come up with a nominal measure, we know almost nothing empirical about the concept.

- Recall that the ordinal scale ranks the entities according to some criterion. An ordinal scale is more powerful than a nominal scale because it orders the entities. The ordering might be “greater than”’, “better than” or “more complex”.

- Some example of variables which are often turned into dummy variables:

1. education: high school, some college, college, post-graduate work
2. uninformed, informed
3. treatment, control
4. Watergate
5. voucher

# Presidential Administration as an Indicator Variable

- Since our dataset ranges from: January 1953 to June 2000, it includes data from 9 presidential administrations.

- We expect that the mean level of approval varies with presidential administration, and that our other independent variables do not explain this variation—see Presidential Approval Figure.

- Why can't we add a dummy variable for each of the 9 presidential administrations in our data to the following model:

$$\text{Approval}_t = \alpha + \beta_1 \text{Unemp}_t + \beta_2 \text{Infl}_t + \epsilon_t?$$

- We can't, because we would have the same variable in our model twice.

- Recall that the intercept takes on the value 1 for each and every observation—it is a constant.

- Hence, the intercept is a linear combination of all of the 9 presidential dummy variables. For every observation $i$:

$$\begin{aligned} 1 &= \text{Eisenhower} + \text{Kennedy} + \text{Johnson} + \text{Nixon} + \text{Ford} + \text{Carter} + \\ &\quad \text{Regan} + \text{Bush} + \text{Clinton} \\ &= \text{intercept} \end{aligned}$$

- Because of this issue, we have to leave out one of the nine administration indicator variables.

- Leaving one of the indicator variables out, allows us to interpret the remaining coefficients relative to the indicator left out.

# Results of the 5 Models (coefficients only)

| Variables | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
|---|---|---|---|---|---|
| Intercept | 69.30 | 65.03 | 69.78 | 9.86 | 13.76 |
| Unemp | −2.28 | | −0.93 | −0.13 | −0.60 |
| Infl | | −2.30 | −2.14 | −0.31 | −0.35 |
| $\text{App}_{t-1}$ | | | | 0.86 | 0.83 |
| Eisenhower | | | | | 1.30 |
| Kennedy | | | | | 2.50 |
| Johnson | | | | | −0.86 |
| Nixon | | | | | −0.28 |
| Ford | | | | | 2.70 |
| Carter | | | | | 1.14 |
| Regan | | | | | 1.93 |
| Bush, GHW | | | | | 1.83 |

Compare with the findings previously presented based on summary data.

# Standard Errors

- Since we know what the distribution of $\beta$ is, we can conduct hypothesis tests as outlined in the Hypothesis Testing section.

- We are usually interested in testing the hypothesis that $\beta = 0$. Hence, we write $H_0 : \beta = 0$.

- The alternative or maintained hypothesis is $H_1 : \beta \neq 0$.

- The test statistic which we use to conduct this test is often called the $t-$statistic.

- It is called the $t-$statistic because for small sample sizes we use the $t-$distribution (see Section 8-2 in Wonnacott and Wonnacott). But in this class we use the normal approximation.

- To test the null hypothesis that $\beta = \beta_0$ we calculate the $t-$statistic:

$$t = \frac{\beta - \beta_0}{s_{\hat{\beta}}},$$

where $s^2_{\hat{\beta}} = \frac{s^2}{\sum_{i=1}^{n} x_i^2}$, and $s^2 = \frac{\sum \hat{e}_i^2}{n-k}$.

- For more information about this formula, see the Estimation Sigma and Sampling Variance of Beta sections.

- When $n - k$ is larger than 30 and we assume that the null hypothesis is correct, the test statistic $t$ follows a normal distribution with mean zero and variance 1.

- If the test statistic $t$ is significantly larger than we should expect under the null hypothesis, we reject the null hypothesis.

- Since we know that 95% of the normal distribution falls within 1.96 standard units of its mean, we obtain a 95% confidence interval for $\beta$ by:

$$\hat{\beta} \quad \pm \quad 1.96 * s_{\hat{\beta}} \tag{277}$$

- Recall (from the Hypothesis Testing section) that the size of a test is the probability that the test statistic will reject the null hypothesis when it is true—i.e., $P(H_0^R | H_0^T)$. The size of a test is also called its significance level.

- Hence, if we decide that the size of our test should be 0.05, we can reject the null hypothesis of $\beta = \beta_0$, if the absolute value of $t$ is $\geq 1.96$

- There is nothing magical about a 95% confidence interval, or a hypothesis test of power 0.05. We could easily be interested in a test of size .1. In which case, our critical value is no longer 1.96, but 1.645.

# Full Results of the 5 Models A

| Variables | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
|---|---|---|---|---|---|
| Intercept | $69.30^{***}$ | $65.03^{***}$ | $69.78^{***}$ | $9.86^{***}$ | $13.76^{***}$ |
| | (1.89) | (0.70) | (1.63) | (1.66) | (2.06) |
| Unemp | $-2.28^{***}$ | | $-0.93^{**}$ | $-0.13$ | $-0.60^{**}$ |
| | (0.32) | | (0.29) | (0.14) | (0.21) |
| Infl | | $-2.30^{***}$ | $-2.14^{***}$ | $-0.31^{***}$ | $-0.35^{***}$ |
| | | (0.14) | (0.15) | (0.09) | (0.13) |
| $\text{App}_{t-1}$ | | | | $0.86^{***}$ | $0.83^{***}$ |
| | | | | (0.02) | (0.02) |

Significance Codes: if p-value $\approx 0$, $^{***}$; if p-value $< 0.001$, $^{***}$; if p-value $< 0.001$, $^{**}$; if p-value $< 0.01$, $^{*}$; if p-value $< 0.05$, $^{\$}$.

Standard errors in parenthesizes.

# Full Results of the 5 Models B

| Variables | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
|---|---|---|---|---|---|
| Eisenhower | | | | | $1.30^{\$}$ |
| | | | | | (0.75) |
| Kennedy | | | | | $2.50^{*}$ |
| | | | | | (1.05) |
| Johnson | | | | | −0.86 |
| | | | | | (0.84) |
| Nixon | | | | | −0.28 |
| | | | | | (0.87) |
| Ford | | | | | $2.70^{*}$ |
| | | | | | (1.36) |
| Carter | | | | | 1.14 |
| | | | | | (1.24) |
| Regan | | | | | $1.93^{*}$ |
| | | | | | (0.90) |
| Bush, GHW | | | | | $1.83^{\$}$ |
| | | | | | (0.96) |

# How to Chose a Model?

- There are many criterion by which we can select a model.
- A natural choice would be to choose models which have a small sum of squared errors:

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2. \tag{278}$$

- Another way to consider the sum of squared errors is the root mean squared error:

$$\text{RMSE} = \sqrt{\frac{1}{n-k}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \tag{279}$$

- The RMSE is the standard error of the regression. And is related to prediction error.

- A natural way to consider the sum of squared errors is $R^2$.

# $R^2$ and $\bar{R}^2$

For each observation, we can break down the difference between $Y_i$ and its mean $\bar{Y}$ as follows:

$$(Y_i - \bar{Y}) = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y}) \tag{280}$$

Squaring both sides and summing over all observations (1 to $n$), we obtain

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\bar{Y} - \hat{Y}_i)^2 \tag{281}$$

$$\begin{aligned}\text{Variation in } Y &= \text{Residual variation} + \text{Explained variation}\\ \text{Total sum of squares} &= \text{Error sum of squares} + \text{Regression sum of squares}\\ \text{TSS} &= \text{ESS} + \text{RSS}\end{aligned}$$

$R^2$ is defined as:

$$\begin{aligned}
R^2 &= \frac{\mathrm{RSS}}{\mathrm{TSS}} && (282)\\
&= \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} && (283)\\
&= 1 - \frac{\sum_{i=1}^{n}\hat{\epsilon}_i^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} && (284)
\end{aligned}$$

$R^2$ measures the proportion of the variation in $Y$ which is “explained” by the multiple regression equation.

$R^2$ is often used as a goodness-of-fit statistic and to compare the validity of regression results under alternative specifications of the independent variables in the model.

There are some problems with $\mathrm{R}^2$:

- $\mathrm{R}^2$ is sensitive to the number of independent variables included in the regression model. The addition of more independent variables to the regression equation can never lower $\mathrm{R}^2$ and is likely to raise it. This occurs because the addition of a new explanatory variables does not alter TSS but is likely to increase RSS.

- $\mathrm{R}^2$ does not account for the number of degrees of freedom.

- One idea is to use variances, not variations, thus (in part) accounting for the number of independent variables in the model. The correction is based on the fact that variance equals variation divided by degrees of freedom.

$\bar{R}^2$ or corrected $R^2$ is defined as

$$\bar{R}^2 = 1 - \frac{\widehat{var}(\epsilon)}{\widehat{var}(Y)} \tag{285}$$

Where

$$\widehat{var}(\epsilon) = s^2 \tag{286}$$

$$= \frac{\sum_{i=1}^{n} \hat{\epsilon}_i^2}{n-k} \tag{287}$$

$$\widehat{var}(Y) = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n-1} \tag{288}$$

It is important to note that:

$$R^2 = 1 - \frac{s^2}{\widehat{var}(Y)} \frac{n-k}{n-1} \tag{289}$$

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k} \tag{290}$$

Examination of Equation 290 reveals that:

1. If $k = 1$, then $R^2 = \bar{R}^2$.

2. If $k$ is greater than 1, then $R^2 \geq \bar{R}^2$.

3. $\bar{R}^2$ can be negative.

It is very important NOT to use $R^2$ to compare the validity of alternative regression models when the dependent variable varies from regression to regression.

$R^2$ can be misleading in part because it chooses models which are too large. This is also true of $\bar{R}^2$ even though $\bar{R}^2$ is obviously likely to choose smaller models than $R^2$.

# Fit Summaries of the 5 Model

| Variables | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
|---|---|---|---|---|---|
| RMSE | 11.48 | 9.95 | 9.87 | 4.93 | 4.90 |
| $R^2$ | 0.084 | 0.311 | 0.324 | 0.831 | 0.836 |
| $\bar{R}^2$ | 0.083 | 0.310 | 0.321 | 0.831 | 0.833 |
| F$-$statistic | $52.38_{1,568}$ | $256.9_{1,568}$ | $135.8_{2,567}$ | $931.5_{3,566}$ | $259.3_{11,558}$ |
| p-value of F | 1e-12 | 2e-16 | 2e-16 | 2e-16 | 2e-16 |

# $F$ Statistic

The $F$ statistic calculated by most regression programs can be used in the multiple regression model to test the significance of the $R^2$ statistic.

The $F$ statistic with $k-1$ and $n-k$ degrees of freedom allows us to test the hypothesis that none of the explanatory variables helps explain the variation of $Y$ about its mean.

The $F$ statistic tests the joint hypothesis that $\beta_2 = \beta_3 = \cdots = \beta_k = 0$.

It can be shown that:

$$F_{k-1,n-k} = \frac{RSS}{ESS}\,\frac{n-k}{k-1} \tag{291}$$

$$= \frac{R^2}{1-R^2}\,\frac{n-k}{k-1} \tag{292}$$

If the null hypothesis is true, then we would expect RSS, $R^2$, and therefore $F$ to be close to 0. Thus, a large value of the $F$ statistic is a rationale for rejecting the null hypothesis.

An $F$ statistic not significantly different from 0 lets us conclude that the explanatory variables do little to explain the variation of $Y$ about its mean.

The $F$ test of the significance of a regression equation may allow for rejection of the null hypothesis even though none of the regression coefficients are found to be significant according to individual $t$ tests.

This situation may arise if the independent variables are highly correlated with each other.

# The $F$ Distribution

The $F$ distribution is named in honor of the English statistician Sir Ronald Fisher (1890-1962).

The $F$ distribution has a skewed shape and ranges in value form 0 to infinity.

Let $X$ be the sum of the squares of $n_1$ independently distributed normal random variables (with mean 0).

Let $Z$ be the sum of the squares of $n_2$ independently distributed normal random variables (with mean 0).

Under the null hypothesis that the variance of $X$ and $Z$ are equal, the ratio $\frac{X/n_1}{Z/n_2}$ is distributed according to a $F$ distribution with $n_1$ and $n_2$ degrees of freedom.

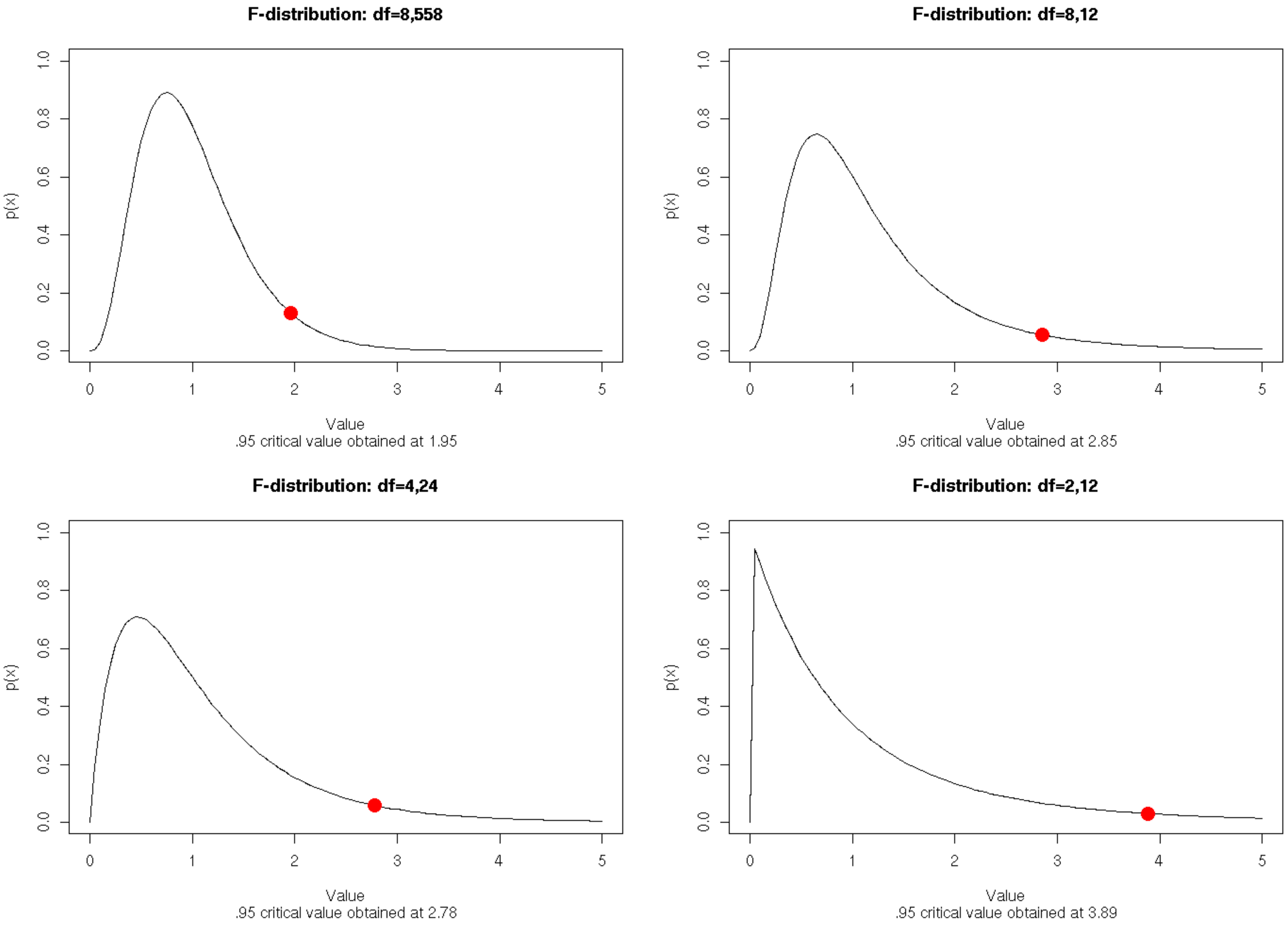
See the F-distribution Figure which plots a number of F-distributions, with differing degrees of freedom.

Note that unlike the normal distribution, the shape of the F-distribution radically changes depending on its parameters.

**R** has various functions associated with the $F-$distribution, which parallel its functions related to the normal distribution: `rf`, `df`, `df`, `qf`.

See page 328 of W&W for more information.

# Figure 3: Graphs of the F−distribution



F-distribution: df=8,558

F-distribution: df=8,12

F-distribution: df=4,24

F-distribution: df=2,12

# The General $F$ Statistic

Consider the following unrestricted multiple regression model:

$$Y = \beta_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon. \tag{293}$$

We wish to test whether a subset $q$ of the regression coefficients is jointly equal to zero.

$$Y = \beta_1 + \beta_2 X_2 + \cdots + \beta_{k-q} X_{k-q} + \beta_{k-q+1} X_{k-q+1} + \cdots + \beta_k X_k + \epsilon. \tag{294}$$

The restricted model, which would be correct if all of the last $q$ variables really where equal to 0, is:

$$Y = \beta_1 + \beta_2 X_2 + \cdots + \beta_{k-q} X_{k-q} + \epsilon \tag{295}$$

The null hypothesis is that $\beta_{k-q+1} + \cdots + \beta_k = 0$.

The test is based on the error (or residual) sum of squares associated with the unrestricted and restricted models.

The relevant $F$ statistic is equal to:

$$F_{q,n-k} = \frac{(ESS_R - ESS_{UR})/q}{ESS_{UR}/(n-k)} \tag{296}$$

$$= \frac{\left[\sum_{i=1}^{n}(Y_{R,i} - \hat{Y}_R)^2 - \sum_{i=1}^{n}(Y_{UR,i} - \hat{Y}_{UR})^2\right]/q}{\left[\sum_{i=1}^{n}(Y_{UR,i} - \hat{Y}_{UR})^2\right]/(n-k)} \tag{297}$$

The $F$ statistic can be written in a simpler form:

$$F_{q,n-k} = \frac{(R^2_{UR} - R^2_R)/q}{(1 - R^2_{UR})/(n-k)} \tag{298}$$

If we seek to determine if the extra variables in Model 5 (the model with the lags and indicator variables) significantly improve the fit when compared to the simpler Model 4 (the model with only the lag), we do the following:

q is equal to the degrees of of freedom of Model 4 minus the degrees of freedom of Model 5: $q = 566 - 558 = 8$. Model 5 has 8 more parameters than Model 4.

And for Model 5, $n - k = 570 - 12 = 558$.

$$\begin{aligned} F_{q,n-k} &= \frac{(R^2_{UR} - R^2_R)/q}{(1 - R^2_{UR})/(n-k)} \\ &= \frac{(.836 - .831)/8}{(1 - .836)/(558)} \\ &= 2.126524 \end{aligned}$$

Which translates into a p-value of 0.0317. What at conventional test levels, is significant.