

# Econometrics 461

---

LECTURE NOTES

**Craig Schulman**

TEXAS A&M UNIVERSITY | DEPARTMENT OF ECONOMICS

4228 TAMU | COLLEGE STATION, TX 77843



Updated January 16, 2023

Econometrics 461 | Lecture Notes

All rights reserved.

No part of this publication may be reproduced, copied, or transmitted in any form without the prior permission of the author. Requests for permission should be directed to:

Craig Schulman  
Texas A&M University | Department Of Economics  
4228 TAMU | College Station, TX 77843

**ECONOMETRICS 461 | LECTURE NOTES**  
**CRAIG SCHULMAN**

**TABLE OF CONTENTS**

Section 0. Introduction .....	1
Section 1. Data and Data Descriptions.....	2
<b>A. Variables and Values in Lists.....</b>	<b>2</b>
<b>B. Random Variables, Populations and Random Samples.....</b>	<b>3</b>
<b>C. Scales of Measurement – Data Types .....</b>	<b>4</b>
<b>D. Numeric Scales.....</b>	<b>4</b>
<b>E. Graphs to Describe Data.....</b>	<b>5</b>
<b>F. Frequency Distributions for Numerical Data.....</b>	<b>10</b>
<b>G. Example Problems.....</b>	<b>16</b>
Section 2. Measures of Central Tendency, Variability and Co-Movement .....	18
<b>A. Measures of Central Tendency .....</b>	<b>18</b>
<b>B. Distance Measures of Variability.....</b>	<b>23</b>
<b>C. Variance and Standard Deviation .....</b>	<b>25</b>
<b>D. Measures of Co-Movement.....</b>	<b>27</b>
<b>E. Example Problems.....</b>	<b>30</b>
Section 3. Probability and the Normal Probability Distribution.....	33
<b>A. Probability and Probability Distribution Functions .....</b>	<b>33</b>
<b>B. Binomial Distribution.....</b>	<b>36</b>
<b>C. Poisson Distribution .....</b>	<b>38</b>
<b>D. Hypergeometric Distribution .....</b>	<b>39</b>
<b>E. Exponential Distribution.....</b>	<b>40</b>
<b>F. The Normal Distribution .....</b>	<b>42</b>
<b>G. Normal Approximation to the Binomial Distribution.....</b>	<b>46</b>
<b>H. Sampling Distributions.....</b>	<b>47</b>
<b>I. Normal Probability Plots.....</b>	<b>49</b>
<b>J. Example Problems.....</b>	<b>52</b>
Section 4. Estimation and Hypothesis Testing .....	55
<b>A. Sampling Distributions.....</b>	<b>55</b>

<b>B. Confidence Intervals: Single Sample</b> .....	56
<b>C. Hypothesis Testing: One Sample Tests</b> .....	58
<b>D. Confidence Intervals: Two Sample</b> .....	62
<b>E. Hypothesis Testing: Two Sample Tests</b> .....	66
<b>F. P-Values</b> .....	68
<b>G. Example Problems</b> .....	69
Section 5. Analysis of Variance.....	73
<b>A. One-Way Analysis of Variance</b> .....	73
<b>B. Two-Way Analysis of Variance</b> .....	77
Section 6. Linear Regression.....	83
<b>A. The Bivariate Regression Model</b> .....	83
<b>B. Correlation Analysis</b> .....	87
<b>C. Bivariate Regression with Multiple Groups</b> .....	89
<b>D. Alternative Tests for Differences in Sub-Sample Correlations</b> .....	91
<b>E. ANOVA as Multiple Regression</b> .....	93
Section 7. Nonparametric Statistical Tests.....	97
<b>A. Introduction</b> .....	97
<b>B. Chi-Square Goodness-of-Fit Test</b> .....	97
<b>C. Chi-Square Test of Homogeneity/Independence</b> .....	99
Section 8. Introduction to Time Series Analysis .....	102
Section 9. Classical Probability Appendix.....	103
<b>A. Basic Definitions</b> .....	103
<b>B. Conditional Probabilities</b> .....	104
<b>C. Bayes Theorem</b> .....	107

## Section 0. INTRODUCTION

---

These notes are intended to accompany lectures for Econometrics 461: Introduction to Economic Data Analysis. Economic Data Analysis concerns data (numeric and categorical), statistics, and statistical inference. My approach to these topics is decidedly applied in nature. Statistics necessarily involves a certain amount of mathematics. This course will cover the formal mathematical models of statistics and statistical inference. However, the focus will not be on derivations and proofs, but rather how these models and methods are used in practice. Throughout these notes you will find “how-to” examples for using Microsoft Excel®. For better or worse, Excel has become the ubiquitous standard software tool for many types of data analysis, simulations, and data reporting in business and applied economics. There are certainly more powerful software tools available (e.g., SAS, Stata, SQL based software, SAP, R, Python, etc.) but many entry level jobs open to Economics majors will require the use of Excel in some form, which is why I emphasize its use.

You should find in these notes all the essential concepts and mathematical formulas that will be covered during the semester. In addition, I have included example problems and solutions in each substantive section like what you will see in your homework sets and on exams.

**Section 1. DATA AND DATA DESCRIPTIONS**

“In God we trust. All others must use data.”

**A. Variables and Values in Lists**

Economic data analysis is fundamentally about data. Webster defines data as follows:

Data: Information, especially information organized for analysis or used as the basis for decision making.<sup>1</sup>

This definition is obviously quite broad as data/information can take many forms: Sales levels measured as quantities or in dollar terms, the price of a share of a company’s stock, responses to a questionnaire regarding voting preferences or satisfaction levels, and simple categorizations of items such as the make of an automobile are all examples of data that can be used for analysis or as a basis for decision making. How such data is organized can greatly simplify the process of analysis. Shown below is an example of county level labor force data organized in an Excel workbook from the U.S. Department of Labor.

LAUS Code	State FIPS Code	County FIPS Code	County Name/State Abbreviation	Year	Labor Force	Employed	Unemployed	Unemployment Rate (%)
CN0100100000000	01	001	Autauga County, AL	2015	25,308	23,981	1,327	5.2
CN0100300000000	01	003	Baldwin County, AL	2015	87,316	82,525	4,791	5.5
CN0100500000000	01	005	Barbour County, AL	2015	8,625	7,854	771	8.9
CN0100700000000	01	007	Bibb County, AL	2015	8,490	7,929	561	6.6
CN0100900000000	01	009	Blount County, AL	2015	24,352	23,036	1,316	5.4
CN0101100000000	01	011	Bullock County, AL	2015	4,773	4,402	371	7.8
CN0101300000000	01	013	Butler County, AL	2015	9,142	8,452	690	7.5
CN0101500000000	01	015	Calhoun County, AL	2015	46,051	42,837	3,214	7.0

These data are organized as variables in columns (such as ‘County Name/State Abbreviation’) with the values following in rows as a list. Data could also be organized with variables in rows and values in successive columns such as the following extract of annual balance sheet information for Wal-Mart Stores.

<sup>1</sup> Webster’s II New Riverside University Dictionary, Houghton Mifflin Company, 1984.

Report Date	01/31/2002	01/31/2003	01/31/2004	01/31/2005	01/31/2006	01/31/2007
Currency	USD	USD	USD	USD	USD	USD
Audit Status	Not Qualified	Not Qualified	Not Qualified	Not Qualified	Not Qualified	Not Qualified
Consolidated	Yes	Yes	Yes	Yes	Yes	Yes
Scale	Thousands	Thousands	Thousands	Thousands	Thousands	Thousands
Cash & cash equivalents	2,161,000	2,758,000	5,199,000	5,488,000	6,414,000	7,373,000
Receivables, net	2,000,000	2,108,000	1,254,000	1,715,000	2,662,000	2,840,000
Inventories at replacement cost	22,749,000	25,056,000	0	0	0	0
Inventories less: LIFO reserve	135,000	165,000	0	0	0	0
Inventories at LIFO cost	22,614,000	24,891,000	0	0	0	0
Inventories	0	0	26,612,000	29,447,000	32,191,000	33,685,000
Prepaid expenses & other current assets	1,471,000	726,000	1,356,000	1,841,000	2,557,000	2,690,000
Current assets of discontinued operations	0	0	0	0	0	0
Total current assets	28,246,000	30,483,000	34,421,000	38,491,000	43,824,000	46,588,000
Land	10,241,000	11,228,000	12,699,000	14,472,000	16,643,000	18,612,000

## B. Random Variables, Populations and Random Samples

The labor force and Wal-Mart balance sheet data shown above help introduce the notion of random variables, populations, and samples. Consider the labor force data. The number of people in the labor force, the number employed, unemployed and the unemployment rate are each variables that might be analyzed. A random variable takes on different values with certain probabilities. For any particular variable, the **population** is the collection of all possible outcomes for that variable. A **sample** is a subset of observations of a particular variable taken from the population. A **random sample** is a sample such that each observation in the population is equally likely of being selected, the selection of any one observation does not influence the selection of any other, and every possible sample of a particular size is equally likely of being selected. Because populations are usually quite large and there are costs associated with data collection and analysis, we usually deal with samples. The purpose of using random samples is to try to prevent drawing incorrect inferences about the nature of a population based on the sample being analyzed.

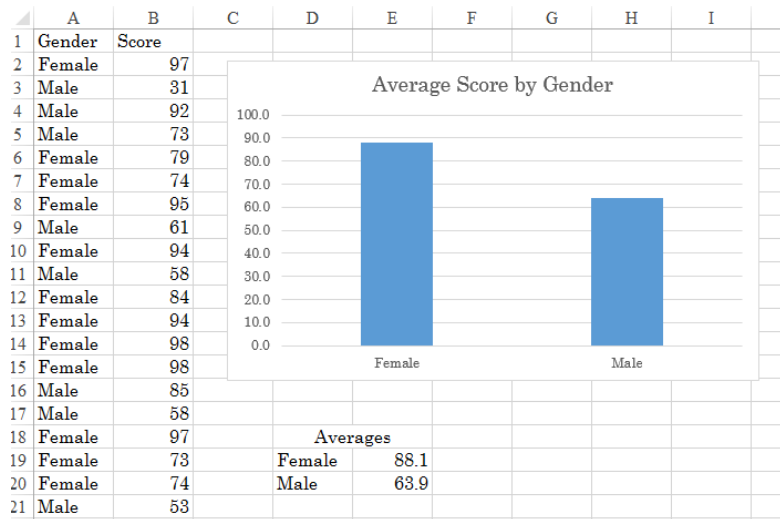
A numerical measure that describes some aspect of a population is called a **parameter**. A numerical measure that describes some aspect of a sample is called a **statistic** or **sample statistic**.



### C. Scales of Measurement – Data Types

As illustrated by the labor force data example, data can take on different types. The first four columns of the labor force data contain what are called **Category Variables** that merely represent mutually exclusive groups. While the values of the State and County “FIPS” code in these data are numeric, the actual values are arbitrarily assigned and have no meaning beyond identifying the specific group to which the data belong. Another example is the following sample of exam scores with a category variable for the Gender of the student.

Also shown in this example is an Excel “Column” chart of the average score by gender. The category variable is arranged on the horizontal axis and the numerical values for the average are shown on the vertical axis. By way of contrast, in an Excel “Bar” chart, categories are displayed on the vertical axis and the numerical values are shown on the horizontal axis.



### D. Numeric Scales

While there is only one type of scale for category variables, there are three types of numeric scales:

**Ordinal Scales:** A variable measured as an ordinal scale are often rankings that convey the order of what value comes first (smallest or largest), second, third, and so on. For example, in a race involving 6 runners, a list of which runners came in first, second, third, etc. would be ordered from the smallest (fastest) time to the largest. However, the order of the finish (the ranking) would not tell you how much faster first place was compared to second, for that you would need to compare the actual finish times. Responses to satisfaction surveys are another example of an ordinal variable.

**Interval Scales:** Variables measured on an interval scale provide information on the rank and *difference* between measurements from an *arbitrary* zero point on the scale. Temperature measured in °F is a classic example. If you are told that the temperature in Dallas is 90°F and that the temperature in Chicago is 30°F, you know that the difference is 60°F. However, that does not mean that is 3 times warmer in Dallas than it is in Chicago because the zero point on °F is an arbitrary designation.

**Ratio Scales:** Variables measured on a ratio scale provide information on the rank and distance from a natural zero point, with the ratio of two values having meaning. A person that is 50 years old is twice the age of someone who is 25. If per capita personal income in Mississippi is \$35,000 and in Texas it is \$47,000 personal income in Texas is about 34% higher than that in Mississippi.

### E. Graphs to Describe Data

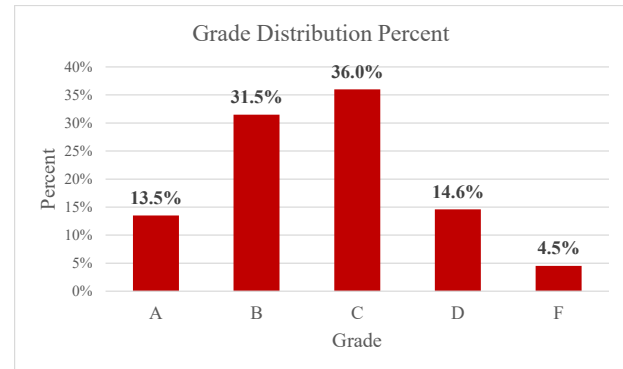
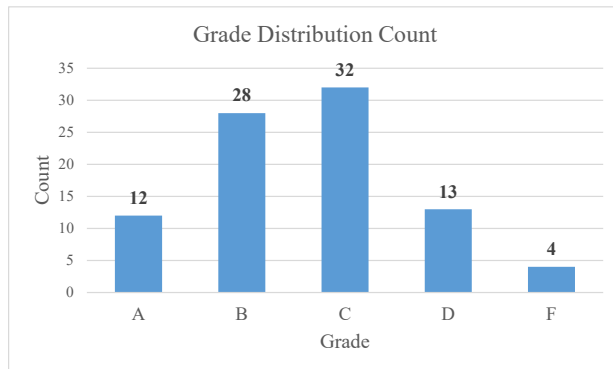
Categorical variables can be described with frequency tables and graphs such as bar charts and pie charts. These types of graphs are quite simple to construct and can provide a very powerful visual description of the underlying data.

A **frequency distribution** tabulates counts (or the number of observations) of all the possible outcomes for the variable being studied. For example, a distribution of student grades could be shown as in the following table and bar chart:

The first column shows the possible outcomes – grades – called classes or groups. The second column shows the number of students (count) that earned each grade. And the third column shows the **relative frequency** expressed as a percentage of the total number of outcomes.

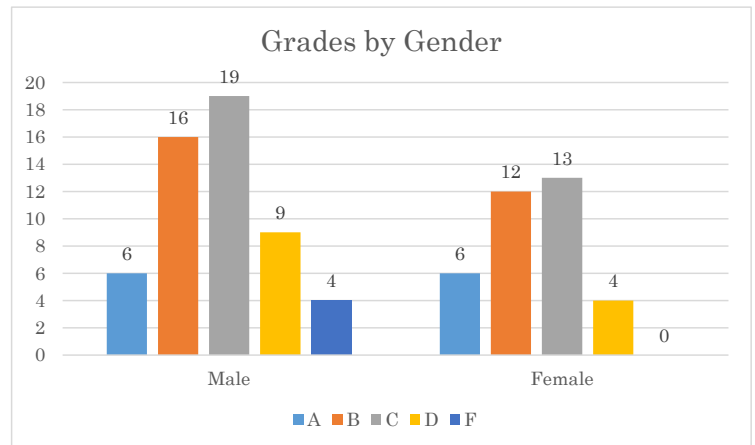
Grade	Students	Percent
A	12	13.5%
B	28	31.5%
C	32	36.0%
D	13	14.6%
F	4	4.5%
<b>Totals</b>	<b>89</b>	<b>100.0%</b>

A bar chart of the distribution shows either the frequency counts or the relative frequency percentage with the classes (grades) displayed on the horizontal axis.



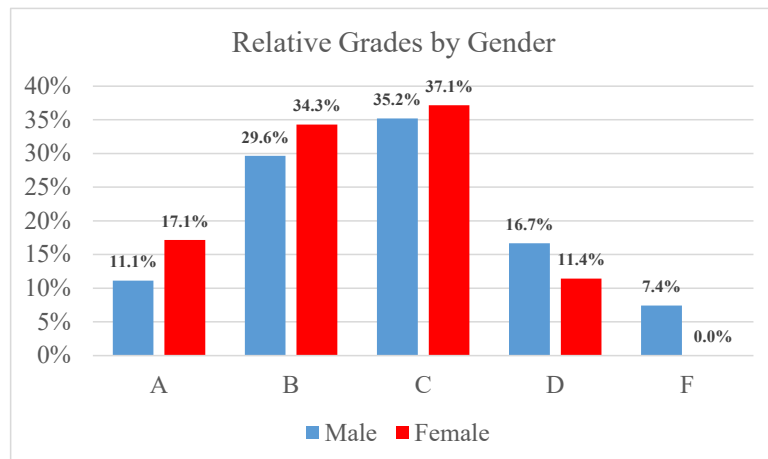
Quite often, we wish to describe relationships among multiple dimensions of categorical variables. Voting patterns by income categories, education level, gender, and race, for example. A **contingency table** (also known as a cross table or crosstab) can be used to summarize outcomes for two or more categorical variables. Using the grade distribution example from above, we can show differences in how grades are distributed by gender:

Grade	Male	Female	Total
A	6	6	12
B	16	12	28
C	19	13	32
D	9	4	13
F	4	0	4
<b>Totals</b>	<b>54</b>	<b>35</b>	<b>89</b>

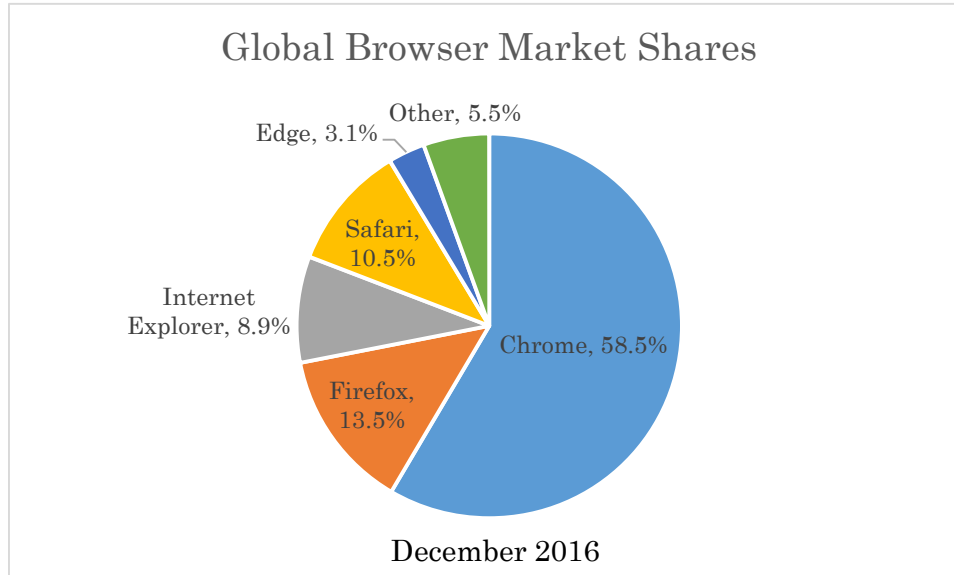


Converting the grade frequency counts in the table above to percentages, we can also present these data as relative frequencies to make direct comparisons easier:

Grade	Male	Female
A	11.1%	17.1%
B	29.6%	34.3%
C	35.2%	37.1%
D	16.7%	11.4%
F	7.4%	0.0%

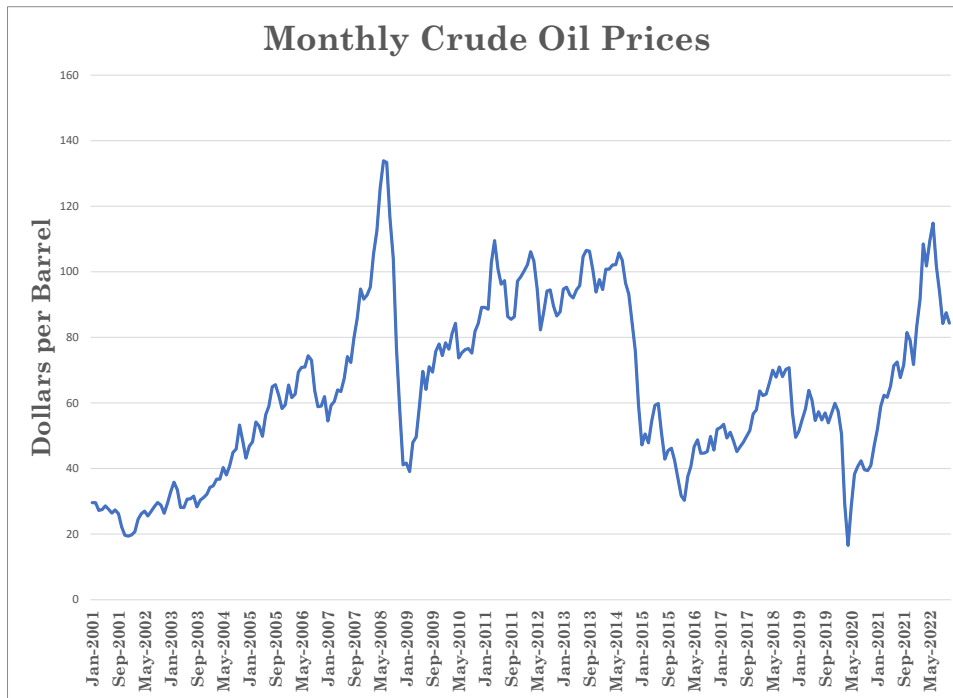


If we want to describe the proportion of frequencies in each category, a **pie chart** is a useful and visually appealing tool for this purpose. The following chart shows the global market share in December 2016 for different internet browsing software as reported by <http://gs.statcounter.com/>

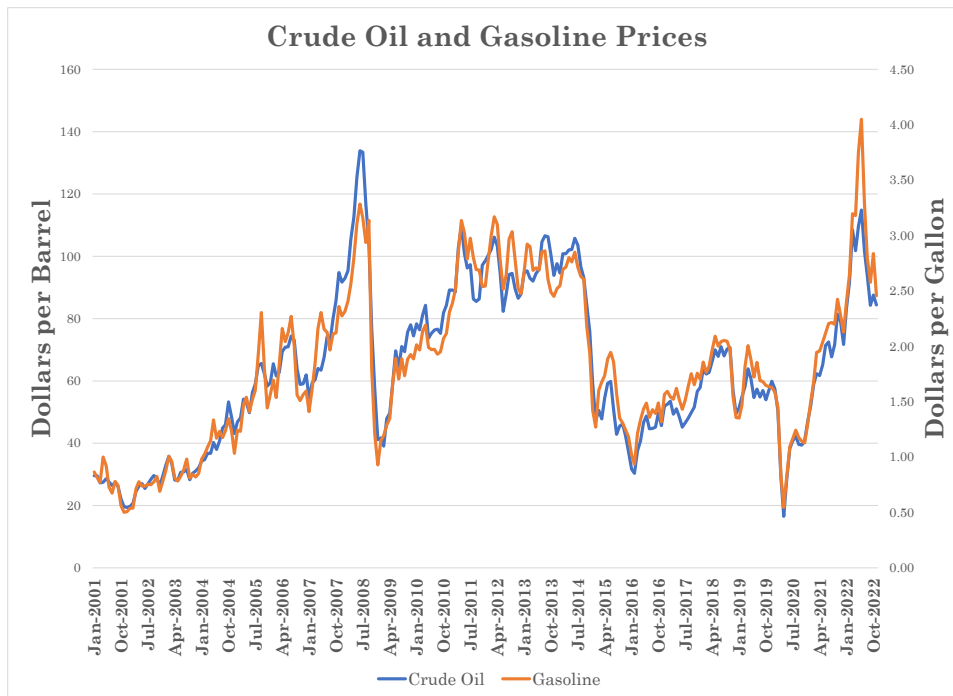
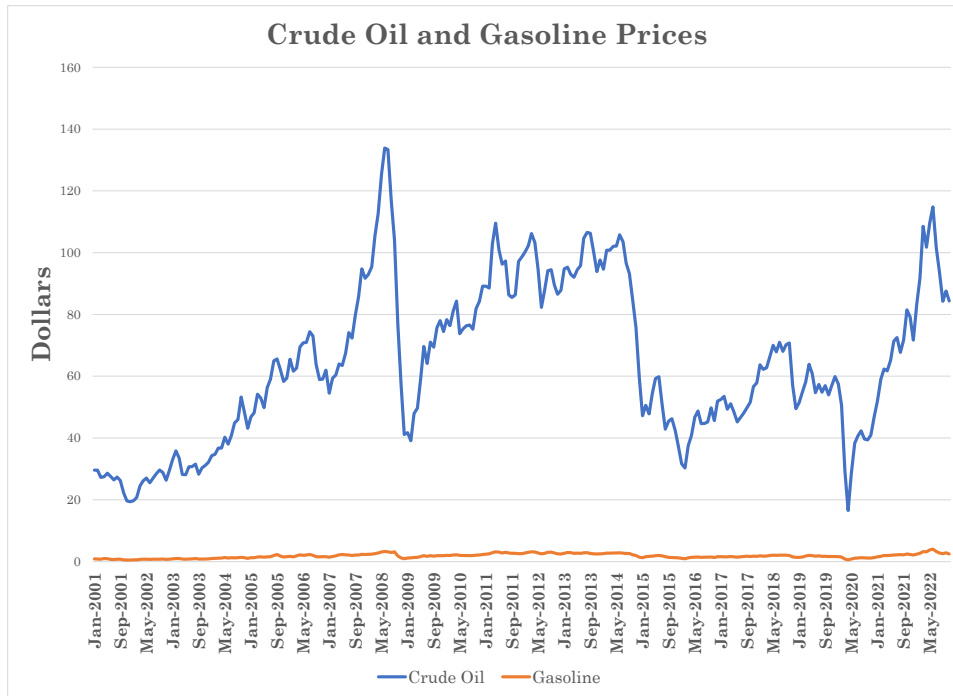


### Time Series Data

The labor force and grades data shown above are known as **cross-sectional** data. There is no natural order to the data, or any particular sub-sample taken from these data. Another type of data that is very common in economics and business is **time-series** data where variables are observed at different intervals in time and thus naturally ordered. In the Wal-Mart balance sheet data shown above, the value for the variable “Cash and cash equivalents” in 2002 naturally precedes the value for 2003. We can illustrate how such variables move over time in a **time-series plot** (also known as a line chart). The following chart shows monthly crude oil prices measured in dollars per barrel from January 2001 through November 2022 as reported by the Energy Information Administration ([www.eia.gov](http://www.eia.gov)).



We can show how two or more variables move together over time by adding additional series to the chart. However, if the series are measured in different scales, it may help to use a secondary vertical axis to illustrate a more meaningful association. The following two charts show crude oil prices measured in dollars per barrel and gasoline prices measured in dollars per gallon. The first chart has both variables measured on the left-hand vertical axis. The second chart measures crude oil prices in dollars per barrel on the left-hand axis and gasoline prices in dollars per gallon on the right-hand axis.

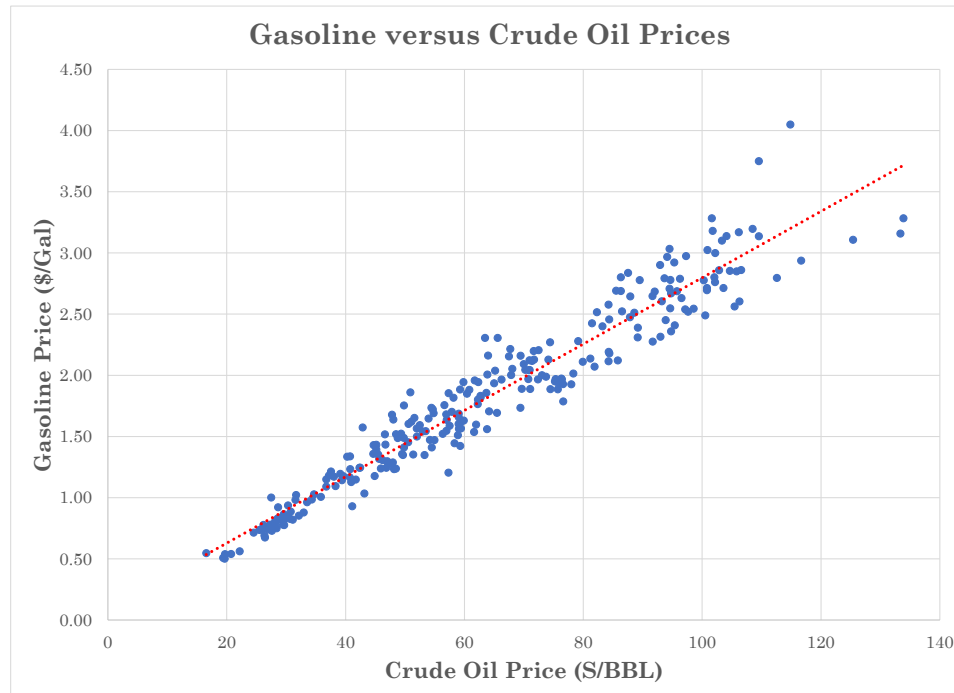


Some of the special features of time series data and its uses are introduced in Section 8.

### X-Y Graphs and Co-movement among Variables

While the foregoing time-series plot clearly shows how gasoline and crude oil prices move together over time, it does not illustrate the association (or **co-movement**) between the *levels* of the two variables. To show co-movement in levels, we would

want to use an X-Y graph that plots one of the two variables on the vertical axis (the “Y” variable) and the other on the horizontal axis (the “X” variable) such as in the following chart of the gasoline and crude oil prices:



Each “dot” in this chart represents one monthly X-Y pair of the gasoline-crude oil data. The dashed line represents the linear association (a “trend line”) between the two variables.

**Note regarding Date Values in Excel:** Excel stores date values, such as 8/30/2016, as an integer serial number representing the number of days beginning January 1, 1900 (1/1/1900 is stored as the integer value of 1). So 8/30/2016 is stored as the integer 42,612.

#### F. Frequency Distributions for Numerical Data

A frequency distribution is effectively a count of the number of observations from a particular sample that fall in different ranges. For a given sample of values of a numeric variable, we can identify the following values:

**Minimum:** Smallest value in the sample

**Maximum:** Largest value in the sample

**Range:** Maximum minus the Minimum

**Sample Size** (usually denoted  $n$ ): the number of individual observations in the sample.

To construct a frequency distribution, you first decide how many equally wide “classes” to break the range of data into. For some variables, there may be somewhat standard break points to sub-divide the data. For example, with grades from a 100 point scale, standard break points might be 60, 70, 80, and 90. For

measures of dollar income, you might want to use increments of \$10,000 starting with the first multiple of \$10,000 observed in the sample and including enough classes to encompass all the values in the sample. If there are no natural break points, a rough guide is determined by the number of observations in the sample as shown in the following table:

Sample Size	Number of Classes
Fewer than 50	5 – 7
50 to 100	7 – 8
101 to 500	8 – 10
501 to 1,000	10 – 11
1,000 to 5,000	11 – 14
More than 5,000	14 – 20

Given the number of classes, next determine the *class width*:

Eq. 1:1

$$w = \text{Class Width} = \frac{\text{Maximum} - \text{Minimum}}{\text{Number of Classes}}$$

Be sure to round the class width up to the next significant digit. For example, if your data are integer values such as 23, 46, 57, and the calculation above results in a class width of 7.2 (or any other decimal value), round it up to 8. If your data has a single decimal point, round up the class width to the next tenth (for example, 10.72 to 10.77 is rounded up to 10.8). The class break points are then given by:

Class	Value
1	Minimum + Class Width
2	Class 1 + Class Width
3	Class 2 + Class Width
...	...

The next step is then to count the number of observations in the sample that fall within each class. There are a number of ways to simplify this task in Excel. If you want to control exactly how the class break points are treated, use the COUNTIFS function. Whether you are using standardized or calculated break points for the various classes, you must take care that classes completely cover the range of the data and that there is no overlap in classes. This involves a choice in how you treat the class break points. For example, when counting the number of grades from a 100 point scale, you would want to include in the first class the count of the number of



grades up to but less than 60, the second class would be the count of grades from 60 up to but less than 70, etc. The following example shows how the “>=” and “<” operators in the COUNTIFS function allow you to control whether the break points are or are not included in the class:

	D	E	F	G	H	I	J	K	L	M
	Exam									
4	89									
								Cumulative		
								Percent		
3	66	Class	At least	But Less	Frequency	Percent		Percent		
1	69			Than	Count	Frequency		Frequency		
3	69	1	0	60	=COUNTIFS(\$D\$2:\$D\$99,">=" &G4,\$D\$2:\$D\$99,"<" &H4)					
3	56	2	60	70	20	20.4%		39.8%		
3	50	3	70	80	16	16.3%		56.1%		
3	67	4	80	90	19	19.4%		75.5%		
3	74	5	90	101	24	24.5%		100.0%		
4	55	Total			98					

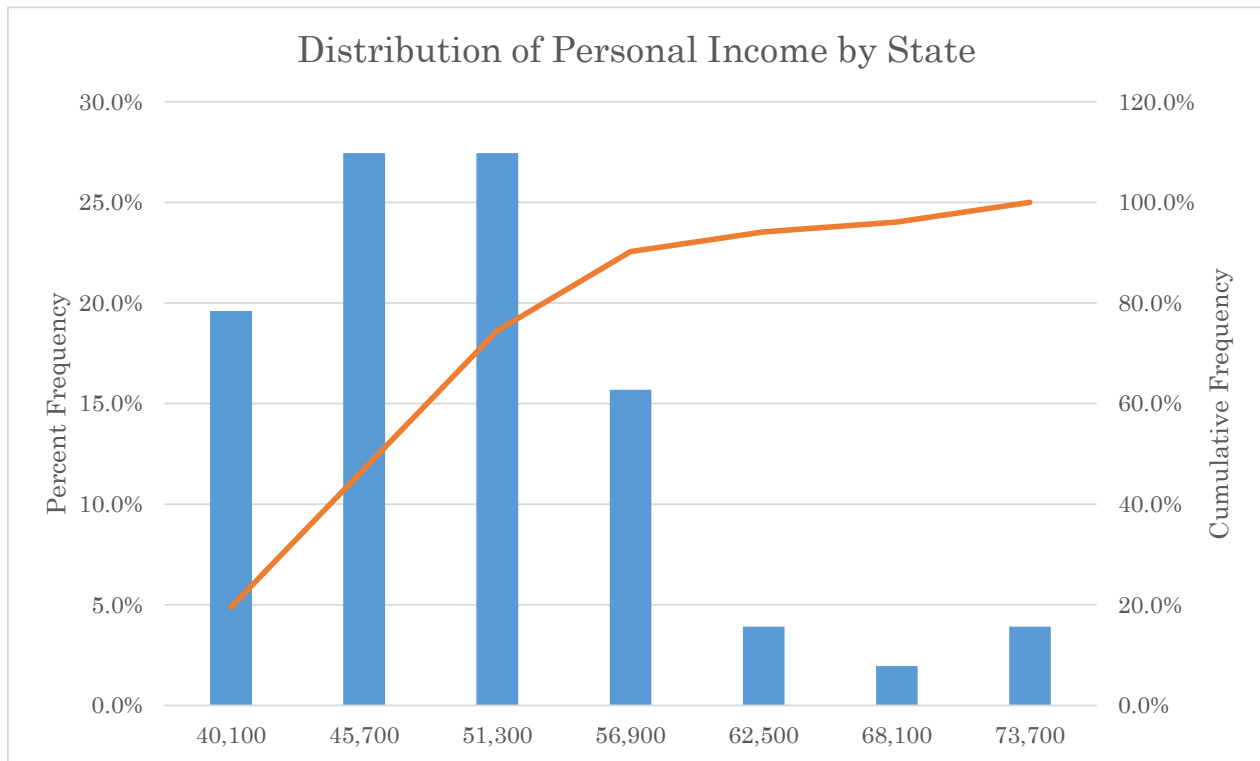
The Percent Frequency is simply the frequency count divided by the sample size. The Cumulative Percent Frequency sums up the individual class Percent Frequencies. In the example above, 56.1% of the grades were less than 80.

As an alternative example where the break points are calculated, I downloaded State per capita personal income data for 2015 from the Bureau of Economic Analysis ([www.bea.gov](http://www.bea.gov)) that includes the fifty U.S. States plus the District of Columbia for 51 total observations.

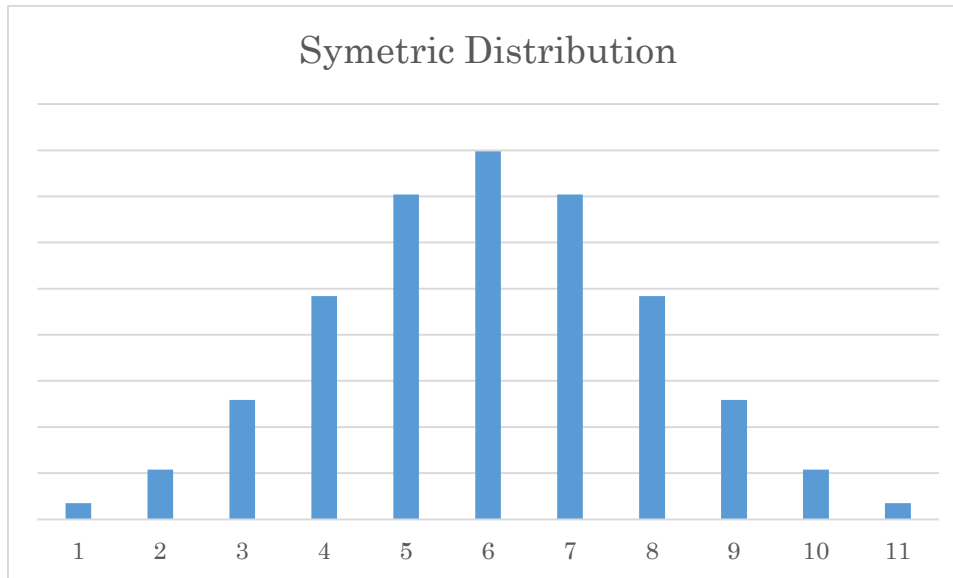
Minimum	34,771								
Maximum	73,302								
Range	38,531								
Sample Size	51								
Classes	7								
Interval Size	5504.4286								
Rounded	5,600								
Class	No more than	Up to	Frequency	Percent Frequency	Cumulative	Percent Cumulative			
1	34,500	40,100	=COUNTIFS(\$C\$4:\$C\$54,">" &F15,\$C\$4:\$C\$54,"<=" &G15)						
2	40,100	45,700	14	27.5%	24	47.1%			
3	45,700	51,300	14	27.5%	38	74.5%			
4	51,300	56,900	8	15.7%	46	90.2%			
5	56,900	62,500	2	3.9%	48	94.1%			
6	62,500	68,100	1	2.0%	49	96.1%			
7	68,100	73,700	2	3.9%	51	100.0%			
Totals			51	100.0%					

With 51 total observations I choose to use 7 classes. The calculated class width of 5,504.4286 was rounded up to 5,600. Starting from a ‘floor’ of 34,500, the first class

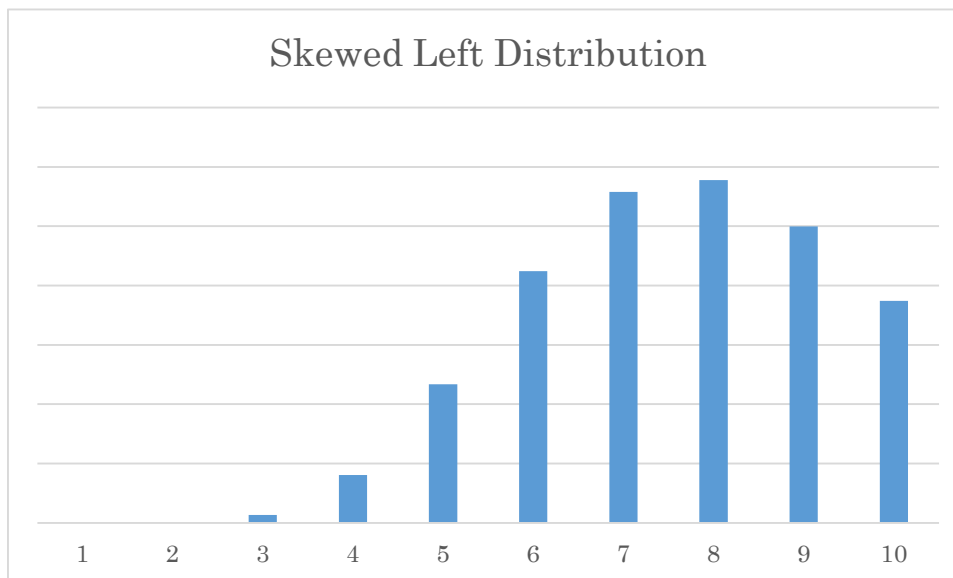
counts the number of States with per capita personal income up to and including 40,100. The second class includes those states with incomes greater than 40,100 and up to and including 45,700, etc. The following bar chart – also known as a **histogram** – shows the Percent Frequency and the Cumulative Percent Frequency as a line scaled on the right-hand axis:



Additional features of the frequency distribution that help to describe the data relate to the **mode** of the distribution – the mode is the class with the largest frequency. Note that there may be more than one mode – called a multi-modal distribution. If observations are roughly equally distributed on each side of a single mode, the distribution is called **symmetric**, like the following:



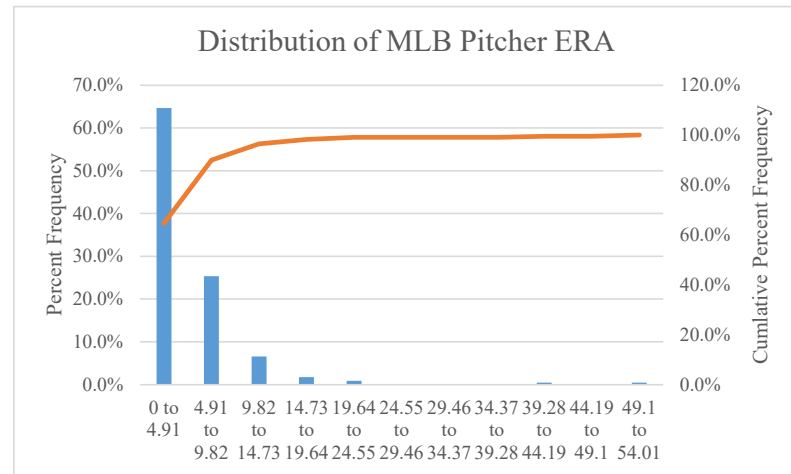
The distribution of personal incomes is **skewed right** – because it has a ‘long right tail.’ An example of a distribution that is **skewed left** is as follows:



The method described above for constructing frequency distributions works well when the data sample under consideration does not contain “extreme” values – observations widely skewed (at the lower and/or the upper end of the distribution) from the bulk of the remaining observations. However, it is very common where these types of extreme values are an issue. For example, consider the following summary statistics for the ERA (a measure of performance) of a sample of 519 Major League Baseball pitchers:

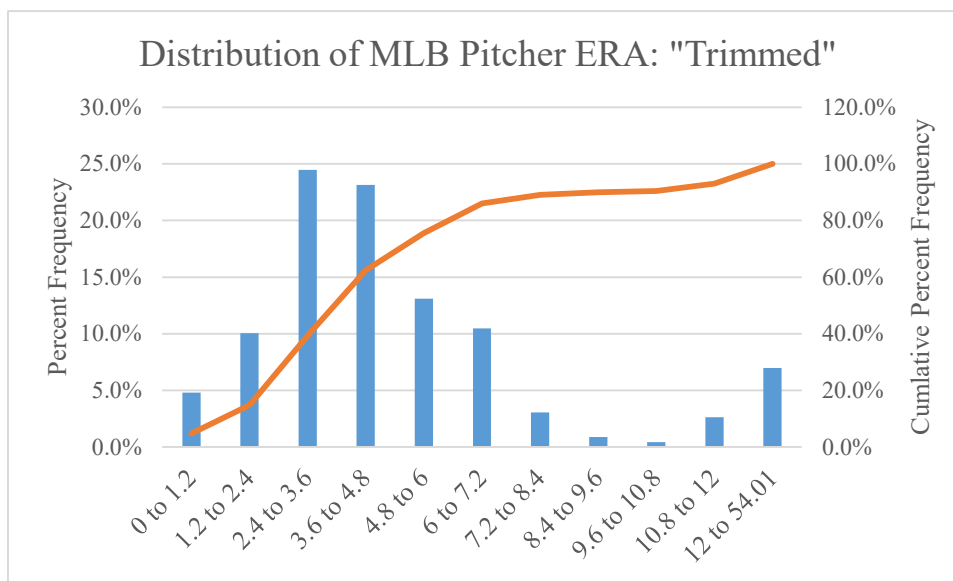
Minimum	0
Maximum	54
Range	54
Sample Size	519
Classes	11
Width	4.909
Rounded	4.91

Using the procedure for determining class widths presented above would result in the following frequency distribution:



Which is not particularly informative as over 80% of the observations fall in the first two classes.

A slightly amended procedure “trims” the distribution by forcing the roughly 5% of “extreme” observations at the upper end of the range into a final class. (See the discussion of quartile and percentile statistics in Section 2.B and the Excel function discussion in Section 10 below.) For this particular sample, roughly 5% of the observation have an ERA value of 12 or more. Setting the class width at 1.2 and letting the final class include all observations with an ERA of 12 or more results in the following frequency distribution chart:



This view is much more informative in showing the breakdown of the sample distributed between and ERA of 0 and 4.8.

**G. Example Problems**

Using the following 22 data points:

-2.1	-5.5	14.2	-1.2	-36.1	10.6	6.4	1.0	-2.6	20.2	-11.0
21.2	9.5	6.7	19.9	-21.0	30.6	-1.6	-13.3	-3.0	21.6	40.2

Construct a frequency distribution table including the cumulative frequency distribution of the data.

Sketch a histogram based on your frequency analysis.

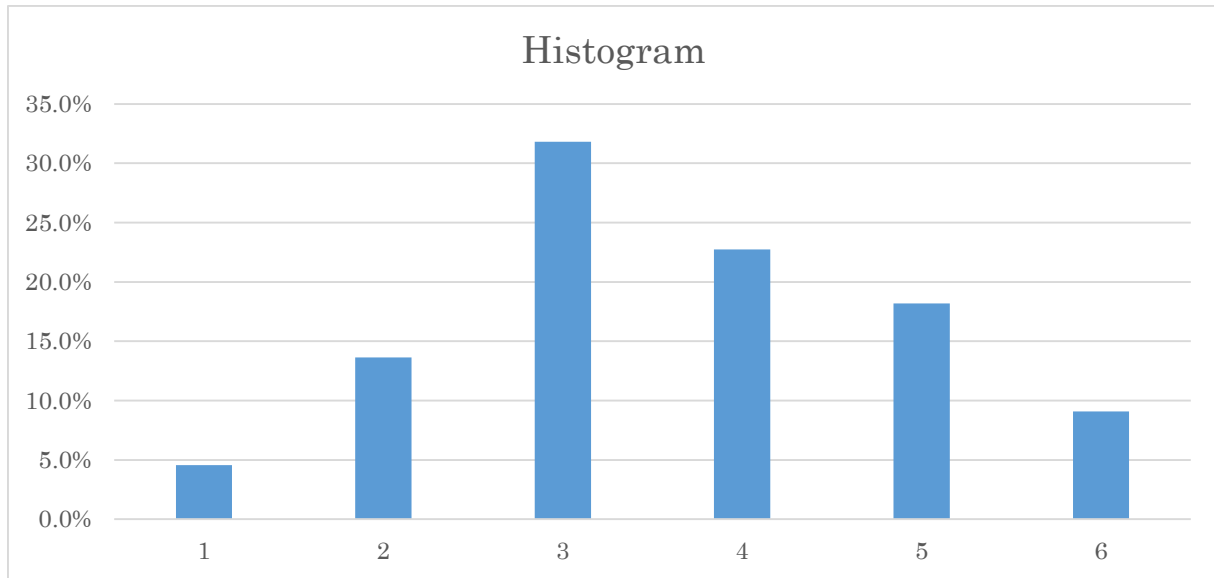
Does this distribution appear to be symmetric, or skewed to the right, or skewed to the left?

Solution: Since there are 22 observations, you should use 5 to 7 classes. I have chosen 6. The minimum is -36.1 and the maximum is 40.2, so the class width is:

$$Width = \frac{40.2 - (-36.1)}{6} \approx 12.72$$

Rounded up to 12.8. This results in frequency counts as follows:

Class	At Least	Less Than	Freq	Percent	Cum
1	-36.1	-23.3	1	4.5%	4.5%
2	-23.3	-10.5	3	13.6%	18.2%
3	-10.5	2.3	7	31.8%	50.0%
4	2.3	15.1	5	22.7%	72.7%
5	15.1	27.9	4	18.2%	90.9%
6	27.9	40.7	2	9.1%	100.0%
Total			22		



The distribution is roughly symmetric to slightly skewed right.

## Section 2. MEASURES OF CENTRAL TENDENCY, VARIABILITY AND CO-MOVEMENT

In analyzing any numerical variable, the frequency distributions we explored in Section 1 help illustrate that with many economic variables, certain values (or ranges of values) are observed more frequently than others. Ultimately, we want to be able to make statements in probability about the likelihood of a particular variable taking on either a specific value or a value in specific range. For a single variable, this will require constructing measures that help indicate whether observations are centered or clustered around a particular value and the degree to which observations deviate around that central value. These are called measures of central tendency and variability. When two or more variables are the focus of the analysis, we will examine measures of co-movement that provide an indication of how observations may cluster together.

### A. Measures of Central Tendency

For a sample of a single random variable, measures of central tendency provide a single value for the “center” of the data. We will examine three different measures of central tendency: the **mean**, **median** and **mode**. In addition, we will examine three different types of measures of the mean: the **arithmetic mean**, the **weighted mean**, and the **geometric mean**.

For a random variable  $X$ , the *population* mean is a *parameter* that we will denote with the Greek letter  $\mu$ . For a given sample of the variable  $X$  with  $n$  observations, the **arithmetic mean** (or simply *mean*) will be denoted  $\bar{X}$  and is defined as:

Eq. 2:1

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n} = \left(\frac{1}{n}\right) \sum_{i=1}^n X_i$$

The mean minimizes the *sum of squared errors*, denoted **SSE**, in the data sample. Again, for a given sample of the variable  $X$  with  $n$  observations, SSE measures how the data vary around some constant value, say  $a$ . If you minimize that function with respect to  $a$ , you get the following:

Eq. 2:2

$$\min_a SSE = \sum_{i=1}^n (X_i - a)^2$$

take the derivative with respect to  $a$ , set it equal to zero and solve for  $a$

Eq. 2:3

$$\begin{aligned} -2 \sum_{i=1}^n (X_i - a) &= 0 \\ na &= \sum_{i=1}^n X_i \end{aligned}$$

$$a = \left(\frac{1}{n}\right) \sum_{i=1}^n x_i = \bar{x}$$

**Note:** You will not be required to derive the minimum SSE result above. However, we will see the notion of minimum SSE statistics throughout the semester and you will see this concept again in detail in your Econometrics 463 course.

While the Eq. 2.1 for the sample mean  $\bar{x}$  looks a bit messy with the summation operator  $\sum$  in practice, functions built into Excel make these calculations quite simple. The Excel function for the arithmetic mean is `=AVERAGE(data range)`. Consider a sample from the labor force data in Section 1 of twenty Texas counties around Houston (Harris County). As shown in below, you can calculate the mean Unemployment Rate for this sample using the AVERAGE function with reference to the values in cells E2:E21

LEFT					=AVERAGE(E2:E21)					
	A	B	C	D	E					
	County Name/State Abbreviation	Labor Force	Employed	Unemployed	Unemploy- ment Rate (%)					
1										
2	Austin County, TX	14,154	13,508	646	4.6					
3	Brazoria County, TX	167,023	159,395	7,628	4.6					
4	Brazos County, TX	105,943	102,351	3,592	3.4					
5	Burleson County, TX	7,695	7,355	340	4.4					
6	Chambers County, TX	17,895	16,945	950	5.3					
7	Fort Bend County, TX	348,961	333,948	15,013	4.3					
8	Galveston County, TX	157,727	149,858	7,869	5.0					
9	Grimes County, TX	11,462	10,844	618	5.4					
10	Hardin County, TX	25,221	23,801	1,420	5.6					
11	Harris County, TX	2,239,426	2,135,626	103,800	4.6					
12	Jefferson County, TX	108,580	101,003	7,577	7.0					
13	Liberty County, TX	31,068	28,934	2,134	6.9					
14	Madison County, TX	5,106	4,887	219	4.3					
15	Matagorda County, TX	17,190	16,024	1,166	6.8					
16	Montgomery County, TX	255,338	244,446	10,892	4.3					
17	San Jacinto County, TX	11,260	10,621	639	5.7					
18	Walker County, TX	22,861	21,693	1,168	5.1					
19	Waller County, TX	21,351	20,318	1,033	4.8					
20	Washington County, TX	15,355	14,616	739	4.8					
21	Wharton County, TX	21,239	20,295	944	4.4					
22										
23										
24	Mean Unemployment Rate	=AVERAGE(E2:E21)								
25										

Note how the formula for the mean in cell B24 shows up in the ‘formula bar.’



**Effect of changing scale on the Sample Mean:** If you rescale a variable – by multiplying and/or adding a constant – this will affect the calculated value of the sample mean. For example, suppose an instructor gives an exam worth a total of 60 points and the mean score is 43. If scores are rescaled to a 100 point basis by multiplying by 10/6, and the instructor then applies a curve by adding 4 points, what is the new mean score? To see the solution, let the original set of scores be represented by the variable X, let a and b be constants, and let the rescaled scores be represented by the variable Y with:  $Y = a + bX$ . Then,

Eq. 2:4

$$\begin{aligned} \bar{Y} &= \sum \frac{(a + bX_i)}{n} \\ &= \frac{na}{n} + b \sum \frac{X_i}{n} \\ &= a + b\bar{X} \end{aligned}$$

Applying this formula to the exam example, the new mean score is about 75.67.

Another measure of central tendency is the **Median**: For a sample of the variable X with  $n$  observations, the median is loosely defined as the middle observation when the data is sorted in order from smallest to largest. That definition is technically true when the number of observations  $n$  is an odd number. For example, with a sample size of 7, when the data are sorted from smallest to largest, the 4<sup>th</sup> observation in the sorted sample is the median:

	A	B	C	D	E	F	G
1				<b>SORTED</b>			
2	Obs	X		Obs	X		
3	1	82		1	55		
4	2	55		2	59		
5	3	94		3	66		
6	4	89		4	79	<== The Median	
7	5	79		5	82		
8	6	66		6	89		
9	7	59		7	94		

So if the sample size  $n$  is an odd number, when the sample is sorted from smallest to largest, the observation number of the median will be:  $\left(\frac{n-1}{2}\right) + 1$ .

However, with a sample size that is an even number, there is no single “middle” observation in the sorted sample. There are a number of different ways to define the median in this case, the most expedient and the one used in Excel, is to average the two numbers that fall in the middle of the sorted sample. For example, for a sample of 10 observations, the median is the average of the 5<sup>th</sup> and 6<sup>th</sup> observations in the sorted sample:

	A	B	C	D
1	Obs	X		
2	1	33		
3	2	55		
4	3	59		
5	4	64		
6	5	66	72.5    Median = (66+79)/2	
7	6	79		
8	7	82		
9	8	89		
10	9	94		
11	10	97		
12				

If the sample size  $n$  is an even number, when the sample is sorted from smallest to largest, the median will be average of observation numbers:  $\left(\frac{n}{2}\right)$  and  $\left(\frac{n}{2}\right) + 1$ .

If the sample has a *symmetric* distribution, the mean and median will be very close to one another. If the sample has a distribution that is *skewed to the right*, the mean will be greater than the median, and if it is *skewed to the left*, the mean will be less than the median.

**The Mode:** The Mode is simply the most frequently observed value within a particular sample. Note, however, that variables can exhibit multiple modes where observations cluster around several different values. The concept of the mode is often applied to the classes in a frequency distribution so that the class with the largest frequency count, or equivalently the largest percentage frequency, is described as the mode. However, you must be careful to recognize that this can depend on how the distribution classes (the break points) are defined. In addition, when a variable exhibits multiple modes, it may be because two or more different types of measures are included in the sample. For example, the heights of students in a sample of 100, 50 men and 50 women, will quite likely have two modes (it is *bimodal*) one for the males and one for the females.

**Weighted Mean:** There are many situations in business and economics where a simple mean provides an inaccurate measure of the population mean when there are substantially different sub-groups within a sample. In such cases a weighted mean may be more appropriate. In a simple mean, each observation  $X_i$  from a sample of size  $n$  on the variable  $X$  is weighted by  $1/n$ . With a weighted mean, each observation is associated with a weight  $w_i$  and the weighted mean, denoted  $\bar{X}_w$ , is calculated as:

Eq. 2:5

$$\bar{X}_w = \frac{\sum_{i=1}^n w_i X_i}{\sum_{i=1}^n w_i}$$

An example where a weighted mean would be appropriate is when you have summary measures on a number of groups (for example, unemployment rates for different counties) and the groups are different sizes (counties have different sized labor forces). For example, in the 20 county sample of labor force data shown above, the simple mean of the unemployment rate is 5.07. When rates are weighted by the size of the labor force, the weighted mean is 4.65. In this example, the weighted mean is less than the simple mean because the counties with the largest labor force tend to have unemployment rates less than that of the simple mean. It is straightforward to calculate a weighted mean in Excel using a combination of the SUMPRODUCT and SUM functions via: =SUMPRODUCT(range for weights, range for X values)/SUM(range for weights) such as the following:

LEFT : <input type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/> fx					
=SUMPRODUCT(B2:B21,E2:E21)/SUM(B2:B21)					
	A	B	C	D	E
	County Name/State Abbreviation	Labor Force	Employed	Unemployed	Unemployment Rate (%)
1					
2	Austin County, TX	14,154	13,508	646	4.6
3	Brazoria County, TX	167,023	159,395	7,628	4.6
4	Brazos County, TX	105,943	102,351	3,592	3.4
5	Burleson County, TX	7,695	7,355	340	4.4
6	Chambers County, TX	17,895	16,945	950	5.3
7	Fort Bend County, TX	348,961	333,948	15,013	4.3
8	Galveston County, TX	157,727	149,858	7,869	5.0
9	Grimes County, TX	11,462	10,844	618	5.4
10	Hardin County, TX	25,221	23,801	1,420	5.6
11	Harris County, TX	2,239,426	2,135,626	103,800	4.6
12	Jefferson County, TX	108,580	101,003	7,577	7.0
13	Liberty County, TX	31,068	28,934	2,134	6.9
14	Madison County, TX	5,106	4,887	219	4.3
15	Matagorda County, TX	17,190	16,024	1,166	6.8
16	Montgomery County, TX	255,338	244,446	10,892	4.3
17	San Jacinto County, TX	11,260	10,621	639	5.7
18	Walker County, TX	22,861	21,693	1,168	5.1
19	Waller County, TX	21,351	20,318	1,033	4.8
20	Washington County, TX	15,355	14,616	739	4.8
21	Wharton County, TX	21,239	20,295	944	4.4
22					
23	<b>Mean Unemployment Rate</b>	<b>5.07</b>			
24	<b>Weighted Mean Unempl. Rate</b>	<b>=SUMPRODUCT(B2:B21,E2:E21)/SUM(B2:B21)</b>			

Another specialized measure of central tendency is the **Geometric Mean**. The Geometric Mean, denoted  $\bar{X}_g$ , is a specialized average used in business and economics with growth rates and rates of return. Instead of adding the values in a sample and dividing by the sample size, the values are multiplied together (you take

the product of the series) and the  $n^{\text{th}}$  root is applied to the result. The geometric mean of  $X$  for a sample of size  $n$  is given by:

Eq. 2:6

$$\bar{X}_g = \sqrt[n]{X_1 \times X_2 \times \dots \times X_n} = \left( \prod_{i=1}^n X_i \right)^{\frac{1}{n}}$$

Where the Greek letter  $\Pi$  is used to denote the product of the values. If  $X_i$  is the periodic gross rate of growth between period  $i-1$  and  $i$ :  $X_i = \left(\frac{P_i}{P_{i-1}}\right)$ , then the geometric mean measures the Average Compound Periodic Return, and the calculation simplifies to:

Eq. 2:7

$$\bar{X}_g = \left(\frac{P_n}{P_0}\right)^{\frac{1}{n}}$$

You can calculate a geometric mean in Excel using the `GEOMEAN(data range)` function.

Before delving into measures of variability, it is important to reiterate the distinction between *population* parameters and *sample* statistics. Population parameters of either central tendency, such as the mean, or of variability such as the variance or standard deviation, are usually unknown. However, assumptions about the basic distribution of a variable allows us to make probability based inferences about a variable based on sample statistics that we can calculate from observed data. The key population and sample statistics and notation that we will use to distinguish them are as follows:

Statistic	Population Parameter Notation	Sample Statistic Notation
Mean	$\mu$	$\bar{X}$
Variance	$\sigma^2$	$s^2$
Standard Deviation	$\sqrt{\sigma^2} = \sigma$	$\sqrt{s^2} = s$
Covariance (between X and Y)	$\sigma_{XY}$	$s_{XY}$
Correlation (between X and Y)	$\rho_{XY}$	$r_{XY}$

## B. Distance Measures of Variability

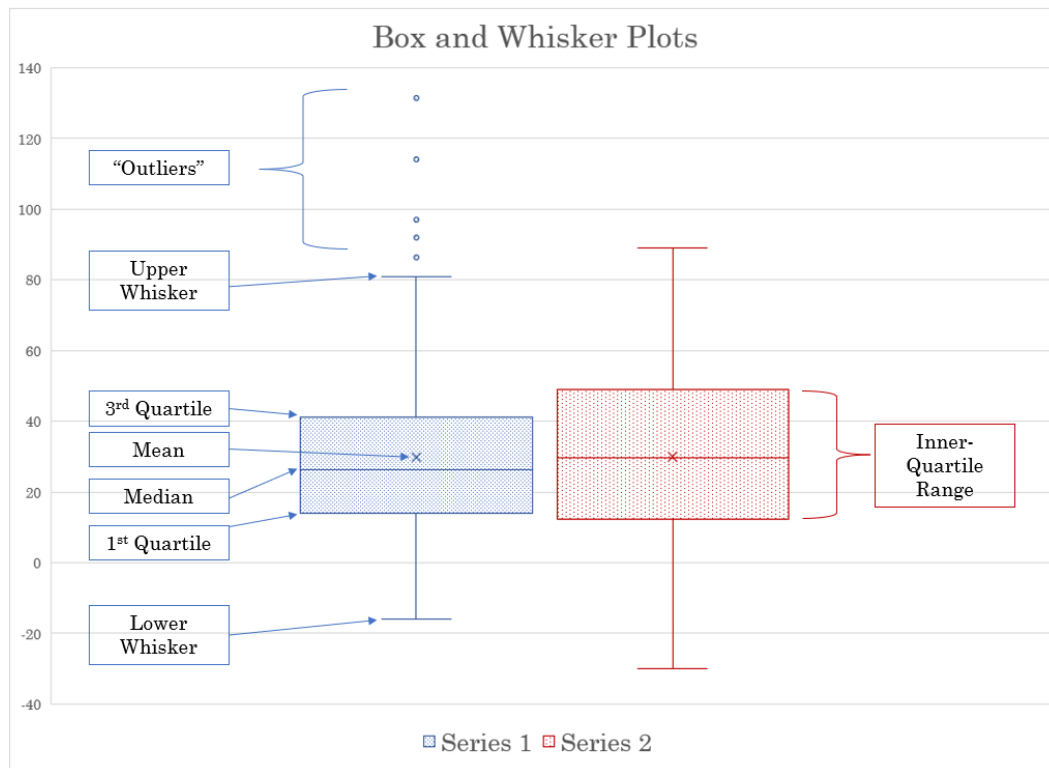
The **Range**: The range in a sample is simply the difference between the maximum observed value and the minimum observed value. The greater the spread of data from the center of the distribution, the larger the range will be. However, the range can be a misleading measure of variability if there are a few extreme observations

(either very large or very small), called *outliers*. To control for this possibility, it is helpful to look at the range of the middle 50% of the data: the *interquartile range*.

**Interquartile Range:** The first quartile, denoted  $Q_1$ , is the value below which 25% of the observations fall. The first quartile is also known as the 25<sup>th</sup> percentile. The third quartile,  $Q_3$ , is the value below which 75% of the observations fall (a.k.a., the 75<sup>th</sup> percentile). The *interquartile range* is then:  $Q_3 - Q_1$ . Note that the second quartile, a.k.a. the 50<sup>th</sup> percentile, is the median. In practice, there are several different methods for calculating quartiles, none of which are definitive. Excel includes two different functions for calculating quartiles: QUARTILE.INC and QUARTILE.EXC (Excel also includes a 'legacy' function QUARTILE that is the same as QUARTILE.INC). For several reasons, I would suggest using the QUARTILE.EXC function. For more information on calculating quartiles see the blog post: <http://datapigtechnologies.com/blog/index.php/why-excel-has-multiple-quartile-functions-and-how-to-replicate-the-quartiles-from-r-and-other-statistical-packages/>.

The PERCENTILE.EXC function is similar to the QUARTILE.EXC function. It returns the value in a data array below which any percentage between 0 and 1 of the sample fall. For example, PERCENTILE.EXC(array,0.9) returns the value in array below which 90% of the observations fall.

**Box-and-Whisker Plots:** A graphical representation of the range, interquartile range, and median that is useful to identify skewed distributions and outliers is a Box-and-Whisker plot, such as the following:



The “Box” for each of the two illustrated series, shows the 1<sup>st</sup> Quartile (bottom of the Box), the Median (solid line in the middle of the Box), and the 3<sup>rd</sup> Quartile (top of the Box). The Mean for each series is shown as an × on the chart.

As shown on the Box for Series 2, the Interquartile Range is illustrated by the height of the Box.

The Lower Whisker shows one of two things – either the Minimum value of the series or the smallest value that is not less than  $Q_1 - 1.5 \times \text{Interquartile Range}$ .

The Upper Whisker is defined similar to the Lower Whisker – either the Maximum value of the series or the largest value that is not greater than  $Q_3 + 1.5 \times \text{Interquartile Range}$ .

Any data points outside the Lower or Upper Whiskers are defined to be “Outliers” or extreme values.

If the Median is close to the middle of the range between the 1<sup>st</sup> and 3<sup>rd</sup> quartiles, and the Mean is very close to the Median, (as with Series 2, above) the distribution will tend to be symmetric.

If the Median is closer to the first quartile, and the Mean is greater than the Median, (as with Series 1 above) the distribution will tend to be skewed right.

If the Median is closer to third quartile, and the Mean is less than the Median, the distribution will tend to be skewed left.

Alternative distance measures of variability may be based on the range between certain percentiles of a data sample. For example, the range between the 90<sup>th</sup> percentile and the 10<sup>th</sup> percentile.

### C. Variance and Standard Deviation

The preceding distance measures of variability all measure ranges based in specific pairs of observations in a sample. Variance and standard deviation statistics are averages of variability across all observations either in the population or in a particular sample. The **population variance**,  $\sigma^2$ , is the sum of the squared differences between each observation and the population mean,  $\mu$ , divided by the population size  $N$ :

Eq. 2:8

$$\sigma^2 = \frac{\sum_{i=1}^n (X_i - \mu)^2}{N}$$

The **sample variance**,  $s^2$ , is the sum of the squared differences between each observation and the *sample* mean,  $\bar{X}$ , divided by the sample size  $n$ , minus 1:

Eq. 2:9

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

The  $(n - 1)$  in the denominator is called the **degrees of freedom**, and is used instead of simply the sample size to make the sample variance  $s^2$  an **unbiased** estimator of the population variance: that is, on average,  $s^2$  will equal the population variance. A computational shortcut that is handy if you are calculating a sample variance manually is as follows:

Eq. 2:10

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

**Standard deviation**, for either the population or the sample, is the square root of the variance. Thus, the population standard deviation is:

Eq. 2:11

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^n (X_i - \mu)^2}{N}}$$

And the sample standard deviation is:

Eq. 2:12

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

The Excel functions for variance and standard deviation are summarized in the following table.

Function	Result
VAR.P	Population variance – uses N in the denominator
VAR.S	Sample variance – using n-1 in the denominator
STDEV.P	Population standard deviation – the square root of VAR.P
STDEV.S	Sample standard deviation – the square root of VAR.S

**Changing Scales:** If you add (or subtract) a constant from a variable X, there is NO effect on the variance or standard deviation since the mean will change by the same constant. However, if you multiply by a constant  $k$ , the resulting variance will be  $k^2$  times the original variance, and the standard deviation will be  $k$  times the original standard deviation.

The importance of the standard deviation is illustrated by the **Empirical Rule**: for any variable that has approximately a normal bell-shaped distribution, approximately 68% of the observations will lie within the interval  $\mu \pm \sigma$ , approximately 96% of the observations will lie within the interval  $\mu \pm 2\sigma$ , and almost all the observations will lie within the interval  $\mu \pm 3\sigma$ .

**Coefficient of Variation:** With measures of the standard deviation on different variables, it may be tempting to make a direct comparison among the standard deviation statistics to address the question of which variable exhibits more or less variability about its mean. However, if the variables have different means, such a comparison does not have much meaning. To remove these scale effects, we can compare the **coefficient of variation**, denoted **CV**, which is the ratio of a variable's standard deviation to its mean (either population or sample measures) *provided the means are positive*, and is usually expressed as a percentage. Thus, the population coefficient of variation is:

Eq. 2:13

$$CV = \frac{\sigma}{\mu}$$

And the sample coefficient of variation is:

Eq. 2:14

$$CV = \frac{s}{\bar{X}}$$

**Z-Score:** A **z-score** is a standardized value that indicates the number of standard deviations a specific data value is from the mean. It can be positive (value is greater than the mean) negative (value is less than the mean) or zero (equal to the mean):

Eq. 2:15

$$z = \frac{X_i - \mu}{\sigma}$$

The **z-score** can also be calculated based on the sample measures of the mean and standard deviation:

Eq. 2:16

$$z_s = \frac{X_i - \bar{X}}{s}$$

For a given sample, if all the observations are converted to z-scores, it is straightforward to show that the sample mean of the z-scores is zero and its standard deviation is one.

#### D. Measures of Co-Movement

The measures of central tendency discussed above help us understand how the observations of a random variance will cluster around a central point while measures of variability help us understand the extent of variation. We now turn to how two random variables tend to move together – or co-movement. The base measure of co-movement in a pair of random variables X and Y, is their **covariance**. The sample measure of the covariance between X and Y is:



Eq. 2:17

$$S_{XY} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{n - 1}$$

The covariance provides a measure of the tendency of the variable X to move together with the variable Y. If the variable Y tends to increase while the variable X increases, the covariance will be positive. If Y tends to decrease as X increases, the covariance will be negative. Beyond the **sign** or direction of the relationship, however, the covariance is not very informative because it is subject to scale effects. If the variable X is multiplied by a constant  $k_X$ , then the covariance will change by  $k_X$ . Similarly for a change in the scale of Y. To eliminate scale effects, we use the **correlation** between X and Y – the covariance of X and Y divided by their standard errors:

Eq. 2:18

$$r_{XY} = \frac{S_{XY}}{S_X S_Y}$$

Since changing the scale of X and/or Y will change both the covariance and the standard deviations by the same scale, the correlation eliminates any scale effects. Moreover, by the nature of its definition, the correlation coefficient is bound on the interval **[-1, 1]**. If observations of X and Y fall *exactly* on a downward sloping straight line, then the correlation will be -1. If the line is upward sloping, the correlation will be 1.

The Excel functions for calculating a covariance are CONVARIANCE.P (population covariance that uses N in the denominator) and COVARIANCE.S (the sample covariance that uses n-1 in the denominator as in Eq. 2:17). To calculate correlations in Excel, you can use of combination of the COVARIANCE.S and STDEV.S functions to replicate Eq. 2:18, or use the CORREL function.

Consider two variables X and Y with a positive correlation such as that shown in the chart below. The slope and intercept of a straight line of “best fit” between X and Y will depend on the scale of measurement of the two variables. However, the correlation coefficient has a scale free relation to the line of “best fit” through the z-scores of X and Y. For a given value of X,  $X_i$ , the z-score is:

Eq. 2:19

$$z_x = \frac{(X_i - \bar{X})}{s_x}$$

For that particular value of X, the predicted z-score for Y that falls along the straight line of “best fit” is:

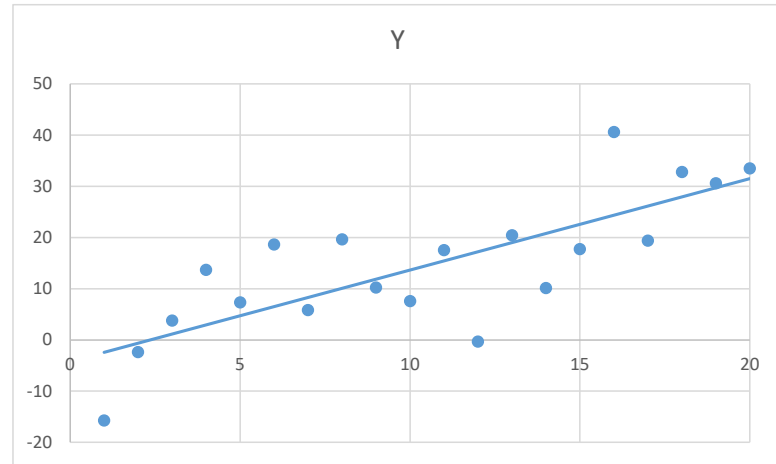
*Eq. 2:20*

$$\hat{z}_y = r_{xy}z_x$$

Where the “hat” denotes a predicted value. Given the sample correlation coefficient  $r_{xy}$ , and the calculated z-score for  $X_i$ , we can solve for a predicted  $Y_i$  with:

*Eq. 2:21*

$$\hat{Y}_i = (\hat{z}_y s_y) + \bar{Y} = (r_{xy} z_x s_y) + \bar{Y}$$



**E. Example Problems**

The following data represent a sample of 12 test scores (out of a total of 50 possible points) ordered from lowest to highest:

Obs.	1	2	3	4	5	6	7	8	9	10	11	12
Grade	21	25	26	27	29	31	35	36	38	39	40	41

Calculate the sample mean, median, and standard deviation of this data.

mean            32.33  
 median        33.00  
 stddev        6.71

Suppose the instructor wants to rescale the scores to be out of 100 by multiplying by 2, and then adds a curve of 8 points. Calculate the sample mean and standard deviation for the rescaled and curved grades. How does these relate to the original mean and standard deviation?

New Mean    72.67  
 New stddev   13.41  
 New Mean = (Old Mean \* 2) + 8  
 New Std. Dev. = Old Std. Dev \*2

.....  
 An investor buys shares in company CTS at a purchase price of 63.25 and observes the end of month prices in each of twelve months as shown in the table below:

Purchase Price	63.25	
End of Month	Price	Gross Return
Jan	66.73	1.055
Feb	65.66	0.984
Mar	66.91	1.019
Apr	69.59	1.040
May	71.26	1.024
Jun	73.61	1.033
Jul	72.80	0.989
Aug	73.96	1.016
Sep	73.22	0.990
Oct	72.12	0.985
Nov	75.15	1.042
Dec	76.20	1.014

Calculate the geometric mean of the Gross Return data to get the Average Compound Monthly (Gross) Return.

$$\bar{X}_g = \sqrt[12]{1.055 \times 0.984 \times 1.019 \times \dots \times 1.014} = 1.0156$$

OR

$$\bar{X}_g = \left( \frac{76.20}{63.25} \right)^{\frac{1}{12}} = 1.0156$$

.....

An investor purchases stock in two different companies, ABC and XYZ. Over a period of time, the following summary statistics of the two company's stock prices are observed:

	ABC	XYZ
Mean Price	11.34	33.66
Std. Dev. Price	3.107	6.811
Covariance	-16.129	

Which of the two company's stock prices exhibits more variability? Explain the basis for your answer.

Coefficient of Variation:  $\frac{S}{\bar{X}}$

ABC = 0.27

XYZ = 0.20

So ABC is more variable relative to the mean.

Calculate the correlation coefficient for the two company's stock prices.

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{-16.129}{11.34 \times 33.66} = -0.762$$

Suppose the stock prices are rescaled to account for number of shares purchased. ABC's prices are rescaled by multiplying by 100 and XYZ's prices are rescaled by multiplying by 150.

What happens to the covariance of the stock prices from this rescaling?

$$S_{xy} = \frac{\sum(100X_i - 100\bar{X})(150Y_i - 150\bar{Y})}{n - 1} = 100 \times 150 \times Old Cov$$

Covariance is scaled by a factor of 15,000

Would this rescaling change your answer to part (a) above? Explain why or why not.

No because both Std. Dev. and Mean are scaled by the same factor in this case.

Suppose the stock price for company ABC goes up by 1.5 standard deviations. Based on your calculated correlation coefficient from part (b), what is expected to happen to XYZ's stock price?

Expect XYZ's stock price to go down based on the negative correlation by:

$$Z_{XYZ} = r_{xy} \times Z_{ABC} = -0.762 * 1.5 = -1.14$$

Standard deviations.

## Section 3. PROBABILITY AND THE NORMAL PROBABILITY DISTRIBUTION

---

### A. Probability and Probability Distribution Functions

The first two sections dealt with various measures of central tendency, variability, and co-movement. We now turn to the methods by which we can make statements in probability about random variables. A random variable  $X$  is a variable that takes on values out of a defined set of possibilities. We learned that a key measure of central tendency for  $X$  is the population mean, denoted  $\mu$ . Our sample measure of the mean is the (arithmetic) mean, denoted  $\bar{X}$ . The key measures of variability were the population variance ( $\sigma^2$ ) and standard deviation ( $\sigma$ ) with sample measures denoted as  $s^2$  and  $s$ . These measures give us an indication of how values of  $X$  may cluster around a particular point (the mean) and the degree to which values of  $X$  vary around that point (the standard deviation). What we need is a means of linking specific values of  $X$  to the probability of observing that value – this is the **Probability Distribution Function** or **PDF**, denoted  $f(X)$ . The PDF maps specific values of the random variable  $X$  to the probability of that value being observed. Thus, for a specific value of  $X$ , say  $X_0$ , the PDF  $f(X_0)$  is such that

*Eq. 3:1*

$$f(X_0) = \text{Probability}[X = X_0] \text{ or } P[X = X_0]$$

**By definition, all probabilities are bound on the interval [0, 1]. A probability of 1 means an event is certain to occur while a probability of zero means it is certain not to occur, and any value in between zero and one indicates the likelihood of observing a particular value. The probability function,  $f(X)$  is therefore also bound on the interval [0, 1].**

The specific form of  $f(X)$  will depend on the nature of the random variable  $X$ . Random variables can be either *discrete* ( $X$  takes on a set of countable values, usually integers) or *continuous* ( $X$  can take on *any* value in an interval). Note that a discrete random variable could involve an infinite number of possibilities (any positive integer, for example) and that a continuous random variable could involve a very narrow range (any fractional value between 1 and 2, for example).

An example of a discrete random variable is the roll of a pair of six-sided dice with the faces numbered 1 to 6 and the outcome is the sum of the two die. As shown in the following table, there are 36 possible combinations of 11 discrete outcomes in the set  $\{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$ . Assuming the dice are “fair” so that the chances of any given number being rolled on one of the individual dies is equally likely, the probabilities of each outcome – the PDF,  $f(X)$  – is as shown in the following table:

Die 1	Die 2						Outcomes: X	Frequency	PDF: f(X)	CDF: F(x)
	1	2	3	4	5	6				
1	2	3	4	5	6	7	2	1/36	1/36	
2	3	4	5	6	7	8	3	2/36	3/36	
3	4	5	6	7	8	9	4	3/36	6/36	
4	5	6	7	8	9	10	5	4/36	10/36	
5	6	7	8	9	10	11	6	5/36	15/36	
6	7	8	9	10	11	12	7	6/36	21/36	
							8	5/36	26/36	
							9	4/36	30/36	
							10	3/36	33/36	
							11	2/36	35/36	
							12	1/36	1	
							Total	36	1	

Since there is only one combination that results in X=2, the PDF f(2)=1/36 (similarly for X=12), while there are 6 combinations that result in X=7 so f(7)=6/36. For a discrete random variable X, the properties of the PDF f(X) are as follows (Xi is any single outcome in the set of possibilities):

Eq. 3:2

$$\begin{aligned}
 f(X_i) &= P[X = X_i] \\
 0 &\leq f(X_i) \leq 1 \\
 \sum_{X_i} f(X_i) &= 1
 \end{aligned}$$

In words,  $f(X_i)$  is the probability that X takes on the specific value  $X_i$ ; the probabilities are, by definition, bound between zero and one; the sum of probabilities over all possible outcomes of X is equal to one. The population measures of the mean  $\mu$  and variance  $\sigma^2$  are defined using **expected values** as functions of the PDF. The expected value of a discrete random variable X, denoted E(X), is defined as:

Eq. 3:3

$$E(X) = \sum_{X_i} X_i f(X_i) = \mu$$

So the population mean  $\mu$  is the probability weighted sum of all possible values of X. If you apply this definition to the dice example above, multiplying each outcome of X by its probability and adding them all together, you will find that the expected value (the mean) is equal to seven. The population variance,  $\sigma^2$  is defined as:

Eq. 3:4

$$Var(X) = E[X - E(X)]^2 = E[X - \mu]^2 = \sum_{X_i} (X_i - \mu)^2 f(X_i) = \sigma^2$$

The standard deviation of  $X$  is simply the square root of the variance. While the PDF gives the probability that  $X$  will take on a particular value, the **Cumulative Distribution Function**, or **CDF**, is denoted  $F(X)$  and is the probability that  $X$  is less than or equal to a particular value. For a discrete random variable, the CDF has the following properties:

Eq. 3:5

$$F(X_i) = P[X \leq X_i]$$

$$F(X_i) = \sum_{X \leq X_i} f(X)$$

$$0 \leq F(X_i) \leq 1$$

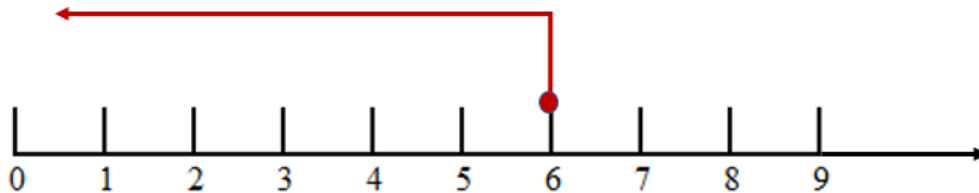
if  $X_0$  and  $X_1$  are such that  $X_0 < X_1$  then  $F(X_0) \leq F(X_1)$

Given the definition of the CDF and since the PDF must sum to 1, if we know the probability  $P[X \leq X_i] = F(X_i)$  then  $P[X > X_i] = 1 - F(X_i)$ . In the dice example above, for example, we know that the probability that  $X \leq 5$  from the CDF is  $10/36$  so the probability that  $X > 5$  is  $1 - 10/36 = 26/36$ . We can also use these properties to find the probability that  $X$  falls in a particular range. For example:

Eq. 3:6

$$\text{if } X_0 \text{ and } X_1 \text{ are such that } X_0 < X_1 \text{ then } P[X_0 < X \leq X_1] = F(X_1) - F(X_0)$$

For discrete probability distributions, the distinction between strict inequalities, for example,  $X < 5$ , versus weak inequalities, for example,  $X \leq 5$ , is important. A problem that asks for a probability that  $X$  is less than 7, does not include 7, it is 6 or less:



Other examples include:

At least 2 but less than 7: Includes 2, 3, 4, 5, and 6, but does not include 7

More than 6: Includes 7 or more

For continuous random variables, the same basic properties of the PDF and CDF remain but the summations in the definitions above are replaced with integrals to measure the area under a continuous function. For example:

Eq. 3:7

$$E(X) = \int Xf(X)dX = \mu$$

$$Var(X) = E(X - \mu)^2 = \int (X - \mu)^2 f(X)dX = \sigma^2$$



With a continuous random variable, the probability that the variable takes on any specific value technically devolves to zero: it is impossible to know whether a variable is arbitrarily close to some specific value or equal to that value. As a result, there is no real distinction between strict and weak inequalities with continuous random variables.

The results related to rescaling a random variable by applying a linear function adding a constant “a” and multiplying by a constant “b” that we covered previously, continue to hold in this context. For the random variable X and constants, a and b, for example:

*Eq. 3:8*

$$\begin{aligned} E(a + bX) &= a + bE(X) = a + b\mu \\ \text{Var}(a + bX) &= b^2\text{Var}(X) = b^2\sigma^2 \end{aligned}$$

## B. Binomial Distribution

A Binomial random variable is one that can take on one of only two values: a “success” or a “failure.” The properties of such a random variable are as follows:

A fixed number of observations, n; for example, 13 tosses of a coin; 11 cell phones taken from a production line.

Two mutually exclusive and collectively exhaustive categories; “Heads” or “tails” on the toss of a coin; “Defective” or “not defective” for a given cell phone;

Constant probability of “success” for each observation;

Observations are independent: the outcome of one observation does not affect the outcome of another.

The form of the Binomial Distribution is derived from the number of possible successes that are possible in n independent experiments. The number of sequences with x successes in n independent experiments is given by:

*Eq. 3:9*

$$C_x^n = \frac{n!}{x!(n-x)!}$$

Where  $n! = n(n-1)(n-2) \dots$ , and  $0! = 1$

Let P be the probability of a success for any single observation. Then the Probability Distribution Function, or PDF,  $P(x)$  – the probability of x successes in n trials – for the Binomial Distribution is:

*Eq. 3:10*

$$P(x) = \frac{n!}{x!(n-x)!} P^x (1-P)^{n-x}$$

The PDF is also often denoted as  $f(x)$ .

The Cumulative Distribution Function,  $F(x)$ , is

Eq. 3:11

$$F(x_0) = P[x \leq x_0] = \sum_{x=0}^{x_0} \frac{n!}{x!(n-x)!} P^x (1-P)^{n-x}$$

In Excel, we can use the BINOM.DIST function to handle these calculations:

$$=BINOM.DIST(\text{number\_s}, \text{trials}, \text{probability\_s}, \text{cumulative})$$

Where “number\_s” corresponds to  $x$ , “trials” to  $n$ , “probability\_s” to  $P$ , and “cumulative” is a True/False so 0 results in the PDF and 1 results in the CDF.

### Example Problem:

Based on Fall 2015 enrollment statistics, 59% of A&M students in the College of Liberal Arts are female.

- a. Suppose you select a random sample of 6 Liberal Arts students. What is the probability that 4 of those chosen are female?

Here the probability of “success” is given as 0.59

Eq. 3:12

$$P[x = 4] = \frac{6!}{4!(6-4)!} (0.59)^4 (1-0.59)^{(6-4)} \cong 0.3055$$

In Excel: =BINOM.DIST(4,6,0.59,0)

- b. Suppose you again select a random sample of 6 Liberal Arts students. What is the probability that 3 or less of those chosen are female?

Eq. 3:13

$$P[x \leq 3] = \sum_{x=0}^3 \frac{6!}{x!(6-x)!} (0.59)^x (1-0.59)^{(6-x)} \cong 0.4764$$

In Excel: =BINOM.DIST(3,6,0.59,1)

- c. Suppose you again select a random sample of 6 Liberal Arts students. What is the probability that 2 or more of those chosen are female?

Eq. 3:14

$$P[x \geq 2] = 1 - P[x \leq 1] = 1 - \sum_{x=0}^1 \frac{6!}{x!(6-x)!} (0.59)^x (1-0.59)^{(6-x)} \cong 0.9542$$

In Excel: =1-BINOM.DIST(1,6,0.59,1)

---

### C. Poisson Distribution

The Poisson probability distribution can be used to model the number of occurrences (or successes) of a certain event in a given continuous interval such as time, spatial area, or length.

The number of trucks arriving at a warehouse in a given week.

The number of failures in a computer system in a given day.

The number of defects in a large roll of sheet metal.

The number of customers to arrive at a coffee bar in a given time interval.

The assumptions of the Poisson distribution are as follows:

Assume that an interval is divided into a large number of equal subintervals so that the probability of the occurrence of an event in any subinterval is small.

The probability of the occurrence of an event is constant for all subintervals.

There can be no more than one occurrence in each subinterval.

Occurrences are independent – an occurrence in one subinterval does not have an effect on the probability of an occurrence in another subinterval.

The random variable  $X$  follows the Poisson distribution if it has the following probability distribution:

**Error! Bookmark not defined.** *Eq. 3:15*

$$P(x) = \frac{e^{-\lambda} \lambda^x}{x!} \text{ for } x = 0, 1, 2, \dots$$

Where:

$P(x)$  = the probability of  $x$  successes over a given time or space given  $\lambda$ .

$\lambda$  = the mean (or expected) number of successes per time or space unit,  $\lambda > 0$ .

$e = 2.71828$  (the base for natural logarithms).

In Excel, we can use the POISSON.DIST function:

=POISSON.DIST(x,mean,cumulative)

Where “ $x$ ” is as used above, “mean” is for  $\lambda$ , and “cumulative” is a True/False so 0 gives the PDF and 1 gives the CDF.

---

#### Example Problem:

Suppose customers arrive at the Evan’s Library Starbucks® at an average of 4 every five minutes. Assume that arrivals are independent with a constant arrival rate,

and that arrivals follow the Poisson distribution, with  $X$  denoting the number of arrivals in a given five-minute period and mean  $\lambda = 4$ .

- a) Find the probability that 2 or fewer customers arrive in a five minute period.

The probability that  $X$  is 2 or less,  $P(X \leq 2) = P(X=0) + P(X=1) + P(X=2)$ . With  $\lambda = 4$ , then

$$P(X = 0) = \frac{e^{-4}4^0}{0!} \cong 0.0183$$

$$P(X = 1) = \frac{e^{-4}4^1}{1!} \cong 0.0733$$

$$P(X = 2) = \frac{e^{-4}4^2}{2!} \cong 0.1465$$

So  $P(X \leq 2) = 0.0183 + 0.0733 + 0.1465 = 0.2381$

In Excel: =POISSON.DIST(2,4,1)

- b) Find the probability that more than 3 customers arrive in a five minute period.

The probability that  $X$  is more than 3,  $P(X > 3) = 1 - P(X \leq 3)$ . We found  $P(X \leq 2)$  in part a, above, and

$$P(X = 3) = \frac{e^{-4}4^3}{3!} \cong 0.1954$$

So  $P(X > 3) = 1 - P(X \leq 3) = 1 - (0.0183 + 0.0733 + 0.1465 + 0.1954) = 0.5665$

In Excel: =1 - POISSON.DIST(3,4,1)

#### D. Hypergeometric Distribution

The Binomial distribution discussed above assumes that items are drawn independently with the probability of selecting any item being constant. In practice, these assumptions can be met if a small sample is drawn from a large population. There are many applied problems, however, that posit the selection of a group of items from a relatively small population. Drawing from a small population is a situation of sampling without replacement. This implies that the probability of selection changes after each succeeding selection.

Suppose a random sample of  $n$  objects is drawn from a group of  $N$  objects,  $S$  of which are successes. The distribution of the number of successes,  $X$ , in the sample follows the Hypergeometric distribution if its probability distribution is given by:

*Eq. 3:16*

$$P(x) = \frac{C_x^S C_{n-x}^{N-S}}{C_n^N} = \frac{S!}{x!(S-x)!} \times \frac{(N-S)!}{(n-x)!(N-S-n+x)!} \frac{N!}{n!(N-n)!}$$

Where  $x$  can take on integer values from the larger of 0 and  $[n - (N - S)]$  and the smaller of  $n$  and  $S$ .

In Excel, we can use HYPGEOM.DIST( $x,n,S,N,cumulative$ )

### Example problem:

A financial analyst is given a list of 14 corporate bonds. Out of this list, 5 of the bonds would subsequently be downgraded. Suppose the analyst randomly selected 3 bonds from the list. What is the probability that at least 2 of the bonds chosen by the analyst were among those to be downgraded?

Here, the population size  $N=14$ , the number of success in the population  $S=5$ , and the sample size  $n=3$ . We want  $P(X \geq 2) = 1 - P(X \leq 1) = 1 - P(X=0) - P(X=1)$ :

$$P(x = 0) = \frac{5!}{0!(5-0)!} \times \frac{(14-5)!}{(3-0)!(14-5-3+0)!} \frac{14!}{3!(14-3)!} \cong 0.2308$$

$$P(x = 1) = \frac{5!}{1!(5-1)!} \times \frac{(14-5)!}{(3-1)!(14-5-3+1)!} \frac{14!}{3!(14-3)!} \cong 0.4945$$

So,  $P(X \geq 2) = 1 - 0.2308 - 0.4945 = 0.2747$

In Excel:  $=1 - \text{HYPGEOM.DIST}(1,3,4,14,1)$

## E. Exponential Distribution

The Binomial, Poisson, and Hypergeometric distributions are *discrete* distributions in that outcomes are countable (even though they by infinitely countable such as the number of grains of sand in a beach). The Exponential Distribution Function is a continuous distribution in that outcomes can take on any value greater than zero. It can be used to model the length of time between occurrences of an event.

Time between trucks arriving at a warehouse.

Time between customers calling a helpline.

An Exponential random variable  $T$  ( $t > 0$ ) has a probability distribution function,  $f(t)$  as follows:

Eq. 3:17

$$f(t) = \lambda e^{-\lambda t}$$

Where:

$\lambda$  is the mean number of **occurrences per unit**  $t$  (a time dimension or space dimension)

$t$  is the number of units (time or space)

$e$  is the nature number = 2.71828 ...

The mean number of **units per occurrence** is given by  $1/\lambda$

Recall from above that with continuous random variables, the probability that the variable takes on any specific value devolves to zero. Thus, in terms of statements of probability for a continuous random variable, we are only interested in ranges. For example, the probability the  $t < 5$  or the probability that  $3 < t < 5$ . Moreover, because  $t$  is continuous, we can ignore the distinction between strict and weak inequalities:  $P[t < 5]$  is the same as  $P[t \leq 5]$ . Probabilities related to ranges of a random variable are addressed with the Cumulative Distribution Function (CDF).

The Cumulative Distribution Function  $F(t)$  for the exponential distribution is given by:

Eq. 3:18

$$F(t_0) = P[t \leq t_0] = 1 - e^{-\lambda t_0}, t > 0$$

### Example Problem:

For Cheryl's Burger-Max restaurant, assume customer arrivals during the "lunch rush" follow the exponential distribution and that, on average, there are 45 customer arrivals per hour.

- a) What is the probability that more than 2 minutes will elapse between customer arrivals?

Here  $\lambda=45$  and  $t$  is measured in hours, so 2 minutes is  $(2/60)$  hours.

$$P[t > 2] = 1 - P[t \leq 2] = 1 - \left(1 - e^{-(45)\left(\frac{2}{60}\right)}\right) = 1 - 0.7779 = 0.2231$$

In Excel: =1 - EXPON.DIST((2/60),45,1)

- b) What is the probability that 3 minutes or less will elapse between customer arrivals?

$$P[t \leq 3] = 1 - e^{(-45)\left(\frac{3}{60}\right)} = 0.8946$$

In Excel: =EXPON.DIST((3/60),45,1)

- c) What is the probability that between 1.5 minutes and 2.5 minutes will elapse between customer arrivals?

$$P[t \leq 2.5] - P[t \leq 1.5] = \left(1 - e^{-(45)\left(\frac{2.5}{60}\right)}\right) - \left(1 - e^{-(45)\left(\frac{1.5}{60}\right)}\right) =$$

$$P[t \leq 2.5] - P[t \leq 1.5] = e^{-(45)\left(\frac{1.5}{60}\right)} - e^{-(45)\left(\frac{2.5}{60}\right)} = 0.1713$$

In Excel: =EXPON.DIST((2.5/60),45,1) – EXPON.DIST((1.5/60),45,1)

## F. The Normal Distribution

The Normal Probability distribution is a symmetric bell-shaped distribution that is widely observed in nature and economics. A continuous random variable  $X$  with mean (expected value)  $\mu$  and variance  $\sigma^2$  follows the normal probability distribution if the PDF of  $X$  has the following mathematical form:

Eq. 3:19

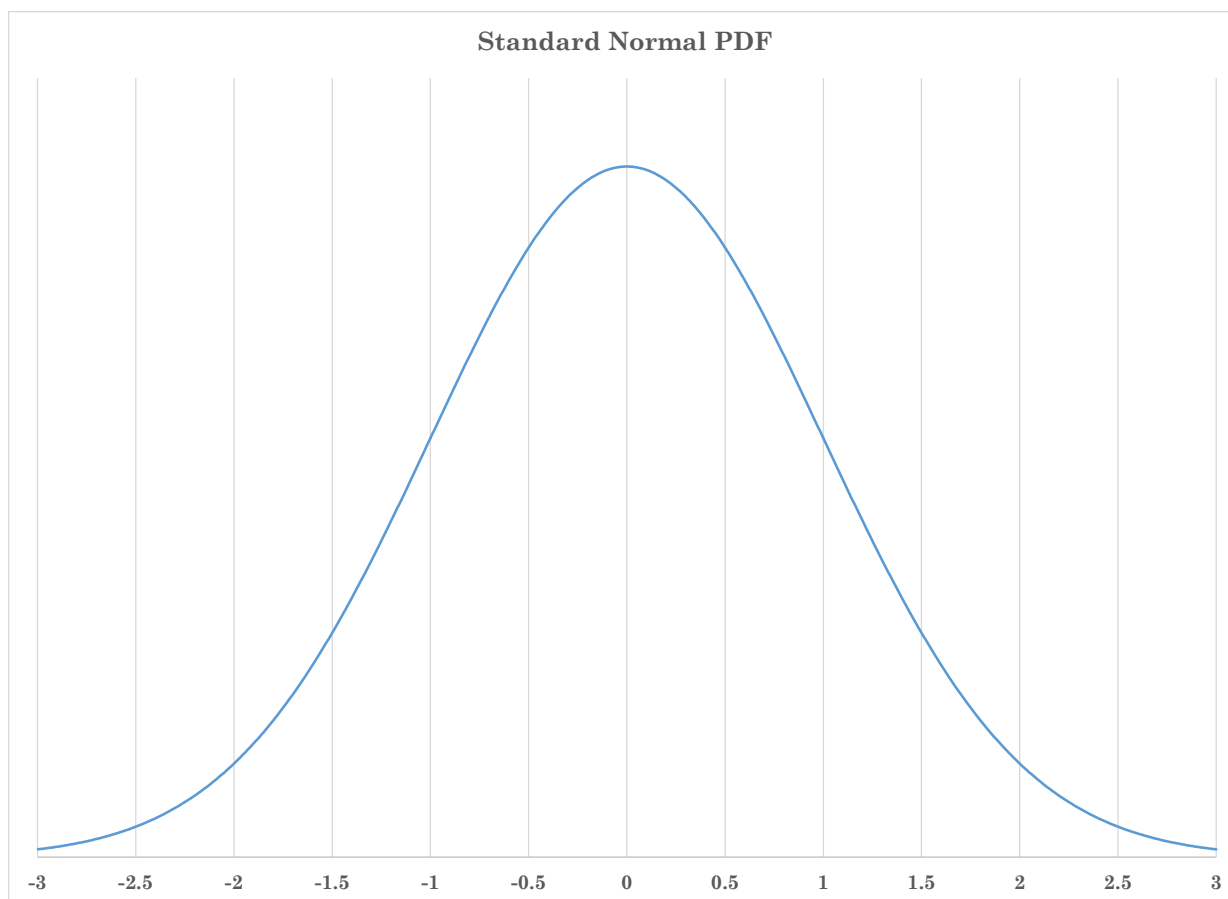
$$f(X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X-\mu)^2}{2\sigma^2}}$$

And is denoted  $X \sim N(\mu, \sigma^2)$ . Because of the mathematical form of the distribution, the integration required to derive the mean and variance cannot be directly solved, it must be computed numerically for different specific values of  $\mu$  and  $\sigma^2$ . However, because of the properties of linear functions of a random variable, any Normal random variable  $X$  can be **standardized** as its  $Z$ -score to get a Normal variable with a mean of zero and a standard deviation of 1. Thus, if  $X \sim N(\mu, \sigma^2)$ , then:

Eq. 3:20

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

(Apply the formulas in Eq. 3.8 above with  $a = -\mu/\sigma$  and  $b = 1/\sigma$  to show this.) So instead of having to compute probabilities for every value of  $\mu$  and  $\sigma^2$ , we can transform values of  $X$  into  $Z$ -scores and get the same probabilities from this standardized distribution. The PDF of the Standard Normal has the familiar symmetric bell-shaped distribution as shown in the following chart. The CDF is the area under the curve.



The probabilities for the Standard Normal CDF have been tabulated and are attached at the end of these notes. However, one of the benefits of modern computing is that almost all statistics software, including Excel, has algorithms for computing probabilities for the Normal Distribution.

Given the properties of the PDF and CDF above, for a given value of  $X$ ,  $X_0$ , we calculate  $Z_0 = \frac{X_0 - \mu}{\sigma}$ , and:

*Eq. 3:21*

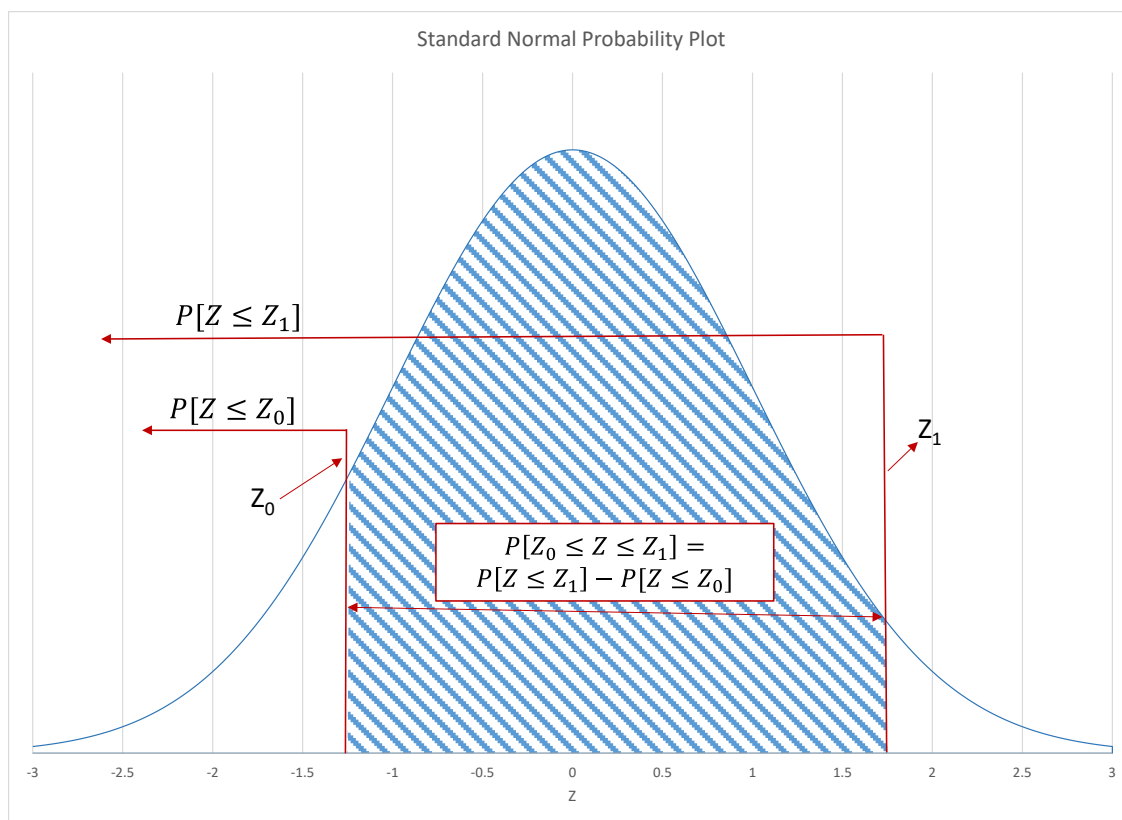
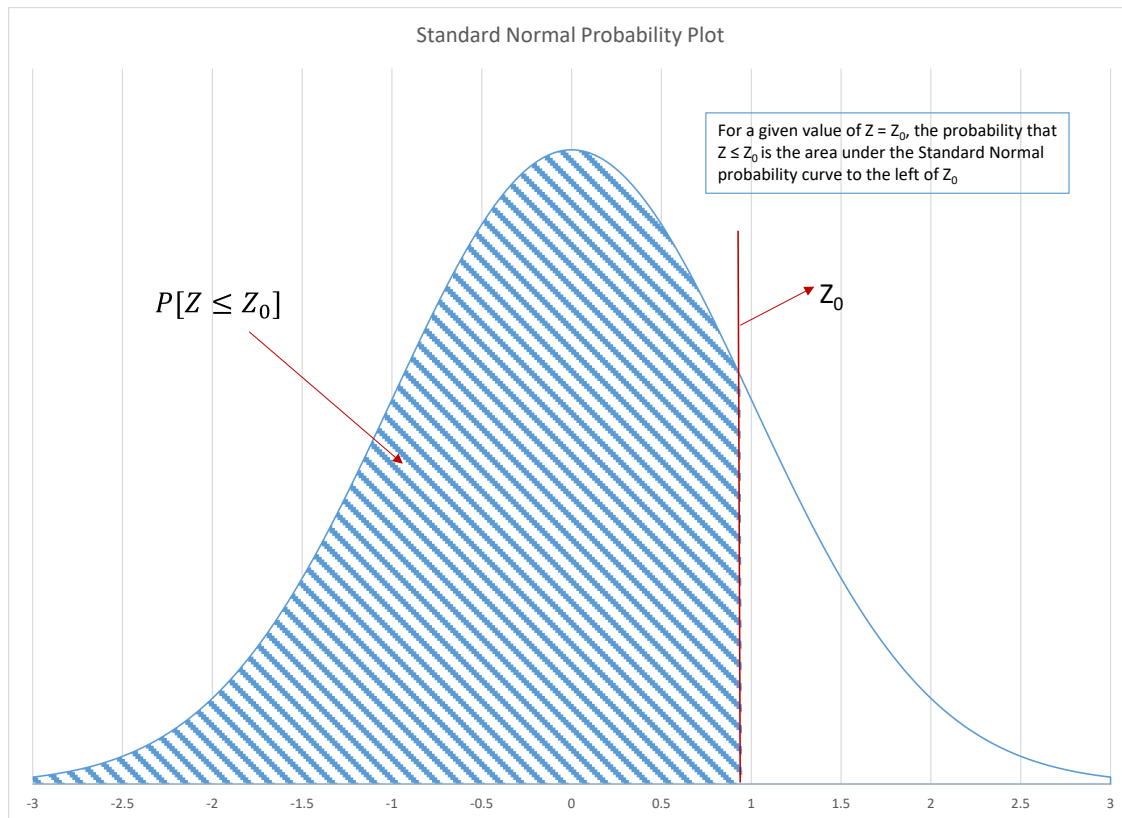
$$P[X \leq X_0] = P[Z \leq Z_0] = F(Z_0)$$

$$P[X_0 \leq X \leq X_1] = P[Z_0 \leq Z \leq Z_1] = F(Z_1) - F(Z_0)$$

In Excel, we can use `=NORM.S.DIST(Z,1)` to calculate probabilities for a given value of  $Z$  (again, the “1” tells Excel to use the cumulative distribution or CDF).

In addition, we can use `=NORM.S.INV(P)` in Excel to calculate  $Z$  for a given probability  $P$ .



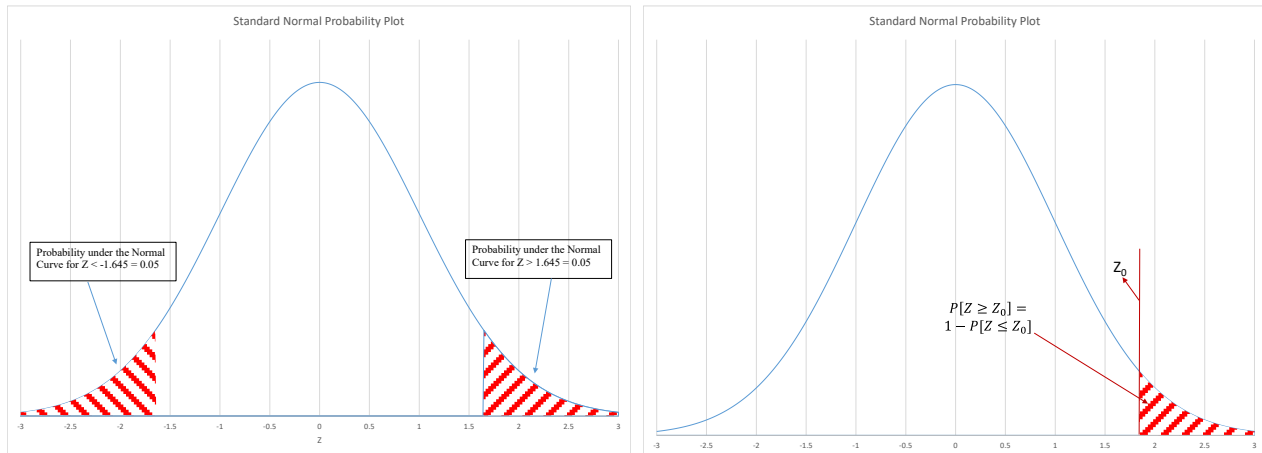


Additionally, because the Normal distribution is symmetric:

Eq. 3:22

$$P[Z \leq -Z_0] = F(-Z_0) = 1 - F(Z_0) = 1 - P[Z \leq Z_0]$$

$$P[Z \geq Z_0] = 1 - P[Z \leq Z_0]$$



For example, suppose  $X$  is distributed as a Normal random variable with  $\mu = 10$  and  $\sigma^2 = 4$  so  $\sigma = 2$ ;  $X \sim N(10, 4)$ . Then to get the probability  $P[X \leq 13.06]$ :

Eq. 3:23

$$Z_{13} = \frac{13.06 - 10}{2} = 1.53$$

In the attached table of Standard Normal probabilities, the first column is the value of  $Z$  to one decimal place and the body of the table shows the probabilities for successive values of the second decimal place of  $Z$ . For this example, go down to the  $Z$  value of 1.5 and the 0.03 column tells us that  $P[X \leq 13.06] = P[Z \leq 1.53] = 0.9370$ . Alternatively, in Excel: `=NORM.S.DIST(1.53,1)` would give the same result.

Because of symmetry, this also tells us that:

Eq. 3:24

$$P[X \geq 13.06] = P[Z \geq 1.53] = 1 - P[Z \leq 1.53] = 1 - 0.9370 = 0.0630$$

Similarly, to find  $P[X \leq 7.5]$ , calculate  $Z_{7.5} = -1.25$  and

Eq. 3:25

$$P[Z \leq -1.25] = 1 - P[Z \leq 1.25] = 1 - 0.8944 = 0.1056$$

You can verify this in Excel in that:

$$\text{NORM.S.DIST}(-1.25,1) = 1 - \text{NORM.S.DIST}(1.25,1)$$

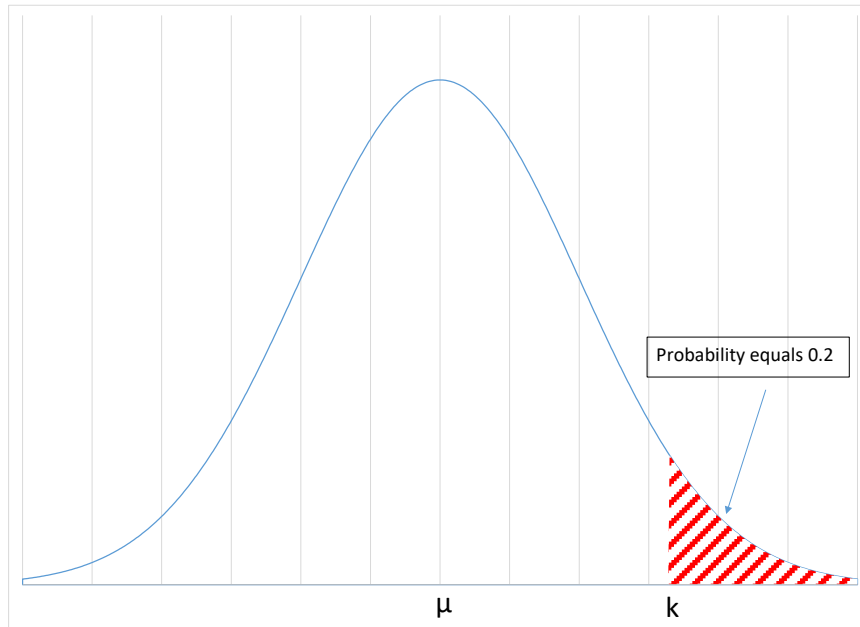
Note also that symmetry implies that  $P[Z \geq -1.25] = P[Z \leq 1.25] = 0.8944$ .

This type of problem can also be worked the other direction in that given a particular probability you could find the value of  $X$  above, below, or in an interval consistent

with that probability. For example, given  $X \sim N(10, 4)$ , (that is,  $X$  is distributed as a Normal random variable with mean  $\mu = 10$ , and variance  $\sigma^2 = 4$ ) you are told that the probability is 0.2 that  $X$  is greater than some number  $k$ . Thus, we want to find the number  $k$  such that:

Eq. 3:26

$$P[X > k] = 0.2$$



Because of symmetry, this information tells us that the probability that  $X$  is less than  $k$  is  $1 - 0.2 = 0.8$ . In Excel, use `NORM.S.INV(0.8)` to find  $Z \cong 0.8416$ . Then

Eq. 3:27

$$Z = 0.8416 = \frac{k - \mu}{\sigma} = \frac{k - 10}{2}$$

solving for  $k \cong 11.6832$

### G. Normal Approximation to the Binomial Distribution

As discussed above, a **Binomial** random variable is one that can only take on one of two values: for example, the flip of a coin coming up heads or tails, or among a sample of people whether an individual is male or female. One of the two events is defined as a “success” and for the given population, a success will occur with probability  $P$ . When sampling from a Binomial distribution, if the sample is “large” enough (the sample size  $n$  is such that  $nP(1 - P) > 5$ ) then the distribution of the total number of successes in the sample,  $X$ , is approximately Normal with:

Eq. 3:28

$$E(X) = \mu = nP$$

$$Var(X) = \sigma^2 = nP(1 - P)$$

We can make statements in probability about  $X$  by calculating Z-scores using  $nP$  for  $\mu$  and  $nP(1 - P)$  for  $\sigma^2$ . Similarly, dividing through the equations above by the sample size  $n$ , we can make statements in probability about the proportion of successes, which is approximately Normal with:

Eq. 3:29

$$E\left(\frac{X}{n}\right) = \mu = P$$

$$\text{Var}\left(\frac{X}{n}\right) = \sigma^2 = \frac{P(1 - P)}{n}$$

So that we calculate Z-scores using  $P$  for  $\mu$  and  $\frac{P(1-P)}{n}$  for  $\sigma^2$ .

## H. Sampling Distributions

The foregoing discussion relates to making probability statements about specific values of a Normal random variable. To make probability statements about sample statistics, such as the sample mean  $\bar{X}$ , we need to know the expected value (mean) and variance (and thus the standard deviation) of that particular sample statistic. For a random sample of size  $n$  for the variable  $X \sim N(\mu, \sigma^2)$  where the observed values of  $X$  are assumed to be independent from one another, the sample mean has the following distribution properties:

Eq. 3:30

$$E(\bar{X}) = E\left(\frac{\sum X_i}{n}\right) = \left(\frac{\sum E(X_i)}{n}\right) = \frac{n\mu}{n} = \mu$$

$$\text{Var}(\bar{X}) = \text{Var}\left(\frac{\sum X_i}{n}\right) = \left(\frac{\sum \text{Var}(X_i)}{n^2}\right) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Given these properties, we can make statements in probability about the sample mean  $\bar{X}$  by calculating Z-scores using the population mean  $\mu$  and the standard deviation of the *sample* mean  $\frac{\sigma}{\sqrt{n}}$ . For example, suppose you are told the random variable  $X$  follows a Normal distribution with mean  $\mu=15$  and variance  $\sigma^2=400$ . A random sample for  $X$  of size  $n=16$  is obtained. What is the probability that  $\bar{X} \leq 17$ ?

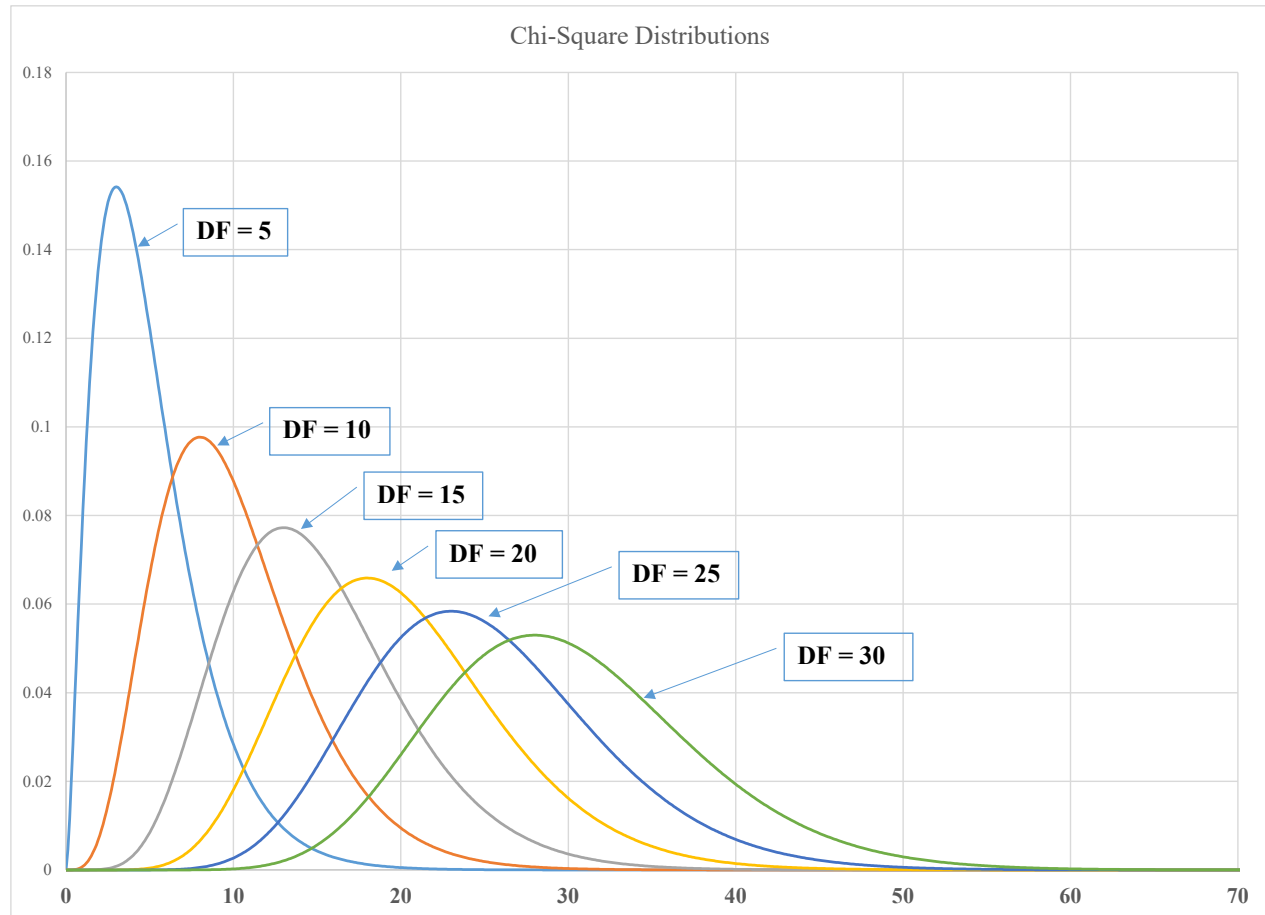
Eq. 3:31

$$Z = \frac{17 - 15}{\sqrt{400/16}} = 0.4$$

From the Normal probability table,  $P[Z \leq 0.4] = 0.6554$ .

To get the sampling distribution of the sample variance  $S^2$ , we need to introduce a probability distribution related to the Normal. If  $Z \sim N(0, 1)$  then  $Z^2$  is distributed as a Chi-Squared random variable with 1 degree of freedom. This is denoted  $\chi^2_{(1)}$ . If  $Z_1, Z_2, \dots, Z_n$  are independently distributed  $N(0, 1)$  then  $\sum Z_i^2 \sim \chi^2_{(n)}$  (Chi-Squared with  $n$  degrees of freedom). The Chi-Squared distribution is strictly positive (because it is

based on  $Z^2$ ), it is not symmetric, and it changes shape based on the degrees of freedom:



Given a random variable  $X \sim N(\mu, \sigma^2)$ , the sampling distribution of the sample variance  $S^2$  based on a sample of  $n$  observations has the following distribution properties:

Eq. 3:32

$$E(S^2) = \sigma^2$$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

That is, on average, the sample variance  $S^2$  is equal to the true population variance  $\sigma^2$ , and the ratio  $\frac{(n-1)S^2}{\sigma^2}$  is distributed as a Chi-Square random variable with  $(n-1)$  degrees of freedom.

For example, suppose you are told that the random variable  $X$  follows the Normal distribution with a **standard deviation**  $\sigma = 20$ . A random sample of  $n = 35$  observations is obtained. What is the probability of finding a sample variance  $S^2$  less than 300?

Given the sampling distribution of the sample variance, calculate:

Eq. 3:33

$$\chi^2_{calc} = \frac{(n-1)S^2}{\sigma^2} = \frac{(35-1) \times 300}{20^2} = 25.5$$

Next, in Excel use =CHISQ.DIST(25.5,35-1,1) to get  $\cong 0.1471$ . Because all probabilities must sum to 1, the probability that  $S^2 > 300$  is  $1 - 0.1471 = 0.8529$ .

To use the Normal approximation for the Binomial distribution to make probability statements about a sample proportion  $\hat{P}$  (X “successes” in a sample of size n gives a sample proportion  $\hat{P} = X/n$ ), the mean and variance are the same as those given in Eq. 3:29, above, but we use the sample proportion  $\hat{P}$  for the mean and in the calculation of the standard deviation. For example, suppose you are told that out of a sample of 500 voters, 52% ( $\hat{P}$ ) say they intend to vote for Candidate M. What is the probability that Candidate M gets 50% or more of the votes?

Eq. 3:34

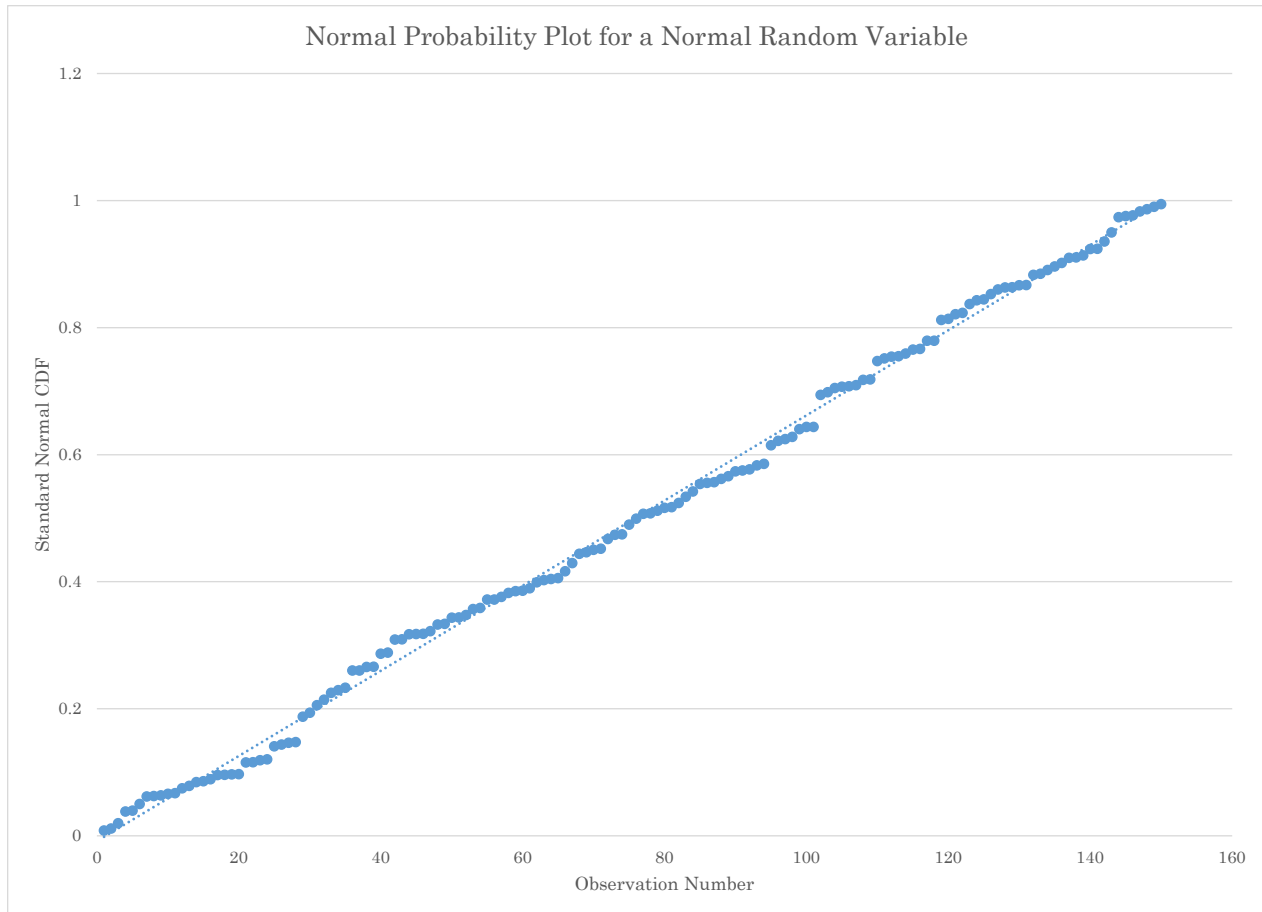
$$Z = \frac{0.50 - 0.52}{\sqrt{\frac{0.52(1 - 0.52)}{500}}} = -0.895$$

Since the Z is negative but we want to know the probability that the proportion is *greater* than 50%, we get the answer from  $P[Z \leq 0.895] \approx 0.815$ .

In the preceding discussion of the Normal Distribution, I have tried to be consistent in using weak inequalities ( $\leq$ ) in the probability statements. However, with continuous probability distributions, there is no real difference between strict and weak inequalities. For a continuous distribution such as the Normal, the probability of any *specific* value of the random variable is technically zero – think of the distinction of a variable being equal to 1.5 versus 1.5000001 or 1.49999999. Thus, statements such as  $P[Z \leq 0.895]$  and  $P[Z < 0.895]$  are essentially the same.

### I. Normal Probability Plots

The Normal Probability Distribution is the most widely used probability model in statistics. In practice however, it is not uncommon to deal with random variables whose behavior is quite different than the Normal model. If we apply the Normal Probability model in such a situation, it will most likely lead to incorrect inferences about the nature of the population. Thus, we need some method to assess whether the variable we are analyzing approximately follows the Normal Distribution in its behavior. The chart below shows a Normal Probability Plot for 150 observations on a random variable known to Normal.



The vertical axis in this chart is the value of the Standard Normal CDF for each observation in the data. Constructing this type of chart in Excel is quite straightforward. First, sort the data from smallest to largest and calculate a Z-Score for each observation by subtracting the sample mean and dividing by the sample standard deviation. Next use the `NORM.S.DIST(Z-Score,1)` function to get the value of the Standard Normal CDF for each observation. Finally, plot these results against the (sorted) observation number. If the data are close to Normal, the plot will appear similar to that above with the observations falling approximately along a straight line. By way of contrast, the County Labor Force data discussed in Section 1 has a distribution that is skewed sharply to the right. The Normal Probability Plot for this variable is as follows.





**J. Example Problems**

Let the random variable  $X$  follow a Normal distribution with mean  $\mu = 50$  and variance  $\sigma^2 = 64$ .

1. Find the probability that  $X$  is greater than 60.

*Answer:* Z-Score  $\implies Z_{score} = \frac{60-50}{8} = 1.25$

$P[Z > Z\text{-Score}] \implies 1 - P[Z < 1.25] \cong 0.1056$

2. Find the probability that  $X$  is greater than 35 and less than 62.

*Answer:* Z-Score1 for  $X$  greater than 35  $\implies Z_{score1} = \frac{35-50}{8} = -1.875$

Z-Score2 for  $X$  less than 62  $\implies Z_{score2} = \frac{62-50}{8} = 1.5$

$P[Z\text{-Score1} < Z < Z\text{-Score2}] \implies P[Z < 1.5] - P[Z < -1.875] \cong 0.9028$

3. Find the Probability that  $X$  is less than 55.

*Answer:* Z-Score  $\implies Z_{score} = \frac{55-50}{8} = 0.63$

$P[Z < Z\text{-Score}] \implies P[Z < 0.63] \cong 0.73$

4. The probability is 0.2 that  $X$  is greater than what number?

*Answer:* Z-Score  $\implies Z_{score}$  such that  $P[Z < Z_{score}] = 0.80$   $Z_{score} \cong 0.8416$

Solve for  $X \implies$  Solve  $\frac{X-50}{8} = 0.8416$  for  $X \cong 56.7330$

5. The probability is 0.05 that  $X$  is outside a symmetric interval about the mean between what number  $K$  from  $\mu \pm K$ ?

*Answer:* Z-Score  $\implies Z_{score}$  such that  $P[Z < Z_{score}] = 0.975$   $Z_{score} \cong 1.96$

Solve for  $K \implies$  Solve  $\frac{\mu+K-\mu}{\sigma} = \frac{K}{8} = 1.96$  for  $K \cong 15.6797$

.....  
It is known that 10% of all items produced by a particular manufacturing process are defective. From the very large output of a single day, 400 items are selected at random.

6. What is the probability that at least 35 of the 400 selected items are defective (the number of defectives items  $X$  is greater than or equal to 35)?

*Answer:* Expected mean defective items?  $\implies E(X) = P \times n = 0.1 \times 400 = 40$

Variance of the number of defective items?  $\implies Var(X) = n \times P \times (1 - P) = 36$

Z-Score  $\implies Z_{score} = \frac{35-40}{6} \cong -0.8333$

$P[Z > Z\text{-Score}] \implies 1 - P[Z < -0.8333] \cong 0.7977$

7. What is the probability that between 40 and 50 of the selected items are defective?

Answer: Z-Score1 for X greater than 40  $\implies Z_{score1} = \frac{40-40}{6} = 0$

Z-Score2 for X less than 50  $\implies Z_{score2} = \frac{50-40}{6} \cong 1.6667$

$P[Z\text{-Score1} < Z < Z\text{-Score2}] \implies P[Z < 1.667] - P[Z < 0] \cong 0.4522$

.....  
Let the random variable X follow a Normal distribution with mean  $\mu = 200$  and variance  $\sigma^2 = 625$ . A random sample of  $n = 50$  is obtained.

8. What are the mean and variance of the sample mean,  $\bar{X}$ ?

Answer: Mean  $\implies E(\bar{X}) = \mu = 200$  Variance  $\implies Var(\bar{X}) = \frac{\sigma^2}{n} = 12.5$

9. What is the probability that  $\bar{X}$  is greater than 204?

Answer: Z-Score for  $\bar{X}$  greater than 204  $\implies Z_{score} = \frac{204-200}{\sqrt{625/50}} \cong 1.1314$

$P[Z > Z\text{-Score}] \implies 1 - P[Z < 1.1314] \cong 0.1289$

10. What is the probability that  $\bar{X}$  is between 198 and 211?

Answer: Z-Score1 for  $\bar{X}$  greater than 198  $\implies Z_{score1} = \frac{198-200}{\sqrt{625/50}} \cong -0.5657$

Z-Score2 for  $\bar{X}$  less than 211  $\implies Z_{score2} = \frac{211-200}{\sqrt{625/50}} \cong 3.1127$

$P[Z\text{-Score1} < Z < Z\text{-Score2}] \implies P[Z < 3.1127] - P[Z < -0.5657] \cong 0.7133$

11. What is the probability that the sample variance  $S^2 < 500$ ?

Answer: Calculate  $\chi^2_{calc} = \frac{(49)(300)}{625} = 39.2$ . In Excel, use =CHISQ.DIST(39.2,49,1) to get  $\sim 0.1596$

12. What is the probability that the sample variance  $S^2 > 800$ ?

Answer: Calculate  $\chi^2_{calc} = \frac{(49)(800)}{625} = 62.72$ . In Excel, use =1-CHISQ.DIST(62.72,49,1) to get  $\sim 0.0901$ .

13. What is the probability that the sample variance is between 500 and 800?

Answer: Given the values for  $\chi^2_{calc} = \frac{(49)(300)}{625} = 39.2$ , and  $\chi^2_{calc} = \frac{(49)(800)}{625} = 62.72$ , from above, in Excel, use =CHISQ.DIST(62.72,49,1)-CHISQ.DIST(39.2,49,1) to get  $\sim 0.7504$

.....

The patient mix for a large group of hospitals is such that 46% of the patients have some type of government sponsored health insurance. A random sample of 200 patients is obtained.

14. What are the mean and variance of the sample proportion of patients with government sponsored health insurance  $\hat{P}$ -Hat?

*Answer:* Mean  $\implies E(\hat{P}) = P = 0.46$  Variance  $\implies Var(\hat{P}) = \frac{P(1-P)}{n} \cong 0.001242$

15. What is the probability that the sample proportion of patients with government sponsored health insurance  $\hat{P}$ -Hat is greater than 50%

*Answer:* Z-Score for  $\hat{P}$ -Hat greater than 50%  $\implies Z_{score} = \frac{0.5-0.46}{\sqrt{0.001242}} \cong 1.135$

$P[Z > Z\text{-Score}] \implies 1 - P[Z < 1.135] \cong 0.1282$

16. What is the probability that the sample proportion of patients with government sponsored health insurance  $\hat{P}$ -Hat is between 42% and 48%?

*Answer:* Z-Score1 for  $\hat{P}$ -Hat greater than 42%  $\implies Z_{score1} = \frac{0.42-0.46}{\sqrt{0.001242}} \cong -1.135$

Z-Score2 for  $\hat{P}$ -Hat less than 48%  $\implies Z_{score} = \frac{0.48-0.46}{\sqrt{0.001242}} \cong 0.5675$

$P[Z\text{-Score1} < Z < Z\text{-Score2}] \implies P[Z < 0.5675] - P[Z < -1.135] \cong 0.5866$

## Section 4. ESTIMATION AND HYPOTHESIS TESTING

### A. Sampling Distributions

We introduced the notion of Sampling Distributions in Section 3, above, but the topic bears repeating. Given a random sample of a random variable  $X$  that has a constant mean  $\mu$  and constant variance  $\sigma^2$  (denoted  $X \sim (\mu, \sigma^2)$ ) the **Sampling Distribution** is the PDF of the various sample statistics such as the sample mean, sample variance, and sample proportion. Given the random variable  $X \sim (\mu, \sigma^2)$ , the sample mean  $\bar{X}$  has the following distribution properties:

Eq. 4:1

$$\begin{aligned}\bar{X} &= \frac{\sum X_i}{n} \\ E(\bar{X}) &= \mu \\ \text{Var}(\bar{X}) &= \frac{\sigma^2}{n} \\ Z &= \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim (0, 1)\end{aligned}$$

The **Central Limit Theorem** states that for a random variable  $X \sim (\mu, \sigma^2)$ , as the sample size  $n$  becomes “large” then  $Z$  as defined above is approximately Normally distributed. Thus, for *any* random variable with a constant mean and variance, we can make statements in probability about the sample mean based on the Normal probability distribution when we have a sufficiently large sample (usually, 20 or more observations is a sufficiently large sample).

If  $X$  is a binomial random variable that takes on a value of one (a “success”) with probability  $P$  and is zero with probability  $(1 - P)$ , the sample proportion  $\hat{P} = \sum X / n$  has the following distribution properties:

Eq. 4:2

$$\begin{aligned}E(\hat{P}) &= P \\ \text{Var}(\hat{P}) &= \frac{P(1 - P)}{n} \\ Z &= \frac{\hat{P} - P}{\sqrt{P(1 - P)/n}} \sim N(0, 1) \text{ if } n \text{ is "large"}\end{aligned}$$

To get the sampling distribution of the sample variance  $S^2$ , we need to introduce a probability distribution related to the Normal. If  $Z \sim N(0, 1)$  then  $Z^2$  is distributed as a Chi-Squared random variable with 1 degree of freedom. This is denoted  $\chi^2_{(1)}$ . If  $Z_1, Z_2, \dots, Z_n$  are independently distributed  $N(0, 1)$  then  $\sum Z_i^2 \sim \chi^2_{(n)}$  (Chi-Squared with  $n$  degrees of freedom). The sample variance then has the following distribution properties:

Eq. 4:3

$$E(S^2) = \sigma^2$$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

The Chi-Square distribution is not symmetric so you must be careful to identify whether you are looking at lower tail versus upper tail probabilities. A copy of the Chi-Square probability tables are attached at the end of these notes. In Excel, the CHISQ.DIST(*stat, degrees-of-freedom, cumulative*) returns the “left tail” probability (the probability of finding a value less than “stat”) under the Chi-Square distribution if cumulative is set to yes (1). CHISQ.DIST.RT returns the “right tail” probability.

## B. Confidence Intervals: Single Sample

The forgoing distribution properties allow us to make probability statements about the various sample statistics similar to those we constructed for Normal random variables. In particular, we can construct **Confidence Intervals** for the various sample statistics. For the random variable  $X \sim (\mu, \sigma^2)$  where the population variance  $\sigma^2$  is known, given a sufficiently large sample of size  $n$ , we can construct the Z-score for the sample mean:

Eq. 4:4

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Now if we choose a **Confidence Level**,  $(1 - \alpha)$  where  $\alpha$  is the chosen significance level between zero and one: for example, for a 90% confidence level,  $\alpha=0.1$  or 10%. Next, split  $\alpha$  in half so that there is a probability of  $\alpha/2$  in each tail of the Normal distribution. Finding the Z value associated with  $\alpha/2$  then allows us to set up the following:

Eq. 4:5

$$P \left[ -Z_{\frac{\alpha}{2}} \leq Z \leq Z_{\frac{\alpha}{2}} \right] = 1 - \alpha$$

$$P \left[ -Z_{\frac{\alpha}{2}} \leq \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq Z_{\frac{\alpha}{2}} \right] = 1 - \alpha$$

Rearranging the terms inside the brackets, we can say that there is a probability of  $(1 - \alpha)$  that the population mean  $\mu$  is in the interval:

Eq. 4:6

$$\bar{X} - \frac{\sigma}{\sqrt{n}} Z_{\frac{\alpha}{2}} \leq \mu \leq \bar{X} + \frac{\sigma}{\sqrt{n}} Z_{\frac{\alpha}{2}}, \text{ or}$$

$$\mu \text{ is in the interval: } \bar{X} \pm \frac{\sigma}{\sqrt{n}} Z_{\frac{\alpha}{2}}, \text{ where}$$

$$\text{the Upper Confidence Limit: } UCL = \bar{X} + \frac{\sigma}{\sqrt{n}} Z_{\frac{\alpha}{2}}$$

the Lower Confidence Limit:  $LCL = \bar{X} - \frac{\sigma}{\sqrt{n}} Z_{\frac{\alpha}{2}}$ , and

the Margin of Error:  $ME = \frac{\sigma}{\sqrt{n}} Z_{\frac{\alpha}{2}}$

For example, suppose the random variable X has a known variance  $\sigma^2=144$ . With a random sample of 25 observations, if we wanted to construct of 90% confidence interval, we need to find the value of Z such there is 5% in the upper right tail:

$P\left[Z \geq Z_{\frac{\alpha}{2}}\right] = 0.05$ , or  $1 - P\left[Z \leq Z_{\frac{\alpha}{2}}\right] = 0.95$ . From the Normal probability table  $Z_{\alpha/2} \approx 1.645$ . Because of symmetry of the Normal this also tells us that  $P\left[Z \leq -Z_{\frac{\alpha}{2}}\right] = 0.05$ .

Therefore, there is a 90% probability that the population mean is in the interval:

Eq. 4:7

$$\begin{aligned} & \bar{X} \pm \frac{12}{\sqrt{25}} 1.645 \\ & = \bar{X} \pm (2.4)(1.645) \\ & = \bar{X} \pm 3.948 \end{aligned}$$

If the population variance for our random variable X is **unknown**, we need to use yet another distribution related to the Normal: the **t-distribution**. The t-distribution is a ratio of a Standard Normal random variable and the square root of a Chi-Square random variable and makes use of the sampling distribution of the sample variance in Eq. 4:3 above. Like the Normal distribution, the t-distribution is symmetric and gets closer and closer to the Normal as the sample size increases. Replacing the population standard deviation  $\sigma$  in our Z-score with the sample standard deviation S, we get the  $t_{Stat}$ :

Eq. 4:8

$$t_{stat} = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{(n-1)}$$

A t-distribution with (n-1) degrees of freedom. The interval for a  $(1 - \alpha)\%$  confidence level is then given by:

Eq. 4:9

$$\bar{X} \pm \frac{S}{\sqrt{n}} t_{(n-1), \alpha/2}$$

Thus, we replace the population standard deviation with the sample standard deviation and use values out of the table of probabilities for the t-distribution rather than the standard Normal distribution. A copy of the probability tables for the t-distribution is attached at the end of these notes.

For the sample proportion from a binomial random variable  $\hat{P}$  (see Eq. 4:2, above) if the sample size n is “large” ( $n\hat{P}(1 - \hat{P}) > 5$ ) then the  $(1 - \alpha)\%$  confidence interval is calculated from:

Eq. 4:10

$$\hat{p} \pm Z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

For a confidence interval of the sample variance,  $S^2$ , we need the Chi-Square values for the Upper tail corresponding to a probability of the  $(1 - \alpha/2)$ , and the Lower tail Chi-Square value corresponding to a probability of  $\alpha/2$ . The confidence interval is calculated from:

Eq. 4:11

$$\frac{(n - 1)S^2}{\chi_{(n-1),Upper Tail}^2} \leq \sigma^2 \leq \frac{(n - 1)S^2}{\chi_{(n-1),Lower Tail}^2}$$

Note that the Chi-Square value from the “Upper Tail” will be associated with the LCL and the value from the “Lower Tail” will be associated with the UCL.

### C. Hypothesis Testing: One Sample Tests

The preceding discussion of Confidence Intervals is directly related to formal **Hypothesis Testing**. The hypothesis to be tested might specify a specific single value, “the mean of variable X is 7,” or any other specific value, including zero, for example. Or it could involve the relationship between the sample statistics from two (or more) random variables, “the mean for X is twice as large as the mean for Y,” for example. In either case, we start by setting up a **Null Hypothesis**. For the specific value example, we would state the Null, denoted  $H_0$ , as:

Eq. 4:12

$$H_0: \mu_X = 7$$

We will then use the same information we used to construct our confidence intervals above to reach a conclusion on whether the **Reject** or **Fail to Reject** the Null Hypothesis (a Null Hypothesis is *never* “accepted”). In drawing our conclusion, we have to compare our Null to an **Alternative Hypothesis**, denote  $H_A$ . The Alternative can be **one-sided** or **two-sided**. Continuing with the specific value example:

Eq. 4:13

$$\begin{aligned} \text{One sided Alternatives: } & H_A: \mu_X > 7, \text{ or } H_A: \mu_X < 7 \\ \text{Two sided Alternative: } & H_A: \mu_X \neq 7 \end{aligned}$$

One-sided Alternatives are sometimes termed “directional” hypotheses. The two-sided alternative is basically, “the Null Hypothesis is not true.”

In constructing our confidence intervals, we chose a probability  $\alpha$  that defined our Confidence Level:  $1 - \alpha$ ;  $\alpha$  is also often called the “significance level” of the test. In the context of hypothesis testing, the value of  $\alpha$  is the probability of Rejecting the

Null Hypothesis when in fact the Null is true. This is called a Type I error and we control the probability of a Type I error by choosing the value of  $\alpha$ . If we Fail to Reject the Null when in fact the Null is false, this is known as a Type II error and has a probability  $\beta$ . The **Power** of a test is  $(1 - \beta)$ . There is a tradeoff between the significance level of a test and its power, as  $\alpha$  gets smaller  $\beta$  will get larger.

Decision	States of Nature	
	Null Hypothesis is True	Null Hypothesis is False
Fail to Reject $H_0$	Correct Decision $Prob = 1 - \alpha$	Type II Error $Prob = \beta$
Reject $H_0$	Type I Error $Prob = \alpha$	Correct Decision $Prob = 1 - \beta$

From the confidence intervals we defined above, we use the same information for calculate Z-Statistics or t-Statistics depending on what information is available, that I will denote  $Z_{calc}$  and  $t_{calc}$ . To continue the specific value example, suppose X has a known population variance  $\sigma^2$  and the sample size is n. Then to test  $H_0: \mu_x = 7$

Eq. 4:14

$$Z_{calc} = \frac{\bar{X} - 7}{\frac{\sigma}{\sqrt{n}}}$$

To test this hypothesis against the two-sided Alternative  $H_A: \mu_x \neq 7$ , we choose a significance level  $\alpha$  and compare  $Z_{calc}$  to  $Z_{\alpha/2}$ .  $Z_{\alpha/2}$  is called the “critical” Z. Our decision rule regarding the Null is:

Eq. 4:15

$$\begin{aligned} \text{Reject } H_0 \text{ if } Z_{calc} > Z_{\frac{\alpha}{2}} \text{ or } Z_{calc} < -Z_{\frac{\alpha}{2}} \\ \text{Fail to Reject if } -Z_{\frac{\alpha}{2}} < Z_{calc} < Z_{\frac{\alpha}{2}} \end{aligned}$$

That is, we reject  $H_0$  if  $Z_{calc}$  is in “one of the tails” of the Normal distribution. If we have a one-sided alternative, then instead of splitting our significance into two tails, we place all the probability in one tail and compare  $Z_{calc}$  to  $Z_{\alpha}$  for the Alternative  $H_A: \mu_x > 7$ , and to  $-Z_{\alpha}$  for the Alternative  $H_A: \mu_x < 7$ . Our decision rule regarding the Null is:

Eq. 4:16

$$\begin{aligned} \text{For } H_A: \mu_x > 7, \text{ Reject } H_0 \text{ if } Z_{calc} > Z_{\alpha}, \text{ otherwise Fail to Reject} \\ \text{For } H_A: \mu_x < 7, \text{ Reject } H_0 \text{ if } Z_{calc} < -Z_{\alpha}, \text{ otherwise Fail to Reject} \end{aligned}$$

It bears repeating that this applies to *any* specific value hypothesis. Instead of comparing  $Z_{calc}$  to a critical Z,  $Z_{\alpha}$  or  $Z_{\alpha/2}$ , we can rearrange our test statistic to



calculate a critical  $\bar{X}$  denoted  $\bar{X}_{crit}$ . To test  $H_0: \mu_X = \mu_0$ , where  $\mu_0$  is any specific value, assuming the variance of  $X$  is known, then:

Eq. 4:17

$$\begin{aligned} \text{for } H_A: \mu_X > \mu_0: \bar{X}_{crit} &= \mu_0 + Z_\alpha \frac{\sigma}{\sqrt{n}} \\ &\text{Reject } H_0 \text{ if } \bar{X} > \bar{X}_{crit} \\ \text{for } H_A: \mu_X < \mu_0: \bar{X}_{crit} &= \mu_0 - Z_\alpha \frac{\sigma}{\sqrt{n}} \\ &\text{Reject } H_0 \text{ if } \bar{X} < \bar{X}_{crit} \\ \text{for } H_A: \mu_X \neq \mu_0: \bar{X}_{crit} &= \mu_0 \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \\ \text{Reject } H_0 \text{ if } \bar{X} > \mu_0 + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, &\text{ or } \bar{X} < \mu_0 - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \end{aligned}$$

Note that with the two-sided alternative there is an “upper” critical value and a “lower” critical value.

Suppose you are given a sample of  $n=40$  observations of the random variable  $X$  and told that the population variance is known:  $\sigma^2 = 30$ . Test the hypothesis that the population mean is equal to 12 versus the alternative that it is greater than 12:  $H_0: \mu_X = 12$  vs.  $H_A: \mu_X > 12$ , at the  $\alpha=0.10$  level of significance. Although we are not given the value of the sample mean  $\bar{X}$ , we have enough information to calculate the critical value  $\bar{X}_{crit}$ . Since this is an “upper” one-sided alternative we find  $Z_\alpha \approx 1.282$ , and

Eq. 4:18

$$\bar{X}_{crit} = \mu_0 + Z_\alpha \frac{\sigma}{\sqrt{n}} = 12 + 1.282 \sqrt{\frac{30}{40}} \approx 13.12$$

Thus, our decision rule would be to reject the null if the sample mean is greater than the critical value: Reject if  $\bar{X} > \bar{X}_{crit} = 13.12$

Altering the example slightly, suppose you are told the sample mean  $\bar{X} = 10.9$  and are asked to test  $H_0: \mu_X = 12$  vs.  $H_A: \mu_X < 12$ , at the  $\alpha=0.05$  level of significance. With this “lower” one-sided alternative, we find  $-Z_\alpha \approx -1.645$ , and

Eq. 4:19

$$Z_{calc} = \frac{10.9 - 12}{\sqrt{30/40}} \approx -1.27$$

Since  $-1.27 > -1.645$ ,  $Z_{calc} > -Z_\alpha$ , and we Fail to Reject the null hypothesis.

For a specific test of a population proportion from a “large” sample of size  $n$ , the null hypothesis would take the form  $H_0: P = P_0$ , and given the sample proportion  $\hat{P}$

$$Z_{calc} = \frac{\hat{P} - P_0}{\sqrt{P_0(1 - P_0)/n}}$$

Note that the variance in the denominator of this statistic is based on the hypothesized value of the proportion,  $P_0$ . We then compare  $Z_{calc}$  to  $Z_\alpha$  or  $-Z_\alpha$  for one-sided tests, or to  $\pm Z_{\alpha/2}$  for a two-sided test. Or, for a given “critical”  $Z_{crit}$  ( $Z_\alpha$  or  $Z_{\alpha/2}$ ) we can compare  $\hat{P}$  to a critical  $\hat{P}_{crit}$

Eq. 4:20

$$\hat{P}_{crit} = P_0 \pm Z_{crit} \sqrt{\frac{P_0(1 - P_0)}{n}}$$

Use the “+” in the above for an upper tailed one-sided test, use the “-” for a lower tailed one-sided test, and use both for a two-sided test. For example, you are told that out of a sample of 400 items from a production line, 25 are defective. Test that hypothesis that the population proportion of defective items is 4% against a two-sided alternative at the  $\alpha=0.01$  level of significance. The sample proportion  $\hat{P} = \frac{400}{25} = 0.0625$ , and

Eq. 4:21

$$Z_{calc} = \frac{0.0625 - 0.04}{\sqrt{(0.04)(0.96)/400}} = 2.296$$

Since this is a two-sided alternative, we find  $Z_{0.005} \approx 2.58$  and since

Eq. 4:22

$$-2.58 < 2.29 < 2.58$$

We Fail to Reject the Null. Note that if our significance level had been  $\alpha=0.05$ , then  $Z_{\alpha/2}=1.96$  and we would have Rejected the null hypothesis.

As with our Confidence Intervals, the detail of how to calculate the test statistic depends on what information is known or must be estimated. To test the mean of a random variable  $X$  with an **unknown** variance, given the sample mean  $\bar{X}$ , the sample variance  $S^2$ , and the sample size  $n$ , we calculate a t-statistic,  $t_{calc}$

Eq. 4:23

$$H_0: \mu_X = \mu_0$$

$$t_{calc} = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

Then compare  $t_{calc}$  to  $t_{(n-1, \alpha)}$  or  $t_{(n-1, \alpha/2)}$  depending on whether it is a one-sided or two-sided alternative. For example, if  $n=25$  and  $\alpha=0.05$ , we compare  $t_{calc}$  to  $t_{(24, 0.05)}=1.711$  for a one-sided alternative and to  $t_{(24, 0.025)}=2.064$  for a two-sided alternative.

For tests of the variance of a random variable  $X$  we calculate a Chi-Square statistic:

Eq. 4:24

$$H_0: \sigma^2 = \sigma_0^2$$

$$\chi_{calc}^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

Compare to  $\chi_{(n-1, \alpha)}^2$  or  $\chi_{(n-1, \frac{\alpha}{2})}^2$

#### D. Confidence Intervals: Two Sample

We now turn to intervals for the difference between the sample statistics for two random variables. Suppose you have a random variable X with mean  $\mu_X$  and variance  $\sigma_X^2$ :  $X \sim (\mu_X, \sigma_X^2)$ ; and the random variable Y with mean  $\mu_Y$  and variance  $\sigma_Y^2$ :  $Y \sim (\mu_Y, \sigma_Y^2)$ . The nature of the sampling distributions for the difference between the *sample* statistics of these two random variables will depend on whether the samples are *dependent* or *independent*.

##### Dependent Samples

A dependent sample is such that each specific observation in the sample for X can be linked with a specific observation in the sample for Y. Consider the following example. A sample of n students is given an aptitude test and the scores are recorded. Call this initial set of scores the random variable X. The *same* set of students then participate in a test preparation course, after which they retake the aptitude test. Call the recorded set of scores for this second round of testing the random variable Y. The score for each specific student in the first round of testing,  $X_i$ , can be linked to that student's score on the second round of testing,  $Y_i$ . Define the *difference* between the two test scores as:  $d_i = X_i - Y_i$ . In this example, I have defined the difference as “before” minus “after” but it can be defined the other way as well. The difference variable  $d_i$  has the following properties:

Eq. 4:25

$$\bar{d} = \frac{\sum d_i}{n} = \frac{\sum (X_i - Y_i)}{n}$$

$$S_d^2 = \frac{\sum (d_i - \bar{d})^2}{n-1}$$

$$E(\bar{d}) = \mu_X - \mu_Y$$

$$Var(\bar{d}) = \frac{1}{n}(\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}) = \frac{\sigma_{\bar{d}}^2}{n}$$

Given the sample mean of the difference variable,  $\bar{d}$  and its sample variance  $S_d^2$ , and assuming the sample size is large enough to apply the Central Limit Theorem, the  $(1 - \alpha)\%$  confidence interval with significance for the difference in the population means  $(\mu_X - \mu_Y)$  is given by:

Eq. 4:26

$$\bar{d} - \frac{S_d}{\sqrt{n}} t_{(n-1), \frac{\alpha}{2}} \leq \mu_X - \mu_Y \leq \bar{d} + \frac{S_d}{\sqrt{n}} t_{(n-1), \frac{\alpha}{2}} \text{ or}$$

$$\bar{d} \pm \frac{S_d}{\sqrt{n}} t_{(n-1), \frac{\alpha}{2}}$$

$$\text{where the margin of error: } ME = \frac{S_d}{\sqrt{n}} t_{(n-1), \frac{\alpha}{2}}$$

Note that the interval is calculated using the sample standard deviation of the difference variable and uses the t-distribution with degrees of freedom based on the number of *pairs* of observations,  $n - 1$ .

### Independent Samples

Suppose we have two random variables X and Y with means and variances as defined in (1) above. Now, however, suppose that the samples for X and Y are independent of one another. To continue with a modified version of the testing example above, suppose a sample of  $n_X$  students who are Economics majors are administered the aptitude test and scores are recorded as the random variable X. A second sample of  $n_Y$  students who are Sociology majors are also administered the test and their scores recorded as the random variable Y. If the population variances of the two variables are **known**, then the difference in the sample means,  $\bar{X} - \bar{Y}$ , has the following properties:

Eq. 4:27

$$E(\bar{X} - \bar{Y}) = \mu_X - \mu_Y$$

$$Var(\bar{X} - \bar{Y}) = \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}$$

And the  $(1 - \alpha)\%$  confidence interval for the difference in the population means is given by:

Eq. 4:28

$$(\bar{X} - \bar{Y}) - \left( \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}} \right) Z_{\frac{\alpha}{2}} \leq \mu_X - \mu_Y \leq (\bar{X} - \bar{Y}) + \left( \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}} \right) Z_{\frac{\alpha}{2}}$$

$$(\bar{X} - \bar{Y}) \pm \left( \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}} \right) Z_{\frac{\alpha}{2}}$$

$$ME = \left( \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}} \right) Z_{\frac{\alpha}{2}}$$

Now suppose that the population variances for our two independent samples are **unknown but assumed to be equal**, so:  $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ . The variance of the difference in the sample means is now:

Eq. 4:29

$$Var(\bar{X} - \bar{Y}) = \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y} = \sigma^2 \left( \frac{1}{n_X} + \frac{1}{n_Y} \right)$$

Since the (assumed equal) population variance  $\sigma^2$  is unknown, we need a sample estimate. Given the sample variance for X and Y,  $S_X^2$  and  $S_Y^2$ , as an estimate of the (assumed equal) variance  $\sigma^2$  we calculate a **pooled** sample variance  $S_P^2$  as:

Eq. 4:30

$$S_P^2 = \frac{(n_x - 1)S_X^2 + (n_y - 1)S_Y^2}{n_x + n_y - 2}$$

And the  $(1 - \alpha)\%$  confidence interval for the difference in the population means is given by:

Eq. 4:31

$$\bar{X} - \bar{Y} \pm \left( \sqrt{\frac{S_P^2}{n_x} + \frac{S_P^2}{n_y}} \right) t_{(n_x+n_y-2), \frac{\alpha}{2}}$$

So we use a pooled estimate of the population variance and the t-distribution with degrees of freedom based on the sum of the sample sizes:  $(n_x + n_y - 2)$ .

Now suppose that the population variances for our two independent samples are **unknown but not assumed to be equal**. The variance of the difference in the sample means is now:

Eq. 4:32

$$\text{Var}(\bar{X} - \bar{Y}) = \frac{\sigma_X^2}{n_x} + \frac{\sigma_Y^2}{n_y}$$

Given the sample variances for X and Y,  $S_X^2$  and  $S_Y^2$ , the  $(1 - \alpha)\%$  confidence interval for the difference in the population means is similar to that in (9) above, but the extra uncertainty introduced by the unequal variances requires a complicated adjustment to the degrees of freedom for the t-distribution.

Eq. 4:33

$$\bar{X} - \bar{Y} \pm \left( \sqrt{\frac{S_X^2}{n_x} + \frac{S_Y^2}{n_y}} \right) t_{(m, \frac{\alpha}{2})}$$

Where the degrees of freedom,  $m$  is calculated from:

Eq. 4:34

$$m = \frac{\left( \frac{S_X^2}{n_x} + \frac{S_Y^2}{n_y} \right)^2}{\frac{\left( \frac{S_X^2}{n_x} \right)^2}{n_x - 1} + \frac{\left( \frac{S_Y^2}{n_y} \right)^2}{n_y - 1}}$$

Rounded down to the next integer. Obviously, this is a messy calculation but the Excel T.TEST function includes an option for the two-sample unequal variances case (see the Excel help for the T.TEST function for a description).

For the difference in two sample proportions,  $\hat{P}_X$  and  $\hat{P}_Y$ , assuming the sample sizes are “large” the  $(1 - \alpha)\%$  confidence interval for the difference is:

Eq. 4:35

$$\hat{P}_X - \hat{P}_Y \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{P}_X(1 - \hat{P}_X)}{n_X} + \frac{\hat{P}_Y(1 - \hat{P}_Y)}{n_Y}}$$

Summary of Confidence Interval formulas:

Statistic	Confidence Interval
Sample Mean, Known Variance	$\bar{X} \pm \frac{\sigma}{\sqrt{n}} Z_{\frac{\alpha}{2}}$
Sample Mean, Unknown Variance	$\bar{X} \pm \frac{S}{\sqrt{n}} t_{(n-1), \alpha/2}$
Sample Proportion	$\hat{P} \pm Z_{\alpha/2} \sqrt{\frac{\hat{P}(1 - \hat{P})}{n}}$
Sample Variance	$\frac{(n-1)S^2}{\chi_{(n-1), Upper Tail}^2} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi_{(n-1), Lower Tail}^2}$
Two Sample Means, Dependent Samples	$\bar{d} \pm \frac{S_d}{\sqrt{n}} t_{(n-1), \frac{\alpha}{2}}$
Two Sample Means, Known Variances	$(\bar{X} - \bar{Y}) \pm \left( \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}} \right) Z_{\frac{\alpha}{2}}$
Two Sample Means, Unknown Variances assumed Equal	$\bar{X} - \bar{Y} \pm \left( \sqrt{\frac{S_p^2}{n_X} + \frac{S_p^2}{n_Y}} \right) t_{(n_X+n_Y-2), \frac{\alpha}{2}}$ <p>Where the pooled variance is calculated as</p> $S_p^2 = \frac{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}{n_X + n_Y - 2}$

<p>Two Sample Means, Unknown Variances not assumed Equal</p>	$\bar{X} - \bar{Y} \pm \left( \sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}} \right) t_{(m, \frac{\alpha}{2})}$ <p>Where the degrees of freedom m is from</p> $m = \frac{\left( \frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y} \right)^2}{\frac{\left( \frac{S_X^2}{n_X} \right)^2}{n_X - 1} + \frac{\left( \frac{S_Y^2}{n_Y} \right)^2}{n_Y - 1}}$
<p>Sample proportion when nx and ny are "large"</p>	$\hat{P}_X - \hat{P}_Y \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{P}_X(1 - \hat{P}_X)}{n_X} + \frac{\hat{P}_Y(1 - \hat{P}_Y)}{n_Y}}$

**E. Hypothesis Testing: Two Sample Tests**

For two sample means, **dependent sample with unknown variance**. Given the difference variable  $d_i = X_i - Y_i$ , the sample mean and standard deviation of the difference variable,  $\bar{d}$  and  $S_d$ , and sample size n,

Eq. 4:36

$$H_0: \mu_X - \mu_Y = 0$$

$$t_{calc} = \frac{\bar{d}}{S_d/\sqrt{n}}$$

Then compare  $t_{calc}$  to  $t_{(n-1, \alpha)}$  or  $t_{(n-1, \alpha/2)}$  depending on whether it is a one-sided or two-sided alternative.

For two sample means, **independent samples with known variances**

Eq. 4:37

$$H_0: \mu_X - \mu_Y = 0$$

$$Z_{calc} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}}$$

Then compare to  $Z_\alpha$  or  $Z_{\alpha/2}$

For two sample means, **independent samples with unknown variances assumed to be equal**

Eq. 4:38

$$H_0: \mu_X - \mu_Y = 0$$

$$Pooled\ Variance: S_p^2 = \frac{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}{(n_X + n_Y - 2)}$$

$$t_{calc} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_P^2}{n_X} + \frac{S_P^2}{n_Y}}}$$

Then compare  $t_{calc}$  to  $t_{(n_X+n_Y-2, \alpha)}$  or  $t_{(n_X+n_Y-2, \frac{\alpha}{2})}$ .

For two sample means, **independent samples with unknown variances not assumed to be equal**

$$H_0: \mu_X - \mu_Y = 0$$

$$t_{calc} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}}$$

Then compare  $t_{calc}$  to the critical t with degrees of freedom set to:

$$m = \frac{\left(\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}\right)^2}{\frac{\left(\frac{S_X^2}{n_X}\right)^2}{n_X - 1} + \frac{\left(\frac{S_Y^2}{n_Y}\right)^2}{n_Y - 1}}$$

Rounded down to the next integer.

For two sample proportions from **“large” independent samples**, since under the null hypothesis the two proportions are equal, we calculate a pooled proportion from the two samples for the variance estimate in the denominator of the test statistic.

Eq. 4:39

$$H_0: P_X = P_Y$$

$$\text{Pooled sample Proportion: } P_0 = \frac{n_X \hat{P}_X + n_Y \hat{P}_Y}{n_X + n_Y}$$

$$Z_{calc} = \frac{\hat{P}_X - \hat{P}_Y}{\sqrt{\frac{P_0(1-P_0)}{n_X} + \frac{P_0(1-P_0)}{n_Y}}}$$

Then compare to  $Z_\alpha$  or  $Z_{\alpha/2}$

For a test of the equality of the variances from two samples, we need to introduce another probability distribution related to the Normal and Chi-Squared distributions. The F Distribution has a “numerator” and “denominator” degrees of freedom. For a given level of significance,  $\alpha$ , in the attached table of “Upper Critical Values of the F Distribution,” the numerator degrees of freedom is read across columns and the denominator degrees of freedom is read down rows.

Eq. 4:40

$$H_0: \sigma_X^2 = \sigma_Y^2 \text{ vs. } H_A: \sigma_X^2 > \sigma_Y^2$$



$$F_{calc} = \frac{S_X^2}{S_Y^2}$$

Where  $S_X^2$  is the larger of the two sample variances so that  $F_{calc} \geq 1$ . Compare to  $F_{(n_X-1, n_Y-1, \alpha)}$ .

For example, suppose you are given a sample for the variable X with  $n_X = 10$  and  $S_X^2 = 12.3$ , and for the variable Y,  $n_Y = 12$  and  $S_Y^2 = 5.6$ , then

Eq. 4:41

$$F_{calc} = \frac{12.3}{5.6} = 2.196$$

From the F table with  $\alpha=0.05$ ,  $F_{(9,11,0.05)} = 2.896 > F_{calc}$  so we would Fail to Reject the null hypothesis that the two variances are equal.

## F. P-Values

The preceding discussion of hypothesis testing compared calculated test statistics to “critical” values derived from the sampling distribution of the hypothesis in question (normal, t-distribution, Chi-Square distribution, or F distribution.) An equivalent decision rule involves calculating the probability in the “tail(s)” of the distribution for a given calculated test statistic – the **p-value** for the test statistic – and comparing the p-value to the chosen level of significance,  $\alpha$ . The decision rule is to Reject the Null Hypothesis if the p-value is **less than**  $\alpha$ , and Fail to Reject the Null Hypothesis if the p-value is **greater than**  $\alpha$ . In this sense, the p-value is the smallest significance level at which the null hypothesis can be rejected.

To illustrate, consider the example of testing a hypothesis regarding the mean of a normal random variable with known variance. If the alternative hypothesis is an upper (right-tailed) alternative, for the calculated Z-Score,  $Z_c$ , the p-value is the probability  $P[Z > Z_c]$  and can be calculated using “=1-NORM.S.DIST( $Z_c$ ,1)” in Excel. If the alternative hypothesis is a lower (left-tailed) alternative, we need  $P[Z < Z_c]$  and use “=NORM.S.DIST( $Z_c$ ,1)” in Excel.

For a two-sided (or two-tailed) alternative hypothesis, the p-value is the probability given by  $2 \times P[Z > |Z_c|]$  and can be calculated using “=1-NORM.S.DIST(ABS( $Z_c$ ),1)” in Excel.

### G. Example Problems

Scores on an aptitude test are known to follow a normal distribution with a standard deviation of 32.4 points. A random sample of 12 test scores had a mean score of 189.7 points.

1. Find an 80% confidence interval for the population mean for this sample.

$$LCL \Rightarrow LCL = \bar{X} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \cong 189.7 - (1.2816) \left( \frac{32.4}{\sqrt{12}} \right) \cong 177.7136$$

$$UCL \Rightarrow UCL = \bar{X} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \cong 189.7 + (1.2816) \left( \frac{32.4}{\sqrt{12}} \right) \cong 201.6864$$

2. Based on the sample results, a confidence interval for the population mean is found extending from 171.4 to 208 points. Find the confidence level of this interval.

$$\text{Margin of Error (ME)} \Rightarrow ME = UCL - \bar{X} = 208 - 189.7 = 18.3$$

(or  $ME = \bar{X} - LCL$ )

$$\text{Z-Score } (Z_{\alpha/2}) \Rightarrow \text{Solve } ME = Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \quad 18.3 = Z_{\frac{\alpha}{2}} \frac{32.4}{\sqrt{12}} \rightarrow Z_{\frac{\alpha}{2}} \cong 1.96$$

$$\text{Confidence Level} \Rightarrow P \left[ Z > Z_{\frac{\alpha}{2}} = 1.96 \right] = 0.025 \text{ so Confidence Level} = 0.95$$

**Note:** The information and questions 14 and 15 could also be applied to the case of an unknown variance (e.g. the sample standard deviation is 32.4). In the foregoing answers, you would replace  $Z_{\alpha/2}$  with  $t_{\alpha/2, (n-1)}$ .

.....

A sample of 24 months of stock return data for “Company X” is obtained and the sample variance of returns is calculated to be  $S^2 \cong 0.0035$ . Assume the data are drawn from a normal distribution with unknown variance.

3. Calculate the LC and UCL for a 90% confidence interval of the population variance of monthly returns.

$$LCL \Rightarrow LCL = \frac{(n-1)S^2}{\chi^2_{(n-1, 1-\frac{\alpha}{2})}} \cong \frac{(23)(0.0035)}{35.172} \cong 0.00229$$

$$UCL \Rightarrow UCL = \frac{(n-1)S^2}{\chi^2_{(n-1, \frac{\alpha}{2})}} \cong \frac{(23)(0.0035)}{13.091} \cong 0.00615$$

.....

A sample of 25 students is administered an aptitude test before and after they completed a test prep course. The “After” minus “Before” scores of the students yielded the following test statistics:  $\bar{d} = 24.72$ ,  $S_d^2 = 4317.21$

4. Calculate the LCL and UCL for a 90% CI ( $\alpha = 0.1$ ) for the mean difference in scores.

$$\text{Need } \bar{d} \pm t_{(n-1, \frac{\alpha}{2})} \sqrt{\frac{S_d^2}{n}} = 24.72 \pm (1.7109) \times \sqrt{\frac{4317.21}{25}}$$

$$LCL \Rightarrow LCL \cong 2.2371$$

$$UCL \Rightarrow UCL \cong 47.2029$$

.....

The results from independent random sampling from two normally distributed populations is provided in the following table

Variable	Sample Size n:	Sample Mean	Known Variance
X	81	140	25
Y	100	120	14

5. Find a 95% confidence interval for the difference between the means in these two populations (the mean of X minus the mean of Y):

$$\text{Need } (\bar{X} - \bar{Y}) \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}} = 20 \pm (1.96) \times \sqrt{\frac{25}{81} + \frac{14}{100}}$$

LCL ==> LCL  $\cong$  18.6872

UCL ==> UCL  $\cong$  21.3128

.....  
 Data from two independent samples of the annual returns for a sample of 12 Technology firms and a sample of 14 Finance firms are collected. The population variance of annual returns is unknown but assumed to be equal across the two samples and the following sample statistics are calculated.

	Mean	Variance	Sample Size
Technology Firms	0.1141	0.00022	12
Finance Firms	0.1034	0.00021	14

Calculate the LCL and UCL for a 90% CI ( $\alpha = 0.1$ ) for the difference in mean returns between the two samples.

Need the pooled variance estimate from  $S_p^2 = \frac{(n_x-1)S_x^2 + (n_y-1)S_y^2}{(n_x+n_y-2)} \cong 0.00022$

Need  $(\bar{X} - \bar{Y}) \pm t_{(n_x+n_y-2, \frac{\alpha}{2})} \sqrt{\frac{S_p^2}{n_x} + \frac{S_p^2}{n_y}} \cong 0.0107 \pm (1.7109) \times \sqrt{\frac{0.00022}{12} + \frac{0.00022}{14}}$

LCL ==> LCL  $\cong$  0.0007

UCL ==> UCL  $\cong$  0.0206

.....  
 In a random sample of 120 large retailers, 85 used regression analysis as a method of forecasting. In an independent random sample of 163 small retailers, 78 used regression as a method of forecasting. Calculate the LCL and UCL for a 98% confidence interval of the difference between the two population proportions.

Based on the information provided,  $\hat{P}_X = \frac{85}{120} = 0.708$  and  $\hat{P}_Y = \frac{78}{163} = 0.479$

Need  $(\hat{P}_X - \hat{P}_Y) \pm Z_{\frac{\alpha}{2}} \times \sqrt{\frac{\hat{P}_X(1-\hat{P}_X)}{n_x} + \frac{\hat{P}_Y(1-\hat{P}_Y)}{n_y}} \cong 0.2298 \pm (2.326)(0.057)$

LCL ==> LCL  $\cong$  0.097

UCL ==> UCL  $\cong$  0.362

.....  
 A random sample of size n=25 is obtained from a normal population with variance  $\sigma^2 = 625$ , and the sample mean is computed. Test the null hypothesis  $H_0: \mu = 100$

versus the alternative hypothesis  $H_1: \mu > 100$  with  $\alpha = 0.05$ . Compute the critical value  $\bar{X}_c$  and state your decision rule for this hypothesis:

Given that this is a right-tailed one-sided test and  $\alpha = 0.05$ , the critical  $Z_\alpha = 1.645$  (this is calculated in Excel using the `NORM.S.INV(0.95)` or by interpolating from the Standard Normal Probability table.) The critical value  $\bar{X}_c$  is found by comparing  $Z_\alpha$  to the calculated Z-Score  $Z_c$ :

$$Z_c = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}$$

The standard error of the sample mean is calculated as:  $\sigma_{\bar{X}} = \sqrt{\frac{\sigma^2}{25}} = \sqrt{\frac{625}{25}} = 5$ , then

$$Z_c = \frac{\bar{X} - 100}{5}$$

Setting this equal to  $Z_\alpha$  and solving for  $\bar{X}$  gives  $\bar{X}_c = 5 \times 1.645 + 100 = 108.224$  and the decision rule is "Reject  $H_0$  if  $\bar{X} > 108.224$ "

Note: If the alternative hypothesis had been  $H_1: \mu < 100$  with  $\alpha = 0.05$ , we would have used  $Z_\alpha = -1.645$  (because this is a left-tailed alternative) and the resulting decision rule would be "Reject  $H_0$  if  $\bar{X} < 91.776$ "

.....  
A random sample is obtained from a normal population with variance  $\sigma^2 = 400$ , and the sample mean is computed to be 76.5. Consider the null hypothesis  $H_0: \mu = 80$  versus the alternative hypothesis  $H_1: \mu < 80$ . For a sample size  $n=40$ , compute the p-value and state your decision if  $\alpha = 0.1$ .

Since the alternative hypothesis is a lower (left-tailed) alternative, to get the p-value we need  $P[Z < Z_c]$ . The calculated Z-Score is  $Z_c = \frac{76.5-80}{\sqrt{400/40}} \cong -1.1068$

Using `NORM.S.DIST(Zc,1)` gives  $P[Z < Z_c] \cong 0.1341$ . Since the p-value is greater than  $\alpha = 0.1$ , we Fail to Reject the Null Hypothesis.

Note that is the sample size is  $n=70$ ,  $Z_c \cong -1.4642$  resulting in a p-value of 0.0716, so in this case we would Reject the Null Hypothesis.

.....  
You are provided a sample of  $n=23$  test scores. The sample mean is calculated to be  $\bar{X} = 81.5$  and the sample variance is calculated to be  $S^2 = 14.7$ . Assuming the sample is drawn from a Normal population with unknown variance, provide a test of the null hypothesis  $H_0: \mu = 80$  against each of the following alternative hypotheses and significance levels:

a.  $H_1: \mu \neq 80$  with  $\alpha = 0.05$

The calculated t-statistic is given by:  $t_{calc} = \frac{\bar{X}-\mu}{\sqrt{S^2/n}} = \frac{81.5-80}{\sqrt{14.7/23}} \cong 1.876$

With degrees of freedom equal to 22, the critical t for this two-tailed test is  $t_{(n-1, \alpha/2)} = 2.074$  which is greater than  $t_{calc}$  so we Fail to Reject the Null Hypothesis.

**b.**  $H_1: \mu > 80$  with  $\alpha = 0.05$

The calculated t-statistic remains the same, but because this is now a one-tailed test, the critical t is  $t_{(n-1, \alpha)} = 1.717$  which is less than  $t_{calc}$  so we Reject the Null Hypothesis.

Alternatively, the probability  $P[t > t_{calc}]$  can be calculated using “1-T.DIST( $t_{calc}, 22$ )” giving a one-tailed p-value of about 0.037 which implies a two-tailed p-value of about 0.074 ( $2 \times 0.037$ ). For part (a), since the two-tailed p-value is greater than  $\alpha$ , you Fail to Reject the null, and for part (b), you Reject the null because the p-value is less than  $\alpha$ .

More examples for this section will be provided through homework sets and the mid-term exam study guide.

## Section 5. ANALYSIS OF VARIANCE

---

Our discussion of hypothesis testing has thus far been limited to comparing sample statistics between at most two samples. There are many applications in business and economics that posit potential differences across many more than two groups identifiable in a population. Consider the following example: You have a sample of test scores from students at 15 different schools and you would like to know if average test scores differ across schools. One possibility would be to do pair-wise tests for differences in the mean scores for each possible pair of schools. However, with 15 schools, this would involve 105 difference pair-wise comparisons.

The statistical tool known as Analysis of Variance, or ANOVA, provides a basis for testing for differences in the population means among any number of groups. As with the two variable tests we have covered previously, the nature of these ANOVA tests will depend on the properties of the sample data. Conceptually, the basis of an ANOVA is a comparison of the degree of variability (the variance) in the overall sample assuming some null hypothesis is true (that is, assuming all the groups in the sample have the same overall mean) versus the degree of variability when each group in the sample is allowed to take on its own mean.

### A. One-Way Analysis of Variance

Suppose that you want to compare the means of  $K$  populations, *each of which is assumed to have the same variance*. Independent random samples of  $n_1, n_2, \dots, n_K$  observations are taken from these populations. Let  $x_{ij}$  denote the  $j^{\text{th}}$  observation from the  $i^{\text{th}}$  population. Formally, we wish to test the following hypothesis:

Eq. 5:1

$$H_0: \mu_1 = \mu_2 = \dots = \mu_K$$

$$H_A: \mu_i \neq \mu_j \text{ for at least one pair, } \mu_i, \mu_j$$

The sample means for the  $K$  groups can be expressed as follows:

Eq. 5:2

$$\bar{x}_i = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i} \quad (i = 1, 2, \dots, K)$$

Let  $n$  denote the total number of observations:

Eq. 5:3

$$n = \sum_{i=1}^K n_i$$

The overall mean across *all*  $K$  groups, denoted  $\bar{\bar{x}}$ , can be expressed as:

Eq. 5:4

$$\bar{\bar{x}} = \frac{\sum_{i=1}^K \sum_{j=1}^{n_i} x_{ij}}{n} = \frac{\sum_{i=1}^K n_i \bar{x}_i}{n}$$

The degree of variability for each group – *when each group is allowed to take on its own mean* – is measured by its **Sum of Squares**

Eq. 5:5

$$SS_i = \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

The total within-group variability, or *within group sum of squares*, denoted SSW, is then given by:

Eq. 5:6

$$SSW = \sum_{i=1}^K SS_i = \sum_{i=1}^K \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

SSW, then, is the total observed variability in the sample *when each group is allowed to take on its own mean*. Next, we need a measure of the degree of variability in the sample based on the null hypothesis that all the means are equal. This measure is known as the *between group sum of squares*, denoted SSG and calculated by:

Eq. 5:7

$$SSG = \sum_{i=1}^K n_i (\bar{x}_i - \bar{\bar{x}})^2$$

If the null hypothesis is, in fact, true, SSG should be close to zero. Another sum of squares that is often calculated and presented as part the ANOVA results is known as the *total sum of squares*, denoted SST as calculated by:

Eq. 5:8

$$SST = \sum_{i=1}^K \sum_{j=1}^{n_i} (x_{ij} - \bar{\bar{x}})^2$$

And it can be shown that:

Eq. 5:9

$$SST = SSW + SSG$$

To formulate a test statistic, we need to know the distributional properties of our sum of squares measures. Under the assumptions that the population variances are equal and the population distributions are normal, SSW is distributed as Chi-Square with  $(n - K)$  degrees of freedom and SSG is Chi-Square with  $(K - 1)$  degrees of freedom. We can use these two measures to calculate an F-Statistic in the following form

Eq. 5:10

$$F_{calc} = \frac{SSG/(K-1)}{SSW/(n-K)} \sim F_{(K-1),(n-K)}$$

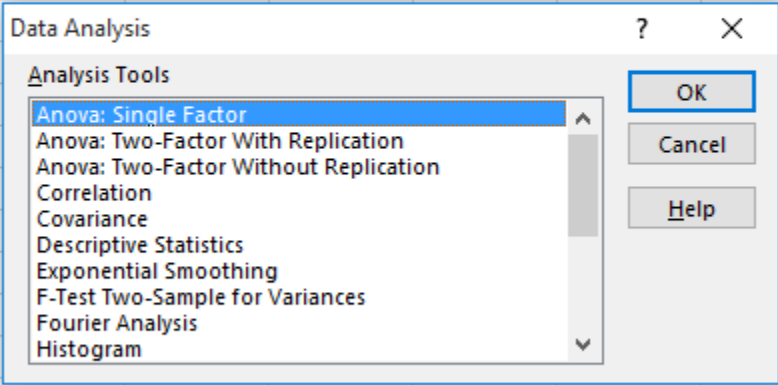
That is,  $F_{calc}$  is distributed as an F-statistic with numerator degrees of freedom equal to  $(K-1)$  and denominator degrees of freedom equal to  $(n-K)$ . For a chosen level of significance  $\alpha$ , we compare our calculated test statistic,  $F_{calc}$  to the *critical*  $F_{(K-1),(n-K),\alpha}$  and reject the null hypothesis that the means are equal if  $F_{calc} > F_{(K-1),(n-K),\alpha}$ .

Suppose, for example, an instructor administers the same exam to 5 different sections of students with the following results:

Section	Students	Average
S1	21	74.7
S2	20	73.9
S3	12	71.8
S4	22	79.4
S5	23	73.8
Overall	98	75.0

While it is clear that section S4 scored the highest on average and section S3 the lowest, the instructor wants to test whether the exam means are significantly different across sections. With each section's scores arranged in consecutive columns, in Excel, the instructor uses the Data Analysis Add-In for a "ANOVA: Single Factor" which looks like the following:

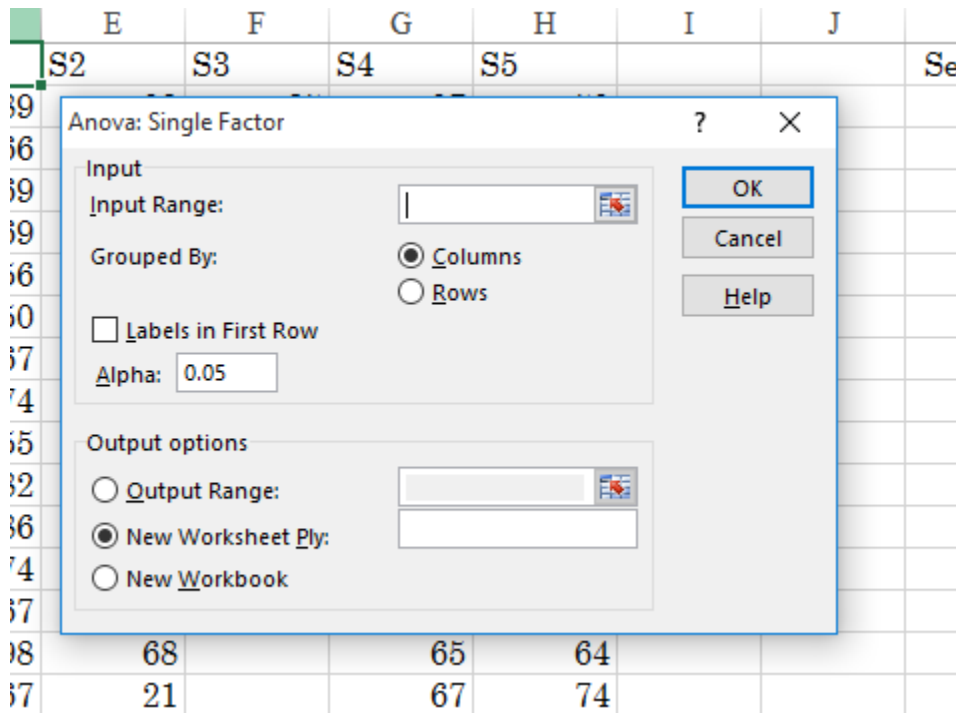
D	E	F	G	H	I	J	K
S1	S2	S3	S4	S5			Section
89	92	85	97	59			S1
66	97	72	98	46			S2
69							S3
69							S4
56							S5
50							
67							
74							
55							
82							
86							
74	27	59	90	94			
67	69		78	86			
66	66		66	66			



The screenshot shows the 'Data Analysis' dialog box in Excel. The 'Analysis Tools' list includes: Anova: Single Factor (selected), Anova: Two-Factor With Replication, Anova: Two-Factor Without Replication, Correlation, Covariance, Descriptive Statistics, Exponential Smoothing, F-Test Two-Sample for Variances, Fourier Analysis, and Histogram. The dialog box has 'OK', 'Cancel', and 'Help' buttons.



Choosing “OK” then brings up the following dialog box:



Where you can choose the data range (“Input Range”), indicate whether the data is arranged in rows or columns, indicate whether the first row or column contains variable names (“Labels in First Row” or “Labels in First Column”), choose the significance level  $\alpha$  (“Alpha”) for the test, and where you want the output. For our example, this will generate the following output

Anova: Single Factor

SUMMARY

<i>Groups</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>
S1	21	1568	74.67	200.53
S2	20	1478	73.90	503.36
S3	12	862	71.83	301.24
S4	22	1747	79.41	325.78
S5	23	1698	73.83	252.97

ANOVA

<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
Between Groups	606.152	4	151.538	0.4811	0.7495	2.4696
Within Groups	29294.756	93	314.997			
Total	29900.908	97				

In this example, there are 5 groups ( $K=5$ ) and a total of 98 observations ( $n=98$ ), from Eq. 5:10 above, the value of “ $F$ ” in the ANOVA table is then calculated from:

Eq. 5:11

$$F_{calc} = \frac{SSG/(K-1)}{SSW/(n-K)} = \frac{606.152/4}{29294.756/93} = 0.4811$$

The “ $F$  crit” in the ANOVA table is the value of  $F_{\alpha,(K-1),(n-K)}$  with  $\alpha=0.05$ . Since  $F_{calc} < F_{crit}$  we Fail to Reject the null hypothesis that the group (section) means are equal to one another. To find “ $F$  crit” in Excel for any specific level of significance  $\alpha$ , numerator degrees of freedom  $DF_1$ , and denominator degrees of freedom  $DF_2$ , you would use “=F.INV.RT( $\alpha,DF_1,DF_2$ )”

The “ $P$ -value” in the ANOVA table, is the probability in the right-hand tail of the  $F$  distribution starting at the calculated  $F$  statistic  $F_{calc}$ . An equivalent way to evaluate the null hypothesis is to Reject the null if the  $P$ -value  $< \alpha$ , and Fail to Reject the null otherwise. In our example, since the  $0.7495 > 0.05$  we Fail to Reject the null. To find a  $P$ -value in Excel for a particular calculated  $F_{calc}$  with numerator degrees of freedom  $DF_1$ , and denominator degrees of freedom  $DF_2$ , you would use “=F.DIST.RT( $F_{calc},DF_1,DF_2$ )”

For exam purposes on this type of problem, I would give you Summary and ANOVA tables similar to those above, **but leave the values for “ $F$ ,” “ $P$ -value,” and “ $F$  crit” blank**. It would be up to you to take the values out of the ANOVA table to calculate the appropriate  $F_{calc}$  and find the appropriate “ $F$  crit” from the probability tables.

## B. Two-Way Analysis of Variance

In a One-Way ANOVA, we are testing for differences in the population means across a single dimension – across the different Groups in the sample. In a Two-Way ANOVA, we allow for the possibility that the population means may differ in at least two ways – across the different Groups, and across different Blocks. Suppose for example, that you have a set of 5 automobiles (the Groups) that you want to test for differences in fuel consumption as measured by miles-per-gallon (MPG), among a set of 7 specific drivers (the Blocks). In the language of ANOVA, each unique combination of a Group type and Block type is referred to as a “cell” in the experimental design. The number of ways that we can assess whether there are differences in the population means within this design depends on whether we have a single observation or multiple observations for each Group (car) Block (driver) combination – that is, single or multiple observations per cell.

If there is a single observation per cell, this is called a **Two-Way ANOVA Without Replication**. With a single observation per cell in our car/driver example, the data may look like the following:

	A	B	C	D	E	F
1		Car-1	Car-2	Car-3	Car-4	Car-5
2	Driver-1	18.3	14.95	13.65	16.42	18.73
3	Driver-2	21.15	16.07	17.31	18.17	20.98
4	Driver-3	15.97	14.44	13.55	18.3	21.6
5	Driver-4	19.97	18.56	18.59	21.43	18.72
6	Driver-5	17.29	13.27	19.15	20.45	17.51
7	Driver-6	16.66	18.92	16.24	18.3	19.07
8	Driver-7	19.14	20.42	13.52	16.55	16.71
9						

As before, let K denote the number of Groups (cars, in our example, so K=5) and let H denote the number of Blocks (drivers, in our example, so H=7). Then the total number of observations is n=KH and we can define three different Sum of Squares measures and associated degrees of freedom (DF):

Eq. 5:12

$$\begin{aligned}
 \text{between Groups: } SSG &= H \sum_{i=1}^K (\bar{x}_i - \bar{\bar{x}})^2 \quad \text{with } DF_G = K - 1 \\
 \text{between Blocks: } SSB &= K \sum_{j=1}^H (\bar{x}_j - \bar{\bar{x}})^2 \quad \text{with } DF_B = H - 1 \\
 \text{error: } SSE &= \sum_{i=1}^K \sum_{j=1}^H (x_{ij} - \bar{x}_i - \bar{x}_j + \bar{\bar{x}})^2 \quad \text{with } DF_E = (K - 1)(H - 1)
 \end{aligned}$$

Where  $\bar{x}_i$  is the sample mean for each Group type (cars) across Blocks (drivers, this is also called the “column means”),  $\bar{x}_j$  is the sample mean for each Block type across Groups (called the “row means”) and as before,  $\bar{\bar{x}}$  is the overall mean.

To test the null hypothesis that the population means are the same across Groups (cars) we would compare SSG to SSE with the following F-statistic:

Eq. 5:13

$$F_{calc} = \frac{SSG / (K - 1)}{SSE / (K - 1)(H - 1)} \quad \text{vs. } F_{\alpha, (K-1), (K-1)(H-1)}$$

And to test the null hypothesis that the population means are the same across Blocks (drivers) we would compare SSB to SSE with the following F-statistic:

Eq. 5:14

$$F_{calc} = \frac{SSB / (H - 1)}{SSE / (K - 1)(H - 1)} \quad \text{vs. } F_{\alpha, (H-1), (K-1)(H-1)}$$

In Excel, we would use the “ANOVA: Two Factor without Replication” option in the Data Analysis ToolPack and in our car/driver example the output would look like the following:

	A	B	C	D	E	F	G
1	Anova: Two-Factor Without Replication						
2							
3	<i>SUMMARY</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>		
4	Driver-1	5	82.05	16.410	4.676		
5	Driver-2	5	93.68	18.736	5.081		
6	Driver-3	5	83.86	16.772	10.527		
7	Driver-4	5	97.27	19.454	1.564		
8	Driver-5	5	87.67	17.534	7.339		
9	Driver-6	5	89.19	17.838	1.711		
10	Driver-7	5	86.34	17.268	7.078		
11							
12	Car-1	7	128.48	18.354	3.456		
13	Car-2	7	116.63	16.661	7.091		
14	Car-3	7	112.01	16.001	6.013		
15	Car-4	7	129.62	18.517	3.451		
16	Car-5	7	133.32	19.046	3.048		
17							
18							
19	ANOVA						
20	<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
21	Rows (Blocks SSB)	34.53264	6	5.75544	1.330491	0.282282	2.508189
22	Columns (Groups SSG)	48.08478	4	12.0212	2.778953	0.049844	2.776289
23	Error (SSE)	103.8192	24	4.325801			
24							
25	Total	186.4366	34				
26							

Given these results, we would Reject the null that population means are equal across Cars (Groups) at significance level  $\alpha=0.05$  because  $F_{calc} = 2.7789 > 2.7762 = F_{crit}$  (or equivalently because the p-value =  $0.0498 < 0.05 = \alpha$ ) and we would Fail to Reject the null that the population means are equal across Drivers (Blocks) because  $F_{calc} = 1.330 < 2.508 = F_{crit}$  (equivalently because the p-value  $> \alpha$ ).

If we have multiple observations for each Block (in our car/driver example, multiple observations of a particular driver in a particular car), this is called a **Two Way ANOVA With Replication**. In Excel, an example of how your may be arranged is as follows:

	Car-1	Car-2	Car-3	Car-4	Car-5
Driver-1	15.22	15.65	13.38	20.2	19.16
	14.44	18.36	18.77	14.76	18.08
	18.82	18.07	15.19	17.26	19.46
	19.23	15.93	22.45	15.53	20.24
	17.06	18.37	15.56	17.8	21.26
	17.95	17.39	15.21	14.53	18.21
	18.61	18.07	17.78	17.03	18.73
	18.3	18.29	17.73	17.8	20.79
	17.84	16.4	14.1	16.36	18.4
	16.94	17.93	18.54	19.43	16.73
Driver-2	18.97	15.22	19.24	15.05	19.53
	18.73	16.68	14.73	19.26	17.79
	17.75	16.55	21.24	21.97	15.87
	20.01	16.23	15.05	16.12	19.27
	20.78	18.13	15.09	15.6	20.26
	16.44	17.31	21.8	18.64	20.14
	16.97	18.68	19.43	18.93	20.34
	22.64	20.1	19.72	20.14	19.34
	20.8	19.28	21.88	21.69	17.45
	21.49	21.09	14.74	19.27	17.55
Driver-3	18.88	17.36	17.78	14.53	18.03
	18.36	18.51	18.41	17.58	21.36

In this example, we have 10 observations on each driver for each car. Importantly, in the Excel ToolPack for this type of ANOVA, you *must* have the same number of observations per “cell” (car/driver combination). The notation for how to calculate the specific sums of squares gets very messy, so I will present the summary names and their degrees of freedom. Let  $m$  be the number of observations per cell. Then we can define four different sum of squares measures:

Eq. 5:15

$$\begin{aligned}
 & \textit{between Groups: SSG with } DF_G = K - 1 \\
 & \textit{between Blocks: SSB with } DF_B = H - 1 \\
 & \textit{Interaction: SSI with } DF_I = (K - 1)(H - 1) \\
 & \textit{error: SSE with } DF_E = HK(m - 1)
 \end{aligned}$$

While the notion of there being differences in the population means across Groups (cars) or Blocks (drivers) is somewhat intuitive, the **Interaction** effect relates to the possibility that some combinations of the Group and Block effects may be different than others.

To test the null hypothesis that the population means are the same across Groups (cars) we would compare SSG to SSE with the following F-statistic:

Eq. 5:16

$$F_{calc} = \frac{SSG/(K-1)}{SSE/HK(m-1)} \text{ vs. } F_{\alpha,(K-1),HK(m-1)}$$

To test the null hypothesis that the population means are the same across Blocks (drivers) we would compare SSB to SSE with the following F-statistic:

Eq. 5:17

$$F_{calc} = \frac{SSB/(H-1)}{SSE/HK(m-1)} \text{ vs. } F_{\alpha,(H-1),HK(m-1)}$$

And to test whether or not there are Interaction effects, we would compare SSI to SSE with the following F-statistic:

Eq. 5:18

$$F_{calc} = \frac{SSI/(K-1)(H-1)}{SSE/HK(m-1)} \text{ vs. } F_{\alpha,(K-1)(H-1),HK(m-1)}$$

In Excel, we would use the “ANOVA: Two Factor without Replication” option in the Data Analysis ToolPack and in our car/driver example the ANOVA Table output would look like the following:

ANOVA							
Source of Variation	SS	df	MS	F	P-value	F crit	
Sample (Blocks SSB)	51.18255	6	8.530426	2.279631	0.036101	2.127402	
Columns (Groups SSG)	75.00415	4	18.75104	5.010939	0.000629	2.400311	
Interaction	131.9571	24	5.498213	1.469317	0.074958	1.551998	
Within (Error SSE)	1178.737	315	3.742021				
Total	1436.88	349					

With these results, the hypotheses and conclusions for  $\alpha=0.05$  are as follows:

Population means equal across Groups (cars): Reject because  $5.01 > 2.4$  (equivalently, because the P-Value is less than  $\alpha$ ).

Population means equal across Blocks (drivers): Reject because  $2.796 > 2.127$  (equivalently, because the P-Value is less than  $\alpha$ ).

Interaction effects are not present: Fail to Reject because  $1.469 < 1.552$  (equivalently, because the P-Value is greater than  $\alpha$ ).

Note that these conclusions are for a significance level of  $\alpha=0.05$ . If the significance level were set at a smaller value, say  $\alpha=0.01$ , we would still Reject the first hypothesis above because the P-value 0.0006 is less than  $\alpha$ . However, we would Fail

to Reject the second hypothesis because the P-value 0.036 is now greater than  $\alpha$ . Similarly, if we set  $\alpha=0.1$ , we would Reject the hypothesis that there are no Interaction effects because it's P-value of 0.075 is less than  $\alpha$ .

## Section 6. LINEAR REGRESSION

Up to this point, other than examining some degree of association via a correlation, our examination of the relationship between two variables has been limited to comparing their means. Linear Regression is a method of “explaining” the observed variation in some variable  $Y$  that is believed to be related in some way to the variable  $X$ .  $Y$  is denoted the “Dependent” variable;  $X$  is denoted the “Explanatory” variable. We begin by expressing  $Y$  as a function of  $X$ :  $Y = f(X)$

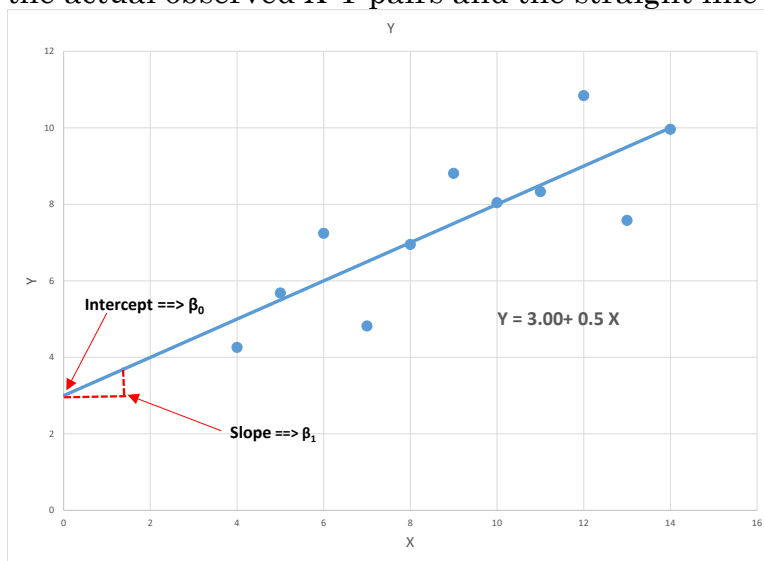
### A. The Bivariate Regression Model

When there is a single right-hand-side variable, the regression model is known as a bivariate model. Linear Regression assumes that the relationship between  $Y$  and  $X$  is linear in the population parameters  $\beta_0$  and  $\beta_1$  and a random error term  $\varepsilon$

Eq. 6:1

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

$\beta_0$  is the intercept;  $\beta_1$  is the slope;  $\varepsilon$  is the error term and is the difference between the actual observed  $X$ - $Y$  pairs and the straight line – the regression line.



Given this structure, we need an estimator (a rule) that uses the observed values of the variables  $Y$  and  $X$  to calculate sample estimates of the population parameters  $\beta_0$  and  $\beta_1$ . Recall that for a single random variable with a population mean  $\mu$ , our estimator for the population mean was the sample mean  $\bar{X}$ .

Letting  $Y_i$  and  $X_i$  represent individual observations out of a sample of size  $n$  of the variables  $Y$  and  $X$ , the assumptions of the Linear Regression model are as follows:

- The relationship between  $Y$  and  $X$  is linear in the population parameters  $\beta_0$  and  $\beta_1$  and a random error term  $\varepsilon$ :  $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
- The values of  $X$  are either fixed or a random variable that is independent of the error term  $\varepsilon_i$



- c. The random error term  $\varepsilon_i$  is distributed with a mean of zero and a constant variance  $\sigma^2$ :  $\varepsilon_i \sim (0, \sigma^2)$

The estimator that we will use to calculate sample estimates for the population parameters  $\beta_0$  and  $\beta_1$  and a random error term  $\varepsilon$  is the Least Squares Procedure which is based on finding the sample estimators  $b_0$  (for  $\beta_0$ ) and  $b_1$  (for  $\beta_1$ ) that minimize the Sum of Squared Errors – SSE – defined as follows:

*Eq. 6:2*

$$\begin{aligned} SSE &= \sum e_i^2 \\ &= \sum (Y_i - (b_0 + b_1 X_i))^2 \\ &= \sum (Y_i - \hat{Y}_i)^2 \end{aligned}$$

Taking derivatives of SSE with respect to  $b_0$  and  $b_1$ , setting them equal to zero and solving for  $b_0$  and  $b_1$  gives (after some manipulation) the following Least Squares estimators of  $\beta_0$  and  $\beta_1$

*Eq. 6:3*

$$b_1 = \frac{b_0 = \bar{Y} - b_1 \bar{X}}{\frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}}$$

The Least Squares estimator of the random error term is:

*Eq. 6:4*

$$e_i = Y_i - b_0 - b_1 X_i = Y_i - \hat{Y}_i.$$

Note that so long as the regression equation includes an intercept term,  $b_0$ , the Least Squares procedure will give the result:  $\sum e_i = 0$ . The Least Squares Estimator for the variance of  $\varepsilon$ ,  $\sigma^2$  is denoted  $S^2$  and given by:

*Eq. 6:5*

$$S^2 = \frac{SSE}{n - 2}$$

Given the assumptions of the Least Squares procedure in (4) above, the Least Squares Estimators  $b_0$  and  $b_1$  are Unbiased:

*Eq. 6:6*

$$E(b_0) = \beta_0$$

$$E(b_1) = \beta_1$$

And among the class of linear estimators, they have the minimum variance. Thus they are termed BLUE (Best Linear Unbiased Estimators).

As I noted above, the goal of Linear Regression is to “explain” the observed variation in the variable Y. A measure of the degree of variation in Y that is explained by the regression is known as the Coefficient of Determination or the regression  $R^2$ . The observed variation in Y can be broken down as follows:

Eq. 6:7

$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum e_i^2$$

$$SST = SSR + SSE$$

The regression  $R^2$  is then defined as:  $R^2 = 1 - \frac{SSE}{SST}$ . So long as the regression equation includes an intercept term, the  $R^2$  is bound between zero and one and gives the proportion of the variation in Y explained by the regression equation. Note that in a regression model with only one right-hand-side variable X, the regression  $R^2$  is simply the squared sample correlation between X and Y.

Expanding our assumptions concerning the random error term  $\varepsilon$  to include that it is Normally distributed, we can derive the distributional properties of the estimators  $b_0$  and  $b_1$ .

Eq. 6:8

$$b_0 \sim N\left(\beta_0, \left\{\frac{1}{n} + \frac{\bar{X}^2}{(n-1)\sum(X_i - \bar{X})^2}\right\}\sigma^2\right)$$

$$b_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum(X_i - \bar{X})^2}\right)$$

For the sample estimators of the variances of  $b_0$  and  $b_1$ , simply replace  $\sigma^2$  with  $S^2$  so that:

Eq. 6:9

$$S_{b_0}^2 = \left\{\frac{1}{n} + \frac{\bar{X}^2}{(n-1)\sum(X_i - \bar{X})^2}\right\}S^2$$

$$S_{b_1}^2 = \frac{S^2}{\sum(X_i - \bar{X})^2}$$

We can then test hypotheses concerning the intercept and slope parameters using the t-distribution. For example, to test the hypothesis that the slope parameter is equal to a constant k, we have the following:

Eq. 6:10

$$H_0: \beta_1 = k$$

$$t_{calc} = \frac{b_1 - k}{S_{b_1}}$$

We then compare  $t_{calc}$  to the value of the t distribution with (n-2) degrees of freedom and a significance value of  $\alpha/2$  for a two-tailed test or  $\alpha$  for a one tailed test.

While the various formulas given above may appear somewhat intimidating, in practice we rely on standard statistical software packages to do all the calculations. The results produced by the Excel regression Add-on package look like the following:

SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.8164
R Square	0.6665
Adjusted R Square	0.6295
Standard Error	1.2366
Observations	11

ANOVA					
	<i>Df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	27.510	27.510	17.990	0.002
Residual (SSE)	9	13.763	1.529		
Total	10	41.273			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	3.000	1.125	2.667	0.026	0.456	5.544
X1	0.500	0.118	4.241	0.002	0.233	0.767

The “R Square” in the table above is the  $R^2$  discussed in Eq. 6:9 above and for this example, indicates that the regression explains 66.65% of the observed variation in Y. From the ANOVA table, the regression SSE is 13.763 and  $S^2=1.529$ . The estimated value for  $b_0$  and  $b_1$  are the values in the “Coefficients” column for “Intercept” and “X1” respectively. The values for  $S_{b_0}$  and  $S_{b_1}$  are given in the “Standard Error” column. The values in the “t Stat” and “P-value” columns are to test the specific hypotheses that the intercept and slope parameters  $\beta_0$  and  $\beta_1$  are equal to zero against a two-tailed alternative. Looking at these values for X1, the t stat of 4.241 and P-value of 0.002 indicates that we would reject the null that  $\beta_1$  is equal to zero at a 95% or 99% confidence level and would have to set  $\alpha$  less than 0.2% before we would fail to reject.

The regression output above can also be used to test specific hypotheses regarding the size of  $\beta_0$  or  $\beta_1$  such as that in Eq. 6:10 above. For example, suppose we wish to test the hypothesis  $H_0: \beta_1 = 0.35$  versus the alternative hypothesis  $H_A: \beta_1 > 0.35$  at a

confidence level of 90% ( $\alpha=10\%$ ). Using the information for  $b_1$  (the Coefficient for  $X_1$ ) and  $S_{b_1}$  (the Standard Error of  $b_1$ ) from the regression results table and Eq. 6:10 we get:

Eq. 6:11

$$t_{calc} = \frac{0.5 - 0.35}{0.118} = 1.273$$

The degrees of freedom for this test is 9 (the number of observations  $n$  minus 2) which gives a (one tailed) critical  $t$  of  $t_{(9,0.1)} = 1.383 > t_{calc}$  so we would Fail to Reject the Null Hypothesis.

## B. Correlation Analysis

Sample covariance and correlation statistics as measures of co-movement between two random variables  $X$  and  $Y$  were introduced in Section 2.D., Eq. 2:17 and Eq. 2:15 Eq. 2:18 The sample correlation coefficient  $r_{XY}$  (the estimator for the population correlation coefficient  $\rho_{XY}$ ) is bound on the interval  $[-1, 1]$  and is calculated based on the sample covariance between  $X$  and  $Y$ ,  $S_{XY}$  and the sample standard deviations of  $X$  and  $Y$ ,  $S_X$  and  $S_Y$ , as follows:

Eq. 6:12

$$r_{XY} = \frac{S_{XY}}{S_X S_Y}$$

Negative values of  $r_{XY}$  imply an inverse relationship between  $X$  and  $Y$  with higher values of  $X$  being associated with lower values of  $Y$ . Positive values of  $r_{XY}$  imply an upward sloping relationship – higher  $X$ , higher  $Y$ . (Note that this is a measure of **association** and says little if anything about **causation**). Values close to -1 or 1 imply that observed  $X$ - $Y$  pairs fall almost exactly on a straight line with values closer to zero suggesting little if any statistical association between the variables.

The sample correlation coefficient between  $X$  and  $Y$  is directly related to the estimated regression slope coefficient  $b_1$  from the Bivariate Regression model discussed above, as follows:

Eq. 6:13

$$b_1 = \frac{S_Y}{S_X} r_{XY}$$

Therefore, a test of the hypothesis  $H_0: \beta_1 = 0$  is a direct test of the hypothesis  $H_0: \rho_{XY} = 0$ . However, we can also construct a test statistic based on sample correlation:

Eq. 6:14

$$t_{calc} = \frac{r_{XY}\sqrt{n-2}}{\sqrt{1-r_{XY}^2}}$$

Versus  $t_{(n-2,\alpha)}$  for a one tailed test and  $t_{(n-2,\frac{\alpha}{2})}$  for a two tailed test. For example, suppose you have a sample of  $n=25$  observations on the variables X and Y, and the calculated sample correlation coefficient is  $r_{XY} = 0.36$ . Then  $t_{calc}$  from Eq. 6:14 is:

Eq. 6:15

$$t_{calc} = \frac{0.36\sqrt{25-2}}{\sqrt{1-0.36^2}} = 1.851$$

For  $\alpha=5\%$ , since  $t_{(23,0.025)} = 2.064$ , we would fail to reject  $H_0: \rho_{XY} = 0$  against a two-tailed alternative hypothesis. However, since  $t_{(23,0.05)} = 1.711$ , we would reject  $H_0$  in favor of  $H_A: \rho_{XY} > 0$ .

The t-statistic presented in **Error! Bookmark not defined.**Eq. 6:14 is a straightforward and robust test of the null hypothesis  $H_0: \rho_{XY} = 0$ . However, for values other than zero, the sampling distribution of this test statistic becomes highly skewed. To deal with this issue, we can use what is known as **Fisher's z<sub>r</sub> transformation**, as follows:

Eq. 6:16

$$z_r = \frac{1}{2} \ln \left[ \frac{1+r_{XY}}{1-r_{XY}} \right]$$

Where  $\ln$  represents the natural logarithm. The sampling distribution of  $z_r$  is Normal with a variance of  $\frac{1}{n-3}$ , where  $n$  is the sample size. Thus an  $\alpha\%$  confidence interval for  $z_r$  can be computed as follows:

Eq. 6:17

$$z_r \pm Z_{\alpha/2} \sqrt{\frac{1}{n-3}}$$

And a test of the null hypothesis  $H_0: \rho_{XY} = \rho_0$  can be conducted using the calculated Z statistic:

Eq. 6:18

$$Z_{calc} = \frac{z_r - z_{\rho_0}}{\sqrt{\frac{1}{n-3}}}$$

Where  $z_r$  is the Fisher transformed value of  $r_{xy}$  and  $z_{\rho_0}$  is the Fisher transformed value of  $\rho_0$  from the null hypothesis. The value of  $Z_{calc}$  is compared to  $Z_{\alpha/2}$  for a two-tailed test and to  $Z_{\alpha}$  for a one tailed test.

For example, suppose you have a sample of  $n=45$  observations on the variables X and Y, and the calculated sample correlation coefficient is  $r_{XY} = 0.68$ . To test the null hypothesis  $H_0: \rho_{XY} = 0.5$ , use Eq. 6:16 and Eq. 6:18 to get:

Eq. 6:19

$$z_r = \frac{1}{2} \ln \left[ \frac{1 + 0.68}{1 - 0.68} \right] = 0.8921$$

$$z_{\rho_0} = \frac{1}{2} \ln \left[ \frac{1 + 0.5}{1 - .05} \right] = 0.5493$$

$$Z_{calc} = \frac{0.8921 - 0.5493}{\sqrt{\frac{1}{45 - 3}}} = 1.813$$

For  $\alpha=5\%$ , since  $Z_{0.025}=1.96$ , we would fail to reject the null versus a two tailed alternative. However, since  $Z_{0.05}=1.645$ , we would reject  $H_0$  in favor of  $H_A: \rho_{XY} > 0.5$ .

### C. Bivariate Regression with Multiple Groups

The Bivariate Regression model presented in Section 6.A, above, provides a straightforward method for testing whether there is a significant (that is, non-zero) correlation between two variables X and Y. As discuss above, the intercept term  $\beta_0$  is interpreted as the **mean value of Y after accounting for the effect of X**. The slope term  $\beta_1$  is interpreted as the **marginal effect on Y of a one unit increase in X – if X increases by one unit, Y changes by  $\beta_1$  units**. However, suppose there are multiple distinct subgroups identifiable in the data. An extension of the Bivariate Model can be specified to test for differences in the mean value of Y (after accounting for the effects of X) and the marginal effect of X on Y (that is, the correlation between X and Y) among the different subgroups in the data.

To allow for (potential) differences in the mean of Y and the degree of correlation between X and Y among subgroups, we make use of what are known as a **dummy** variables – a 0/1 (yes/no) variables equal to one if a particular X-Y pair is part of a specific subgroup, and zero otherwise. For example, suppose you have 3 distinct subgroups in a set of data. (This formulation can accommodate any number of subgroups provided that there are a sufficient number of observations – e.g. 10 or more – in each subgroup). Choose one of the groups as a “reference” group. While the interpretation of the resulting individual regression coefficients will change depending on which group is chosen as the reference group, the overall regression

results will not, so the choice of reference group is arbitrary. For this example of 3 subgroups, let group 3 be the reference group. Define  $D_1$  to be a dummy variable equal to one if an observation is part of group 1 and zero otherwise. Similarly for a dummy variable  $D_2$ . Next, create the variables  $X \times D_1$  and  $X \times D_2$ . In Excel, such data could look something like the following:

	A	B	C	D	E	F	G
1	Group	Y	X	D1	D2	X*D1	X*D2
2	1	91.59604	38.186	1	0	38.186	0
3	2	126.2713	36.419	0	1	0	36.419
4	3	131.698	27.741	0	0	0	0
5	1	141.7304	44.632	1	0	44.632	0
6	2	157.6186	34.383	0	1	0	34.383
7	3	173.1076	40.456	0	0	0	0
8	...						
9	More data lines						
10							
11							

Thus, when  $D_1=D_2=0$ , the observation relates to Group 3. Next, using the  $D_1$ ,  $D_2$ ,  $X \times D_1$ , and  $X \times D_2$ , variables, specify the regression equation:

Eq. 6:20

$$Y = \beta_0 + \beta_1 X + \alpha_1 D_1 + \alpha_2 D_2 + \gamma_1 (X \times D_1) + \gamma_2 (X \times D_2) + \varepsilon$$

This is the “Unrestricted” regression equation. It is unrestricted in the sense that both the intercept terms (the means of Y) and slope terms (the correlations between Y and X) are allowed to differ among all the subgroups. For purposes of testing whether the slope terms are equal across the subgroups, the key statistic that we need from this regression is the sum of squared errors,  $SSE_U$  and its degrees of freedom  $DF_U$ . Estimating the regression in Excel, these statistics are reported in the regression ANOVA table as the “Residual” df and SS as in the example below:

ANOVA

	df	SS	MS	F	Significance F
Regression	1	12010.4	12010.4	20.725	9.016E-06
Residual	208	120537.8	579.509		
Total	209	132548.2			

When  $D_1=D_2=0$  in Eq. 6:20 for Group 3, the regression equation reduces to the simple Bivariate Regression:

Eq. 6:21

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

So  $\beta_0$  measures the intercept for Group 3 and  $\beta_1$  measures the slope for Group 3. When  $D_1=1$  and  $D_2=0$ , for Group 1, the regression equation becomes:

Eq. 6:22

$$Y = (\beta_0 + \alpha_1) + (\beta_1 + \gamma_1)X + \varepsilon$$

So that the intercept and slope for Group 1 are  $\beta_0 + \alpha_1$  and  $\beta_1 + \gamma_1$ , respectively. Last, when  $D_1=0$  and  $D_2=1$ , the intercept and slope for Group 2 are,  $\beta_0 + \alpha_2$  and  $\beta_1 + \gamma_2$ , respectively.

To test the null hypothesis that the slope terms are equal across all the groups, the formal hypothesis becomes  $H_0: \gamma_1 = \gamma_2 = 0$ . Imposing the null hypothesis restriction on our regression equation in Eq. 6:20, results in a “Restricted” regression equation:

Eq. 6:23

$$Y = \beta_0 + \beta_1 X + \alpha_1 D_1 + \alpha_2 D_2 + \varepsilon$$

Estimating this restricted regression equation and taking the sum of squared errors and degrees of freedom as  $SSE_R$  and  $DF_R$ , we can calculate an F-statistic to test our null hypothesis:

Eq. 6:24

$$F_{calc} = \frac{(SSE_R - SSE_U)/(DF_R - DF_U)}{SSE_U/DF_U}$$

This  $F_{calc}$  is then compared to the critical value  $F(\alpha, DF_R - DF_U, DF_U)$ . Note that this framework is flexible enough to allow for the joint test of the slopes and/or intercept terms or any subset of either or both.

#### D. Alternative Tests for Differences in Sub-Sample Correlations

In Section 6.B, above, we introduced Fisher’s  $z_r$  transformation (Eq. 6:16). This transformation of the sample correlation coefficient can also be used to test for the equality of correlations across two or more subgroups. For two subgroups, suppose you have the sample correlation coefficient for the first group  $r_1$  based on  $n_1$  observations, and similarly for the second group,  $r_2$  based on  $n_2$  observations. First calculate the  $z_r$  transformations as:

Eq. 6:25

$$z_{r_1} = \frac{1}{2} \ln \left[ \frac{1 + r_1}{1 - r_1} \right] \quad z_{r_2} = \frac{1}{2} \ln \left[ \frac{1 + r_2}{1 - r_2} \right]$$

Then, to test the null hypothesis that the population correlation coefficients for these two subgroups are equal to one another:  $H_0: \rho_1 = \rho_2$ , construct the calculated Z-statistic:



Eq. 6:26

$$Z_{calc} = \frac{z_{r_1} - z_{r_2}}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

The  $Z_{calc}$  in Eq. 6:24 is then compared to  $Z_\alpha$  for a one-tailed alternative or  $Z_{\alpha/2}$  for a two-tailed alternative.

For example, suppose you have two independent subgroup samples with sample correlation coefficients, sample sizes, and  $z_r$  transformations as follows:

Group 1 Sample Correlation: $r_1$	0.582
Group 1 Sample Size: $n_1$	123
Group 1 $z_r$ transformation: $z_{r_1}$	0.6655
Group 2 Sample Correlation: $r_2$	0.413
Group 2 Sample Size: $n_2$	117
Group 2 $z_r$ transformation: $z_{r_2}$	0.4392

Then from Eq. 6:26, the calculated Z-statistic is:

Eq. 6:27

$$Z_{calc} = \frac{0.6655 - 0.4392}{\sqrt{\frac{1}{123 - 3} + \frac{1}{117 - 3}}} = 1.729$$

For  $\alpha=5\%$ , since  $Z_{\alpha/2}=1.96$ , we would fail to reject the null hypothesis of equal population correlations against a two-tailed alternative.

The  $z_r$  transformation can also be used to test for the joint equality among a group of sample correlation coefficients. Suppose you have  $k$  subgroups and wish to test the null hypothesis  $H_0: \rho_1 = \rho_2 = \dots = \rho_k$  against the alternative hypothesis that  $H_0$  is not true. Based on the  $z_r$  transformations of the sample correlation coefficients, and subgroup sample sizes, we can calculate the chi-square test statistic as follows:

Eq. 6:28

$$\chi^2_{calc} = \sum_{j=1}^k [(n_j - 3)z_{r_j}^2] - \frac{[\sum_{j=1}^k (n_j - 3)z_{r_j}]^2}{\sum_{j=1}^k (n_j - 3)}$$

We compare this to the upper chi-square value with  $k - 1$  degrees of freedom at the chosen  $\alpha$  level. While this equation looks a bit intimidating, it is fairly straightforward to calculate in Excel. Suppose you have four subgroups with sample correlation statistics, sample sizes and  $z_r$  transformations as follows:

Group	Sample Corr. Coeff.	Sample Sizes	$z_r$ trans.	$n_j-3$	$(n_j - 3)z_{r_j}^2$	$(n_j - 3)z_{r_j}$
1	0.245	73	0.2501	70	4.378	17.506
2	0.351	86	0.3666	83	11.154	30.426
3	0.277	67	0.2844	64	5.178	18.204
4	0.581	91	0.6640	88	38.795	58.429
Sums				305	59.505	124.565

The calculated chi-square statistic from Eq. 6:28 is then:

Eq. 6:29

$$\chi_{calc}^2 = 59.505 - \frac{(124.565)^2}{305} = 8.631$$

At  $\alpha=5\%$ , since the upper  $\chi^2$  critical value with 3 degrees of freedom is 7.815, we would reject the null hypothesis that the four groups have the same population correlation.

### E. ANOVA as Multiple Regression

The preceding discussion involving a regression with a single right-hand-side variable is known as a bivariate regression. When additional X variables are added to the equation, it is called a multiple regression. The formulas for the Least Squares parameters in a multiple regression can get very messy in summation form – in particular, the formulas change every time you add another right-hand-side variable – and we will not go into that level of detail. However, I do want to go into a special version of multiple regression for ANOVA.

Consider the Car-Driver two way ANOVA from the ANOVA lecture outline where there were 5 different cars and seven different drivers and the variable Y is the gas mileage for the car-driver combinations. Define the variable  $C_1$  to be an indicator variable equal to one if the observed Y is associated with car 1 and zero otherwise. Similarly define the variables  $C_2$ ,  $C_3$ , and  $C_4$ , for cars 2-4, and  $D_1, \dots, D_6$  for drivers 1-6. Note that we need only four “car” variables to categorize the 5 cars because if  $C_1=C_2=C_3=C_4=0$  we know that the observation is associated with car 5. Similarly, we only need 6 driver variables to categorize the 7 drivers. The regression equation is then:

Eq. 6:30

$$Y = \beta_0 + \beta_1 C_1 + \beta_2 C_2 + \beta_3 C_3 + \beta_4 C_4 + \beta_5 D_1 + \beta_6 D_2 + \beta_7 D_3 + \beta_8 D_4 + \beta_9 D_5 + \beta_{10} D_6 + \varepsilon$$

Since all the right-hand-side variables are zero-one indicator variables, the intercept term  $\beta_0$  has a specific interpretation. When all the “C” and “D” variables are equal to zero it implies we are looking at the value of Y for car 5, driver 7, so  $\beta_0$  is the mean

mileage for car 5, driver 7. The coefficients on the other “C” and “D” variables measure how different those combinations are compared to car 5, driver 7. In this context then, car 5, driver 7 are called the “reference” categories. For this example, the data in Excel would look like the following:

Y	C1	C2	C3	C4	D1	D2	D3	D4	D5	D6
18.3	1	0	0	0	1	0	0	0	0	0
21.15	1	0	0	0	0	1	0	0	0	0
15.97	1	0	0	0	0	0	1	0	0	0
19.97	1	0	0	0	0	0	0	1	0	0
17.29	1	0	0	0	0	0	0	0	1	0
16.66	1	0	0	0	0	0	0	0	0	1
19.14	1	0	0	0	0	0	0	0	0	0
14.95	0	1	0	0	1	0	0	0	0	0
16.07	0	1	0	0	0	1	0	0	0	0

Including all 4 of the car variables and all 6 of the driver variables, the Excel regression output is as follows:

**SUMMARY OUTPUT**

---

*Regression Statistics*

Multiple R	0.6657
R Square	0.4431
Adjusted R Square	0.2111
Standard Error	2.0799
Observations	35

**ANOVA**

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	10	82.617	8.262	1.910	0.094
Residual (SSE)	24	103.819	4.326		
Total	34	186.437			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	18.598	1.1660	15.9501	0.0000	16.1912	21.0042
C1	-0.691	1.1117	-0.6219	0.5398	-2.9859	1.6031
C2	-2.384	1.1117	-2.1447	0.0423	-4.6788	-0.0898
C3	-3.044	1.1117	-2.7383	0.0115	-5.3388	-0.7498
C4	-0.529	1.1117	-0.4754	0.6388	-2.8231	1.7659
D1	-0.858	1.3154	-0.6523	0.5204	-3.5729	1.8569
D2	1.468	1.3154	1.1160	0.2755	-1.2469	4.1829
D3	-0.496	1.3154	-0.3771	0.7094	-3.2109	2.2189
D4	2.186	1.3154	1.6618	0.1096	-0.5289	4.9009
D5	0.266	1.3154	0.2022	0.8415	-2.4489	2.9809
D6	0.57	1.3154	0.4333	0.6687	-2.1449	3.2849

In order to test the hypotheses that there are no differences across cars or drivers, we estimate two additional “restricted” versions of the regression – one that forces

all the car effects to be zero (it leaves out the car variables) and another that forces all the driver effects to be zero. The general form of the test statistic for these hypotheses is as follows:

*Eq. 6:31*

$$F_{calc} = \frac{(SSE_R - SSE_U)/(DF_R - DF_U)}{SSE_U/DF_U} \sim F(DF_R - DF_U, DF_U)$$

Where  $SSE_U$  is the SSE from the “unrestricted” regression above,  $DF_U$  is the degrees of freedom for the unrestricted regression,  $SSE_R$  is the SSE from a regression that “restricts” either the car effects or the driver effects to be zero. For this example, the ANOVA tables from the two “restricted” regressions are as follows:

ANOVA		No Driver Effects			
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	4	48.085	12.021	2.607	0.055
Residual (SSE)	30	138.352	4.612		
Total	34	186.437			

ANOVA		No Car Effects			
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	6	34.533	5.755	1.061	0.409
Residual (SSE)	28	151.904	5.425		
Total	34	186.437			

To test the null hypothesis of no driver effects we would then use:

*Eq. 6:32*

$$F_{calc} = \frac{(138.352 - 103.819)/(30 - 24)}{103.819/24} = 1.33$$

At a significance value of  $\alpha=0.05$ , the critical F for 6 and 24 degrees of freedom is 2.508 so we would fail to reject the null that there are no driver effects. To test for car effects we would use:

*Eq. 6:33*

$$F_{calc} = \frac{(151.904 - 103.819)/(28 - 24)}{103.819/24} = 2.779$$

Since the 5% critical F for 4 and 24 degrees of freedom is 2.776, we would reject the null hypothesis of no car effects.

While compared to the simple two-way ANOVA with no replication, the foregoing may seem a bit cumbersome, the utility of using a multiple regression framework for ANOVA is that you can include any number of different category groups and there is no requirement that you have an equal number of observations in repeated designs.

## Section 7. NONPARAMETRIC STATISTICAL TESTS

---

### A. Introduction

The confidence intervals and statistical tests presented in Section 4 – Section 6 can be described as **parametric** statistical tests in the sense that we made specific assumptions about one or more of the population parameters that characterize the underlying probability distributions of for which the tests were used. These types of tests are entirely appropriate when the level of measurement of the variables under consideration are interval or ratio data (see Section 1.C and Section 1.D). However, we are quite often confronted with questions related to **categorical** data – does a particular observation for a variable fall in a particular category? To address these types of questions, we employ what are known as **nonparametric** or **distribution free** statistical tests. Two such tests are presented in this section: the Chi-Square Goodness-of-Fit test, and the Chi-Square Independence test.

### B. Chi-Square Goodness-of-Fit Test

In a wide range of applications, the data analyst is confronted with the question of whether a set of observed categorical data are consistent with (or “fit”) a particular distribution. That is, in the underlying population represented by a sample, are the observed cell frequencies (the number of observations that fall in a particular category) different from the expected cell frequencies. The general form of the test begins with a set of category frequency counts, as follows:

Category	Observed Frequency Count
$C_1$	$O_1$
$C_2$	$O_2$
...	...
$C_K$	$O_K$
Total Obs.	n

Where  $C_i$  is one of  $K$  mutually exclusive categories,  $O_i$  is the observed frequency count for category  $i$ , and  $n$  is the total number of observations. The assumptions underlying the chi-square goodness-of-fit test are as follows: a) Categorical/nominal data representing mutually exclusive categories are used in the analysis; b) The data represent a sample of  $n$  independent observations; c) The expected frequency of each category (or cell) is 5 or greater.

If there are  $K$  different categories of outcomes,  $O_i$  is the number of observed outcomes for category  $i$ , and  $E_i$  is the number of expected outcomes for category  $i$  (based on the distribution postulated in the null hypothesis), the null hypothesis to be tested is,  $H_0: O_i = E_i$  for all  $K$  categories, versus the alternative hypothesis  $H_A: O_i \neq E_i$  for at least one category. The following chi-square statistic then provides a test of the null hypothesis that the data are consistent with (“fit”) the hypothesized distribution:

Eq. 7:1

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

This is distributed as  $\chi^2_{(K-1)}$  and is always evaluated as a one-tailed (upper or right hand value) alternative. Thus we would reject the null hypothesis if  $\chi^2_{calc} > \chi^2_{K-1,\alpha}$  (upper). For example, suppose you have data that reflect 5 different categories of outcomes, with observed and expected outcomes as follows:

Category	Observed	Expected	$\frac{(O_i - E_i)^2}{E_i}$
1	35	50	4.50
2	40	30	3.33
3	20	12	5.33
4	1	6	4.17
5	4	2	2.00
	$\chi^2_{calc} = \sum_{i=1}^K \frac{(O_i - E_i)^2}{E_i} =$		19.33

For  $\alpha = 0.05$ , the upper critical value  $\chi^2_4 = 9.49$  so we would reject the null hypothesis.

This type of test can be used for many types of discrete probability distributions. As another example, suppose you roll a six-sided die 120 times and wish to test whether the die is “fair.” With a fair die, the probability of any individual outcome  $\{1, 2, \dots, 6\}$  on a particular roll is  $1/6$ . Thus, with a sample of  $n=120$  rolls, the expected frequency of each possible outcome under the null hypothesis of a “fair” die is:  $nP = 120 \times \frac{1}{6} = 20$ . Using the framework above, suppose you observe outcome frequencies as shown in the following table:

Category	Observed Frequency ( $O_i$ )	Expected Frequency ( $E_i$ )	$\frac{(O_i - E_i)^2}{E_i}$
1	20	20	0
2	14	20	1.8
3	18	20	0.2
4	17	20	0.45
5	22	20	0.2
6	29	20	4.05
Totals	120	120	6.70

Thus, from Eq. 7:1,  $\chi_{calc}^2 = 6.70$  versus  $\chi_{(5,0.05)}^2 = 11.07$ . We would therefore fail to reject the null of a “fair” die.

### C. Chi-Square Test of Homogeneity/Independence

The Chi-Square Goodness-of-Fit Test presented above can also be adapted for a test of **homogeneity** or **independence** in categorical data. In the underlying population(s) represented by a sample(s), are the observed cell frequencies (the number of observations that fall in a particular category) different from the expected cell frequencies. Both of these tests have the same basic form but a slightly different predicate in terms of the nature of the data.

In a test of homogeneity, there are **k** independent random samples and within each sample, there are **m** mutually exclusive categories. For example, there are  $k=8$  different Recitation sections as part of ECMT 461, and each student in each section can be categorized into one of  $m=4$  different classifications: {U1, U2, U3, U4}. The Chi-Square test of homogeneity could then be used to test whether the distribution of classifications are statistically different across the sections.

In a test of independence, there is a single overall random sample that can be categorized along two different dimensions. This is a type of two-way contingency table. For example, suppose a random sample of size  $n$  can be categorized according to Factor A with  $k$  distinct outcomes and Factor B with  $m$  distinct outcomes of the form:



		Factor A				Row Totals
		A <sub>1</sub>	A <sub>2</sub>	...	A <sub>k</sub>	
Factor B	B <sub>1</sub>	O <sub>11</sub>	O <sub>12</sub>	...	O <sub>1k</sub>	n <sub>B1</sub>
	B <sub>2</sub>	O <sub>21</sub>	O <sub>22</sub>	...	O <sub>2k</sub>	n <sub>B2</sub>
	...	...	...	...	...	...
	B <sub>m</sub>	O <sub>m1</sub>	O <sub>m2</sub>	...	O <sub>mk</sub>	n <sub>Bm</sub>
Column Totals		n <sub>A1</sub>	n <sub>A2</sub>	...	n <sub>Ak</sub>	n

Where O<sub>11</sub> is the number of observations (frequency count) that fall into the mutually exclusive combination of factor A<sub>1</sub> and B<sub>1</sub> (or, in the context of a homogeneity test, the number of observations in sample A<sub>1</sub> that fall into category B<sub>1</sub>). And similarly, O<sub>ij</sub> observations in the B<sub>i</sub>, A<sub>j</sub> combination. If Factor A and Factor B are independent (or the B category distribution across samples is homogeneous) then the expected number of observations in each cell E<sub>ij</sub> (row i and column j) is:

Eq. 7:2

$$E_{ij} = \frac{n_{Bi}n_{Aj}}{n}$$

And the calculated chi-square statistic for a test of independence (or homogeneity) is given by:

Eq. 7:3

$$\chi^2_{calc} = \sum_{i=1}^m \sum_{j=1}^k \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

This statistic is distributed as  $\chi^2_{(m-1)(k-1)}$ .

For example, consider the following data related to the number of hours spent in volunteer work per week (Factor A) and volunteer type (Factor B):

Observed Outcomes

Type of Volunteer	1-3 Hours	4-6 Hours	7-9 Hours	Row Total
Community College Students	111	96	48	255
Four-Year College Students	96	133	61	290
Nonstudents	91	150	53	294
<b>Column Total</b>	298	379	162	839

Based on these data, we can calculate Expected Outcomes under the null hypothesis of independence:

Expected Outcomes

Type of Volunteer	1-3 Hours	4-6 Hours	7-9 Hours	Row Total
Community College Students	91	115	49	255
Four-Year College Students	103	131	56	290
Nonstudents	104	133	57	294
Column Total	298	379	162	839

And the  $\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$  elements of the test statistic:

4.607	3.197	0.031	
0.476	0.030	0.447	
1.726	2.225	0.250	
$\chi^2_{calc} = \sum_{i=1}^m \sum_{j=1}^k \frac{(O_{ij} - E_{ij})^2}{E_{ij}} =$			12.991

Since there are 3 of each factor type, this statistic is distributed as  $\chi^2_4$  – chi-squared with 4 degrees of freedom. At  $\alpha=0.05$ , the critical value is 9.488, and we would therefore reject the null hypothesis of independence.

**Section 8. INTRODUCTION TO TIME SERIES ANALYSIS**

---

Coming soon. In production...

## Section 9. CLASSICAL PROBABILITY APPENDIX

---

Classical probability analysis builds on a number of intuitive ideas related to uncertain events.

### A. Basic Definitions

Let's begin with a few basic definitions:

Random Experiment – a process leading to an uncertain outcome.

Basic Outcome – a possible outcome of a random experiment.

Sample Space ( $S$ ) – the collection of all possible outcomes of a random experiment.

Event ( $E$ ) – any subset of basic outcomes from the sample space.

Intersection of Events – if  $A$  and  $B$  are two events in the sample space  $S$ , then the intersection,  $A \cap B$ , is the set of all outcomes in  $S$  that belong to both **A and B**.

Mutually Exclusive Events –  $A$  and  $B$  are mutually exclusive if they have no basic outcomes in common. The set  $A \cap B$  is empty.

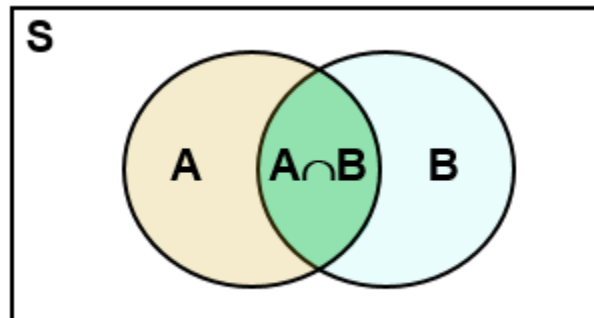
Union of Events – If  $A$  and  $B$  are two events in the sample space  $S$ , then the union,  $A \cup B$ , is the set of all outcomes in  $S$  that belong to either **A or B**.

A group of events are Collectively Exhaustive if their union completely covers the sample space  $S$ .

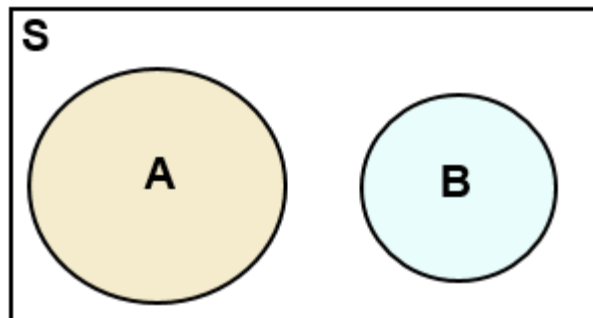
The Complement of an event  $A$  is the set of all outcomes in the sample space that do not belong to  $A$  and is denoted  $\bar{A}$ . Note that for an event  $A$  in the sample space  $S$ ,  $A \cup \bar{A}$  is collectively exhaustive.

Venn diagrams are an intuitive way to illustrate these concepts:

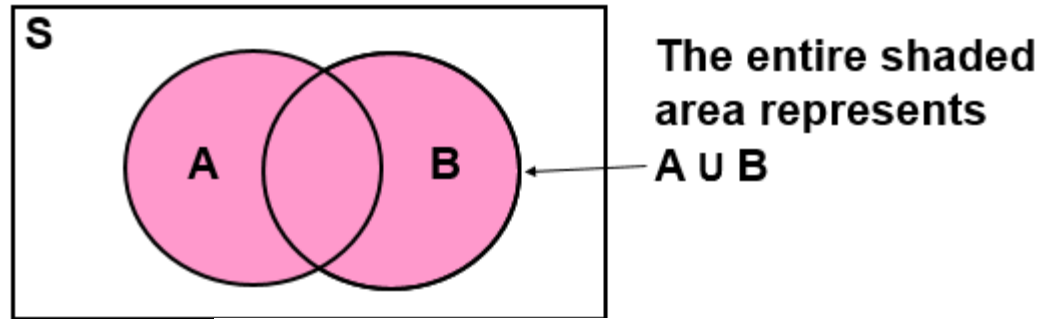
Intersection of events  $A$  and  $B$



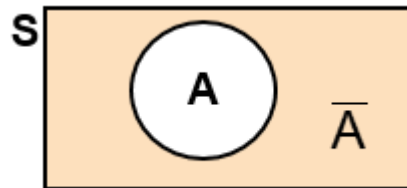
Mutually Exclusive Events



The Union of Events



The Complement of an Event



For uncertain events A and B in the sample space S, there are a number of straightforward probability postulates:

The probability that A occurs is bound on the zero-one interval. That is:  $0 \leq P(A) \leq 1$ .

The probability for the entire sample space S, is 1:  $P(S) = 1$ .

The Complement Rule:  $P(\bar{A}) = 1 - P(A)$ .

Joint Probability:  $P(A \cap B)$ , the probability of the intersection of A and B, is called a joint probability.

The Addition Rule:  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

The probabilities and joint probabilities for two events A and B can be summarized in a Probability Table:

	B	$\bar{B}$	
A	$P(A \cap B)$	$P(A \cap \bar{B})$	$P(A)$
$\bar{A}$	$P(\bar{A} \cap B)$	$P(\bar{A} \cap \bar{B})$	$P(\bar{A})$
	$P(B)$	$P(\bar{B})$	$P(S) = 1.0$

### B. Conditional Probabilities

A Conditional Probability is the probability of one event given that another event has occurred: the probability of A given B is denoted  $P(A|B)$ .

The relationship between joint and conditional probabilities is as follows:

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



**The conditional probability of A given that B has occurred**

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



**The conditional probability of B given that A has occurred**

An additional postulate is known as Statistical Independence. Two events A and B are Statistically Independent if and only if:

*Eq. 9:1*

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

If A and B are independent then:

*Eq. 9:2*

$$P(A|B) = P(A) \text{ if } P(B) > 0$$

$$P(B|A) = P(B) \text{ if } P(A) > 0$$

We can use these postulates to solve a wide range of probability problems. Suppose that of the cars on a used car lot, 70% have air conditioning (AC), 40% have a CD player (CD) and 20% of the cars have both. What is the probability that a car has a CD player given that it has AC?

We can summarize the given information in a probability table

	CD	No CD	Total
AC	0.2		0.7
No AC			
Total	0.4		

We can use the Complement Rule and the fact that the probability of the entire sample space is 1 to fill out the rest of the table:

	CD	No CD	Total
AC	0.2	<b>0.5</b>	0.7
No AC	<b>0.2</b>	<b>0.1</b>	<b>0.3</b>
Total	0.4	<b>0.6</b>	<b>1</b>

We want to find  $P(CD|AC)$  so we use the following:

*Eq. 9:3*

$$P(CD|AC) = \frac{P(CD \cap AC)}{P(AC)} = \frac{0.2}{0.7} = 0.2857$$

Here is another example from a previous exam:

In a recent survey about the direction of the U.S. economy, a sample was asked whether they thought the U.S. economy was headed in the “right direction” or the “wrong direction” (the only two choices) Of those surveyed, 41% of the respondents said that they thought the U.S. economy was headed in the right direction. Males comprised 45% of the sample, and of the males, 36% said the U.S. economy was headed in the right direction. A person is randomly chosen from the survey sample.

- a) What is the probability that the person we select is male that thinks the U.S. economy is headed in the right direction?
- b) Are the events “thinks the U.S. economy is headed in the wrong direction” and “male” statistically independent? Why or why not?
- c) What is the probability that the person we select is female?
- d) Suppose we select a respondent that thinks the U.S. economy is headed in the right direction. What is the probability that the person we select is male?

Let M denote Male, F denote Female, R denote “right direction” and W denote “wrong direction.” Use a probability table to summarize what is given and fill in missing information with the various postulates.

	R	W (or $\bar{R}$ )	Total
M			45%
F (or $\bar{M}$ )			<b>55%</b>
Total	41%	<b>59%</b>	<b>1</b>

In addition, we are given  $P(R|M) = 36\%$

For part (a) we want  $P(M \cap R)$ . Using the relationship between conditional and joint probabilities from above:

Eq. 9:4

$$P(R|M) = \frac{P(M \cap R)}{P(M)}, \text{ rearranging}$$

$$P(M \cap R) = P(R|M)P(M) = (0.36)(0.45) = 0.162$$

Given this result, we can fill in the rest of the probability table.

	R	W (or $\bar{R}$ )	Total
M	<b>16.2%</b>	<b>28.8%</b>	45%
F (or $\bar{M}$ )	<b>24.8%</b>	<b>30.2%</b>	<b>55%</b>
Total	41%	<b>59%</b>	<b>1</b>

For part b:

Statistically independent if  $P(W \cap M) = P(W)P(M)$

From the completed probability table,  $P(W \cap M) = 0.288$

$P(W)P(M) = (0.59)(0.45) = 0.2655$

So the two are not statistically independent.

Alternatively,

$$P(W \cap M) = P(W|M)P(M)$$

$$P(W|M) = 1 - P(R|M) = 0.64, \text{ so}$$

$$P(W \cap M) = (0.64)(0.45) = 0.288$$

So, again, the two are not statistically independent

For part c, from our probability table  $P(F)=0.55$

For part d, looking for  $P(M|R)$

*Eq. 9:5*

$$P(M|R) = \frac{P(M \cap R)}{P(R)} = \frac{0.162}{0.41} = 0.3951$$

### C. Bayes Theorem

1. Let A and B be two events in the sample space. Bayes Theorem states:

*Eq. 9:6*

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}, \text{ and}$$

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

Alternatively, using the probability postulates from above:

*Eq. 9:7*

$$P(A) = P(A \cap B) + P(A \cap \bar{B})$$

$$P(A \cap B) = P(A|B)P(B), \text{ and } P(A \cap \bar{B}) = P(A|\bar{B})P(\bar{B}), \text{ so}$$

$$P(A) = P(A|B)P(B) + P(A|\bar{B})P(\bar{B})$$

Bayes Theorem can thus be restated as:

*Eq. 9:8*

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|\bar{B})P(\bar{B})}, \text{ and}$$

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

More generally, let A be an event in the sample space and let  $E_i$  be the  $i^{\text{th}}$  event of  $k$  mutually exclusive and collectively exhaustive events. Then Bayes Theorem states:

*Eq. 9:9*

$$P(E_i|A) = \frac{P(A|E_i)P(E_i)}{P(A|E_1)P(E_1) + P(A|E_2)P(E_2) + \dots + P(A|E_k)P(E_k)}$$

Several examples help to illustrate.



Example 1: Let the sample space  $S = \{A, \bar{A}, B, \bar{B}\}$

Given  $P(A) = 0.6, P(B|A) = 0.6$ , and  $P(B|\bar{A}) = 0.4$ , what is  $P(A|B)$

First, using the complement rule:

$$\begin{aligned} P(A) = 0.6 &\rightarrow P(\bar{A}) = 0.4 \\ P(B|A) = 0.6 &\rightarrow P(\bar{B}|A) = 0.4 \\ P(B|\bar{A}) = 0.4 &\rightarrow P(\bar{B}|\bar{A}) = 0.6 \end{aligned}$$

Then:

$$\begin{aligned} P(A|B) &= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\bar{A})P(\bar{A})} \\ &= \frac{(0.6)(0.6)}{(0.6)(0.6) + (0.4)(0.4)} = 0.6923 \end{aligned}$$

Example 2: An economics professor finds that 21% of students earn a course grade of A. Of those students who obtain a course grade of A, 67% obtained an A on the midterm examination. Also, 15% of the students who did not obtain a course grade of A, earned an A on the midterm examination.

- (a) If a student earned a course grade of A, what is the probability that a student did not earn an A on the midterm exam?

Let  $A_1$  be the event “earns a course grade of A” and  $A_2$  be the complement of  $A_1$  (did not earn a course grade of A). Let  $B_1$  be the event “earned an A on the midterm exam” and let  $B_2$  be the complement of  $B_1$ . Then:

$$\begin{aligned} \text{Given: } P(A_1) = 0.21 &\rightarrow P(A_2) = 0.79, P(B_1|A_1) = 0.67 \rightarrow P(B_2|A_1) = 0.33, \\ P(B_1|A_2) = 0.15 &\rightarrow P(B_2|A_2) = 0.85 \end{aligned}$$

For part (a), looking for  $P(B_2|A_1) = 0.33$

- (b) For a student that earned an A on the midterm examination, what is the probability that they earned a course grade of A?

Looking for  $P(A_1|B_1)$

$$P(A_1|B_1) = \frac{P(B_1|A_1)P(A_1)}{P(B_1|A_1)P(A_1) + P(B_1|A_2)P(A_2)} = \frac{(0.67)(0.21)}{(0.67)(0.21) + (0.15)(0.79)} \cong 0.5428$$

