

# Statistics Review

Created for St095

Written by John M., Udacity Course Manager



Copyright © 2013 Udacity

UDACITY.COM

Licensed under the Creative Commons Attribution-NonCommercial 3.0 Unported License (the “License”). You may not use this file except in compliance with the License. You may obtain a copy of the License at <http://creativecommons.org/licenses/by-nc/3.0>. Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an “AS IS” BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

*Summer 2013*

## Contents

<b>1</b>	<b>Intro to statistical research methods</b>	<b>7</b>
1.1	Constructs	7
1.2	Population vs Sample	7
1.3	Experimentation	7
<b>2</b>	<b>Visualizing Data</b>	<b>9</b>
2.1	Frequency	9
2.1.1	Proportion	9
2.2	Histograms	9
2.2.1	Skewed Distribution	10
2.3	Practice Problems	12
<b>3</b>	<b>Central Tendency</b>	<b>13</b>
3.1	Mean, Median and Mode	13

<b>3.2</b>	<b>Practice Problems</b>	<b>14</b>
<b>4</b>	<b>Variability</b> .....	<b>17</b>
<b>4.1</b>	<b>Box Plots and the IQR</b>	<b>17</b>
4.1.1	Finding outliers .....	18
<b>4.2</b>	<b>Variance and Standard Deviation</b>	<b>18</b>
4.2.1	Bessel's Correction .....	18
<b>4.3</b>	<b>Practice Problems</b>	<b>18</b>
<b>5</b>	<b>Standardizing</b> .....	<b>19</b>
<b>5.1</b>	<b>Z score</b>	<b>19</b>
5.1.1	Standard Normal Curve .....	19
<b>5.2</b>	<b>Examples</b>	<b>20</b>
5.2.1	Finding Standard Score .....	20
<b>5.3</b>	<b>Practice Problems</b>	<b>21</b>
<b>6</b>	<b>Normal Distribution</b> .....	<b>23</b>
<b>6.1</b>	<b>Probability Distribution Function</b>	<b>23</b>
6.1.1	Finding the probability .....	24
<b>6.2</b>	<b>Practice Problems</b>	<b>25</b>
<b>7</b>	<b>Sampling Distributions</b> .....	<b>27</b>
<b>7.1</b>	<b>Central Limit Theorem</b>	<b>27</b>

7.2	Practice Problems	28
<b>8</b>	<b>Estimation</b>	<b>29</b>
8.1	Confidence Intervals	29
8.1.1	Critical Values	29
8.2	Practice Problems	30
<b>9</b>	<b>Hypothesis testing</b>	<b>31</b>
9.1	What is a Hypothesis test?	31
9.1.1	Error Types	32
9.2	Practice Problems	33
<b>10</b>	<b>t-Tests</b>	<b>35</b>
10.1	t-distribution	35
10.1.1	Cohen's d	35
10.2	Practice Problem	36
<b>11</b>	<b>t-Tests continued</b>	<b>37</b>
11.1	Standard Error	37
<b>12</b>	<b>One-way ANOVA</b>	<b>39</b>
12.1	Anova Testing	39
12.1.1	F-Ratio	39
12.2	Practice Problem	40
<b>13</b>	<b>ANOVA continued</b>	<b>41</b>
13.1	Means	41
13.2	Tukey's HSD	41
13.3	Practice Problems	42
<b>14</b>	<b>Correlation</b>	<b>43</b>
14.1	Scatterplots	43
14.1.1	Relationships in Data	43
14.2	Practice Problems	43
<b>15</b>	<b>Regression</b>	<b>45</b>
15.1	Linear Regression	45

15.2	Practice Problems	45
16	Chi-Squared tests .....	47
16.1	Scales of measurement	47
16.2	Chi-Square GOF test	47
16.2.1	Chi-Square test of independence .....	48
16.3	Practice Problem	48
17	Acknowledgements .....	49



# 1 — Intro to statistical research methods

## 1.1 Constructs

**Definition 1.1 — Construct.** A construct is anything that is difficult to measure because it can be defined and measured in many different ways.

**Definition 1.2 — Operational Definition.** The operational definition of a construct is the unit of measurement we are using for the construct. Once we operationally define something it is no longer a construct.

■ **Example 1.1** Volume is a construct. We know volume is the space something takes up but we haven't defined how we are measuring that space. (i.e. liters, gallons, etc.) ■

**R** Had we said volume in *liters*, then this would **not** be a construct because now it is operationally defined.

■ **Example 1.2** Minutes is already operationally defined; there is no ambiguity in what we are measuring. ■

## 1.2 Population vs Sample

**Definition 1.3 — Population.** The population is *all* the individuals in a group.

**Definition 1.4 — Sample.** The sample is *some* of the individuals in a group.


**Definition 1.5 — Parameter vs Statistic.** A *parameter* defines a characteristic of the population whereas a *statistic* defines a characteristic of the sample.

■ **Example 1.3** The mean of a population is defined with the symbol  $\mu$  whereas the mean of a sample is defined as  $\bar{x}$  ■

## 1.3 Experimentation

**Definition 1.6 — Treatment.** In an experiment, the manner in which researchers handle subjects is called a treatment. Researchers are specifically interested in how different treatments might yield differing results.

**Definition 1.7 — Observational Study.** An observational study is when an experimenter watches a group of subjects and does not introduce a treatment.

 A survey is an example of an observational study

**Definition 1.8 — Independent Variable.** The independent variable of a study is the variable that experimenters choose to manipulate; it is usually plotted along the x-axis of a graph.

**Definition 1.9 — Dependent Variable.** The dependent variable of a study is the variable that experimenters choose to measure during an experiment; it is usually plotted along the y-axis of a graph.

**Definition 1.10 — Treatment Group.** The group of a study that receives varying levels of the independent variable. These groups are used to measure the effect of a treatment.

**Definition 1.11 — Control Group.** The group of a study that receives no treatment. This group is used as a baseline when comparing treatment groups.

**Definition 1.12 — Placebo.** Something given to subjects in the control group so they think they are getting the treatment, when in reality they are getting something that causes no effect to them. (e.g. a Sugar pill)

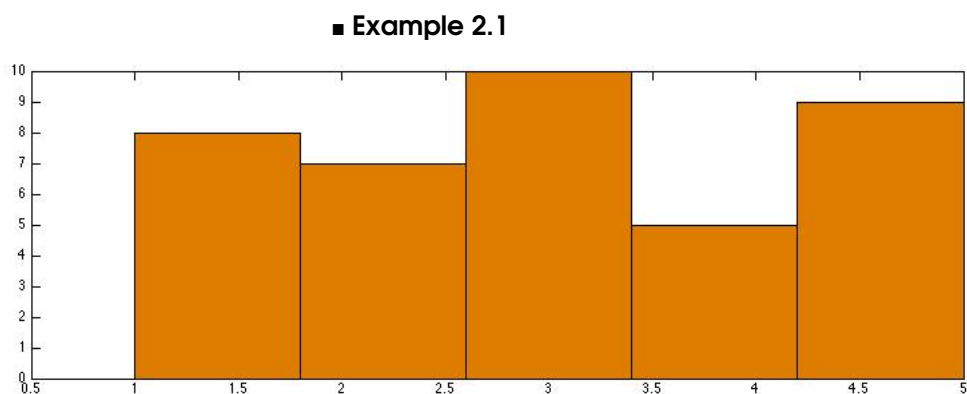
**Definition 1.13 — Blinding.** Blinding is a technique used to reduce bias. Double blinding ensures that both those administering treatments and those receiving treatments do not know who is receiving which treatment.



## 2 — Visualizing Data

### 2.1 Frequency

**Definition 2.1 — Frequency.** The frequency of a data set is the number of times a certain outcome occurs.



This histogram shows the scores on students tests from 0-5. We can see no students scored 0, 8 students scored 1. These counts are what we call the frequency of the students scores. ■


#### 2.1.1 Proportion

**Definition 2.2 — Proportion.** A proportion is the fraction of counts over the total sample. A proportion can be turned into a percentage by multiplying the proportion by 100.

■ **Example 2.2** Using our histogram from above we can see the proportion of students who scored a 1 on the test is equal to  $\frac{8}{39} \approx 0.2051$  or 20.51% ■

### 2.2 Histograms

**Definition 2.3 — Histogram.** is a graphical representation of the distribution of data, discrete intervals (bins) are decided upon to form widths for our boxes.

 Adjusting the bin size of a histogram will compact (or spread out) the distribution.

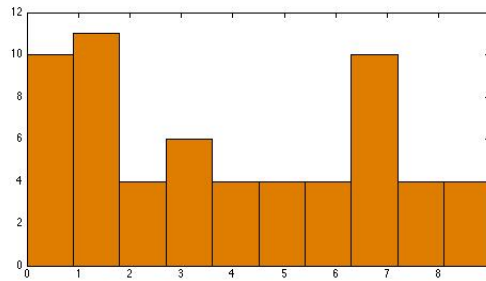


Figure 2.1: histogram of data set with bin size 1

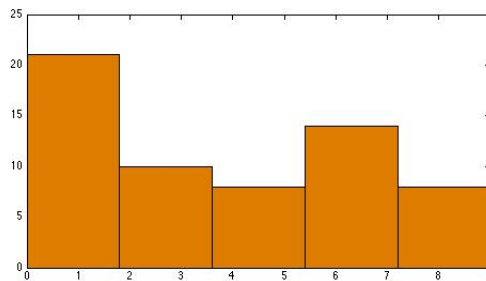


Figure 2.2: histogram of data set with bin size 2

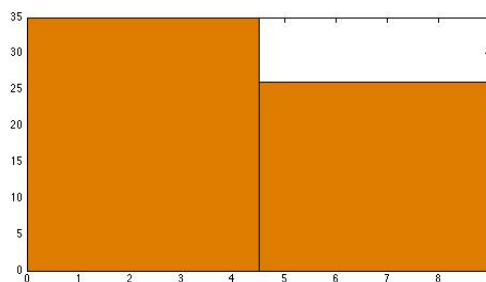


Figure 2.3: histogram of data set with bin size 5

### 2.2.1 Skewed Distribution

**Definition 2.4 — Positive Skew.** A positive skew is when outliers are present along the right most end of the distribution

**Definition 2.5 — Negative Skew.** A negative skew is when outliers are present along the left most end of the distribution

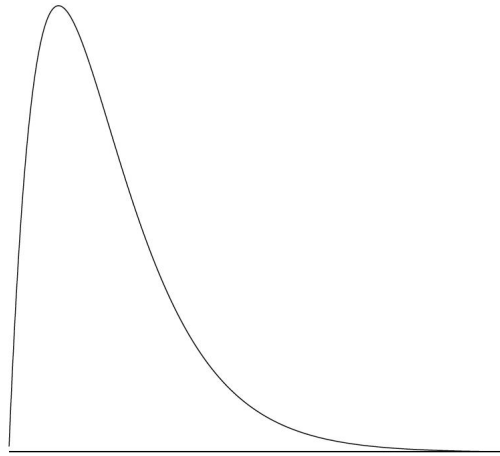


Figure 2.4: positive skew

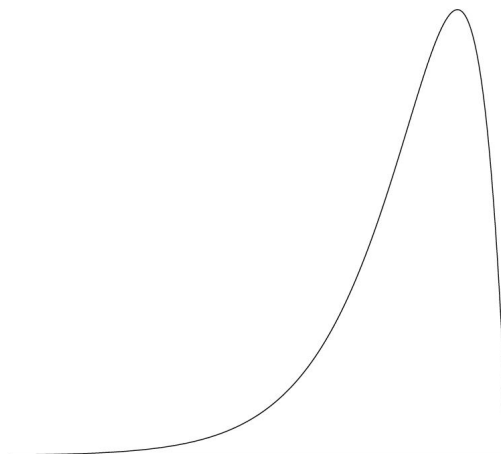


Figure 2.5: negative skew

## 2.3 Practice Problems

**Problem 2.1** Kathleen counts the number of petals on all the flowers in her garden, create a histogram and describe the distribution of flower petals on Kathleen's flowers. Use a bin size of 2.

15	16	17
16	21	22
15	16	15
17	16	22
14	13	14
14	15	15
14	15	16
10	19	15
15	22	24
25	15	16

Table 2.1: Kathleens petal counts

**Problem 2.2** What number of petals seems most prominent in Kathleen's garden? What happens if we change the bin size to 5?

**Problem 2.3** What does the skew in Kathleen's flower petal distribution seem to indicate?



## 3 — Central Tendency

### 3.1 Mean, Median and Mode

**Definition 3.1 — Mean.** The mean of a dataset is the numerical average and can be computed by dividing the sum of all the data points by the number of data points:

$$\bar{x} = \frac{\sum_{i=0}^n x_i}{n}$$

**R** The mean is heavily affected by outliers, therefore we say the mean is *not* a robust measurement.

**Definition 3.2 — Median.** The median of a dataset is the datapoint that is directly in the middle of the data set. If two numbers are in the middle then the median is the average of the two.

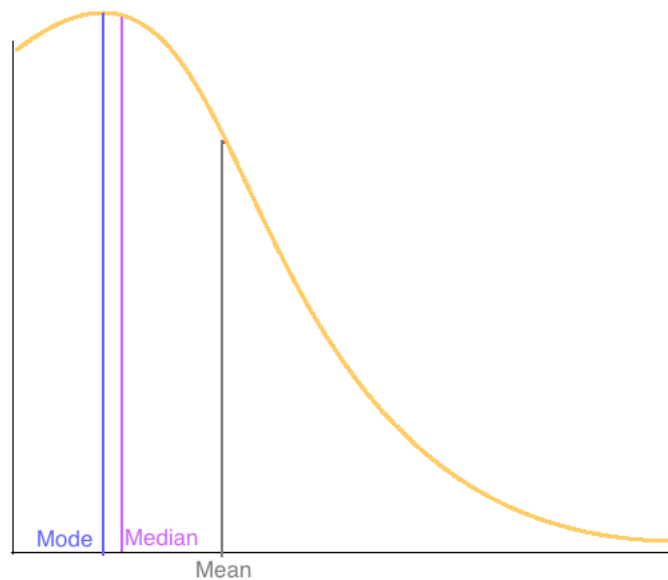
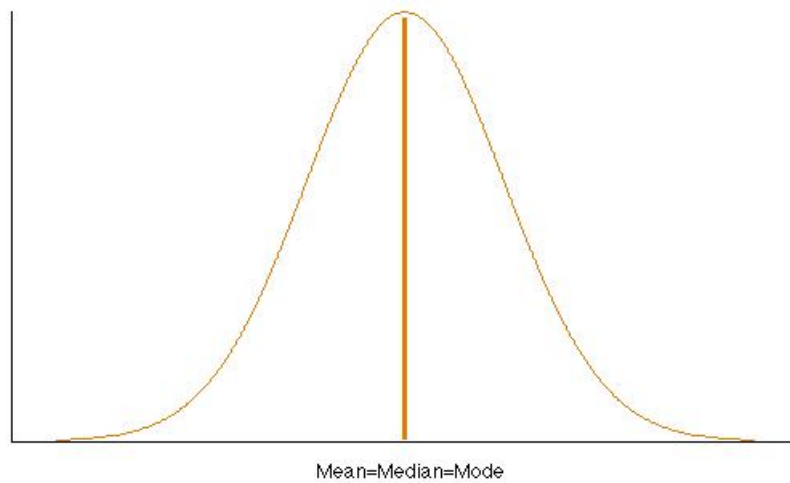
1. The data set is odd  $n/2 =$  the position in the data set the middle value is
2. The data set is even  $\frac{x_k + x_{k+1}}{n}$  gives the median for the two middle data points

**R** The median is robust to outliers, therefore an outlier will not affect the value of the median.

**Definition 3.3 — Mode.** The mode of a dataset is the datapoint that occurs the most frequently in the data set.

**R** The mode is robust to outliers as well.

**R** In the normal distribution the mean = median = mode.



### 3.2 Practice Problems

**Problem 3.1** Find the mean, median and mode of the data set

**Problem 3.2** A secret club collects the following monthly income data from its members. Find the mean, median, and mode of these incomes. Which measure of center would best describe this distribution?

15	16	17
16	21	22
15	16	15
17	16	22
14	13	14
14	15	15
14	15	16
10	19	15
15	22	24
25	15	16

Table 3.1: Problem 1

\$2500	\$3000	\$2900
\$2650	\$3225	\$2700
\$2740	\$3000	\$3400
\$2500	\$3100	\$2700

Table 3.2: Incomes





## 4 — Variability

### 4.1 Box Plots and the IQR

A box plot is a great way to show the 5 number summary of a data set in a visually appealing way. The 5 number summary consists of the minimum, first quartile, median, third quartile, and the maximum

**Definition 4.1 — Interquartile range.** The Interquartile range (IQR) is the distance between the 1st quartile and 3rd quartile and gives us the range of the middle 50% of our data. The IQR is easily found by computing:  $Q3 - Q1$

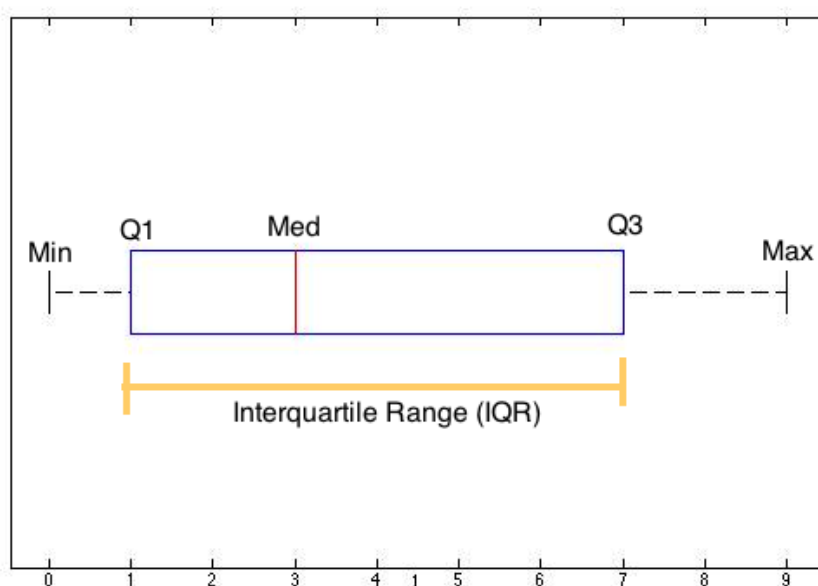


Figure 4.1: A simple boxplot

### 4.1.1 Finding outliers

**Definition 4.2 — How to identify outliers.** You can use the IQR to identify outliers:

1. Upper outliers:  $Q3 + 1.5 \cdot IQR$
2. Lower outliers:  $Q1 - 1.5 \cdot IQR$

## 4.2 Variance and Standard Deviation

**Definition 4.3 — Variance.** The variance is the average of the squared differences from the mean. The formula for computing variance is:

$$\sigma^2 = \frac{\sum_{i=0}^n (x_i - \bar{x})^2}{n}$$

**Definition 4.4 — Standard Deviation.** The standard deviation is the square root of the variance and is used to measure distance from the mean.

**R** In a normal distribution 65% of the data lies within 1 standard deviation from the mean, 95% within 2 standard deviations, and 99.7% within 3 standard deviations.

### 4.2.1 Bessel's Correction

**Definition 4.5 — Bessel's Correction.** Corrects the bias in the estimation of the population variance, and some (but not all) of the bias in the estimation of the population standard deviation. To apply Bessel's correction we multiply the variance by  $\frac{n}{n-1}$ .

**R** Use Bessel's correction primarily to estimate the population standard deviation.

## 4.3 Practice Problems

**Problem 4.1** Make a box plot of the following monthly incomes

\$2500	\$3000	\$2900
\$2650	\$3225	\$2700
\$2740	\$3000	\$3400
\$2500	\$3100	\$2700

Table 4.1: Incomes

**Problem 4.2** Find the standard deviation of the incomes.

**Problem 4.3** What is a better descriptor of the distribution the box plot, or the mean and standard deviation? Why?

## 5 — Standardizing

### 5.1 Z score

**Definition 5.1 — Standard Score.** Given an observed value  $x$ , the Z score finds the number of Standard deviations  $x$  is away from the mean.

$$Z = \frac{x - \mu}{\sigma}$$

#### 5.1.1 Standard Normal Curve

The standard normal curve is the curve we will be using for most problems in this section. This curve is the resulting distribution we get when we standardize our scores. We will use this distribution along with the Z table to compute percentages above, below, or in between observations in later sections.

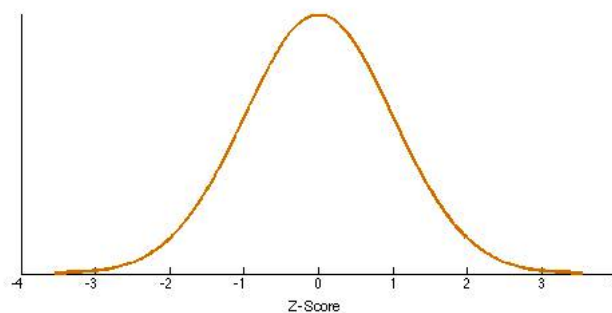


Figure 5.1: The Standard Normal Curve

## 5.2 Examples

### 5.2.1 Finding Standard Score

■ **Example 5.1** The average height of a professional basketball player was 2.00 meters with a standard deviation of 0.02 meters. Harrison Barnes is a basketball player who measures 2.03 meters. How many standard deviations from the mean is Barnes' height?

First we should sketch the normal curve that represents the distribution of basketball player heights.

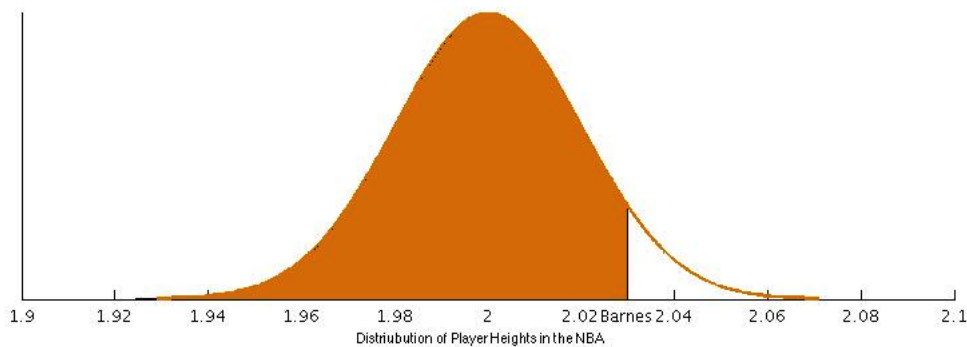


Figure 5.2: Notice we place the mean height 2.00 right in the middle and make tick marks that are each 1 standard deviation or 0.02 meters away in both directions.

Next we should compute the standard score (i.e. z score) for Barnes' height. Since  $\mu = 2.00$ ,  $\sigma = 0.02$ , and  $x = 2.03$  we can find the z-score

$$\frac{x - \mu}{\sigma} = \frac{2.03 - 2.00}{0.02} = \frac{0.03}{0.02} = 1.5$$

**R** Finding 1.5 as the z score tells us that Barnes' height is 1.5 standard deviations from the mean, that is  $1.5\sigma + \mu = \text{Barnes' Height}$

■

■ **Example 5.2** The average height of a professional hockey player is 1.86 meters with a standard deviation of 0.06 meters. Tyler Myers, a professional hockey, is the same height as Harrison Barnes. Which of the two is taller in their respective league?

To find Tyler Myers standard score we can use the information:  $\mu = 1.86$ ,  $\sigma = 0.06$ , and  $x = 2.03$ . This results in the standard score:

$$\frac{x - \mu}{\sigma} = \frac{2.03 - 1.86}{0.06} = \frac{0.17}{0.06} = 2.833$$

Comparing the two z-scores we see that Tyler Myers score of 2.833 is larger than Barnes' score of 1.5. This tells us that there are more hockey players shorter than Myers than there are basketball players shorter than Barnes'.

■

### 5.3 Practice Problems

Find the Z-score given the following information

**Problem 5.1**  $\mu = 54, \sigma = 12, x = 68$

**Problem 5.2**  $\mu = 25, \sigma = 3.5, x = 20$

**Problem 5.3**  $\mu = 0.01, \sigma = 0.002, x = 0.01$

**Problem 5.4** The average GPA of students in a local high school is 3.2 with a standard deviation of 0.3. Jenny has a GPA of 2.8. How many standard deviations away from the mean is Jenny's GPA?

**Problem 5.5** Jenny's trying to prove to her parents that she is doing better in school than her cousin. Her cousin goes to a different high school where the average GPA is 3.4 with a standard deviation of 0.2. Jenny's cousin has a GPA of 3.0. Is Jenny performing better than her cousin based on standard scores?

**Problem 5.6** Kyle's score on a recent math test was 2.3 standard deviations above the mean score of 78%. If the standard deviation of the test scores were 8%, what score did Kyle get on his test?





## 6 — Normal Distribution

### 6.1 Probability Distribution Function

**Definition 6.1 — Probability Distribution Function.** The probability distribution function is a normal curve with an area of 1 beneath it, to represent the cumulative frequency of values.

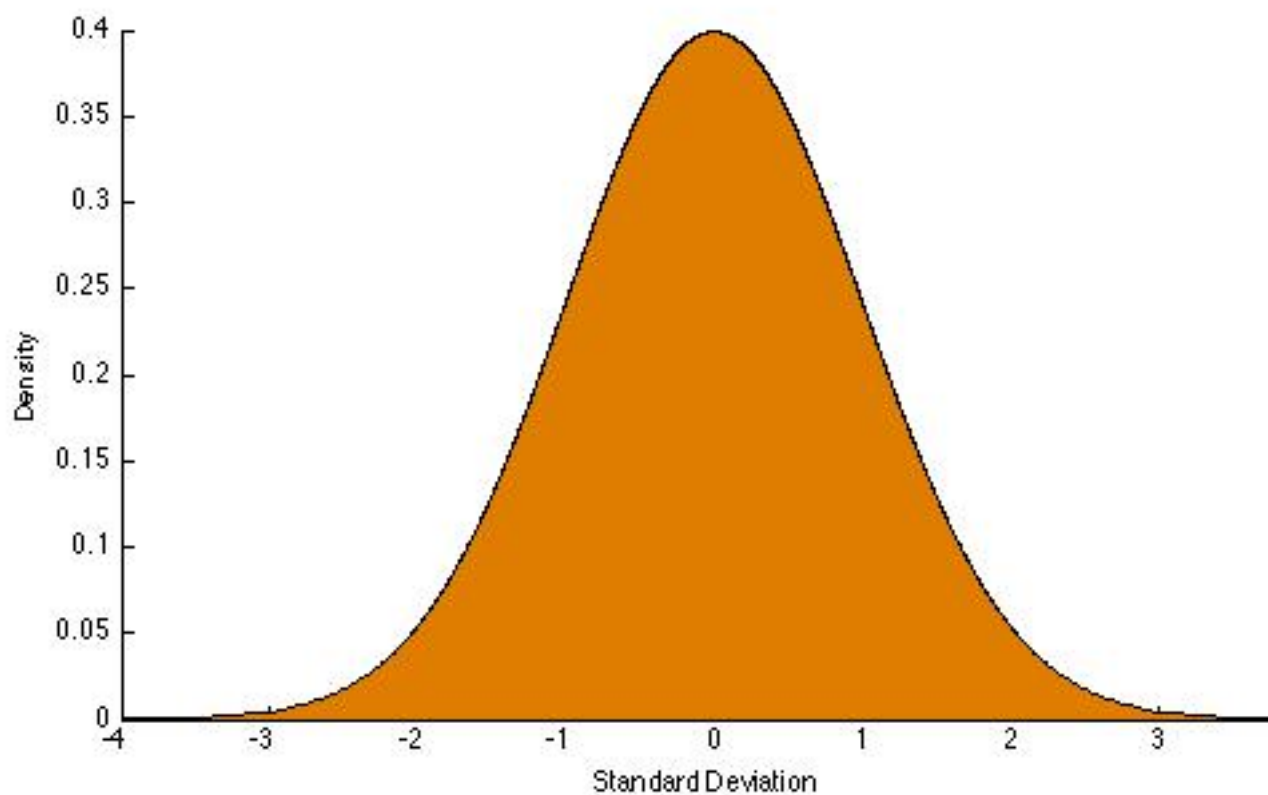


Figure 6.1: The area beneath the curve is 1

### 6.1.1 Finding the probability

We can use the PDF to find the probability of specific measurements occurring. The following examples illustrate how to find the area below, above, and between particular observations.

■ **Example 6.1** The average height of students at a private university is 1.85 meters with a standard deviation of 0.15 meters. What percentage of students are shorter or as tall as Margie who stands at 2.00 meters.

To solve this problem the first thing we need to find is our z-score:

$$z = \frac{x - \mu}{\sigma} = \frac{2.05 - 1.85}{0.15} = 1.3$$

Now we need to use the z-score table to find the proportion below a z-score of 1.33.

**R** The z-table only shows the proportion below. In this instance we are trying to find the orange area.

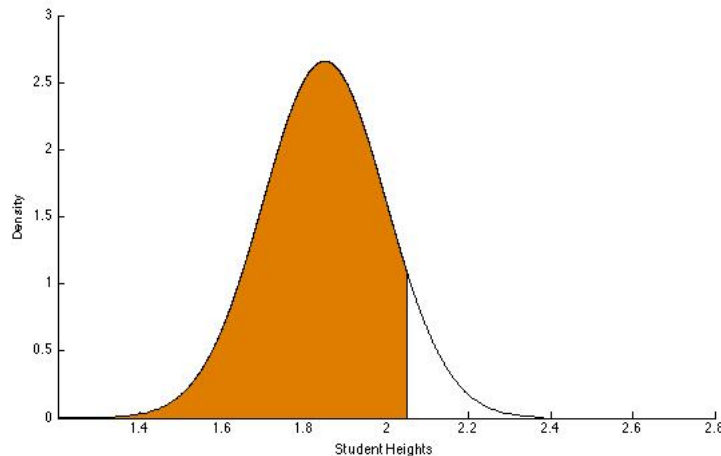


Figure 6.2: 85% is the shaded area

To use the z-table we start in the left most column and find the first two digits of our z-score (in this case 1.3) then we find the third digit along the top of the table. Where this row and column intersect is our proportion below that z-score.

■ **Example 6.2** Margie also wants to know what percent of students are taller than her. Since the area under the normal curve is 1 we can find that proportion:

$$1 - 0.9082 = 0.0918 = 9.18\%$$

■ **Example 6.3** Anne only measures 1.87 meters. What proportion of classmates are between Anne and Margie's heights.

We already know that 90.82% of students are shorter than Margie. So let's first find the percent of students that are shorter than Anne.



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633

Figure 6.3: using the z-table for 1.33

This means that Margie is taller than 90.82% of her classmates.

$$\frac{1.87 - 1.85}{0.15} = 0.1\bar{3}$$

If we use the z-table we see that this z-score corresponds with a proportion of 0.5517 or 55.17%. So to get the proportion in between the two we subtract the two proportions from each other. That is the proportion of people who's height's are between Anne and Margies height is  $90.82 - 55.17 = 35.65\%$ .

■

## 6.2 Practice Problems

**Problem 6.1** In 2007-2008 the average height of a professional basketball player was 2.00 meters with a standard deviation of 0.02 meters. Harrison Barnes is a basketball player who measures 2.03 meters. What percent of players are taller than Barnes?

**Problem 6.2** Chris Paul is 1.83 meters tall. What proportion of Basketball players are between Paul and Barne's heights?

**Problem 6.3** 92% of candidates scored as good or worse on a test than Steve. If the average score was a 55 with a standard deviation of 6 points what was Steve's score?





## 7 — Sampling Distributions

### 7.1 Central Limit Theorem

The Central Limit Theorem is used to help us understand the following facts regardless of whether the population distribution is normal or not:

1. the mean of the sample means is the same as the population mean
2. the standard deviation of the sample means is always equal to the standard error (i.e.  $SE = \frac{\sigma}{\sqrt{n}}$ )
3. the distribution of sample means will become increasingly more normal as the sample size,  $n$ , increases.

**Definition 7.1 — Sampling Distribution.** The sampling distribution of a statistic is the distribution of that statistic. It may be considered as the distribution of the statistic for all possible samples from the same population of a given size.

■ **Example 7.1** We are interested in the average height of trees in a particular forest. To get results quickly we had 5 students go out and measure a sample of 20 trees. Each student returned with the average tree height from their samples.

Sample results : 35.23 , 36.71, 33.21, 38.2, 35.54

If it is known that the population average of tree heights in the forest is 36 feet with a standard deviation of 2 feet. How many Standard errors is the students average away from the population mean?

To solve this problem we first need to find the average of these students averages so

$$\bar{x} = \frac{35.23 + 36.71 + 33.21 + 38.2 + 35.54}{5} = 35.78$$

Now we find our Standard error of the sample:

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{2}{5} = 0.4$$

So now to get the number of standard errors away from the mean our observation is we can use the z-score formula:

$$\frac{35.78 - 36}{0.4} = -0.55$$

So our sample distribution is relatively close to the population distribution! ■

## 7.2 Practice Problems

**Problem 7.1** The known average time it takes to deliver a pizza is 22.5 minutes with a standard deviation of 2 minutes. I ordered pizza every week for the last 10 weeks and got an average time of 18.5 minutes. What is the probability that get this average?

**Problem 7.2** If I continue to order pizzas for eternity what could I expect this average to get close to?



## 8 — Estimation

### 8.1 Confidence Intervals

**Definition 8.1 — Margin of error.** The margin of error of a distribution is the amount of error we predict when estimating the population parameters from sample statistics. The margin of error is computed as:

$$Z^* \cdot \frac{\sigma}{\sqrt{n}}$$

Where  $Z^*$  is the critical z-score for the level of confidence.

**Definition 8.2 — Confidence level.** The confidence level of an estimate is the percent of all possible sample means that fall within a margin of error of our estimate. That is to say that we are some % sure the the true population parameter falls within a specific range

**Definition 8.3 — Confidence Interval.** A confidence interval is a range of values in which we suspect the population parameter lies between. To compute the confidence interval we use the formula:

$$\bar{x} \pm Z^* \cdot \frac{\sigma}{\sqrt{n}}$$

This gives us an upper and lower bound that capture our population mean.

#### 8.1.1 Critical Values

The critical z-score is used to define a critical region for our confidence interval. Observations beyond this critical region are considered observations so extreme that they were very unlikely to have just happened by chance.

## 8.2 Practice Problems

**Problem 8.1** Find a confidence interval for the distribution of pizza delivery times.

Company A
20.4
24.2
15.4
21.4
20.2
18.5
21.5

Table 8.1: Pizza Companies Delivery Times



## 9 — Hypothesis testing

### 9.1 What is a Hypothesis test?

A hypothesis test is used to test a claim that someone has about how an observation may be different from the known population parameter.

**Definition 9.1 — Alpha level ( $\alpha$ ).** The alpha level ( $\alpha$ ) of a hypothesis test helps us determine the critical region of a distribution.

**Definition 9.2 — Null Hypothesis.** The null hypothesis is always an equality. It is a the claim we are trying to provide evidence against. We commonly write the null hypothesis as one of the following:

$$H_0 : \mu_0 = \mu$$

$$H_0 : \mu_0 \geq \mu$$

$$H_0 : \mu_0 \leq \mu$$

**Definition 9.3 — Alternative Hypothesis.** The Alternative hypothesis is result we are checking against the claim. This is always some kind of inequality. We commonly write the alternative hypothesis as one of the following:

$$H_a : \mu_a \neq \mu$$

$$H_a : \mu_a > \mu$$

$$H_a : \mu_a < \mu$$

■ **Example 9.1** A towns census from 2001 reported that the average age of people living there was 32.3 years with a standard deviation of 2.1 years. The town takes a sample of 25 people and finds there average age to be 38.4 years. Test the claim that the average age of people in the town has increased. (Use an  $\alpha$  level of 0.05)

First lets define our hypotheses:

$$H_0 : \mu_0 = 32.3 \text{ years}$$

$$H_a : \mu_0 > 32.3 \text{ years}$$

Now let's identify the important information:

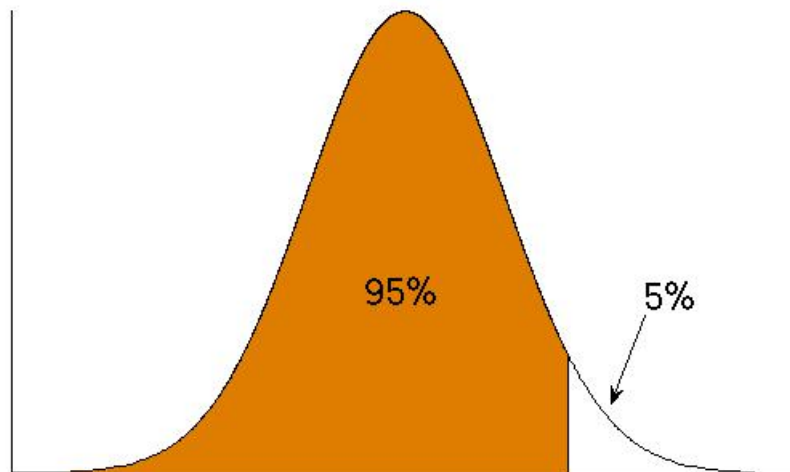
$$\bar{x} = 38.4$$

$$\sigma = 2.1$$

$$n = 25$$

$$SE = \frac{2.1}{\sqrt{25}} = 0.42$$

The last piece of important info we need is our critical value: Finding Z-critical value:



So we look up as close as we can to 95%

So that gives us a Z-crit of 1.64

Once we have all our important information we can now find our test statistic:

$$z\text{-score} = \frac{38.4 - 32.3}{0.42} = 14.5238$$

Since our z-score is much bigger than our z-crit we reject the claim (reject the null) that the average age of people living there was 32.3 years. ■

### 9.1.1 Error Types

**Definition 9.4 — Type I Error.** A Type I Error is when you reject the null when the null hypothesis is actually true. The probability of committing a Type I error is  $\alpha$

**Definition 9.5 — Type II Error.** A Type II Error is when you fail to reject the null when it is actually false. The probability of committing a Type II error is  $\beta$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857

## 9.2 Practice Problems

**Problem 9.1** An insurance company is reviewing its current policy rates. When originally setting the rates they believed that the average claim amount was \$1,800. They are concerned that the true mean is actually higher than this, because they could potentially lose a lot of money. They randomly select 40 claims, and calculate a sample mean of \$1,950. Assuming that the standard deviation of claims is \$500, and set  $\alpha = 0.05$ , test to see if the insurance company should be concerned.

**Problem 9.2** Explain a type I and type II error in context of the problem. Which is worse?



## 10 — t-Tests

### 10.1 t-distribution

The t-Test is best to use when we do not know the population standard deviation. Instead we use the sample standard deviation.

**Definition 10.1 — t-stat.** The t-Test statistic can be computed very similarly to the z-stat, to compute the t-stat we compute:

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

We also have to compute the degrees of freedom (df) for the sample:  $df = n - 1$

Like the Z-stat we can use a table to get the proportion below or between a specific value. T-tests are also great for testing two sample means (i.e. paired t-tests), we modify the formula to become:

$$\frac{(x_2 - x_1) - (\mu_2 - \mu_1)}{\frac{\sqrt{(s_1^2 + s_2^2)}}{n}}$$

#### ■ Example 10.1 ■

#### 10.1.1 Cohen's d

**Definition 10.2 — Cohen's d.** Cohen's d measures the effect size of the strength of a phenomenon. Cohen's d gives us the distance between means in standardized units. Cohen's d is computed by:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s}$$

where  $s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$

## 10.2 Practice Problem

**Problem 10.1** Pizza company A wants to know if they deliver Pizza faster than Company B. The following table outlines there delivery times:

Company A	Company B
20.4	20.2
24.2	16.9
15.4	18.5
21.4	17.3
20.2	20.5
18.5	
21.5	

Table 10.1: Pizza Companies Delivery Times

**Problem 10.2** Use Cohen's d to measure the effect size between the two times.



## 11 — t-Tests continued

### 11.1 Standard Error

**Definition 11.1 — Standard Error.** The Standard error is the standard deviation of the sample means over all possible samples (of a given size) drawn from the population. It can be computed by:


$$SE = \frac{\sigma}{\sqrt{n}}$$

The standard error for two samples can be computed with:

$$SE = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

**Definition 11.2 — Pooled Variance.** Pooled variance is a method for estimating variance given several different samples taken in different circumstances where the mean may vary between samples but the true variance is assumed to remain the same. The pooled variance is computed by using:

$$S_p^2 = \frac{SS_1 + SS_2}{df_1 + df_2}$$

 We can use pooled variance to compute standard error that is:

$$SE = \sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}$$





## 12 — One-way ANOVA

### 12.1 Anova Testing

The grand mean of several data sets is simply the sum of all the data divided by the number of data points. The grand mean is commonly given the symbol  $\bar{x}_G$

**Definition 12.1 — Between-Group Variability.** Describes the distance between the sample means of several data sets and can be computed as the Sum of Squares Between divided by the degrees of freedom between:

$$SS_{between} = n \sum (\bar{x}_k - \bar{x}_G)^2$$

$$df_{between} = k - 1$$

where k is the number samples

**Definition 12.2 — Within-Group Variability.** Describes the variability of each individual sample and can be computed as the Sum of Squares within divided by the degrees of freedom within:

$$SS_{within} = \sum (x_i - \bar{x}_k)^2$$

$$df_{within} = N - k$$

The hypotheses for a typical anova test are:

$$H_0 : \mu_0 = \mu_1 = \dots = \mu_k$$

$$H_a : \text{any of these means differs}$$

#### 12.1.1 F-Ratio

The F-ratio can be found by taking the between-group variability and dividing by the within-group variability. The F-ratio is used in the same way as the t-stat, or z-stat.

## 12.2 Practice Problem

**Problem 12.1** Neuroscience researchers examined the impact of environment on rat development. Rats were randomly assigned to be raised in one of the four following test conditions: Impoverished (wire mesh cage - housed alone), standard (cage with other rats), enriched (cage with other rats and toys), super enriched (cage with rats and toys changes on a periodic basis). After two months, the rats were tested on a variety of learning measures (including the number of trials to learn a maze to a three perfect trial criteria), and several neurological measure (overall cortical weight, degree of dendritic branching, etc.). The data for the maze task is below. Compute the appropriate test for the data provided below. What would be the null hypothesis in this

Impoverished Enriched	Standard Super Enriched
22	17
12	8
19	21
14	7
15	15
11	10
24	12
9	9
18	19
15	12

Table 12.1: Scores

study

**Problem 12.2** What is your F-critical?

**Problem 12.3** What is your F-stat?

**Problem 12.4** Are there any significant differences between the four testing conditions?





## 13 — ANOVA continued

### 13.1 Means

**Definition 13.1 — Group Means.** The group means are the individual mean for each group in an Anova test.

**Definition 13.2 — Mean Squares.** The  $MS_{between}$  and  $MS_{within}$  are computed as:

$$MS_{between} = \frac{SS_{between}}{df_{between}}$$

$$MS_{within} = \frac{SS_{within}}{df_{within}}$$

### 13.2 Tukey's HSD

**Definition 13.3 — Tukey's HSD.** Tukey's HSD allows us to make pairwise comparisons to determine if a significant difference occurs between means. If Tukey's HSD is greater than the difference between sample means then we consider the samples significantly different. Keep in mind that the sample sizes must be equal. Tukey's HSD is computed as:

$$q^* \sqrt{\frac{MS_{within}}{n}}$$

We can also use Cohen's d for multiple comparisons on sample sets. Using Cohen's d we have to compute the value for every possible combination of samples.

**Definition 13.4 —  $\eta^2$ .**  $\eta^2$  (read eta squared) is the proportion of total variation that is due to between group differences.

$$\eta^2 = \frac{SS_{between}}{SS_{within} + SS_{between}} = \frac{SS_{between}}{SS_{total}}$$

**R** The value of  $\eta^2$  is considered large if it is greater than 0.14

### 13.3 Practice Problems

**Problem 13.1** Amy is trying to set-up a home business of selling fresh eggs. In order to increase her profits, she wants to only use the breed of hens that produce the most eggs. She decides to run an experiment testing four different breeds of hens, counting the number of eggs laid by each breed. She purchases 10 hens of each breed for her experiment. What is the studentized range statistic ( $q^*$ ) for this experiment at an alpha level of 0.05?

**Problem 13.2** Amy finds that the  $MS_{within}$  for the first batch of eggs laid by her hens to be 45.25. How far apart do the group means for the different breeds have to be to be considered significant?

**Problem 13.3** Amy also finds that  $SS_{within} = 1629.36$  and  $SS_{between} = 254.64$ . What proportion of the total variation in the number of eggs produced by each breed can be attributed to the different breeds? (Calculate eta-squared)

**Problem 13.4** Using Tukey HSD, are the sample means significantly different?



## 14 — Correlation

### 14.1 Scatterplots

A scatterplot shows the relationship between two sets of data. Each pair of data points is represented as a single point on the plane. The more linear our set of points are the stronger the relationship between the two data sets is.

#### 14.1.1 Relationships in Data

**Definition 14.1 — Correlation coefficient (Pearson's  $r$ ).** The Correlation coefficient, commonly referred to as Pearson's  $r$ , describes the strength of the relationship between two data sets. The closer  $|r|$  is to 1 the more linear(stronger) our relationship. The closer  $r$  is to zero the more scattered(weaker) our relationship. To compute Pearson's  $r$  you can use the formula:

$$r = \frac{\text{Covariance}(x,y)}{S_x \cdot S_y}$$

**R** On a Google Docs spreadsheet we can do

```
=Pearson(start cell for variable x : end cell for variable x,  
          start cell for variable y : end cell for variable y)
```

**Definition 14.2 — Coefficient of Determination( $r^2$ ).** The coefficient of determination is the percentage of variation in the dependent variable ( $y$ ) that can be explained by variation in the independent variable ( $x$ )

### 14.2 Practice Problems

**Problem 14.1** A researcher wants to investigate the relationship between outside temperature and the number of reported acts of violence. For this investigation, what is the predictor ( $x$ ) variable and what is the outcome ( $y$ ) variable?

**Problem 14.2** Given a correlation coefficient of  $-.95$ , what direction is the relationship and how do we know this? What is the strength of this relationship and how do we know this? In

terms of strength and relationship, how does this correlation coefficient differ from one that is .95?

**Problem 14.3** What does it mean if we have a coefficient of determination = .55?

**Problem 14.4** If a researcher found that there was a strong positive correlation between outside temperature and the number of reported acts of violence, does this mean that an increase or decrease in temperature causes an increase or decrease in the number of reported acts of violence? Why or why not?



## 15 — Regression

### 15.1 Linear Regression

**Definition 15.1 — Regression Equation.** The linear regression equation  $\hat{y} = ax + b$  describes the linear equation that represents the "line of best fit". This line attempts to pass through as many of the points as possible.  $a$  is the slope of our linear regression equation and represents the rate of change in  $y$  versus  $x$ .  $b$  represents the  $y$ -intercept.

**R** The regression equation may also be written as  $\hat{y} = bx + a$

The line of best fit helps describe the dataset. It can also be used to make approximate predictions of how the data will behave.

**Corollary 15.1** We can find the linear regression equation with the two following pieces of information:

$$\text{slope} = r \frac{s_y}{s_x}$$

The regression equation passes through the point  $(\bar{x}, \bar{y})$

#### ■ Example 15.1 ■

### 15.2 Practice Problems

**Problem 15.1** Marcus wants to investigate the relationship between hours of computer usage per day and number of minutes of migraines endured per day. After collecting data, He finds a correlation coefficient of 0.86, with  $s_y = 375.55$  and  $s_x = 1814.72$ . The mean hours of computer usage from his data set was calculated to be 4.5 hours and the average number of minutes of migraine was calculated to be 25 minutes. Find the regression line that best fits his data.

**Problem 15.2** Using the line that you calculated above, given 2 hours of computer usage, how many minutes of migraine would Marcus predict to follow?

**Problem 15.3** Marcus coincidentally has a point in his data set that he collected for exactly 2 hours of computer usage. Given that the residual between his observed value for 2 hours of

computer usage and the expected value (as calculated in the previous question) equals 1.89, how many minutes of migraine did Marcus observe for that point in his data set?



## 16 — Chi-Squared tests

### 16.1 Scales of measurement

**Definition 16.1 — Ordinal Data.** There is a clear order in the data set but the distance between data points is unimportant.

**Definition 16.2 — Interval Data.** Similar to an ordinal set of data in that there is a clear ranking, but each group is divided into equal intervals

**Definition 16.3 — Ratio Data.** Similar to interval data except there exists an absolute zero.

**Definition 16.4 — Nominal Data.** This is the same as qualitative data, where we differentiate between items or subjects based only on their names and/or categories and other qualitative classifications they belong to.

Type of Data	Example	Data
Ordinal	Ranks in a race	1st, 2nd, 3rd
Interval	Temperature in Celsius	$-10^{\circ} - 0^{\circ}$ , $1^{\circ} - 10^{\circ}$ , $11^{\circ} - 20^{\circ}$
Ratio	Percentage correct on test	$0 - 10\%$ , $11 - 20\%$ , $21 - 30\%$
Nominal	Shirt Colors	Red, Blue, Yellow, White

Table 16.1: Examples of different scales of measurement

#### ■ Example 16.1



### 16.2 Chi-Square GOF test

The Chi-Square GOF test allows us to see how well observed values match expected values for a certain variable. In particular we compare the frequencies of our data sets.

### 16.2.1 Chi-Square test of independence

This variation of the Chi-Square test is used to determine if 2 nominal variables are independent. In particular we use the marginal totals.

### 16.3 Practice Problem

**Problem 16.1** A poker-dealing machine is supposed to deal cards at random, as if from an infinite deck. In a test, you counted 1600 cards, and observed the following: table[h]

Suit	Count	Card counts
Spades	404	
Hearts	420	
Diamonds	400	
Clubs	376	

Could it be that the suits are equally likely? Or are these discrepancies too much to be random?





## 17 — Acknowledgements

Thanks to Katie Kormanik, Sean Laraway and Ronald Rogers for creating the class. Also special thanks to Mathias Legrand and <http://www.LaTeXTemplates.com> for providing the template for this packet. Extra special thanks to fellow course managers Kathleen C. and Steven J. for helping with the packet!

