Constructs
Population vs Sample
Experimentation

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 μ 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.

R 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.

Frequency
Proportion
Histograms
Skewed Distribution
Practice Problems

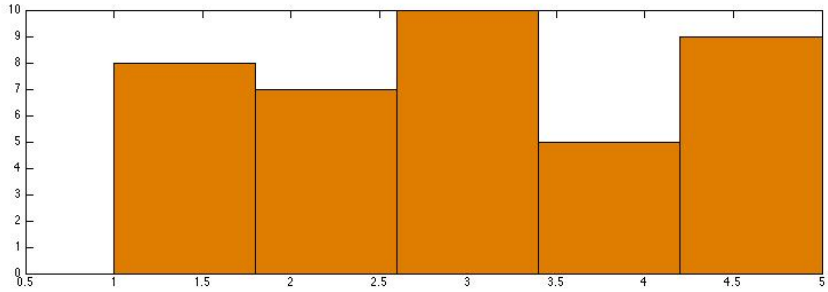
UDACITY

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.

■ **Example 2.1**



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.

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

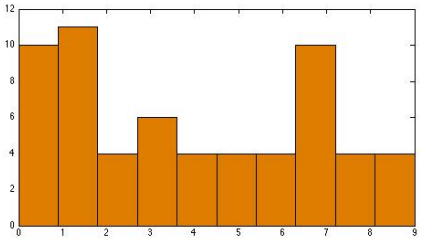


Figure 2.1: histogram of data set with bin size 1

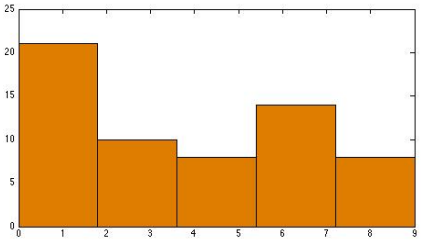


Figure 2.2: histogram of data set with bin size 2

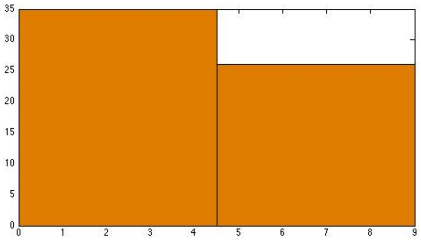


Figure 2.3: histogram of data set with bin size 5

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

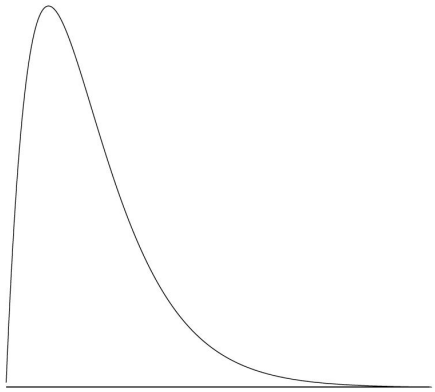


Figure 2.4: positive skew

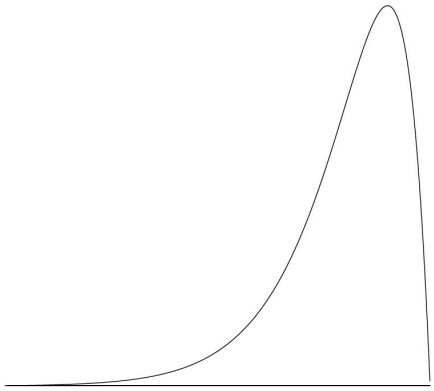


Figure 2.5: negative skew

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?

Mean, Median and Mode
Practice Problems

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.

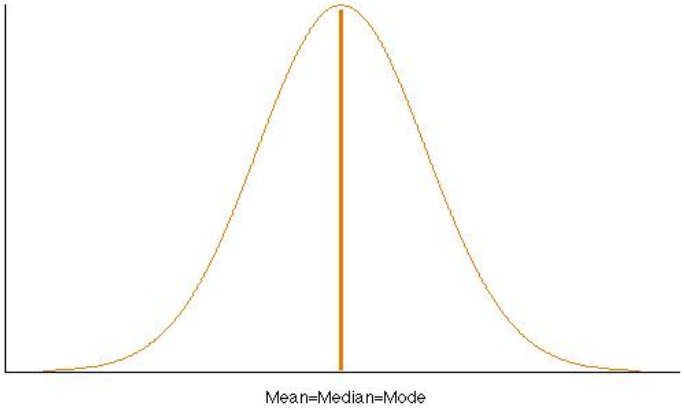
1. The data set is odd $n/2 =$ the position in the data set the middle value is
2. The data set is even $\frac{x_k + x_{k+1}}{n}$ gives the median for the two middle data points

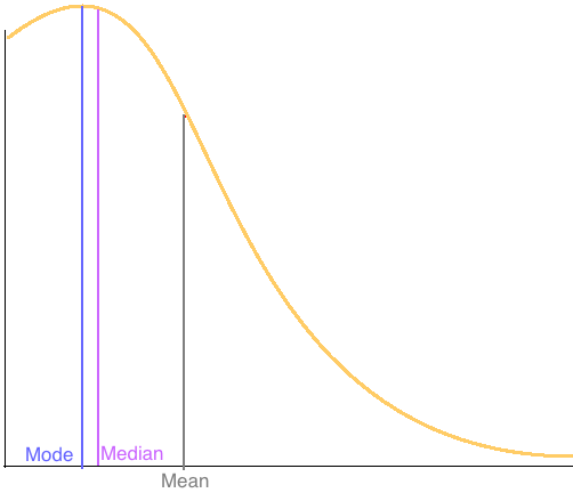
R The median is robust to outliers, therefore an outlier will not affect the value of the median.

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

Box Plots and the IQR
Finding outliers
Variance and Standard Deviation
Bessel's Correction
Practice Problems

UDACITY

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: $Q3 - Q1$

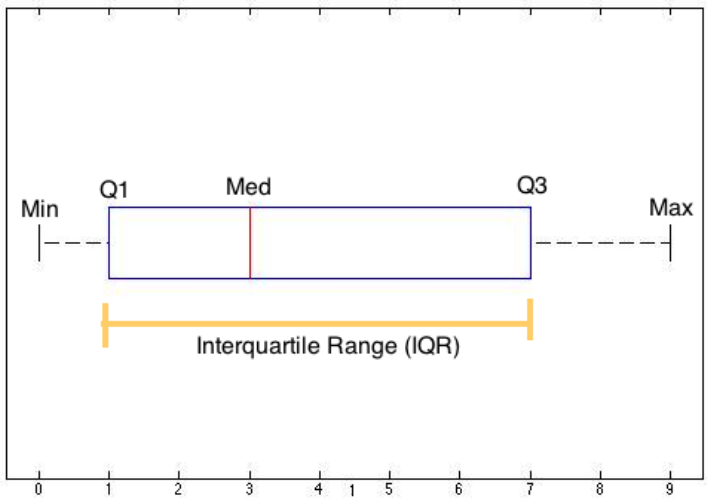


Figure 4.1: A simple boxplot

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?

Z score
Standard Normal Curve
Examples
Finding Standard Score
Practice Problems

UDACITY

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.

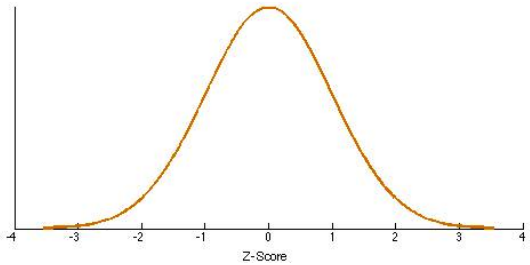


Figure 5.1: The Standard Normal Curve

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.

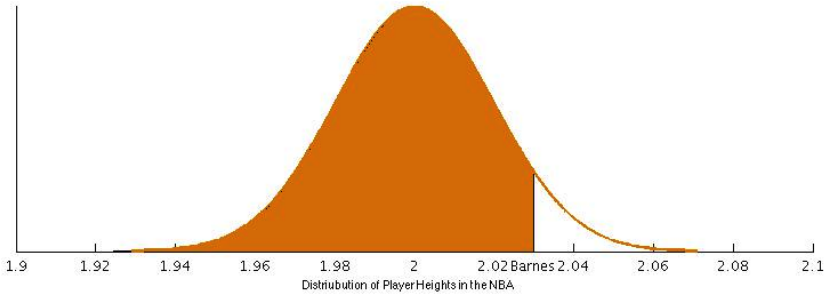


Figure 5.2: Notice we place the mean height 2.00 right in the middle and make tick marks that are each 1 standard deviation or 0.02 meters away in both directions.

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 $1.5\sigma + \mu$ =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 professionally 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?

UDACITY

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.

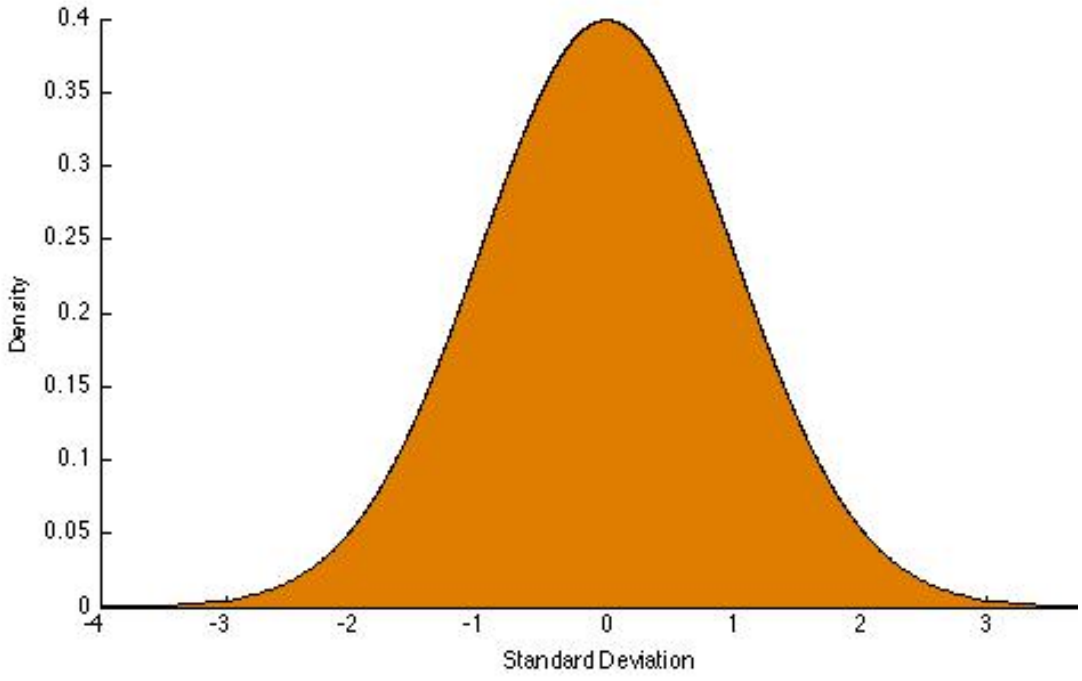


Figure 6.1: The area beneath the curve is 1

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:

$$z = \frac{x-\mu}{\sigma} = \frac{2.05-1.85}{0.15} = 1.\bar{3}$$

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.

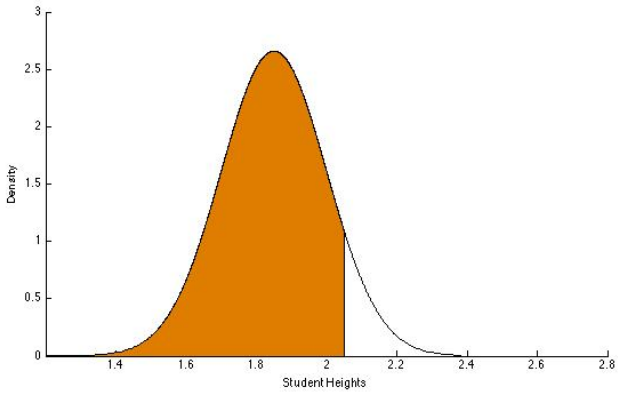


Figure 6.2: 85% is the shaded area

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 Margies heights.

We already know that 90.82% of students are shorter that Margie. So lets 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	0.906	0.9082	.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?

Central Limit Theorem
Practice Problems

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?