Chapter 1: Introduction to Statistics

Misleading Information:

- Surveys and advertising claims can be biased by unrepresentative samples, biased questions, inappropriate comparisons and errors in data.
- Graphs can be misleading through the use of broken scales, pictographs and errors.
- Be aware of the motives of who is presenting the data.

Statistics: The collection, analysis and interpretation of data

Population: The total collection of individuals or objects under consideration

Parameter: A number that describes a characteristic of a population

Sample: The portion of the population selected for study

Statistic: A number that describes a characteristic of a sample

Descriptive Statistics: The use of numerical and/or visual techniques to summarize data

Inferential Statistic: Draws a conclusion about the population from the sample

Representative Sample: A sample that has the pertinent characteristics of the population in the same proportion as the population

Chapter 2: Organizing and Presenting Data

Variable: Contains information or a property of an object, person or thing

Categorical Data: Data which falls into categories

Numerical Data: Data values (numbers) which are the result of measurements

(Numerical) Continuous Data: Data which can take on any value between two numbers

(Numerical) Discrete Data: Data which can take on only certain values

Distribution: A presentation of data along with the number of times each data value occurs.

Four Levels of Measurement

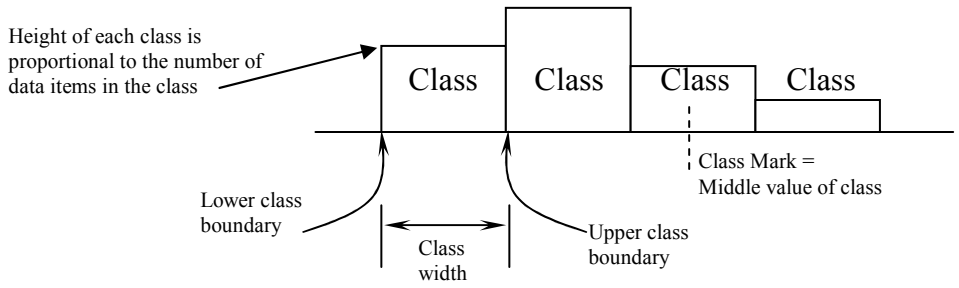
Nominal (Qualitative Data): Names or categories
Ordinal (Qualitative or Quantitative): A category which has an inherent order
Interval (Quantitative): A category in which only the value of the difference between the numbers has a meaning.
Ratio (Quantitative): A category in which both the interval and the ratio have meaning.

Stem and Leaf Plot: A display of data in which the rightmost digits of the data are initially ignored and the numbers represented by the remaining digits are listed down in ascending order. For example,

The values below …….would be written like this

199	19	99
199	20	012
200	21	0
201	22	
202	23	113
210		
231		
231		
233		

Histogram: A bar graph of data whose vertical coordinate shows how many items of data (frequency) fall into each of a series of ranges (classes). Bars must touch each other.



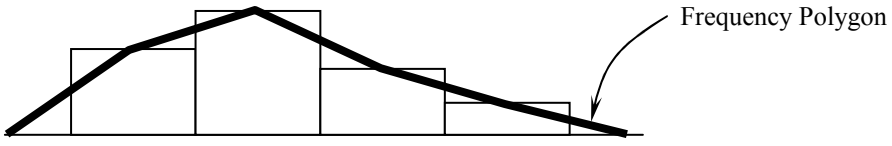
Construct a Histogram:

Determine the class width: $\text{Class Width} = \dfrac{\text{largest data value - smallest data value}}{\text{number of classes}}$

(If Class width turns out to be a whole number increase it by 1. Else, round class width up to next highest whole number.

$$\text{Relative Frequency} = \frac{\text{class frequency}}{\text{total number of data values}}$$

Frequency Polygon: a straight line graph connecting the midpoints of the tops of each class in a histogram. The end points of the graph fall to touch the x-axis at half the class width outside of the histogram. Usually either the histogram or the frequency polygon is drawn – but not both.

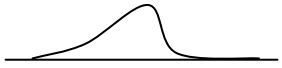


Shapes of Distributions:

Symmetrical

Skewed to Right

Skewed to Left



Chapter 3: Numerical Techniques for Describing Data

$$\mu = \textbf{Population Mean} = \frac{\textbf{Sum of all data values}}{\textbf{Number of Data values}} = \frac{\sum x}{N}$$

$$\bar{x} = \textbf{Sample Mean} = \frac{\textbf{Sum of all SAMPLE values}}{\textbf{Total number of all SAMPLE values}} = \frac{\sum x}{n}$$

Adding a constant to each member of a population or sample increases the mean by that value.

Multiplying each member of a population or sample by a constant multiplies the mean by that value.

Median: The middle value in a sorted list of data values. If the number of values is an odd number, the median is the middle value. If the number of values in even, the median is the average of the two middle value.

Mode: The most frequently occurring data value.

σ = Population Standard Deviation = $\sqrt{\frac{\sum (x-\mu)^2}{N}}$ = The square root of the average of the squared deviations from the mean of the population. (It is roughly the average deviation from the mean.)

s = Sample Standard Deviation = $\sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$ = The square root of the average of the squared deviations from the mean of the sample. (It is roughly the average deviation from the mean.)

The Empirical Rule States					
	68%	of data values fall within	1 σ	of	μ
	95%	of data values fall within	2 σ	of	μ
	99.7%	of data values fall within	3 σ	of	μ

The mean, median and mode are measures of central tendency of the data.
The standard deviation measures the extent to which data spreads about the mean.

x = raw score: the actual data value (e.g. dollars, feet, inches, degrees, etc.)
z = z-score: the distance of a data value from the mean in units of standard deviations

To compute one from the other, $z = \frac{x - \mu}{\sigma}$ $\qquad x = z\sigma + \mu$

PR(x) = Percentile Rank of x = Percentage of data values less than x.

$$PR = \frac{B + (1/2)E}{N} 100 \text{ (rounded to the nearest whole number)}$$

Where…

B is the number of data values less than x

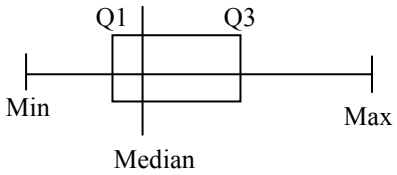
E is the number of data values equal to x (including x) and

N is the total number of data values

Percentile P(x) = Percentile of x = The data value that has x % of the data values below.
(Example: Joe's height is 72" which makes him taller than 89% of his class.
The Percentile Rank or 72" = PR(72) = 89 % and
the 89th Percentile = P(89%) = 72"

The minimum is the smallest data value
The Q1 point is the 25th Percentile, that is, 25% of the data values are below it.
The median is the 50th Percentile, that is, 50% of the data values are below it.
The Q3 point is the 75th Percentile, that is, 75% of the data values are below it.
The maximum is the largest data value

The box and whisker plot illustrates the above with a figure.



Interquartile Range = IQR = Q3 – Q1 = distance between Q3 and Q1

An outlier is any data value greater than
$3 \times$ IQR greater than Q3 or $3 \times$ IQR less than Q1

Chapter 4: Linear Regression and Correlation

Given: Two continuous variables (e.g. Weight of Car vs Miles/Gallon)

Are these variables linearly correlated?

Ho: No Linear Correlation between Car Weight and Miles/Gallon

H_a: Car Weight and Miles/Gallon are linearly correlated with

$$\rho > 0, \qquad \rho < 0, \qquad or \quad \rho \neq 0,$$

Perform a LinRegTTest which will calculate the correlation coefficient (**r**) as well as the values of **a** and **b** in the linear regression equation **y = a + bx.** The linear regression equation allows you to predict value of the dependent variable (**y**, Miles/Gallon) when you substitute the a value for the independent variable, (**x,** Weight of Car).

(You will need to enter your data values, Car Weights and associated Miles/Gallon, into L1 and L1).

The calculator will also return the p-value. If $p < \alpha$, reject Ho.
The calculator also returns $\mathbf{r^2}$, the Coefficient of Determination, which tells you proportion of the variance explained by the independent variable (essentially how well the regression equation predicts the value of the dependent variable from the independent variable.)

Chapter 5: Probability

Sample Space: The total number of possible outcomes of an experiment.

Definition: Classical (a priori) Probability :
If an experiment or a situation has a number (n) of possible outcomes, all equally likely, then the probability that any one outcome will occur is 1/n. (n = sample space). Then for equally likely events,

$$P(A) = \frac{number\ of\ ways\ event\ A\ can\ ocurr}{Total\ number\ of\ outcomes\ in\ the\ samples\ pace}$$

Definition: Relative Frequency (a posteriori) determination of probability
Perform an experiment many times. The probability of Event A occurring is

$$P(A) = \frac{number\ of\ \ times\ event\ A\ ocurred}{Total\ \ number\ of\ \ times\ experiment\ was\ repeated}$$

An outcome that has a probability of zero can never occur.
An outcome that has a probability of one will occur every time.

The probability is always a number between 0 and 1. ($0 \le p \le 1$)
If only two outcomes are possible in an experiment, A and B, then P(B) = 1 – P(A)

Addition Rule (for Mutually Exclusive Events): An event can be defined in an experiment as getting any one of a number of outcomes, for example, the event that "I win" can be defined as getting an even number when tossing a die (that is, getting the number 2, 4 or 6.) For a 6-sided fair die, this probability is P(even number) = P(2) + P(4) + P(6) = 1/6 + 1/6 + 1/6 = 3/6 = ½

Addition Rule (for Mutually Exclusive Events): If two events, A and B, are mutually exclusive, then the probability that event **A or B** will occur is the sum of their probabilities, P(A or B) = P(A) + P(B).

Multiplication Rule (for Independent Events): Two events are independent when the outcome of one has nothing to do with the outcome of another. For example, in tossing two dice, the outcome of the second die has nothing to do with the outcome of the first die. The event of getting two 6s is P(6&6) = P(6) x P(6) = 1/6 x 1/6 = 1/12.

Multiplication Rule (for Independent Events): If two events, A and B are independent, then the probability that both **A and B** will occur is the product of their probabilities, P(A & B) = P(A) x P(B)

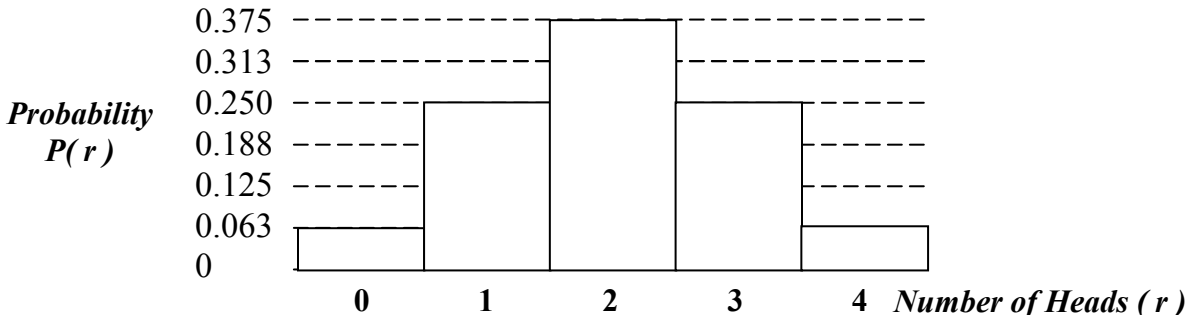
Chapter 6: Random Variables and Discrete Probability Distributions

If we toss a coin 4 times and we assume

1. that the coin is fair (P(Head) = ½) and
2. that each flip has no effect on any other flip (the flips are INDEPENDENT),

then the probabilities of getting 0, 1, 2, 3 or 4 heads is shown below as a

Discrete Probability Distribution..



The values of each of the probabilities above was calculated from the formula

$$P(r\ heads) = 4Cr(0.5)^r(1-0.5)^{4-r} \quad \ldots \text{ for r = 0, 1, 2, 3 and 4}$$

- -

In general, in a series of (n) independent trials,
where the outcome of each trial is considered either a success (S) or a failure (F),
and the probability of success in a single trial is p,
then the probability of having r successes in n trials is,

$$P(r\ Successes) = (nCr)\ (p)^r\ (1-p)^{n-r}$$

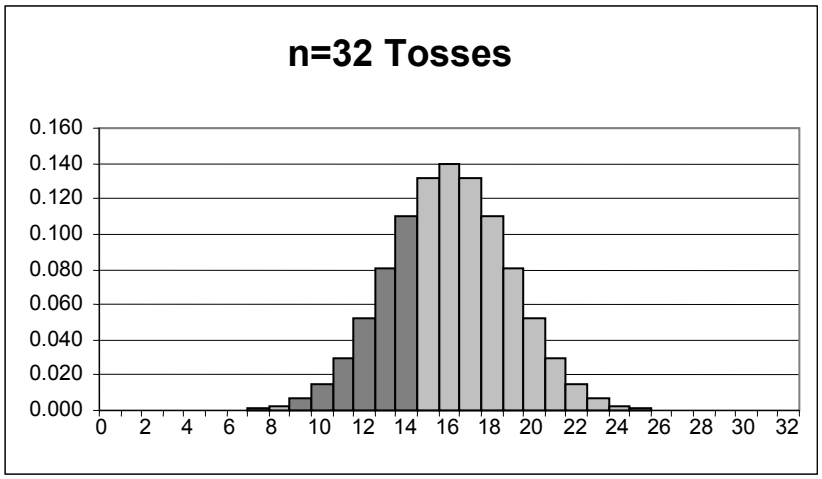
$$\text{where} \quad nCr = \frac{n!}{(r!)(n-r)!}$$

Example: Using the Addition Rule, the probability of having less than 2 Heads in 4 tosses would be calculated as P(0 Heads) + P(1 Head) = .063 + .250 = .313

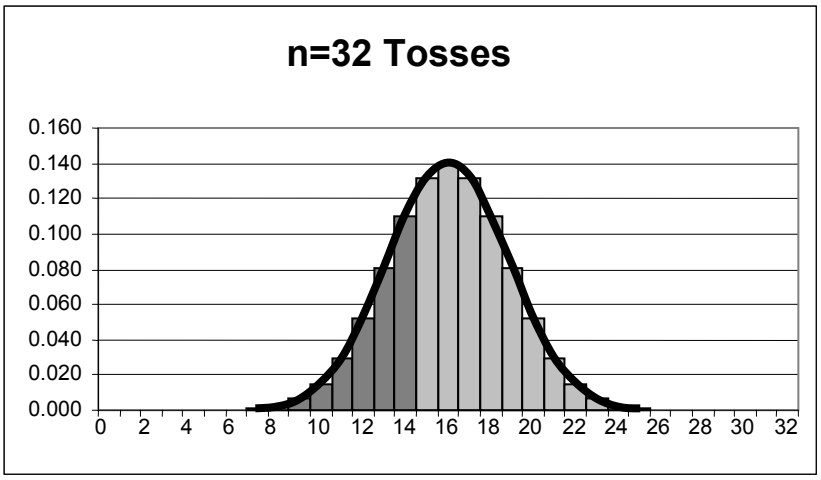
Chapter 7: Continuous Probability Distributions

When the number of tosses (trials) becomes large, say 32, and we wish to know the probability of getting say 14 or fewer Heads, it becomes cumbersome to add all the probabilities from 0 to 14, P(0) + P(1) + P(2) + … + P(14).

Note that the sum from 0 to 14 is actually the area under the curve from 0 to 14, (shown dark grey).



Therefore, we approximate the above discrete probability distribution with the normal curve and calculate the area under the normal curve, which is much easier.



The mean μ and standard deviation σ of the normal approximation to a binomial probability distribution is $\mu = np$ and $\sigma = \sqrt{np(1-p)}$. However, this approximation may only be used if np > 5 and n(1-p) > 5.

To find the area under the Normal Distribution between any two values, we enter the z-scores of those values into the calculator. In the above problem to find the probability of getting between 0 and 14 Heads in 32 tosses of a fair coin, we calculate

$\mu = np = 32 \times 0.5 = 16$ and $\sigma = \sqrt{np(1-p)} = \sqrt{32 \times 0.5 \times (1-0.5)} = 2.828$.

We need the z-scores of 0 and 14, but to include all the area of the binomial distribution between those two values, we need to use and extra 0.5 outside that range. That is, we use the range from -0.5 to 14.5.

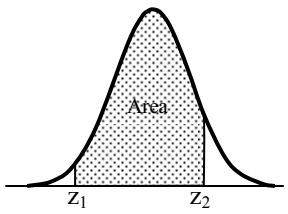
The z-scores are $z_{left} = \dfrac{-0.5-16}{2.828} = -5.834$ and $z_{right} = \dfrac{14.5-16}{2.828} = -0.530$

Then we use normalcdf(-5.834, -0.530) = 0.298 which is the probability that we will get between 0 Heads and 14 Heads when we toss a fair coin 32 times.

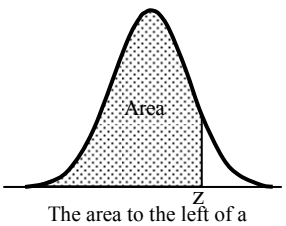
The total area under the normal curve is 1 (one).

Areas are always expressed as decimals.

If the area to the left of a z-score is known, then invnormal(area) gives the z-score at the <u>right</u> side of the area.

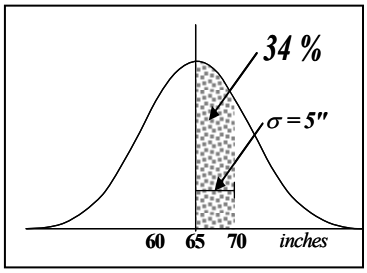


The area between 2 z-scores is normalcdf (z_1, z_2)



The area to the left of a z-score is normalcdf (-E99, z)

Many kinds of data can be described with a normal distribution, for example, the heights of students, the IQ's of students, the lifetimes of tires.



The curve at right shows that the heights of students are normally distributed around a mean of 65 inches with a standard deviation of 5 inches. This means that 34% (or a proportion of 0.34) of the students have heights between 65 and 70 inches. It also means that the probability is 0.34 that a student, chosen at random from a population of students, will be between 65 and 70 inches tall.

Chapter 8: The Sampling Distribution of the Mean

Notation:

Population	**Sample**
μ is the mean of a population	$\bar{x}$ is the mean of a sample
σ is the standard deviation of the population	**s** is the standard deviation of the sample
N is the number of items (people or things) in the population	**n** is the number of items (people or things) in the sample

If we take every possible sample of size n from a population, and calculate the mean of each of those samples, the mean of all those sample means, $\mu_{\bar{x}}$, will be equal to the mean of the population, μ.

The standard deviation of the distribution of all those sample means is called the **standard error of the mean** and is equal to the population standard deviation divided by the square root of the sample size, that is, $\frac{\sigma}{\sqrt{n}}$.

In mathematical notation $\mu_{\bar{X}} = \mu$ and $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$.

$\sigma_{\bar{X}}$ is called the standard error of the mean.

If the population is normally distributed, the sample means will be normally distributed.

The **Central Limit Theorem** states that regardless of the shape of the population distribution, the sampling distribution of the mean approaches a Normal Distribution as the sample size n becomes large. (Generally $n > 30$)

Proportions

Proportion = Fraction of the population with a certain characteristic.

$p = X / N$ = the population proportion
X = number of occurrences of those characteristic members in the population
N = population size

$\hat{p} = x / n$ = the sample proportion
x = number of occurrences of those characteristic members in the sample
n = sample size

The sampling distribution of the proportion can be approximated by a normal distribution if

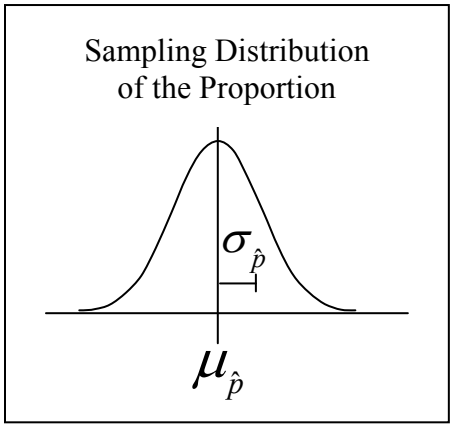
$$np > 5 \quad and \quad n(1-p) > 5$$

The mean of that distribution equals the population proportion.

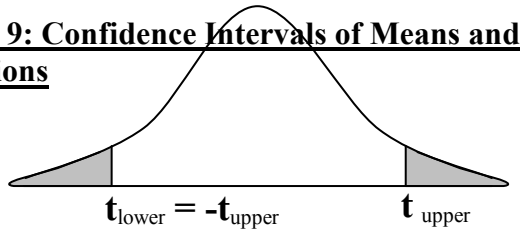
$$\mu_{\hat{p}} = p$$

The standard deviation (standard error of the proportion) of that distribution is

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



Chapter 9: Confidence Intervals of Means and Proportions



Confidence Intervals			
99%	98%	95%	90%

df =(n-1)	t(99.5%)	t(99%)	t(97.5%)	t(95%)
1	63.657	31.821	12.706	6.314
2	9.925	6.965	4.303	2.920
3	5.841	4.541	3.182	2.353
4	4.604	3.747	2.776	2.132
5	4.032	3.365	2.571	2.015
6	3.707	3.143	2.447	1.943
7	3.499	2.998	2.365	1.895
8	3.355	2.896	2.306	1.860
9	3.250	2.821	2.262	1.833
10	3.169	2.764	2.228	1.812
11	3.106	2.718	2.201	1.796
12	3.055	2.681	2.179	1.782
13	3.012	2.650	2.160	1.771
14	2.977	2.624	2.145	1.761
15	2.947	2.602	2.131	1.753
16	2.921	2.583	2.120	1.746
17	2.898	2.567	2.110	1.740
18	2.878	2.552	2.101	1.734
19	2.861	2.539	2.093	1.729
20	2.845	2.528	2.086	1.725
21	2.831	2.518	2.080	1.721
22	2.819	2.508	2.074	1.717
23	2.807	2.500	2.069	1.714
24	2.797	2.492	2.064	1.711
25	2.787	2.485	2.060	1.708
26	2.779	2.479	2.056	1.706
27	2.771	2.473	2.052	1.703
28	2.763	2.467	2.048	1.701
29	2.756	2.462	2.045	1.699
30	2.750	2.457	2.042	1.697
35	2.724	2.438	2.030	1.690
40	2.704	2.423	2.021	1.684
45	2.690	2.412	2.014	1.679
50	2.678	2.403	2.009	1.676
60	2.660	2.390	2.000	1.671
80	2.639	2.374	1.990	1.664
100	2.626	2.364	1.984	1.660
200	2.601	2.345	1.972	1.653
500	2.586	2.334	1.965	1.648
Normal	2.576	2.326	1.960	1.645

Methods of Estimation

Estimate of the Population Mean (μ)

σ known → **ZInterval (stats)**

σ : pop std dev
n : sample size
$\bar{x}$: sample mean
C-level: Confidence level

Confidence Interval

$$\bar{x} - z_c \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_c \frac{\sigma}{\sqrt{n}}$$

$$\text{Margin of Error} = E = z_c \frac{\sigma}{\sqrt{n}}$$

Estimate of the Population Mean (μ)

σ unknown → **TInterval (stats)**

s : sample std dev
n : sample size
$\bar{x}$: sample mean
C-level: Confidence level

Confidence Interval

$$\bar{x} - t_c \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_c \frac{s}{\sqrt{n}}$$

$$\text{Margin of Error} = E = t_c \frac{s}{\sqrt{n}}$$

Estimate of the Population Proportion (p)

$n\hat{p} > 5$ and $n(1-\hat{p}) > 5$ → **1-PropZInt**

x : no. of cases in sample
n : sample size
$\hat{p} = x/n$
C-level: Confidence level

Confidence Interval

$$\hat{p} - z_c(s_{\hat{p}}) < p < \hat{p} + z_c(s_{\hat{p}})$$

$$\text{...where } s_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\text{Margin of Error} = E = z_c(s_{\hat{p}})$$

Chapter 10: Introduction to Hypothesis Testing

Null Hypothesis: Ho	Alternative Hypothesis Ha	Hypothesis Type
Always try to reject this statement …	… so that you can accept this statement as true.	
Joe's tires last about 50,000 miles	Joe's tires last less than 50,000 miles	Directional
The average college student works 20 hrs/week	The average college student works > 20 hrs/week	Directional
The average Whoopie bar contains 180 calories	The average Whoopie bar does not contain 180 calories	Non-directional
The unemployment rate is 5.5%	The unemployment rate less than 5.5%	Directional

Hypothesis Testing Procedure

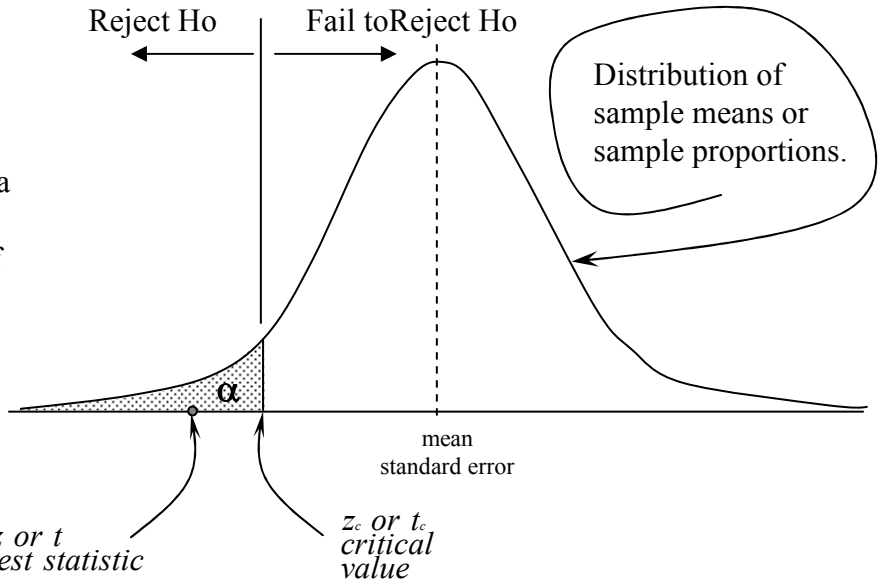
1. Formulate both hypotheses
2. Determine the model to test the null hypothesis
3. Formulate the decision rule
4. Analyze the sample data.
5. State the conclusion

Type I Error:	The error you make when you incorrectly reject the null hypothesis.
Level of Significance, α,	The probability of making a type I error that you are willing to accept when you test.
p-value, p,	Based on the test statistic, the p-value is the probability that you will reject the null hypothesis even though it is true, thus making a type I error.

Diagram Descriptors:

Using the example of a one-tail test with a Normal distribution

1. Normal or t-distribution curve
2. Mean and standard error of the mean z_c, critical value, (the z or t-score of the data value beyond which you will reject Ho).
3. α, significance (shaded area to the left of zc)
4. z, test statistic, (the z or t-score of the experimental result).
5. p-value, area under the curve beyond the test statistic, z.



Chapter 11: Hypothesis Testing Involving One Population

Formulate the null and alternative hypotheses.

Look for **directional words** (e.g. greater, more, less)to determine whether you need a **1TT** or a **2TT**?

Determine important information:

If the problem gives the population standard deviation, σ, the distribution of sample means is **normal.**
If the problem gives the sample standard deviation, **s**, the distribution of sample means is **t-distributed.**

Use α to determine the critical value (z_c *or* t_c) for rejection of Ho.

Use the data to determine the test statistic, (z *or* t) If the test statistic is further from the mean than the critical value, REJECT Ho	**…or…**	Use the test statistic to determine the p-value. If the p-value is less than α, REJECT Ho

or …Using the Calculator:

Hypothesis test involving a mean

For a **normal** distribution, perform a **Z-Test** to determine the p-value. If $p < \alpha$, reject H_o.

For a **t-distribution** perform a **T-Test** to determine the p-value. If $p < \alpha$, reject H_o.

Hypothesis test involving a proportion

First check that np>5 and n(1-p)>5

Perform a **1-PropZTest** to determine the p-value. If $p < \alpha$, reject H_o.

Chapter 13: Hypothesis Test Involving Two Population Means

Given: Two populations and a sample from each. $\bar{x}_1$ **and** $\bar{x}_2$

Mean of Sample_1 = $\bar{x}_1$ **Sample Std Dev of Sample_1 =** S_1

Mean of Sample_2 = $\bar{x}_2$ **Sample Std Dev of Sample_2 =** S_2

Test whether the means of the two populations are different, that is,
(Test whether we can reject the null hypothesis at an α)

Ho: $\mu_1 - \mu_2 = 0$

H_a: $\mu_1 - \mu_2 > 0$ **or** **H_a:** $\mu_1 - \mu_2 < 0$

Perform a 2-SampTTest and calculate the p-value. If $p < \alpha$, reject H_o.

Chapter 14: Chi-Square

Chi-Square determines whether there is a dependence between two categorical variables.

Given: Two variables of categorical data (e.g. ProLife/ProChoice vs. Gender)

Are these variables independent?

(If you reject Ho you can conclude that the variables are dependent, but you can never prove them independent.)

	Pro Life	**Pro Choice**
Male	**A**	**B**
Female	**C**	**D**

…where A,B,C and D are the numbers of people in each category.

Ho: Gender and Choice are independent

H_a: Gender and Choice are NOT independent (i.e. they are dependent)

Perform a $\chi 2$-Test and calculate the p-value. If $p < \alpha$, reject H_o.
(You will need to set up Matrix A and enter your observed data before doing the test.). .
