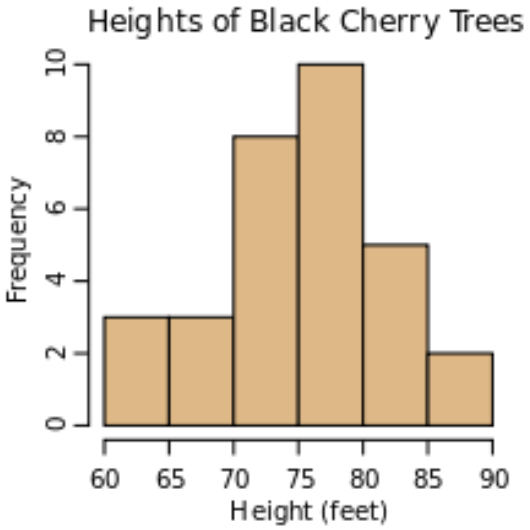
Final Lecture

Histograms and Frequency Tables

Histograms show the frequency of an event occurring. They can be used for plotting the frequency of a quantitative variable when the quantitative variable is separated into ranges, or bins. Here we have the height of black cherry trees separated into the ranges: 60-65, 65-70, 70-75, 75-80, 80-85, 85-90. We can refer to these categories as trees that are 60 feet, 65 feet, 70 feet, 75 feet, 80 feet, and 85 feet.



Taken from Wikipedia

To find the mean black cherry tree height we need to use the Excel function SUMPRODUCT. What this function is used for is to find the sum of the product of frequency times height for each group.

Height (feet)	Frequency
60	3
65	3
70	8
75	10
80	5
85	2

So, the mean black cherry tree height is:
(60*3 +65*3 + 70*8 + 75*10 + 80*5 + 85*2) / (3+3+8+10+5+2) = 72.7419

Scatterplots

Question 1

A survey was conducted to study the relationship between the annual income of a family and the amount of money the family spends on entertainment. Data were collected from a random sample of 280 families from a certain metropolitan area. ✔ Points: *10 out of 10*

Which of the following would be a meaningful display of the data from this study?

- **A.** Side-by-side boxplots
- **B.** A two-way table
- **C.** A pie chart
- **D.** A scatterplot
- **E.** A histogram

Feedback

Correct. A scatterplot is the appropriate display of quantitative relationship (in other words, to display the relationship between a quantitative explanatory variable and a quantitative response variable).

Boxplots

Question 2

In order to study whether there is a relationship between gender and age at marriage, 50 couples were randomly selected and the age of the bride and groom were recorded. ✔ Points: *10 out of 10*

Which of the following would be a meaningful display of the data from this study?

- **A.** A scatterplot
- **B.** A pie chart
- **C.** A two-way table
- **D.** Side-by-side boxplots
- **E.** A histogram

Feedback

Correct. Side-by-side boxplots are the appropriate display for comparing several groups of quantitative data (in other words, for displaying the relationship between a categorical explanatory variable in this case “gender," and a quantitative response variable, in this case "age at marriage.")

Outliers and Interquartile Range (IQR)

Outliers are any data points that are smaller than Q1 - 1.5*IQR

And outliers are any data points that are larger than Q3 + 1.5*IQR

Find Q1 using the Excel function QUARTILE.INC(selected array of data, 1)

Find Q3 using the Excel function QUARTILE.INC(selected array of data, 3)

IQR = Q3-Q1

Two-Way Tables

A two-way table can be used to compare two categorical variables.

Which game console do you prefer?

	Males	Females
Xbox	60 %	55 %
PlayStation	40 %	45 %
	100 %	100 %

Question 4

Which of the tables is the appropriate table of conditional percentages to discover if the region where one lives affects whether or not one has health insurance?

Points: *10 out of 10*

table A

Region	Uninsured	Insured	Total
Northeast	12.6%	87.4%	100%
Midwest	12.0%	88.0%	100%
South	18.2%	81.8%	100%
west	17.4%	82.6%	100%

table B

Region	Uninsured	Insured	Total
Northeast	2.3%	16.2%	18.5%
Midwest	2.7%	19.6%	22.3%
South	6.6%	29.5%	36.1%
West	4.0%	19.1%	23.1%
Total	15.6%	84.4%	100%

table C

Region	Uninsured	Insured
Northeast	15.0%	19.2%
Midwest	17.1%	23.3%
South	42.1%	35.0%
West	25.8%	22.6%
Total	100%	100%

A. (•) Table A
B. ( ) Table B
C. ( ) Table C

Feedback

Correct. This table reports the conditional percentages for each value of the explanatory variable

Explanatory and Response Variables

The explanatory variable is the variable that explains why some response can occur.

If we are interested in the best amount of rain for crop growth, then the amount of rain may explain how much crops grow. So rain is the explanatory variable and crop growth is the response variable.

If we are interested in the ability of eating enough vitamin D to prevent catching a cold, then the amount of vitamin D we eat might explain whether we catch a cold as a response. So, vitamin D is the explanatory variable and catching a cold is the response.

In conditional percentage tables the one above in question 4, the conditional percentages are across from the explanatory variable. Here, they are interested in whether where someone lives has an effect on whether or not they have health insurance. They believe that the location where someone lives (AKA region) might explain whether someone has health insurance or not. So, region is the explanatory variable and health insurance is the response variable. This is why the correct answer is A, because the conditional percentages (that add up to 100%) are across from region, which is the explanatory variable.

Question 55

High blood pressure is unhealthy. Here are the results of one of the studies that link high blood pressure to death from cardiovascular disease. The researchers classified a group of white males aged 35 to 64 as having low blood pressure or high blood pressure, then followed the subjects for 5 years. The following two-way table gives the results of the study:

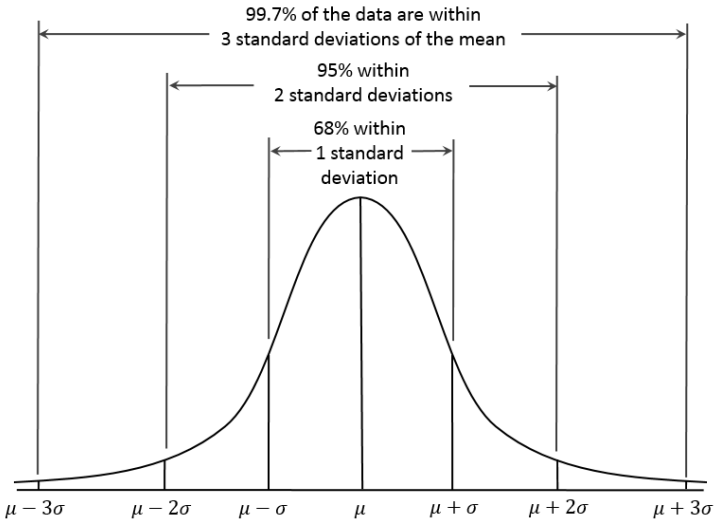
✔ Points: *10 out of 10*

	Blood pressure		
Cardiovascular death?	Low	High	Total
Yes	21	55	76
No	2655	3283	5938
Total	2676	3338	6014

In this example, which of the following would be appropriate to calculate?

A. Conditional row percentages
B. Conditional column percentages
C. The correlation coefficient r
D. The five-number summary of both variables

Standard Deviation Rule



The distribution of income is approximately
normal in shape with a mean of 60 and a standard deviation of 10.
According to the standard deviation rule, 95%
% of people have an IQ between 40 and 80.

Mean	60
Standard Deviation	10

Between 40 and 80

How many standard deviations is 40 from the mean? What about 80?

Mean + 1 St Dev	70
Mean - 1 St Dev	50
Mean + 2 St Dev	80
Mean - 2 St Dev	40

This means that 80 is 2 standard deviations away from the mean
Also, 40 is 2 standard deviations away from the mean

By the 68-95-99.7 rule, 95% of the data fall within 2 standard deviations away from the mean

p is for Probability

$p(x) = 0$: event will never occur

$p(x) = .5$: there's a 50/50 chance that the event will occur

$p(x) = 1$: event will definitely occur

The Law of Large Numbers

:as the number of trials increases, the empirical probability gets closer and closer to the theoretical probability.

empirical probability: the one you find in your sample and is defined by the relative frequency

theoretical probability: the true probability that you are trying to infer about the population

Sample Space

The space, which is the list of all possible outcomes. It includes all possible combinations.

The following is the sample space for choosing two committee members out of three women and two men

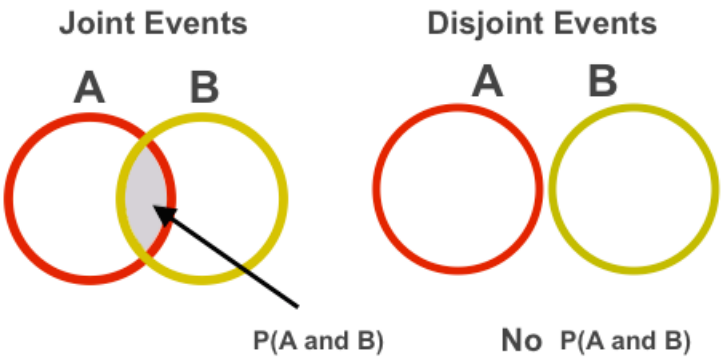
W1W2 W2W3 W1W3 W1M1 W2M1 W3M1 W1M2 W2M2 W3M2 M1M2

Disjoint Events

Two events are disjoint if they are mutually exclusive. I cannot be both tall and short at the same time, so if we are looking at height in two different groups, then those two groups are disjoint.

Venn diagrams for disjoint events have no overlapping region in the center.

P(A and B) = 0 for Disjoint Events



Independent Events

If one event does not depend on the other, then the events are independent.

If we are taking a random sample of individuals, then an event for each individual is independent of the event for another individual.

Lets say we are pulling marbles out of a bag. If we put each marble we take out of the bag back in before pulling one out again then by replacing the marble we have conducted our test "with replacement" and so we have independent events.

P(A and B) = P(A)*P(B) for Independent Events

For any event

$$\boldsymbol{P(A\ or\ B) = P(A) + P(B) - P(A\ and\ B)}$$

For **disjoint events** we know that P(A and B)=0 so,
$P(A\ or\ B) = P(A) + P(B) - P(A\ and\ B)$
$P(A\ or\ B) = P(A) + P(B) - 0$
$P(A\ or\ B) = P(A) + P(B)$

For **independent events** we know that P(A and B)=P(A)*P(B) so,
$P(A\ or\ B) = P(A) + P(B) - P(A\ and\ B)$
$P(A\ or\ B) = P(A) + P(B) - P(A) * P(B)$

$$P(\textbf{neither } A \textbf{ nor } B) = 1 - P(A \textbf{ or } B)$$

When they ask you about events that happen **at least once** it is often useful to use

$$P(A) + P(not\ A) = 1$$

Then,

$$P(at\ least\ once) = 1 - P(never)^{number\ of\ trials}$$

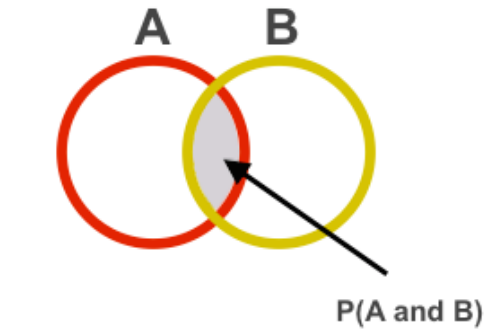
Lets say a car dealership makes cars that sometimes break. The probability of a new car breaking is .05. If they release 20 cars, what is the probability that at least 1 breaks?

$$P(at\ least\ 1\ car\ breaks) = 1 - P(no\ car\ breaks)^{number\ of\ trials}$$

$$P(at\ least\ 1\ car\ breaks) = 1 - (1 - .05)^{20}$$

Conditional Probability

Conditional probability considers the probability of one event, given another event.



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

Bayes Rule

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

If we do not know whether some events A and B are independent, then we cannot conclude that

$$P(A\ and\ B) = P(A) * P(B) \qquad only\ for\ independent\ events$$

$$P(A\ and\ B) = P(B|A) * P(A) \qquad otherwise$$

Four ways to test independence is to ask whether

1) $P(A|B) = P(A)$
2) $P(A|B) = P(A|not\ B)$
3) $P(B|A) = P(B)$
4) $P(A\ and\ B) = P(A) * P(B)$

Lets say we are picking marbles out of a bag. There are 5 red marbles, 3 blue marbles, and 2 yellow marbles.

What is the probability that we pick out one red and then one blue marble?

If the marble is replaced into the bag, then the events are independent. Then,

$$P(red\ and\ blue) = P(red) * P(blue) = \frac{5}{10} * \frac{3}{10}$$

If the marble is not replaced into the bag, then the events are not independent. Then,

$$P(red\ and\ blue) = P(red) * P(blue|red) = \frac{5}{10} * \frac{3}{9}$$

We know that we picked the red marble out the first time. So, we now have only 9 marbles left in the bag.

Random Variable

A random variable assigns a unique numerical value to the outcome of a random experiment.

Discrete variables can be found by answering the question, – how many?
Continuous variables can be found by answering the question, – how much?

The mean of a discrete random variable, or the expected value is

$$\mu_x = x_1p_1 + x_2p_2 + \cdots + x_np_n = \sum_{i=1}^{n} x_ip_i$$

The variance of a discrete random variable is

$$\sigma_x^2 = (x_1 - \mu_x)^2p_1 + (x_2 - \mu_x)^2p_2 + \cdots + (x_n - \mu_x)^2p_n = \sum_{i=1}^{n} (x_i - \mu_X)^2p_i$$

The standard deviation of a discrete random variable is

$$\sigma_x = \sqrt{\sigma_x^2}$$

Binomial Random Variable

A binomial random variable comes from a binomial experiment. A binomial experiment has a fixed number of trials, two possible outcomes: either success or failure, a fixed success rate in each trial, and all trials in a binomial experiment are independent from one another.

Lets say you are flipping a coin 3 times and are calling getting tails a success.

$$The\ number\ of\ possible\ outcomes\ with\ k\ successes\ out\ of\ n\ trials = \frac{n!}{k!(n-k)!}$$

$$\frac{n!}{k!\,(n-k)!} = \frac{3!}{2!\,(3-2)!} = \frac{3*2*1}{2*1\,(1)} = \frac{6}{2} = 3$$

The three possible outcomes: TTH HTT THT

$$The\ probability\ of\ getting\ k\ successes\ out\ of\ n\ trials\ with\ probability\ p\ for\ success$$

$$P(X=k) = \frac{n!}{k!\,(n-k)!}p^k(1-p)^{(n-k)} \qquad k = 0\,,1\,,\ldots.n$$

Here, p is the probability of success, n is the total number of trials, k is the number of successes

In Excel =FACT(n)/((FACT(k)*FACT(n-k)))*(p^k)*(1-p)^(n-k)
=BINOM.DIST(k,n,p,0) for exactly k
=BINOM.DIST(k,n,p,1) for less than
=1-BINOM.DIST(k,n,p,1) for greater than

The mean of a binomial random variable, or the expected value is

$$\mu_x = np$$

The variance of a binomial random variable is

$$\sigma_x^2 = np\,(1-p)$$

The standard deviation of a binomial random variable is

$$\sigma_x = \sqrt{\sigma_x^2}$$

Normal Random Variable

$$z - score = \frac{value - mean}{standard\ deviation}$$

$$z - score = \frac{X - \mu_x}{\sigma_x}$$

The z-score is the number of standard deviations that a value is away from the mean.

For example, lets say that the average number of hours a student spends studying outside of class is 4 per day with a standard deviation of 2. Lets say someone studies for 5 hours. The z-score for this value is,

$$z - score = \frac{X - \mu_x}{\sigma_x} = \frac{5 - 4}{2} = \frac{1}{2}$$

This means that 5 hours of studying is .5 standard deviations above the mean.

The probability of having a value that is some number of standard deviations above or below the mean is P(z-score). We can find it using Excel.

$P(z < Z1)$ In Excel: **NORM.S.DIST(Z1,1)**

$P(z > Z2)$ In Excel: **NORM.S.DIST(-Z2,1)**

Given the mean and the standard deviation, we can use Excel to find the probability of having a value of our normal random variable, X.

$P(X > X1)$ In Excel: **1- NORM.DIST(X1, μ, σ, 1)**

$P(X < X2)$ In Excel: **NORM.DIST(X2, μ, σ, 1)**

The z-score with an area of the probability density curve below (or above) A can be found using Excel.

$z - score(below\ A)$ In Excel: **NORM.S.INV(A)**

$z - score(above\ A)$ In Excel: **-NORM.S.INV(A)**

The value of a normal random variable X that is found with a given probability can be determined using Excel.

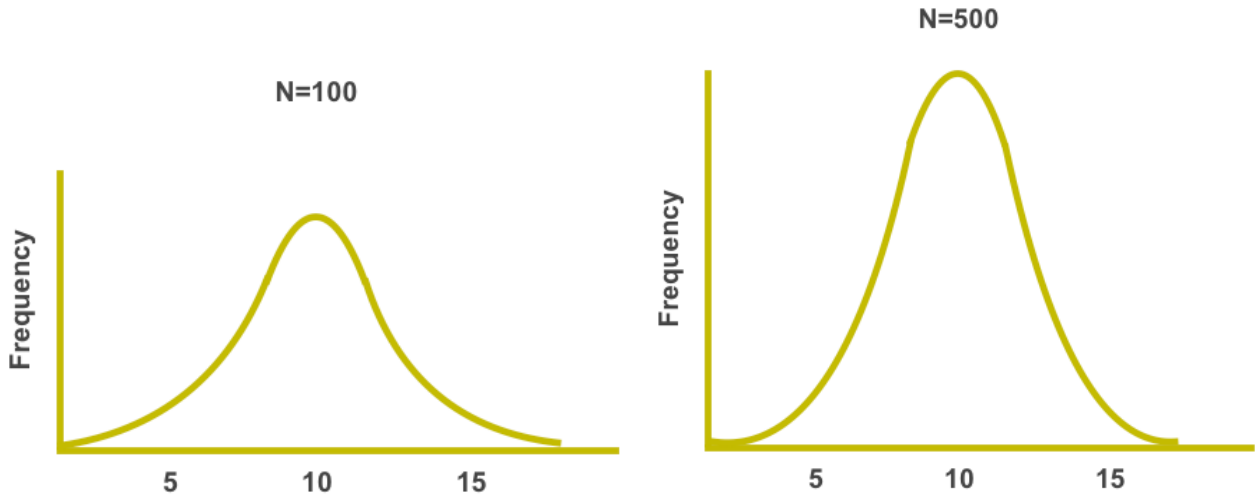
$X\ is\ less\ than\ what\ with\ probability\ P, \mu, \sigma$ In Excel: **NORM.INV(P, μ, σ)**

Statistics and Parameters

A statistic is a value that describes the sample, while a parameter describes the population.

The population mean is denoted μ and the population standard deviation is denoted σ.
The sample mean is denoted $\bar{x}$ and the sample standard deviation is denoted s.

The population proportion is p, while the sample proportion is $\hat{p}$.



In both cases the center is at 10. This is 10 heads out of 20 coin flips.

The spread is larger for N=100 than for N=500. (larger spread means larger standard deviation)

Both of these distributions look approximately normal in shape.

The standard deviation of a sample proportion is,

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

The mean of a sample proportion is,

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

The z-score for a proportion,

$$z - score = \frac{X - \mu_x}{\sigma_x} = \frac{X - \mu_{\hat{p}}}{\sigma_{\hat{p}}}$$

Sample Means

If many samples are taken, then each will have its own mean $\bar{x}$. The mean of the sample means is the population mean, μ. The standard deviation of the sample means is the standard error,

$$standard\ deviation(\bar{x}) = standard\ error \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

For sample proportions the distribution is approximately normal if

$$np \geq 10 \qquad and \qquad n(1-p) \geq 10$$

For sample means the distribution is always approximately normal if

$$n > 30$$

The central limit theorem: the distribution of a large sample size is approximately normal, regardless of the population distribution.
(Large means greater than 30)

Sample and Population Proportions

Using a confidence level to describe the probability that a random sample estimates the population proportion.

We are able to use a sample proportion to infer about a population proportion if:

There are at least 10 successes and 10 failures

and $np \geq 10$ and $n(1-p) \geq 10$

What do we do when we don't know the population standard deviation, σ ?

We estimate it with the sample standard deviation, s!

standard error of a sample mean when we know σ $\frac{\sigma}{\sqrt{n}}$

standard error of a sample mean when we don't know σ $\frac{s}{\sqrt{n}}$

In this case we use a t-test instead of a z-score. These are very similar.

when we know σ $z = \frac{\bar{x} - \mu_{\bar{x}}}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}}$

when we don't know σ $t = \frac{\bar{x} - \mu_{\bar{x}}}{\frac{s}{\sqrt{n}}} = \frac{\bar{x} - \mu_{\bar{x}}}{s_{\bar{x}}}$

The number of degrees of freedom for the t-test is called df.

$$df = n - 1$$

The standard error of the sampling distribution of sample proportions:

standard error of a sample proportion $\sqrt{\frac{p(1-p)}{n}}$

We are 95% confident that the following interval contains the population proportion:

95% confidence interval $p \pm 1.96\sqrt{\frac{p(1-p)}{n}}$

The standard error of the sampling distribution of sample means:

standard error of a sample mean $\frac{\sigma}{\sqrt{n}}$

We are 95% confident that the following interval contains the population mean:

95% confidence interval $\bar{x} \pm 1.96\ \frac{\sigma}{\sqrt{n}}$

The standard error of the sampling distribution of sample means when σ is unknown:

standard error of a sample mean when we don't know σ $\frac{s}{\sqrt{n}}$

We are 95% confident that the following interval contains the population mean:

95% confidence interval $\bar{x} \pm t_c \frac{s}{\sqrt{n}}$

Here, $(t_c\ \frac{s}{\sqrt{n}})$ is the margin of error

Hypothesis Testing

Hypothesis testing is testing a falsifiable supposition, or claim.

	We Reject H_0. (accept H_a)	We Fail to Reject H_0 (not enough evidence to accept H_a)
H_0 is true.	Type I Error	Correct Decision
H_0 is false. (H_a is true)	Correct Decision	Type II Error

Hypothesis Test for a Population Proportion

A hypothesis test for a population proportion tests the hypothesis that some population proportion is equal to a particular value.

The **normal distribution** is an appropriate model for a sampling distribution of a sample proportion if there are at least 10 expected successes and 10 expected failures.

$$np \geq 10 \text{ and } n(1-p) \geq 10$$

Given the population proportion, p_o and the sample proportion, $\hat{p}$:

The standard deviation of a sample proportion is,

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

The z-score for a proportion,

$$z-score = \frac{\hat{p}-p_o}{\sigma_{\hat{p}}}$$

H_o: p = .5

H_a: p > .5

We want the area above the z-score because our alternative hypothesis, H_a is p>.5. So, we use "1-NORM.S.DIST."

$$p-value = 1 - NORM.S.DIST(z-score, 1) = \ 1 - NORM.S.DIST(6.3246,1) = 1.2695E-10$$

Hypothesis Test for a Population Mean

A hypothesis test for a population proportion tests the hypothesis that some population mean is equal to a particular value.

The **normal distribution** is an appropriate model for a sampling distribution of a sample mean if the population is normal. If we do not know if the population is normal, the **normal distribution** is still an appropriate model for a sampling distribution of a sample mean if the sample size is large enough,

$$n > 30$$

Given the population mean, μ_o and the sample mean, $\hat{\mu}$:

The standard deviation of a sample mean is,

If we know the population standard deviation σ

$$\sigma_{\hat{\mu}} = \frac{\sigma}{\sqrt{n}}$$

If we only know the sample standard deviation s

$$\sigma_{\hat{\mu}} = \frac{s}{\sqrt{n}}$$

The z-score for a sample mean,

$$z - score = \frac{\hat{\mu} - \mu_o}{\sigma_{\hat{\mu}}}$$

The t-score for a sample mean,

$$t - score = \frac{\hat{\mu} - \mu_o}{\sigma_{\hat{\mu}}}$$

Lets find the p-value for our bus fare example. Lets assume we know the standard deviation for bus fare is $.25.

H_o: μ = \$2.00

H_a: μ < \$2.00

We sample 100 busses and find that the average bus fare is \$1.95. We choose a significance level of $\alpha = .05$.

$$\mu_o = \$2.00 \qquad \hat{\mu} = \$1.95 \qquad n = 100$$

$$\sigma_{\hat{\mu}} = \frac{\sigma}{\sqrt{n}} = \frac{.25}{\sqrt{100}} = .025$$

$$z - score = \frac{\hat{\mu} - \mu_o}{\sigma_{\hat{\mu}}} = \frac{1.95 - 2.00}{.025} = -2$$

In our case, z=-2. We use the NORM.S.DIST function to find the p-value. We want the area below the z-score because our alternative hypothesis, H_a was μ <\$2.00. So, we use "=NORM.S.DIST(z-score,1)."

$$p - value = NORM.S.DIST(z - score, 1) = \quad NORM.S.DIST(-2,1) = .02275$$

If we don't know the standard deviation of the population we instead use the standard deviation of the sample, s.

In our bus fare example, lets say the standard deviation of $.25 was of the sample. In this case,

$$\sigma_{\hat{\mu}} = \frac{s}{\sqrt{n}} = \frac{.25}{\sqrt{100}} = .025$$

$$t - score \;\; = \;\; \frac{\hat{\mu} - \mu_o}{\sigma_{\hat{\mu}}} = \frac{1.95 - 2.00}{.025} = -2$$

Then, we use the T.DIST function to find the p-value. We want the area below the t-score because our alternative hypothesis, H_a was μ <$2.00. So, we use "=T.DIST(t-score,n-1,1)."

$$p - value = T.DIST(t - score, n - 1,1) = \;\; T.DIST(-2{,}99,1) = .0241$$

Sampling Distribution of the Difference Between Sample Proportions

The **normal distribution** is an appropriate model for a sampling distribution of a difference between sample proportions if there are at least 10 expected successes and 10 expected failures in both samples.

$n_1p_1 \geq 10$ and $n_1(1-p_1) \geq 10$ and $n_2p_2 \geq 10$ and $n_2(1-p_2) \geq 10$

Given the population proportions, p_{o1} and p_{o2} and the sample proportions, $\widehat{p_1}$ and $\widehat{p_2}$:

The standard deviation of the difference between sample proportions is,

$$\sigma_{\hat{p}} = \sqrt{\frac{p_{o1}(1-p_{o1})}{n_1} + \frac{p_{o2}(1-p_{o2})}{n_2}}$$

The z-score for a difference between proportions,

$$z - score \;\; = \;\; \frac{(\widehat{p_1} - \widehat{p_2}) - (p_{o1} - p_{o2})}{\sigma_{\hat{p}}}$$

Confidence Interval for a Difference Between Two Population Proportions

The confidence interval has the following general form

$$statistic \pm margin\ of\ error$$

Now, our statistic is the difference between two sample proportions. We thus use the following to estimate the difference in population proportions:

$$difference\ between\ sample\ proportions \pm margin\ of\ error$$

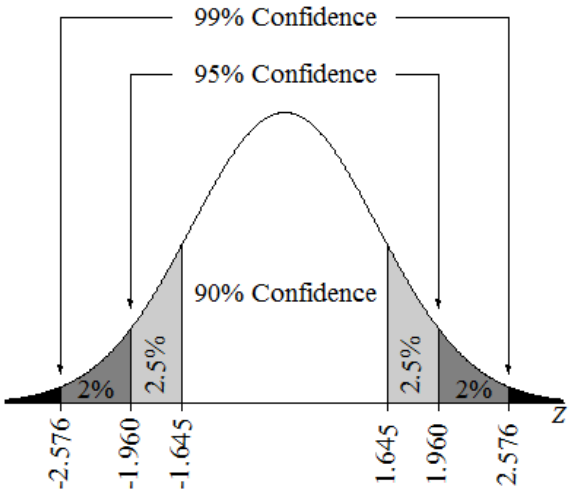
$$difference\ between\ sample\ proportions \pm Z_c * \sigma_{\hat{p}}$$

The standard error for sample proportions is,

$$\sigma_{\hat{p}} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

So, the confidence interval for a difference in sample proportions is,

$$(p_2 - p_1) \pm Z_c * \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$



Confidence Level	Critical Value Z_c
90%	1.645
95%	1.960
99%	2.576

Taken from OLI

Hypothesis Testing for a Difference Between Two Population Proportions

The general form of the null hypothesis for the difference between two population proportions is,

H_o: p_1-p_2 = 0 AKA $p_1 = p_2$

The alternative hypothesis can have one of three forms,

H_a: p_1-$p_2 \neq 0$ AKA $p_1 \neq p_2$

H_a: p_1-$p_2 < 0$ AKA $p_1 < p_2$

H_a: p_1-$p_2 > 0$ AKA $p_1 > p_2$

We create a pooled proportion from both sample proportions.

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

Then we use the pooled proportion to estimate the standard error

$$\sigma_{\hat{p}} = \sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}}$$

The z-score for a difference between proportions,

$$z - score = \frac{(\widehat{p_1} - \widehat{p_2}) - (p_{o1} - p_{o2})}{\sigma_{\hat{p}}}$$

But, the null hypothesis states that, $p_{o1} - p_{o2} = 0$. So, the z-score reduces to,

$$z - score = \frac{(\widehat{p_1} - \widehat{p_2})}{\sigma_{\hat{p}}}$$

Then, we use the Excel function **NORM.S.DIST** to get the p-value