MEANS --------- $t_{crit} = t_{LS/\#\ of\ tails}(DF)$ ---------

Nearly Normal Assumption
-Data comes from dist. that is uni-modal & symmetric
-With large enough samples (>40) we are safe to use normal model

One-Sample *t*-Interval

When the assumptions and conditions are met, we're ready to find the **confidence interval for the population mean, μ.** The confidence interval is

$$\bar{y} \pm t^*_{n-1} \times SE(\bar{y}),$$

One-Sample *t*-Test for the Mean

The conditions for the one-sample *t*-test for the mean are the same as for the one-sample *t*-interval. We test the hypothesis $H_0: \mu = \mu_0$ using the statistic

$$t_{n-1} = \frac{\bar{y} - \mu_0}{SE(\bar{y})},$$

where the standard error of $\bar{y}$ is: $SE(\bar{y}) = \frac{s}{\sqrt{n}}$.

When the conditions are met and the null hypothesis is true, this statistic follows a Student's *t*-model with $n - 1$ degrees of freedom. We use that model to obtain a P-value.

Confidence Interval for the Difference Between Two Means

When the conditions are met, we're ready to find a **two-sample *t*-interval** for the difference between means of two independent groups, $\mu_1 - \mu_2$. The confidence interval is

$$(\bar{y}_1 - \bar{y}_2) \pm t^*_{df} \times SE(\bar{y}_1 - \bar{y}_2),$$

where the standard error of the difference of the means is

$$SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

The critical value t^*_{df} depends on the particular confidence level and on the number of degrees of freedom.

Two-Sample *t*-Test

When the appropriate assumptions and conditions are met, we test the hypothesis

$$H_0: \mu_1 - \mu_2 = \Delta_0,$$

where the hypothesized difference Δ_0 is almost always 0. We use the statistic

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - \Delta_0}{SE(\bar{y}_1 - \bar{y}_2)}.$$

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

When the null hypothesis is true, the statistic can be closely modelled by a Student's *t*-model with a number of degrees of freedom given by a special formula. We use that model to compare our *t*-ratio with a critical value for *t* or to obtain a P-value.

Pooled *t*-Test and Confidence Interval for the Difference Between Means

The conditions for the **pooled *t*-test** for the difference between the means of two independent groups are the same as for the two-sample *t*-test, with the additional assumption that the variances of the two groups are the same. We test the hypothesis

$$H_0: \mu_1 - \mu_2 = \Delta_0,$$

where the hypothesized difference Δ_0 is almost always 0, using the statistic

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - \Delta_0}{SE_{pooled}(\bar{y}_1 - \bar{y}_2)}.$$

The standard error of $\bar{y}_1 - \bar{y}_2$ is

$$SE_{pooled}(\bar{y}_1 - \bar{y}_2) = s_{pooled}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

where the pooled variance is

$$s^2_{pooled} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}.$$

When the conditions are met and the null hypothesis is true, we can model this statistic's sampling distribution with a Student's *t*-model with $(n_1 - 1) + (n_2 - 1)$ degrees of freedom. We use that model to obtain a P-value for a test or a margin of error for a confidence interval.

The corresponding **pooled *t* confidence interval** is

$$(\bar{y}_1 - \bar{y}_2) \pm t^*_{df} \times SE_{pooled}(\bar{y}_1 - \bar{y}_2),$$

where the critical value t^* depends on the confidence level and is found with $(n_1 - 1) + (n_2 - 1)$ degrees of freedom.

Paired *t*-Test

When the conditions are met, we're ready to test whether the mean paired difference is significantly different from a hypothesized value (called Δ_0). We test the hypothesis

$$H_0: \mu_d = \Delta_0,$$

where the *d*'s are the pairwise differences and Δ_0 is almost always 0.

We use the statistic

$$t_{n-1} = \frac{\bar{d} - \Delta_0}{SE(\bar{d})},$$

where s_d is the mean of the pairwise differences, *n* is the number of pairs, and

$$SE(\bar{d}) = \frac{s_d}{\sqrt{n}},$$

where s_d is the standard deviation of the pairwise differences.

Paired *t*-Confidence Interval

When the conditions are met, we're ready to find the confidence interval for the mean of the paired differences. The confidence interval is

$$\bar{d} \pm t^*_{n-1} \times SE(\bar{d}),$$

where the standard error of the mean difference is $SE(\bar{d}) = \frac{S_d}{\sqrt{n}}$.

The critical value t^* from the student's *t*-model depends on the particular confidence level you specify and on the degrees of freedom, $n - 1$, which is based on the number of pairs, *n*.

Wilcoxin

$Ho: M_d = 0$	$Ho: M_d = 0$	-Dependent Samples
$Ha: M_d < 0$	$Ha: M_d \neq 0$	-Sample of Diff is NOT Normally Distributed

Mann-Whitney

$Ho: M_1 - M_2 = 0$	$Ho: M_1 - M_2 = 0$	-Independent Samples
$Ha: M_1 - M_2 < 0$	$Ha: M_1 - M_2 \neq 0$	-One or both of the samples is NOT normally distributed

Required Sample Size (always round up)

-proportion $n = \frac{Z_{\alpha/2}^2 \hat{p}\hat{q}}{E^2}$

*$\hat{p}, \hat{q}$ and E are always entered as decimals
*If $\hat{p}$ and $\hat{q}$ are not given use values of 0.5

-means (σ known) $n = \frac{Z_{\alpha/2}^2 \sigma^2}{E^2}$

-means (σ unknown) $n = \frac{T_{\alpha/2}^2 s^2}{E^2}$

The Kruskal-Wallis test

The **Kruskal-Wallis test** extends the Wilcoxon rank-sum (Mann-Whitney) test in order to deal with more than two samples.

H_0: All the samples come from the same distribution.

H_A: At least one of the samples comes from a distribution that is shifted higher or lower than the others.

Binomial

$Ho: p = 0.3$	$Ho: p = 0.4$
$Ha: p < 0.3$	$Ha: p \neq 0.4$

$P(X=x) = nCr * p^x q^{n-x}$

n=# of trials (sample size)
p=probability of a success
q=probability of a failure
q=1-p
x=# of successes in "n" trials

Reject Ho if $p < \alpha$

HYPOTHESIS TESTING

STEP 1: State Hypothesis for Parameter
H_0: Null Hypothesis
H_A: Alternate Hypothesis

STEP 2: Calculate the Test Statistic

STEP 3: Calculate the Critical Value

STEP 4: Compare & Conclusion
-Test Statistic vs Critical Value
-P-value vs Level of Significance

	Two-Sided	One-Sided RIGHT	One-Sided LEFT
How we write it in this book	$H_0: p = p_0$ $H_A: p \neq p_0$	$H_0: p = p_0$ $H_A: p > p_0$	$H_0: p = p_0$ $H_A: p < p_0$
How some people write it to spell out the details	No change, i.e., $H_0: p = p_0$ $H_A: p \neq p_0$	$H_0: p \leq p_0$ $H_0: p > p_0$	$H_0: p \geq p_0$ $H_0: p < p_0$
Practical example	Is the proportion of "up" days on the stock market different from the proportion of "down" days?	Is there a home field advantage?	Arc customers returning fewer items this year than the 3% they returned last year?
p_0 in the example	0.5	0.5	0.03

Decision or Conclusion	State of Nature H_0	State Of Nature H_A
H_0	Correct Decision	Type II Error: E_{II} $P[E_{II}] = \beta$
H_A	Type I Error: E_I $P[E_I] = \alpha$	Correct Decision

If P-value < LS: REJECT H_0 **Two-Tailed & Left – Tailed: If {|Test Statistic|} > {Critical Value}: REJECT H_0** **Right Tailed: If {Test Statistic} > {Critical Value}: REJECT H_0**

PROPORTIONS

Z_{crit}: z @ LS/ # of tails (from table)

Confidence Interval for the Difference Between Two Proportions

The confidence interval for the difference between two proportions is

$$(\hat{p}_1 - \hat{p}_2) \pm z \times SE(\hat{p}_1 - \hat{p}_2),$$

where z^* is the critical value and

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}.$$

One-Proportion *z*-Interval

When the conditions are met, we are ready to find the confidence interval for the population proportion, *p*. The confidence interval is $\hat{p} \pm z^* \times SE(\hat{p})$, where the standard deviation of the proportion is estimated by $SE(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}}$, and z^* is the critical value discussed in Section 11.3.

Binomial Formula

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

This starts the count of number of ways event can occur. This ends the count of number of ways event can occur. This deletes duplications. This is the probability of success for x trials. This is the probability of failure for the x trials.

Common Z_{crit} Values

Alpha	"<" in Ha 1-Tailed	">" in Ha 1-Tailed	"≠" in Ha 2-Tailed
0.2	-0.84	0.84	1.28
0.1	-1.28	1.28	1.645
0.05	-1.645	1.645	1.96
0.02	-1.645	2.055	2.325
0.01	-2.325	2.325	2.575
0.005	-2.575	2.575	2.81

Two-Proportion *z*-Test

Testing whether the difference between two proportions is equal to a given number, K.

In order to test
$H_0: p_1 - p_2 = K$
$H_A: p_1 - p_2 \neq K$
we calculate the test statistic:

$$z = \frac{\hat{p}_1 - \hat{p}_2 - k}{SD(\hat{p}_1 - \hat{p}_2)}$$

where

$$SD(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}.$$

We then obtain the corresponding P-value from the table for the Normal distribution.

Two-Proportion *z*-Test for equal proportions

Testing whether two proportions are equal.

In order to test
$H_0: p_1 - p_2 = 0$
$H_A: p_1 - p_2 \neq 0$
we calculate the test statistic:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{SD(\hat{p}_1 - \hat{p}_2)}$$

where

$$SD(\hat{p}_1 - \hat{p}_2) = \sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

and

$$p = \frac{x_1 + x_2}{n_1 + n_2} \text{ and } \bar{q} = 1 - \bar{p}.$$

One-Proportion *z*-Test

The conditions for the one-proportion *z*-test are the same as for the one-proportion *z*-interval. We test the hypothesis $H_0: p = p_0$ using the statistic

$$z = \frac{(\hat{p} - p_0)}{SD(\hat{p})}.$$

We use the hypothesized proportion to find the standard deviation:

$SD(\hat{p}) = \sqrt{\frac{p_0 q_0}{n}}$. When the conditions are met and the null hypothesis is true, this statistic follows the standard Normal model, so we can use that model to obtain a P-value.

- **Independence Assumption:** The two samples are independent of each other.
- **Randomization Condition:** The people in each sample were selected at random.
- **10% Condition:** The sample is less than 10% of the population of the two countries.
- **Success/Failure Condition:** $n_1p_1 > 10$; $n_1q_1 > 10$; $n_2p_2 > 10$; $n_2q_2 > 10$.

CHI SQUARE

Assumptions & Conditions
-Counted Data Condition: The data must be counted for the categories of the categorical variable
-Independence Assumption & Randomization Condition
-Sample Size Condition → Expected Cell Frequency >= 5

GOF VS Homogeneity (Independence)

GOF:

df = r -1
H_0: The distribution is the same
H_A: The distribution is different

Homogeneity (Independence):

Df = (r-1)(c-1)
H_0: Homogeneous
H_A: Not Homogeneous

H_0: Independent
H_A: Not Independent

Standardized Residuals

$$\frac{Obs - Exp}{\sqrt{Exp}}$$

-Residuals become 'z-scores' and can now be checked for significance using z values

How to Find Expected Values

In a contingency table, to test for homogeneity, we need to find the expected values when the null hypothesis is true. To find the expected value for row *i* and column *j*, we take

$$Exp_{ij} = \frac{Total_{row\ i} \times Total_{Col\ j}}{Table\ Total}.$$

Here's an example:

Suppose we ask 100 people, 40 men and 60 women, to name their magazine preference, *Sports Illustrated*, *Cosmopolitan*, or *The Economist*, with the following result:

Magazine Preference

	Sports Illustrated	Cosmopolitan	Economist	Total
Men	25	5	10	**40**
Women	10	45	5	**60**
Total	35	50	15	**100**

Then, for example, the expected value under homogeneity for *Men* who prefer *The Economist* would be

$$Exp_{13} = \frac{40 \times 15}{100} = 6.$$

Performing similar calculations for all cells gives the expected values:

	Sports Illustrated	Cosmopolitan	Economist	
Men	14	20	6	40
Women	21	30	9	60
	35	50	15	100

The Chi-Square Calculation

Here are the steps to calculate the chi-square statistic:

1. **Find the expected values.** These come from the null hypothesis model. Every null model gives a hypothesized proportion for each cell. The expected value is the product of the total number of observations times this proportion. (The result need not be an integer.)
2. **Compute the residuals.** Once you have expected values for each cell, find the residuals, $Obs - Exp$.
3. **Square the residuals.** $(Obs - Exp)^2$
4. **Compute the components.** Find $\frac{(Obs - Exp)^2}{Exp}$ for each cell.
5. **Find the sum of the components.** That's the chi-square statistic, $\chi^2 = \sum_{all\ cells} \frac{(Obs - Exp)^2}{Exp}$.
6. **Find the degrees of freedom.** It's equal to the number of cells minus one.
7. **Test the hypothesis.** Large chi-square values mean lots of deviation from the hypothesized model, so they give small P-values. Look up the critical value from a table of chi-square values, such as Table X in Appendix B, or use technology to find the P-value directly.

The steps of the chi-square calculations are often laid out in tables, as in Table 16.3. Use one row for each category, and columns for observed counts, expected counts, residuals, squared residuals, and the contributions to the chi-square total:

	Observed	Expected	Residual $= (Obs - Exp)$	$(Obs - Exp)^2$	Component $= \frac{(Obs - Exp)^2}{Exp}$
Monday	192	193.369	−1.369	1.874	0.0097
Tuesday	189	202.582	−13.582	184.471	0.9106
Wednesday	202	203.695	−1.695	2.873	0.0141
Thursday	199	200.607	−1.607	2.582	0.0129
Friday	218	199.747	18.253	333.172	1.6680

Table 16.3 Calculations for the chi-square statistic in the trading days example.

ANOVA - Analysis of Variance

One Way Analysis of Variance

H_0 : All means are equal

H_a : At least one mean differs from the others

Source	SS	DF	MS	F
Model	SST	I-1	MST	MST/MSE
Error	SSE	N-I	MSE	
Total	SSTotal	N-1		

I - # of sample groups, N – sum total of all sample sizes

Two Way Analysis of Variance (ANOVA)

A Test for Interaction

- The first step is to see if there is any interaction between the two factors
- H_0: No interaction between factors A and B
  H_a: A and B interact
- If H_0 is true then the individual measurements are modeled as

$$X_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$

(X_{ijk}: k[th] individual Measurement from the i[th] level of the first factor and the j[th] level of the 2[nd] factor; μ: Mean Response; α_i: Impact of Factor A; β_j: Impact of Factor B; ε_{ijk}: Random Variation (or residual))

- If H_0 is false then we have a different model that includes an interaction term

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

Deriving a Test Statistic

- Our test statistic is therefore

$$F = \frac{MSAB}{MSE} \sim F_{(a-1)(b-1),ab(r-1)}$$

- If we have determined that there is no significant interaction we can test for the main effects of each separately…

- Step 3: If Interaction is **not** Present, then with caution, **do** all the further analysis on the Factor Level Means to find out if the they are all Equal, or some of them not Equal.

Meaningful Main Effects?

- The main effects in our example were significant while the interaction term was not significant.
- If the interaction had been significant then the main effects tests are less easily interpretable as they no longer represent the entire impact of each factor.

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Sum of Squares	F-statistic
Factor A	a-1	SSA	MSA = SSA/dfa	MSA/MSE
Factor B	b-1	SSB	MSB = SSB/dfb	MSB/MSE
AB interaction	(a-1)(b-1)	SSAB	MSAB = SSAB/dfab	MSAB/ MSE
Error	ab(r-1) or N – ab	SSE	MSE = SSE/dfe	
Total	N-1	SSTotal		

ANOVA-BLOCKED

Assumptions
-See 1-Way ANOVA
**Review BLOCKED ANOVA table in the book for detailed calculations.
2 Possible Tests

Test #1-Was Blocking Needed?
$Ho: \mu_{b1} = \mu_{b2} = ... \mu_{bn}$
Ha: not all block means equal
$F_{stat} = MS_{block}/MS_{err}$

Test #2-Factor Means
$Ho: \mu_{a1} = \mu_{a2} = ... \mu_{an}$
Ha: not all factor means equal
$F_{stat} = MS_{facA}/MS_{err}$

***If in test #1 we DO NOT reject Ho we must run a 1-WAY ANOVA to complete test #2, otherwise, you can proceed with test #2 using the output.

For both tests Reject Ho if $|F_{stat}| > |F_{crit}|$ or $p < \alpha$.
There will be 2 F-values and/or p-values in minitab.

$df_{block} = b-1$ $df_{fac} = c-1$ $df_{err} = (b-1)*(c-1)$ $df_{tot} = N-1$ b = # of rows

The Model

- We proceed essentially *as if* this were a two factor analysis with *no interaction term.*
- Thus our model is…

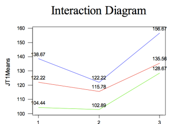
$$X_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

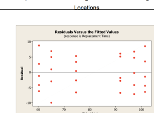
- where ε_{ij} is a random variable with constant variance σ^2
- We divide the total variation SSTotal into variation due to the treatment (column), SSB, the variation due to the block (row), SSA and the random variation due to "noise", SSE: SSTotal = SSB + SSA + SSE
- Degrees of freedom are N-1 (for SSTotal), b-1 (for SSB), a-1 (for SSA) and (a-1)(b-1) (for SSE) where a is the number of blocks – that is the number of observations in each treatment level – and b is the number of treatment levels
- We are only interested in the main effect of the column variable

Example

- Suppose we are interested in testing the impact of gasoline types (regular, premium, ethanol) on gas mileage.
- Candidate for one way ANOVA as we have one factor – gasoline type – and one response variable – gas mileage.
- But a potentially confounding factor that impacts gas mileage is vehicle type
- If we did a completely randomized design, our samples for each factor may have different types of vehicle which would bring into question any conclusion we might draw
- One way of controlling for this is to divide the population of cars up by size and randomly assign ***one*** car from each size to each gasoline type (essentially creating a 2[nd] factor)
- This way we insure that the size of the car does not impact on the outcome.
- This is called a ***randomized block design***.

- Please notice the non-parallelness (skewness) of the three lines. This indicates some Interaction.





- Independence Condition:
  As before, there is no test to verify this assumption, but we can check the randomization condition. The observations within *each treatment group* must be independent of one another.
- Equal Variance Condition:
  In two-way ANOVA (as in one-way ANOVA), the variances of all treatment groups are required to be equal.
  The residuals after fitting *both* effects are what we pool for the error mean square, so it's their variance that we need to examine.
  We check this assumption with the similar variance condition.
- Normality Condition:
  In two-way ANOVA we need to make sure that the underlying errors follow a Normal model.
  To do so, we check a corresponding nearly Normal condition with a Normal probability plot or histogram of the residuals.

7

1-WAY ANOVA

Assumptions
-Samples Normal
-Equal Variances
-Observations Ind

$Ho: \mu_1 = \mu_2 = \mu_3 = ... \mu_n$
Ha: not all means equal
***Review 1 WAY ANOVA table in the book for detailed calculations.

$F_{stat} = MS_{fac}/MS_{err}$
Reject Ho if $|F_{stat}| > |F_{crit}|$ or $p < \alpha$.
Both F_{stat} and p come from minitab.

$df_f = k-1$
$df_e = N-k$
$df_t = N-1$
k=# rows

2-WAY ANOVA

Assumptions
-See 1-Way ANOVA
**Review 2 WAY ANOVA table in the book for detailed calculations.
3 Possible Tests

Test #1-Interaction
Ho: Factors A and B do not Interact
Ha: Factors A and B Interact
$F_{stat} = MS_{Int}/MS_{err}$

Test #2-Factor A Means
$Ho: \mu_{1a} = \mu_{2a} = ... \mu_{na}$
Ha: not all means equal
$F_{stat} = MS_{facA}/MS_{err}$

Test #3-Factor B Means
$Ho: \mu_{1b} = \mu_{2b} = ... \mu_{nb}$
Ha: not all means equal
$F_{stat} = MS_{facB}/MS_{err}$

For all 3 tests Reject Ho if $|F_{stat}| > |F_{crit}|$ or $p < \alpha$.
There will be 3 F-values and/or p-values in minitab.

$df_{facA} = r-1$ $df_{facB} = c-1$ $df_{int} = (r-1)*(c-1)$ $df_{err} = r*c*(n-1)$ $df_{tot} = N-1$

Bonferroni used to identify where the actual differences between means are.

Step 1 - Determine number of rows (r), number of columns (c) and number of observations in each samples being compared (n)
Step 2 - Calculate "J" (the number of possible comparisons) J=[(rc)*(rc-1)]/2 s=SQRT(MSE)
Step 3 - Calculate alpha/2(J) and the df = dfe
Step 4 - Get T-value from table at a 1 tailed significance level of alpha/2(J) and dfe (pick closest value)
Step 5 - Interval = $(x\_bar_1 - x\_bar_2)$ +/- $t_{\alpha/2J}$ * s *SQRT($1/n_1 + 1/n_2$)
critical difference = margin of error = $t_{\alpha/2J}$ * s *SQRT($1/n_1 + 1/n_2$)
Rule - if the interval contains "0", there is no difference between the means being compared
OR - if |actual diff| < |critical diff|, there is no difference between the means

-Used for 1-WAY ANOVA and 2-WAY ANOVA, do not use for BLOCKED ANOVA
-For 2-WAY ANOVA there are 3 types of possible comparisons, row comparisons, column comparisons and cell comparisons, all steps will be exactly the same except step 1. The values of "r", "c" "n" are determined based on the type of comparison.

CHOOSING THE RIGHT TEST

Are we interested in means or proportions?

- Means → Measurement Data
  - How many samples are there?
    - One → Can we link the samples in some way?
      - Dependent
      - Independent
    - Two
    - Three or more
- Proportions → Categoric Data
  - How many samples are there?
    - One
    - Two
    - Three or more

Measurement Data, One Sample

Normality Assumption Warranted?

- Yes → Is σ known?
  - Known → One sample z-test → $H_0: \mu = \mu_0$, $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$
  - Unknown → One sample t-test → $H_0: \mu = \mu_0$, $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$
- No → Wilcoxon Signed Rank Test

Measurement Data, Two *Independent* Samples

Is the normality assumption Warranted?

- Yes → Can we assume equal variance?
  - No → Unpooled two Sample t-test → $H_0: \mu_1 - \mu_2 = 0$, $t = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_{\nu}$
  - Yes → Pooled two Sample t-test → $H_0: \mu_1 - \mu_2 = 0$, $t = \frac{\bar{x}_1 - \bar{x}_2 - 0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$
- No → Mann-Whitney U test

Measurement Data, Two *Dependent* Samples

Is the normality assumption warranted?

- Yes → Matched Pairs Test → $H_0: \mu_d = 0$, $t = \frac{\bar{d} - 0}{s_d/\sqrt{n}} \sim t_{n-1}$
- No → Wilcoxon Signed Rank Test

Measurement Data, Three or More Samples

One or Two Factors?

- One → One Way Analysis of Variance → $H_0: \mu_1 = \mu_2 = \ldots = \mu_j$, $F = \frac{MSA}{MSE} \sim F_{A-1,N-1}$
  - Non parametric alternative for a one way ANOVA: Kruskal-Wallis Test
- Two → Two Way Analysis of Variance → First, test for an interaction: H_0: No Interaction, $F = \frac{MSAB}{MSE} \sim F_{(A-1)(B-1),AB(n-1)}$
  - Accept H_0? No → Done
  - Accept H_0? Yes → Main Effect of Row Factor: H_0: Row Factor has no effect, $F = \frac{MSA}{MSE} \sim F_{A-1,AB(n-1)}$; Main Effect of Column Factor: H_0: Column Factor has no effect, $F = \frac{MSB}{MSE} \sim F_{B-1,AB(n-1)}$

Categoric Data, One Sample

$np > 10$, $n(1-p) > 10$

- One sample Proportion test → $H_0: p_1 = p_0$, $z = \frac{\hat{p}_1 - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim N(0,1)$
- Binomial Distribution
  - Determine the probability of getting the result you did or something that gives even more evidence for the alternative using the binomial distribution.

Categoric Data, Two *Independent* Samples

Does the Null Hypothesis assume proportions equal?

- Pooled two sample Proportion t-test → $H_0: p_1 - p_2 = 0$, $z = \frac{\hat{p}_1 - \hat{p}_2 - 0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$, where $\hat{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{n_1\hat{p}_1 + n_2\hat{p}}{n_1 + n_2}$
- Unpooled two sample Proportion test → $H_0: p_1 - p_2 = p_0$, $z = \frac{\hat{p}_1 - \hat{p}_2 - 0}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \sim N(0,1)$

Categoric Data, Three or More Samples

Chi-square test → H_0: No relationship, $\chi^2 = \sum_{all\,cells}\left(\frac{(Observed - Expected)^2}{Expected}\right) \sim \chi^2_{(row-1)\times(col-1)}$

Using *p* variables (predictors) to predict another (response variable)

Multiple Regression

- $H_0: \beta_1 = 0$ → Tests whether the jth predictor adds anything to a model ***that already contains all the others.*** → $t = \frac{b_j - 0}{SE(b_j)} \sim t_{n-k-1}$
- H_0: Predictors don't help at all → Tests whether the model as a whole helps predict the response variable - a significant result only tells you that it helps predict y but does not tell you how well it predicts y. → $F = \frac{MSR}{MSE} \sim F_{k,n-k-1}$

Using one variable to predict another

Simple Regression

- $H_0: \beta_1 = 0$ → ANOVA → $F = \frac{MSR}{MSE} \sim F_{1,n-2}$
- $H_0: \rho = 0$ → t-test → $t = \frac{b_1 - 0}{s/\sqrt{\sum(x_i - \bar{x})^2}} \sim t_{n-2}$

These are three ways of testing the same thing - ***does the one variable help predict the other in a linear fashion?***

- $t = \frac{r - 0}{\sqrt{\frac{1 - r^2}{n-2}}} \sim t_{n-2}$

REGRESSION

Recognizing Violations of the Assumptions

1. Non-linearity
   - The residual plot will have a curve to it
2. Non-constant variance (heteroscedasticity)
   - The spread of the residuals will change with the fitted value
3. Non-normality of the errors
   - If there are more outliers than would be expected by the empirical rule then try removing them
   - If there is skewness, try variance-stabilizing transformations

An Example of Data Transformations

- Often if the true relationship between X and Y is non-linear it is possible to "linearize" the data by transforming either the X or the Y or both
- For example, consider $Y = \alpha_0(\alpha_1)^X\epsilon$
- If we take the log of both sides we get

$$\log(Y) = \log(\alpha_0\alpha_1^X\varepsilon) = \log\alpha_0 + X\log\alpha_1 + \log\varepsilon$$

and thus the log of Y is linearly related to X

```
The regression equation is
fx = 42.9 + 0.325 midt                                   <- Regression line

Predictor      Coef  SE Coef      T      P
Constant     42.907    4.087  10.50  0.000   } The "value" of
midt        0.32528  0.06042   5.38  0.000   } each coefficient

S = 11.4724   R-Sq = 20.3%   R-Sq(adj) = 19.6%
   ^ Estimate of σ   ^ Total variation explained by the model
Analysis of Variance

Source          DF       SS       MS      F      P   } ANOVA test to
Regression       1   3815.1   3815.1  28.99  0.000   } determine if
Residual Error 114  15004.2    131.6                  } slope of the
Total          115  18819.3                           } population
                                                      } regression line
Unusual Observations                                  } is non-zero.

Obs  midt     fx    Fit  SE Fit  Residual  St Resid
 46    77  40.00  67.85    1.27    -27.85    -2.44R   } R denotes a
 54    73  33.57  66.76    1.17    -33.19    -2.91R   } point with a
 59    60  29.29  62.42    1.11    -33.14    -2.90R   } large residual
 60    74  19.29  67.12    1.20    -47.84    -4.19R   } X denotes a
108    74  42.14  67.12    1.20    -24.98    -2.19R   } point whose X
111    69  41.43  65.32    1.09    -23.89    -2.09R   } value makes it
114    18  52.86  48.69    3.06      4.17     0.38 X  } influential.
115    51  33.57  59.53    1.37    -25.96    -2.28R
116    17  44.29  48.33    3.13     -4.04    -0.37 X
```

Simple Linear Regression

Population Model: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

Y_i = value of the dependent variable on i^{th} observation

X_i = known value of the independent variable on the i^{th} observation

ε_i = a r.v. or random error such that

- $\mu(\varepsilon_i) = 0$ (equally likely to be above as below the regression line)
- $\sigma^2(\varepsilon_i) = \sigma^2 \ \forall i$ (constant variance)
- ε_i's are independent, normally distributed

We want to estimate β_0 and β_1 to get the "best" regression line possible.

Idea: Have an independent variable *X* that you believe can help predict some other variable *Y*. What we're going to do is estimate β_0 and β_1 in order to get an *estimate* of the regression line

$$Y_i = b_0 + b_1 X_i$$

This is an *estimate* of what the true regression line would be if we could measure the whole population. This in turn in an *approximation* of the true interaction between *X* and *Y*.

Hypothesis Testing - Is there a Linear Relationship?

$H_0: \beta_1 = 0$ What does this mean?

- no *linear* relationship between *X* and *Y*
- *X* is not a useful *linear* predictor of *Y*

Test Statistic: $t = \frac{b_1 - 0}{SE(b_1)}$

- $H_a: \beta_1 > 0$ Reject if $t > t_{\alpha,n-2}$
- $H_a: \beta_1 < 0$ Reject if $t < -t_{\alpha,n-2}$
- $H_a: \beta_1 \neq 0$ Reject if $t > t_{\alpha/2,n-2}$ or $t < -t_{\alpha/2,n-2}$

Hypothesis Test

$H_0: \beta_1 = 0$ (X has no explanatory info for predicting Y)

Test Statistic: $F = \frac{MSR}{MSE}$ has a F-distribution with 1 and $n-2$ degrees of freedom - *if H_0 is true.*

Reject H_0 if $F > F_{\alpha,1,n-2}$ or use P-values.

(Note: $F_{1,n-2} = t^2_{n-2}$)

Two Key Intervals

1. C.I. for the mean of Y at a given value of X
2. Prediction interval (P.I.) for a future observation of Y at a given value of X

For a given x^*, which interval will be bigger?

- P.I. since the mean varies less than a given observation.

Same point estimate for both:

$$\hat{y}_{x^*} = b_0 + b_1 x^*$$

C.I. for $\mu(Y \mid x^*)$

$$\left(b_0 + b_1 x^*\right) \pm t_{\alpha/2,n-2} \times s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

P.I. for $Y \mid x^*$

$$\left(b_0 + b_1 x^*\right) \pm t_{\alpha/2,n-2} \times s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

Coefficient of Determination

R^2 = "Coefficient of Determination" $s = \sqrt{MSE}$

= Proportion of total variation explained by the model

$$= \frac{SSR}{SSTotal} \quad \left(\text{or } 1 - \frac{SSE}{SSTotal}\right)$$

(In simple linear regression, r^2 - the correlation coefficient - is equal to R^2)

Computation of R^2_{adj}

R^2_{adj} = 1 - (SSE/[n - (k + 1)])/(SSTotal/(n - 1))

= 1 - (MSE) / (SSTotal/(n - 1))

= 1 – (Error Variance / Total Variance)

Multiple Regression

- a single dependent variable Y
- *multiple* independent predictor (or explanatory) variables $X_1, X_2, \ldots, X_k$

The Model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik} + \varepsilon_i$$

Assumptions are the same as before:

- Y_i's are
  - normally distributed,
  - independent, and
  - have constant variance σ^2.
- Y is linearly related to the predictors

Two Tests

- We used Minitab to determine the appropriate sample regression line $y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k$
- Minitab also gives 2 types of tests
  - F-test (ANOVA table) that tests the complete model
  - T-tests (one for each predictor) that tests the usefulness of each predictor *given that all the other predictors are already in the model.*

ANOVA Table

Source	SS	df	MS	F
Model	SSR	k	MSR	MSR/MSE
Error	SSE	n-k-1	MSE	
Total	SSTotal	n-1		

F-Test (ANOVA)

- The F-test looks very much like an ANOVA and compares the full model $y = \beta_0 + \beta_1 x_1 + \ldots \beta_k x_k$ against a model that includes only the constant $y = \beta_0$
- In other words, it is testing the hypothesis $H_0: \beta_1 = \ldots = \beta_k = 0$, H_a: At least one β_i is not zero

T-test (for each predictor)

- The t-test (for the first predictor) is testing the full model $y = \beta_0 + \beta_1 x_1 + \ldots \beta_k x_k$ against a model that contains all the predictors except predictor one $y = \beta_0 + \beta_2 x_2 + \ldots \beta_k x_k$

Thus, it is testing whether predictor x_1 improves the predictive power of a model that already contains all the others.

Multicollinearity

- *Multicollinearity* occurs when at least one predictor is or is close to being a linear combination of the other predictor variables.
- *Pairwise collinearity* occurs when two predictors are highly correlated (as in previous slide)
- Multicollinearity will affect our interpretation of the coefficients but does not affect our prediction.
- We test using the Variation Inflation Factors (VIF) and say that any VIF bigger than 10 constitutes significant multi-collinearity.
- In such a case, we would most likely get rid of one or more predictors. $VIF_j = \frac{1}{1 - R_j^2}$

Introducing Categoric Data

- Suppose you have a categoric predictor with k levels
- Introduce k-1 dummy variables (see slides on qualitative predictors)
- Essentially creating k separate regression lines where the assumption is that the impact of the numeric predictors is independent of the level of the categoric predictor
- If interaction is suspected, need to include interaction predictors as in the example in the slides on qualitative predictors

Stepwise Regression

1. Find the simple regression model that maximizes $|t_{stat}|$, provided its p-value < α;
2. Add a second variable if it has the highest $|t_{stat}|$, provided its p-value < α;
3. Having added a variable, consider dropping an existing variable with the smallest $|t_{stat}|$, if its p-value > α;
4. Keep adding variables until no new variable meets the p-value < α criterion.

Mallows' Cp Statistic

- We choose the set of p predictors that give the smallest Cp statistic subject to the constraint that Cp is approximately p.

- Relatively few predictors, to keep the model simple
- A relatively high R^2, indicating that much of the variability in *y* is accounted for by the regression model
- A relatively small value of σ, the standard deviation of the residuals, indicating that the magnitude of the errors is small
- Relatively small P-values for the *F*- and *t*-statistics, showing that the overall model is better than a simple summary with the mean and that the individual coefficients are reliably different from zero
- No cases with extraordinarily high leverage that might dominate and alter the model
- No cases with extraordinarily large residuals, and Studentized residuals that appear to be nearly Normal. Outliers can alter the model and certainly weaken the

Multiple Regression or Simple Regression Assumptions
-Error Terms Norm Dist.
-Constant Variance
-Error Terms Ind.
-Linear Model

Test #1-Significance of Model
Ho: $B_1=B_2=B_3\ldots=B_n=0$
Ha: Atleast one $B_n \neq 0$
$F_{stat} = MS_{reg}/MS_{err}$
Reject Ho if $|F_{stat}| > |F_{crit}|$ or $p < \alpha$.

Test #2-Significance of Variable in Model
Ho: $B_n = 0$ Ha: $B_n \neq 0$
$T_{stat} = \frac{b_n - B_n}{S_{bn}}$ $df = df_e$
Reject Ho if $|T_{stat}| > |T_{crit}|$ or $p < \alpha$
-All numbers come from minitab.
-This test can be repeated for each independent variable.
CI (coefficient) = $b_n +/- T_{\alpha/2} * s_{bn}$

CI = fit +/- $T_{\alpha/2}$ *SE_{fit} (interval for average of all observations) $df_{reg} = k$ $df_{err} = n-k-1$
PI = fit +/- $T_{\alpha/2}$ *SQRT($SE_{fit}^2 + S^2$) (interval for 1 observation) $df_{tot} = n-1$
-Regression Equation Given in Output as $\hat{y} = c + B_1x_1 + B_2x_2 + \ldots + B_nx_n$
-Residual= actual value - value predicted by model by subbing independent variables into the regression equation.
-The best model will be the one with the highest R^2 or R^2_{adj}, the lowest S and the lowest number of ind. variables.

Variance Inflation Factor (VIF) Multicollinearity is a problem caused by highly correlated independent variables. It is a problem if VIF>10.
VIF = $1/(1-R^2)$

ANOVA and REGRESSION
$R^2 = 1-(SS_{err}/SS_{tot})$
$R^2_{adj} = 1-(MS_{err}/MS_{tot})$
s=SQRT(MSE) $MS = \frac{SS}{df}$
SS_e = SUM[$(n_i-1)*s_i^2$]

Manual Simple Regression Calculations:
Slope = $b_1 = r * (s_x / s_y)$
Intercept = $b_0 = \bar{y} - b_1\bar{x}$
Regression equation: $\bar{y} = b_0 + b_1\bar{x}$
r = coefficient of correlation
s_x = standard dev. of x
s_y = standard dev. of y
$\bar{x}$ = mean of x values
$\bar{y}$ = mean of y values