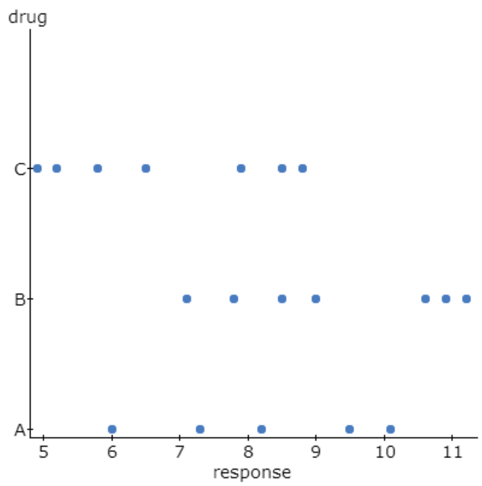
ST 517 Note Outline 8: Inference for More than Two Groups

Lecture 8.1: Introduction to Analysis of Variance

Example: We wish to compare three drugs for treating some disease. A quantitative response is measured such that a smaller value indicates a more favorable response. A total of 19 patients are randomly assigned to one of the three drug (treatment) groups. A dotplot and summary of the data is provided below. Based on these, can you say that the average response is different for the three drugs?



Group	n	Mean	Std Dev
Drug A	5	8.2	1.7
Drug B	7	9.3	1.6
Drug C	7	6.8	1.6
Overall	19	8.1	1.9

Recall: Pooled two sample t-test

H_0: $\mu_1 = \mu_2$
H_A: $\mu_1 \neq \mu_2$

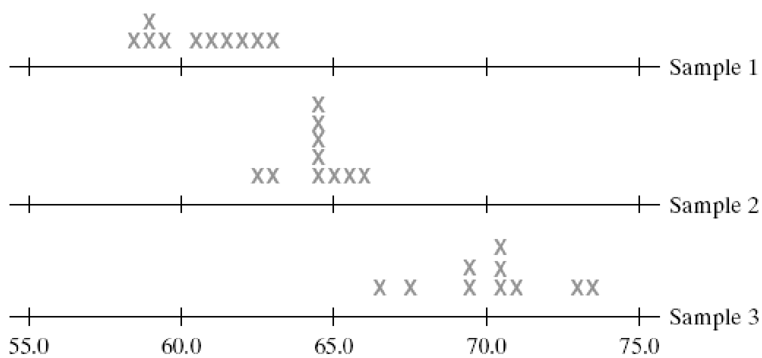
$$t = \frac{(\bar{y}_1 - \bar{y}_2) - \mu_0}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$$

where $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$ is an estimate of the common population variance

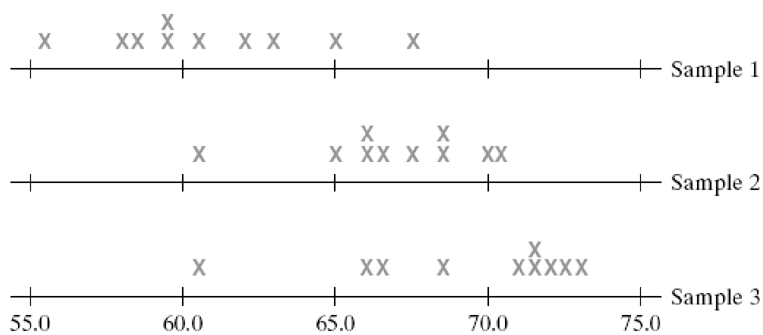
- Idea: Extend this to testing 3 or more groups

Basic Idea: Which of the following scenarios provides more evidence that at least one of the population means is different from the others?

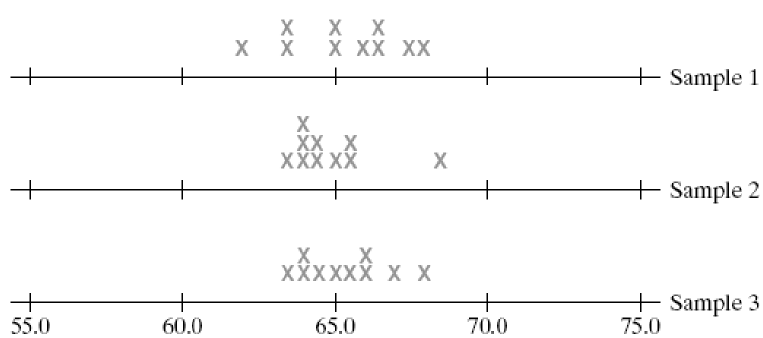
Scenario 1:



Scenario 2:



Scenario 3:



Analysis of Variance (ANOVA/AOV)

- How different are group means, relative to how much variability we expect?
  - Variability with-in each group as well as between the groups

- Notation
  - y_{ij} - j th individual observation in group i
  - $\bar{y}_i$ - average for group i
  - $\bar{y}_{..}$ - grand mean for all observations
  - t - number of groups
  - n_i - sample size for group i
  - $n = \sum_i n_i$ - total sample size

Measure of Overall Variability

- How far is each observation from the grand mean?

$$\mathrm{SS}_{total} = \sum_{i,j} \left(y_{ij} - \bar{y}.. \right)^2$$

Measure of Within Group Variability

- How far is each observation from its group mean?

$$\mathrm{SSE} = \mathrm{SS}_{error} = \sum_{i,j} \left(y_{ij} - \bar{y}_i. \right)^2$$

$$\mathrm{SSE} = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \ldots + (n_t - 1)s_t^2$$

Measure of Between Group Variability

- How far is each sample mean from the grand mean?

$$\mathrm{SS}_{groups} = \sum_i n_i \left(\bar{y}_i - \bar{y}.. \right)^2$$

Degrees of Freedom

Recall:

- One sample t-test: $n-1$ degrees of freedom
- Two sample t-test: $n-2$ degrees of freedom (where $n = n_1 + n_2$)

This scenario:

- __________ degrees of freedom for error
- __________ degrees of freedom for groups
- Notice:

Mean Squares

- Divide each Sum of Squares by the appropriate degrees of freedom

$$MS_{groups} = \frac{SS_{groups}}{(t-1)}$$

$$MSE = \frac{SSE}{(n-t)}$$

ANOVA Table

Source	Degrees of Freedom	Sum of Squares	Mean Squares	F-Statistic	p-value
Between Samples (Groups)					**(Provided by computer)**
Within Samples (Error)					
Total					

Lecture 8.2: Hypothesis testing for ANOVA—the F-test

- When to use:

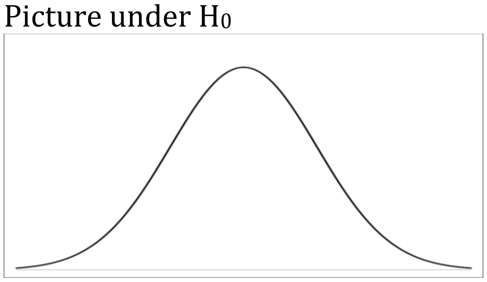
- Parameter: (see step 1 of hypothesis test)

Step 1: Identify research question, population, parameter

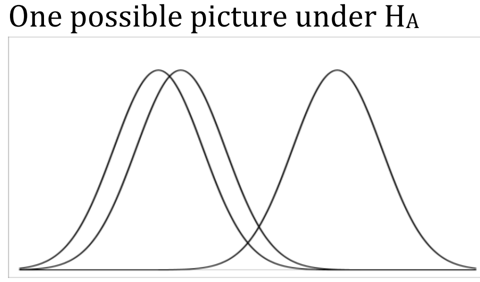
- Research question for ANOVA F-test always of form: Is there a difference between groups on average?
- More than two populations
- One parameter (μ_i) for each population

Step 2: Establish null, alternative hypothesis

Picture under H_0



One possible picture under H_A



Step 3: Identify type of test, check conditions

- Type of test:
- Conditions required for this test to be valid

Step 4: Calculate test statistic

- Test statistic:

$$F = \frac{MS_{groups}}{MSE}$$

  - New interpretation:

Step 5: Identify null distribution, calculate p-value

- Null distribution:
  - Numerator df =
  - Denominator df =
- p-value always represents proportion of null distribution that is greater than or equal to the test statistic
  - No need to be concerned about two-sided vs. one-sided

Step 6: Make a decision about the null hypothesis

- Decision rule of thumb: p-value $\leq \alpha$, reject H_0

Step 7: State conclusion in context of the alternative hypothesis

- If we rejected H_0: Conclude that there is enough evidence to support H_A.
- If did <u>not</u> reject H_0: Conclude that there is <u>not</u> enough evidence to support H_A.

The F Distribution

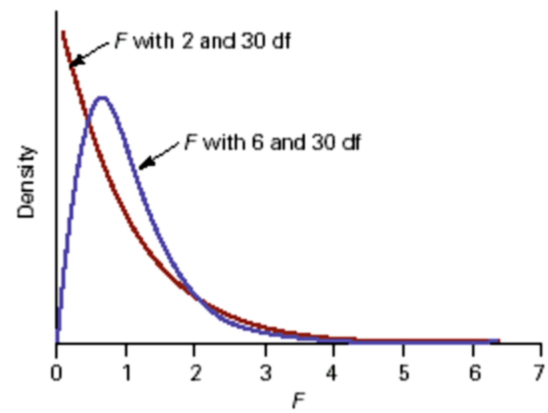
- Models the ratio of two random variables that follow the Chi-square distribution
- Most commonly used for statistical inference, for example when conducting inference for more than 2 population means

- PDF:

$$\frac{1}{\mathrm{B}\left(\frac{\nu_1}{2};\frac{\nu_2}{2}\right)}\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} y^{\frac{\nu_1}{2}-1}\left(1+\frac{\nu_1}{\nu_2}y\right)^{-\frac{\nu_1+\nu_2}{2}} \qquad \text{where } \mathrm{B}(x,y)=\int_0^1 t^{x-1}(-t)^{y-1}dt$$

- Generally right skewed (recall: values of F-statistic cannot be negative)

- Depends on two parameters:



- If you know ν_1 and ν_2, you know everything there is to know about that F-distribution

- Mean of F-distribution:

$$\frac{\nu_2}{\nu_2-2} \text{ for } \nu_2 > 2$$

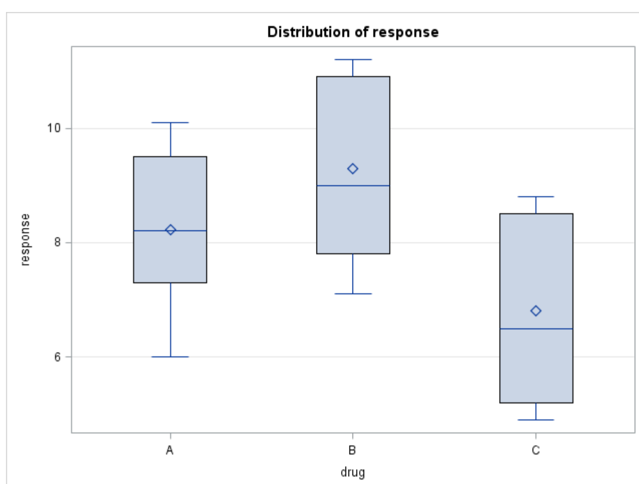
- Variance of F-distribution:

$$\frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)} \text{ for } \nu_2 > 4$$

- We will focus on probabilities (p-values) under the F-distribution, and use technology to find these for us

Drug Example:

- Research question:
- Population:
- Parameter:
- Hypotheses:
- Type of test:
- Conditions:
    - Random samples:
    - Independent samples:
    - Normal populations:
        - Test valid as long as populations are at least reasonably symmetric
        - Can think about context (e.g. distribution of response for all individuals who take Drug A)
        - Can also check histograms, qq-plots, or side-by-side boxplots for evidence of skew (need to check for each group separately)
    - Equal population variances:
        - Check side-by-side boxplots; width of boxes or range (min to max) should be similar



Drug Example (continued):

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	21.98147368	10.99073684	4.19	0.0345
Error	16	41.98800000	2.62425000		
Corrected Total	18	63.96947368			

- Test statistic:

- Null distribution:

- P-value:

- Decision:

- Conclusion:

Lecture 8.3: ANOVA—Multiple Comparisons

- You have enough evidence to say that the group means are significantly different
- But which groups are different from each other?
- Consider all possible pairs of means (combinations of two groups)
    - Compute a **confidence interval** for the difference between the two means and see if 0 falls in the interval or not
    - Perform a **test of hypotheses** to assess if the two means differ significantly
- **Tukey's procedure** is common multiple comparisons procedure using confidence intervals

Drug Example:

Comparisons significant at the 0.05 level are indicated by *.**

drug Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
B - A	1.0800	-1.3676	3.5276	
B - C	2.5000	0.2657	4.7343	***
A - B	-1.0800	-3.5276	1.3676	
A - C	1.4200	-1.0276	3.8676	
C - B	-2.5000	-4.7343	-0.2657	***
C - A	-1.4200	-3.8676	1.0276	

Why not just do t-tests to compare the means in the first place?

- Number of tests becomes large if number of groups is large

- Chance of a type I error

Lecture 8.4: Introduction to Chi-square Tests

3 Types of Chi-square Tests

1. ____: Test to assess if a particular discrete model is a good fitting model for a categorical variable in a single population
    - Categorical response variable; no explanatory variable
    - Extension of 1-sample z-test for a population proportion—considers all categories rather than just focusing on one

2. ____: Test to assess if the distribution for one categorical variable is the same for two or more populations
    - Categorical response variable; categorical explanatory variable that is used to define two or more populations
    - Extension of 2-sample z-test; similar in spirit to ANOVA, but for proportions rather than means

3. ____: Test to assess if there is a relationship between two categorical variables in a single population
    - Categorical response variable; categorical explanatory variable
    - Later we will learn about simple linear regression, which has a similar goal but for continuous variables rather than categorical

- All based on

Example: For each description, determine the type of Chi-square test that should be used.

a. Are you satisfied with your overall appearance? A random sample of 150 women were surveyed. Their answer to this question (Yes or No) was recorded along with their age category (1 = under 30, 2 = 30 to 50, and 3 = over 50). The goal is to test if there is a significant relationship between age and satisfaction with appearance.

b. A random sample of 928 adults were asked to "Name the one place you would want to go for vacation if you had the time and the money." Answers to this question are categorized as: the Caribbean, Europe, Hawaii, and Other. A travel agency would like to assess if the distribution of desired vacation place has changed a model that had been established previously.

c. An article had the headline "For adults, chicken pox vaccine may stop shingles". A clinical trial was conducted in which 420 subjects were randomly assigned to receive the chicken pox vaccine or a placebo vaccine. The side effect of swelling around the injection site was categorized as: major, minor, or no swelling. The goal is to determine if the distribution of swelling is the same for those who received the vaccine and those who received the placebo.

Basic Idea

- Data consists of observed counts (n_i)
- Compute expected counts under the null hypothesis (E_i)
- Compare observed and expected counts using the Chi-square test statistic:

$$\chi^2 = \sum_i \left[\frac{(n_i - E_i)^2}{E_i}\right]$$

  - Interpretation: Measures how close observed counts are to expected counts under H_0
  - If this distance is large, we have support for the alternative H_A
  - Follows the Chi-square distribution

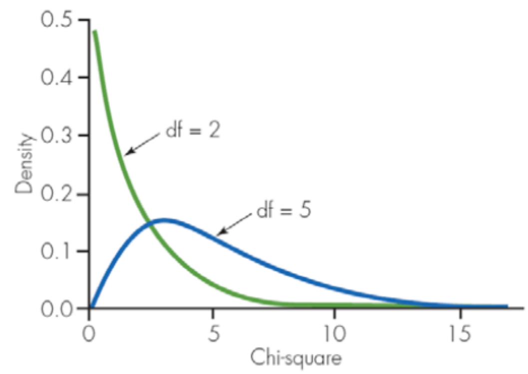
Chi-Square Distribution

- Models the sum of squared random variables that follow the Normal distribution
- Most commonly used for statistical inference, for example when conducting inference for variances, multiple proportions, or counts in contingency tables

- PDF:

$$\frac{1}{2^{\nu/2}\Gamma(\nu/2)} y^{(\nu/2)-1} e^{-y/2} \qquad \text{where } \Gamma(y) = \int_0^\infty y^{x-1} e^{-x} dx$$

- Generally right skewed (recall: values of squared variables cannot be negative)
- Depends on one parameter:



- If you know ν, you know everything there is to know about that Chi-square distribution

- Mean of Chi-square distribution:

- Variance of Chi-square distribution:

- We will focus on probabilities (p-values) under the Chi-square distribution, and use technology to find these for us

Lecture 8.5: Three Types of Chi-square Tests

Test of Goodness of Fit

- When to use: To assess if a particular discrete model is a good fitting model for a categorical variable in a single population

- Steps 1, 6, and 7 are the same as always
- Hypotheses:

- Conditions:

- Test statistic:

$$\chi^2 = \sum_i \left[\frac{(n_i - E_i)^2}{E_i}\right]$$

  - Where:

- Null distribution:

- p-value always represents proportion of null distribution that is greater than or equal to the test statistic
  - No need to be concerned about two-sided vs. one-sided

Example: A travel agency would like to assess if the distribution of desired vacation place has changed from the model of 3 years ago. A random sample of 928 adults were asked to "Name the one place you would want to go for vacation if you had the time and the money." The table below displays the model for the distribution of desired vacation place 3 years ago and the observed results based on the recent poll. Based on the table, can you say that vacation preferences are different today than they were 3 years ago?

	Hawaii	**Europe**	**Caribbean**	**Other**
Percent 3 years ago	12%	40%	18%	30%
Count from survey	119	394	130	285
Percent from survey	13%	42%	14%	31%

a. State the hypotheses using the appropriate statistical notation.

b. Calculate the Chi-square test statistic based on this data.

c. The p-value is 0.02. Write a sentence interpreting this value.

Test of Homogeneity

- When to use: To assess if the distribution for one categorical variable is the same for two or more populations

- Steps 1, 6, and 7 are the same as always
- Hypotheses:

- Conditions:
  - Random sample
  - At least 80% of the expected counts are greater than 5 and none are less than 1

- Test statistic:

$$\chi^2 = \sum_{i,j} \left[\frac{(n_{ij} - E_{ij})^2}{E_{ij}} \right]$$

  - Where:

- Null distribution:

- p-value always represents proportion of null distribution that is greater than or equal to the test statistic
  - No need to be concerned about two-sided vs. one-sided

Example: An article had the headline "For adults, chicken pox vaccine may stop shingles". A clinical trial was conducted in which 420 subjects were randomly assigned to receive the chicken pox vaccine or a placebo vaccine. Some side effects of interest were swelling and rash around the injection site. Results for the swelling side effect are shown below. Use this data to test if the distribution of swelling (major, minor, or no swelling) is the same for those who received the vaccine and those who received the placebo. [Note: Output continues on next page; a third page is provided for additional space.]

The FREQ Procedure

Table of trt by swelling				
trt	**swelling**			
	major	**minor**	**none**	**Total**
placebo	16	32	142	190
vaccine	54	42	134	230
Total	70	74	276	420

Example (Side Effects; continued)

Statistics for Table of trt by swelling

Statistic	DF	Value	Prob
Chi-Square	2	18.5707	<.0001
Likelihood Ratio Chi-Square	2	19.5565	<.0001
Mantel-Haenszel Chi-Square	1	17.6963	<.0001
Phi Coefficient		0.2103	
Contingency Coefficient		0.2058	
Cramer's V		0.2103	

Example (Side Effects; continued)

Test of Independence

- When to use: To assess if there is an association between two categorical variables in a single population

- Steps 1, 6, and 7 are the same as always
- Hypotheses:

- Conditions:
    - Random sample
    - At least 80% of the expected counts are greater than 5 and none are less than 1

- Test statistic:

$$\chi^2 = \sum_{i,j} \left[\frac{(n_{ij} - E_{ij})^2}{E_{ij}} \right]$$

    - Where:

- Null distribution:

- p-value always represents proportion of null distribution that is greater than or equal to the test statistic
    - No need to be concerned about two-sided vs. one-sided
    - Exact value found using software

Example: Are you satisfied with your overall appearance? A random sample of 150 women were surveyed. Their answer to this question (Yes or No) was recorded along with their age category (1 = under 30, 2 = 30 to 50, and 3 = over 50). SAS Output of the data is presented below. Use this output to test if there is a significant relationship between age and satisfaction with appearance. [Note: Output continues on next page; a third page is provided for additional space.]

Frequency
Expected

Table of satisfied by age				
satisfied	**age**			
	1	**2**	**3**	**Total**
No	10 15.36	29 18.88	9 13.76	48
Yes	38 32.64	30 40.12	34 29.24	102
Total	48	59	43	150

Example (Appearance Satisfaction; continued):

Statistics for Table of satisfied by age

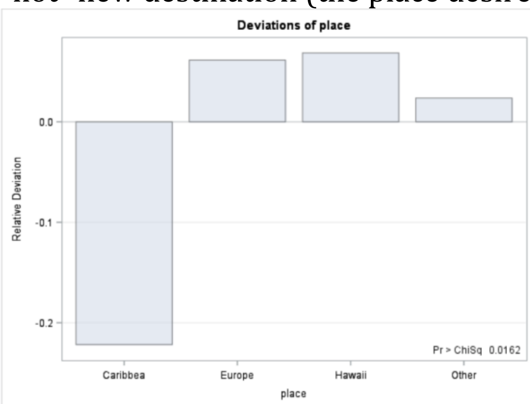
Statistic	DF	Value	Prob
Chi-Square	2	13.1493	0.0014
Likelihood Ratio Chi-Square	2	13.0387	0.0015
Mantel-Haenszel Chi-Square	1	0.0181	0.8930
Phi Coefficient		0.2961	
Contingency Coefficient		0.2839	
Cramer's V		0.2961	

Example (Appearance Satisfaction; continued):

Lecture 8.6: Chi-square—Multiple Comparisons

- You have enough evidence to say that the proportions/cell counts are significantly different from what was expected under the null hypothesis, but which are different?
- Consider the **deviations** (or **residuals**)—the difference between observed and expected counts—for which groups is n_i very different from E_i?
- Note: this is not all possible pairwise comparisons (as with ANOVA); it is looking more globally for differences or trends

Example (Desired Vacation Place): p-value = 0.016; there is evidence that the distribution of desired vacation destinations is different than it was 3 years ago…so which place is the "hot" new destination (the place desired more than expected)? Which place fell out of favor?



Example (Side Effects):

Frequency / Expected

Table of trt by swelling				
	swelling			
trt	major	minor	none	Total
placebo	16 31.667	32 33.476	142 124.86	190
vaccine	54 38.333	42 40.524	134 151.14	230
Total	70	74	276	420

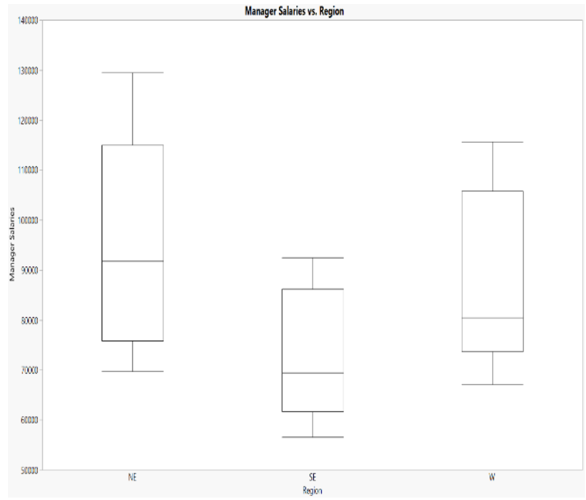
Example (Appearance Satisfaction):

Frequency / Expected / Deviation / Cell Chi-Square

Table of satisfied by age				
	age			
satisfied	1	2	3	Total
No	10 15.36 -5.36 1.8704	29 18.88 10.12 5.4245	9 13.76 -4.76 1.6466	48
Yes	38 32.64 5.36 0.8802	30 40.12 -10.12 2.5527	34 29.24 4.76 0.7749	102
Total	48	59	43	150

Lecture 8.7: Additional Examples

Additional Example 1: Is there a difference in the average annual salaries for marketing managers in different regions on the United States? A sample of 24 states was randomly selected from each of three regions (8 states from the Northeast, 8 from the Southeast and 8 from the West) and mean annual salaries for marketing managers were retrieved from the U.S. Bureau of Labor Statistics.



a. Comment on the plausibility of the assumptions for conducting an ANVOA.

b. Fill in the missing values in the ANOVA table.

Source	Degrees of Freedom	Sum of Squares	Mean Squares	F-Statistic	p-value
Model			1.1484e+9		0.0466
Error					
Total		9439543504			

c. Would it be reasonable to conduct a multiple comparisons procedure? State yes or no and explain your reasoning. If it is reasonable, use the output below to determine which groups are different. If it's not, explain why not.

salary Comparison	Difference in Means	Simultaneous 95% Confidence Limits	
NE – SE	23188.75	1194.5	45182.97
NE – W	9021.25	-12973.0	31015.47
W – SE	14167.50	-8464.4	36799.36

For each example below…

- Determine and conduct the appropriate type of Chi-square test
    - p-value is provided
    - Need to fill in all other steps to the test, including interpreting and drawing a well-labeled picture of p-value
- When appropriate, conduct multiple comparisons to identify groups that appear to be different than expected.

Additional Example 2: We would like to learn if ice cream preference (vanilla, chocolate, or strawberry) is the same for preschool boys and girls. [p-value = 0.4211]

	Boys	Girls	Total
Vanilla	25	26	51
Chocolate	30	23	53
Strawberry	20	26	46
Total	75	75	150

Additional Example 3: We are interested in learning about traffic patterns at a busy exit on a toll road. At this exit there are four toll booths, and we would like to know if they are used equally often. [p-value = 0.6962]

	Booth 1	Booth 2	Booth 3	Booth 4
Observed # cars	26	20	28	26

Additional Example 4: We are interested in learning if there a relationship between smoking habits (non-smoker, moderate smoker, or heavy smoker) and whether or not workers at a particular factory experience hypertension. [p-value = 0.0007]

		Smoking Status			
		Non	Moderate	Heavy	Total
Hypertension Status	Yes	21	36	30	87
	No	48	26	19	93
	Total	69	62	49	180