Chi-Square & ANOVA

Assignment 2: Chi-Square & ANOVA

ALY6015

Prepared by: Heejae Roh
Presented to: Professor Paromita Guha
Date: Jan 22th, 2023

PART 0. INTRODUCTION

In this project I learn how to use chi-square and ANOVA testing to solve the problems. To study chi-square and ANOVA test, in examples, I usually assume that all assumptions are met. However, for later use in real-life problem, I think about the assumptions and conditions about chi-square and ANOVA. I focus on the steps that I have to take to execute chi-square and ANOVA testing. There are 5 chi-square, 3 one-way ANOVA, and 2 two-way ANOVA testing. In each testing, I concentrate on differences and guidelines of each method in R program. In addition to this, I try to understand the structure of table that I have to create for each method.

PART 1. ANALYSIS

PART 1-1. chi-square

Understanding:

1. Chi-square can be used for tests concerning frequency distributions. The chi-square distribution can be used to test the homogeneity of proportions. For example, is the proportion of high school seniors who attend college immediately after graduating the same for the northern, southern, eastern, and western parts of the United States (Bluman, 2017).
2. The assumptions of the Chi-square include: The data in the cells should be frequencies, or counts of cases rather than percentages or some other transformation of the data. The levels (or categories) of the variables are mutually exclusive (Biochem, 2013).
3. The four assumptions of a chi-square Test (ZACH, 2021)
    - Assumption 1: Both variables are categorical.
    - Assumption 2: All observations are independent.
    - Assumption 3: Cells in the contingency table are mutually exclusive.
    - Assumption 4: Expected value of cells should be 5 or greater in at least 80% of cells.

Section 11-1. 6 Blood Types

Step 0 Finding a Key-Metrics.

This is about blood type and want to compare frequency distributions

A medical researcher wishes to see if hospital patients in a large hospital have the same blood type distribution as those in the general population. The distribution for the general population is as follows: type A, 20%; type B, 28%; type O, 36%; and type AB = 16%. He selects a random sample of 50 patients and finds the following: 12 have type A blood, 8 have type B, 24 have type O, and 6 have type AB blood.

At α = 0.10, can it be concluded that the distribution is the same as that of the general population?

Frequency	Type A	Type B	Type O	Type AB	Total
Observed	12	8	24	6	50
Expected	20%	28%	36%	16%	100%

Step 1 Hypothesis.

Null: The proportion of people in each blood type is as follows: type A, 20%; type B, 28%; type O, 36%; and type AB, 16% (Claim)

- H0: Type A=0.2; Type B=0.28; Type O=0.36; and type AB = 0.16

Alternative: The distribution is not the same as stated in the null hypothesis.

Step 2 Find the critical value.

- The p-value is α = 0.10

Step 3 Compute the test value.

Chi-squared test for given probabilities		
data: observed [1] 12 8 24 6		
X-squared = 5.4714	df = 3	p-value = 0.1404

Step 4 Make the decision.

- There is not enough evidence to reject null hypothesis since 0.1404 (p-value) > 0.10

Step 5 Summarize the results.

- There is not enough evidence to reject the claim that the proportion of people in each blood type is as follows: type A, 20%; type B, 28%; type O, 36%; and type AB = 16%

Section 11-2. 8 On-Time Performance by Airlines

Step 0 Finding a Key-Metrics.

This is about frequency distribution about on-time performance by airlines

Records of 200 randomly selected flights for a major airline company showed that 125 planes were on time; 40 were delayed because of weather, 10 because of a National Aviation System delay, and the rest because of arriving late.

Action	% of Time	Sample
On time	70.8	125
National Aviation System delay	8.2	10
Aircraft arriving late	9.0	25
Other (because of weather and other conditions)	12.0	40
Total	100	200

At α = 0.05, do these results differ from the government's statistics?

Step 1 Hypothesis.

Null: The proportion of airlines on-time performance is as follows: On time, 70.8%; National Aviation System delay, 7.2%; Aircraft arriving late, 9.0%; and Other (because of weather and other conditions), 12.0% (Claim)

- H0: On time=0.708; National Aviation System delay=0.082; Aircraft arriving late=0.09; and Other (because of weather and other conditions) = 12.0

Alternative: The distribution is not the same as stated in the null hypothesis.

Step 2 Find the critical value.

- The significance level of degrees of freedom 3 in α = 0.05 is 7.815

Step 3 Compute the test value.

Chi-squared test for given probabilities		
data: observed2 [1] 125 10 25 40		
X-squared = 17.832	df = 3	p-value=0.0004763

Step 4 Make the decision.

- There is enough evidence to reject null hypothesis since 17.832 (X-squared) > 7.815 (by table)

Step 5 Summarize the results.

- There is enough evidence to reject the claim that the proportion of airlines on-time performance is: On time, 70.8%; National Aviation System delay, 7.2%; Aircraft arriving late, 9.0%; and Other (because of weather and other conditions), 12.0%

Section 11-2. 8 Ethnicity and Movie Admission

Step 0 Finding a Key-Metrics.

This is about frequency distribution about movie admissions related to ethinicity

Are movie admissions related to ethnicity? A 2014 study indicated the following numbers of admissions (in thousands) for two different years.

	Caucasian	Hispanic	African American	Other
2013	724	335	174	107
2014	370	292	152	140

At the 0.05 level of significance, can it be concluded that movie attendance by year was dependent upon ethnicity?

Step 1 Hypothesis.

Null: The movie attendance by year is independent of the ethnicity (claim).
Alternative: The movie attendance by year is dependent of the ethnicity.

Step 2 Find the critical value.

- The p-value is $\alpha = 0.05$

Step 3 Compute the test value.

Chi-squared test for given probabilities		
data: observed2 [1] 125 10 25 40		
X-squared = 60.144	df = 3	p-value=5.478e-13

Step 4 Make the decision.

- There is enough evidence to reject null hypothesis since 05.478e-13 (p-value) < 0.05

Step 5 Summarize the results.

- There is enough evidence to reject the claim that the movie attendance by year is independent of the ethnicity

Section 11-2. 10 Women in the Military

Step 0 Finding a Key-Metrics.

This is about frequency distribution of a relationship between rank and branch of the Armed Forces in women group

Action	Officers	Enlisted
Army	10,791	62,491
Navy	7,816	42,750
Marine Corps	932	9,525
Air Force	11,819	54,344

At α = 0.05, is there sufficient evidence to conclude that a relationship exists between rank and branch of the Armed Forces?

Step 1 Hypothesis.

Null: The rank by Armed Force is independent of branch in women group.
Alternative: The rank is dependent of branch of the Armed Force in women group (claim).

Step 2 Find the critical value.

- The p-value is α = 0.05

Step 3 Compute the test value.

Chi-squared test for given probabilities		
data: observed2 [1] 125 10 25 40		
X-squared = 654.27	df = 3	p-value < 2.2e-16

Step 4 Make the decision.

- There is enough evidence to reject null hypothesis since 2.2e-16 (p-value) < 0.05

Step 5 Summarize the results.

- There is not enough evidence to reject the claim that the rank by Armed Force is independent of branch in women group

PART 1-2. one-way ANOVA

Understanding:

1. When an F test is used to test a hypothesis concerning the means of three or more populations, the technique is called analysis of variance. The one-way analysis of variance test is used to test the equality of three or more means using sample variances (Bluman, 2017).
2. Use a one-way ANOVA when you have collected data about one categorical independent variable and one quantitative dependent variable. The independent variable should have at least three levels (Rebecca, 2020)
3. There are three primary assumptions in ANOVA: (Penn state college, n.d)
    - Assumption 1: The responses for each factor level have a normal population distribution.
    - Assumption 2: These distributions have the same variance.
    - Assumption 3: The data are independent.

Scheffé test understanding:

1. The Scheffé test can be used to determine whether individual means differ, or whether an average one group of means differs from the average of another group of means (Will, 2021).
2. The Scheffé test can be used with unequal sample sizes between groups. The Scheffé test provides a more sensitive test of non-pairwise comparisons than some other post hoc testing procedures (StatTrek, n.d.).

Difference between the Scheffe's test and Tukey's test:

Generally, Tukey and Scheffé tests are more conservative. They find it harder to see differences and generally give the same result. In relation to the differences (Ouorou, 2019):

1. In pairwise comparisons, Tukey test is based on studentized range distribution while Scheffe is based in F distribution
2. Tukey's test is very rigorous, controlling the type I error very well, but favors the type II error
3. The Scheffe test allows comparing any contrast between means and allows different number of observations per treatment (Ouorou, 2019)

Section 12-1. 8 Sodium Contents of Foods

Step 0 Finding a Key-Metrics.

This is about three independent variables (type of food) related to sodium, which is one quantitative dependent variable

Condiments	Cereals	Desserts
270	260	100
130	220	180
230	290	250
180	290	250
80	200	300
70	320	360
200	140	300
		160

At the 0.05 level of significance, is there sufficient evidence to conclude that a difference in mean sodium amounts exists among condiments, cereals, and desserts?

Step 1 Hypothesis.

Null: There is no difference in mean of sodium amount among condiments, cereals, and desserts.
H0: μ1 = μ2 = μ3
Alternative: At least one mean is different from the others (claim).

Step 2 Find the critical value.

- The p-value is $\alpha = 0.05$

Step 3 Compute the test value.

Summary(anova)					
	Df	Sum Sq	Mean Sq	F Value	Pr(>F)
food	2	27544	13772	2.399	0.118
Residuals	19	109093	5742		

Step 4 Make the decision.

- There is not enough evidence to reject null hypothesis since 0.118 (p-value) > 0.05

Step 5 Summarize the results.

- There is not enough evidence to reject the claim that there is no difference in mean of sodium amount among condiments, cereals, and desserts.

Section 12-2. 10 Sales for Leading Company

Step 0 Finding a Key-Metrics.

This is about one quantitative variable which is sales of three independent variables as companies

Cereal	Chocholate Candy	Coffee
578	311	261
320	106	185
264	109	302
249	125	689
237	173	

The sales in millions of dollars for a year of a sample of leading companies are shown. At α = 0.01, is there a significant difference in the means?

Step 1 Hypothesis.

Null: There is no difference in sales of cereal, chocolate candy, and coffee company.
H0: μ1 = μ2 = μ3
Alternative: At least one mean is different from the others (claim).

Step 2 Find the critical value.

- Since k= 3, N= 14, and α = 0.01,
- d.f.N= k-1 = 3-1 = 2,
- d.f.D= N-k = 18-3 = 15
- The critical value is 6.36

Step 3 Compute the test value.

Summary(anova)					
	Df	Sum Sq	Mean Sq	F Value	Pr(>F)
company	2	103770	51885	2.172	0.16
Residuals	11	262795	23890		

Step 4 Make the decision.

- There is not enough evidence to reject null hypothesis since 2.172 (F-value) < 6.36

Step 5 Summarize the results.

- There is not enough evidence to reject the claim that there is no difference in sales of cereal, chocolate candy, and coffee company.

Section 12-2. 12 Per-Pupil Expenditures

Step 0 Finding a Key-Metrics.

This is about one quantitative which is expenditures per pupil for three categorical variables as three sections of the country

Eastern third	Middle third	Western third
4946	6149	5282
5953	7451	8605
6202	6000	6528
7243	6479	6911
6113		

The expenditures (in dollars) per pupil for states in three sections of the country are listed. Using $\alpha = 0.05$, can you conclude that there is a difference in means?

Step 1 Hypothesis.

Null: There is no difference in per pupil expenditures for states in three sections of the country.
H0: μ1 = μ2 = μ3
Alternative: At least one mean is different from the others (claim).

Step 2 Find the critical value.

- The p-value is $\alpha = 0.05$

Step 3 Compute the test value.

Summary(anova)					
	Df	Sum Sq	Mean Sq	F Value	Pr(>F)
states	2	1244588	622294	0.649	0.543
Residuals	10	9591145	959114		

Step 4 Make the decision.

- There is not enough evidence to reject null hypothesis since 0.543 (p-value) > 0.05

Step 5 Summarize the results.

- There is not enough evidence to reject the claim that there is no difference in per pupil expenditures for states in three sections of the country.

PART 1-3. two-way ANOVA

Understanding:

1. The two-way ANOVA is an extension of the one-way analysis of variance; it involves two independent variables. The independent variables are also called factors. The two-way ANOVA technique is used to determine if there is a significant difference in the main effects or interaction (Bluman, 2017).
2. A two-way ANOVA is used to estimate how the mean of a quantitative variable changes according to the levels of two categorical variables. Use a two-way ANOVA when you want to know how two independent variables, in combination, affect a dependent variable.
3. Assumptions of the two-way ANOVA (Rebecca, 2022)
   - Assumption 1: Homogeneity of variance (among all group)
   - Assumption 2: Independence of observations (not dependent on one another)
   - Assumption 3: Normally-distributed dependent variable

Tukey Test:

1. The Tukey test can also be used after the analysis of variance has been completed to make pairwise comparisons between means when the groups have the same sample size. The symbol for the test value in the Tukey test is q (Bluman, 2017).
2. The Tukey's honestly significant difference test (Tukey's HSD) is used to test differences among sample means for significance. The Tukey's HSD tests all pairwise differences while controlling the probability of making one or more Type I errors (David, n.d.).

Section 12-3. 10 increasing Plant Growth

Step 0 Finding a Key-Metrics.

This is about plant growth according to the levels of plant food & light, which is two categorical variables.

	Grow – light 1	Grow – light 2
Plant food A	9.2, 9.4, 8.9	8.5, 9.2, 8.9
Plant food B	7.1, 7.2, 8.5	5.5, 5.8, 7.6

Can an interaction between the two factors (plant food A/B, and Grow-light ½) be concluded? Is there a difference in mean growth with respect to light? With respect to plant food? Use $\alpha = 0.05$.

Step 1 Hypothesis.

1. **The hypotheses for interaction are stated as follows.**

Null Hypothesis: There is no interaction effect between type of plant food used and type of light used on plant growth.
Alternative Hypothesis: There is an interaction effect between type of plant food used and type of light used on plant growth.

2. **The hypotheses regarding plant food are stated as follows.**

Null Hypothesis: There is no difference in means of heights of plants grown using different foods.
Alternative Hypothesis: There is a difference in means of heights of plants grown using different foods.

3. **The hypotheses regarding light are stated as follows:**

Null Hypothesis: There is no difference in means of heights of plants grown in different light.
Alternative Hypothesis: There is a difference in means of heights of plants grown in different light.

Step 2 Find the critical value.

- The p-value is $\alpha = 0.05$
- Factor A (Plant food): d.f.N. = 2 – 1 = 1, Factor B(light): d.f.N. = 2 – 1 = 1
- Interaction (A x B): (2 – 1)(2 – 1)= 1
- Within (error): d.f.D. = ab(n-1) = 2*2(3 – 1) = 8

Step 3 Compute the test value.

Summary(anova)

	Df	Sum Sq	Mean Sq	F Value	Pr(>F)
light	1	1.920	1.920	3.681	0.09133 .
food	1	12.813	12.813	24.562	0.00111 **
light:food	1	0.750	0.750	1.438	0.26482
Residuals	8	4.173	0.522		

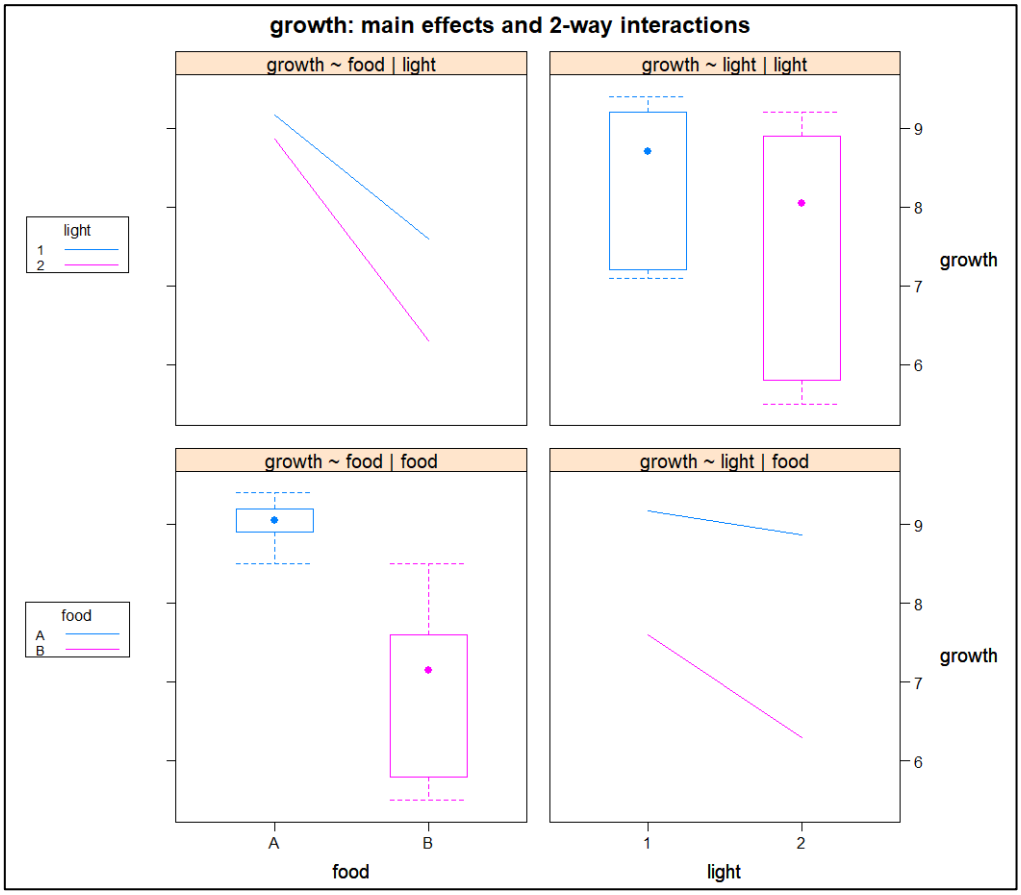
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Step 4 Make the decision.

- Hypothesis 1: Interaction, there is no interaction effect between type of plant food used and type of light used on plant growth. since 0.264 (p-value) > 0.05
- Hypothesis 3: Light, there is no difference in means of heights of plants grown in different light. since 0.09133 (p-value) > 0.05
- Hypothesis 2: Plant food, there is a difference in means of heights of plants grown using different foods. since 0.00111 (p-value) < 0.05

Step 5 Summarize the results.

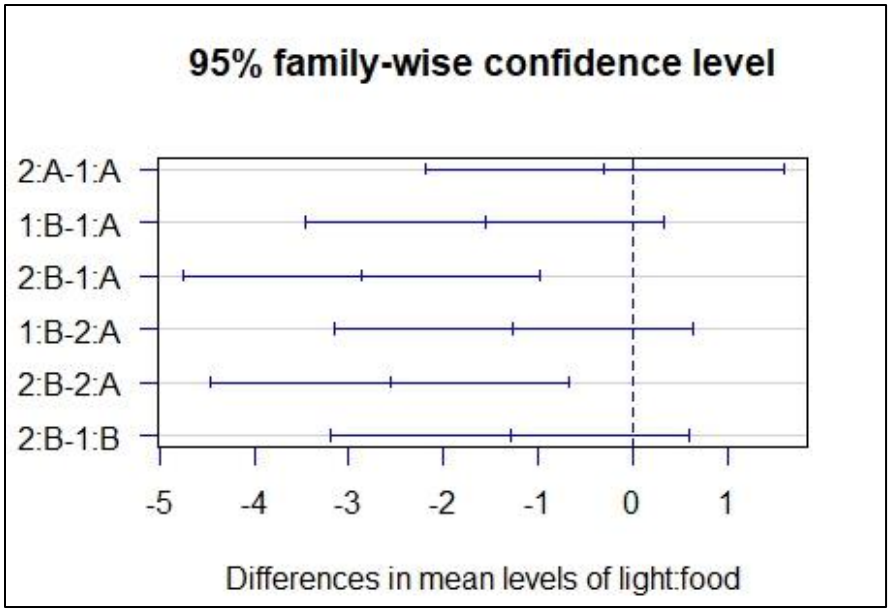
- Since the interaction hypothesis and null hypothesis about light are not rejected, it can be concluded that the plant food difference does affect the growth of plant



Interpretation:

1. There is a difference of the growth by the type of food (left-below)
2. There is no difference of the growth by the type of light (right-above)

Post-hoc, Tukey test



Interpretation:

1. 2 (light) & B (food) – 1 (light) & A (food) is significant
2. 2 (light) & B (food) – 2 (light) & A (food) is significant
   Since those does not cover 0 in there 95% CI

PART 2. ANALYSIS with BASEBALL DATASET

PART 2-1. SUMMARY of EDA

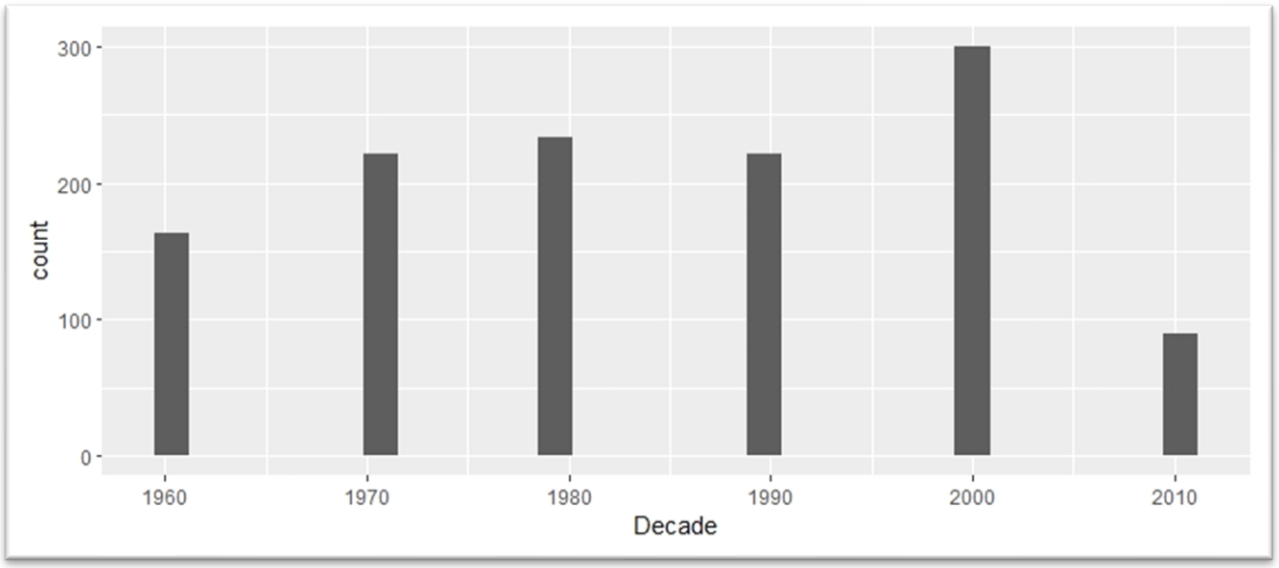
Headtail of Data 1

	Team	League	Year	RS	RA	W	OBP	SLG	BA	Playoffs	G	Decade	…
1	ARI	NL	2012	734	688	81	0.33	0.42	0.26	0	162	2010	…
2	ATL	NL	2012	700	600	94	0.32	0.39	0.25	1	162	2010	…
3	BAL	AL	2012	712	705	93	0.31	0.42	0.25	1	162	2010	…
…													
1230	SFG	NL	1962	878	690	103	0.34	0.44	0.28	1	165	1960	…
1231	STL	NL	1962	774	664	84	0.34	0.39	0.27	0	163	1960	…
1232	WSA	AL	1962	599	716	60	0.31	0.37	0.25	0	162	1960	…

Descriptive Analysis of primary data

	n	mean	sd	median	trimmed	mad	min	max	range	skew
Team*	1232	18.93	10.61	20.00	18.76	13.34	1.00	39.00	38.00	0.06
League*	1232	1.5	0.5	1.5	1.5	0.74	1.00	2.00	1.00	0.00
Year	1232	1988.96	14.82	1989	1989.32	19.27	1962	2012	50.00	-0.15
RS	1232	715.08	91.53	711.00	713.34	90.44	463	1009	546	0.17
RA	1232	715.08	93.08	709	712.44	91.92	472	1103	631	0.30
W	1232	80.90	11.46	81	81.12	11.86	40	116	76	-0.18
…						…				
Decade	1232	1984.4	15.27	1980	1984.58	14.83	1960	2010	50	-0.09

Histogram of decades



Explanation:

1. This dataset has results of each baseball team in major league from 1962 to 2012
2. I made Decade table from Year by deleting the Reminders
3. This is the table of wins by decade

Table						
Decade	1960	1970	1980	1990	2000	2010
wins	13267	17934	18926	17972	24286	7289

Trends and interesting stuffs about dataset:

1. There are a lot of information about each team by year in MLB
2. There is RankSeason which mean rank but there are many <NA> in there. Playoffs is binary which is playoffs yes or no
3. There are many variables that can made by this dataset, for example, win rate can be made with G (number of game) and W (number of win)
4. In decade 2010, there is just 2010-2012, so the number of games and wins are smaller than other decades
5. In decade 1960, there is not 1960-1962, because the dataset does not cover. Therefore, the number of games and wins are smaller than other decades except 2010

PART 2-2. chi-square with BASEBALL DATA

Wins by decade

Step 0 Finding a Key-Metrics.

This is about numbers of wins of all team by decade

If there is a difference in the number of wins by decade? Use α = 0.05.

Step 1 Hypothesis.

Null Hypothesis: There is no difference in number of wins by decade (claim).
Alternative Hypothesis: There is difference in number of wins by decade.

Step 2 Find the critical value.

- The p-value is α = 0.05

Step 3 Compute the test value.

Chi-squared test for given probabilities						
data: observed	[1] 13267	17934	18926	17972	24286	7289
X-squared = 9989.5		df = 5		p-value < 2.2e-16		

Step 4 Make the decision.

- There is difference in the number of wins by decade.
  since 2.2e-16 (p-value) < 0.05

Step 5 Summarize the results.

- Since the null hypothesis is rejected, it can be concluded that there is difference in the number of wins by decade.

PART 3. ANALYSIS with CROP DATASET

PART 3-1. SUMMARY of EDA

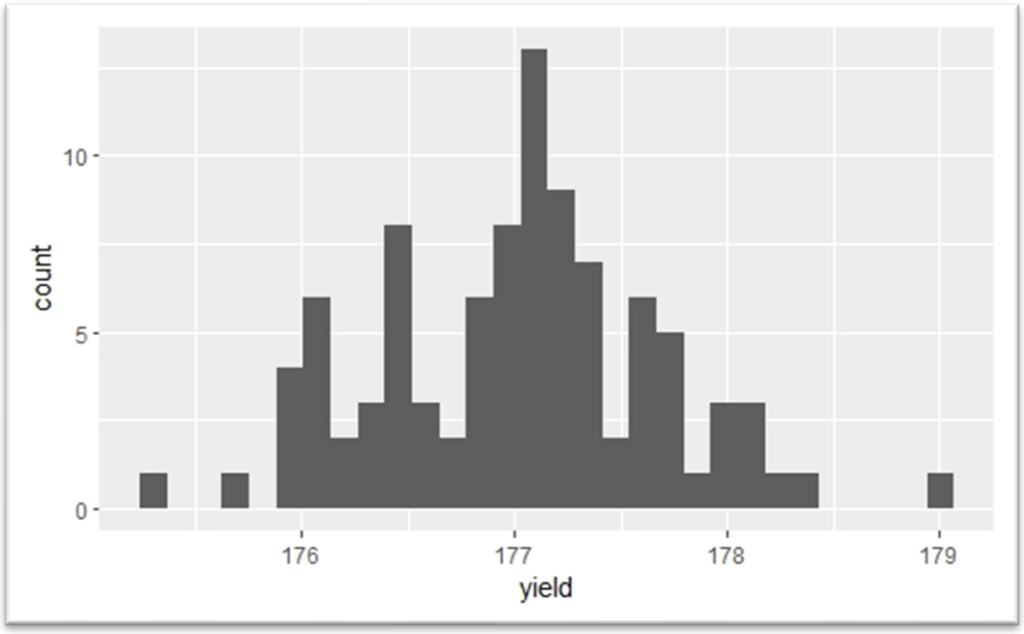
Headtail of Data 1

	density	block	fertilizer	yield
1	1	1	1	177.23
2	2	2	1	177.55
3	1	3	1	176.41
…				
94	1	1	3	178.14
95	1	3	3	177.69
96	2	4	3	177.12

Descriptive Analysis of primary data

	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
density	96	1.5	0.5	1.5	1.5	0.74	1.00	2.00	1.0	0.00	-2.02	0.05
block	96	2.5	1.12	2.5	2.5	1.48	1.00	4.00	3.0	0.00	-1.39	0.11
fertilizer	96	2.00	0.82	2.00	2.00	1.48	1.00	3.00	2.0	0.00	-1.53	0.08
yield	96	177.02	0.66	177.06	177.01	0.68	175.36	179.06	3.7	0.11	0.01	0.07

Histogram of yield



Explanation:

1. This dataset has dataset about crop yield and factors that might affect the crop yield
2. There are two density type, 4 block, and 3 fertilizer type
3. Above is the histogram of yield

Trends and interesting stuffs about dataset:

1. The yield's mean is 1772.02 and median is 177.06
2. In this dataset, I can compare the 2 density, 4 block, and 3 fertilizer type
3. The yield is only continuous variable and other variables are all categorized

PART 3-2. two-way ANOVA with CROP DATA

increasing Plant Growth

Step 0 Finding a Key-Metrics.

This is about density of crops. The dependent variables are yield. Fertilizer and density are the independent variables.

Perform a Two-way ANOVA test using yield as the dependent variable and fertilizer and density as the independent variables. Is there reason to believe that fertilizer and density have an impact on yield? Use $\alpha = 0.05$.

Step 1 Hypothesis.

1. **The hypotheses for interaction are stated as follows.**

Null Hypothesis: There is no interaction effect between density and fertilizer on yield.
Alternative Hypothesis: There is an interaction effect between density and fertilizer on yield.

2. **The hypotheses regarding density are stated as follows.**

Null Hypothesis: There is no difference in means of yield using different density.
Alternative Hypothesis: There is a difference in means of yield using different density.

3. **The hypotheses regarding fertilizer stated as follows:**

Null Hypothesis: There is no difference in means of yield using different fertilizer.
Alternative Hypothesis: There is a difference in means of yield using different fertilizer.

Step 2 Find the critical value.

- The p-value is $\alpha = 0.05$
- Factor A: d.f.N. (density) = 2 – 1 = 1, Factor B (fertilizer): d.f.N. = 3 – 1 = 2
- Interaction (A x B): (2 – 1)(3 – 1)= 2
- Within (error): d.f.D. = ab(n-1) = 2*3(16 – 1) = 90

Step 3 Compute the test value.

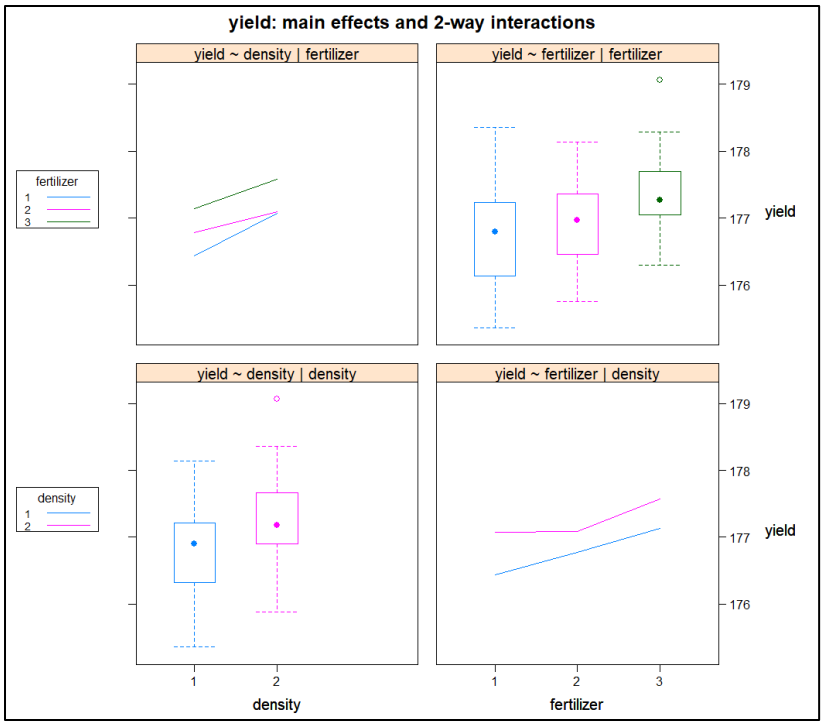
Summary(anova)					
	Df	Sum Sq	Mean Sq	F Value	Pr(>F)
density	1	5.122	5.122	15.195	0.000186 ***
fertilizer	2	6.068	3.034	9.001	0.000273 ***
density:fertilizer	2	0.428	0.214	0.635	0.532500
Residuals	90	30.337	0.337		
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

Step 4 Make the decision.

- Hypothesis 1: Interaction, there is no interaction effect between density and fertilizer on yield. since 0.5325 (p-value) > 0.05
- Hypothesis 2: Density, there is difference in means of yield in different density. since 0.000186 (p-value) < 0.05
- Hypothesis 3: Fertilizer, there is a difference in means of yield in different fertilizer. since 0.000273 (p-value) < 0.05

Step 5 Summarize the results.

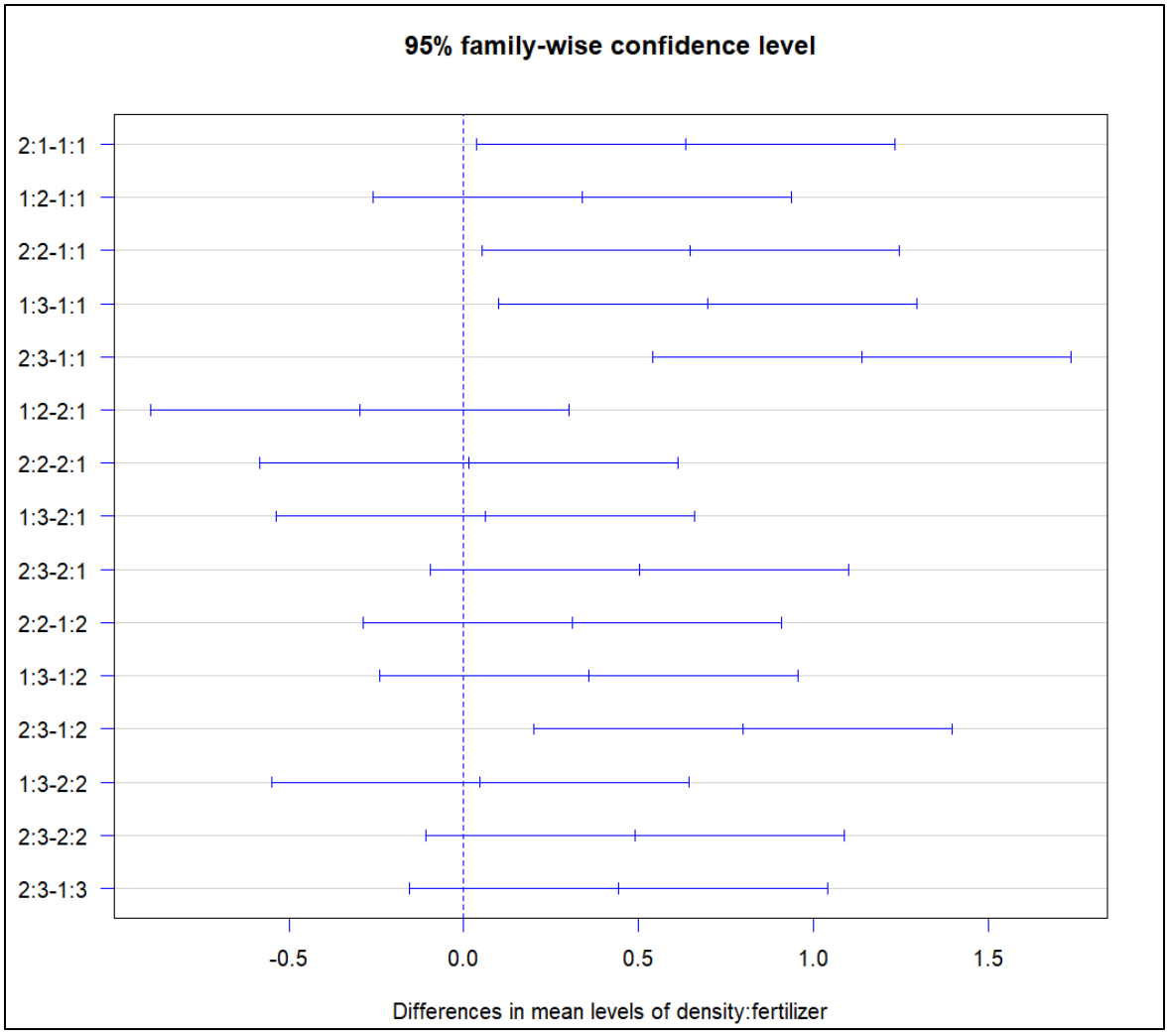
- Since the interaction null hypothesis is not rejected, it can be concluded that the density and fertilizer does affect the yield



Interpretation:

1. There is a difference of the yield by density type (left-below)
2. There is a difference of the yield by fertilizer type (right-above)

Post-hoc, Tukey test



Interpretation:

1. I can recognize which is significant in specific categories
2. 2 (density) & 1 (fertilizer) – 1 (density) & 1 (fertilizer) is significant
3. 2 (density) & 2 (fertilizer) – 1 (density) & 1 (fertilizer) is significant
4. 1 (density) & 3 (fertilizer) – 1 (density) & 1 (fertilizer) is significant
5. 2 (density) & 3 (fertilizer) – 1 (density) & 1 (fertilizer) is significant
6. 2 (density) & 3 (fertilizer) – 1 (density) & 2 (fertilizer) is significant
   Since those confidence level covers 0

Post-hoc, Scheffe Test

Posthoc multiple comparisons of means: Scheffe Test 95% family-wise confidence level				
$density				
	diff	lwr.ci	upr.ci	pval
2-1	0.461956	0.05869865	0.8652134	0.0140 *
$fertilizer				
	diff	lwr.ci	upr.ci	pval
2-1	0.1761687	-0.31771870	0.6700561	0.9147
3-1	0.5991256	0.10523815	1.0930130	0.0073 **
3-2	0.4229569	-0.07093057	0.9168443	0.1431

Posthoc multiple comparisons of means: Scheffe Test				
	diff	lwr.ci	upr.ci	pval
2-1 – 1:1 .	0.63489351	-0.063568788	1.3333558	0.0998
1:2 – 1:1	0.33868995	-0.359772347	1.0371522	0.7421
…		…		…
2:3 – 1:1 ***	1.13713411	0.438671815	1.8355964	6.1e-05
…		…		…
2:3 – 1:3	0.44112356	-0.257338735	1.1395859	0.4695
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Interpretation:

1. As can be seen in this table, in fact, the result value is similar to the Tukey test
2. However, I think, Scheffe was a more comfortable way to see the numbers in "R". Since I can see the p-value is the table directly.
3. 2 (density) & 2 (fertilizer) – 1 (density) & 1 (fertilizer) is significant
4. 1 (density) & 3 (fertilizer) – 1 (density) & 1 (fertilizer) is significant
5. 2 (density) & 3 (fertilizer) – 1 (density) & 1 (fertilizer) is significant
6. 2 (density) & 3 (fertilizer) – 1 (density) & 2 (fertilizer) is significant

RESULT & INTERPRETATIONS

1. In this project, I did a lot of chi-square, one-way ANOVA, and two-way ANOVA. In the process, I learn what the assumptions are and under what conditions each test was performed.
2. Chi-square is used for tests concerning frequency distributions. One-way ANOVA is executed to test the equality of three or more means using sample variances. Two-way ANOVA is used to determine if there is a significant difference in the main effects or interaction
3. I learned several ways to find critical values. Among them, it was confirmed that the p-value comparison method is convenient because it can be applied to all tests.
4. I also learn about the tests that can be done after conducting the two-way ANOVA. I find out that there is a Tukey test and a Scheffe test for post-hoc, and that they do almost the same thing. I also look for differences between the two.

CONCLUSION

I learn how to compare frequencies and compare various types of variances at once through chi-square and one-way & two-way ANOVA. Through the process of analyzing the results of chi-square and ANOVA, I can understand more clearly what each test meant. I also conduct the tests using real data. And after executing the ANOVA test, I learned how to figure out more detailed values and draw and visualize graphs.

REFERENCES

Bluman, Allan. (2017). Elementary statistics: a step by step approach 10th edition. McGraw-Hill.

Kabacoff, Robert I. (2015). R in Action 2nd Edition. Manning Publisher.

TechVidvan. (n.d). R Matrix – How to create, name and modify matrices in R?. Retrieved from https://techvidvan.com/tutorials/r-matrix/

DataFlair. (n.d). Chi-Square Test in R | Explore the Examples and Essential concepts!. Retrieved from https://data-flair.training/blogs/chi-square-test-in-r/

ZACH. (2020, October 21). How to Perform a Chi-Square Goodness of Fit Test in R. STATOLOGY. Retrieved from https://www.statology.org/chi-square-goodness-of-fit-test-in-r/

STHDA. (n.d.). Two-Way ANOVA Test in R. STATOLOGY. Retrieved from http://www.sthda.com/english/wiki/two-way-anova-test-in-r

Shaun Turney. (2022, May 31). Chi-Square (X^2) Table | Examples & Downloadable Table. Scribbr. Retrieved from https://www.scribbr.com/statistics/chi-square-distribution-table/

Biochem Med. (2013, June 15). The Chi-square test of independence. PMC. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900058/

ZACH. (2021, August 14). The Four Assumptions of a Chi-Square Test. STATOLOGY. Retrieved from https://www.statology.org/chi-square-test-assumptions/

PennState. (n.d.). Assumptions for One-Way ANOVA Test. Retrieved from https://online.stat.psu.edu/stat500/lesson/10/10.2/10.2.1

Rebecca Bevans. (2022, March 6). One-way ANOVA | When and How to Use It (With Examples). Retrieved from https://www.scribbr.com/statistics/one-way-anova/

Rebecca Bevans. (2020, March 20). Two-Way ANOVA | Examples & When To Use It. Retrieved from https://www.scribbr.com/statistics/two-way-anova/

ZACH. (2022, March 28). How to Fix in R: Invalid Graphics State (3 Solutions). Retrieved from https://www.statology.org/r-invalid-graphics-state/

Will Kenton. (2021, May 30). Scheffé Test. Investopedia. Retrieved from https://www.investopedia.com/terms/s/scheffes-test.asp

StatTrek. (n.d.). Scheffé's Test for Multiple Comparisons. Retrieved from https://stattrek.com/anova/follow-up-tests/scheffe

Ouorou Ganni Mariel Guera. (2019, August 2). What are the major differences between scheffe and tukey post hoc tests? ResearchGate. Retrieved from https://www.researchgate.net/post/What-are-the-major-differences-between-scheffe-and-tukey-post-hoc-tests

David Mark Lane. (n.d.). Tukey's Honestly Significant Difference (HSD). researchmethods. Retrieved https://methods.sagepub.com/reference/encyc-of-research-design/n478.xml

R-Codes

```
library(psych)
library(dplyr)
library(tibble)
library(ggpubr)
library(gplots)
library(ggplot2)
library(tidyverse)
library(GGally)
install.packages("DescTools")
library(DescTools)

######### 6. Blood Types #########
####################################

#   Expected | observed
# A    20%   |      12
# B    28%   |      8
# O    36%   |      24
# AB   16%   |      6

#State the hypothesis
#H0: A = 0.20, B=0.28, O=0.36, AB=0.16
#H1: The distribution is not the same as stated in the null hypothesis.

# Set significance level
alpha <-  0.10

# Create a vector of the values
observed <-  c(12, 8, 24, 6)
observed

p <- c(0.20, 0.28, 0.36, 0.16)

result <- chisq.test(x=observed, p = p)
result$statistic # Chi-square test value
result$p.value # Chi-square p-value
result$parameter # degrees of freedom (# of categories - 1)

result

ifelse(result$p.value > alpha, "Fail to reject the null hypothesis", "Reject the null hypothesis")

#### 8. On-time Performance by Airlines ####
```

```
##############################################

#                                    Expected   | observed
# On time                                70.8%   |     125
# National Aviation System delay          8.2%   |     10
# Aircraft arriving late                  9.0%   |     25
# Other (because of weather and other conditions)  12.0%   |     40

#State the hypothesis
#H0: On time = 0.708, NASD=0.082, AAL=0.09, Other=0.12
#H1: The distribution is not the same as stated in the null hypothesis.

# Set significance level
alpha2 <-  0.05

# Create a vector of the values
observed2 <-  c(125, 10, 25, 40)
observed2

p2 <- c(0.708, 0.082, 0.09, 0.12)

result2 <- chisq.test(x=observed2, p = p2)
result2$statistic # Chi-square test value
result2$p.value # Chi-square p-value
result2$parameter # degrees of freedom (# of categories - 1)

result2
ifelse(result2$p.value > alpha2, "Fail to reject the null hypothesis", "Reject the null hypothesis")

#### 8. Ethnicity and Movie Admissions ####
##############################################

#     Caucasian  |  Hispanic  | African American | Other
# 2013   724     |     335   |      174        |  107
# 2014   370     |     292   |      152        |  140

#State the hypothesis
#H0: There is no difference in the number of movie admissions related to ethnicity.
#H1: There is difference in the number of movie admissions related to ethnicity.

## From Agresti(2007) p.39
alpha3 <- 0.05

Movie <- as.table(rbind(c(724, 335, 174, 107), c(370, 292, 152, 140)))
dimnames(Movie) <- list(Year = c(2013, 2014),
                        Ethnicity = c("Caucasian","Hispanic", "African American", "Other"))
```

```
Movie

result3 <- chisq.test(Movie)
result3

ifelse(result3$p.value > alpha3, "Fail to reject the null hypothesis", "Reject the null hypothesis")

#### 10. Women in the Military ####
##########################################

#                    Officers |   Enlisted
# Army          10791     |    62491
# Navy           7816     |    42750
# Marine Corps   932      |    9525
# Air Force     11819     |    54344

#State the hypothesis
#H0: There is no difference in the number of ranks between branch of the Armed Forces.
#H1: There is difference in the number of ranks between branch of the Armed Forces.

## From Agresti(2007) p.39
alpha4 <- 0.05

Rank <- as.table(rbind(c(10791, 62491), c(7816, 42750), c(932, 9525), c(11819, 54344)))
dimnames(Rank) <- list(Branch = c("Army", "Navy", "Marine Corps", "Air Force"),
                       Rank = c("Officers","Enlisted"))
Rank

result4 <- chisq.test(Rank)
result4

ifelse(result4$p.value > alpha4, "Fail to reject the null hypothesis", "Reject the null hypothesis")


#### ANOVA 8. Sodium Contents of Foods ####
##########################################

# Set significance level
alpha5 <-  0.05

# At the 0.05 level of significance, it there?
# Condiment |Cereals | Desserts

# State the hypotheses
# H0: μ1 = μ2 = μ3
# H1: At least one mean is differnt from the others
```

```
# Set the significance level
alpha5 <-  0.05

# Create a data.frame for condiments
condiments <-  data.frame('sodium'=c(270, 130, 230, 180, 80, 70, 200),
'food'=rep('condiment',7), stringsAsFactors = FALSE)

# Create a data.frame for Cereals
cereals <-  data.frame('sodium'=c(260, 220, 290, 290, 200, 320, 140), 'food'=rep('cereal',7),
stringsAsFactors = FALSE)

# Create a data.frame for desserts
desserts <-  data.frame('sodium'=c(100, 180, 250, 250, 300, 360, 300, 160),
'food'=rep('dessert',8), stringsAsFactors = FALSE)

# Combine the data.frames into one
sodium <-  rbind(condiments, cereals, desserts)
sodium$food <-  as.factor(sodium$food)

sodium

# Run the ANOVA test
anova <- aov(sodium ~ food, data = sodium)

# View the model summary
summary(anova)

# Save summary to an object
sodium.summary <-  summary(anova)
sodium.summary

# Degree of freedom
# k-1 : between group variance - numerator
df.numerator <- a.summary[[1]][1, "Df"]
df.numerator

df.denominator <- a.summary[[1]][2, "Df"]
df.denominator

# Extract the F value from the summary
F.value <- a.summary[[1]][1, "F value"]
F.value

# Extract the F value from the summary
p.value <- sodium.summary[[1]][1, "Pr(>F)"]
```

```
p.value

ifelse(p.value > alpha5, "fail to reject the null", "reject the null")

# See differences
TukeyHSD(anova)

#### ANOVA 10. Sales for Leading Companies ####
##############################################

# Set significance level
alpha6 <-  0.01

# At the 0.01 level of significance, it there?
# Cereal | Chocolate Candy | Coffee

# State the hypotheses
# H0: μ1 = μ2 = μ3
# H1: At least one mean is differnt from the others

# Set the significance level
alpha6 <-  0.01

# Create a data.frame for cereal
cereal <-  data.frame('sales'=c(578, 320, 264, 249, 237), 'company'=rep('cereal',5),
stringsAsFactors = FALSE)

# Create a data.frame for Chocolate Candy
chocolate.candy <-  data.frame('sales'=c(311, 106, 109, 125, 173),
'company'=rep('chocolate.candy',5), stringsAsFactors = FALSE)

# Create a data.frame for cereal
coffee <-  data.frame('sales'=c(261, 185, 302, 689), 'company'=rep('coffee',4), stringsAsFactors
= FALSE)

# Combine the data.frames into one
company <-  rbind(cereal, chocolate.candy, coffee)
company$company <-  as.factor(company$company)

company

# Run the ANOVA test
anova2 <- aov(sales ~ company, data = company)

# View the model summary
summary(anova2)
```

```
# Save summary to an object
company.summary <-  summary(anova2)

# Degree of freedom
# k-1 : between group variance - numerator
df.numerator <- a.summary[[1]][1, "Df"]
df.numerator

df.denominator <- a.summary[[1]][2, "Df"]
df.denominator

# Extract the F value from the summary
F.value <- a.summary[[1]][1, "F value"]
F.value

# Extract the F value from the summary
p.value2 <- company.summary[[1]][1, "Pr(>F)"]
p.value2

ifelse(p.value2 > alpha6, "fail to reject the null", "reject the null")

# See differences
TukeyHSD(anova2)

#### ANOVA 12. Per-Pupil Expenditures ####
##########################################
# At the 0.05 level of significance, it there?
# Eastern third | Middle third | Western third

# State the hypotheses
# H0: μ1 = μ2 = μ3
# H1: At least one mean is differnt from the others

# Set the significance level
alpha7 <-  0.05

# Create a data.frame for eastern
eastern <-  data.frame('expenditures'=c(4946, 5953, 6202, 7243, 6113), 'states'=rep('eastern',5),
stringsAsFactors = FALSE)

# Create a data.frame for middle
middle <-  data.frame('expenditures'=c(6149, 7451, 6000, 6479), 'states'=rep('middle',4),
stringsAsFactors = FALSE)

# Create a data.frame for western
```

```
western <-  data.frame('expenditures'=c(5282, 8605, 6528, 6911), 'states'=rep('western',4), stringsAsFactors = FALSE)

# Combine the data.frames into one
state <-  rbind(eastern, middle, western)
state$states <-  as.factor(state$states)

state

# Run the ANOVA test
anova3 <- aov(expenditures ~ states, data = state)

# View the model summary
summary(anova3)

# Save summary to an object
state.summary <-  summary(anova3)

# Degree of freedom
# k-1 : between group variance - numerator
df.numerator <- a.summary[[1]][1, "Df"]
df.numerator

df.denominator <- a.summary[[1]][2, "Df"]
df.denominator

# Extract the F value from the summary
F.value <- a.summary[[1]][1, "F value"]
F.value

# Extract the F value from the summary
p.value3 <- state.summary[[1]][1, "Pr(>F)"]
p.value3

ifelse(p.value3 > alpha7, "fail to reject the null", "reject the null")

# See differences
TukeyHSD(anova3)


############ two-way ANOVA ############

# making table for two-independents
plant1 <-  data.frame('light'=rep(1,6), 'food'=c("A", "A", "A", "B", "B", "B"), stringsAsFactors = FALSE)
```

```
plant1

plant2 <-  data.frame('light'=rep(2,6), 'food'=c("A", "A", "A", "B", "B", "B"), stringsAsFactors =
FALSE)

plant2

# Combine the data.frames into one
plant <-  rbind(plant1, plant2)
plant
plant3 <-  data.frame('growth'=c(9.2,9.4,8.9,7.1,7.2,8.5,8.5,9.2,8.9,5.5,5.8,7.6), stringsAsFactors
= FALSE)
plant3
plant <- cbind(plant, plant3)
plant

plant$food <- as.factor(plant$food)
plant$light <- as.factor(plant$light)

fit <- aov(growth ~ light*food, data=plant)
summary(fit)

ggboxplot(plant, x="light", y="growth", color="food", palette = c("#00AFBB", "#E7B800"))

ggline(plant, x = "light", y = "growth", color = "food",
       add = c("mean_se", "dotplot"),
       palette = c("#00AFBB", "#E7B800"))

plot(fit, 1)
plot(fit, 2)

attach(plant)
plotmeans(growth ~ interaction(food, light, sep=" "),
          connect=list(c(1,3,5),c(2,4,6)),
          col=c("red", "darkgreen"),
          main = "Interaction Plot with 95% CIs",
          xlab="Food and Light Combination")
detach(plant)

install.packages("HH")
library(HH)
interaction2wt(growth~food*light)

#post-hoc
fit.tukey <- TukeyHSD(fit)
plot(fit.tukey, col="blue", las=1)
```

```
#Setwd in file direction
setwd("C:\\Users\\14083\\Desktop\\0. Winter course in 2023 Northeastern\\ALY
6015\\Module2\\Assignment2\\dataset")

#1. Import dataset using read.csv()
bb <- read.csv("baseball.csv", stringsAsFactors = T,
               header=T)

##### EDA ##############
headtail(bb,5)

# Extract decade from year
bb$Year %% 10
bb$Decade <- bb$Year - (bb$Year %% 10)

bb$Decade

headtail(bb,5)
describe(bb)

dev.off()
#histogram (Decade)
ggplot(bb, aes(Decade)) +
  geom_histogram()

# Create a wins table by summing the wins by decade
wins <- bb %>%
  group_by(Decade) %>%
  summarize(wins = sum(W)) %>%
  as.tibble()

games <- bb %>%
  group_by(Decade) %>%
  summarize(games = sum(G)) %>%
  as.tibble()

### chi-square test
alpha.bb <- 0.05

# Create a vector of the values
observed <- wins$wins
observed

p <- c(1/6, 1/6, 1/6, 1/6, 1/6, 1/6)
```

```
result.bb <- chisq.test(x=observed, p = p)
result.bb$statistic # Chi-square test value
result.bb$p.value # Chi-square p-value
result.bb$parameter # degrees of freedom (# of categories - 1)

result.bb

ifelse(result.bb$p.value > alpha.bb, "Fail to reject the null hypothesis", "Reject the null
hypothesis")

#1. Import dataset using read.csv()
crop <- read.csv("crop_data.csv", stringsAsFactors = T,
            header=T)

##### EDA needed ##############
headtail(crop,5)
describe(crop)

#histogram yield
ggplot(crop, aes(yield)) +
  geom_histogram()

# Extract decade from year
crop.sub <- subset(crop, select=c("density", "fertilizer", "yield"))
crop.sub

crop.sub$density <- as.factor(crop.sub$density)
crop.sub$fertilizer <- as.factor(crop.sub$fertilizer)

fit2 <- aov(yield ~ density*fertilizer, data=crop.sub)
summary(fit2)

attach(crop)

#plots
interaction2wt(yield ~ density*fertilizer)

#post-hoc
fit.tukey2 <- TukeyHSD(fit2)
fit.tukey2
plot(fit.tukey2, col="blue", las=1)

#post-hoc2
ScheffeTest(fit2)
detach(crop)
```