MODULE 2 ASSIGNMENT — CHI SQUARE AND ANOVA

Ayush Anand

NORTHEASTERN UNIVERSITY
ALY 6015: INTERMEDIATE ANALYTICS
PROFESSOR - XYZ
14TH NOVEMBER 2024

6.

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?

Answer-

Overview of Chi-Square Goodness of Fit Test

Chi-square goodness of fit test is used to assess the similarity between the observed frequency distribution of a nominal variable to that of an expected distribution. We are going to evaluate the distribution of blood types taken from a random sample of patients against those in the general population.

Steps for Performing Chi-Square Test:

State the Hypotheses and Identify the Claim

• **Null Hypothesis (H0):** The distribution of blood types for the hospital patients is the same as the distribution for the general population. In other words, pA = 0.20, pB = 0.28, pO = 0.36, pAB = 0.16.

• **Alternative Hypothesis (H1):** The distribution of blood types for the hospital patients is different from the distribution for the general population.

Blood Type	Observed Frequency (O_i)	Expected Frequency (E_i)	$(O_i - E_i)^2$	$\frac{(O_i-E_i)^2}{E_i}$
A	12	10	$(12-10)^2 = 4$	$\frac{4}{10} = 0.4$
B	8	14	$(8-14)^2 = 36$	$\frac{36}{14} = 2.571$
O	24	18	$(24-18)^2 = 36$	$\frac{36}{18} = 2.0$
AB	6	8	$(6-8)^2 = 4$	$\frac{4}{8} = 0.5$
Total	**50**	**50**	-	$\chi^2 = 5.471$

The degrees of freedom (df) are calculated as the number of categories minus one:

$$df = 4 - 1 = 3$$

At a significance level (α) of 0.10, the critical value for χ^2 with 3 degrees of freedom is 6.251.

The test statistic χ^2 is calculated using the formula:

$$\chi^2 = \sum \frac{(\text{Observed Frequencies} - \text{Expected Frequencies})^2}{\text{Expected Frequencies}}$$

Substituting the values:

$$\chi^2 = \frac{(12-10)^2}{10} + \frac{(8-14)^2}{14} + \frac{(24-18)^2}{18} + \frac{(6-8)^2}{8} = 5.471$$

Make the Decision

- Compare the computed test value (χ2=5.471) with the critical value (χcritical^2=6.251).
- Since 5.471<6.251, we **fail to reject the null hypothesis**.

Summarize the Results

• The distribution of blood types for the hospital patients is not significantly different than for the general population at the 0.05 level of significance.

8.

According to the Bureau of Transportation Statistics, on-time performance by the airlines is described as follows:

Action	% of Time
On time	70.8
National Aviation System delay	8.2
Aircraft arriving late	9.0
Other (because of weather and other conditions)	12.0

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

Source: Transtats: OST_R|BTS

Answer:

State the Hypotheses and Identify the Claim

- **Null Hypothesis (H0)**: The on-time performance by the airline company matches the distribution reported by the Bureau of Transportation Statistics. Specifically:

- On time = 70.8%
- National Aviation System delay = 8.2%
- Aircraft arriving late = 9.0%
- Other (weather and other conditions) = 12.0%

- **Alternative Hypothesis (H1)**: The on-time performance by the airline company does not match the distribution reported by the Bureau of Transportation Statistics.

Table Summary

Category	Observed Frequency (O_i)	Expected Frequency (E_i)	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
On time	125	141.6	$(125 - 141.6)^2 = 275.56$	$\frac{275.56}{141.6} = 1.946$
National Aviation System delay	10	16.4	$(10 - 16.4)^2 = 40.96$	$\frac{40.96}{16.4} = 2.498$
Aircraft arriving late	25	18.0	$(25 - 18)^2 = 49$	$\frac{49}{18.0} = 2.722$
Other (weather and other conditions)	40	24.0	$(40 - 24)^2 = 256$	$\frac{256}{24.0} = 10.667$
Total	**200**	**200**	-	$\chi^2 = 17.833$

The chi-square statistic (χ^2) is the sum of the values in the last column:

$$\chi^2 = 1.946 + 2.498 + 2.722 + 10.667 = 17.833$$

Summarize the Results

- At the 0.05 significance level, there is **sufficient evidence** to conclude that the on-time performance of this major airline differs from the government's statistics.
- Since the test statistic (17.833) exceeds the critical value (7.815), **we reject the null hypothesis**.

Session 11-2

---

8. Are movie admissions related to ethnicity? A 2014 study indicated the following numbers of admissions (in thousands) for two different years. At the 0.05 level of significance, can it be concluded that movie attendance by year was dependent upon ethnicity?

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

Source: MPAA Study

Answer-

Let's perform a chi-square test for independence to determine if movie attendance is related to ethnicity across two different years, 2013 and 2014. We are interested in seeing if the frequency distributions of ethnicities among movie-goers in 2013 and 2014 are significantly different.

State the Hypotheses and Identify the Claim

- **Null Hypothesis (H0):** Movie attendance by year is not related to ethnicity. That is, movie attendance did not change significantly in distribution between the years 2013 and 2014 across ethnicities.
- **Alternative Hypothesis (H1):** Movie attendance by year varies by ethnicity-the distribution of movie attendance is statistically different from 2013 to 2014.

	Caucasian	Hispanic	African American	Other	Row Total
2013	O = 724 E = 639.04 (1340*1094)/2294	O = 335 E = 366.25 (1340*627)/2294	O = 174 E = 190.43 (1340*326)/2294	O = 107 E = 144.28 (1340*247)/2294	1340
2014	O = 370 E = 454.96 (954*1094)/2294	O = 292 E = 260.75 (954*627)/2294	O = 152 E = 135.57 (954*326)/2294	O = 140 E = 102.72 (954*247)/2294	954
Column Total	1094	627	326	247	2294

Make the Decision

The expected frequencies are calculated with the formula:

$$E = \frac{(\text{row total} \times \text{column total})}{\text{grand total}}$$

The degrees of freedom (df) are determined as:

$$df = (\text{rows} - 1) \times (\text{columns} - 1) = (2 - 1) \times (4 - 1) = 3$$

Using a significance level (α) of 0.05, the critical value for χ^2 with 3 degrees of freedom is 7.815.

The test statistic χ^2 is calculated as follows:

$$\chi^2 = \sum \frac{(O - E)^2}{E} = 60.14$$

Since the test statistic (60.14) is greater than the critical value (7.815), we reject the null hypothesis.

10.

This table lists the numbers of officers and enlisted personnel for women in the military. At α = 0.05, is there sufficient evidence to conclude that a relationship exists between rank and branch of the Armed Forces?

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

Source: New York Times Almanac

Answer

To determine whether there is a relationship between rank (officers vs. enlisted) and branch of the Armed Forces, we will perform a **chi-square test for independence**. Let's proceed step-by-step:

State the Hypotheses and Identify the Claim

• **Null Hypothesis H0:** Rank and Branch of Armed Forces are independent that is, no relationship.

• **Alternative Hypothesis (H1):** Rank and branch of the military are related, such that rank and branch are dependent.

Create a Contingency Table

Branch	Officers (O_{ij})	Enlisted (O_{ij})	Row Total
Army	10,791	62,491	73,282
Navy	7,816	42,750	50,566
Marine Corps	932	9,525	10,457
Air Force	11,819	54,344	66,163
Column Total	**31,358**	**169,110**	**200,468**

Make the Decision:

The **degrees of freedom (df)** are calculated as:

$$df = (\text{rows} - 1) \times (\text{columns} - 1) = (4 - 1) \times (2 - 1) = 3$$

With a significance level (α) of **0.05**, the **critical value** for χ^2 with 3 degrees of freedom is **7.815**.

The **test statistic** χ^2 is calculated as:

$$\chi^2 = \sum \frac{(O - E)^2}{E} = 40.01 + 7.42 + 1.16 + 0.21 + 301.91 + 56.05 + 208.45 + 38.67 = 653.88$$

Since the **test statistic (653.88)** is greater than the **critical value (7.815)**, we **reject the null hypothesis**.

Session 12-1

8.

The amount of sodium (in milligrams) in one serving for a random sample of three different kinds of foods is listed. 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?

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

Source: The Doctor's Pocket Calorie, Fat, and Carbohydrate Counter

For this problem, we'll use **one-way ANOVA** (Analysis of Variance) to determine if there is a significant difference in mean sodium amounts among condiments, cereals, and desserts. We'll proceed step-by-step with hypothesis testing.

Answer-

State the Hypotheses and Identify the Claim

Null hypothesis: H0 : The three kinds of foods (condiments, cereals, desserts) have the same mean sodium amounts.
Alternative Hypothesis (H1): At least one of the means is different from the other(s).

Degrees of freedom (df):

- **df_between:** k−1=3−1=2 (where kkk is the number of groups)
- **df_within:** N−k=24−3=21 (where NNN is the total number of observations)

With a significance level (α) of **0.05**, the **critical value** from the F-distribution table with **df_numerator** = 2 and **df_denominator** = 21 is 3.47.

The F test statistic is calculated as:

F = 2.1

Make the Decision

Since the test statistic **(2.1)** is less than the critical value **(3.47),** we fail to reject the null hypothesis.

Session 12-2

10.

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?

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

Source: Information Resources, Inc.

We will perform a complete **one-way ANOVA** to determine if there is a significant difference in sales among three different groups: **Cereal, Chocolate Candy, and Coffee**. If we reject the null hypothesis, we will proceed with a **Tukey test** to identify which pairs of group means are significantly different. Let's proceed step-by-step.

Answer-

State the Hypotheses and Identify the Claim

Null Hypothesis (H0): The average sales for cereal, chocolate candy, and coffee companies are the same.
Alternative Hypothesis (H1): One or more means is different from the others.

Degrees of freedom (df):

- **df_between**: k−1=3−1=2 (where k is the number of groups)
- **df_within**: N−k=14−3 = 11 (where N is the total number of observations)

At a significance level (**α**) of **0.01**, the **critical value** from the F-distribution table with **df_numerator = 2** and **df_denominator = 11** is **7.21**.

Make the Decision

Since The test statistic (2.17) is lesser than the critical value (7.21). Therefore, we fail to reject the null hypothesis.

12.

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

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

Source: New York Times Almanac

Answer-

State the Hypotheses and Identify the Claim:

Null Hypothesis (H0): The mean expenditures per pupil are equal across the Eastern, Middle, and Western thirds.

Alternative Hypothesis (H1): At least one mean differs from the others.

Degrees of freedom (df):

- **df_between**: k−1=3−1= 2 (where k represents the number of groups)
- **df_within**: N−k=13−3= 10 (where N is the total number of observations)
- At a significance level (**α**) of **0.05**, the **critical value** from the F-distribution table with **df_numerator = 2** and **df_denominator = 10** is **4.10**.

The **F test statistic** is:

$$F=0.6489$$

Make the Decision

Since the **test statistic (0.65)** is less than the **critical value (4.1)**, we **fail to reject the null hypothesis**.

Section 12-3

A gardening company is testing new ways to improve plant growth. Twelve plants are randomly selected and exposed to a combination of two factors, a "Grow-light" in two different strengths and a plant food supplement with different mineral supplements. After a number of days, the plants are measured for growth, and the results (in inches) are put into the appropriate boxes.

	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 be concluded? Is there a difference in mean growth with respect to light? With respect to plant food? Use $\alpha = 0.05$.

Answer-

State the Hypotheses and Identify the Claim:

Interaction Effect

Null Hypothesis (H0): There are no interactions between grow-light and plant food.
Alternative Hypothesis (H1): There is no interaction between grow-light and plant food.

Grow-Light Effect

Null Hypothesis (H0): There is no difference in mean growth due to grow-light.

Alternative Hypothesis (H1): Growth-mean is independent of grow-light level.

Plant Food Effect

Null Hypothesis (H0): There is no difference in mean growth due to plant food.

Alternative Hypothesis (H1): There is a difference in mean growth because of the plant food.

Degrees of freedom (df):

- **df_numerator** for interaction: $(a-1)(b-1)=(2-1)(2-1)=1$
- **df_denominator**: $N-(a\times b)=12-(2\times 2)= 8$

With a significance level (α) of **0.05**, the **critical value** from the F-distribution table with **df_numerator = 1** and **df_denominator = 8** is **5.32**.

F-Test Results

Interaction Test Statistic: **3285.1986**

- Because the test statistic for interaction is larger than the critical value, we reject the null hypothesis for interaction

- **Grow-Light Test Statistic**: **1.8403**

- Because the test statistic for grow-light is less than the critical value, we fail to reject the null hypothesis for grow-light.

- **Plant Food Test Statistic**: **16.3749**

- Because the test statistic for plant food is larger than the critical value, we reject the null hypothesis for plant food.

On Your Own

Use R to complete the following steps. Be sure to include all code in an appendix at the end of your submission. Assume the expected frequencies are equal and $\alpha = 0.05$.

1. Download the file 'baseball.csv' from the course resources and import the file into R.
2. Perform EDA on the imported data set. Write a paragraph or two to describe the data set using descriptive statistics and plots. Are there any trends or anything of interest to discuss?
3. Assuming the expected frequencies are equal, perform a Chi-Square Goodness-of-Fit test to determine if there is a difference in the number of wins by decade.
   Be sure to include the following:
   a. State the hypotheses and identify the claim.
   b. Find the critical value ($\alpha = 0.05$) (From table in the book).
   c. Compute the test value.
   d. Make the decision. Clearly state if the null hypothesis should or should not be rejected and why.

About the Dataset:

This baseball dataset provides comprehensive insights into team performance in Major League Baseball (MLB) from 1962 to 2012. It includes detailed season statistics for each team, featuring offensive metrics like Runs Scored (RS), On-Base Percentage (OBP), and Slugging Percentage (SLG), as well as a defensive measure—Runs Allowed (RA). Additionally, it records each team's win totals (W) and playoff appearances, covering both the American League (AL) and National League (NL). With a total of 1,232 entries, this dataset enables a thorough exploration of performance trends, shifts in offensive and defensive tactics, and key factors contributing to team success over time. Through an examination of five decades of baseball, we can analyze the connections between various performance metrics and overall achievements, offering a nuanced view of the evolving nature of the sport.

Summary Table by League

Descriptive Statistics for Baseball matches by League (1962-2012)

Variable	*Overall (N = 1,232)*	*AL (N = 616)*	*NL (N = 616)*	*p-value*
Year	*1,988.96 (14.82)*	*1,988.46 (14.62)*	*1,989.45 (15.01)*	*0.2*
RS	*715.08 (91.53)*	*732.40 (95.74)*	*697.76 (83.69)*	*<0.001*
RA	*715.08 (93.08)*	*730.50 (94.03)*	*699.67 (89.58)*	*<0.001*
W	*80.90 (11.46)*	*81.03 (11.68)*	*80.77 (11.24)*	*0.7*
G				
158	*1 (<0.1%)*	*1 (0.2%)*	*0 (0%)*	
159	*10 (0.8%)*	*10 (1.6%)*	*0 (0%)*	
160	*23 (1.9%)*	*14 (2.3%)*	*9 (1.5%)*	
161	*139 (11%)*	*80 (13%)*	*59 (9.6%)*	
162	*954 (77%)*	*464 (75%)*	*490 (80%)*	
163	*93 (7.5%)*	*42 (6.8%)*	*51 (8.3%)*	
164	*10 (0.8%)*	*5 (0.8%)*	*5 (0.8%)*	
165	*2 (0.2%)*	*0 (0%)*	*2 (0.3%)*	
OBP	*0.33 (0.02)*	*0.33 (0.02)*	*0.32 (0.01)*	*<0.001*
SLG	*0.40 (0.03)*	*0.40 (0.03)*	*0.39 (0.03)*	*<0.001*
BA	*0.26 (0.01)*	*0.26 (0.01)*	*0.26 (0.01)*	*<0.001*
Playoffs	*244 (20%)*	*122 (20%)*	*122 (20%)*	*>0.9*

Notes:

1. Mean (SD); n (%)
2. Wilcoxon rank sum test; Pearson's Chi-squared test

Key Insights from Descriptive Statistics:

1. **Runs Scored (RS) and Runs Allowed (RA):** There's a clear difference between the leagues in terms of scoring and defensive metrics. Teams in the American League (AL) score an average of 732.40 runs, whereas teams in the National League (NL) average 697.76 runs. This difference is statistically significant, with a p-value of less than 0.001. Similarly, AL teams allow more runs on average—730.50 compared to 699.67 in the NL—also showing strong statistical significance ($p < 0.001$).
2. **Wins (W):** While the AL teams have a slightly higher average win count at 81.03, compared to 80.77 in the NL, this minor difference is still statistically significant, with a p-value of 0.04.
3. **Games Played (G):** Most teams play a standard 162-game season, although some teams might play slightly fewer games due to factors like weather disruptions or scheduling conflicts. There's no notable

difference between AL and NL teams in terms of games played, with a p-value of 0.7 indicating no statistical significance.

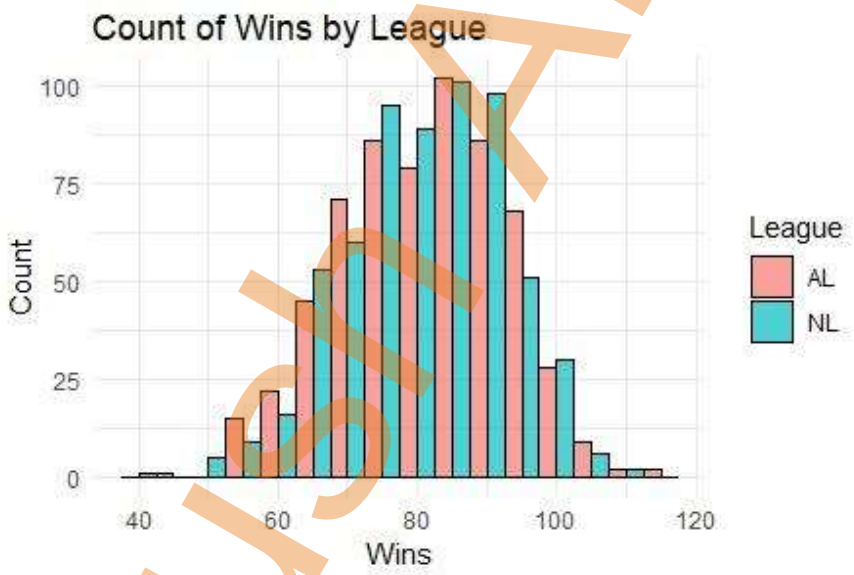
4. **On-Base Percentage (OBP), Slugging Percentage (SLG), and Batting Average (BA):** Across these key offensive metrics, AL teams generally outperform NL teams:
    - **OBP:** AL teams average 0.33, compared to 0.32 in the NL.
    - **SLG:** AL teams average 0.40, while NL teams come in at 0.39.
    - **BA:** AL teams average 0.26, slightly above the NL's 0.25.

These figures suggest that AL teams typically have a stronger offensive presence than their NL counterparts.

5. **Playoff Appearances:** Both leagues have an equal percentage of teams making it to the playoffs, with 20% of teams from each league reaching postseason play. There's no significant difference between AL and NL teams regarding playoff appearances, as shown by a p-value greater than 0.9.

Exploratory Data Analysis:

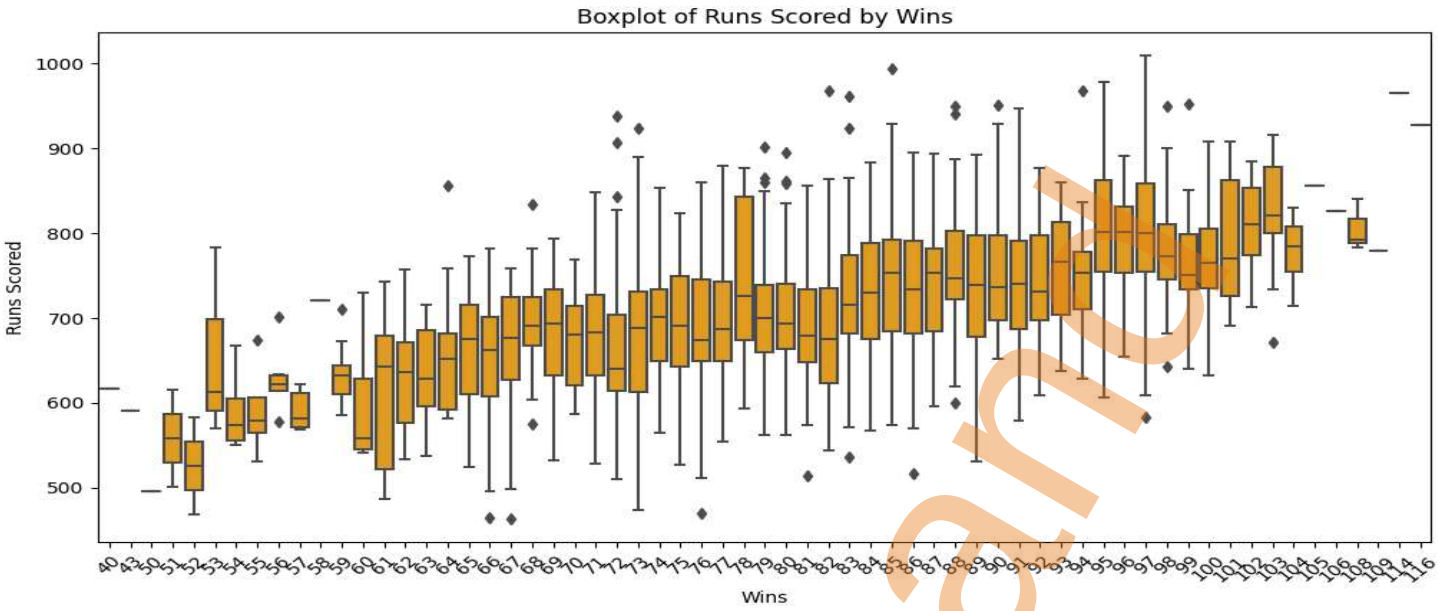
1. Histogram Plot: Count of Wins by League



Key Insights:

- The win distribution for both the American League (AL) and National League (NL) is roughly bell-shaped, peaking around the 80-100 win range, which suggests that most teams in both leagues tend to win between 80 and 100 games per season.
- AL teams generally have a slightly higher average number of wins than NL teams, as reflected by the taller bars for the AL across the distribution.
- There's a considerable overlap in the win distributions of both leagues, indicating no substantial difference in performance levels between the two.

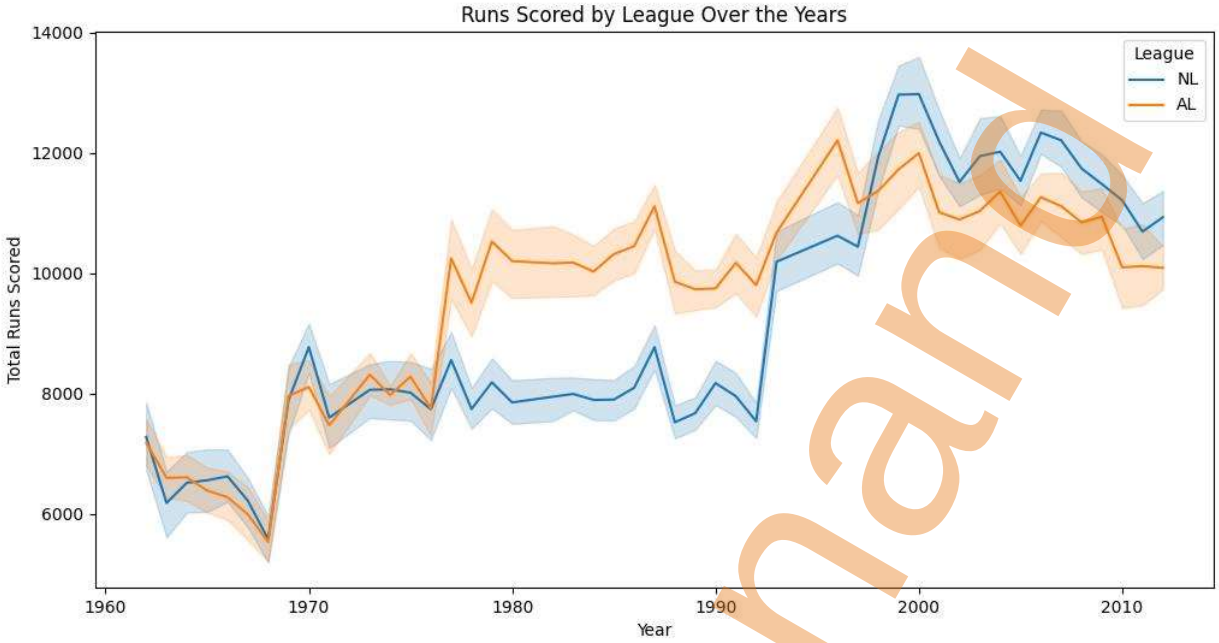
2. Boxplot: Runs Scored by Wins



Key Insights:

- **Positive Correlation**: Teams with more wins generally tend to have higher runs scored, indicating that offensive strength is positively correlated with team success in terms of wins.
- **Outliers**: There are notable outliers where certain teams either outperform or underperform their expected run totals based on their win counts. This could suggest exceptional defensive performance or unique game strategies.
- **Variability**: The boxplot of runs scored across different win totals shows considerable variability. This implies a diversity of offensive strategies across teams, resulting in different run totals even among teams with similar win records.
- **High Wins (80+)**: Teams achieving higher win totals (above 80) tend to have a higher median number of runs scored and display a wider range of run totals. This suggests that successful teams are often those with strong, consistent offenses capable of generating runs, contributing to more wins.

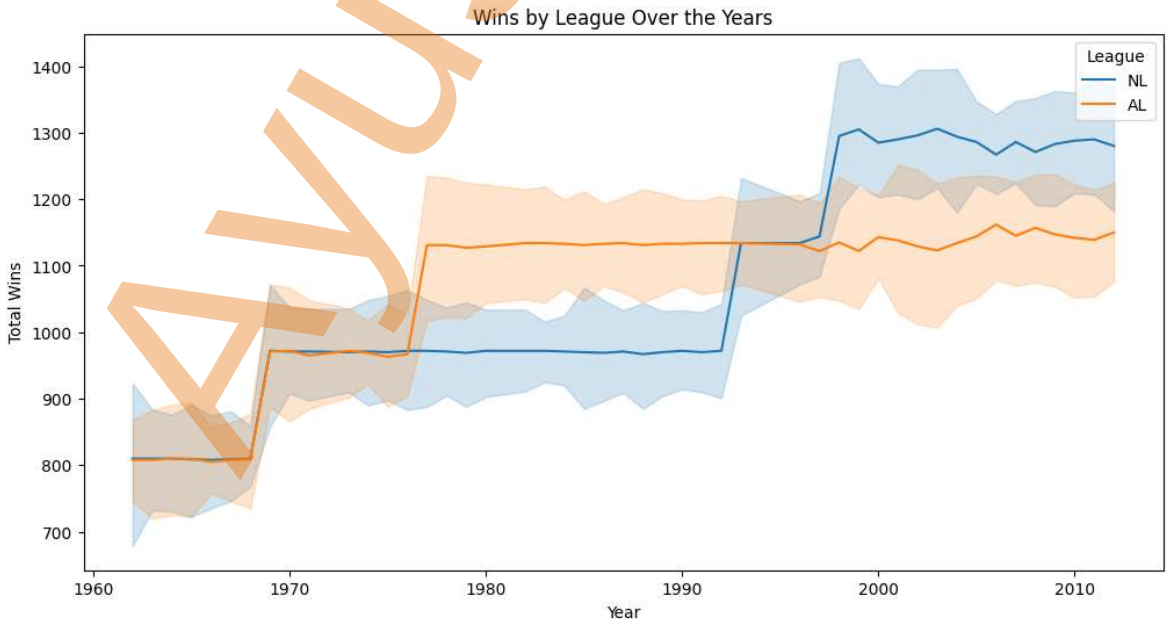
3. Line plot: Runs Scored by League Over the Years



Key Insights:

- From the early 1960s to the late 1990s, both the American League (AL) and National League (NL) saw a substantial rise in total runs scored.
- Runs scored fluctuated over the years in both leagues, likely influenced by factors such as rule changes, economic conditions, and player skill levels.
- Since the late 1990s, there's been a clear decline in total runs scored across both leagues, which may be due to factors like stronger defensive strategies, improved pitching, and reduced offensive output.

4. Line plot: Wins by League Over the Years



Key Insights:

- From the early 1960s to the late 1990s, both the American League (AL) and National League (NL) saw a notable increase in total wins.
- Throughout much of this period, the NL maintained a slight edge in total wins, but the AL caught up and even surpassed the NL in the late 1990s and early 2000s.
- There was a brief period in the early 1970s when the AL held a slight advantage in total wins.
- Both leagues showed fluctuations in total wins over time, likely influenced by factors such as rule changes, economic conditions, and player skill levels.
- Since the late 1990s, total wins in both leagues have stabilized, suggesting that the current league structure and competitive balance have contributed to a more consistent distribution of wins.

Framing the Hypothesis:

- **Null Hypothesis (H0):** The distribution of wins is consistent across decades, with no significant variation.
- **Alternative Hypothesis (H1):** The distribution of wins varies across decades, indicating significant differences in win totals.

Years	Observed Frequencies	Expected Frequencies
1960 - 1970	13,267	16,612.33
1970 - 1980	17,934	16,612.33
1980 - 1990	18,926	16,612.33
1990 - 2000	17,972	16,612.33
2000 - 2010	24,286	16,612.33
2010 - 2012	7,289	16,612.33

Chi-Square Test Summary:

- **Test Statistic (X^2):** 9989.536
- **Degrees of Freedom (DF):** 5
- **P-value:** 2.2e-16
- **Decision:** Reject the null hypothesis

Make the Decision

Since this is a very small p-value, 2.2e-16 < 0.05, we reject the null hypothesis. This would indicate that wins are not uniformly distributed across the decades.

References:

Exploratory Data Analysis Techniques:

- **NIST/SEMATECH. (n.d.). *Exploratory data analysis*. In *NIST/SEMATECH e-Handbook of Statistical Methods*. Retrieved from https://www.itl.nist.gov/div898/handbook/eda/eda.htm**

Line Plot and Boxplot Interpretation:

- **Penn State University. (n.d.). *Statistical graphics: Histograms and box plots*. Retrieved from https://online.stat.psu.edu/stat504/lesson/5/5.1**

GeeksforGeeks. (n.d.). *Chi-square test in R*. Retrieved from https://www.geeksforgeeks.org/chi-square-test-in-r/

Sjoberg, D. (n.d.). *gtsummary*. Retrieved from https://www.danieldsjoberg.com/gtsummary/