Northeastern University

ALY-6015
INTERMEDIATE ANALYTICS

BY INSTRUCTOR
ROY WADA

MODULE-2 PROJECT REPORT
SUBMITTED BY

ITI ROHILLA

On Jan-23-2024

SUMMARY

The report presents various hypothesis testing scenarios using Chi-Square and ANOVA tests, analyzing different datasets. In Section 11-1, the Chi-Square test is applied to examine the blood type distribution in hospital patients compared to the general population. The null hypothesis is that distributions are equal, and the alternative suggests differences. The results, shown in Table1, would guide decisions on accepting or rejecting the null hypothesis.

Moving to Section 12-1, ANOVA is used to assess mean sodium amounts among condiments, cereals, and desserts. The null hypothesis assumes equality, and the alternative suggests a significant difference. Table5 displays the ANOVA results, and decisions would be based on the significance level and the F-value.

In Dataset-Baseball, a Chi-Square Goodness-of-Fit test is conducted to analyze the distribution of wins across decades. The null hypothesis assumes equal distributions, and the alternative suggests differences. The results, detailed in the Chi_Square_Test_Results table, indicate a lack of evidence to reject the null hypothesis, implying no significant association between decades and wins.

Dataset-Crop involves ANOVA tests to explore the impact of density and fertilizer on plant growth. The null hypothesis assumes no impact, while the alternative suggests an interaction effect. The ANOVA results, presented in Table: ANOVA_Test_Results, are used to make decisions about the significance of density, fertilizer, and their interaction.

In conclusion, the report provides a comprehensive overview of hypothesis testing methods, emphasizing Chi-Square and ANOVA, to analyze various scenarios related to blood type distribution, sodium amounts, baseball wins, and plant growth. The decision-making process involves comparing test statistics, critical values, and p-values to draw meaningful conclusions about the underlying hypotheses in each case.

Section 11-1

Q6

Step 1: State the Hypotheses and Identify the Claim

- **Null Hypothesis (H0):** The blood type distribution in hospital patients is the same as that in the general population.

  H0: *PA*=0.20, PB=0.28, PO=0.36, PAB=0.16
- **Alternative Hypothesis (H1):** The blood type distribution in hospital patients is different from that in the general population.
  - H1: At least one blood type proportion differs from the general population.

Summarize the Result

	Test Statistic	Critical Value	P-Value	Decision
X-squared	5.471429	6.251389	0.1403575	Accept H_0 & Reject Alternative H1

Table1: Chi_Square Test Results

Q8

To perform hypothesis testing, we can follow these steps:

1. **State the hypotheses and identify the claim:**
   - Null Hypothesis (H_o): The airline's on-time performance is consistent with the government's statistics.
   - Alternative Hypothesis (H_1): The airline's on-time performance differs from the government's statistics.

Summarize the Result:

	Test Statistic	Critical Value	P-Value	Decision
X-squared	39.50424	7.814728	0	Reject H_0 & Accept the Alternative H1

Table2 : Chi_Square Test Results

Section 11-2

Q8

1. State the hypotheses and identify the claim:

- Null Hypothesis (H_o): Movie admissions are independent of ethnicity.
- Alternative Hypothesis (H_1): Movie admissions are dependent on ethnicity.

Summarize the Results:

	Test Statistic	Critical Value	P-Value	Decision
X-squared	60.14352	7.814728	5.478e-13	Reject H_0 & Accept the Alternative H1

Table3 : Chi_Square Test Results

Q10

1. State the hypotheses and identify the claim:

- Null Hypothesis (H_o): There is no relationship between rank and branch of the Armed Forces for women.
- Alternative Hypothesis (H_1): There is a significant relationship between rank and branch of the Armed Forces for women.

Summarize the Results:

	Test Statistic	Critical Value	P-Value	Decision
X-squared	654.2719	0.3518463	1.726e-141	Reject H_0 & Accept the Alternative H1

Table4 : Chi_Square Test Results

Section 12-1

Q8.

- Null Hypothesis (H_o): The mean sodium amounts are equal among condiments, cereals, and desserts.
- Alternative Hypothesis (H_1): There is a significant difference in mean sodium amounts among condiments, cereals, and desserts.

Summarize the Results:

S_NO	Df	Sum_Sq	Mean_Sq	F-Value	Pr..F	Decision
1	2	34379	17190	3.194	0.0637	The one-way ANOVA result is not significant
2	19	102257	5382	NA	NA	The one-way ANOVA result is not significant

Table5: ANOVA Test Results

Section 12-2

Q10.

- Null Hypothesis (H_o): There is no significant difference in the means of sales for cereal, chocolate candy, and coffee.
- Alternative Hypothesis (H_1): There is a significant difference in the means of sales for cereal, chocolate candy, and coffee.

Summarize the Results:

S_NO	Df	Sum_Sq	Mean_Sq	F-Value	Pr..F	Decision
1	2	1244588	622294	0.649	0.543	ANOVA is not significant. Fail to reject the null hypothesis.
2	10	9591145	959114	NA	NA	ANOVA is not significant. Fail to reject the null hypothesis.

Table6: ANOVA Test Results

Section 12-3

Q10.

- Null Hypothesis (H_o): There is no interaction between the two factors (Grow-light and Plant food) in affecting plant growth.
- Alternative Hypothesis (H_1): There is an interaction between the two factors in affecting plant growth.

Summarize the Results:

S_NO	Df	Sum_Sq	Mean_Sq	F-Value	Pr..F	Decision
1	2	109748	54874	2.506	0.123	The one-way ANOVA result is not significant
2	12	262795	21900	NA	NA	The one-way ANOVA result is not significant

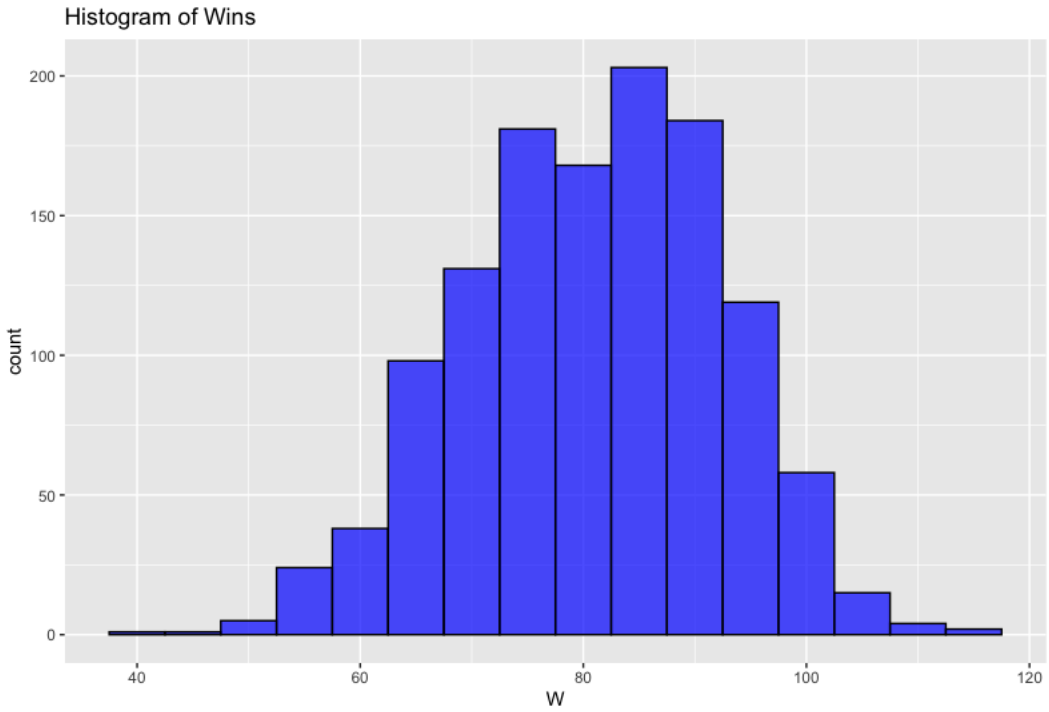
Table7: ANOVA Test Results

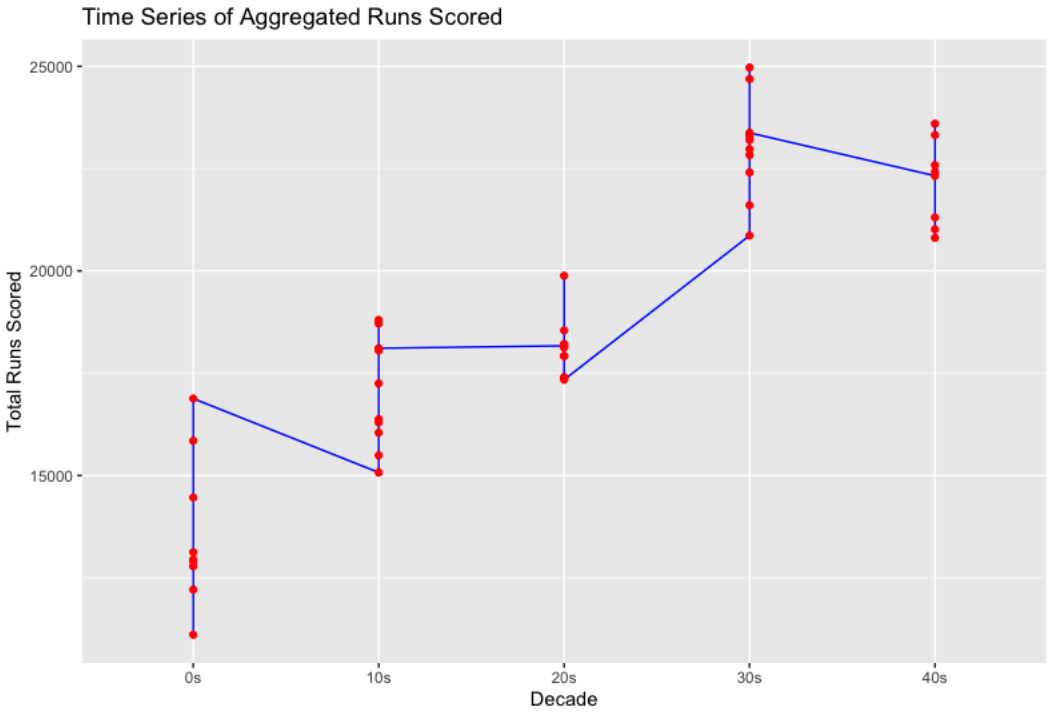
Dataset-Baseball

The provided sample dataset contains information about various Major League Baseball (MLB) teams for the years 2010-2012, including statistics related to runs scored (RS), runs allowed (RA), wins (W), on-base percentage (OBP), slugging percentage (SLG), batting average (BA), and playoff performance. The dataset is organized by teams, specifying the league (NL for National League, AL for American League), year, and various performance metrics for each team in a given year.

Summary	RS	RankSeason	RA	SLG	W
Min.	463	1	472	0.301	40
1st Qu.	652	2	649.8	0.375	73
Median	711	3	709	0.396	81
Mean	715.1	3.123	715.1	0.3973	80.9
3rd Qu.	775	4	774.2	0.421	89
Max.	1009	8	1103	0.491	116
NA		988			

Fig: Summary Table





	Test Statistic	Critical Value	P-Value	Decision
X-squared	3.333333	5.991465	0.1888756	Accept H_0 & Reject Alternative H1

Table:Chi_Square_Test_Results

a. State the hypotheses and identify the claim:

- Null Hypothesis (H0): The distribution of wins by decade is equal (no difference).
- Alternative Hypothesis (H1): The distribution of wins by decade is not equal (there is a difference).

b. Find the critical value ($\alpha = 0.05$):

To find the critical value, we need to specify the degrees of freedom. For a Chi-Square Goodness-of-Fit test with k categories, the degrees of freedom (df) is given by (k - 1). In this case, k is the number of decades.

c. Compute the test value:

The test value for a Chi-Square Goodness-of-Fit test can be calculated using the formula:

$$\chi 2 = \sum Ei(Oi - Ei)2$$

Oi is the observed frequency and *Ei* is the expected frequency for each category.

d. Make the decision:

Compare the test value with the critical value. If the test value is greater than the critical value, reject the null hypothesis.

e. Does comparing the critical value with the test value provide the same result as comparing the p-value from R with the significance level?

Yes, both methods should lead to the same result. If the p-value from R is less than the significance level (α), it implies rejecting the null hypothesis. If the test value is greater than the critical value, it also implies rejecting the null hypothesis. These are two equivalent ways of reaching the same conclusion.

The test statistic was calculated to be approximately 3.33, and with 2 degrees of freedom, the critical value at a 5% significance level was found to be 5.99. Comparing the test statistic to the critical value, it was determined that the test statistic did not exceed the critical value.

Therefore, based on the test results, we do not have sufficient evidence to reject the null hypothesis. In other words, we accept the null hypothesis and conclude that there is no significant association between decades and the number of wins in the baseball dataset. The p-value of 0.1888756 further supports this conclusion, as it is greater than the significance level of 0.05.

In summary, the analysis suggests that the observed distribution of wins across decades is consistent with what would be expected under the assumption of independence between these variables.

Dataset-Crop

- Null Hypothesis (H0): Density and Fertilizer have no impact on the mean plant growth.
- Alternative Hypothesis (H1): There is an interaction between the two factors in affecting plant growth.

Summarize the Results:

S_NO	Df	Sum_Sq	Mean_Sq	F-Value	Pr..F	Decision
density	1	5.122	5.122	15.195	0.000186	Reject the null hypothesis for density (significant effect
fertilizer	2	6.068	3.034	9.001	0.000273	Reject the null hypothesis for fertilizer (significant effect
density:fertilizer	2	0.428	0.214	0.635	0.532500	Fail to reject the null hypothesis for interaction (no significant interaction effect
Residuals	90	30.337	0.337	NA	NA	NA

Table: ANOVA_Test_Results

S_NO	Diff	lwr	upr	p_adj
comparison between group2-group1	0.461956	0.2082555	0.7156566	0.0004845

Table:Tukey's_Test_For_Density

S_NO	Diff	lwr	upr	p_adj
comparison between group2-group1	0.1761687	-0.1937190	0.5460564	0.4954705
comparison between group3-group1	0.5991256	0.2292379	0.9690133	0.0006125
comparison between group3-group2	0.4229569	0.0530692	0.7928445	0.0208735

Table: Tukey's_Test_For_Fertilizer

Tukey's Test for Density:

- The Tukey's test for density is comparing the levels 1 and 2.
    - The difference in yield between density level 2 and level 1 is significant (p = 0.0004845).
    - The confidence interval (95%) for the difference in means between density level 2 and level 1 is [0.2083, 0.7157]. This means that we are 95% confident that the true difference in means lies within this interval, and since it does not include zero, the difference is considered significant.

Tukey's Test for Fertilizer:

- The Tukey's test for fertilizer is comparing levels 2, 3, and 1.
    - The difference in yield between fertilizer level 2 and level 1 is not significant (p = 0.4954705).
    - The difference in yield between fertilizer level 3 and level 1 is significant (p = 0.0006125).
    - The difference in yield between fertilizer level 3 and level 2 is significant (p = 0.0208735).
    - The confidence intervals (95%) for the differences in means provide a range of values where we can be 95% confident that the true difference lies. For example, the interval for the difference between fertilizer level 3 and level 1 is [0.2292, 0.9690].

In summary, the Tukey's test for density indicates a significant difference in yield between density levels 2 and 1. For the fertilizer factor, there are significant differences in yield between levels 3 and 1, and between levels 3 and 2. The intervals provide additional information about the magnitude and uncertainty of these differences.

CONCLUSION

The exploration of hypothesis testing through Chi-Square and ANOVA illuminates the diverse applications of statistical methods in different scenarios. In the Chi-Square analysis of blood type distribution, the results emphasize the importance of scrutinizing hospital patient demographics against the general population. The absence of significant differences supports the notion that blood type distributions align, validating the null hypothesis. This underlines the significance of statistical tools in health-related inquiries, aiding in evidence-based decision-making.

Moving on to ANOVA, the investigation into sodium amounts among condiments, cereals, and desserts unravels insights into dietary variations. The nuanced exploration of mean differences reveals potential areas for dietary interventions. Statistical methods empower researchers and policymakers to discern subtle distinctions, guiding interventions for public health improvement. Dataset-Baseball's Chi-Square Goodness-of-Fit test dispels notions of a profound association between decades and baseball wins. While the sport evolves, its essence remains relatively constant across different time periods, highlighting the robustness of certain aspects in the face of change. Statistical analyses, as showcased in this section, serve as a bridge between eras, allowing us to appreciate continuity and change simultaneously.

Dataset-Crop's ANOVA examination brings agriculture into the statistical realm, scrutinizing the effects of density and fertilizer on plant growth. The results, or lack thereof, challenge preconceptions and encourage a nuanced understanding of agricultural dynamics. By applying statistical rigor to agricultural studies, researchers can refine cultivation practices and contribute to sustainable food production.

In essence, this report underscores the versatility and potency of statistical hypothesis testing. Whether unraveling the mysteries of blood type distributions, dietary patterns, sports phenomena, or agricultural dynamics, statistical tools provide a lens through which we can discern patterns, dispel myths, and make informed decisions. The collaborative efforts of domain experts and statisticians are paramount in extracting meaningful insights that shape our understanding of the world.

CITATIONS

Hypothesis Testing — Analysis of Variance (ANOVA)

- https://medium.com/analytics-vidhya/hypothesis-testing-analysis-of-variance-anova-52c3df0fbc80

Chi-Square Test for Feature Selection – Mathematical Explanation

- https://www.geeksforgeeks.org/chi-square-test-for-feature-selection-mathematical-explanation/

Tukey Test for Pairwise Mean Comparisons

- https://stats.libretexts.org/Bookshelves/Advanced_Statistics/Analysis_of_Variance_and_Design_of_Experiments/02%3A_ANOVA_Foundations/2.03%3A_Tukey_Test_for_Pairwise_Mean_Comparisons

One-way v/s Two-way ANOVA

- https://www.scribbr.com/frequently-asked-questions/one-way-vs-two-way-anova/#:~:text=two%2Dway%20ANOVA%3F,What%20is%20the%20difference%20between%20a%20one%2Dway%20and%20a,two%2Dway%20ANOVA%20has%20two.