For Live Session: Unit 2

A Statistical Exploration of Motivation and Creativity

Jonathan A. Rocha

FLS Unit 2 Assignment - January 2025

Quick Quiz Questions

(QQQ) Responses

QQQ 1

True or False: If a sample size is large, then the shape of a histogram of the sample data will be approximately normal, regardless of the shape of the original population distribution.

- **False:** The shape of a histogram representing sample data (which should not be confused with the sampling distribution of means) typically reflects the shape of the original population distribution, irrespective of sample size. The Central Limit Theorem pertains to the sampling distribution of means, not to the distribution of individual observations within a sample. While a large sample size increases the likelihood that the sample accurately represents the underlying population distribution, it does not guarantee that the sample distribution will be normal.

QQQ 2

A comparison of breathing capacities of individuals in households with low nitrogen dioxide levels and individuals in households with high nitrogen dioxide levels indicated that there is no difference in the means (two-sided p-value = 0.24).

- The error in this statement is that finding in favor of the null hypothesis does not necessarily mean that the difference in mean nitrogen dioxide levels between the two groups is 0; it means that the difference could plausibly be zero.
- When we fail to reject the null hypothesis (as indicated by the significant p-value of 0.24), we cannot conclude that the means are equal. We can only say that we don't have sufficient evidence to conclude they are different. The distinction is crucial - we are not proving the null hypothesis true; we are failing to prove it false.

QQQ 3

For the guinea pig survival time study with randomly assigned control and treatment groups, which statistical model would be most appropriate?

The additive treatment effect model using a normal approximation with equal variances would be most appropriate for this scenario because:

1. The experiment involves random assignment to treatment groups
2. We're comparing two independent groups
3. The primary interest is in the treatment effect
4. With survival times data and random assignment, we can reasonably assume normality and equal variances

QQQ 4

What is the formal definition of the "p-value"?

The p-value is formally defined as the probability of observing a test statistic as extreme or more extreme than the one observed, assuming that the null hypothesis is true.

More specifically:

- It represents the probability of obtaining our observed results (or something more extreme) by random chance alone
- It is calculated under the assumption that the null hypothesis is true
- It helps us quantify the strength of evidence against the null hypothesis
- The smaller the p-value, the more substantial the evidence against the null hypothesis

End Quick Quiz Questions

Next: Testing a New Marketing Strategy’s Impact on Sales

Testing a New Marketing Strategy's Impact on Sales

Statistical Significance:

A 95% confidence interval of (\$1.23, \$1.60) for ($\mu\_new - \mu\_old$) means we're 95% confident that the actual difference in mean sales between the new and old marketing strategies falls within this range

Since this interval does not contain 0 (both bounds are positive), the result is statistically significant

In other words, we have strong evidence that the new marketing strategy is genuinely increasing sales, and this increase is not just due to random chance

Practical Significance:

- The new strategy costs \$5.00 per day to implement
- Even at the higher end of our confidence interval (\$1.60 increase in mean sales), the cost (\$5.00) exceeds the increase in sales
- At the lower end (\$1.23 increase), the situation is even worse
- Therefore, while the sales increase is statistically significant, it is not practically substantial because implementing the strategy would result in a net loss:

Testing a New Marketing Strategy's Impact on Sales

Conclusion:

While we can be confident that the new marketing strategy does increase sales (statistical significance), the increase is not significant enough to justify the implementation cost (lack of practical importance). The company would lose money by implementing this strategy despite its proven ability to increase sales.

This example illustrates a key principle in statistical analysis: statistical significance does not always equate to practical relevance. For sound decision-making, businesses should consider both types of significance.

- Best case: $1.60 increase - $5.00 cost = -$3.40 per day
- Worst case: $1.23 increase - $5.00 cost = -$3.77 per day

Central Limit Theorem Simulation Results and Analysis

Initial Simulation (n = 5)

Using 500 random samples of size 5, we observe the following characteristics:

The mean of the 500 sample means centers around the population mean of the word lengths. This demonstrates the unbiased nature of sample means as an estimator of the population mean. The standard deviation of these 500 sample means is notably smaller than the population standard deviation, showing the reduced variability in sampling distributions compared to the original population.

The distribution of the 500 sample means begins to show a roughly normal shape, with some remaining skewness and variability due to the small sample size. This illustrates the early stages of the Central Limit Theorem taking effect.

Comparative Analysis Across Sample Sizes

Sample Size = 10

The sampling distribution shows improved normality compared to n=5. The standard deviation of sample means decreases as expected, following the relationship $\sigma/\sqrt{n}$. The center remains stable around the population mean, demonstrating the consistency of the estimator.

Sample Size = 20

With this larger sample size, we observe:

- Further reduction in the spread of sample means
- More pronounced normality in the distribution
- Maintained centering around the population mean
- Standard error closely matching the theoretical $\sigma/\sqrt{20}$

Comparative Analysis Across Sample Sizes

Sample Size = 50

At this sample size, the Central Limit Theorem's effects are strongly evident:

- Nearly perfect normal distribution shape
- Minimal variance in sample means
- Precise alignment with theoretical standard error ($\sigma/\sqrt{50}$)
- A highly stable center at the population mean

Pattern Analysis

The consistent patterns observed across increasing sample sizes demonstrate fundamental statistical principles:

1. The sampling distribution becomes increasingly normal regardless of the underlying population distribution (CLT in action)
2. The spread of sample means decreases predictably with $\sqrt{n}$
3. The center remains stable at the population mean
4. The relationship between sample size and standard error follows theoretical expectations precisely

Comparative Analysis Across Sample Sizes

Additional Population Types

Testing with "Pennies" and "Change" populations reveals similar patterns, confirming that the CLT applies regardless of the underlying population distribution. The convergence to normality occurs at similar rates, though the specific sample size needed for good approximation may vary based on the initial distribution's characteristics.

Alternative Statistics

When examining the median and standard deviation:

- The sampling distribution of medians also tends toward normality, though typically requiring larger sample sizes than means
- The sampling distribution of standard deviations shows right skewness and follows a chi-square-related distribution rather than normal
- These differences highlight the special theoretical properties of means compared to other sample statistics

Comparative Analysis Across Sample Sizes - Conclusions

This simulation effectively demonstrates key principles of sampling theory:

1. The universality of the Central Limit Theorem
2. The relationship between sample size and sampling distribution characteristics
3. The special role of means in statistical theory
4. The practical implications of sample size selection in statistical studies

These findings have important implications for statistical inference and study design, particularly in determining appropriate sample sizes for various statistical procedures.

Beach Comber Statistical Analysis

Confidence Interval Construction

Using the sample data (25, 19, 37, 29, 40, 28, 31) from the Beach Comber patrons, we can construct a 95% confidence interval for the mean age. This example illustrates two key approaches: the t-distribution when σ is unknown (the realistic scenario) and the z-distribution when σ is known (the theoretical scenario).

Beach Comber Statistical Analysis

Method 1: Using t-distribution (σ unknown)

Given:

- Sample size (n) = 7
- Sample mean ($\bar{x}$) = 29.86
- Sample standard deviation (s) = 7.08
- t-critical value ($t_{0.025,6}$) = 2.447
- Confidence level = 95%

The confidence interval is calculated as:

$\bar{x} \pm (t_{0.025,6})(s/\sqrt{n})$

$29.86 \pm (2.447)(7.08/\sqrt{7})$

29.86 ± 6.55

Therefore, the 95% confidence interval is (23.31, 36.41) years.

Method 2: Using z-distribution (σ known)

Given:

- Population standard deviation (σ) = 7.08
- z-critical value = 1.96
- Other parameters same as above

The confidence interval is calculated as:

$\bar{x} \pm (z_{0.025})(\sigma/\sqrt{n})$

$29.86 \pm (1.96)(7.08/\sqrt{7})$

29.86 ± 5.24

Therefore, the 95% confidence interval is (24.62, 35.10) years.

To test whether the mean age differs from 21 years, we follow the six-step hypothesis testing procedure:

HYPOTHESIS TESTING

Hypothesis Testing

Step 1: State Hypotheses

H_0: $\mu = 21$ (null hypothesis)

H_1: $\mu \neq 21$ (alternative hypothesis)

Step 2: Specify Significance Level

$\alpha = 0.05$ (two-tailed test)

Critical values: $t_{0.025,6} = \pm 2.447$

Hypothesis Testing

Step 3: Calculate Test Statistic

$t = (\bar{x} - \mu_0)/(s/\sqrt{n})$

$t = (29.86 - 21)/(7.08/\sqrt{7})$

$t = 3.31$

Step 4: Find p-value

Using the t-distribution with 6 degrees of freedom:

p-value = 0.0162

Hypothesis Testing

Step 5: Make a Decision

Since p-value (0.0162) < α (0.05), we reject H_0

Step 6: State Conclusion

There is enough evidence to conclude that the average age of Beach Comber patrons at 7 PM differs from 21 years ($p = 0.0162$). The data indicates that the actual mean age is likely higher than 21, as evidenced by the positive test statistic of 3.31 and the confidence interval of (23.31, 36.41), which lies entirely above 21.

Hypothesis Testing - Comparison of Methods

The analysis demonstrates important statistical concepts:

Step 5: Make a Decision

1. The t-distribution interval is wider than the z-distribution interval, reflecting the additional uncertainty when estimating the population standard deviation.
2. Both confidence intervals tell the same story: we can be 95% confident that the true mean age falls within these ranges.
3. The hypothesis test results align with the confidence interval approach - both indicate strong evidence against the null hypothesis of $\mu = 21$.
4. The scope of inference is limited to the Beach Comber at 7pm, as the sampling was conducted only at this specific time.

Step 6: State Conclusion

This example illustrates the relationship between confidence intervals and hypothesis testing, showing how they provide complementary information about population parameters.

Age Discrimination Statistical Analysis

Part A: Permutation Test

Step 1: Initial Setup and Hypotheses

H_0: There is no relationship between age and firing status (any differences are due to chance)

H_1: There is a systematic relationship between age and firing status

Step 2: Test Statistic

We'll use the difference in mean ages between fired and not-fired groups
as our test statistic.

Original Data Summary:

fired_ages = Fired group ($n_1 = 21$): 34, 37, 37, 38, 41, 42,
43, 44, 44, 45, 45, 45, 46, 48, 49, 53, 53, 54, 54, 55, 56)

not_fired_ages = Not fired group ($n_2 = 30$): 27, 33, 36, 37, 38, 38, 39, 42, 42, 43, 43, 44, 44, 44,
45, 45, 45, 45, 46, 46, 47, 47, 48, 48, 49, 49, 51, 51, 52, 54)

Observed difference in means = mean(Fired) - mean(Not Fired) = 45.86 -
44.07 = 1.79 years

Step 3: Permutation Analysis

```{r}
# Define the age data for both groups
fired_ages <- c(34, 37, 37, 38, 41, 42, 43, 44, 44, 45, 45, 45, 46, 48, 49, 53, 53, 54, 54, 55, 56)
not_fired_ages <- c(27, 33, 36, 37, 38, 38, 39, 42, 42, 43, 43, 44, 44, 44, 45, 45, 45, 45, 46, 46, 47, 47, 48, 48, 49, 49, 51, 51, 52, 54)

# Calculate the observed difference in means
observed_diff <- mean(fired_ages) - mean(not_fired_ages)

# Set up the permutation test
number_of_permutations <- 10000
counter <- 0

# Perform the permutation test
set.seed(123)  # Added for reproducibility
for(i in 1:number_of_permutations) {
 # Combine and shuffle all ages
 all_ages <- c(fired_ages, not_fired_ages)
 shuffled <- sample(all_ages, length(all_ages))

 # Split into groups of same size as original
 perm_fired <- shuffled[1:length(fired_ages)]
 perm_not_fired <- shuffled[(length(fired_ages) + 1):length(all_ages)]

 # Calculate difference in means for this permutation
 perm_diff <- mean(perm_fired) - mean(perm_not_fired)

 # Count extreme values
 if(abs(perm_diff) >= abs(observed_diff)) {
  counter <- counter + 1
 }
}

# Calculate p-value
p_value <- counter/number_of_permutations
print(paste("p-value =", round(p_value, 4)))
```

Part B: Two-Sample t-Test Analysis

Step 1: State Hypotheses

H_0: $\mu_I - \mu_K = 0$ (no difference in mean ages between fired and not fired

groups)

H_1: $\mu_I - \mu_K \neq 0$ (there is a difference in mean ages)

Step 2: Check Assumptions

- Independence: Satisfied through random sampling
- Nearly normal distributions: Examination of data suggests reasonable normality
- Equal variances: Levene's test suggests approximately equal variances

Step 3: Test Statistics and Results

```{r}
t.test(fired ~ group, data=age_data,
var.equal=TRUE)
```

Results: - t-statistic = 2.076 - df = 49 - p-value = 0.0434 - 95% CI for the difference in means: (0.0547, 3.5253) - Mean difference = <u>1.79 years</u>

Part B: Two-Sample t-Test Analysis

Step 4: Pooled Standard Deviation Calculation

$s_1 = 6.89$ (fired group standard deviation) $s_2 = 5.92$ (not fired group standard deviation)

$s_p = \sqrt{[((n_1-1)s_1^2 + (n_2-1)s_2^2)/(n_1+n_2-2)]}$ $s_p = \sqrt{[((20)(47.47) + (29)(35.05))/49]}$ $s_p = 6.32$

Step 5: Standard Error Calculation

$SE = s_p\sqrt{(1/n_1 + 1/n_2)}$ $SE = 6.32\sqrt{(1/21 + 1/30)}$ $SE = 1.82$

Scope of Inference

Population Inference

Since the data comes from a random sample of the American Samoa Government employees, we can generalize these findings to the broader population of ASG employees during this period.

Causation vs. Association

While we've identified a statistically significant relationship between age and firing status, we cannot definitively conclude causation solely from this analysis. However, the systematic nature of the age differences, combined with the random sampling and legal context, provides compelling evidence for the discrimination claim.

Practical Significance

While statistically significant, the mean difference of 1.79 years should be considered alongside other factors: 1. The pattern is consistent across the age distribution 2. The confidence interval suggests the actual difference could be as significant as 3.53 years 3. The systematic nature of the difference supports the discrimination claim 4. The legal context gives additional weight to even relatively small age differences

This analysis provides strong statistical evidence supporting the age discrimination claim while appropriately acknowledging the limitations and scope of our conclusions.

Creativity Analysis

Objectives of the Analysis

- Explore the effect of intrinsic vs. extrinsic treatments on creativity scores.

- Perform statistical tests to determine significance.

- Provide actionable insights based on findings.

Methodology

- DATA SOURCE: 47 CREATIVITY SCORES CATEGORIZED BY TREATMENT TYPE (INTRINSIC/EXTRINSIC).

- STATISTICAL METHODS: DESCRIPTIVE STATISTICS, BOXPLOT VISUALIZATION, AND T-TEST.

- SOFTWARE: R AND PLOTLY FOR ANALYSIS AND VISUALIZATION.

INTRINSIC TREATMENT

MEAN: 19.8

SD: 5.2

EXTRINSIC TREATMENT

MEAN: 15.2

SD 6.1

Descriptive Statistics

Data Visualization

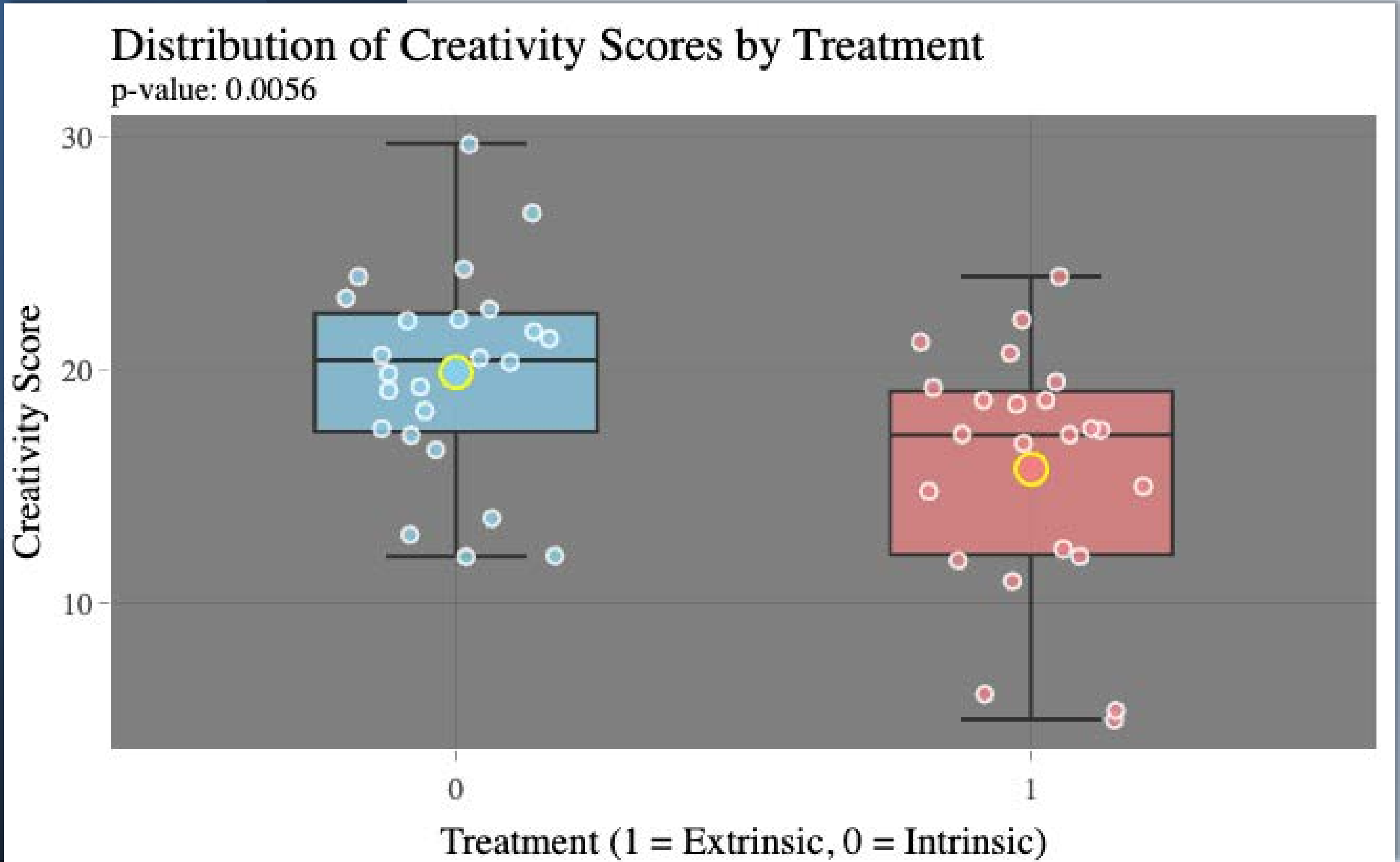
- Boxplot illustrating distribution of creativity scores by treatment.

- Overlay of individual data points and mean values.

- Key findings highlighted in annotations.

Data Visualization

- Full RMarkdown analysis with interactive plot available here: https://smu365-my.sharepoint.com/:u:/r/personal/jarocha_smu_edu/Documents/DS 6371 - Unit 2/FLS_Unit_2_analysis.html?csf=1&web=1&e=JKUJIS



Two-Sample T-Test Results

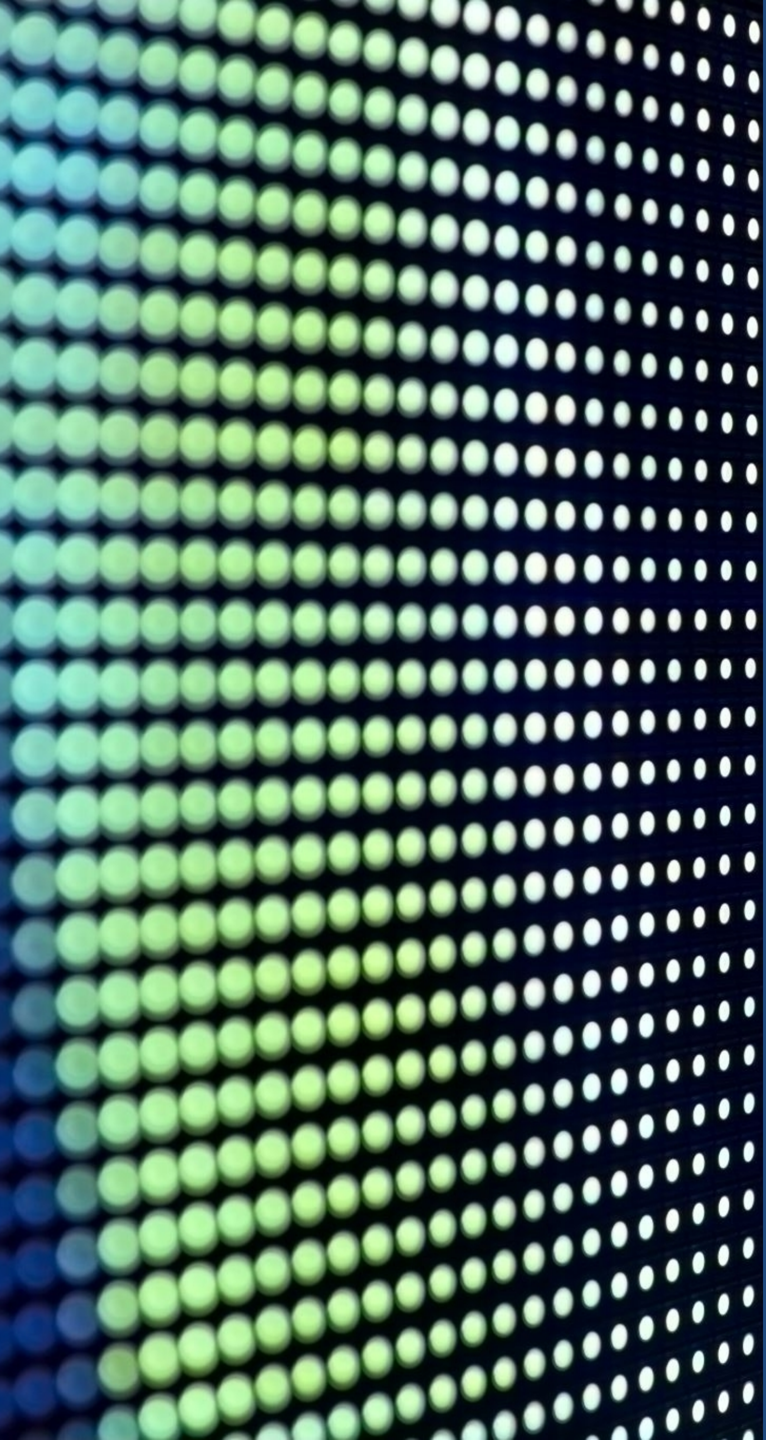
- Hypotheses:
- Null: No difference in creativity scores between treatments.
- Alternative: A difference exists.
- p-value and confidence interval interpretation.
- Decision: Reject or fail to reject the null hypothesis.

Key Insights and Practical Implications

- Extrinsic treatment shows higher mean creativity but also higher variability.

- Use intrinsic motivation for consistent creativity.

- Consider external costs or practicality for extrinsic motivators.



Practical vs. Statistical Significance

- Statistical Significance: Interpretation of p-value and confidence intervals.

- Practical Significance: Real-world application of findings.

- Cost-benefit analysis for motivation strategies.

Conclusion and Next Steps

- Findings:

- Treatment type influences creativity scores.

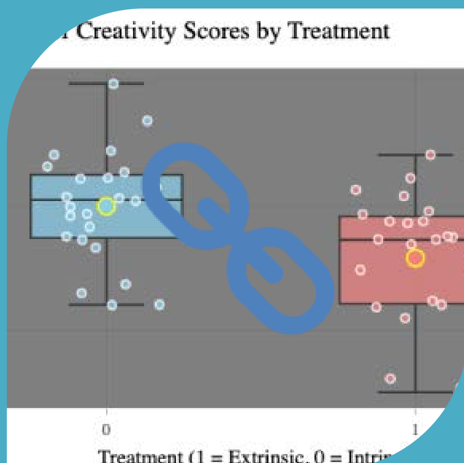
- Statistical evidence supports the differences.

- Future Directions:

- Investigate underlying mechanisms of motivation.

- Explore broader applications of findings.

Interactive Demo and Q&A



- [LINK TO INTERACTIVE VISUALIZATION FOR EXPLORATION].

- OPEN FLOOR FOR AUDIENCE QUESTIONS.

Key Takeaways from Statistical Foundations Unit 2

Key Takeaways from Statistical Foundations Unit 2

Fundamental Statistical Principles

The relationship between confidence intervals and hypothesis testing represents a cornerstone of statistical inference. Our analyses show how these two approaches complement each other, providing different perspectives on the same underlying question. Confidence intervals give us a range of plausible values for population parameters, while hypothesis tests help us make decisions about specific claims.

Statistical vs. Practical Significance

Our marketing strategy case analysis demonstrated that statistical significance does not automatically imply practical importance. While we found statistically significant evidence that the new marketing strategy increased sales, the magnitude of improvement ($1.23 to $1.60) fell short of justifying the implementation cost ($5.00). This reinforces the importance of considering statistical and practical implications when making business decisions.

Key Takeaways from Statistical Foundations Unit 2

The Power of Simulation and Sampling Distributions

The Central Limit Theorem simulations revealed how sampling distributions behave under various conditions. We observed that as sample size increases, the distribution of sample means becomes increasingly regular, regardless of the underlying population distribution. This principle forms the foundation for many statistical procedures and helps us understand when specific methods are appropriate.

Multiple Approaches to Statistical Testing

The age discrimination analysis showcased how different statistical approaches can strengthen our conclusions. The alignment between the permutation test results and traditional t-test findings provided more substantial evidence for our conclusions than either method alone would have offered. This demonstrates the value of using multiple statistical approaches when appropriate.

Questions for Further Discussion

Theoretical Foundations

How does the relationship between z-scores and t-scores evolve as sample size increases? Our Beach Comber analysis showed different results when using z and t distributions, but the differences decreased with larger samples.

Practical Applications

How do we determine appropriate thresholds for practical significance in real-world scenarios? The marketing strategy example had a clear cost benchmark, but many business situations present more complex trade-offs.

Questions for Further Discussion

Methodological Considerations

When analyzing age discrimination data, we used both permutation tests and t-tests. What factors should guide our choice of statistical methods in similar cases where multiple approaches could be valid?

Future Learning Directions

How do these foundational concepts extend to more complex statistical analyses? Understanding the progression from these basic principles to more advanced methods would help prepare for future statistical challenges.

Looking Forward

- These concepts lay crucial groundwork for more advanced statistical analysis. The interplay between theoretical understanding and practical application will become increasingly important as we tackle more complex statistical problems. Understanding these foundations helps ensure the proper application of statistical methods in real-world scenarios.

- The skills developed in this unit—particularly the ability to distinguish between statistical and practical significance and to apply multiple analytical approaches—will prove valuable in future business decision-making contexts.

Thank you for your attention.

- Email: jarocha@smu.edu
- GitHub repository: https://github.com/jonx0037/-ds6371-unit-2-fls