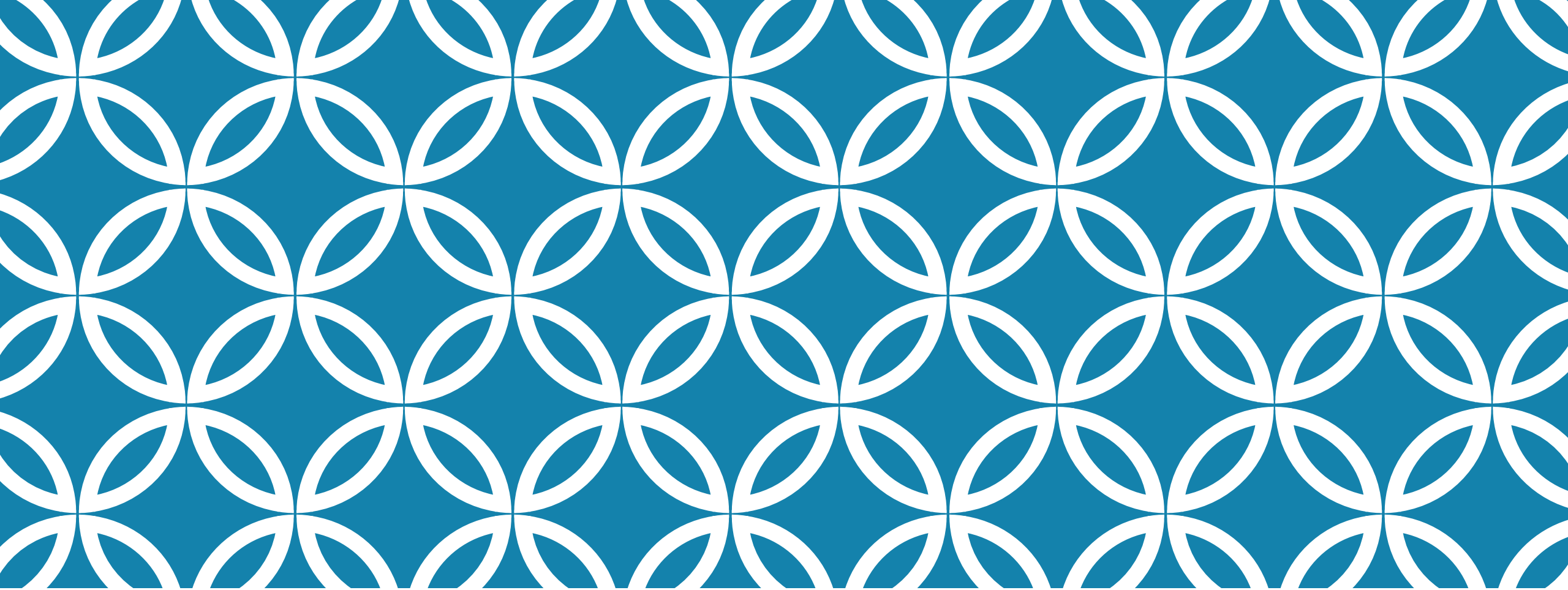

BASICS OF HYPOTHESIS TESTING

Sharath Srivatsa

WHAT IS HYPOTHESIS TESTS

1. Hypothesis tests are a statistical analysis technique.
2. Used to make a statistical decision on a sample of data
3. These tests are conducted to test one or more statistical propositions (hypotheses)
4. Hypothesis testing is used in scientific research, business analysis, medicine, engineering and many other fields.
5. These tests are an important tool for determining whether data differ in a statistically significant way or support a hypothesis

Hypothesis testing is a formal procedure within inferential statistics to test a claim about the population using sample data

It helps determine whether observed differences or relationships in the sample are ***likely due to chance or represent true effects in the population****.*

TWO DIFFERENT APPROACHES TO HYPOTHESIS TESTING BASED ON DATA

- **Parametric Testing**
  - Assumptions about the data
    - Compared groups are normally distributed or distributed in a symmetrical bell curve
    - Compared groups have similar or same variances
  - ***Shapiro Wilk*** and ***Kolmogorov-Smirnov*** tests are applied for the assumption of normality
    - The Kolmogorov-Smirnov Test is more powerful for large sample sizes and is often used with large data sets
    - The Shapiro-Wilk Test is a more sensitive test, especially for small sample sizes
  - ***Levene's test*** is applied for variance homogeneity

- **Non-parametric Testing or Distribution-free Tests**
  - If there is data that is not normally distributed
  - If there is a very small sample size
  - If there is ordinal, ranked data or outliers are present
  - If the variance is not homogeneous

- **Statistical Significance *p-value***
  - ***Statistical tests output p-value or significance***
  - ***An important significance is 95% or p-value of 0.05***
  - ***Any other significance can be adopted depending on problem being solved***

Paired Test	Unpaired Test
A statistical test that compares the means and standard deviations of two related samples	A statistical test that compares the means and standard deviations of two unrelated or independent samples
Dependent test	Independent test

COMMONLY USED SAMPLE STATISTICS, CATEGORIZED BASED ON THEIR FUNCTION

Measures of Central Tendency:

- **Mean:** The average value of the data.
- **Median:** The middle value when the data is sorted.
- **Mode:** The most frequent value in the data.

Measures of Dispersion:

- **Variance:** The average of the squared differences from the mean.
- **Standard Deviation:** The square root of the variance, representing the typical deviation from the mean.
- **Range:** The difference between the maximum and minimum values.

Interquartile Range (IQR): The difference between the 75th percentile (Q3) and the 25th percentile (Q1), representing the spread of the middle 50% of the data.

Specific sample statistics used in hypothesis testing will depend on the nature of the data and the research question being investigated

Measures of Relative Standing:

- **Percentiles:** The value below which a certain percentage of observations fall.

Z-scores: The number of standard deviations an observation is away from the mean.

Measures of Association:

Correlation Coefficient (Pearson's r, Spearman's rho): Measures the strength and direction of the linear relationship between two variables.

Other Sample Statistics:

- **Proportions:** The fraction or percentage of occurrences in a particular category.
- **Counts:** The number of observations in a particular category or group.

Sample Size (n): The total number of observations in the sample.

PEARSON'S CORRELATION VS SPEARMAN CORRELATION

Pearson's correlation

Pearson's correlation (named after Karl Pearson) ***is used to show linear relationship between two variables***

Pearson's Correlation returns a value ***between [-1, 1], with 1 meaning full positive correlation and -1 full negative correlation***

Pearson's Correlation ***uses mean and standard deviation in the calculation***, which implies that it is a parametric method and ***it assumes a Gaussian-like distribution for the data***, hence used in ***Parametric Hypothesis Testing***

Spearman Correlation

Spearman correlation (named after Charles Spearman) is the ***non-parametric version of the Pearson's correlations and evaluates the monotonic relationship***

A ***monotonic relationship*** is a relationship ***as the value of one variable increases, the value of the other variable increases or decreases but not at a constant rate like in a linear relationship*** where the rate of increase or decrease will be constant

Used when the relationship ***between the two variables are non-linear and variables have non-Gaussian distribution***, hence used in ***Non-parametric Hypothesis Testing***

Similar to Pearson's Correlation, ***Spearman also returns a value between [-1,1] for full negative correlation and full positive correlation, respectively***

HYPOTHESIS TESTING STEPS

Step 1: Defining the NULL and Alternate Hypothesis

Step 2: Fixing the level of significance(commonly the Alpha value)

Step 3: Performing the statistical test

Step 4: Concluding the result based on P-value

Performing the statistical tests:

Widely used Parametric tests are the ***paired or unpaired T-test, ANOVA, F-test, and Z-test***

Widely used Non-parametric tests are the ***Chi-square, Mann-Whitney U-test, and Kruskal-Wallis H-test***

ONE TAILED AND TWO TAILED TESTS

One-Tailed Hypothesis Test

- **Directional:** The alternative hypothesis specifies a direction of the effect or difference. It predicts that the parameter of interest is either greater than or less than the value specified in the null hypothesis.
- **Critical Region:** The rejection region is located entirely in one tail of the sampling distribution (either the left tail or the right tail).
- **Example:**
- Null hypothesis (H0): The new drug is not more effective than the standard drug.
- Alternative hypothesis (H1): The new drug is more effective than the standard drug.

Two-Tailed Hypothesis Test

- **Non-directional:** The alternative hypothesis simply states that there is a difference between the parameter of interest and the value specified in the null hypothesis, without specifying the direction of the difference.
- **Critical Regions:** The rejection regions are located in both tails of the sampling distribution.
- **Example:**
- Null hypothesis (H0): There is no difference in the mean scores of two groups.
- Alternative hypothesis (H1): There is a difference in the mean scores of the two groups.

Choosing the Right Test:

- **One-tailed Test:** Appropriate when you have a strong prior belief or theoretical reason to expect the effect or difference to be in a specific direction.

Two-tailed Test: Generally preferred when you are exploring the possibility of a difference in either direction or when you have no specific directional hypothesis.

The choice between one-tailed and two-tailed tests should be made before conducting the analysis, based on the research question and prior knowledge

PARAMETRIC TESTS

- **T-test** is used to determine if there is a significant difference ***between the means of two groups when variance is unknown***
- **Z-test** is used to determine if there is a significant difference ***between the means of two groups when variance is unknown and sample size is large***
- **Analysis of variance (ANOVA)** is used to determine if there is a significant difference ***between the means of more than two groups***
- **F-test** is used to determine if there is a significant difference ***between variances of two samples or the ratio of variances between multiple samples***

NON-PARAMETRIC TESTS

- The **Chi-Squared test** is commonly ***used to test the hypothesis of independence between two variables*** or if the difference between the two is due to chance or a relationship exists
  - For example, it can be used to determine whether there is a relationship between gender and political affiliation or whether there is a difference in the frequency of certain diseases among different age groups
- **Mann-Whitney U test** is used to compare differences between two independent groups ***when the dependent variable is either ordinal or continuous, but not normally distributed***
- **Kruskal-Wallis H test** is used to determine if there are statistically significant differences ***between two or more groups of an independent continuous or ordinal variables***
  - It is considered the **nonparametric alternative to the one-way ANOVA**, and an extension of the Mann-Whitney U test to allow the comparison of more than two independent groups

CASE STUDY 1

A university professor gave online lectures instead of face-to-face classes due to Covid-19. Later, he uploaded recorded lectures to the cloud for students who followed the course asynchronously (those who did not attend the lesson but later watched the records). However, he believes that the students who attend class at the class time and participate in the process are more successful. Therefore, he recorded the average grades of the students at the end of the semester.

Conduct the hypothesis testing to check whether the professor's belief is statistically significant by using a 0.05 significance level to evaluate the null and alternative hypotheses

1. Parametric or Non-parametric test?
2. Pair or Unpaired test?
3. One-tailed or two tailed test?

CASE STUDY 2

A pediatrician wants to see the effect of formula consumption on the average monthly weight gain (in gr) of babies. For this reason, she collected data from three different groups. The first group is exclusively breastfed children (receives only breast milk), the second group is children who are fed with only formula and the last group is both formula and breastfed children

According to this information, conduct the hypothesis testing to check whether there is a difference between the average monthly gain of these three groups by using a 0.05 significance level

1. Parametric or Non-parametric test?
2. Hypothesis?

CASE STUDY 3

A human resource specialist working in a technology company is interested in the overwork time of different teams. To investigate whether there is a difference between overtime of the software development team and the test team, she selected 17 employees randomly in each of the two teams and recorded their weekly average overwork time in terms of an hour.

According to this information, conduct the hypothesis testing to check whether there is a difference between the overwork time of two teams by using a 0.05 significance level

1. Parametric or Non-parametric test?
2. Hypothesis?

MORE CASE STUDIES REFERENCE

https://medium.com/towards-data-science/hypothesis-testing-with-python-step-by-step-hands-on-tutorial-with-practical-examples-e805975ea96e

Q&A

THANK YOU