Day 8

02 June 2024 18:39

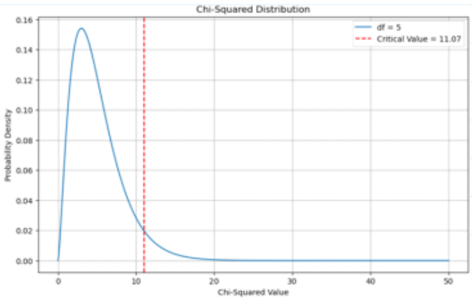
Df implications - chi2

Shape and Spread:

- With 1 degree of freedom (df = 1), the chi-squared distribution is highly skewed to the right.
- As the degrees of freedom increase, the distribution becomes less skewed.
- For df > 30, the chi-squared distribution approximates a normal distribution.
- Higher degrees of freedom lead to a broader spread of the distribution, meaning the distribution has a larger range of values.

Critical Values:

- critical value for the chi-squared statistic, which is used to determine statistical significance, depends on the degrees of freedom and the desired significance level (e.g., 0.05).
- For higher degrees of freedom, the critical value increases.
- means that with more degrees of freedom, you need a larger chi-squared statistic to reject the null hypothesis.

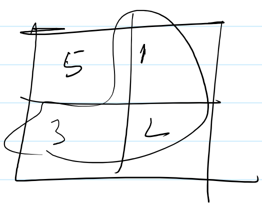


Assumptions

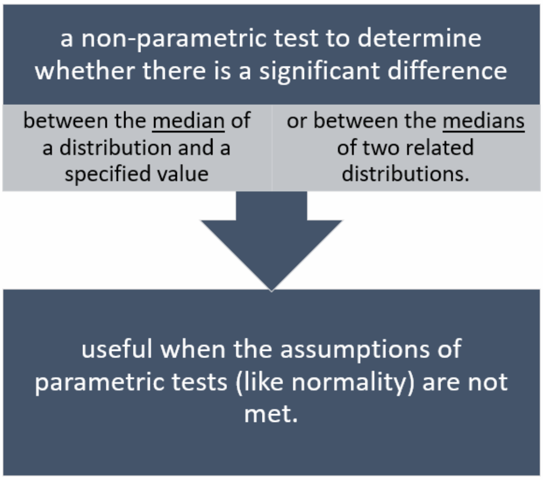
- Categorical data
- Random sampling
- Categories ME
- For chi2 - larger DF
- Expected freq > 5

Types of chi2

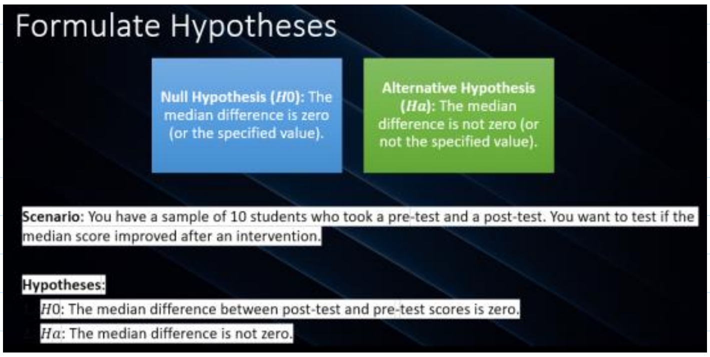
Test for Independence	Goodness-of-Fit Test	Test for Homogeneity
• assesses whether two categorical variables are independent of each other. • Testing if there is an association between gender (male/female) and preference for a product (like/dislike)	• determines if a sample data fits a population with a specific distribution. • compares the observed frequencies in each category with the expected frequencies based on a theoretical distribution. • Testing if a six-sided die is fair (each side has an equal probability of landing).	• tests whether different samples come from the same population or different populations. • Testing if different brands of cereal have the same proportion of customers who prefer them.



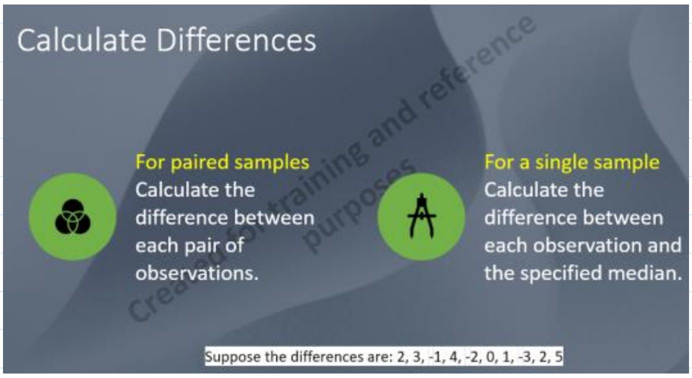
Sign test



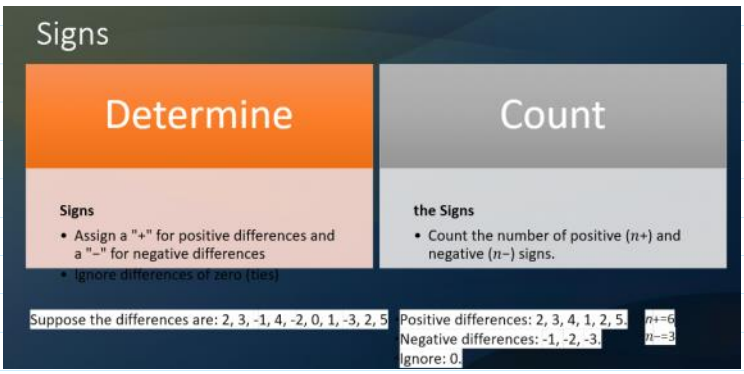
a non-parametric test to determine whether there is a significant difference
between the median of a distribution and a specified value
or between the medians of two related distributions.
useful when the assumptions of parametric tests (like normality) are not met.



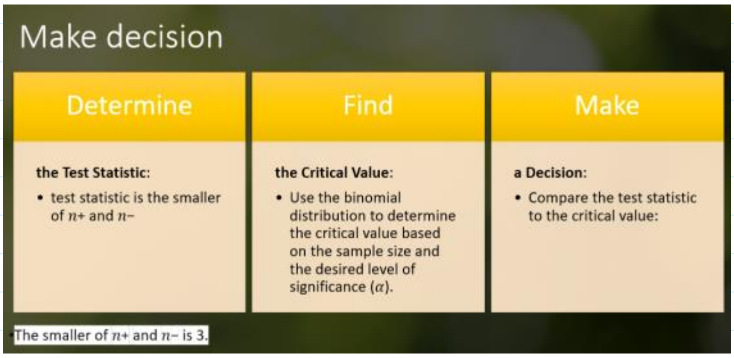
Formulate Hypotheses
Null Hypothesis ($H0$): The median difference is zero (or the specified value).
Alternative Hypothesis (Ha): The median difference is not zero (or not the specified value).
Scenario: You have a sample of 10 students who took a pre-test and a post-test. You want to test if the median score improved after an intervention.
Hypotheses:
$H0$: The median difference between post-test and pre-test scores is zero.
Ha: The median difference is not zero.

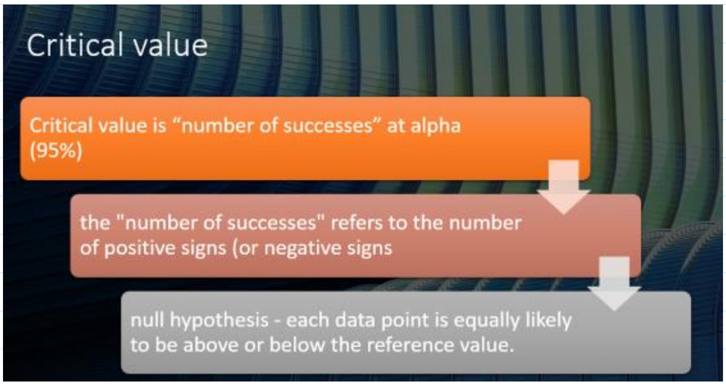


Calculate Differences
For paired samples
Calculate the difference between each pair of observations.
For a single sample
Calculate the difference between each observation and the specified median.
Suppose the differences are: 2, 3, -1, 4, -2, 0, 1, -3, 2, 5



Signs
Determine
Signs
• Assign a "+" for positive differences and a "−" for negative differences
• Ignore differences of zero (ties)
Count
the Signs
• Count the number of positive ($n+$) and negative ($n-$) signs.
Suppose the differences are: 2, 3, -1, 4, -2, 0, 1, -3, 2, 5
Positive differences: 2, 3, 4, 1, 2, 5.
Negative differences: -1, -2, -3.
Ignore: 0.
$n+=6$
$n-=3$





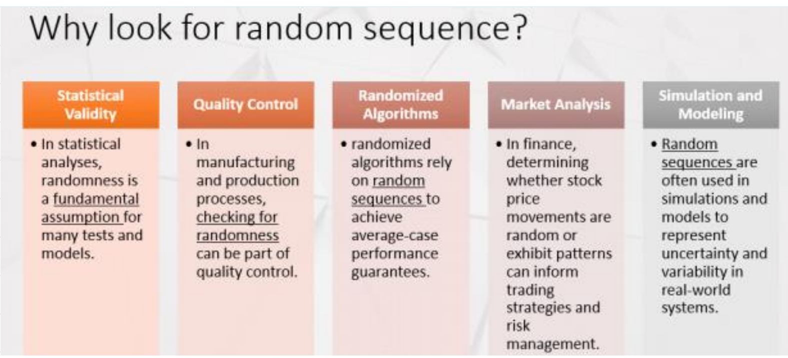
- A student tells her parents that the median rental rate for a studio apartment in Portland is $700. Her parents are skeptical and believe the rent is different.
- A random sample of studio rentals is taken from the internet; prices are listed below.
- Test the claim that there is a difference using $\alpha = 0.10$. Should the parents believe their daughter?

- Data : [700, 650, 800, 975, 855, 785, 759, 640, 950, 715, 825, 980, 895, 1025, 850, 915, 740, 985, 770, 785, 700, 925]

- A professor believes that a new online learning curriculum is <u>increasing</u> the median final exam score from the previous year, which was 75.
- A random sample of final exam scores were collected for students that went through the new curriculum.
- Test to see if the new curriculum is effective using $\alpha=0.05$

- Data = [78, 100, 75, 64, 87, 80, 72, 91, 89, 70, 82, 76]

Runs test



Why look for random sequence?

Statistical Validity	Quality Control	Randomized Algorithms	Market Analysis	Simulation and Modeling
• In statistical analyses, randomness is a fundamental assumption for many tests and models.	• In manufacturing and production processes, checking for randomness can be part of quality control.	• randomized algorithms rely on random sequences to achieve average-case performance guarantees.	• In finance, determining whether stock price movements are random or exhibit patterns can inform trading strategies and risk management.	• Random sequences are often used in simulations and models to represent uncertainty and variability in real-world systems.

Run - toss

- A run is defined as a succession of similar events preceded and followed by a different event.
- E.g. in a sequence of tosses of a coin, we may have
  - H T T H H T T T H T
- first toss is preceded and the last toss is followed by a "no event".
- sequence has 6 runs,
  - first with a length of 1,
  - Second , third with length 2
  - fourth length 3
  - fifth and sixth length 1

Example

If a sequence of numbers have too few runs, it is unlikely a real random sequence.

E.g. 0.08, 0.18, 0.23, 0.36, 0.42, 0.55, 0.63, 0.72, 0.89, 0.91

the sequence has 1 run, an up run.

not likely a random sequence.

Example

- Asimple statistical test of the random-walk theory is a runs test. For daily data, a run is defined as a sequence of days in which the stock price changes in the same direction.
- For example, consider the following combination of upward and downward price changes: ++−−+−+−−−++.
- A + sign means that the stock price increased, and a − sign means that the stock price decreased.
- 7 runs, in 12 observations

Created for training and reference purposes

Count the Number of Increases and Decreases

Number of increases (n1) = 10

Number of decreases (n2) = 10

Created for training and reference purposes

Calculate the Expected Number of Runs (E[R]) and Variance (Var[R])

$$E[R] = \frac{2n_1n_2}{n_1+n_2} + 1 = \frac{2\cdot10\cdot10}{10+10} + 1 = \frac{200}{20} + 1 = 11$$

$$\text{Var}[R] = \frac{2n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)} = \frac{2\cdot10\cdot10\cdot(2\cdot10\cdot10-10-10)}{(10+10)^2(10+10-1)} = \frac{200\cdot180}{400\cdot19} = \frac{36000}{7600} = 4.7368$$

Calculate the Test Statistic (Z), CV

$$Z = \frac{R-E[R]}{\sqrt{\text{Var}[R]}} = \frac{10-11}{\sqrt{4.7368}} = \frac{-1}{2.1766} \approx -0.4593$$

why we use standard normal distribution for critical values

- use of the standard normal distribution for critical values in the runs test is rooted in
  - the Central Limit Theorem (CLT)
  - the properties of large sample approximations.

Z test

What is Z test

used to determine whether two population means are different when the population variances are known

the sample size is large enough to assume normality.

particularly useful for hypothesis testing in situations where the sample size is relatively large.

Assumptions

1. **Normality**:
   1. population from which the samples are drawn should be normally distributed.
   2. for large samples, the Central Limit Theorem (CLT) allows the use of the Z-test even if the population distribution is not perfectly normal.
2. **Known Population Variance**:
   1. population variances should be known.
   2. If the population variances are unknown and estimated from the sample data, a t-test is typically used instead.
3. **Independence**: samples should be independently drawn from the population.
4. **Large Sample Size**:
   1. sample size should be sufficiently large (typically n > 30) for the Central Limit Theorem to hold, which justifies the normal approximation.
5. **Equal Variances (for Two-Sample Z-test)**:
   1. When comparing two samples, it is assumed that the population variances are equal.

One-Sample Z-Test

Used to determine whether the mean of a single sample is different from a known population mean

Formula:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Two-Sample Z-Test

- Used to compare the means of two independent samples to see
  - if they come from the same population or
  - if their means are significantly different.

Formula:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Feature	Z-test	t-test
Sample Size	Large (typically n > 30)	Small (typically n < 30)
Population Variance	Known	Unknown
Distribution	Normal distribution (standard normal)	t-distribution (Student's t-distribution)
Use Case	Compare sample mean to population mean or two sample means	Compare sample mean to population mean or two sample means

- Qs : A teacher claims that the mean score of students in his class is greater than 82 with a standard deviation of 20.
- If a sample of 81 students was selected with a mean score of 90 then check if there is enough evidence to support this claim at a 0.05 significance level.

Eigen things
- Values
- Vectors

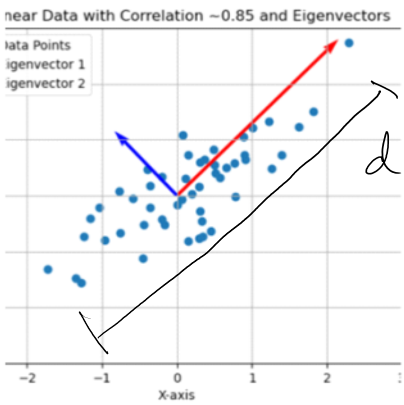
What is eigen decomposition?

also known as spectral decomposition, is a process in linear algebra where a matrix is decomposed into its eigenvalues and eigenvectors.

useful in solving differential equations, quantum mechanics, vibration analysis, and principal component analysis in statistics.

intuitive examples of eigen decomposition

Principal Component Analysis (PCA):	eigen decomposition is used to find the principal components of a dataset.
	Each principal component (eigenvector) represents a direction in the feature space along which the data varies the most.
	The corresponding eigenvalue represents the amount of variance explained by that principal component.
	By sorting the eigenvectors based on their eigenvalues, PCA helps in reducing the dimensionality of the data while preserving most of its variance.



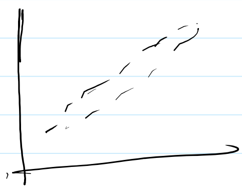
d_1 = eigen value
"magnitude"

D = 2 cols ; how many new axes = 2 ; x_1', x_2'
3 cols , = 3 , x_1', y_2', x_3'
3000 cols , = 3000 x_1', x_2' .. x_{3000}'

D = 2 cols ⇒ 2 new axes → 2 new vectors. (eigen vector)
3 cols ⇒ 3 new eigen vectors
3000 cols ⇒ 3000 eig vector.

each eig vector ⇒ eig value
⇒ spread of data along that eig vector.

if cols (3) have linear tendencies;



eig value (e_1) = 100 ✓
eig value (e_2) = 50 ✓
eig value (e_3) = 10 ✓

e_1 = 10/total ⇒ 37%
e_2 = 100 ⇒ 24%
e_3 = 50 ⇒ 19%
e_4 = ⇒ 15%
⇒ .0001
⇒ .002
e_{3000}

95%

??

D_{3000}
⇓
D_4

1024 x 1024

D.

100,000

"essence of data."
e_1, e_2, e_3, e_4
"understanding data"
"latent data"

intuitive examples of eigen decomposition

- **Image Compression**
  - to compress an image while maintaining its key features.
  - An image can be represented as a matrix, where each entry corresponds to a pixel's intensity.
  - Applying eigen decomposition to the image's covariance matrix gives you

intuitive examples of eigen decomposition

- **Image Compression**
  - to compress an image while maintaining its key features.
  - An image can be represented as a matrix, where each entry corresponds to a pixel's intensity.
  - Applying eigen decomposition to the image's covariance matrix gives you eigenvectors (principal components).
  - By keeping only the top eigenvectors, you can reconstruct the image with fewer components, thus compressing the data while preserving important features.

5

02-06-2024

intuitive examples of eigen decomposition

Google's PageRank Algorithm

- **Scenario**: You want to rank web pages based on their importance.

Vibrations of a Mechanical System

- **Scenario**: You want to understand how a structure (e.g., a bridge) will vibrate under certain conditions.

Facial Recognition

- **Scenario**: You want to identify a person based on their facial features.

Sample	G1	G2	G3	G4	G5	G6	...	G100000
S1	12.3	5.6	8.9	21.1	16.7	9.2		...
S2	15.2	3.8	9.7	18.3	10.5	7.1		...
S3	10.9	6.5	7.2	22.0	14.8	11.0		...

Examples - high-dimensional data (gene)

- Imagine a dataset where each row represents a biological sample (e.g., tissue sample or individual patient), and each column corresponds to the expression level of a specific gene.
- The gene expression values are obtained through techniques like microarrays or RNA sequencing, providing a numeric measure of how active each gene is in a particular sample.

Image	Pixel1	Pixel2	Pixel3	...	Pixel9999	Pixel10000
I1	255	200	150		100	50
I2	100	120	80		200	180
I3	40	60	90		120	180

Examples - high-dimensional data (images)

- A real-world image dataset could have much larger dimensions, especially if dealing with high-resolution images.
- For example, a color image with a resolution of 512x512 pixels would result in 786,432 (3 channels for RGB) or 1,572,864 (4 channels for RGBA) features.

Video	Frame1 (Pixel1, Pixel2, ..., Pixel10000)	Frame2 (Pixel1, Pixel2, ..., Pixel10000)	...	Frame4 (Pixel1, Pixel2, ..., Pixel10000)
V1	(255, 200, ..., 50)	(100, 120, ..., 180)	...	(80, 90, ..., 150)
V2	(150, 180, ..., 50)	(200, 220, ..., 120)	...	(60, 80, ..., 200)
V3	(40, 60, ..., 180)	(120, 140, ..., 80)	...	(90, 100, ..., 120)

Examples - high-dimensional data (video)

Video data introduces an additional dimension of complexity compared to images, as it involves a sequence of frames over time.

Here's a simplified example using a hypothetical scenario with three video samples (V1, V2, V3), each consisting of four frames and 10,000 pixels in each frame:

Examples - high-dimensional data (text)

Review Comment 1: "The movie was fantastic, with great acting and a compelling storyline."
Review Comment 2: "I found the plot a bit predictable, but the performances were excellent."

Review	movie	fantastic	great	acting	compelling	storyline	found	plot	bit	predictable	performances	excellent
Comment 1	1	1	1	1	1	1	0	0	0	0	0	0
Comment 2	1	0	0	0	0	0	1	1	1	1	1	1

In real-world scenarios, the vocabulary size can be much larger, resulting in a correspondingly higher-dimensional space.

Vocabulary:

movie
fantastic
great
acting
compelling
storyline
found
plot
bit
predictable
performances

Example Matrix

- **Find the Eigenvalues**
  - Eigen values are scalars λ such that there exists a non-zero vector v (the eigenvector) where $Av=\lambda v$
  - Characteristic Equation: To find the eigenvalues, we solve the characteristic equation det($A-\lambda I$)=0, where I is the identity matrix.

Let's consider a simple 2x2 matrix:

$$A = \begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix}$$

$$A - \lambda I = \begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 4-\lambda & 1 \\ 2 & 3-\lambda \end{bmatrix}$$

Determinant Calculation

- Calculate the determinant of the matrix A−λI.

$$\det(A - \lambda I) = (4-\lambda)(3-\lambda) - (1 \cdot 2) = \lambda^2 - 7\lambda + 10$$

$$\lambda^2 - 7\lambda + 10 = 0$$

$$(\lambda - 5)(\lambda - 2) = 0$$

So, the eigenvalues are $\lambda_1 = 5$ and $\lambda_2 = 2$.

For each eigenvalue, find the corresponding eigenvector v such that $Av=\lambda v$.

Find the Eigenvectors

1. Eigenvector for $\lambda_1 = 5$:

$$A\mathbf{v} = 5\mathbf{v}$$

$$\begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 5 \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$$

$$\begin{bmatrix} 4v_1 + v_2 \\ 2v_1 + 3v_2 \end{bmatrix} = \begin{bmatrix} 5v_1 \\ 5v_2 \end{bmatrix}$$

This gives us the system of equations:

$$4v_1 + v_2 = 5v_1 \implies v_2 = v_1$$

$$2v_1 + 3v_2 = 5v_2 \implies 2v_1 = 2v_2 \implies v_2 = v_1$$

Thus, the eigenvector corresponding to $\lambda_1 = 5$ is any scalar multiple of $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$

v