Recap

06 April 2023 19:43

Session 2 → $\boxed{HT}$

- → Why
- → What
- → Null and alternate
- → Z-test
- → 1 tailed vs 2 tailed
- → Type 1 vs type 2 error

significance level (α)

Session 2

- P-values
- t-test

P-value

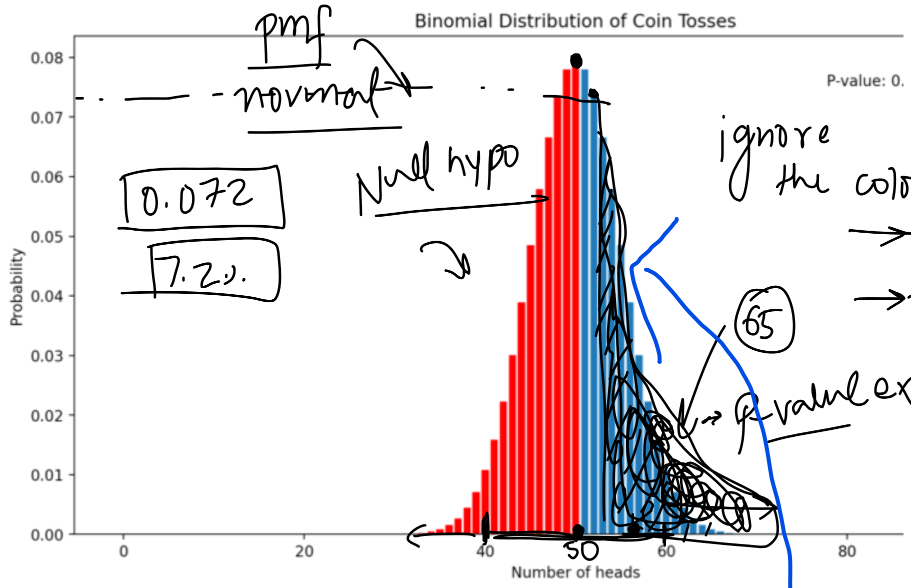
06 April 2023 06:48

of 53 → 53 head 0.07 $P(H>53)$ p_value p, n

experiment → 1 coin → 100 times toss

binomial distribution #heads

P-value is the probability of getting a sample as or more extreme(having more evidence against H0) than our own sample given the Null Hypothesis(H0) is true.



ignore the colors

→ H_0: $P(H) = P(T)$

→ H_a: $P(H) > P(T)$

p-value exp → 100 times → 53 heads

exp → 100 times

P = 0.3

30 times

53 H

In simple words p-value is a measure of the strength of the evidence against the Null Hypothesis that is provided by our sample data.

Null hyp

100 exp → 80 times → 0

2 times

This Area is Pvalue

Experiment - Toss a coin 100 times and count the heads.
do a hypothesis test which says coin is bias towards heads

H0: Coin is fair: P(H) = P(T)
H1: P(H) > P(T)
Ab jab humne 100 bar coin uchala isme 53 baar heads aaye , based on single exp we cannot conclude coin is rigged read the definition of P Value - and note that jo hamara sample hai usme 53 heads aye with probability of 0.07 , ab baki times jo hum 100 bar coin uchalenge unme 53 se zada head aane ka combined probability is P- Value (P(H)>53 = P-Value) Hence this area is P-value.

Understanding through visualization tool. Ex1 - use tool and make number of heads = 53 , p value will be = 0.30 . which basically means Probability of getting 53 or more extreme values ka sum

Interpretation - Experiment ye hai Coin ko 100 bar uchalna , aur ye experiment kitni baar kr rhe ho 100 baar , agar apka P-value 0.3 hai iska mtlb jo 100 baar experiment kiye sme se 30 times , Head ka value 53 ya use extreme (zada) hai

Ex 2 - 60 heads , P value = 0.0284 , interpretion - exp 100 baar krne pe 2 times aisa hoga jisme head ki value 60 ya use zada hogi , Now is the time to refer to chat in the folder ,(agar aapke sample me 60 ya zada aara hai jiska probability hone ka only 2 times , iska mtlb apka coin bias hai kyumki itna km probability hone ke baad bhi aoka 60 values aari hai , iska mtlb you proved null hypothesis is false

Final Interpretation - Aap Experiment kr rhe ho coin ko 100 baar uchalne ka and count kr rhe ho head , Aur aap Ye experiment kr rhe ho 100 baar, agar aapka p-value 0.3 hai iska mtlb ye hua

Interpreting p-value

06 April 2023 08:25

reject your Ho

With significance value

$\alpha = 0.05 / 0.01 \rightarrow$ p-value $\leq \alpha$ if p value less than alpha then reject

0.05

Without significance value $\alpha \rightarrow$

1. Very small p-values (e.g., p < 0.01) indicate strong evidence against the null hypothesis, suggesting that the observed effect or difference is unlikely to have occurred by chance alone.
2. Small p-values (e.g., 0.01 ≤ p < 0.05) indicate moderate evidence against the null hypothesis, suggesting that the observed effect or difference is less likely to have occurred by chance alone.
3. Large p-values (e.g., 0.05 ≤ p < 0.1) indicate weak evidence against the null hypothesis, suggesting that the observed effect or difference might have occurred by chance alone, but there is still some level of uncertainty.
4. Very large p-values (e.g., p ≥ 0.1) indicate weak or no evidence against the null hypothesis, suggesting that the observed effect or difference is likely to have occurred by chance alone.

α

Pvalue approach
Since , we have population std we will conduct Z test ,and do hypothesis using P values
calculate Z - stat - ie 4.10. IN p - value we do not find critical point(alpha approach or rejection region approach)
In this you will find the area to right side of Z stat value which is 4.10 ie use Z table which is 1 - 0.9999 = 0.0001
which is the p- value. Considering we have alpha value = 0.05 we will compare if p value less than or equal alpha whcih is right , hence we have strong evidence against null hypothesis , hence we will reject null hypothesis.

two tailed test lays example first calculate Z stat which is -1.26 since we do not know the direction we will calculate for bth sides ie area for -1.26 and 1.26 then add them hence we get Pvalue = area1 + area2 . then compare with Significance level (find 1 side area and multiple by 2 since it is symmetrical)
P value is 0.26 which is greater than 0.05 hence we cannot reject the null hypothesis

P-value in context of Z-test

06 April 2023 07:08

1) rejection region approach 2) p-value approach

Suppose a company is evaluating the impact of a new training program on the productivity of its employees. The company has data on the average productivity of its employees before implementing the training program. The average productivity was 50 units per day. After implementing the training program, the company measures the productivity of a random sample of 30 employees. The sample has an average productivity of 53 units per day and the pop std is 4. The company wants to know if the new training program has significantly increased productivity.

$\mu = 50$ $n = 30$ $\bar{x} = 53$

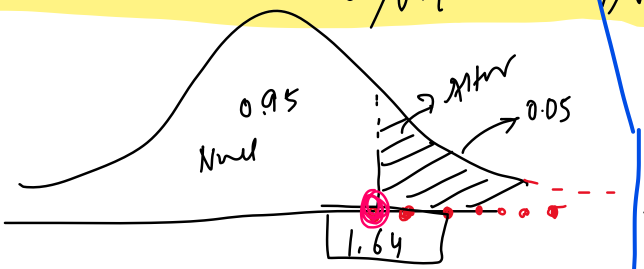
$\sigma = 4$ $\alpha = 0.05$

$H_0: \mu = 50$

$H_a: \mu > 50$

$$\text{Z-stat} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{53 - 50}{4/\sqrt{30}} = \frac{3 \times \sqrt{30}}{4} = \boxed{4.10}$$ $\boxed{1.7}$

reject



$1 - \boxed{0.95}$ → $\boxed{0.05}$

p-value

$\boxed{z = 4.10}$

p-value → critical point

$\boxed{0.999} = \boxed{0.0001}$

p-value < 0.05

reject Null hypo

Suppose a snack food company claims that their Lays wafer packets contain an average weight of 50 grams per packet. To verify this claim, a consumer watchdog organization decides to test a random sample of Lays wafer packets. The organization wants to determine whether the actual average weight differs significantly from the claimed 50 grams. The organization collects a random sample of 40 Lays wafer packets and measures their weights. They find that the sample has an average weight of 49 grams, with a pop standard deviation of 5 grams.

$\mu = 50$ $n = 40$ $\bar{X} = 49$ → $H_0: \mu = 50$

$\sigma = 5$ $\alpha = 0.05$ $H_a: \mu \neq 50$

p-value 2 tailed $\boxed{0.206} > 0.05$

$$z = \frac{49 - 50}{5/\sqrt{40}} = \frac{-\sqrt{40}}{5} = \boxed{-1.26}$$

$\boxed{\text{p-val}}$ $\boxed{0.206}$

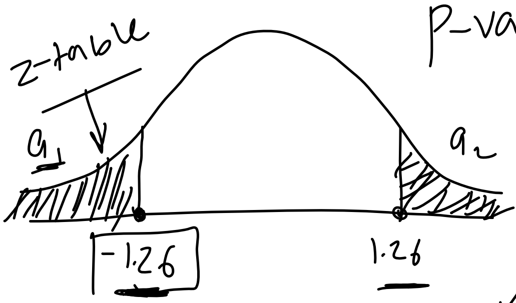
$\text{p-value} = a_1 + a_2$

$5/\sqrt{40}$ $\quad$ 5

$\boxed{\text{p-val}}$ $\boxed{0.206}$

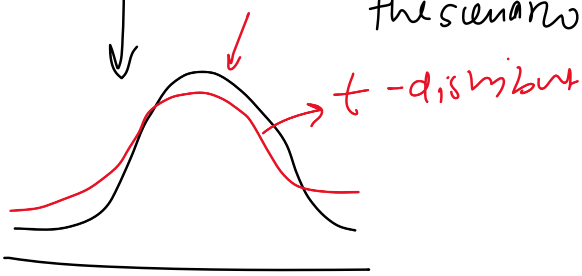
p-value $= a_1 + a_2$

$a_1 = 0.103 \times 2$



Z test → σ ✓

t-test ← smaller sample sizes → s ✓ ; σ ✗ most of the scenario



T-tests

06 April 2023 14:14

A t-test is a statistical test used in hypothesis testing to compare the means of two samples or to compare a sample mean to a known population mean. The t-test is based on the t-distribution, which is used when the population standard deviation is unknown and the sample size is small.

There are three main types of t-tests:

1 sample $\rightarrow \bar{x} \rightarrow \mu$

One-sample t-test: The one-sample t-test is used to compare the mean of a single sample to a known population mean. The null hypothesis states that there is no significant difference between the sample mean and the population mean, while the alternative hypothesis states that there is a significant difference.

$\rightarrow \sigma$ X

Independent two-sample t-test: The independent two-sample t-test is used to compare the means of two independent samples. The null hypothesis states that there is no significant difference between the means of the two samples, while the alternative hypothesis states that there is a significant difference.

test A $\rightarrow$ popl

1 class $\rightarrow$ test B $\rightarrow$ pop

Paired t-test (dependent two-sample t-test): The paired t-test is used to compare the means of two samples that are dependent or paired, such as pre-test and post-test scores for the same group of subjects or measurements taken on the same subjects under two different conditions. The null hypothesis states that there is no significant difference between the means of the paired differences, while the alternative hypothesis states that there is a significant difference.

Single Sample t-test

06 April 2023 14:14

t-test

σ × s ✓

$\frac{\text{lays}}{40 \text{ lays}}$ → 50gm

A one-sample t-test checks whether a sample mean differs from the population mean.

Assumptions for a single sample t-test

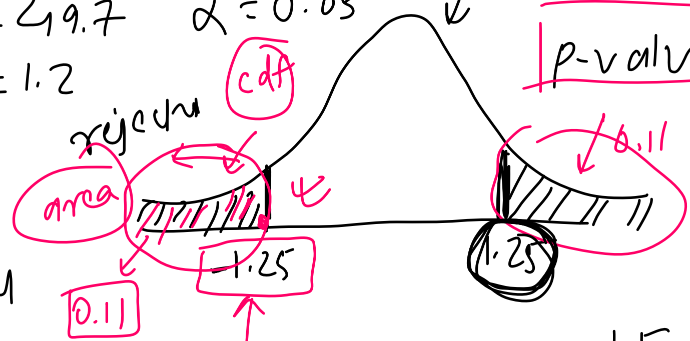
→ sample normally distri

1. Normality - Population from which the sample is drawn is normally distributed
2. Independence - The observations in the sample must be independent, which means that the value of one observation should not influence the value of another observation.
3. Random Sampling - The sample must be a random and representative subset of the population.
4. Unknown population std - The population std is not known.

Suppose a manufacturer claims that the average weight of their new chocolate bars is 50 grams, we highly doubt that and want to check this so we drew out a sample of 25 chocolate bars and measured their weight, the sample mean came out to be 49.7 grams and the sample std deviation was 1.2 grams. Consider the significance level to be 0.05

$\mu = 50 \quad n = 25 \quad \bar{x} = 49.7 \quad \alpha = 0.05$

$s = 1.2$



p-value = 0.22

> 0.05

Null reject ✗

$H_0: \mu = 50$

$H_a: \mu \neq 50$

assuming it is normal

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{49.7 - 50}{1.2/\sqrt{25}} = \frac{-0.3 \times 5}{1.2} = \frac{-1.5}{1.2} = -1.25$$

$df = n - 1 = 24$

$H_0: \boxed{\mu = 35}$ → 25 sample age → normal

$H_a: \mu < 35$

$\boxed{\mu = 35}$ σ ✗ → t-test

$\bar{x}, s, \alpha = 0.05$

Shapiro-Wilk test

p-value < 0.05 not normal

p-value > 0.05 normal

IN T table same approach p- value = left area + right area

Since we do not have value for -1.25 , we will use CDF to find area till -1.25 , refer to the code in the folder , P-value will be 0.22, since it is greater than alpha(0.05) we cant reject the null hypothesis

Python Case Study 1

06 April 2023 17:27

Single T-test

In Python notebook , we will have titanic dataset , now we have to do hypothesis testing that population mean of age is less than 35 , not 35.
So h0 : null hypothesis is mean = 35
h1:alternative hypothesis: mean<35

Now we took 25 data for our sample. Now to do t-test there are certain assumptions in which one is normal distribution, now to check normality there is one test which is called Shapirio wilk test , which give p-value , if p value is less than 0.05 than we can say distribution is not normal if it is greater than we can say it is normal.
Since in our case

Ex of Independent 2 sample t-test - Titanic data me avg age of male > avg age of female

Do not ignore the assumptions. 1. Independence , 2. normailty
3 is interesting which is equal variance , the variance of two population should be approaximately equal.

In the example of desktop and mobile , null hypothesis will be H0 : avg time of desktop = avg time of mobile
h1: desktop not equal mobile

Independent 2 sample t-test

06 April 2023 14:15

An independent two-sample t-test, also known as an unpaired t-test, is a statistical method used to compare the means of two independent groups to determine if there is a significant difference between them.

titanic

males avg age (28)

females avg age (25)

Assumptions for the test:

1. Independence of observations: The two samples must be independent, meaning there is no relationship between the observations in one group and the observations in the other group. The subjects in the two groups should be selected randomly and independently.
2. Normality: The data in each of the two groups should be approximately normally distributed. The t-test is considered robust to mild violations of normality, especially when the sample sizes are large (typically n ≥ 30) and the sample sizes of the two groups are similar. If the data is highly skewed or has substantial outliers, consider using a non-parametric test, such as the Mann-Whitney U test.
3. Equal variances (Homoscedasticity): The variances of the two populations should be approximately equal. This assumption can be checked using F-test for equality of variances. If this assumption is not met, you can use Welch's t-test, which does not require equal variances.
4. Random sampling: The data should be collected using a random sampling method from the respective populations. This ensures that the sample is representative of the population and reduces the risk of selection bias.

QQ, hist → shapiro-wilk

A B

$\sigma_A^2 = \sigma_B^2$

levene

levene

p-value < 0.05

$\sigma_A^2 \neq \sigma_B^2$

p value > 0.05

$\sigma_A^2 = \sigma_B^2$

reject my H_0

Suppose a website owner claims that there is no difference in the average time spent on their website between desktop and mobile users. To test this claim, we collect data from 30 desktop users and 30 mobile users regarding the time spent on the website in minutes. The sample statistics are as follows:

→ avg time

desktop users = [12, 15, 18, 16, 20, 17, 14, 22, 19, 21, 23, 18, 25, 17, 16, 24, 20, 19, 22, 18, 15, 14, 23, 16, 12, 21, 19, 17, 20, 14]

mobile_users = [10, 12, 14, 13, 16, 15, 11, 17, 14, 16, 18, 14, 20, 15, 14, 19, 16, 15, 17, 14, 12, 11, 18, 15, 10, 16, 15, 13, 16, 11]

avg time

$H_0 = \mu_d = \mu_m$

$H_a = \mu_d \neq \mu_m$

Desktop users:

- Sample size (n1): 30
- Sample mean (mean1): 18.5 minutes
- Sample standard deviation (std_dev1): 3.5 minutes

Mobile users:

- Sample size (n2): 30 → 25
- Sample mean (mean2): 14.3 minutes
- Sample standard deviation (std_dev2): 2.7 minutes

check assumptions

We will use a significance level (α) of 0.05 for the hypothesis test.

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

t-statistic

this is the formula for two sample t-test(independent)

where x1 = 18.5 , x2 = 14.3 , sample_std1 = 3.5 , sample_std2 = 2.7 ,n1 = 30 , n2 = 30

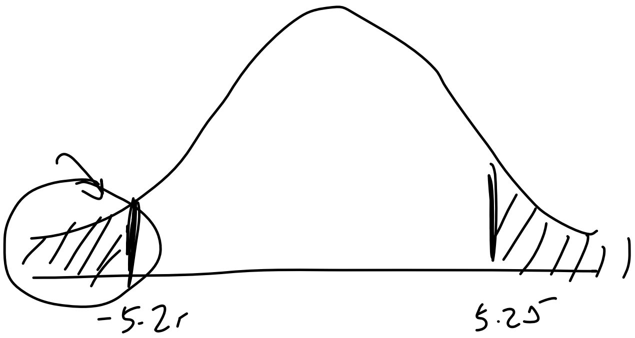
$$t = \frac{18.5 - 14.3}{\quad} = \frac{4.2}{\quad} = \boxed{5.25}$$

$$t = \frac{\ldots - \ldots}{\sqrt{\frac{(3.5)^2}{30} + \frac{(2.7)^2}{30}}} = \frac{4.2}{\sqrt{\frac{19.54}{30}}} = \boxed{5.25}$$

the t-statistic value is 5.25 , calculate p value.

Note that in case of 2 smple degree of freedom will be n1 + n2 - 2

refer to the code for 2 sample



$\boxed{n-1}$ t

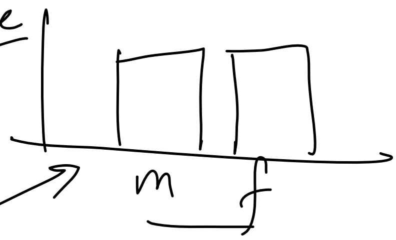
$n_1 + n_2 - 2$

$30 + 30 - 2 = 58$

$\boxed{\mu_m > \mu_f}$

~~age~~ $\boxed{\mu_m = \mu_f}$

age



age and gender

pop

$H_0: \mu_m = \mu_f$

$H_1: \mu_m > \mu_f$

$\alpha = 0.05$

Python Case Study 2

06 April 2023 17:27

example for paired test - Consider a class of datascience have 5 students we took a test and the marks for the students were following - , we took extra classes and tehn again took test , so the assumption says diff between these 2 samples should be normal , and the pairs of samples must bee independent ie marks of A do not relate to B

Paired 2 sample t-test

06 April 2023 14:21

paired 2 sample test is used when there is some relation between 3 groups ie they are not independent , Generally use this test in scenario of before after test

A paired two-sample t-test, also known as a dependent or paired-samples t-test, is a statistical test used to compare the means of two related or dependent groups.

Common scenarios where a paired two-sample t-test is used include:

1. **Before-and-after studies**: Comparing the performance of a group before and after an intervention or treatment.
2. **Matched or correlated groups**: Comparing the performance of two groups that are matched or correlated in some way, such as siblings or pairs of individuals with similar characteristics.

	I	II	d
A	50	55	-5
B	60	60	0
C	76	66	10
D	40	60	-20
E	25	100	-75

normal

Assumptions

1. Paired observations: The two sets of observations must be related or paired in some way, such as before-and-after measurements on the same subjects or observations from matched or correlated groups.
2. Normality: The differences between the paired observations should be approximately normally distributed. This assumption can be checked using graphical methods (e.g., histograms, Q-Q plots) or statistical tests for normality (e.g., Shapiro-Wilk test). Note that the t-test is generally robust to moderate violations of this assumption when the sample size is large.
3. Independence of pairs: Each pair of observations should be independent of other pairs. In other words, the outcome of one pair should not affect the outcome of another pair. This assumption is generally satisfied by appropriate study design and random sampling.

Let's assume that a fitness center is evaluating the effectiveness of a new 8-week weight loss program. They enroll 15 participants in the program and measure their weights before and after the program. The goal is to test whether the new weight loss program leads to a significant reduction in the participants' weight.

$\mu_{dif} = \mu_{bef} - \mu_{af} = 0$

$H_0: \mu_{before} = \mu_{after}$

$H_1: \mu_{before} > \mu_{after}$

Before the program:
[80, 92, 75, 68, 85, 78, 73, 90, 70, 88, 76, 84, 82, 77, 91]
After the program:
[78, 93, 81, 67, 88, 76, 74, 91, 69, 88, 77, 81, 80, 79, 88]

Significance level (α) = 0.05

> 0.05

name	wt before	wt after	diff
A	80	78	x_1
B	92	93	x_2
C	75	.	x_3
.	.	.	.
K	91	88	x_{15}

diff

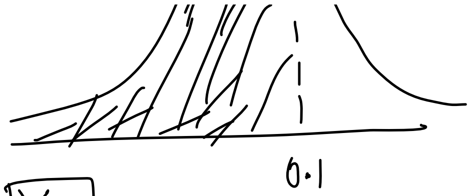
normal dist

$\bar{X}_{diff}$ S_{diff}

Check for assumptions. if assumptions are true , then find the mean of Diff and sample std.

$t = \bar{X}_{diff}$ 15

$t = \frac{\bar{X}_{diff}}{S_{diff}/\sqrt{n}}$ 15



$0.54 > 0.05 \rightarrow \boxed{X}$

Note that formula for paired t test is X bar diff(sample mean of diff) - mew diff (before - after) , since in null hypothesis we assume , before and after training result is same , hence the mew diff will be zero divided by sample std/root(n)

refer to the code