Z Test - 2

Agenda

1) Recap
2) Power of Test
3) Z Test - 2 Sample
4) Z Test Proportions

⇒ **Recap** 1) Hypothesis Testing

a) H_0 & H_a
b) Choose a distribution
c) Define which tailed test to use
d) Calculate P value
e) Compare P value to α.

CLT → SM distⁿ → is always normal

$$\Rightarrow N\left(\mu_{pop}, \frac{\sigma_{pop}}{\sqrt{n}}\right)$$

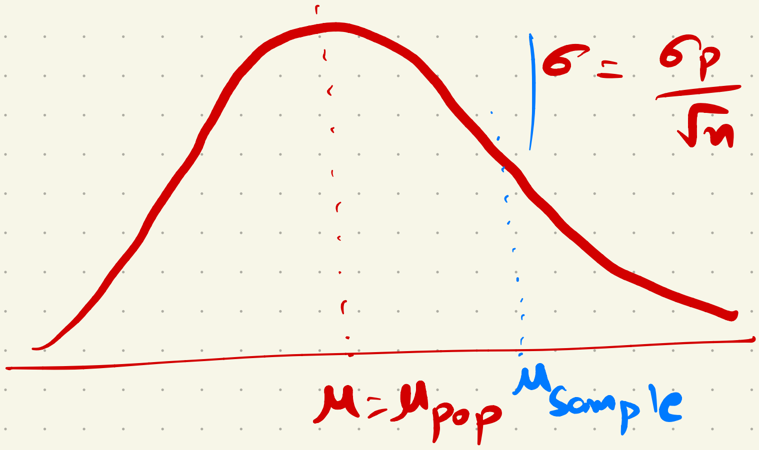
⇒ So datapts → → find SM distⁿ

$n=50$

μ_{pop}

$\sigma_{pop} \approx \sigma_{sample}$ } → this distribution to be used.

Mean/avg of 50 datapts → $\underline{\mu_{sample}}$



⇒ μ_{sample} → find the z value → 1−cdf
↳ p-value

⇒ Critical value →

That specific μ_{sample} value such that corresponding p-value $\equiv \alpha$.

$$1 - cdf(x) \equiv \alpha$$

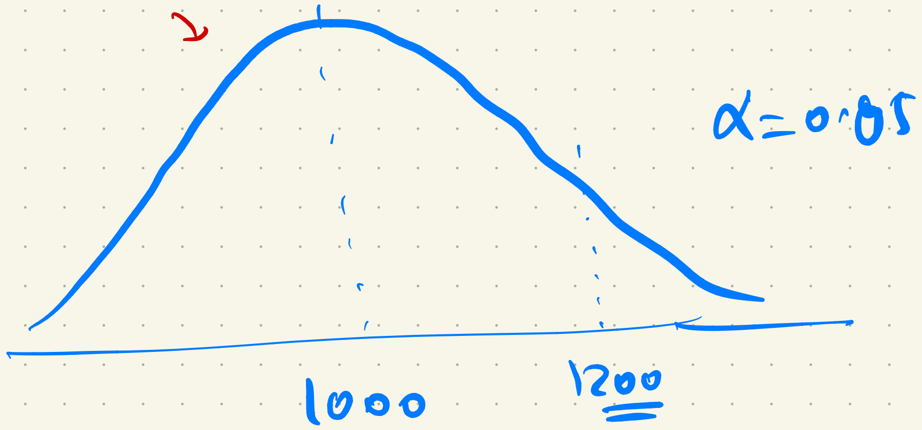
$$1 - \alpha = cdf(x)$$

$$x = ppf(1-\alpha)$$

Is this x the required value → xx
↳ z value

$$x = \frac{(x_0^{\prime}) - \mu}{\sigma/\sqrt{n}}$$

Critical value $\to x_0 = \boxed{\mu + \text{ppf}(1-\alpha)\frac{\sigma}{\sqrt{n}}}$

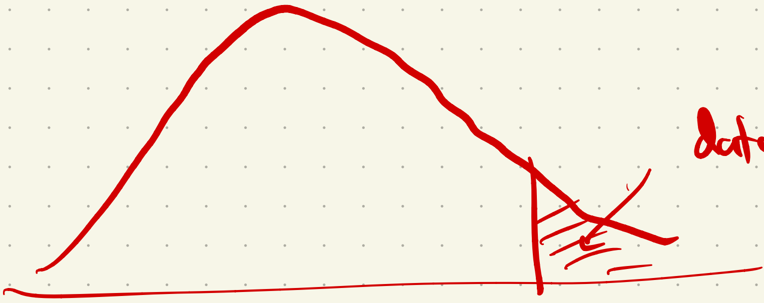


1200 → 1230
→ p-value < 0.05

⇒ Any value on the extreme side of critical value
↓
p-value < α
or
You should the reject H_0.

⇒ P-value → P(extreme data | H_0 is true)

we use distribution → where is H_0 true



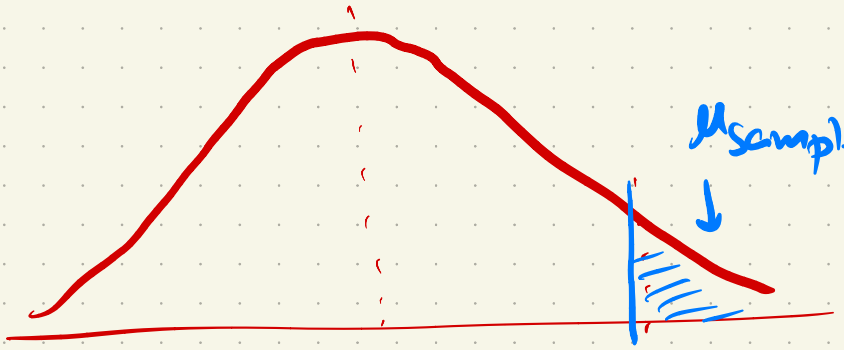
data lies on extreme region of the H_0 dist'

↓

reject it.

⇒ Distribution of H_0 → see if data lies in extreme.

⇒



μ_{sample} lies here.

⇒ CI → range of population mean based on data.

Sample → SM = 65.5 inch

$$\underbrace{65.5}_{[60 \; - \; 70]}$$

→ 95% sure that actual population mean will lie in this range

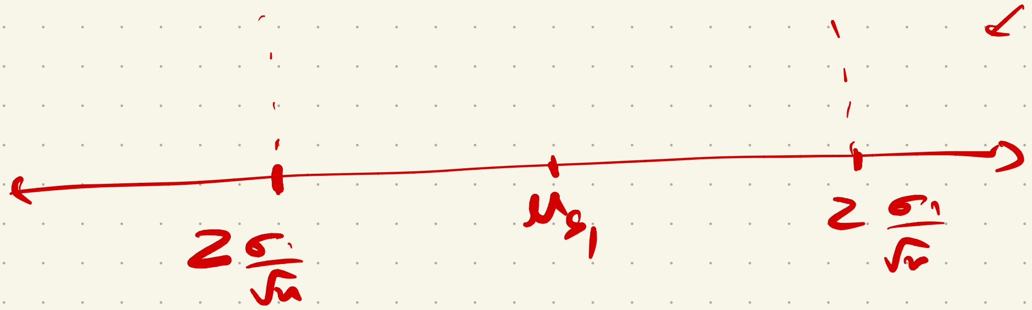
---

⇒ CI ⇒ $\mu_{sample} \pm Z \frac{\sigma}{\sqrt{n}}$

95% confiden → norm.ppf(   )

⇒ Case 1 → I collect 50 data pts → μ_s, σ_{sample}

Case 2 → — " 50 dat pts → μ_2, σ_2



If my actual population mean is outside this range $\longrightarrow$ 5% chance

$\Rightarrow$ If for any sample I calculate a CI and my actual population mean is outside this CI $\longrightarrow$ we can reject my H_0.

$\Rightarrow$ Cohen's d $\longrightarrow$ $|\mu_{sample} - \mu_{pop}|$

$\Rightarrow$ original method $\longrightarrow$ Based on μ_{pop} define a dist^n of SMs and check if your actual μ_{sample} lies in the extreme region.

CI method $\longrightarrow$ Based on the μ_{sample} $\longrightarrow$ Calculate a CI $\longrightarrow$ If your actual μ_{pop} is outside CI $\longrightarrow$ reject H_0.

⇒

		Inno	Criminal ← Reality
Judge says	Inn	True -ve	False -ve (Type 2)
	Crim	False +ve (Type I)	True +ve

$\alpha = 0.05$

p-value
Chances that we see extreme data given H_0 is true

⇒ optimizing for Type I error → α

H_0 to be true → extreme region is the criminals.

$\underbrace{95\%}_{\text{innocent}} - \underbrace{5\%}_{\text{criminal}}$ ← α

⇒ Hypothesis testing → α → max allowed False +ve.

⇒ What about Type 2 Error →

① Type II Error → β

- α & β are both very low → ideal test
  - α ↓ False +ve
  - β ↓ False -ve

- Till now we are just optimizing on α →
  ↳ reducing False +ve below α.

- Power of test = $1-\beta$

⇒ Both α should be low
$1-\beta$ should be high }

⇒ Factors which impact power →

1) Sample size (n) →
power will boosted

$\alpha + \beta \neq 1$

2) Significance level (α) →

$\alpha \rightarrow$ 0'05 $\rightarrow$ 0·01 $\rightarrow$ (decrease)

power $\rightarrow$ 0·9 $\rightarrow$?? decrease

- Type I Error $\downarrow$ Type II Error $\uparrow$

3) Variability (σ) $\rightarrow$

$$\sigma \propto \frac{1}{P}$$

4) Effect Size $\rightarrow$

$$d = \frac{\text{Sample mean} - \text{Pop mean}}{\text{Sample std dev.}}$$

Cohen's d

value

$\Rightarrow$ Q Power of test is 82% $\rightarrow$ error

what would be type II value

= 1 - power

= 18%

⇒ Higher power means lower probability of Type II Error.

Q. Is α & β complimentary →

α & β are inversely propotional

If $\alpha \uparrow$ $\beta \downarrow$

$\alpha \downarrow$ $\beta \uparrow$

$\alpha = 0.05$ $\not\rightarrow$ $\beta = 0.05$ ✗

$\alpha = 0.01$ $\beta = 0.09$ ✗

$\beta = 0.08$

0.07

⇒ μ_{pop} against 1 Sample

σ_{pop}

H_0 → $\mu_{pop} = x$

⇒ Historical Sales avg → μ_{pop} → calculate if there is a change or not

⇒ What if μ_{pop} is not there ??

⇒ A → M_1 B → M_2

eg

⇒ Compare → time to recover

⇒ M_1 = [13, 11, 12, ... 10, 9, 7, . . . 21]
(50 people)

M_2 = [12, 13, 15, ... 17, 18, . . . 30]
(40 people)

⇒ μ_{M_1} = 14.5 days | 14.5 days

μ_{M_2} = 15.2 days | 27.8 days

⇒
⇒ Just checking if μ_1 & μ_2 are significantly different

① $H_0 \rightarrow \mu_1 = \mu_2$

$H_a \rightarrow \mu_1 \neq \mu_2$

② If we assume that data follows normal dist^n.

③ Tailed test $\rightarrow$ Two tailed

④ Calculate P value $\rightarrow$

$$Z = \frac{x - \mu}{\sigma}$$

propotional distance $\Rightarrow$

$$\Rightarrow \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1, \sigma_2}}$$

$$\boxed{Z^* = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}}$$

Test statistic for 2 Sample Z Test

$\Rightarrow N(0,1) \rightarrow \left.\begin{matrix} \mu = 0 \\ \sigma = 1 \end{matrix}\right\} \rightarrow$

⇒ 

cdf(z) ⇒ % value ≤ the given x.

$$z = \frac{x - \mu}{\sigma}$$

⇒ Calculate corresponding p value using z.

If p-value < α → reject H_0.

Q

$\mu_1 = 2.5$ $\sigma_1 = 0.5$ $n_1 = 50$

$\mu_2 = 3$ $\sigma_2 = 0.4$ $n_2 = 30$

$$z = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Z Test → assume normal

p value = 1

One Sample Z Test → population against 1 Sample

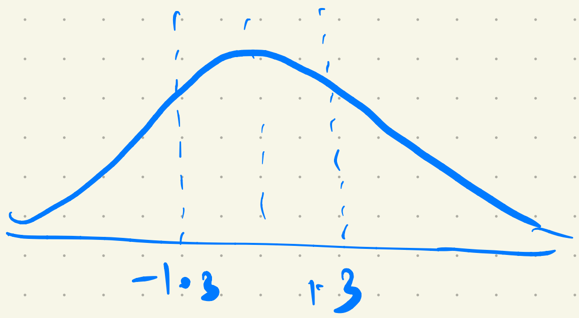
Two Sample Z Test → Samp1 & Sample 2

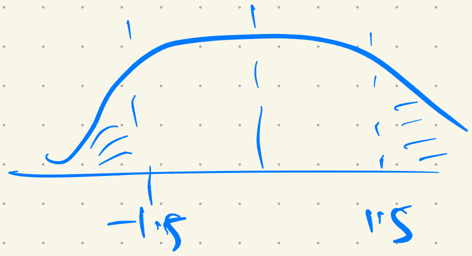
One Sample	Two Sample
$Z = \frac{\mu_{samp} - \mu_{pop}}{\sigma_{pop}/\sqrt{n}}$	$Z = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$
↓	↓
Calculate P_valu	Calculate P_value using Normal Dist

Right tailed test ⇒ 1 − norm.cdf(z)

Left tailed test ⇒ norm.cdf(z)

Two Tailed test ⇒ 2 × (1 − norm.cdf(|z|))





Compare P_value against α.

⇒ Conditions for 2 Sample Z Test

① Population must have finite μ & σ.

② n_1 & n_2 should not be very small

$n_1, n_2 \geq 30$

③ Data should be normally distributed **

④ Data should be Continuous.

Determine whether there is a statistically significant difference in the average heights of plants grown with fertilizer X and fertilizer Y.
Group A (Fertilizer X): Heights = [14, 16, 13, 17, 12, 18]
Group B (Fertilizer Y): Heights = [18, 19, 16, 17, 15, 20]
Significance Level (α): 0.1

$H_0 \Rightarrow \mu_A = \mu_B$

$H_a \Rightarrow \mu_A \neq \mu_B$

$z_1 = -2.03$

p_val = 0.04 $< \alpha$

⇒ 2-proportion Test →

Q. If my weekly sales have increased.

Q. Have my conversion increase.

$$\text{Conv Rate} = \frac{\text{Total people ordering}}{\text{Total people visiting}}$$ [% of Total]

↓

proportion of Total

Q. Company launches a new product

A) Historically 70% people liked the App

B) After launch → In a sample of 120 people

73% liked it. {80/120}

⇒ $H_0 = \quad p = 0.7$

$H_a = \quad p \neq 0.7$

$$\Rightarrow Z = \frac{\text{Sample Proportion} - \text{Pop prop}}{\sqrt{\dfrac{\text{pop prop}\,(1 - \text{pop prop})}{n}}}$$

$$= \frac{0.73 - 0.7}{\sqrt{\frac{(0.7)(0.3)}{120}}}$$

$$\boxed{Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}}$$

p = original propo

$\hat{p}$ = Sample propo

A fast-food restaurant claims that 80% of their customers prefer their new burger over the old one.
In a random sample of 100 customers, 85 said they preferred the new burger.
What is the null and alternative hypothesis?

$H_0 \rightarrow p = 0.8$

$H_a \rightarrow p \neq 0.8$ $\quad$ $p > 0.8$

One sample proportion $\rightarrow$ population propo

Two sample Z-proportion Test

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

$$\hat{p}_1 = x_1/n_1$$

$$\hat{p}_2 = x_2/n_2$$

⇒ 1^{st} Sample 80/120 liked it

2^{nd} Sample 81 out of 130 liked it

$$\hat{p}_1 = 80/120 \qquad \hat{p}_2 = 81/130$$

$$\hat{p} = \frac{80 + 81}{120 + 130}$$

$$\boxed{Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}}$$

A company introduces a new feature in its mobile App that allows users to subscribe to a premium service.
They want to evaluate if the introduction of this feature has led to an increase in the number of premium users.
They collect data from two different time periods: before the feature was introduced (Group A) and after the feature was introduced (Group B).
Which test should you use to determine if the new feature has significantly increased in the number of premium users?

Conclusion

→ Calculate Z Score →

One sample → $\frac{X - \mu_{pop}}{\sigma_{pop}/\sqrt{n}}$

Two Sample → $\frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

One Sample propor = $\frac{\hat{p}_1 - p}{\sqrt{\frac{p(1-p)}{n}}}$

Two Sample Proportion $= \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$

$\hat{p} = \dfrac{x_1+x_2}{n_1+n_2}$

When smokers smoke, nicotine is transformed into cotinine, which can be tested.

The average cotinine level in a group of 50 smokers was 243.5 ng ml.

Assuming that the standard deviation is known to be 229.5 ng ml, at 95 % confidence, test the assertion that the mean cotinine level of all smokers is equal to 300.0 ng ml.

$H_0 = \mu = 300$

$\sigma = 229.5$

$\mu_{samp} = 243.5$

$n = 50$

$$Z \text{ score} = \frac{\mu_{sample} - \mu_{pop}}{\sigma/\sqrt{n}} = \frac{243.5 - 300}{229.5/\sqrt{50}} =$$

2 Tailed Test $\longrightarrow$

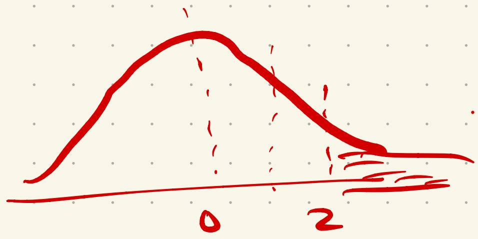
$$\text{p-value} = 2 * (1 - \text{norm.cdf}(|z|))$$

Q You are drawing from a random variable that is normally distributed with mean 0 and standard deviation 1, X ~ N(0, 1), once a day. Approximately what is the expected number of days you have to draw until you get a value greater than 2?

A

$\mu = 0$

$\sigma = 1$



$= 1 - \text{norm.cdf}\left(\frac{2-0}{1}\right)$

$= 0.04$

$G = \frac{1}{0.04}$

A french cake shop claims that the average number of pastries they can produce in a day exceeds 500.

The average number of pastries produced per day over a 70 day period was found to be 530. Assume that the population standard deviation for the pastries produced per day is 125.

Test the claim using a z-test with the critical z-value = 1.64 at the alpha (significance level) = 0.05, and state your interpretation.

$\rightarrow$ Critical z-value = 1.64 at $\alpha = 0.05$

Calculate z-score $\Rightarrow$

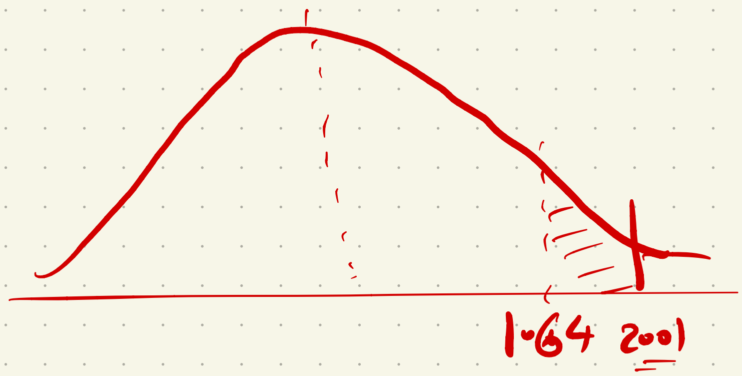
$n = 70$

$\mu_{sam} = 530$

$\mu = 500$

$\sigma = 125$

$z\ \text{score} = \dfrac{530 - 500}{\frac{125}{\sqrt{70}}} = 2.01$



x value → p-value is $= \alpha$

x → Calculate z → p-value $= \alpha$

$z = \text{norm.ppf}(1-\alpha)$

$= \underline{1.64}$

as 2.01 falls in critical region →

reject H_0

$2.01 > 1.64$ H_0 is rejected

Collab Link : https://colab.research.google.com/drive/1aW3qp3ZwVJiaoRMuXadpFbnndDqYplrB?usp=sharing