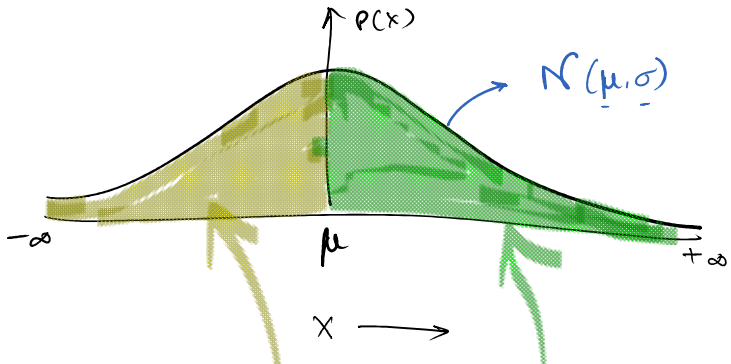
09 October 2022 18:56

Properties of Normal Distribution



① N.D. is a 2-parameter distrib^n. → mean & Std.dev. of the distrib^n.

② N.D. is Symmetric around the mean.

$$\int_{-\infty}^{\mu} P(x)\,dx = \int_{\mu}^{+\infty} P(x)\,dx = 0.5$$

Area under the curve on both sides of the mean is same = 0.5.

③ mean = median = mode value of N.D.

mean ⇒ Normal average

median ⇒ Centre point of any distrib^n (after sorting)

→ 50th Percentile ⇒ Q_2 (Second Quartile)

mode ⇒ Most frequent value of a distrib^n. (which value repeats most no. of times)

$$X = \{5, 10, 12, 15, 15, 20, 25, 28, 30, 35\}$$

$$\text{mean} = \frac{5 + 10 + 12 + 15 + 15 + 20 + 25 + 28 + 30 + 35}{10}$$

median → $\left(\frac{n+1}{2}\right)^{th}$ value if n is odd.

→ average of $\frac{n}{2}^{th}$ & $\left(\frac{n+2}{2}\right)^{th}$ value. → if n is even

where n = no. of datapoints.

$\frac{10}{2}$ = 5th value $\qquad$ $\frac{12}{2}$ = 6th value.

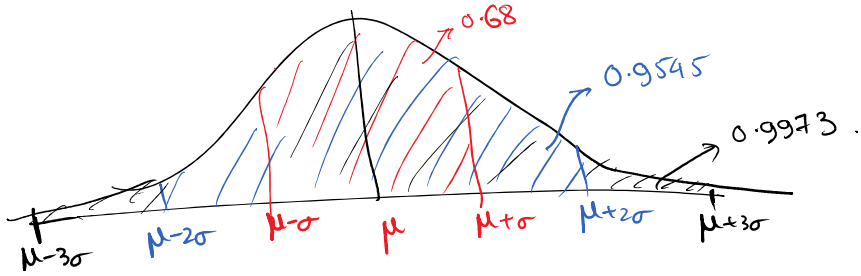
Here the average of 5th & 6th value = $\frac{15+20}{2}$ = 17.5

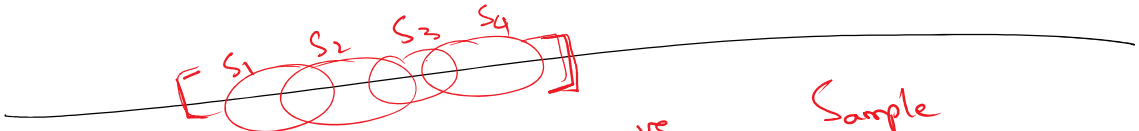
④ VV Imp :

$\mu - \sigma \leq x \leq \mu + \sigma \Rightarrow$ Area under the ND curve = 0.6828

$\mu - 2\sigma \leq x \leq \mu + 2\sigma \Rightarrow$ " = 0.9545

$\mu - 3\sigma \leq x \leq \mu + 3\sigma \Rightarrow$ " = 0.9973





Population vs Sample

(μ, σ, σ^2) are called as Population Parameters.

$(\bar{x}, s, s^2)$ are called as Sample characteristics

mean = μ

Std. dev^n = $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2}$

N = Popl^n size.

Variance = $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$

as $n \to N$, $s \to \sigma$

mean = $\bar{x}$

Std. dev = $s = \sqrt{\frac{1}{(n-1)}\cdot\sum(x_i-\bar{x})^2}$

n = sample size

Variance = $s^2 = \frac{1}{(n-1)}\sum(x_i-\bar{x})^2$

(n-1) → Bessel Correction.

dof = n-k

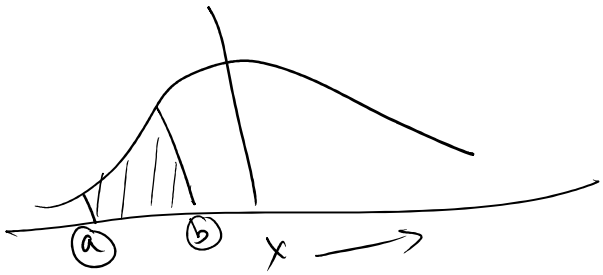
as $n \to \infty$, N ,

Bessel ...

dof = n-k

ddof = diff. between n & dof

= n − (n−k) = k.



Area under the curve between a & b

$$= \int_a^b p(x)\,dx$$

Sum of Probs of all the values that X can take between a & b.

$$P(a \le x \le b) = \int_a^b p(x)\,dx$$

Cumulative Distribution function (CDF)

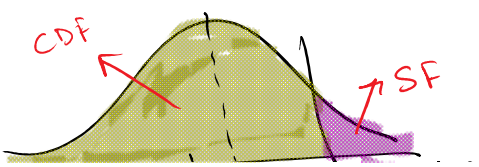
↳ represents the Prob. that a random variable X takes a value less than or equal to a.

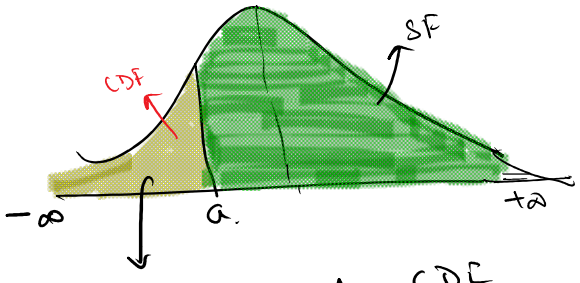
Discrete RV

$$P(x \le a) = \sum_{k=0}^{a} P(x=k) = P(x=0) + P(x=1) + P(x=2) + \dots + P(x=a)$$

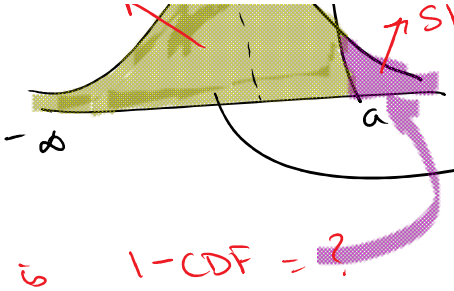
Cont. RV

$$P(x \le a) = \int_{-\infty}^{a} p(x)\,dx$$





entire area covered = CDF.

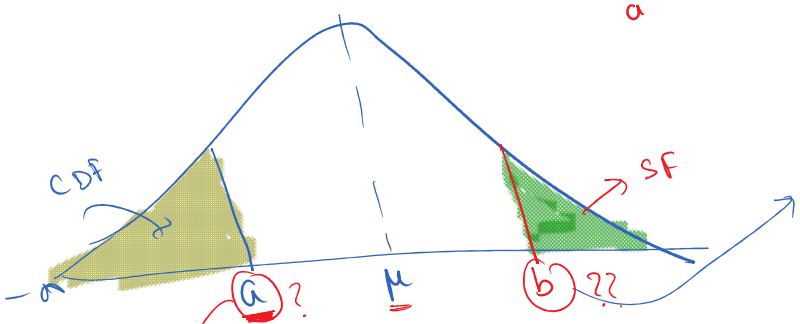


entire area covered = CDF.

What is 1-CDF = ?

SF = Survival function.

$$\int_a^{+\infty} p(x)\,dx$$



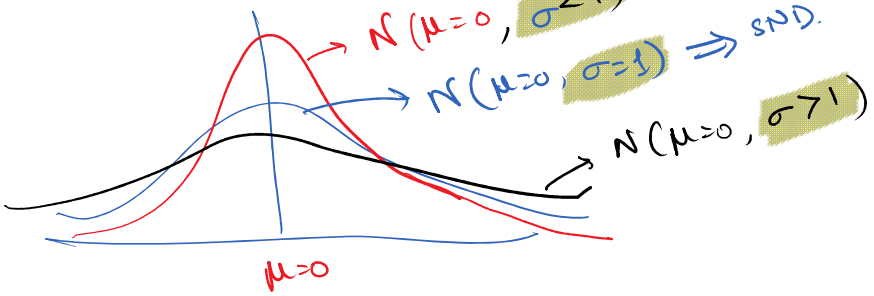
use the isf (inverse of the SF)

Until which value of x, the area under the curve = 0.2 ?

use the PPF (inverse of CDF)

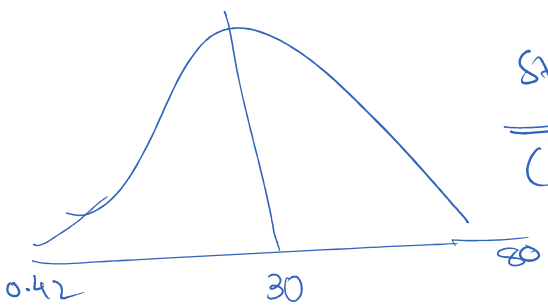
Standard Normal Distribution.

SND. → ND which has $\mu=0$, $\sigma=1$.



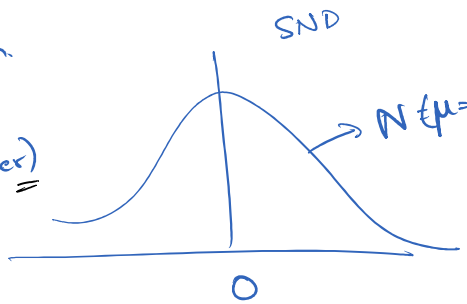
As the $\sigma\uparrow$, the distribution becomes more flatter / Broader

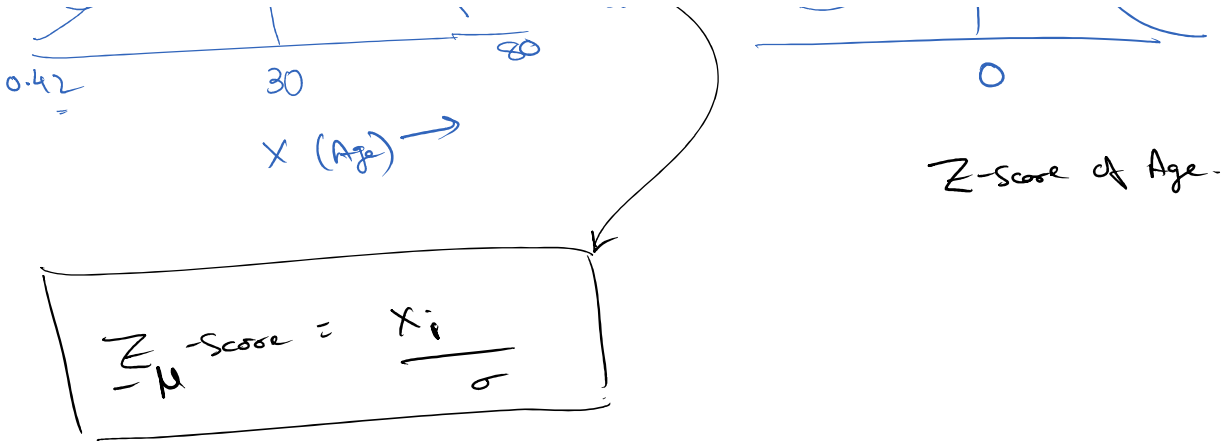
$\sigma\downarrow$, Narrower



Standardisation.

(StandardScaler) sklearn





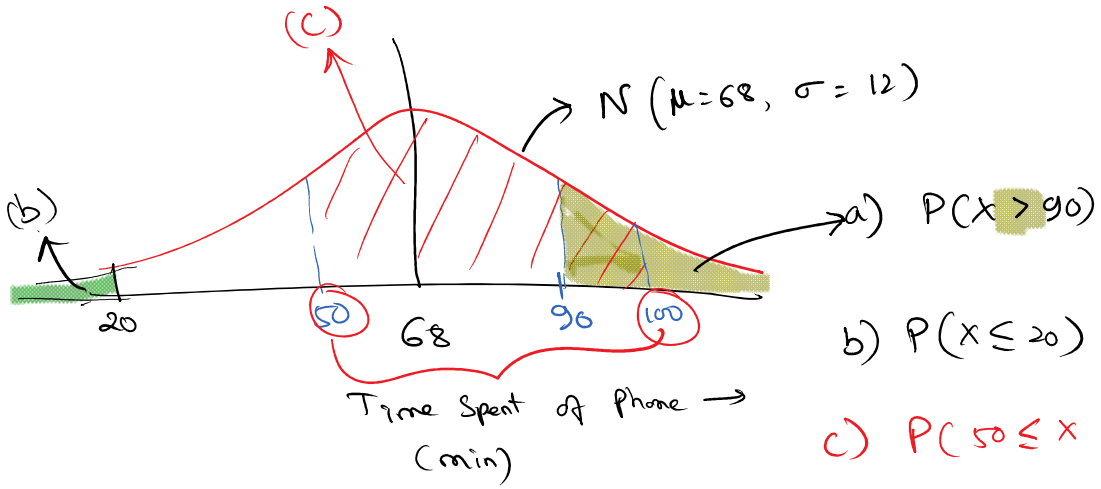
Exercise:

According to a survey on use of smart phones in India, the smart phone users spend 68 minutes in a day on average in sending messages and the corresponding standard deviation is 12 minutes. Assume that the time spent in sending messages follows a normal distribution.

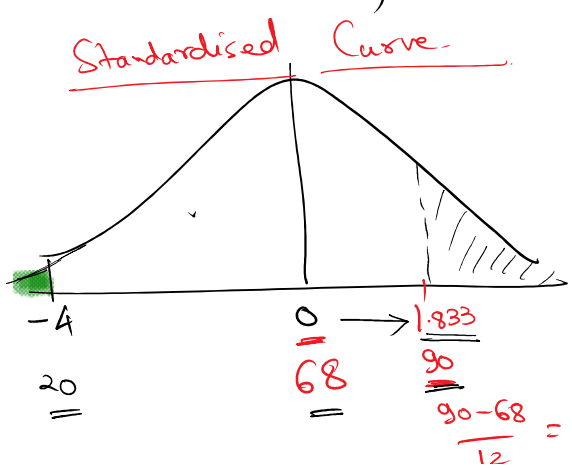
(a) What proportion of the smart phone users are spending more than 90 minutes in sending messages daily?

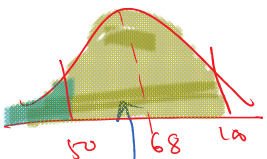
(b) What proportion of customers are spending less than 20 minutes?

(c) What proportion of customers are spending between 50 minutes and 100 minutes?



a

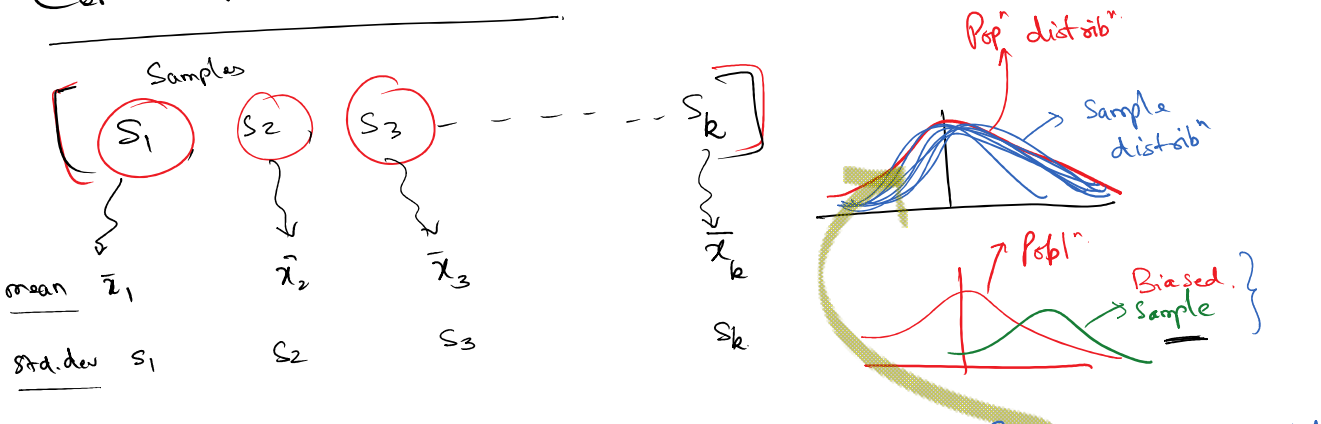




(Area till 100 − Area till 50)

$\frac{}{12} = 1.855$

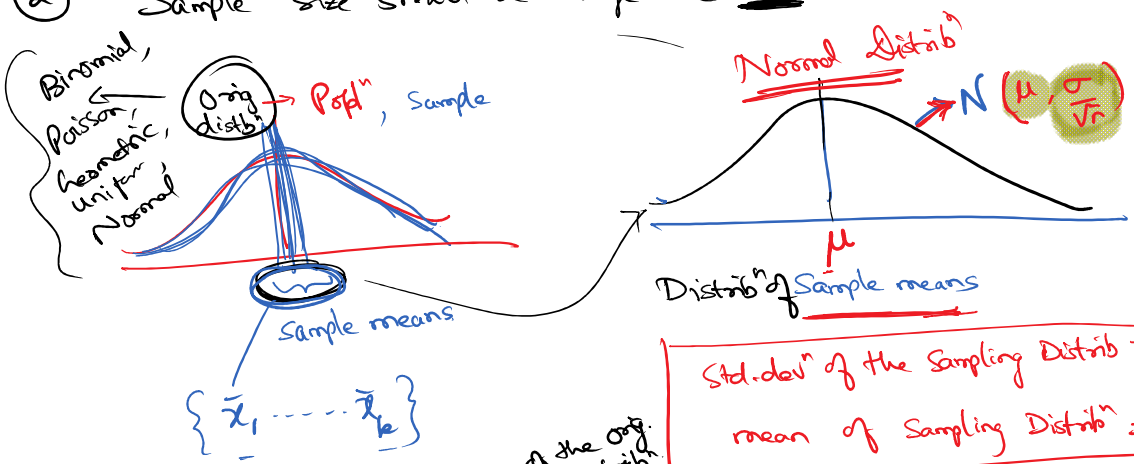
Central Limit Theorem



Assumptions: ① These are IID Samples (Independent & Identically Distributed)

basically means that each sample has been drawn independent of each other → Sampling with replacement ⇒ Bootstrap Sampling.

② Sample size should be large ($N > 30$)



Std.dev^n of the Sampling Distrib = $\frac{\sigma}{\sqrt{n}}$

mean of Sampling Distrib^n = μ.

According to CLT: Irrespective of the orig. Popl^n distrib^n, The distrib^n of Sampling means follows Normal Distrib^n

① with its mean equal to the Popl^n mean.

$$\frac{\bar{x}_1 + \bar{x}_2 + \bar{x}_3 + \dots + \bar{x}_k}{k} = \mu$$

② its Std. dev equal to $\frac{\sigma}{\sqrt{n}}$ ↓ n = Sampling size.

$\sqrt{n}$

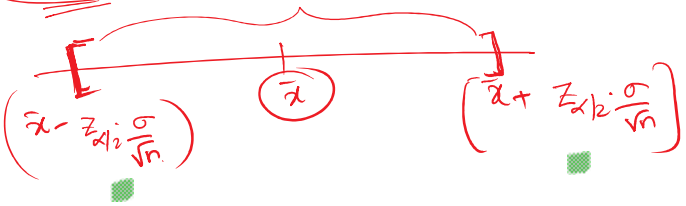
Standard Error in the estimation of the Popl^n mean from the sample means $= \frac{\sigma}{\sqrt{n}}$

$[S_1, S_2, S_3 \text{ ---------- } S_k]$

a) We can estimate the Popl^n parameters. $\Rightarrow$ Confidence Interval Estimation

$S_1 \rightarrow \bar{x}, s_1$

e.g. Popl^n mean lies somewhere $\left[\bar{x} \pm Z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right]$



b) You can Hypothesize a value of Popl^n mean, from the Sample mean & try to validate this number.

e.g. $\bar{x} = 100$, $\mu_0 > 110$ (RT), $\mu_0 < 95$ (LT), $\mu_0 = 110$ (2T)

HYPOTHESIS TESTING

Step 1 : Define H_0 & H_A.

H_0 → Null Hypothesis → formulated in a way to "Nullify" your H_A.

H_A → Alternate Hypothesis. → Whatever You claim against the "Norm" or common understanding.

(equality sign → H_0)

Step 2 : Determine which test ? Z-test vs t-test, χ^2-test etc...

If $n < 30$, t-test.

If $n < 30$, t-test.

If std. dev of popl^n (σ) is unknown → t-test.

Otherwise go for Z-test.

HT. for popl^n means.

Step 3 : Cal the Test Statistic.

$$Z_{stat} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \qquad \mu_0 = \text{Hypothesized popl}^n \text{ mean}$$

$$t_{stat} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Step 4 : Cal the Critical value. Z_{crit}, t_{crit}. (based on α)

$(1-\alpha) \Rightarrow$ confidence Level

$\alpha \rightarrow$ Significance level.

↳ min^m evidence needed in favor of H_0

Step 5 : Decision Making.

Compare Z_{stat} with Z_{crit} & conclude depending

- Left-Tailed
- Right-Tailed
- 2-Tailed

Cal p-value & compare with α.

p-value (prob/evidence in favor of H_0)

Hence, if $p > \alpha$, "Fail to Reject H_0"

$p < \alpha$, $\Rightarrow$ Reject H_0.

actual evidence presented by the sample

$p \leq \alpha \Rightarrow$ cumulative prob. (sf, cdf)

Accused of murder.

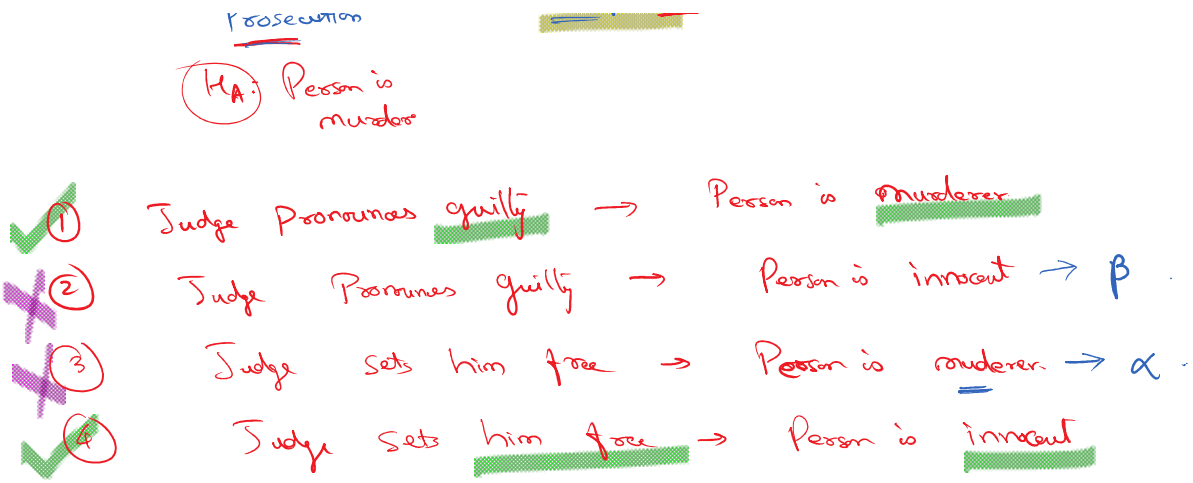
You (Judge)

Prosecution

(H_A: Person is

Defense.

H_0: Person is innocent.



Type-I and Type-II Errors

Type I Error: Conditional probability of rejecting a null hypothesis when it is true is called Type I Error or False Positive (falsely believing that the claim made in alternative hypothesis is true). The significance value a is the value of Type I error.

$$\text{Type I Error} = \alpha = P(\text{Rejecting null hypothesis} \mid H_0 \text{ is true})$$

Type II Error: Conditional probability of failing to reject a null hypothesis (or retaining a null hypothesis) when the alternative hypothesis is true is called Type II Error or False Negative (falsely believing that there is no relationship).

$$\text{Type II Error} = \beta = P(\text{Retain null hypothesis} \mid H_0 \text{ is false})$$

$$\text{Power of the test} = 1 - \beta = 1 - P(\text{Retain null hypothesis} \mid H_0 \text{ is false})$$

$$\text{Alternatively the power of test} = 1 - \beta = P(\text{Reject null hypothesis} \mid H_0 \text{ is false})$$

	Condition for rejection of null hypothesis H_0	
Type of Test	**Condition**	**Decision**
Left-tailed test	Z-statistic < Critical value	Reject H_0
	Z-statistic ≥ Critical value	Retain H_0
Right-tailed test	Z-statistic > Critical value	Reject H_0
	Z-statistic ≤ Critical value	Retain H_0
Two-tailed test	$\lvert Z\text{-statistic}\rvert > \lvert\text{Critical Value}\rvert$	Reject H_0
	$\lvert Z\text{-statistic}\rvert \leq \lvert\text{Critical Value}\rvert$	Retain H_0