Subject: Probability & Statistics Date:

Probability: A measure of how likely an event is to occur

☆ Sum of Probabilities
- Disjoint events: $P(A \cup B) = P(A) + P(B)$ Mutually exclusive
- joint events: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ Non-mutually exclusive

Independence: When the occurrence of one event does not affect the probability of the occurrence of another event.

Product rule for Independent events: $P(A \cap B) = P(A) \cdot P(B)$

Conditional Probability: Calculating the probability of an event happening given that another event has already happened.

e.g:

$$P(HH \mid 1st\ is\ H) = \frac{1}{2} \qquad P(HH \mid 1st\ is\ T) = \frac{0}{2}$$

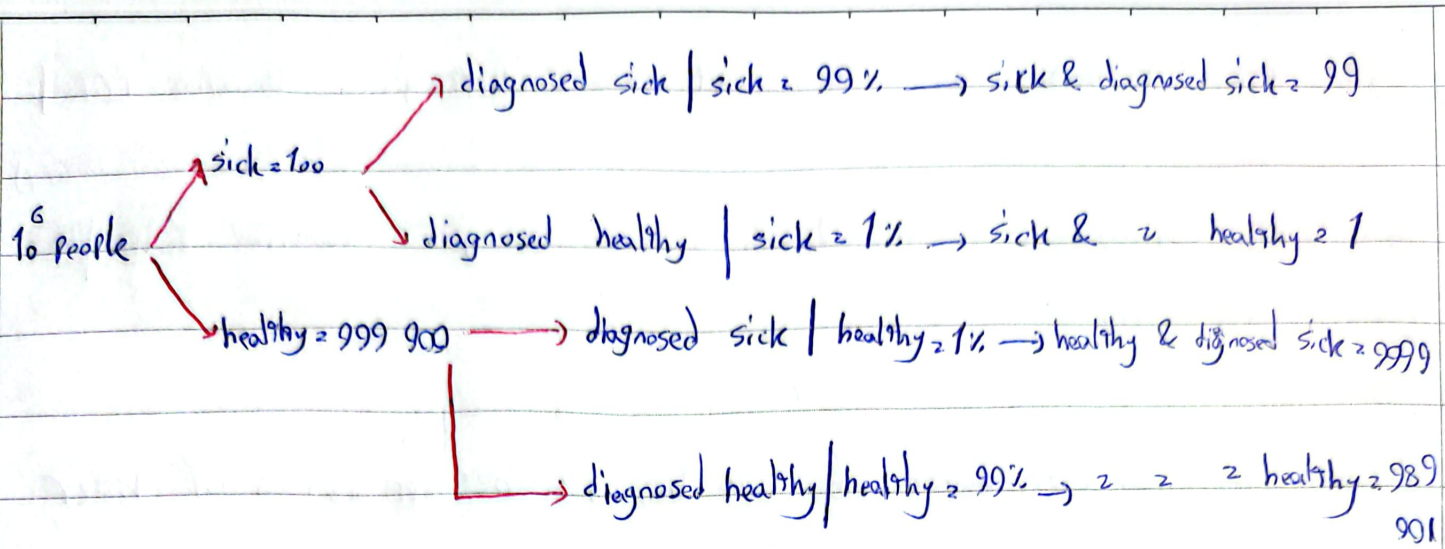
The general Product Rule:

$$P(A \cap B) = P(A) \cdot P(B|A)^{*}$$

* when independent: $P(B|A) = P(B)$

Bayes theorem: We have 1 mil People in the population and th illnes only affects 1 of every 10^5 People. The test that the doctor gives you is 99% effective (out of every 100 sick People, 99 would diagnosed sick and 1 would diagnosed healthy and out of 100 healthy People, 99 = = healthy = = = = sick)
You went to the doc and you tested sick! Are you really sick or not?

Date: Subject:



$$P(\text{sick} \mid \text{diagnosed sick}) = \frac{\text{sick and diagnosed sick}}{\text{diagnosed sick}} = \frac{99}{99+9999} = 0.0098$$

→ A = sick B = diagnosed sick

Bayes formula

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A) \times P(A)}{P(A \cap B) + P(A' \cap B)} = \frac{P(B|A) \times P(A)}{P(B|A) \times P(A) + P(B|A') \times P(A')}$$

$P(B-A) + P(A \cap B) = P(B)$

$P(B|A) = 99\%$ $P(A) = 0.01\%$ $P(A') = 99.99\%$ $P(B|A') = 1\%$

→ $P(A|B) = 0.0098$

Prior: The original probability that you can calculate not knowing anything else $P(A)$

Event: Gives us information about the probability E

Posterior: $P(A|E)$

e.g: Prior = P(sick) event = diagnosed positive Posterior = P(sick | diag pos)

* Posterior is a much more accurate probability

Subject: Date:

e.g2: Prior: $P(\text{Spam})$ event: Email contains 'Lottery' Posterior: $P(\text{Spam}|\text{lottery})$

e.g3: Prior: $P(\text{Spam})$ event: Email contains 'winning' Posterior: $P(\text{spam}|\text{winning})$

How to combine these two events?

Naive assumption (Foundation of naive bayes): the appearances of lotter and winning are 'independent (even though they are not!)

$$P(\text{Spam}|\text{lottery \& winning}) = \frac{P(\text{Spam}) \cdot P(\text{lotter \& winning}|\text{Spam})}{P(\text{Spam}) \cdot P(\text{lottery \& winning}|\text{spam}) + P(\text{ham}) P(\text{lottery \& winning}|\text{ham})}$$

$\xrightarrow{P(A \cap B) = P(A)P(B)}$

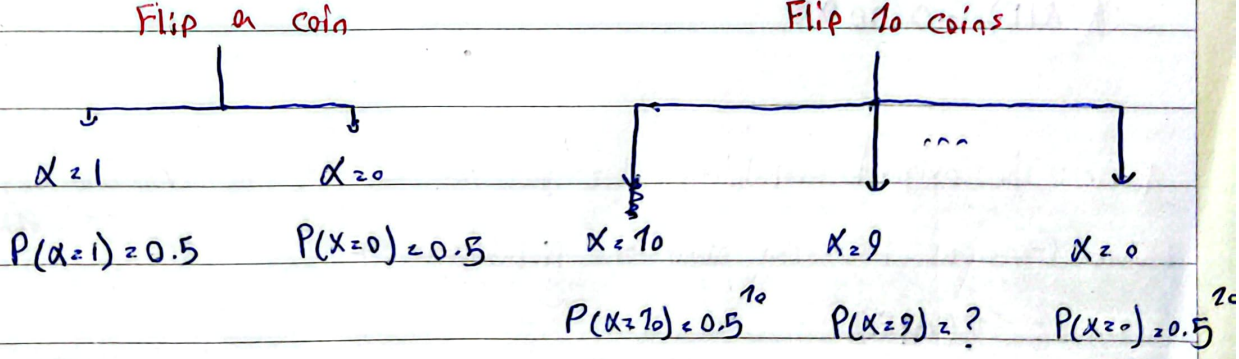
$$\frac{P(\text{Spam}) \cdot P(\text{lottery}|\text{Spam}) \; P(\text{winnigs}|\text{Spam})}{P(\text{Spam}) P(\text{lottery}|\text{spam}) P(\text{winning}|\text{spam}) + P(\text{ham}) P(\text{lottery}|\text{ham}) \; P(\text{winning}|\text{ham})}$$

in general (naive bayes):

$$P(A|B_1, B_2, \dots B_n) = \frac{P(A) \; P(B_1|A) \cdots P(B_n|A)}{P(A) P(B_1|A) \cdots P(B_n|A) + P(\bar{A}) P(B_1|\bar{A}) \cdots P(B_n|\bar{A})}$$

Random variable: Variables that can take many values → e.g: Temperature

X = number of heads



$\star \sum P(X=i) = 1$

Custom random vars: Define X: $P(X=1) = 0.5$ $P(X=-7) = 0.2$

$P(X=\text{TT}) = 0.3$

Date: Subject:

Discrete & Continuous Random Vars:

Discrete: Can take only a countable number of values

* can be finite: Number of head when flipping a coin
* Can be infinite: X: Number of flips until we get our first head

€

Continuous: Can take values on an interval. e.g: Wait time until the next bus arrives → 1 min? 1.01 min? 1.0001 min? ...

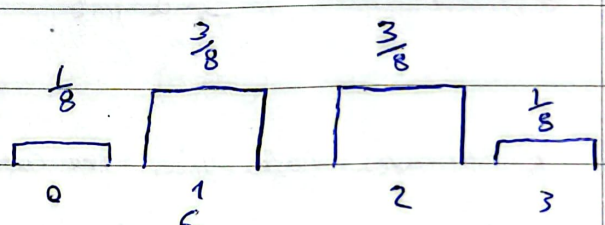
Random Vars vs Deterministic Vars

Deterministic: Fixed outcome. e.g: $X = 2$ or $f(x) = x^2$ (once it's defined, it's like that forever)

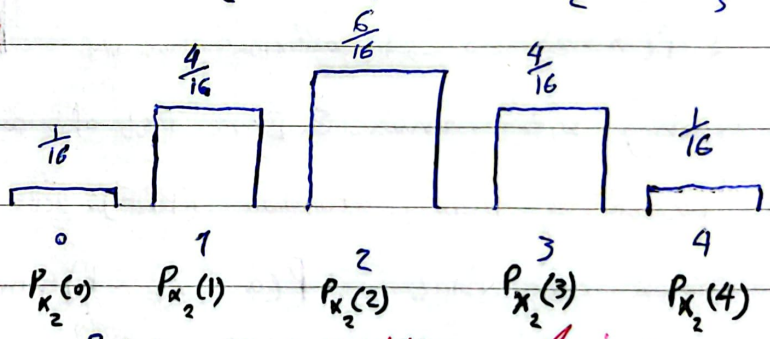
Random: Uncertain outcome. e.g: X = number of defective items in a shipment

Probability Distributions (Discrete)

X_1: number of heads in 3 coin toses



X_2: number of heads in 4 coin toses



$P(X_2 = x)$ $x = 0, 1, 2, 3, 4$ $\xrightarrow{=}$ $P_X(x)$ → (PMF: Probability mass function)

* $P_X(x) \geq 0$ * $\sum_x P_X(x) = 1$

* PMS is for Discrete Vars

Subject: Date:

Binomial Distribution: Number of heads in n coin toses? $P(H) = p$ (heads)

Event: $X = x \rightarrow x$ heads in n toses

$$P_X(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, 2, \ldots \qquad X \sim Binomial(n, p)$$

EX: What is the Probability of getting three ones when rolling a dice 5 times?

$P(one) = \frac{1}{6}$ $P(not\ one) = \frac{5}{6}$ $P_X(x) = \binom{5}{3}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^2 = 0.032$

Bernoulli distribution: $X \sim Bernoulli(p)$ p: Probability of success

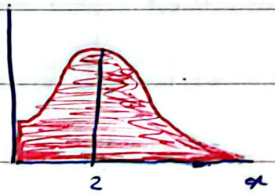
X = Number of heads — Flip a coin → H: $P(X=1) = 0.5$ success; → T: $P(X=0) = 0.5$ Failure

Probability Distribution (Continious)

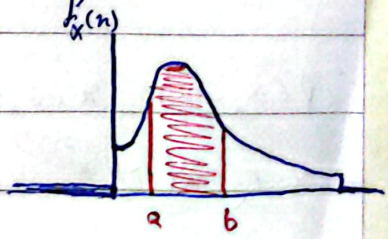
★ Discrete → a list. Continious → an interval

★ Red area (area under the curve) = 1

★ $P(x = 2) = 0$ → In continious, we are working with intervals. ∴ Exact Prob of a given point is zero.



Probability Density Function (PDF): Tells you the rate you accumulate Prob around each Point → $P(a < X < b)$ = area under $f_X(x)$

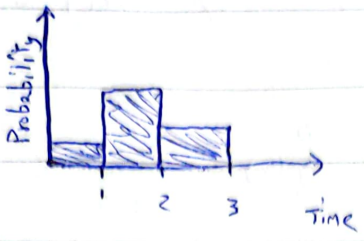


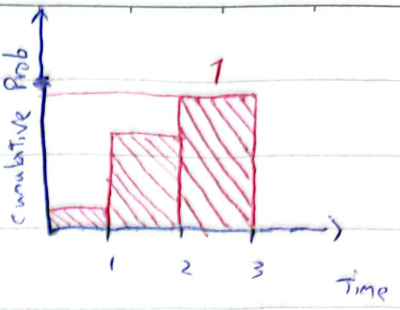
→ $f_X(x)$ needs to satisfy: 1- Defined for all numbers (it can be zero)

2- $f_X(x) \geq 0$ 3- Area under $f_X(x) = 1$

★ PDF only defined for Continuous vars

Cumulative Distribution

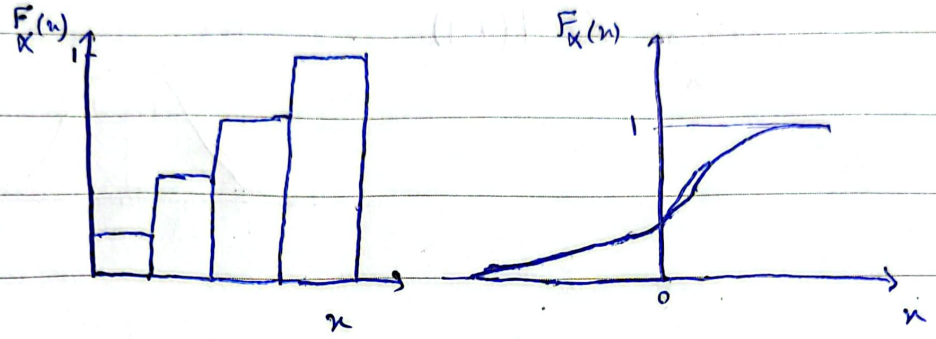




★ Starts at 0 and ends at 1

★ The CDF shows how much prob the variable has accumulated until a certain value

$CDF(x) = P(X \leq x)$ → Defined for every real number



Properties: 1- $0 \leq F_X(x) \leq 1$ 2- Left Point = 0 3- Right Point = 1 4- Never Decreases

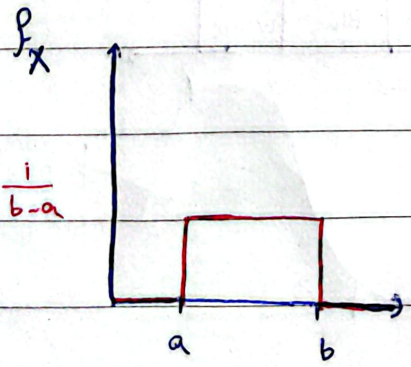
Uniform Distribution: A continuous random var can be modeled with a uniform distribution if all possible values lie in an interval and have same frequency of occurrence

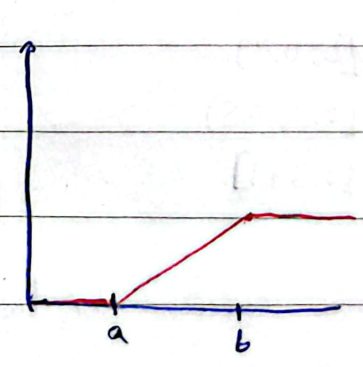
Parameters:

a: beginning of the interval

b: end " " "

$$f_X(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{ow} \end{cases}$$





$$F_X(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x < b \\ 1 & b \leq x \end{cases}$$

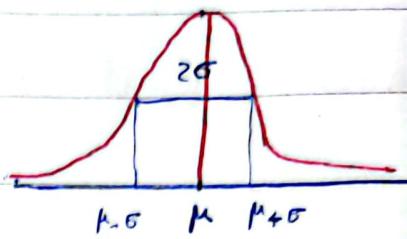
Normal (Gaussian) Distribution: $X \sim N(\mu, \sigma^2)$

Parameters:

μ: center of the bell

σ: spread of the bell

$$f_X(x) = \underbrace{\frac{1}{\sqrt{2\pi}\,\sigma}}_{\text{Scaling constant}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

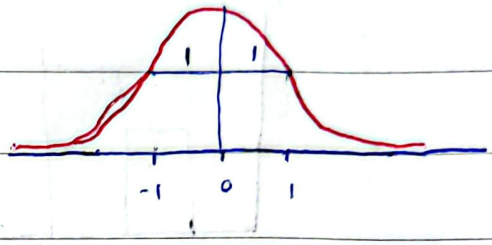


★ Range: All numbers

★ Symmetrical

Standard Normal Distribution: $X \sim N(0,1)$

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

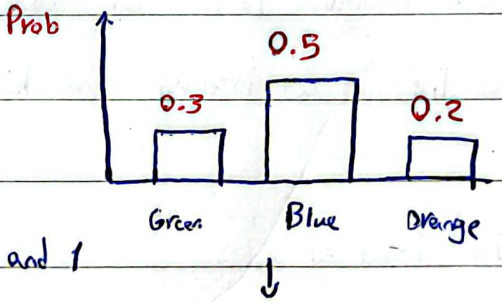


Standardization: Convert to std-Normal dis

e.g: X distributes normally with $\mu = 2$, $\sigma = 2.5$ $\rightarrow Z = \frac{X-\mu}{\sigma} = \frac{X-2}{2.5}$

Sampling from a dist:

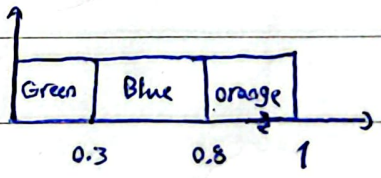
Create a random sample of data that follows a dist



1- Generate a random number between 0 and 1

2- Find out which interval the number belongs to:

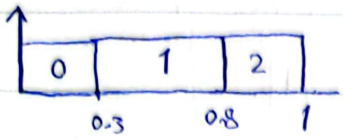
- $[0, 0.3)$
- $[0.3, 0.8)$
- $[0.8, 1]$

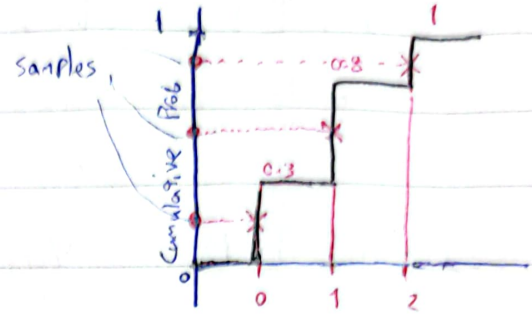


3- Assign the correct color based on the interval.

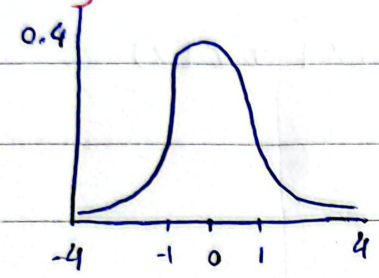
Date: Subject:

Another way:

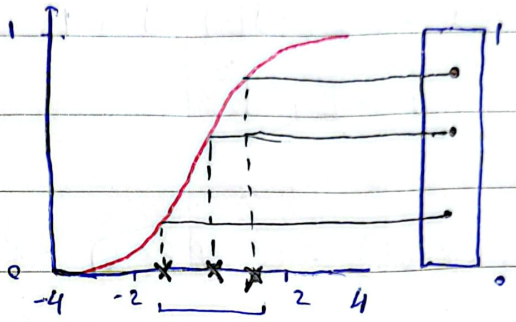


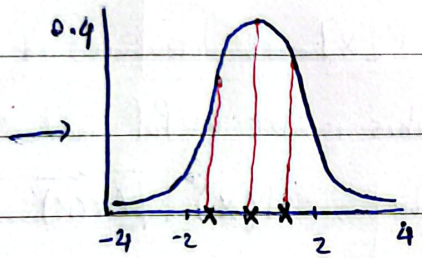


sampling from a normal dist



CDF →





★ Mean = Avg Median = data in the middle (in terms of ordered position)

Mode = the value with highest prob (we can have more than one mode → Multimodal)

★ $E[X]$ = Mean / Balancing point of a distribution → Expected Value

★ $E[X] = \sum_i P_x(x_i) x_i$ (expected value of a var) $E[f(x)] = \sum_i f(x_i) P(x_i)$ (expected value of a function)

★ $E[ax+b] = a \cdot E[x] + b$ (b: constant) $E[b] = b$

Sum of Expectation: $E[\alpha_1 + \alpha_2 + \cdots + \alpha_n] = E[\alpha_1] + E[\alpha_2] + \cdots + E[\alpha_n]$

e.g: Imagine that you have 8 billion unique name in a bag and we have 8 billion People around the world. We are going to travel the world and give each person one piece of paper. What is the expected number of Correct assignments?

Subject: Date:

N People (N ≈ 8 billion)

$E[\text{matches}] = E[X_{\text{Person 1}}] + E[X_{\text{Person 2}}] + \dots + E[X_{\text{Person N}}]$

$= \frac{1}{n} + \frac{1}{n} + \dots + \frac{1}{n} = \boxed{1}$

Variance : $E[X] = \mu$ $\quad Var(X) = E[(X-\mu)^2]$

$Var(X) = E[(X-\mu)^2] = E[X^2 - 2\mu X + \mu^2] = E[X^2] - 2\mu E[X] + E[\mu^2]$

$\rightarrow E[X^2] - 2E[X]E[X] + \mu^2 \rightarrow E[X^2] - 2E[X]^2 + E^2[X]$

$= E[X^2] - E[X]^2$

Properties of the Variance

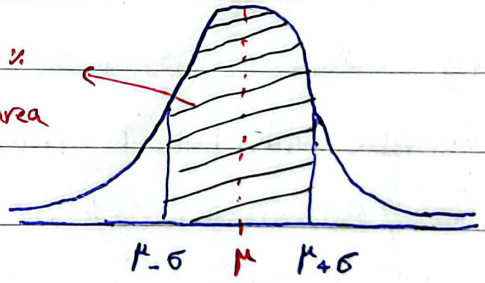
★ $Var(aX+b) = a^2 Var(X)$

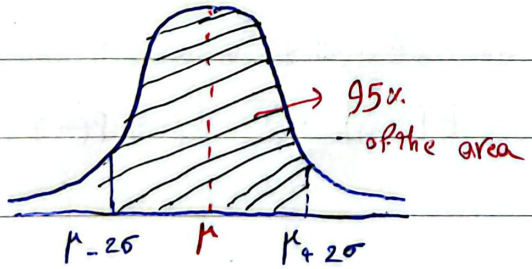
Standard Deviation : Say X is measured in unit Y, E[X] is measured in Y, Var(X) is measured in Y^2 → but sometimes (sometimes) this is not useful and we want to have measurements with the same unit as variable → $Std(X) = \sqrt{Var(X)}$

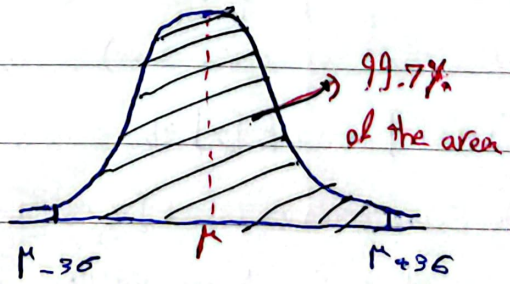
Normal distribution : 68 - 95 - 99.7 Rule

μ : Expected value

σ : Std



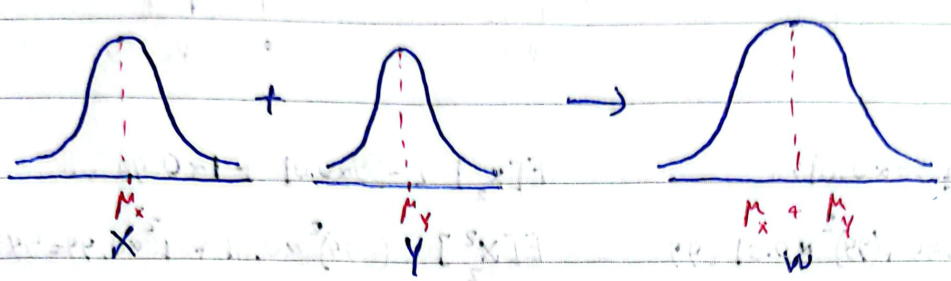




SAMANI

Date: Subject:

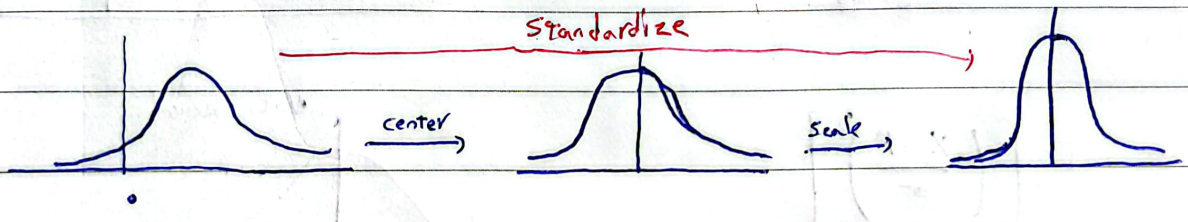
Sum of Gaussians: If $X \sim N(\mu_x, \sigma_x^2)$ and $Y \sim N(\mu_y, \sigma_y^2)$ are independent, and $W = aX + bY \Rightarrow W \sim N(a\mu_x + b\mu_y, a^2\sigma_x^2 + b^2\sigma_y^2)$



★ $\text{Var}(cX) = c^2 \text{Var}(X)$ $\quad \text{std}(cX) = c\,\text{std}(X)$

Standardize a distribution: It is good to have $\mu = 0$ and $\sigma = 1$.

$X \xrightarrow{\text{center}} X - \mu \xrightarrow[\text{scale}]{} \frac{X-\mu}{\sigma}$



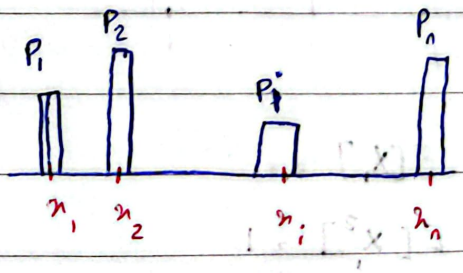
benefits: 1- Comparability between different datasets 2- Simplification of statistical analysis 3- Improved performance of ML models (improve the convergence rate of optimization algorithms and prevent some features from dominating others, leading to improved model performance)

Moment of distribution

$E[X] = p_1 x_1 + \cdots + p_n x_n$

$E[X^2] = p_1 x_1^2 + \cdots + p_n x_n^2$

$E[X^k] = p_1 x_1^k + \cdots + p_n x_n^k$



X - Random var

Subject: Date:

Scenario 1

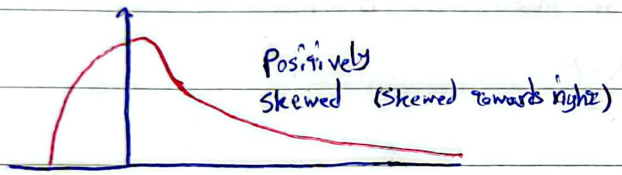
Scenario 2

$E[X_1] = -1 \times 0.99 + 99 \times 0.01 = 0$ = $E[X_2] = -99 \times 0.01 + 1 \times 0.99 = 0$

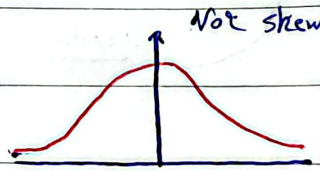
$E[X_1^2] = (-1)^2 \times 0.99 + (99)^2 \times 0.01 = 99$ = $E[X_2^2] = (-99)^2 \times 0.01 + 1^2 \times 0.99 = 99$

$E[X_1^3] = (-1)^3 \times 0.99 + (99)^3 \times 0.01 = 9702$ $\neq$ $E[X_2^3] = (1)^3 \times 0.99 + (-99)^3 \times 0.01 = -9702$

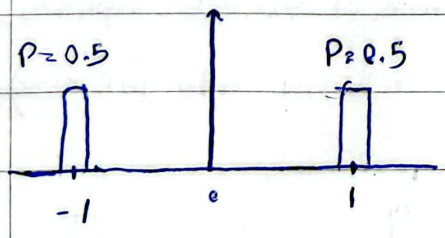
☆ The cube of the var detects if you have numbers that are skewed towards the right or skewed towards the left



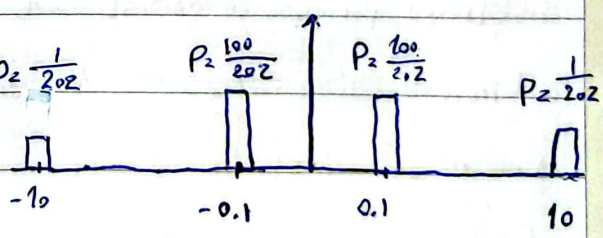
$$\text{Skewness} = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$$



Scenario 1



Scenario 2



$E[X_1] = 0$ = $E[X_2] = 0$

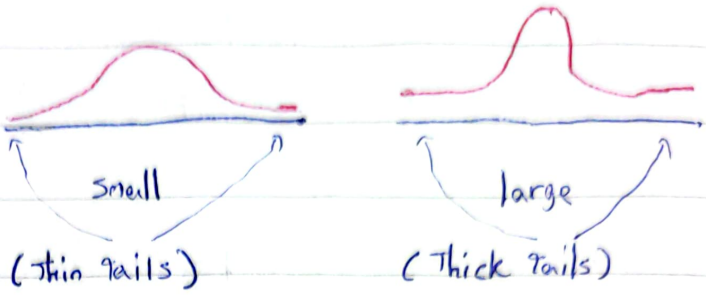
$E[X_1^2] = 1$ = $E[X_2^2] = 1$

$E[X_1^3] = 0$ $(\text{Skew}(X_1) = 0)$ = $E[X_2^3] = 0$ $(\text{Skew}(X_2) = 0)$

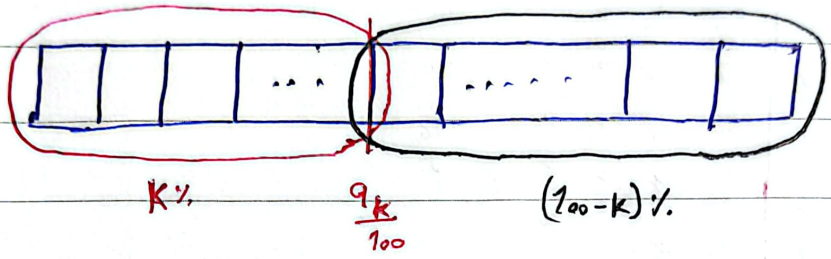
$E[X_1^4] = 1$ $\neq$ $E[X_2^4] = 99.01$

Date: Subject:

Kurtosis: $E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]$



Quantiles: the k% quantile ($q_{k/100}$) is the value that leaves k% of your data to the left and (100-k)% of your data to the right



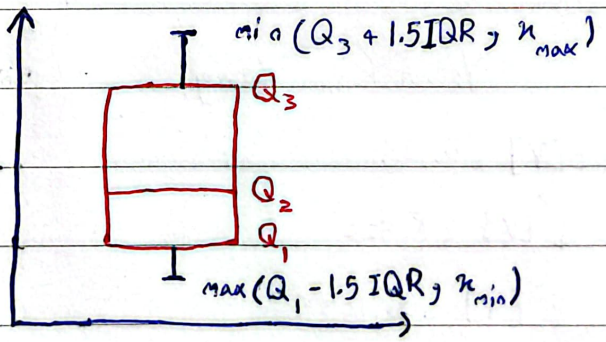
→ some common quantiles: 1- 25% quantile (first quartile - Q1)

2- 50% quantile (median - Q2) . 3- 75% quantile (third quartile - Q3)

★ k% quantile ($q_{k/100}$) is the value such that $P(X \leq q_{\frac{k}{100}}) = \frac{k}{100}$

★ interquartile range (IQR) = $Q_3 - Q_1$

Box-Plots:



Benefits:

- Data is skewed?
- No outlier data? (data were cut at max and min value)
- Analyze dispersion

Joint Distribution

$P_{XY}(x,y) = P(X=x, Y=y)$

$P_{XY}(8,48) = P(X=8, Y=48) = 0$

$P_{XY}(7,46) = \frac{2}{10}$

Age (X) \ Height (Y)	45	46	47	48	49	50
7	$\frac{1}{10}$	$\frac{2}{10}$	0	0	0	0
8	0	0	$\frac{2}{10}$	0	0	0
9	0	0	0	0	$\frac{3}{10}$	$\frac{1}{10}$
10	0	0	0	0	0	$\frac{1}{10}$

Subject: Date:

X = the number rolled on the 1st dice

Y = " " " " " 2nd "

→ X and Y are independent → $P_{XY}(x,y) = P(x).P(y)$

X = the number rolled on the 1st dice Y = sum of two dice

12	0	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$
11	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	0
10	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0
9	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0
8	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0
7	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0
6	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	0
5	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	0	0
4	$\frac{1}{36}$	$\frac{1}{36}$	0	0	0	0
3	$\frac{1}{36}$	0	0	0	0	0
2	$\frac{1}{36}$	0	0	0	0	0
1	0	0	0	0	0	0
	1	2	3	4	5	6

$P(3,7) = \frac{1}{36}$

$P(1,1) = 0$

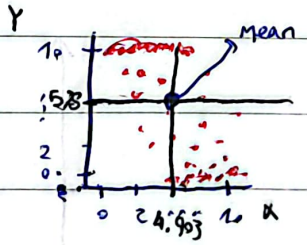
Joint Continuous Distribution:

X: waiting time Y: satisfaction rating

$E[X] = 4.903$ $E[Y] = 5.28$

$Var(X) = E[X^2] - E^2[X] = 8.526$

$Var(Y) = E[Y^2] - E^2[Y] = 10.163$



Marginal Distribution: reduce the dimensions that we don't need anymore

$P_Y(y_j) = \sum_i P_{XY}(x_i, y_j)$

$P_X(x_i) = \sum_j P_{XY}(x_i, y_j)$

Age (X) \ Height (Y)	45	46	47	48	49	50	
7	$\frac{1}{10}$	$\frac{2}{10}$	0	0	0	0	$\frac{3}{10}$
8	0	0	$\frac{2}{10}$	0	0	0	$\frac{2}{10}$
9	0	0	0	0	$\frac{3}{10}$	$\frac{1}{10}$	$\frac{4}{10}$
10	0	0	0	0	0	$\frac{1}{10}$	$\frac{1}{10}$
	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{2}{10}$	0	$\frac{3}{10}$	$\frac{2}{10}$	

SAMANI

Date: Subject:

Discrete Conditional Distribution $f_{Y|X=x}(y) = \frac{P_{XY}(x,y)}{P_X(x)}$

e.g: $P_{Y|X=9}(49) = \frac{3}{4}$

Continuous Conditional Distribution $f_{Y|X=x}(y) = \frac{f_{XY}(x,y)}{f_X(x)}$

Covariance: find the relation between two variables $Cov(X,Y) = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{n}$

$Cov(X,Y) > 0$ → Variables grow together

$Cov(X,Y) \simeq 0$ → variables are independent from each other

$Cov(X,Y) < 0$ → when one of them increases, the other one decreases

★ if the probabilities are not equal → $Cov(X,Y) = \sum P_{XY}(x_i,y_i)(x_i - \mu_x)(y_i - \mu_y)$

★ $Cov(X,Y) = E[XY] - E[X]E[Y]$

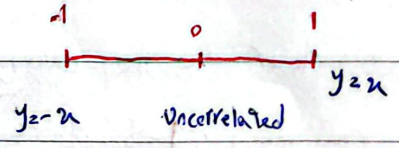
Covariance Matrix Σ:

	X	Y
X	Var(X)	Cov(X,Y)
Y	Cov(X,Y)	Var(Y)

	X	Y	Z
X	Var(X)	Cov(X,Y)	Cov(X,Z)
Y	Cov(X,Y)	Var(Y)	Cov(Y,Z)
Z	Cov(X,Z)	Cov(Z,Y)	Var(Z)

★ $Cov(a,b) = Cov(b,a)$

Correlation Coefficient: $\frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$



Multivariate Gaussian Distribution $f_X(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$

$f_X(x_1,\dots,x_n) = \frac{1}{(2\pi)^{\frac{n}{2}}\,|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$ $\quad x = [x_1, x_2, \dots, x_n]^T$

Subject: Date:

Population (N): The entire group of individuals or elements you want to study which share a common behaviour

Sample (n): Subset of the population you use to draw conclusions about the population as a whole

★ Every dataset you work with in machine learning is a sample

Proportion: $p = \frac{\text{number of items with a given characteristic } (x)}{\text{Population } (n)}$

★ Sample proportion → $\hat{p}$ or p'

Population Variance: $\sigma^2 = \frac{1}{N} \sum (x-\mu)^2$ (Population mean)

$\text{Var}(X)$ (Population size)

Sample Variance: $S^2 = \hat{\text{Var}}(x) = \frac{1}{n-1} \sum (x-\bar{x})^2$ (sample mean)

(sample size)

Law of Large Numbers: As the sample size increases, the average of the sample will tend to get closer to the average of the entire population.

(under certain conditions): 1-Sample is randomly drawn. 2-Sample size must be sufficiently large 3-Independent observations

Central limit theorem: As n (size of sample) increases, the probability distribution becomes closer to a Gaussian Distribution with $\bar{x} \approx \mu = np$ and $\sigma^2 \approx np(1-p)$

SAMANI

Date: Subject:

CTL - Example 2 : Continuous Random Var (Independent) $Y_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

$\rightarrow \mu_{Y_n} = \mu$ → Population $\quad \sigma^2_{Y_n} = \frac{\sigma^2}{n}$ → Population Variance

Maximum likelihood estimation : Find model that most likely produced the data or Maximize P(Data | model)

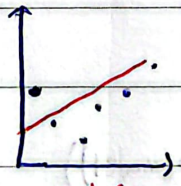
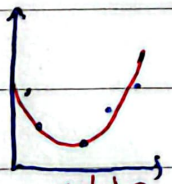
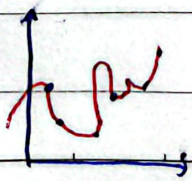
Bernoulli Example: n coins, k heads $\quad X = (X_1, \dots, X_n)$, $X_i \overset{iid}{\sim}$ Bernoulli(p)

Likelihood $L(p; x) = P_p(X = x) = \prod_{i=1}^{n} P_{X_i}(x_i) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$ take log - take derivative and find $\hat{p}$

$\rightarrow \hat{p} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$

Gaussian Example : The best distribution is the one where the mean-variance of the dist is the mean-variance of the sample

Linear Regression Example : Find a line that minimizes the sum of squared distance from the point (covered in Calculus)

		Loss	Equation	Penalty	New loss
Regularization	 Model	7	$y = 4x + 3$	$L_2 = 4^2 + 16$	26
 Data	 model 2	2	$y = 2x^2 - 4x + 5$	$L_2 = 20$	22
	 model 3	0.1	$y = 4x^{10} - 9x^8 - 2x^6 + 3x^5 - 6x^4 - 10x$	$L_2 = 246$	246.1

Subject: Date:

Model: $y = a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x + a_0$

Log-less : ll

L_2 Regularization Error: $a_n^2 + a_{n-1}^2 + \dots + a_1^2$

Regularization parameter : λ

Regularized error : $ll + \lambda(a_n^2 + \dots + a_1^2)$

Maximum Likelihood with Bayes

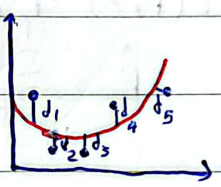
$P(Data|Model) \cdot P(Model) \xrightarrow[\text{logarithms}]{\text{Take}}$ Square-loss + Regularization term

Polynomial regression with regularization

$P(Model\ 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}a_1^2}$

$a_1 x + b$

$a_i \sim N(0,1)$ ☆



$P(Model\ 2) = \prod_{i=1}^{2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}a_i^2}$

$a_1 x^2 + a_2 x + b$

$P(Data|Model_2) = \prod_{i=1}^{5} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}d_i^2}$

$P(Model\ 3) = \prod_{i=1}^{10} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}a_i^2}$

$a_1 x^{10} + \dots + a_{10} x + b$

→ Maximize : $\log(P(Data|Model_2)) + \log(P(Model_2)) \rightarrow -\frac{1}{2}[d_1^2 + \dots + d_5^2 + a_1^2 + a_2^2]$

→ Minimize : $d_1^2 + \dots + d_5^2 + a_1^2 + a_2^2$

Date: Subject:

Frequentist Vs Bayesian Statistics :

Frequentists: 1- Probabilities represent long term frequency of events 2- Concept of likelihood

3- Goal: Find the model that most likely generated the observed data

Bayesians: 1- Probabilities represent the degree of belief (or certainty)

2- concept of prior 3- Goal: Update prior belief based on observations

Maximum A posterior (MAP-Bayesian): Value for the parameter? → choose the one with highest Probability → Make the updated belief (Posterior)

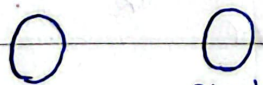
$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

- $P(A|B)$: Posterior: A will happen after considering B evidence
- $P(B|A)$: Likelihood of evidence B appearing, given A happened
- $P(A)$: Prior: A will happen before considering evidence B
- $P(B)$: Probability of evidence B in any circumstances → $P(B|A)P(A) + P(B|A')P(A')$

A: an event you are trying to predict

B: another event, or evidence, that helps refine your prediction

e.g: We have two types of coin and a Mystery coin that is either fair or biased



Fair: $P(H) = 0.5$ Biased: $P(H) = 0.8$

Event to predict: $A \to Y$ takes some value; Y: odds of H for your coin

Evidence: $B \to X$ take some value; X: result of coin flip

Priors: $P(Y=0.5) = 0.75$, $P(Y=0.8) = 0.25$

$$Y = \begin{cases} 0.5 & \text{if coin is fair} \\ 0.8 & \text{if coin is biased} \end{cases} \qquad X = \begin{cases} 0 & \text{if T} \\ 1 & \text{if H} \end{cases}$$

$X = 1$

$$P_{Y|X=1}(0.5) = \frac{P_{X|Y=0.5}(1)\;P_Y(0.5)}{P_X(1)} = \frac{0.5 \cdot 0.75}{0.5 \times 0.75 + 0.8 \times 0.25} \to \text{Posterior}$$

$P(Y=0.5|X=1) = 0.652$

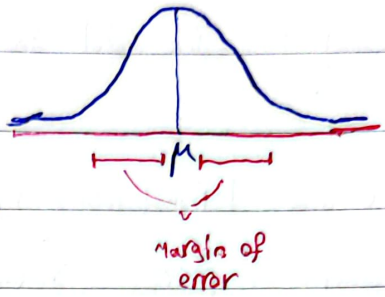
$P(Y=0.8|X=1) = 0.348$ → New Prior

Subject: Date:

Confidence Interval: Interval of values which contains the population parameter (like μ) with some degree of certainty.

$X \sim N(\mu, \sigma^2)$ (unknown, known)

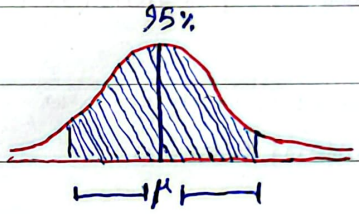
$\bar{X} \sim N(\mu, \sigma^2)$: Describe the prob of selecting different sample means (of size = 1)



★ Confidence level: Prob that your sample mean is within the margin of error

★ significance level (α): 1 − confidence level → confidence level = $1-\alpha$

e.g: $\alpha = 0.05$ → confidence level: 0.95 or 95%.



→ 95% of all randomly generated sample means would fall within the shaded portion of the distribution curve

Confidence interval: $\bar{x} \pm$ margin of error
($\bar{x}$: sample mean)

★ when confidence level is a% → a% of confidence intervals would contain μ

→ what if sample size = k (instead of 1) → $\mu_{\bar{x}} = \mu$ $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{k}}$.

★ Increasing sample size
- → Decreasing margin of error
- → Increasing accuracy ($\bar{x}$ is closer to μ)
- → as margin of error decreases → Confidence interval shrinks

★ Decreasing confidence level → Decreasing margin of error

★ Ideally you want high confidence level and narrow interval → more data would help!

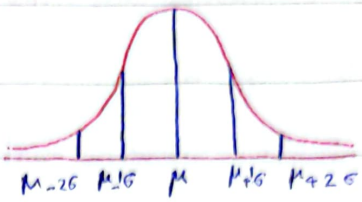
SAMANI

Date: Subject:

Margin of error

$(\mu-1\sigma, \mu+1\sigma) \rightarrow 68\%$

$(\mu-2\sigma, \mu+2\sigma) \rightarrow 95\%$

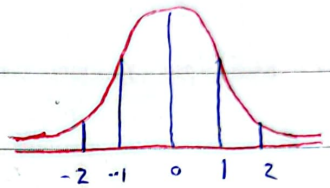


O : z-score

★ Z-distribution (standard normal distribution) : $\frac{X-\mu}{\sigma} = Z \sim N(0,1^2)$

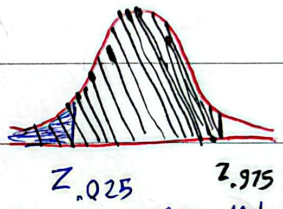
$(-1,1) \rightarrow 68\%$ $(-2,2) \rightarrow 95\%$

★ $(-1.96, 1.96)$ is exact point for 95%



↳ critical values $Z_{.\alpha}$ = it has $\alpha\%$ to it's left

$Z_{.975}$ = 97.5% to it's left



★ if sigaificant level $= \alpha \rightarrow Z_{1-\alpha/2} - Z_{\alpha/2}$: between these two values, lies $(1-\alpha)\%$ of the distribution

But if we had non-standardized normal dist, we multiply these critical values by standard deviation → e.g : 95% : $(\mu-1.96\sigma, \mu+1.96\sigma)$

★ Since $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \rightarrow$ 95% : $(\mu-1.96\frac{\sigma}{\sqrt{n}}, \mu+1.96\frac{\sigma}{\sqrt{n}})$

In general : Margin of Error : $(\mu - Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \mu + Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}})$

So: $\mu - 1.96\frac{\sigma}{\sqrt{n}} < \bar{x} < \mu + 1.96\frac{\sigma}{\sqrt{n}} \rightarrow \bar{x} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}$

Confidence interval: $\bar{x} \pm Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$

Subject: Date:

Confidence interval - Calculation Steps:

1- Find sample mean $\bar{x}$

2- Define a desired confidence level $(1-\alpha)$

3- Get the critical value $(z_{1-\frac{\alpha}{2}})$

4- Find the standard error $(\frac{\sigma}{\sqrt{n}})$

5- Find the margin of error : $z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$

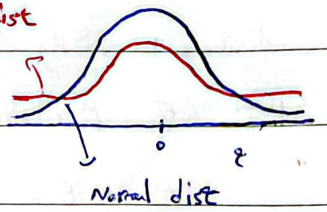
6- Add/subtract the margin of error to the sample mean $\bar{x} \pm z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$ Confidence interval

★ Assumptions: 1- simple random sample 2- Sample size >30 or population is approx normal

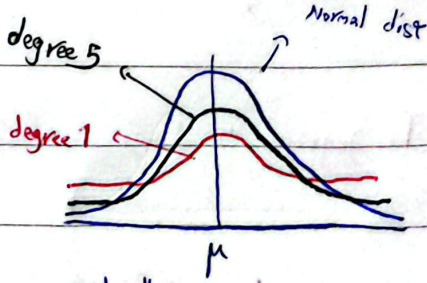
★ $n \geq \left(\frac{z_{1-\frac{\alpha}{2}} \cdot \sigma}{\text{margin of error (MOE)}} \right)^2$: sample size to get the desired MOE

Unknown σ : $\sigma \rightarrow s$ (sample standard deviation) $\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \rightarrow \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$

Student's t dist: t dist normal dist t-dist



$\bar{x} \pm z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \rightarrow \bar{x} \pm t_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$



★ There are several t dist! degree of freedom = sample size $n-1$

★ The larger the number of degrees of freedom → closer to a normal dist

Confidence interval for proportions:

Confidence interval = $\hat{p} \pm$ margin of error

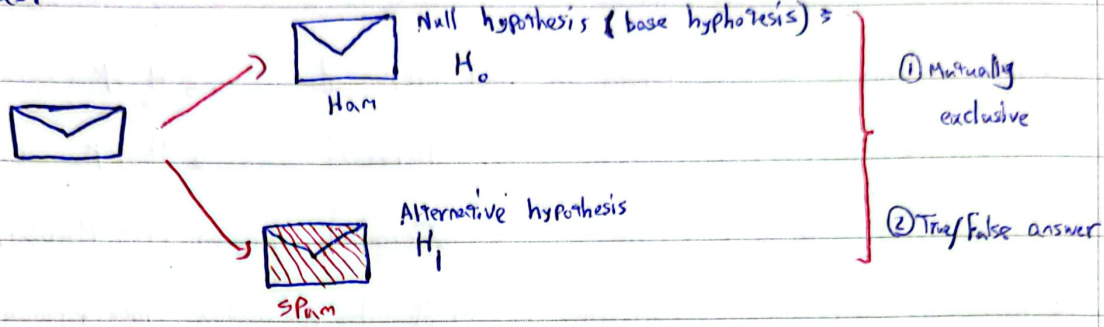
margin of error = $z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

SAMANI

Date: Subject:

Hypothesis testing: Is a way to tell if some belief you have of the population is true or not

e.g: email spam detector



★ If there was alot of evidence that shows that the email is spam → Null hypothesis rejected and alternative is accepted

★ If the evidence gathered is not enough → we are not accepting that the email is ham!

★ we decide between two hypothesis base on: Data - Evidence

Type I and Type II Errors

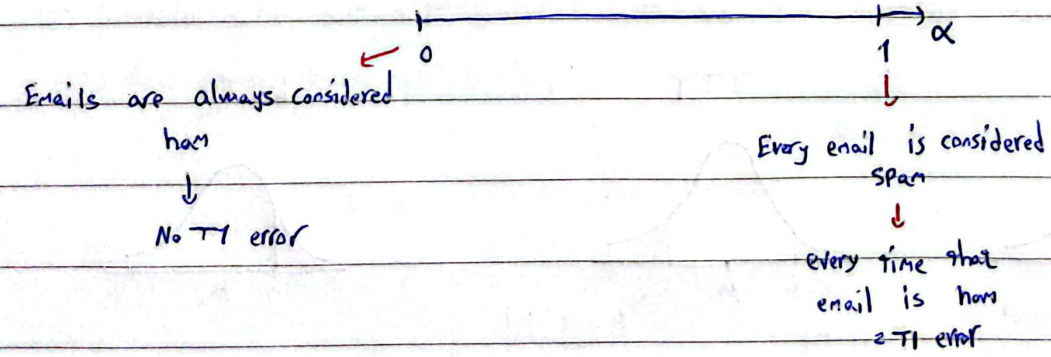
T1 → False positive

T2 → False Negative

Decision	Reality	
	H_0 True	H_0 False
Reject H_0	Type I	✓
Don't Reject H_0	✓	Type II

★ In this case, Type I errors are worse than Type II

Significance level: what is the greatest prob of Type I error you are willing to tolerate?



★ $\alpha = 0.05$ is a good choice

★ ↓ Type I error → ↑ Type 2 error (Too much)

→ $\alpha = \max P(\text{T1 error}) = \max P(\text{Reject } H_0 \mid H_0)$

Data Quality: Data should be Reliable
- → Representative (of the population)
- → Randomized (No bias)
- → Sample size (consider 30 samples or more)

Subject: Date:

e.g: The mean height for 18 yro in the US in the 70s was 66.7 in, what about now?

3 questions → 3 set of hypothesis

① Right-Tailed Test: Population mean increased → $H_0: \mu = 66.7$ vs $H_1: \mu > 66.7$

② Left-Tailed Test: " " decreased → $H_0: \mu = 66.7$ vs $H_1: \mu < 66.7$

③ Two-Tailed Test: " " changed → $H_0: \mu = 66.7$ vs $H_1: \mu \neq 66.7$

★ hypothesis are always formulated in term of the population (like population mean)

★ decisions " based on the observations (samples) Here: sample mean → $\bar{X} = \frac{1}{10}\sum_{i=1}^{10} X_i$

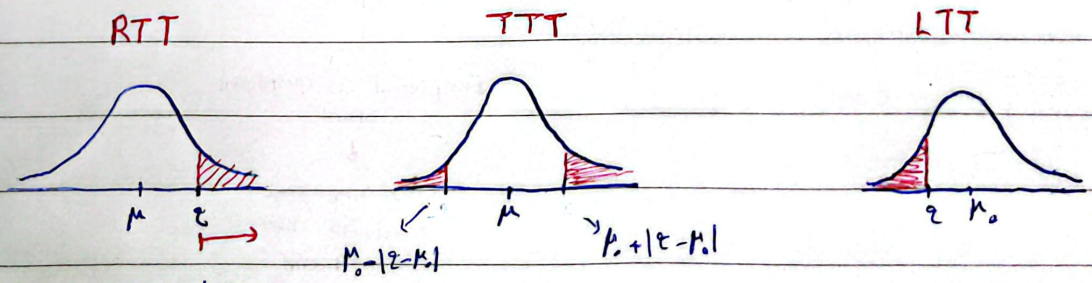
$\bar{X}$ → test statistic: a function of random samples $T(X) = X = (X_1, \dots, X_n)$

① RTT: T1: Determine $\mu > 66.7$ when $\mu = 66.7$
T2: Determine $\mu = 66.7$ " $\mu > 66.7$

② LTT: T1: Determine $\mu < 66.7$ when $\mu = 66.7$
T2: " $\mu = 66.7$ " $\mu < 66.7$

③ TTT: T1: Determine $\mu \neq 66.7$ when $\mu = 66.7$
T2: " $\mu = 66.7$ " $\mu \neq 66.7$

P-values: A p-value is the probability, assuming H_0 is true, that the test statistic takes on a value as extreme as or more extreme than the value observed (t)



$P(T(X) > t \mid H_0)$ $\quad P(|T(X) - \mu_0| > |t - \mu_0| \mid H_0)$ $\quad P(T(X) < t \mid H_0)$

→ if we have $\bar{X} = 68.492$ and the goal is T1 error prob (P-value) $< \alpha$
$\sigma = 3$, $n = 10$ (observed value)
→ If P-value $< \alpha$ → reject H_0 (and accept H_1)
→ If P-value $> \alpha$ → don't reject H_0

SAMAND

Date: Subject:

RTT:

$P(\bar{X} > 68.442 \mid \mu = 66.7) = 0.0332 < \alpha = 0.05$ → reject H_0 (with 5% significance level)

TTT:

$P(|\bar{X} - 66.7| > |68.442 - 66.7| \mid \mu = 66.7) = 0.0663 > \alpha$ → Don't reject H_0

LTT:

$P(\bar{X} < 64.252 \mid \mu = 66.7) = 0.0049 < \alpha$ → reject H_0

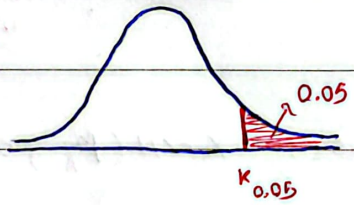
★ If H_0 is true: $\bar{X} \sim N(\mu_0, \frac{3^2}{10})$ or $Z = \frac{\bar{X} - \mu_0}{3/\sqrt{10}} \sim N(0,1)$ (Z-statistic)

Critical Value (K_α): sample that has p-value = α (least extreme sample you could get to reject H_0)

How to Compute: $\alpha = 0.05$ (Height example) $\quad 0.05 = P(\bar{X} > K_{0.05} \mid \mu = 66.7)$

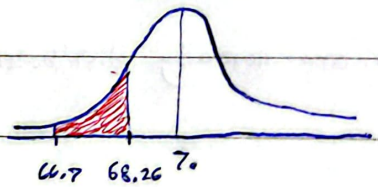
$K_{0.05} = 68.26$

Decision rule: Reject H_0 if $\bar{X} > 68.26$



★ what is the T2 error prob if the true value is $\mu = 70$? if $\mu = 70$, $\bar{X} \sim N(70, \frac{3^2}{10})$

$\beta = P(\bar{X} < 68.26 \mid \mu = 70) = 0.0333$ (H_0 is true)



T2 error: $P(\text{Don't reject } H_0 \mid \mu \in H_1) = \beta$

Power of the test: $P(\text{Reject } H_0 \mid \mu \in H_1)$

} complementary → Power = $1 - \beta$

Steps for Performing Hypothesis testing:

1- State your hypothesis

a) Null hypothesis: the baseline → $H_0: \mu = 66.7$

b) Alternative hypothesis: the statement you want to prove → $H_1: \mu > 66.7$

Subject: Date:

2- Design your test

a) Decide the test statistic to work with → $\bar{X}$

b) Decide the significance level → $\alpha = 0.05$

3- Compute the observed statistic (based on your sample) $\bar{x} = 68.442$

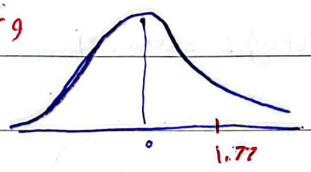
4- Reach a conclusion:

a) if p-value $< \alpha$ → reject H_0

t-dist

if σ is unknown: If H_0 is true → $T = \dfrac{\bar{X} - 66.7}{S/\sqrt{10}} \sim t_9$ (t-dist with $\nu = 9$ (degree of freedom))

repeat above process with t_9:



Independent Two-Sample t-Test: $X \sim N(\mu_A, \sigma_A^2)$ $Y \sim N(\mu_B, \sigma_B^2)$

*Assumptions: 1- All elements in the sample from the two group are different

2- Each element in both samples are independent

3- Populations are normally distributed

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\bar{Y} = \frac{1}{m}\sum_{i=1}^{m} Y_i$$

$$\bar{X} - \bar{Y} \sim N\left(\mu_A - \mu_B, \frac{\sigma_A^2}{n} + \frac{\sigma_B^2}{m}\right)$$

if σ is known:

$$\frac{(\bar{X} - \bar{Y}) - (\mu_A - \mu_B)}{\sqrt{\frac{\sigma_A^2}{n} + \frac{\sigma_B^2}{m}}} \sim N(0,1)$$

if σ is unknown:

Date: Subject:

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_A - \mu_B)}{\sqrt{\frac{S_A^2}{n} + \frac{S_B^2}{m}}} \sim t_v$$

Paired t-Test: Difference between samples: $\bar{D} = \frac{D_1 + D_2 + \dots + D_n}{n}$ $D_i = X_i - Y_i$

$D_i \sim N(\mu_D, \sigma_D^2)$

$$\frac{\bar{D} - \mu_D}{\sigma_D / \sqrt{n}} \sim N(0, 1^2) \xrightarrow{\sigma_D \text{ is unknown}} \sigma_D = S_D \rightarrow T = \frac{\bar{D} - 0}{S_D / \sqrt{n}} \sim t_{n-1}$$

Test for Proportions: • P is the population proportion of individuals in a particular category

• P_0 is the population proportion under the null hypothesis

• x is the observed number of individuals in the sample from the specified category

• n is the sample siz   • $\hat{P} = \frac{x}{n}$ is the sample proportion

$$\rightarrow Z = \frac{\frac{x}{n} - P_0}{\sqrt{P_0(1 - P_0)}} \sqrt{n} \sim N(0, 1)$$

Two Sample Test for Proportions: • $P_1 - P_2$ is the dif in the population proportion between two groups

• x is the observed number of individuals in the sample from the specified category from group 1

• y " " " " " " " " " " " " " 2

• n_1 is the sample size for group 1

• n_2 " " " " " " 2

$$Z = \frac{\frac{X}{n_1} - \frac{Y}{n_2} - 0}{\sqrt{(X + Y)\left(1 - \frac{X + Y}{n_1 + n_2}\right)}} \sqrt{n_1 \cdot n_2} \sim N(0, 1^2)$$