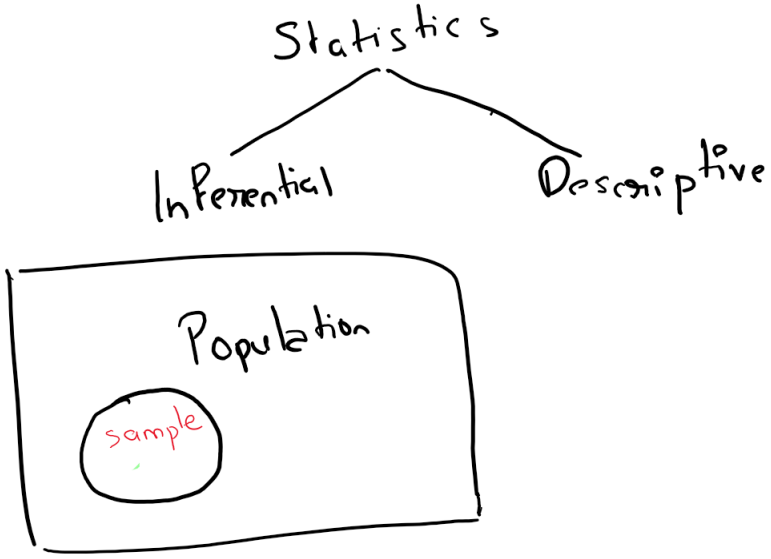
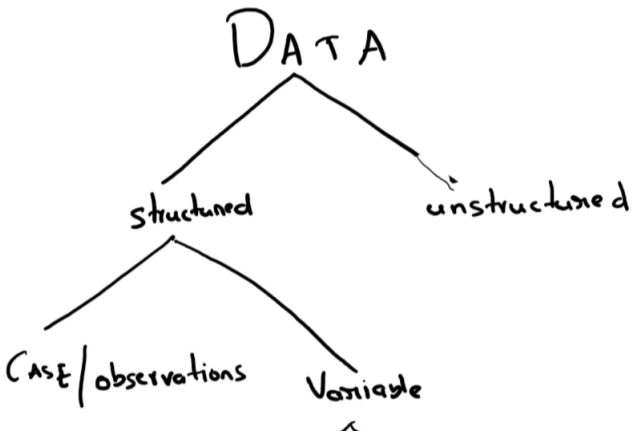
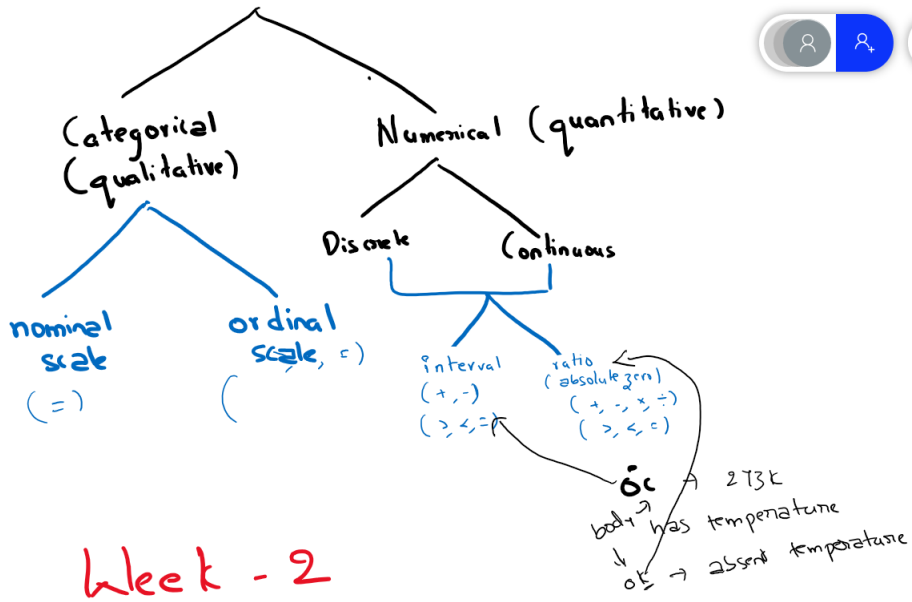
Statistics for Data Science - 1

Notes for end term





Scales of measurement



Week - 2

Week - 2

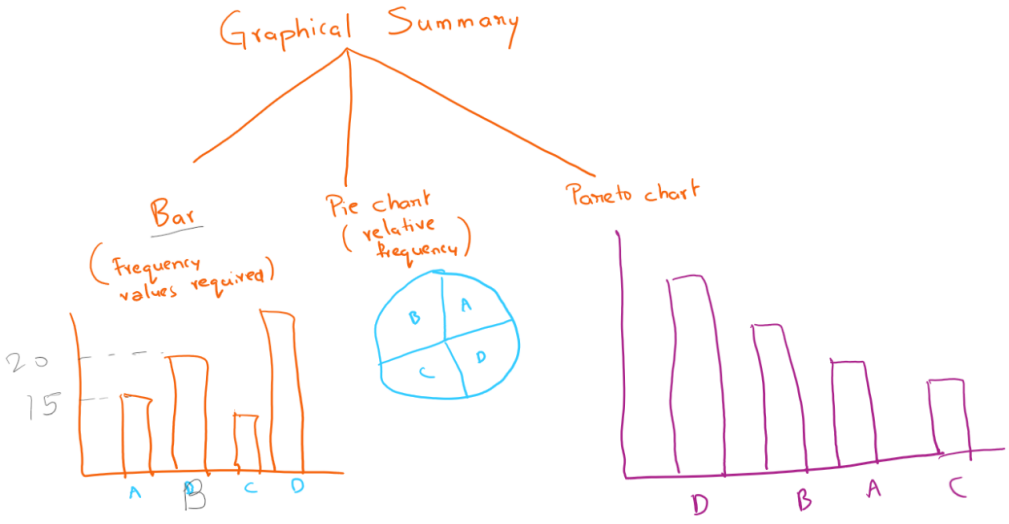
CATEGORICAL VARIABLE

Frequency?

Relative frequency

Frequency table

Categories	Frequency	relative Frequency
A	15	15/60
B	20	20/60
C	15	15/60
D	10	10/60
	60	
Total frequency	50	



Describing categorical data

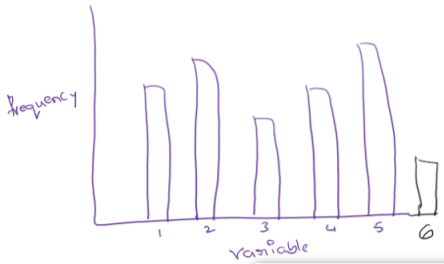
Mode
(ordinal, nominal)
highest frequency category

Median
(ordinal)
Middle value in ordered data ascending or descending order

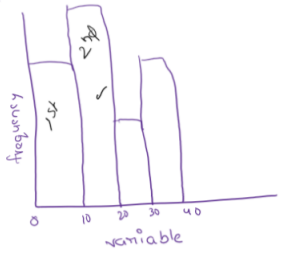
Week 3: Describing numerical variable

Graphical summary

Bar chart
(for discrete less numbered variable)



Histogram
(for larger data)
Data divided into classes

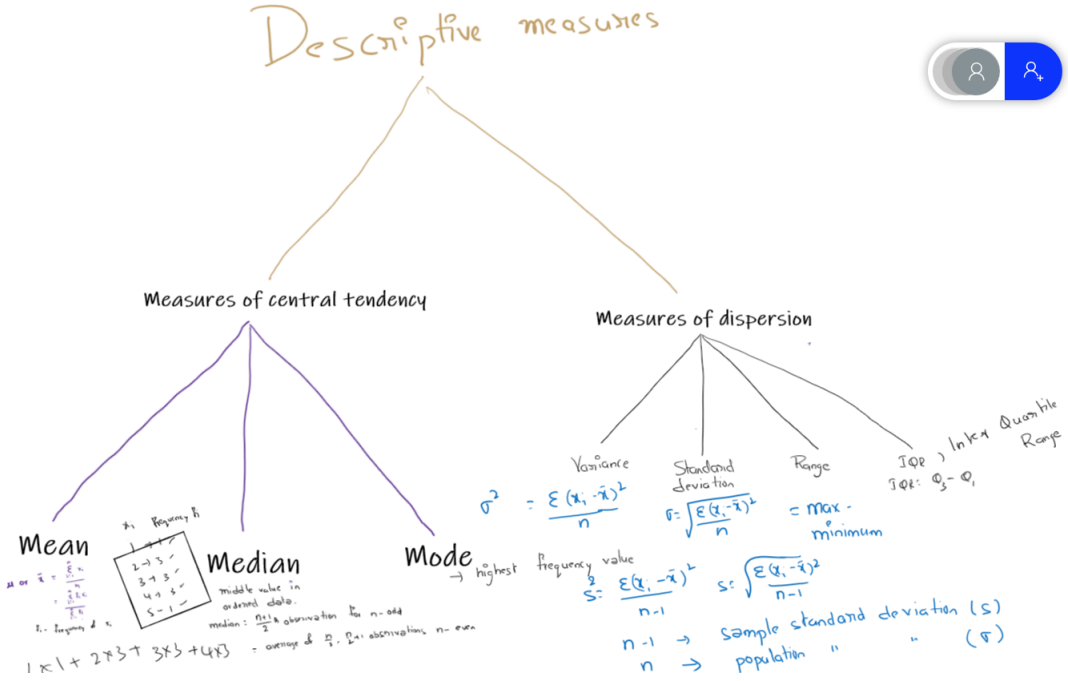


Stem and Leaf plot

Stem	leaf
1	2 9
2	1
3	5
4	1

[0, 10)
[10, 20)

Descriptive measures



Percentiles

p^{th} percentile is np^{th} observation if np is exact integer

" " " next highest integer & np observation if np is not exact integer } ordered ascending

n → total observations

$n \times p$

Q_1 - 25^{th} percentile value

Q_3 - 75^{th} percentile value

$Q_1 = n \times 0.25^{th}$ observation

$Q_3 = n \times 0.75^{th}$ observation

Interquartile range (IQR): $Q_3 - Q_1$

Outliers → extreme values

values less than $Q_1 - 1.5\,IQR$ and values greater than $Q_3 + 1.5\,IQR$

$-8 \quad 2 \quad 3.5 \quad 4.3 \quad 5.1 \quad 6.2 \quad 8.5 \quad 9.1 \quad 19$

↓ outlier

$n = 10$,

$np = 10 \times 0.25 = 2.5$

$Q_1 = 3^{rd}$ observation in ordered data $= 3.5$

$np = 10 \times 0.75 = 7.5$ next highest integer $= 8.5$

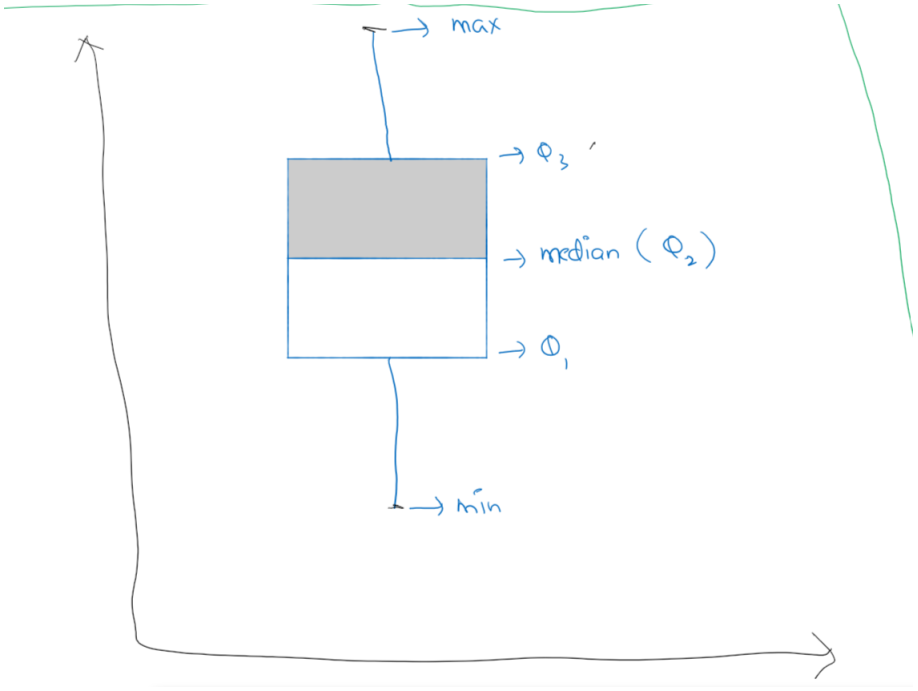
$Q_3 = 8^{th}$ observation $= 8.5$

$IQR = 8.5 - 3.5 = 5$

Outliers values $< Q_1 - 1.5\,IQR = 3.5 - 1.5(5) = -4$

$> Q_3 + 1.5\,IQR = 8.5 + 1.5(5) = 16$

$-8, 19$ are outliers



Week 4

Association between two variables

- two categorical → variate contingency table
- two numerical → scatter plot
- one categorical → two categories

↳ point - bi-serial correlation coefficient (r_{bp}): $\frac{(\bar{y}_0 - \bar{y}_1)\sqrt{p_0 p_1}}{s_x}$

variable 2	employed	unemployed	Total
female	100	50	150
Male	500	150	650
Total	600	200	800

response variable

explanatory variable

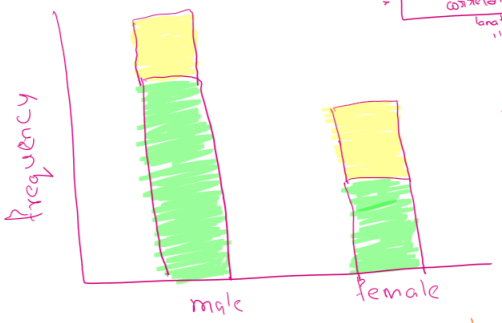
$$cov(x,y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{N} \quad \text{population}$$

$$cov(x,y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{N-1} \quad \leftarrow \text{sample}$$

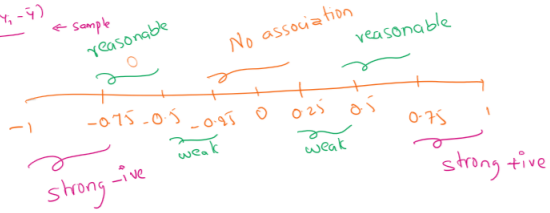
correlation coefficient (r): $\frac{cov(x,y)}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$

$-1 < r < 1$

r is same for sample, population



Stacked bar chart



Week 5 Permutations and combinations

Basic rules of counting

Addition rule

event A → n_1 ways

event B → n_2 ways

A or B in $n_1 + n_2$ ways

Multiplication rule

event A → n_1 ways

event B → n_2 ways

A & B in $n_1 \times n_2$ ways

A or B ...

There are $n_1, n_2, n_3, n_4 \ldots n_i$ in which i events occurs

Any of i events occurs in $n_1 + n_2 + n_3 + n_4 + \ldots n_i$ ways (Addition rule)

All of i events occurs can occur in $n_1 \times n_2 \times n_3 \times n_4 \ldots n_i$ different ways. (Multiplication rule)

Factorials

$n!_0 = n \times (n-1) \times (n-2) \times \ldots \times 2 \times 1$ ways

$n!_0 = n \times (n-1)!_0$, $(n+1)!_0 = (n+1)\, n!_0$

$0!_0 = 1$

... objects = ${}^nP_r = \frac{n!_0}{}$

ROOM

All of ...

Factorials

$n!_0 = n \times (n-1) \times (n-2) \times \ldots \times 2 \times 1$ ways

$n!_0 = n \times (n-1)!_0$, $(n+1)!_0 = (n+1)\, n!_0$

$0!_0 = 1$

Number of possible permutations 'r' objects from 'n' distinct objects = ${}^nP_r = \frac{n!_0}{(n-r)!_0}$

Number of possible permutations 'n' objects from 'n' distinct objects = ${}^nP_n = n!_0$

repetition not allowed

Number of possible permutations of n objects of n_1 objects are of 1 kind: $\frac{n!_0}{n_1!_0}$

Number of possible permutations 'r' objects from 'n' distinct objects = n^r → repetition allowed.

Number of possible permutations 'r' ... 'n' distinct objects in circle: $\frac{{}^nP_r}{r}$ ← repetition

ROOM → 4 letters

Number of objects

Number of possible permutations 'r' from 'n' distinct objects $= n^r$ → repetition allowed.

" " " " 'r' " " 'n' distinct objects in circle: $\frac{{}^nP_r}{r}$ ← repetion not allowed.

Number of possible circular permutation of n distinct objects: $\frac{{}^nP_n}{n} = \frac{n!}{n} = (n-1)$

COMBINATIONS:

Number of ways of choosing 'r' objects from 'n' objects is ${}^nC_r = \frac{n!}{(n-r)!\,r!}$ ${}^nC_r = {}^nC_{n-r}$

Probability

Random experiment

↓

All outcomes → sample space (S) $P(S) = 1$

(E event ⊂ sample space (S) $0 \le P(E) \le 1$

tossing a coin → All outcomes $\{H, \bar{T}\} \Rightarrow S = \{H, \bar{T}\}$

event E → getting head $E = \{H\}$ $E \subset S$.

Union of events: (∪)

A & B are two events.

probability that either A or B occurs is P(A∪B)

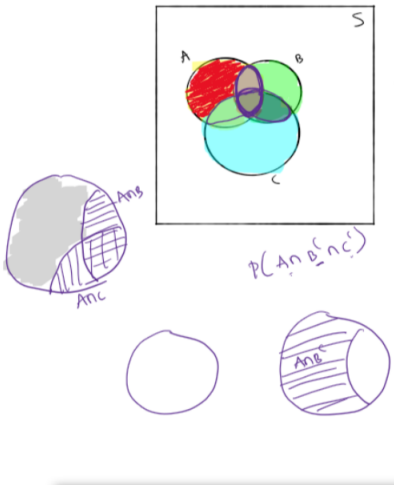
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$ ✓

$\Rightarrow P(A \cap B) = P(A) + P(B) - P(A \cup B)$ ←

null event → φ $P(\phi) = 0$

$\phi = \{\}$ empty set

A∪B



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

$P(\text{event } E) = \frac{n(E)}{n(S)} = \frac{1}{2} = 0.5$

$S = \{H, T\}$

Sign out

Week - 7

① Contingency table

		G (Employed)	H (Unemployed)	Row total
Male	E	a	b	a+b
Female	F	c	d	c+d
Column total		a+c	b+d	a+b+c+d = N

N = total no. of ways
a = E ∩ G
b = E ∩ H
c = F ∩ G
d = F ∩ H

Joint Probabilities

$P(E \cap G) = a/N$
$P(E \cap H) = b/N$
$P(F \cap G) = c/N$
$P(F \cap H) = d/N$

Marginal Probabilities

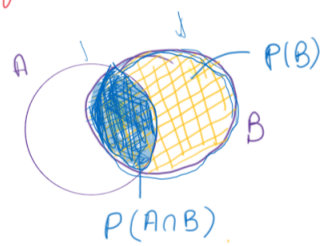
$P(E) = (a+b)/N$
$P(F) = (c+d)/N$
$P(G) = (a+c)/N$
$P(H) = (b+d)/N$

Conditional Probability

$P(E|G) = a/(a+c)$
$P(F|G) = c/(a+c)$
$P(G|E) = a/(a+b)$
$P(H|E) = b/(a+b)$
$P(H|F) = ?$ $d/(c+d)$
$P(G|F) = ?$

② Conditional Probability formula

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



$P(B) > 0$

$$P(A \cap B) = P(A|B)\,P(B)$$

↓

Multiplication rule

③ Mutually Exclusive & Independent Events

Mutually Exclusive ↓ disjoint

$A \cap B = \phi$

$\Rightarrow P(A \cap B) = 0$

Independent Events ↓

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

Let A & B be 2 events for a sample space S.

→ Mutually exclusive but not independent :

① Random exp. : Tossing a coin.

$S = \{H, T\}$ ✓

$P(A) = 1/2$ $A = \{H\}$, $B = \{T\}$ $1/2$

$A \cap B = \phi$

$\Rightarrow P(A \cap B) = 0 \Rightarrow$ A & B are mutually exclusive.

$P(A) = 1/2$, $P(B) = 1/2$, $P(A)\,P(B) = 1/4 \neq 0 = P(A \cap B)$

⇓

not independent

Remark:

If A & B are 2 mutually exclusive events for a sample space S, A & B will be independent if either $P(A) = 0$ or $P(B) = 0$.

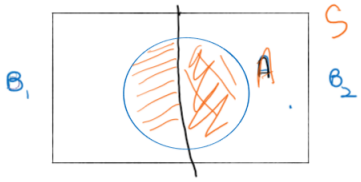
$P(A) \times P(B) = P(A \cap B)$

③ Law of Total Probability

- Mutually Exclusive
- Exhaustive events

① For 2 events

Let B_1 & B_2 be 2 mutually exclusive & exhaustive events for a sample space S.



$A = (A \cap B_1) \cup (A \cap B_2)$

↓ mutually exclusive events

For any event A,

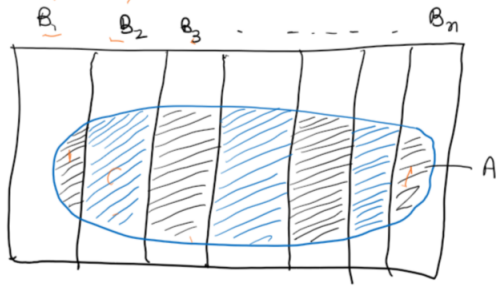
$P(A) = P(A \cap B_1) + P(A \cap B_2)$ (Using addition rule)

$\Rightarrow P(A) = P(A|B_1)\,P(B_1) + P(A|B_2)\,P(B_2)$ (Using Multiplication rule)

↓

∴ Law of Total Probability

② For n events



$$A = (A \cap B_1) \cup (A \cap B_2) \cup \dots \cup (A \cap B_n)$$

$$\Rightarrow P(A) = P[(A \cap B_1) + P(A \cap B_2) + \dots + P(A \cap B_n)]$$

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \dots + P(A|B_n)P(B_n)$$

④ Bayes' Theorem :

$$P(B_i|A) = \frac{P(B_i \cap A)}{P(A)}$$

$$= \frac{P(A|B_i)P(B_i)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}$$

Week 8

Random Variable (Real valued function on the sample space)

- Discrete (Takes at most countable number of values)
- Continuous (When outcomes of random event cannot be counted)

eg. Experiment :- Rolling a fair dic

let X be a discrete random Variable which takes value greater than 3.

$S = \{1, 2, 3, 4, 5, 6\}$

X takes values 4, 5, 6.

Sign out

Probability mass function (p.m.f.)

Let X be a discrete random variable taking n possible values.

$x_1, x_2, x_3, \ldots\ldots, x_n$

p.m.f., $p(x)$ of X is given by $\boxed{p(x_i) = P(X = x_i)}$

X	x_1	x_2	------	x_n
$P(X=x)$	$P(x_1)$	$P(x_2)$	------	$P(x_n)$

$P(X = x_i) = 0 \rightarrow$ for other values of x

Properties of p.m.f :-

1. $p(x_i) \geq 0$ for $i = 1, 2, \ldots, n$
2. $\sum_{i=1}^{n} P(x_i) = 1$

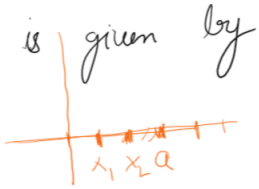
Graph of p.m.f. :

1. Negative skewed
2. Positive skewed
3. Symmetric
4. Constant

Cumulative distribution function (CDF)

CDF, F, of a random variable X is given by

$$F(a) = P(X \leq a)$$



Let X → discrete random variable taking values $x_1, x_2, x_3, \ldots$

$x_1 < x_2 < x_3 < \ldots$

→ CDF, F of X is a step function.

Expectation & Variance of a Random Variable

Expectation

$$E(X) = \sum_{i=1}^{\infty} x_i \, P(X = x_i)$$

→ Also called "long-run-average" value

↓

in repeated independent observations

Variance

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

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

① $Var(cX) = c^2 \, Var(X)$

$Var(X + c) = Var(X)$

$Var(aX + b) = a^2 Var(X)$,

a, b, c are constants

⟹ Expectation of $g(x)$,

$$E[g(x)] = \sum_{x=1}^{\infty} g(x_i) P(X = x_i)$$

$\Rightarrow E[ax+b] = a E(x) + b$ where a, b are constants

$\Rightarrow E[b] = b$

→ X, Y are 2 random variables

$E(X+Y) = E(X) + E(Y)$

X = discrete random variable

$g(X) = 3X + 2$

$= X(X-1)$

→ $Var(X+Y) = Var(X) + Var(Y)$

X, Y are independent Random variables $Var(cX) = c^2$

$Var(X+Y) = V(X-Y) = V(X+(-Y))$

$= V(X) + Var(-Y)$

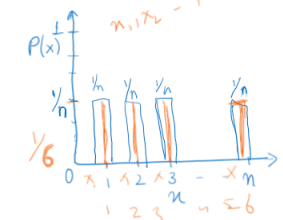
$= V(X) + (Y)$

Discrete Random Variables

Discrete Random Variables

Sign out

1. Uniform
2. Bernoulli
3. Binomial
4. Hypergeometric



$E(X) = \sum_{i=1}^{n} x_i p(x_i)$

$Var(X) = ? \; E(X-\mu)^2$

$X = \begin{cases} p, & X = 1 \\ 1-p, & X = 0 \end{cases}$

$E(X) = p$

$Var(X) = p(1-p)$

pq

→ # trials = 1

→ models numbers of successes in n i.i.d Bernoulli trials

→ n fixed

p — same for all the trials

~ all should be independent.

~ Each trial — Bernoulli

→ without replacement

N ← population size

m — Red ball

$N-m$ — green balls

$P(X = i) = \dfrac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}}$