→ Course — Statistics 67.69 0.7236

① $30

+ 50% $\frac{50}{50}$;

80%(80) + 20%(20) = 68

31.176

31.25

31.01

Data → (Statistics) → Decisions

2000

2000

→ 31.176

→ 30.9523

~~→ 30.909~~

→ 31.0521

→ 31.538

→ 31.25

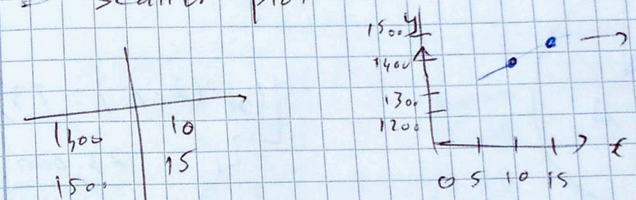
11,00ft² — 88,000

1 ft² — $\frac{88,000}{11,000}$ → 80 $

So:

1500ft² — (80)(1500) $ 83

⇒ Scatter plot:



→ Linear

1.176 x 1700 = 2000

0.9523 x 2100 = 2000

⋮

1.538 x 1300 = 2000

* Linear → that if something is ↑ then it keeps ↑ & rate needs not to be constant.

Size	Price
1700	53,000
2100	65,000
1900	59,000
1300	41,000
1600	50,000

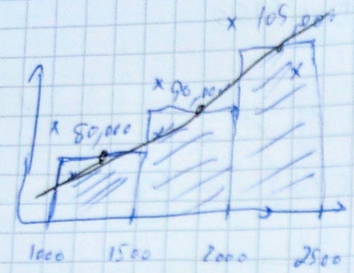
Price = $\boxed{30}$ $/ft² . size

+ $\boxed{2000}$ $

Size	$	
1300	88 000	67.69
1400	72 000	51.43
1600	94 000	58.75
1900	86 000	45.26
2100	112 000	53.33
[illegible]	[illegible]	42.6

S. 22,000 → 105000$ but not trusty result as change is non Linear.

Size	$
1200	88,000
1400	72 K
1600	94 K
1900	86 K
2100	112 K
2300	90 K



* Even though distribution is non linear, we can still use bar graph
* It helps to see global trend for which scatter plot is not useful.

histogram → 1D

bar ~~chart~~ graph → 2D

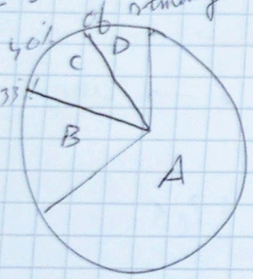
⇒ Good Question

(7/12) A — 175

(2/12) B — 50

(1/12) C — 25

(2/12) D — 50



$(x)(20\%) = 23{,}000$

$x = \frac{23{,}000}{20\%}$

$x = 115{,}000$

And

$B = x - A$

$144{,}000$

$\frac{2}{12}$

$\frac{50}{125}$

$\frac{1}{12}$

5

0.

75

⇒ Simpson's Paradox.

	Male			Female		
	Applied	Admitted	Rate	Applied	Admitted	Rate
to A	900	450	50%	100	80	80%
to B	100	10	10%	900	180	20%
	1000	460	46%	1000	260	26%

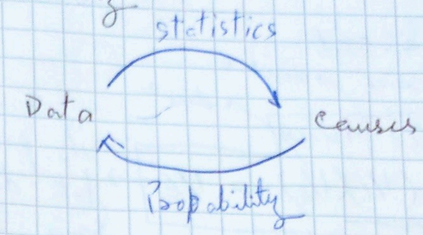
So Males are favored.

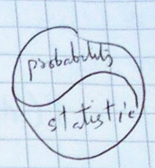
"I don't believe in statistics I didn't doctor myself"

Ch

ctr → click thru rate → analytics

→ probability





(2)

Loaded Coin → where pHead ≠ pTail

→ {Heads, Heads} $p(H) = 0.5$

$P(H,H) = 0.25$ → * prob that both time we'll get H

$\frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} = \frac{2!}{2!(2-2)!} (0.5)^2 (0.5)^0$

$= \boxed{0.25}$

* So 2nd flip depends upon first.
* $(50\%)(50\%) = 25\%$

~~2H, H~~

flip 1	flip 2	
H	H	0.36
H	T	0.24
T	H	0.24
T	T	0.16
H	H	0.25
H	T	0.25
T	H	0.25
T	T	0.25

where $p(H) = 0.6$ (loaded Coin)

$\frac{n!}{(n-k)!\,k!} p^k (1-p)^{n-k}$

$p(H) = 0.5$

→ exactly 1 is head

$\frac{3!}{1!2!} 0.6(0.4)^2$

0.288

Load Coin; H = 0.6

0.6×0.4×0.4

0.4×0.6×0.4

0.4×0.4×0.6

0.288

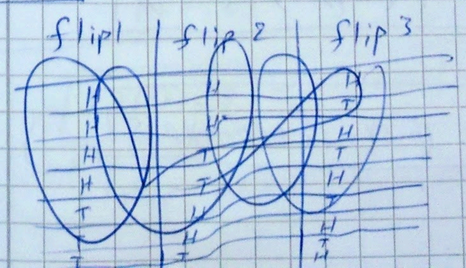
* So Truth tables can be used when you encounter

$\frac{3!}{1!(2!)} (0.6)^1 (0.4)^2$

$(3)(0.6)(0.4)^2$

$\boxed{0.288}$

* at least
* exactly
* at most



flip1	flip2	flip3
H	H	H
H	H	T
H	T	H
H	T	T
T	H	H
T	H	T
T	T	H
T	T	T

exactly 1's Head

fair Coin

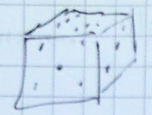
0.5×0.5×0.5

0.5×0.5×0.5

0.5×0.5×0.5

So

$P = 3(0.5)^3 = 0.3$



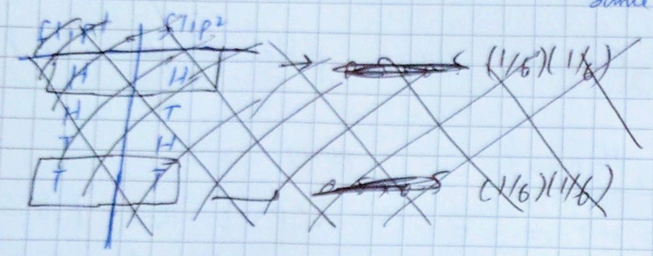
0.096
0.096
0.096

P(die → even) = 50%

# of choice → dice 1 → dice 2

⇒ Throw a fair dice twice

$P(\text{double}) = 6(1/6)(1/6) = 1/6$ → its prob. that both dice will have same value



⇒ Composite event = P.P.P... → independence

→ Conditional probability

P(positive | cancer) → probability of positive given cancer

P(cancer | positive) → probability of cancer given positive

Cancer	Test	P(Pos | Cancer)
Y	P	(0.1)(0.9)
Y	N	(0.1)(0.1)
N	P	(0.9)(0.2)
N	N	(0.9)(0.8)
		$\Sigma P = 1$

y → yes ; P → positive

n → no ; n → negative

P(Cancer) = 0.1

P(>Cancer) = 0.9

P(Pos | Cancer) = 0.9

P(Neg | Cancer) = 0.1

P(Pos | >Cancer) = 0.2

P(Neg | >Cancer) = 0.8

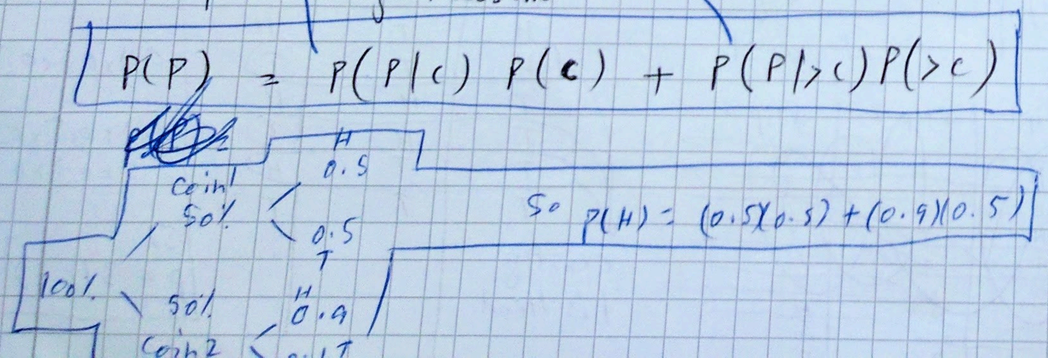
~~P(Pos) = P(Pos | Cancer) P(Cancer)~~

And

P(Positive) = (0.1)(0.9) + (0.9)(0.2)

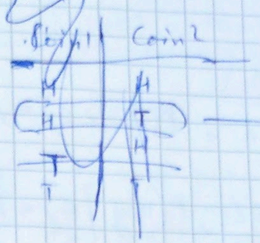
⇒ Total probability Theorem

$$P(P) = P(P|c)\,P(c) + P(P|>c)\,P(>c)$$



So P(H) = (0.5)(0.5) + (0.9)(0.5)

P(T|H)



→ (0.5)(0.5)(0.1)

P(1) = 0.5

P(2) = 0.1

③

P(H | T)

~~0.5 = P(P|c)(0.5) + (1 − P(P|c))(0.5)~~

~~0.5 = 0.5 P(P|c) + 0.5~~

Coin1 ; Coin2 New Flip , flip

P(H | T)

Sol.

Pick	flip1	flip2
1	H	H
1	H	T
1	T	H
1	T	T
2	H	H
2	H	T
2	T	H
2	T	T

chance of picking coin1 ; Since coin is fair

→ (0.5)[(0.5)(0.5)] → 0.125 (P(H), P(T))

chance of picking coin2

→ (0.5)[(0.7)(0.1)] → 0.045 (P(H), P(T))

Loaded Coin

So:

P(H|T) = 0.125 + 0.045 = 0.17

Problem Statement:

Coin1 is fair, coin2 is load (H→90%, T→10%). Find Prob. that coin2 is T given coin is H ⇒ P(H|T)

Note: only 1 type of coin is used at a time

Now Consider Coin1 (H→1, T→0) ; Coin2 (H→0.6, T→0.4)

find P(T,T)

Sol.

P(1,T,T) = (0.5)(0)(0) → Coin1

P(2,T,T) = (0.5)(0.4)(0.4) → Coin2

P(T,T) = 0.08

$$P(\text{Test}) = P(\text{test}|\text{disease})\,P(\text{disease}) + P(\text{test}|>\text{disease})\,P(>\text{disease})$$

Total probability theorem

⇒ Bayes Rule

↳ Thomas Bayes

$P(c) = 1\%$; $P(Pos|c) = 90\%$; $P(Pos|>c) = 10\%$

$P(>c) = 99\%$ $\quad P(neg|>c) = 90\%$; $P(neg|c) = 10\%$

$P(c|Pos) = ?$

Sol.

$$\boxed{P(c|Pos) = \frac{P(Pos|c)\,P(c)}{P(Pos|c)P(c) + P(Pos|>c)\,P(>c)}}$$

prior distribution

posterior distribution

$P(c|Pos) = \frac{(90\%)(1\%)}{(90\%)(1\%) + (10\%)(99\%)}$ $\quad$ $P(c, Pos)$ $\quad$ $P(>c, Pos)$ $\quad$ $P(Pos)$

$$\boxed{P(c|Pos) = 8\%}$$

test is positive if cancer → sensitivity

test is negative if no cancer → specificity

prior probability * evidence → posterior probability

$$\cancel{P(c) = \frac{P(c|Pos)\,P(Pos|c)\,P(c)}{P(Pos|x)} + P(c|Pos)P(Pos|>c)P(>c)}$$

$$\boxed{P(Pos) = P(c, Pos) + P(>c, Pos)}$$

~~P(Pos) =~~

⇒ joint

$P(c, Pos) = P(Pos, c)\,P(c)$

$P(>c, Pos) = P(Pos, >c)\,P(>c)$

⇒ posterior

$$\boxed{P(c|Pos) = \frac{P(Pos, c)\,P(c)}{P(Pos, c)P(c) + P(Pos, >c)\,P(>c)}}$$

$P(c|Neg) = i$

$P(c|Neg) = \frac{P(Neg|c}{P(c, Neg}$

$P(c|Neg) = \frac{P(Neg|c}{P(Neg|c)P(}$

OR

$$\boxed{P(c|Pos) = \frac{P(c, Pos)}{P(c, Pos) + P(>c, Pos)}}$$

where $P(c, Pos) + P(>c, Pos)$ ↳ $P(Pos)$

So:

$\,) \rightarrow$ non-normalized

$| \rightarrow$ normalized

$P(c, Pos) = P(Pos|c)\,P(c)$

$P(>c, Pos) = P(Pos|>c)\,P(>c)$

$P(Pos) = P(c, Pos) + P(>c, Pos)$

And:

$$\boxed{P(c|Neg) = \frac{P(Neg|c)\,P(c)}{P(Neg|c)P(c) + P(Neg|>c)\,P(>c)}}$$

OR $\boxed{\frac{P(c, Neg)}{P(c, Neg) + P(>c, Neg)}}$

Question

$P(c) = 0.001$; $P(Pos|c) = 0.9 \rightarrow$ sensitivity ; $P(Neg|c) = 0.1$

$P(>c) = 0.99$ $\quad P(Neg|>c) = 0.9 \rightarrow$ specificity ; $P(Pos|>c) = 0.1$

find:

$P(c, Neg) = P(Neg|c)\,P(c) = (0.1)(0.001) = 0.0001$

$P(>c, Neg) = P(Neg|>c)\,P(>c) = (0.9)(0.99) = 0.891$

And:

$$\boxed{P(Neg) = P(c, Neg) + P(>c, Neg) = 0.892}$$

So:

$$P(c|Neg) = \frac{P(Neg|c)\,P(c)}{\underbrace{P(Neg|c)\,P(c) + P(Neg|>c)\,P(>c)}_{P(Neg)}} = \frac{0.001}{0.001 + 0.891} = 0.00112$$

$$P(>c|Neg) = \frac{P(Neg|>c)\,P(>c)}{P(Neg)} = \frac{0.891}{0.892} = 0.998$$

$P(c) = 0.1$ $\quad P(>c) = 0.9$

$P(Pos|c) = 0.9$ $\quad P(Neg|c) = 0.1$

$P(Neg|>c) = 0.5$ $\quad P(Pos|>c) = 0.5$

* $P(Neg|c) = 1 - P(Pos|c)$
* $P(Pos|>c) = 1 - P(Neg|>c)$

And:

$P(c, Neg) = P(Neg|c)\,P(c) = 0.01$

$P(>c, Neg) = P(Neg|>c)\,P(>c) = 0.45$

joint

$P(Neg) = P(c, Neg) + P(>c, Neg)$

$P(Neg) = 0.46$

posterior

$P(c|Neg) = \frac{P(c, Neg)}{P(Neg)} = 0.0217$

$P(>c|Neg) = \frac{P(>c, Neg)}{P(Neg)} = 0.9783$

And

$P(c|Pos) = \frac{P(Pos|c)\,P(c)}{P(Pos|c)\,P(c) + P(Pos|>c)\,P(>c)} = \boxed{0.1667}$

$P(>c|Pos) = \frac{P(Pos|>c)\,P(>c)}{P(Pos)} = \boxed{0.833}$

→ Bayes Rule Summary:

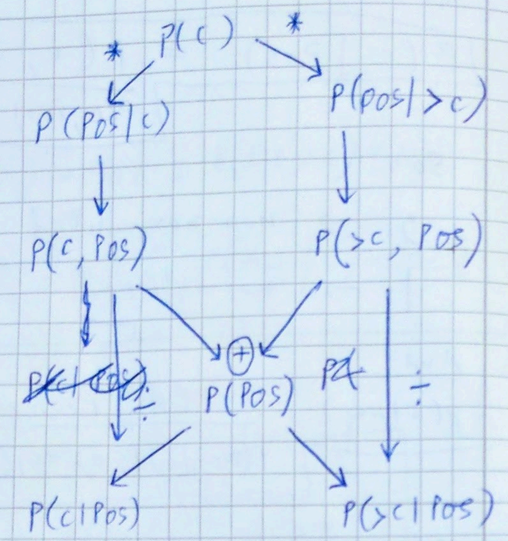
Cancer (c) prior $P(c)$

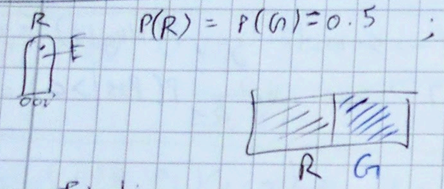
$1 - P(Neg|c)$

$P(Pos|c)$

Test $P(Pos|>c)$

↳ $1 - P(Neg|>c)$





$P(R) = P(G) = 0.5$; $P(\text{see } R | \text{ at } R) = 0.8$; 0.8

$P(\text{see } G | \text{ at } G) = 0.8$; 0.5

$P(\text{see } R | \text{ at } R) = 0.2$; 0.2

$P(\text{see } R | \text{ at } G) = 0.2$; 0.5

Find:

$P(\text{at } R | \text{see } R) = ?$; $P(\text{at } G | \text{see } G) = ?$

So:

$$P(\text{at } R | \text{see } R) = \frac{P(\text{see } R | \text{ at } R)\, P(R)}{P(\text{see } R | \text{ at } R)\, P(R) + P(\text{see } R | >\text{at } R)\, P(>R)}$$

$$= \frac{(0.8)(0.5)}{(0.8)(0.5) + (0.2)(0.5)} = 80\%$$

And:

$$P(\text{at } G | \text{see } G) = \frac{P(\text{see } G | \text{ at } G)\, P(G)}{P(\text{see } G | \text{ at } G)\, P(G) + P(\text{see } G | >\text{at } G)\, P(>G)}$$

$$= \frac{(0.8)(0.5)}{(0.8 \times 0.5) + (0.2 \times 0.5)} = 0.8$$

And:

$$P(\text{at } G | \text{see } R) = \frac{P(\text{see } R | \text{ at } G)\, P(G)}{P(\text{see } R | \text{at } G)\, P(G) + P(\text{see } R | \text{ at } R)\, P(R)} \rightarrow P(>$$

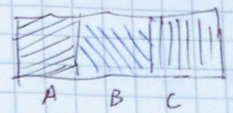
$$= \frac{(0.2 \times 0.5)}{(0.2 \times 0.5) + (0.8 \times 0.5)} = 20\%$$

$$P(A|R) = \frac{P(A,R)}{P(R)} = \frac{P(A,R)}{P(A,R) + P(>A,R)} = \frac{P(A,R)}{P(A,R) + P(B,R) + P(C,R)}$$

$P(A,R) = P(R,A) = P(R|A)\,P(A)$

Question

$P(A) = P(B) = P(C) = 1/3 = 0.333$



$P(R|A) = 0.9 \rightarrow P(G|A) = 0.1$

$P(G|B) = 0.9 \rightarrow R(R|B) = 0.1$

$P(G|C) = 0.9 \rightarrow P(R|C) = 0.1$

$P(A|R) = \cancel{P(R|A)\,R(A)}$ $\rightarrow 1 - (P(R|B) + P(R|C))$

$\cancel{P(R|A)\,R(A) + P(R|>A)\,P(>A)}$ $\rightarrow 1 - P(A)$

$$P(A|R) = \frac{(0.9)(0.333)}{\cancel{(0.9)(0.333) + (}} = \frac{0.2997}{\cancel{0.2997 + (0.8)(0.667)}}$$

$$P(B|R) = \frac{(0.1)(0.333)}{\cancel{(0.1)(0.333)}} = 0.0333$$

$$P(C|R) = (0.1)(0.333) = 0.0333$$

Now, here:

$P(R) = P(A,R) + P(B,R) + P(C,R)$

$= (0.2997) + 2(0.0333)$

$P(R) = 0.3666$

This method can't be used now to find $P(R)$

Now,

$P(A|R) = \frac{0.2997}{0.3666} = \boxed{0.8175}$

$P(B|R) = \frac{0.0333}{0.3666} = \boxed{0.0908}$

$P(C|R) = \frac{0.0333}{0.3666} = \boxed{0.0908}$

$\cancel{P(R|A)\,P(R|B)\,P(R|C)}$

$\cancel{P(R|>A)}$

$P(R|B)P(B) + P(R|>B)P(>B)$

$(0.1)(0.333) + (1)(0.667)$

0.7003

$P(R|C)P(C) + P(R|>C)P(>C)$

$(0.1 \times 0.333) + (1)(0.667)$

0.7003

$R(R|A)R(A) + P(R|>A)R(>A)$

$(0.9)(0.333) + (0.2)(0.667)$

0.4331

$P(\text{home}, \text{rain}) = 0.004$

$P(>\text{home}, \text{rain}) =$

$P(\text{rain}|>\text{home})\,P(>\text{home})$

$(0.3)(0.6)$

Ques. $P(\text{gone}) = 0.6$ $P(\text{rain}|\text{home}) = 0.01 \rightarrow P(>\text{rain}|\text{home}) = 0.9$

$P(\text{home}) = 0.4$ $P(\text{rain}|\text{gone}) = 0.3 \rightarrow P(>\text{rain}|\text{gone}) = 0.$

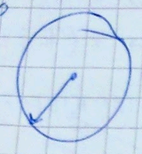
$P(\text{home}|\text{rain}) = ?$

$$P(\text{home}|\text{rain}) = \frac{P(\text{rain}|\text{home})\,P(\text{home})}{P(\text{rain}|\text{home})\,P(\text{home}) + P(\text{rain}|>\text{home})\,P(>\text{home})}$$

$$= \frac{(0.01)(0.4)}{(0.01)(0.4) + (0.3)(0.6)} = \boxed{0.0217} \Rightarrow 2.17\%$$

→ Probability ~~Distribution~~ Density (Continuous):

0 ... 360



⇒ probability that object has $\theta = 180°$

$P_{180°} = 0$

same is true for any angle

$P_{179°} = 0$; P

⇒ This is because here space is continuous unlike coin OR dice

> In continuous space, every outcome has 0 probability

* But we may define probability of range. So:

$P(0 \leq x \leq 180°) = 50\%$

$P(260° < x \leq 290°) = \frac{290 - 260}{360} = 8.33\%$

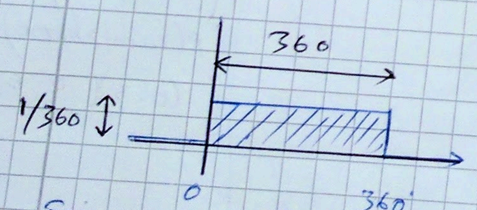
$P(0 < x \leq 1°) = \frac{1}{360} = 0.277\%$

But

$P(0)$ or $P(1)$ ... $= 0$ → Exact probability is always 0 as its continuous case

→ Probability Density:

*



① - Each outcome has equal value

② - Area integrates to 1

So: Area $= \left(\frac{1}{360}\right)(360) = 1.0$

> * In continuous Distribution
> → $p(x)$ → probability → for range
> → $f(x)$ → density → for specific instance

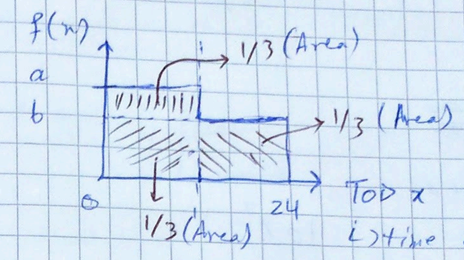
* x = seconds → 0 ... 60
* $p(x) = 0$ → probability for continuous is 0.
* $f(x) = 1/60$ → density is calc. so that Area = 1

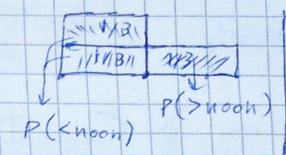
⇒ Question:

$f(x \leq noon) = 2 f(x > noon)$ → density

↳ density being born before Noon

↳ density being born after Noon





$b = \left(\frac{1}{3}\right)\left(\frac{1}{12}\right) = 0.02777$

$a = \left(\frac{2}{3}\right)\left(\frac{1}{12}\right) = 0.05555$

As:

$\boxed{\text{height} = \frac{\text{Area}}{\text{width}}}$

$\frac{1}{24} \rightarrow 0.04166$

$\frac{1}{48} \rightarrow 0.02083$

$\frac{2}{6} \rightarrow \frac{1}{6}$

$\rightarrow \frac{1}{12}$

0.08333

0.04166

12(0.08333)

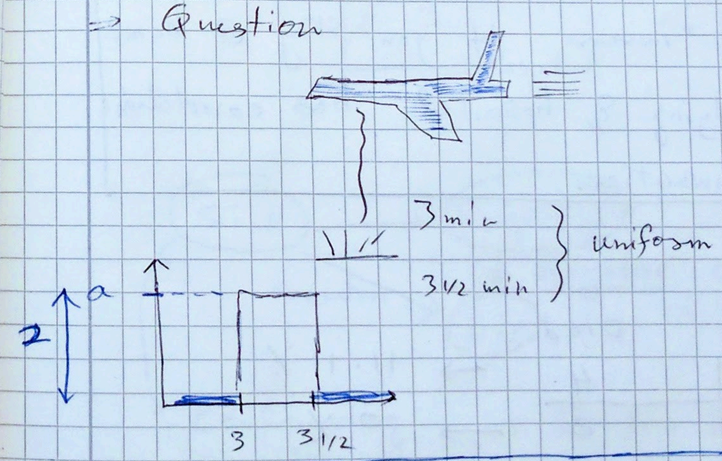
12(0.04166)

1/6

$\left(\frac{1}{6}\right)(12)$

$\frac{0.75}{12}$; $\frac{0.25}{12}$

⇒ Question



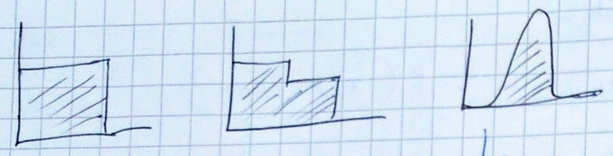
$(0.0083)(0.5)$

$(2)(0.5)$

> Density could be > 1 as long as $A \leq 1$

⇒ Properties of density:

* Non-negative (can be zero)
* needs not to be continuous
* density can be > 1 but $0 \leq$ Area ≤ 1



↳ Uniform density ↳ Step density ↳ guassian density

⇒ Correlation & Causation

hospital	Died	
40	4	10%

home	Died	
8000	20	0.25%

$\frac{10}{0.25}$

So chances of dying in hospital is 40X higher ~~than~~ home

> *even though chance of die in hospital is 40X but this doesn't mean your chances of not dying will increase if you stay @ home
> * because dying & hospital has correlation but not causation.

Now

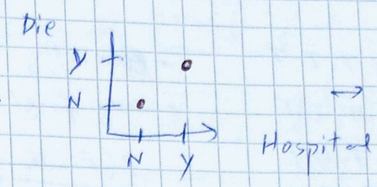
Hospital	Died	
Sick 36	4	→ 11.11%
Healthy 4	0	→ 0%

Home	Died	
Sick 40	20	→ 50%
Healthy 7960	20	→ 0.25%

* it shows since majority of people in hospital come are already sick & in home majority will be healthy.

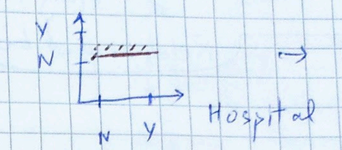
⇒ Correlation ⇒ knowledge of 1 variable gives knowledge of other variable

Case-1 ⇒



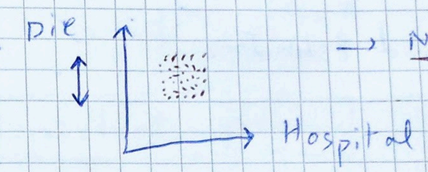
→ Correlation

Case-2 ⇒



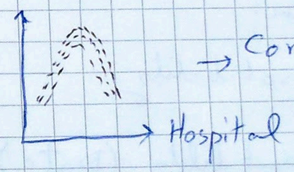
→ Non-correlated

Case-3 ⇒

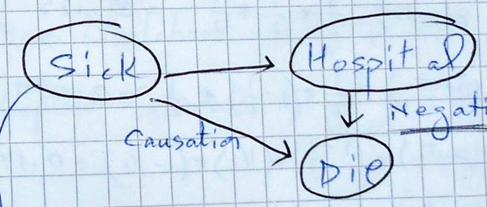


→ Non-correlated
* B/c no matter I go to hospital or not, the die chances are ~~chance~~ same
* So y value is not changing w.r.t x axis.

Case-4 ⇒



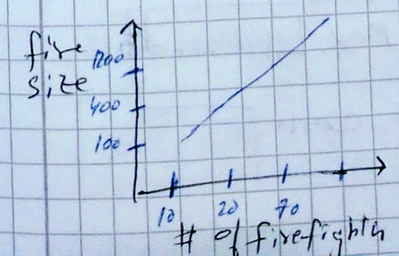
→ Correlated
* B/c y changes w.r.t x



i.e fire is causing more fighters & not other way around.

Confounding Variable → if you don't take into a/c the confounding var, the correlation is then deceiving

> just from data, we can't tell is it relation or causation



Here fire size & # of firefighters are ~~correlated~~ corelated but firefighters are not causing fire but its reverse causation

fire → fighters → reverse causation

⇒ Probability Problem Set:

useless

Flip	Coin
1	H
1	T
2	H
2	T

HTTT → exactly 1 is H = 4

THTT - only first is H = 1

TTHT

TTTH

$\frac{P(\text{Exactly } 1\ H)}{P(\text{only first is } H)} = \frac{4}{1}$

Relative probability

⇒ Estimators:

⇒ Maximum Likelihood Estimator:

$$\underbrace{\frac{1}{N}\sum x_i}_{0 \ldots 1}$$

Estimation problem:

Data → probability

probability → probability (Data)

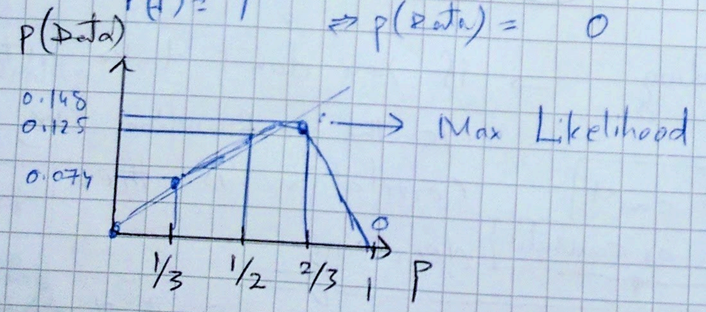
1 0 1

$P(1) = 1/2 \Rightarrow P(\text{Data}) = \frac{1}{2} * \frac{1}{2} * \frac{1}{2} = 0.125$

$P(1) = 1/3 \Rightarrow P(\text{Data}) = \frac{1}{2} * (1 - \frac{1}{2}) * \frac{1}{2} = 0.074$

$P(1) = 2/3 \Rightarrow P(\text{Data}) = (1 - \frac{1}{2}) * (1/2) * (1 - 1/2) = 0.148$

$P(1) = 1 \Rightarrow P(\text{Data}) = 0$



From ~~single~~ coin flips, Max. Likelihood always assume a Loaded coin

* So add fake Data.

Data → Data + Heads + Tails

1	$\frac{1}{1}$	$\frac{1+1}{1+2}$ → w/ fake data
001	$\frac{1}{3}$	$\frac{1+1}{3+2}$
1010	$\frac{2}{4}$	$\frac{2+1}{4+2}$
11	$\frac{2}{2}$	$\frac{2+1}{2+2}$

↳ estimation — Smooth.

tends to ~~converge~~ pull towards 0.5

* MLE after first H assumes all are going to be H

⇒ Dirichlet priors

* Laplacian Estimator → "Method of adding fake data to estimation"

* when we have plenty of data then Laplace ≃ MLE

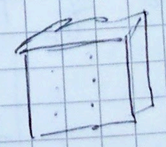
- for sparse data, Laplace is better.

* Max. Likelihood $= \frac{1}{N}(\sum x_i)$

* Laplace Estimator $= \frac{1}{N+K}(1 + \sum x_i)$

N → # of experiments

K → # of outcomes

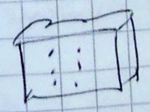


1 2 3 2

MLE $p(1) = \frac{1}{4}$

LAPLACE $p(1) = \frac{1+1}{4+6}$

$\frac{1}{6}$ → P of each



1232

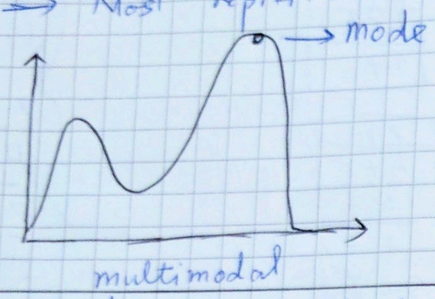
$p(2) = \frac{2}{4}$

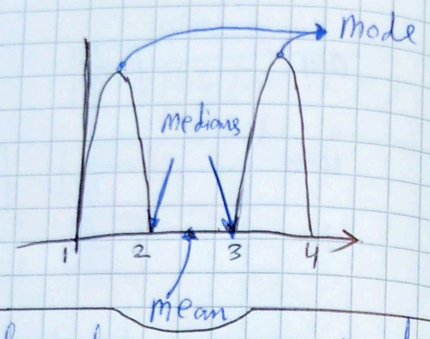
$p(2) = \frac{2+1}{4+6} = \frac{3}{10}$

Basically, just add probability in MLE → Lapla

33 4 4 4 32 32 33

- Mean → Average of Sum
- Median → Middle Value
- mode → Most repititive





* When we have multimodal, always use mode

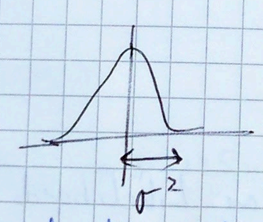
- Variance $(\sigma^2) = \frac{\sum (x - M)^2}{N}$
  - → Spread of a distribution
  - → quadratic deviation

1 2 3 4

$4(0.04) + \frac{\quad}{5}$

So 3 3 3 3 3
- → Mean → 3
- → Variance → 0
- → Deviation → 0

148.4



$$\sigma^2 = \frac{\sum (x_i - M)^2}{N}$$ → needs to calc. M

$$\sigma^2 = \frac{\sum (x_i^2)}{N} - \frac{(\sum x_i)^2}{N^2}$$ → no need of M

$(M)^2$

$\frac{\sum x_i^2}{N}$

V → Variance
S.D → Stand. Deviation
S.V → Stand. Value

⇒ Standarized Data:

$$\frac{\text{Value} - M}{\sigma}$$

⇒ Outliers

① – Quartiles

median

3 3, 4, 5, 6, 6, 7, 7, 8, 8 9 ignored

interquartile

total = $4n + 3$

quartile

2 here

$4n+3$
$4(2)+3 \rightarrow 11$

$4(n)+3$
$4n+3$
$(2)^{th}$

$\left(\frac{n+1}{2}\right)^{th}$

-99 13 17 33 1489

⇒ Binomial Distribution

2 Coins flip — # Head = # Tail

(2)

H	H
H	T
T	H
T	T

3 Coins flip — # Head = # Tail

0

Head ≠ # Tail for odd # of coin flip

⇒ Exactly 1 H for 5 coins using (n)
↳ # of flips

5
- 2 coins — 2
- 3 coins — 3
- ⋮
- 5 coins — 5

⇒ Exactly 2 H for 5 coins

$\frac{5.4}{2}$ H H

↳ To remove redundancy

So:

10 possible

1
H H H
1 H H T
2 H T H
1 H T T
3 T H H
2 T H T
3 T T H
T T T

$(2)^n$

(2)

2 coins → 2, 4
3 coins → 6
4 coins → 8
5 → 8

$(n-1)^k$

$(n)^{k-1}$

$(1)^{2-1}$

$(1)^{k-1}$

H

$\frac{4.3}{2.1}$

$(2)^{n-1}$

$(2)^0 \rightarrow 1$

(2)

H H H H
H H H T
H H T H
H H T T
H T H H
H T H T
H T T H
H T T T
T H H H
T H H T
T H T H
T H T T
T T H H
T T H T
T T T H
T T T T

H

$\frac{3.2}{\quad} \rightarrow 3$ poss

⇒ 5 coin flips — 3 Heads

$\frac{5 \cdot 4 \cdot 3}{3 \cdot 2 \cdot 1}$ H H H

H H H T T
H H T H T
H T T H H } 3

T H H H T T
T T H H H
T H T H H
T H H T H
T H H

So: w/ binomial

5 Coins flips — 1 head.

$\frac{5}{1!} = 5 \rightarrow \boxed{\frac{5!}{1! \cdot 4!}}$

5 Coins flips — 2 head

$\frac{5 \cdot 4}{2!} = 10 \rightarrow \boxed{\frac{5!}{2! \times 3!}}$

5 coin flips — 3 head

$\frac{5 \cdot 4 \cdot 3}{3!} = 10 \rightarrow \boxed{\frac{5!}{(3!)(5-3)!}}$

5 coin flips — 5 head

$\frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{5!} \Rightarrow 1 \rightarrow \boxed{\frac{5!}{5!(5-5)!}}$

10 coins flips — 5 Head

$\frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6}{5!} \Rightarrow 252$

T T H H H		1
T H T H H		2
T H H T H		3
T H H H T		4
H T H H H		
H H T		
H		
H		
H		

$\frac{5}{1!} \times$

factorial:

$n! = n(n-1)(n-2) \dots 1$ — $n-(n-1)$

$\frac{3 \cdot 2}{2!} = \frac{6}{2} = 3!$

$\frac{5 \cdot 4 \cdot 3 \cdot \cancel{2 \cdot 1}}{(3 \cdot 2 \cdot 1)(\cancel{2!})}$

$\frac{5!}{5!(5-5)!}$

⇒ Binomial Distribution:

$\boxed{\frac{n!}{(k!)(n-k)!}}$

formula for Arrangements

125 coins — 3 Head

$\frac{125 \cdot 124 \cdot 123}{3!}$

optimized

⇒ Question:

Flip coin 5 time And:

find $P(\#\text{Heads} = 1) = ?$

Sol.

+ find Arrangement:

$\text{Arrangement} = \frac{n!}{k!(n-k)!} = \frac{5!}{1!(4!)}$

$\text{Arrangement} = \frac{5 \cdot \cancel{4!}}{1! \cdot \cancel{4!}} = 5$

Now

$\boxed{\text{Probability} = (\text{Arrangement})(p)^{\text{coinflips}}}$

Probability ⇒ $P(\#\text{Heads} = 1) = (5)(0.5)^5$

$\boxed{P(\#\text{Heads} = 1) = 0.15625}$

$(0.8)(0.2) + (0.8)(0.2)$

(Arrangements) $\times 0.8$

$p^{\text{coins}} = 0.75$

$(\text{Arr})\left[(pH)^1 \cdot (pT)^{\text{flips} - 1}\right]$

$\frac{2}{1!} = 2$

$2[0.5 \times 0.5]$

$(\text{Arrangements})(p)^{\text{coins}}$

$\frac{(2)(0.5)^2 = 0.2}{(5)(0.5)^5 = 0.15625}$

→ Only valid for Fair Coin.

(10)

H H
H T
T H
T T

(0.5×0.5)
(0.5×0.5)

$P = 0.5$

where

$1/p^{\text{coinflips}} \rightarrow$ Size of Truth table

So

$\boxed{\text{Probability} = \frac{\text{Arrangement}}{\text{Size of table}} = \frac{\frac{n!}{k!(n-k)!}}{1/(p)^{\text{flips}}}}$

$\frac{n!}{k!(n-k)!} \times (p)^{\text{flips}}$

or

$\boxed{\text{Probability} = (\text{Arrangement})(\text{outcomes})^{\text{flips}}}$

↳ 2 for coins

Questions:

Flip coin 5 times

find $P(\#\text{ of Heads} = 3) = ?$

Sol.

$\text{Arrangements} = \frac{5!}{3!(2!)} = \frac{5 \cdot 4 \cdot 3 \cdot \cancel{2!}}{3!(\cancel{2!})} = \frac{5 \cdot 4 \cdot 3}{3!} = 10$

And

$P(\text{Heads} = 3) = (\text{Arrangements})(\text{outcomes})^{\text{flips}} = (10)(1/2)^5$

Questions:

Flip Coin 3 times, $P(\#\text{Head} = 1) = ?$

where $P(\text{Head}) = 0.8$

Sol.

$$\text{Arrangements} = \frac{n!}{k!(n-k)!} = \frac{3!}{1!(2!)} = \frac{3 \cdot 2!}{1 \cdot 2!} = 3$$

Now:

$$\boxed{P(\text{Head}=1) = (\text{Arrangements})(pH)^{H}(pT)^{F-H}}$$

where
$H \rightarrow$ # of Heads
$F \rightarrow$ # of chances or flips

So

$$P(\text{Head}=1) = (3)\left[(0.8)^1(0.2)^{3-1}\right] = 0.096$$

$$\boxed{P(\text{Head}=1)_{H=0.8} = 0.096}$$

So

$$\boxed{P(\text{Head}=1) = \left(\frac{n!}{k!(n-k)!}\right)(pH)^{H}(pT)^{F-H}}$$ or generalized $$\boxed{P(H=k) = \frac{n!}{k!(n-k)!}(p^k)}$$

Question:

Flips $= F = 5$; $P(\#\text{Heads} = 4) = ?$

$pH = 0.8$

Sol.

$$P(H=4) = \left(\frac{5!}{4! \cdot 1!}\right)(0.8)^4(0.2)^{5-4}$$

$$\boxed{P(H=4) = 0.4096}$$

Question:

$F = 5$; $pH = 0.8$; $P(H=3) = ?$

$$P(H=3) = \frac{5!}{3!\,2!}(0.8)^3(0.2)^2 = \boxed{0.2048}$$

Question:

$F = 12$; $pH = 0.8$

$P(H=9) = ?$

Sol.

$$P(H=9) = \frac{12!}{9!\,3!}(0.8)^9(0.2$$

$$P(H=9) = 0.236$$

$P(H=1)$

Arr[(pH)

(3)[(0.8)

H T

(0.8)(0.2)

0.0

$(pH)^H(pT)^F$

$(0.8)^1(0.2)$

Final formula:

$$\boxed{P(k) = \frac{n!}{k!(n-k)!}(p)^k(1-p)^{n-k}}$$

↳ binomial distribution probability

$$\boxed{\text{Arrangements} = \frac{n!}{k!(n-k)!}}$$

↳ binomial arrangement

⇒ Central Limit Theorem

Coin: (0, 1)

$$P(\Sigma x_i = k) = \frac{n!}{(n-k)!(k)!}$$

where $K = (2)$ choices (Here) 0, 1

n↓ K→	0	1	2	3	4
n=1	1	1			
n=2	1	2	1		
n=3	1	3	3	1	
n=4	1	4	6	4	1

→ occurance

$1!/(0)!(1-0)!$

$2!/0!(2-0)!$ $2!/1!(2-1)!$ $2!/2!(2-2)!$

$3!/(0\cdots3)!(3-0\cdots3)!$

$4!/(0\cdots4)!(4-0\cdots4)!$

found using Binomial Distribution w/o probabilities, only arrangements.

→ Pascal Triangle

more than bi

Coin: (0, 1, 2) → binomial can't be used now

	0	1	2	3	4	5	6	7	8
n=1	1	1	1						
n=2	1	2	3	2	1				
n=3	1	3	6	7	6	3	1		
n=4	1	4	10	16	19	16	10	4	1

formulas:

# of Horiz. value = $2^n + 1$

Horiz. value = add 3 values of above row (choices)

~~For n = 1: (1!)/(0...8)!(1 - 0...8)!~~

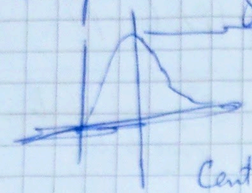
In both cases [Coin(0,1) or Coin(0,1,2)]

⇒ ~~(2) choices~~

* $n = (2)^{\text{choices}}$
* $K = (2)^{\text{choices}} + 1$
* # of Horiz. values $= 2n + 1$
* Horiz. value = add (# of choice) value of above row.

⇒ Summary of Central Limit Theorem:

* for large # of samples say $n > 30$

$\sum_{i=1}^{n} X_i \longrightarrow$  → central value is max Limit

Central Limit Theorem

* In case of binary choices (0, 1) OR (H, T), Binomial ~~Theory~~ Distribution can be used.

* for very large choices, Gaussian Distribution can be used.

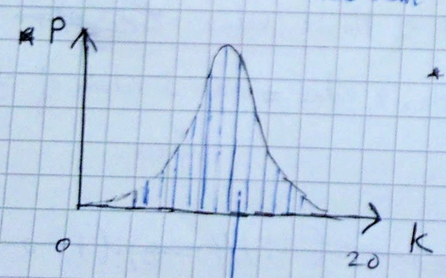
with:

* Binomial Probability Distro. $= \frac{n!}{(n-k)!\,k!} (p)^k (1-p)^{n-k}$

* normal Gaussian Probability Distro. $= \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2}\frac{(x-M)^2}{\sigma^2}}$

⇒ Normal Distribution

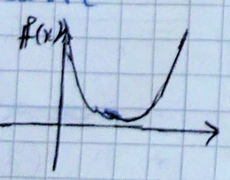
Binomial distr. → central Limit theorem → normal distribution



* max. value of distr. is at $k/2$ so here $20/2$.

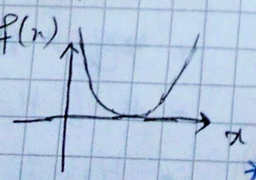
→ bell curve

$f(x) = (x-M)^2$



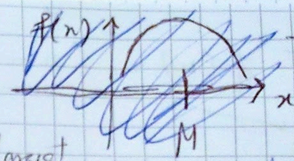
* B/c M is max. value from Central Limit Theorem
* so $f(x)$ is max when $x = 0$ & $f(x) = 0$ when $x = M$

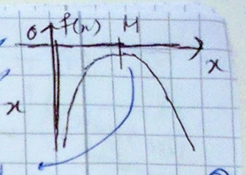
$f(x) \propto \frac{1}{\sigma^2}$ ← $f(x) = \frac{(x-M)^2}{\sigma^2}$



* Here width is controlled by σ^2. Width $\propto \sigma^2$
* High σ^2 Low Confidence

* $f(x) = \frac{-1}{2}\left[\frac{(x-M)^2}{\sigma^2}\right]$





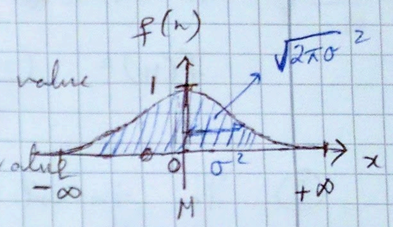
max value = 0

* $f(x) = e^{\frac{-1}{2}\left(\frac{(x-M)^2}{\sigma^2}\right)}$ → its largest when $x = M$

So

$f(M) = 1$ b/c $e^0 = 1$ → max value

$f(-\infty/+\infty) = 0$ b/c $e^{-\infty} = 0$ → min. value



$e^{\infty} \neq 0$; only $e^{-\infty} = 0$

* But in above formula we can use $-\infty$ & $+\infty$ both b/c $(x-M)^2$ so its square × $\frac{-1}{2}$ is always −ve.

* Since area under the curve in above exponential formula is $\sqrt{2\pi\sigma^2}$, we need to normalize it.

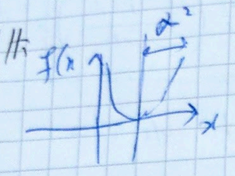
$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2}\frac{(x-M)^2}{\sigma^2}}$

① $f(x) = (x - M)^2$

Like Standard Deviation

② To Control the width

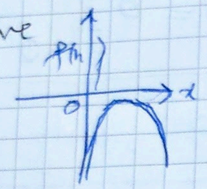
$f(x) = \frac{(x-M)^2}{\sigma^2}$



$f(x) \propto \frac{1}{\sigma^2}$

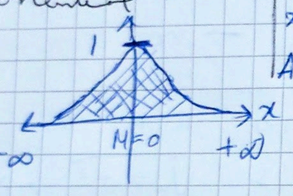
③ To invert the Curve

$f(x) = \frac{-1}{2} \frac{(x-M)^2}{\sigma^2}$

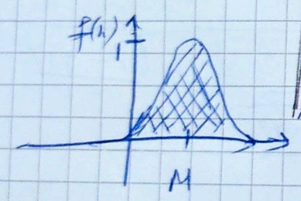


④ Introducing exponential

$f(x) = e^{\frac{-1}{2}\frac{(x-M)^2}{\sigma^2}}$



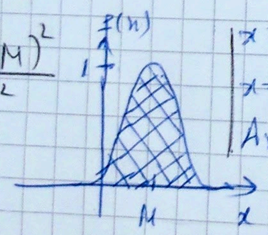
$x = 0;\ f(x) = 1$
$x = \pm\infty;\ f(x) = 0$
$\text{Area} = \sqrt{2\pi\sigma^2}$



$x = M;\ f(x) = 1$
$x = \pm\infty;\ f(x) = 0$
$\text{Area} = \sqrt{2\pi\sigma^2}$

⑤ To normalize it i.e Area ⇒ 1

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



$x = M;\ f(x) = 1$
$x = \pm\infty;\ f(x) = 0$
$\text{Area} = 1$

~~Points of~~ Binomial distribution

Binary possibility	A: p	B: $\frac{n!}{(n-k)!k!} p^k (1-p)^{n-k}$	C: $\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2}\frac{(x-M)^2}{\sigma^2}}$
Single	✓		
Few		✓	
Many			✓

→ normal distribution or Guassian

A → Single Coin flip, B → Many Coins flip, C → Many flips

⇒ Manipulating Normals:

Question:

$M = 60{,}000$; $\sigma = 10{,}000$

raise ⇒ 10,000

new $M = 70{,}000$, new $\sigma = 10{,}000$

$\exp^{\frac{-1}{2}\frac{(x-M)^2}{\sigma^2}} \Rightarrow \exp^{\frac{-1}{2}\frac{(x-60{,}000)^2}{(10{,}000)^2}}$

Now:

$x' = x + 10{,}000$ or $x = x' - 10{,}000$

So:

$\exp^{\frac{-1}{2}\frac{(x'-10000-60{,}000)^2}{(10{,}000)^2}} \Rightarrow \boxed{\exp^{\frac{-1}{2}\frac{(x'-\overset{M'}{70{,}000})^2}{\underbrace{(10{,}000)^2}_{\sigma'}}}}$

Question:

$M = 70{,}000$, $\sigma = 10{,}000$

raise ⇒ double.

$x' = 2(x) \Rightarrow x = \frac{1}{2}(x')$

$\exp^{\frac{-1}{2}\frac{(\frac{1}{2}x' - 70{,}000)^2}{\sigma^2}} \Rightarrow \exp^{\frac{-1}{2}\left[\frac{1}{2^2}\frac{(x'-140{,}000)^2}{\sigma^2}\right]}$

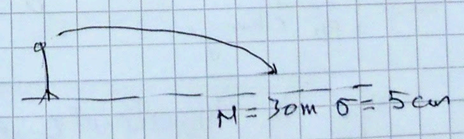
$\boxed{\exp^{\frac{-1}{2}\left[\frac{(x'-140{,}000)^2}{(2\sigma)^2}\right]}}$

So:

new $M = 140{,}000$

new $\sigma = 20{,}000$

Question #:



Improvements:
1- 10% 2- 2m

find new x & σ:

Sol.

⇒ First Improvement:

$x' = 1.1x$

So:

$e^{\frac{-1}{2}\frac{(\frac{x'}{1.1} - 30)^2}{\sigma^2}}$

$e^{\frac{-1}{2}\frac{((x'-33)^2)}{\sigma^2}}$

So:

$M = 33m$

$\sigma = 5.5m$

⇒ 2nd Improvement

$x' = 2 + x$

So:

$e^{\frac{-1}{2}\frac{(x'-2-33)^2}{\sigma^2}}$

So:

$\boxed{M = 35m \quad \sigma = 5.5m}$

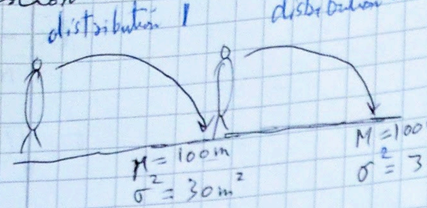
Results after both improvements

$\frac{\sum x_i}{N}$

$\left(\frac{(x_i - M)^2}{N}\right)$

$\sigma = \sqrt{\frac{(x_i - M)^2}{N}}$

Question:



Sol.
These are 2 different guassions, when we combine 2 guass, the results add up. So,

$\mu = 200m$; $\sigma^2 = 60\,m^2$

Note: St. dev. (σ) won't add up. σ would increase by $\sqrt{2}$ not by 2 facts

$\sigma = \sqrt{2}\left(\frac{\sigma_1 + \sigma_2}{2}\right)$

$\sigma = \sqrt{30}$

And

$\boxed{\text{new } \sigma = \sqrt{2}(\sqrt{30}) = 7.74\ m}$

Question:

$N(\mu, \sigma^2) \rightarrow A$.

change: $aA + b$; new μ & σ = ?

Sol.

- new $\mu = a\mu + b$
- new $\sigma^2 = a^2\sigma^2$
- new $\sigma = a\sigma$

* So any change in distribution reflects as it is in median
* In σ^2, multiplicative changes affects by $(\ )^2$ of change but add/subtract doesn't change σ^2 or σ
* for σ, multiplicative changes affect by ( ) of change

Question:

$A \rightarrow N(\mu, \sigma^2)$; $B \rightarrow N(\mu, \sigma^2)$

$A + B \rightarrow$?

Sol.

$\mu_{A+B} = 2\mu$; $\sigma^2_{A+B} = 2\sigma^2$

$\sigma_{A+B} = \sqrt{2}(\sigma)$

Question:

$A \rightarrow N(\mu, \sigma^2)$; $B \rightarrow N(\mu, \sigma^2)$

$A - B \rightarrow$?

Sol.

$\mu_{A-B} = 0$; $\sigma^2_{A-B} = 2\sigma^2$

; $\sigma_{A-B} = \sqrt{2}(\sigma)$

σ^2 is uncertainty and either we add $(A+B)$ or subt. $(A-B)$ two guass, the σ^2 will always add up.

formula to add σ:

$\sigma = \sqrt{2}\left(\frac{\sigma_1 + \sigma_2}{2}\right)$

⇒ Summary:

1 - $(aX + b)$ → $a\mu + b$, $a^2\sigma^2$, $a\sigma$ → change in distribution
(μ, σ^2)

2 - $(X + Y)$ → $\underbrace{\mu + \mu}_{\text{new mean}}$, $\underbrace{\sigma^2 + \sigma^2}_{\text{new } \sigma^2}$, $\underbrace{\sqrt{2}\left(\frac{\sigma + \sigma}{2}\right)}_{\text{new } \sigma}$ → two dists. added
($\mu\sigma^2$, $\mu\sigma^2$)

3 - $(X - Y)$ → $\underbrace{\mu - \mu}_{\text{new mean}}$, $\underbrace{\sigma^2 + \sigma^2}_{\text{new } \sigma^2}$, $\underbrace{\sqrt{2}\left(\frac{\sigma + \sigma}{2}\right)}_{\text{new } \sigma}$ → two dists. subtracted.
($\mu\sigma^2$, $\mu\sigma^2$)

* It's possible that 99% people are smarter than Avg.
* So statement:

Most people are smarter than Avg

is true

⇒ Question. (find Quartile boundaries)

100 87 | 82 78 | 74 72 | 61 58

itse $4n+3$ can't be used as the series is even.

So 25%, 50%, 75%, 100%

⇒ Question **(good Quest.)

100,000 → total users

1,000 → active users

find $P(\geq 2$ active users$)$ for a - 1 sample, b - 10 sample

Sol.

For 1 sample:

$P(1 \text{ active}) = \frac{1{,}000}{100{,}000} = 0.01$

$P(0 \text{ active}) = 1 - 0.01 = 0.99$

So:

$P(\geq 2 \text{ active}) = 1 - 0.01 - 0.99$

$\boxed{P(\geq 2 \text{ active}) = 0}$

for 10 samples

$P(0 \text{ active}) = (0.99)^{10} = 0.904$

$P(1 \text{ active}) = \frac{10!}{1!\,9!}(0.01)(0.99)^9$

$P(1 \text{ active}) = 0.09135$

So: $P(\geq 2 \text{ active}) = 1 - 0.904 - 0.09135$

$\boxed{P(\geq 2 \text{ active}) = 0.00465}$

→ Confidence Intervals: (CI)

* Instead of using exact value of confidence, we use interval w/ tolerance.
So instead of 60%, we say 57% to 63%.

* CI is different from σ^2.

* Sample
$x_1 \dots x_n$
μ
CI

$\left(\frac{\Sigma x_i}{N}\right)$ or mean

Votes

Install 1

A, B → (A=11, B=9) $P(A) = \frac{11}{11+9} = 0.55$

→ 2 (A=55, B=45) $P(A) = 0.55$

→ 3 (A=1100, B=900) $P(A) = 0.55$

Even though they all predict the same result. But Install 3 has more data points hence reliable.

* $\boxed{\text{Size of CI} \propto \frac{1}{\text{Sample Size}}}$
  - → the more samples, the more confidence
  - → But standard Deviation would remain same.
  - → So, more samples doesn't change the deviation.

$\boxed{CI \propto \frac{\sigma}{\sqrt{N}}}$ → depends (shrink or expand) based on sample size.

→ Question:

* flip a fair coin. What μ & σ = ?

Sol.

Since $p(H) = 0.5$; $p(T) = 0.5$

So

$\mu = \boxed{0.5} \rightarrow 1(0.5) + 0(0.5)$ → $p(T)$

$\sigma^2 = \frac{\Sigma(x_i - \mu)^2}{N-1} = \frac{(0.5)(0-0.5)^2 + 1(1-0.5)^2}{1} = \boxed{0.25}$

$p(H)$ (0.5), (0.5)

for 2 samples:

new $\Sigma\mu = 2(0.5) = \boxed{1}$ → 0.5 + 0.5

new $\Sigma\sigma^2 = 2(0.25) = \boxed{0.5}$

* Means add up.

for 10 samples:

new $\Sigma\mu = 10(0.5) = \boxed{5}$

new $\Sigma\sigma^2 = 10(0.25) = \boxed{2.5}$ → 10 distributions are added.

$p = 0.5$ $\mu = 0.5$ $\sigma^2 = 0.25$

	Mean(Σx_i)	VAR(Σx_i)	VAR($\frac{1}{N}\Sigma x_i$)
N=1	0.5	0.25	0.25
N=2	1	0.5	0.125 → $\frac{0.5}{2^2}$
N=10	5	2.5	0.025 → $\frac{2.5}{10^2}$
N=100	50	25	0.0025 → $0.025/100^2$

Mean of Sum — Variance of Sum — Variance of Average

we know $aX \xrightarrow{\text{var}} \sigma^2(a^2)$

* For SD, just take $\sqrt{\ }$ of value

quantile

	CI → S.D(1.96)
N=1	$(\sqrt{0.25})(1.96) \rightarrow 0.98$
N=2	$\sqrt{0.125}(1.96) \rightarrow 0.69$
N=10	$\sqrt{0.025}(1.96) \rightarrow 0.31$
N=100	$\sqrt{0.0025}(1.96) \rightarrow 0.098$

→ finding mean, S.D & CI for Coin.

Question:

$p = 0.1$ (loaded) ; $\boxed{M = p \;;\; \sigma^2 = p(1-p)}$ → Mean & Variance for Coin

$\boxed{CI = 1.96\sqrt{\frac{\sigma^2}{N}}}$ → Confidence Interval for Coin

Empirical ~~Statistical~~ ~~Analytical~~ formula

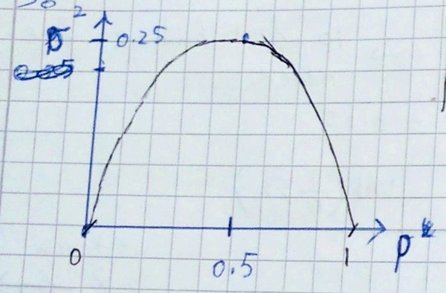
So for $p = 0 \Rightarrow M = 0 ; \sigma^2 = 0$

↳ Coin will come Tail all the time & the results won't vary.

for $p = 1 \Rightarrow M = 1 ; \sigma^2 = 0$

↳ Coin will come Head & the results won't vary.

So:



$\boxed{\text{* } \sigma^2 \text{ is max } (0.25) \text{ for fair coin.}}$

N	σ^2	σ^2/n	CI
1	0.09	$0.09/1^2$	$1.96\sqrt{0.09}$
10	0.9	$0.9/10^2$	$1.96\sqrt{0.9/10^2}$
100	9	$9/100^2$	$1.96\sqrt{9/100^2}$

So:

$p = 0.1 \;;\; \sigma^2 = p(1-p) = 0.09$

$CI = \frac{\sqrt{\sigma^2}}{\sqrt{n}}$

$CI = \frac{\sigma}{\sqrt{n}}$

N	CI
1	0.588
10	0.1859
100	0.0588

→ $1.96\sqrt{0.09/1}$

→ $1.96\sqrt{0.09/10}$

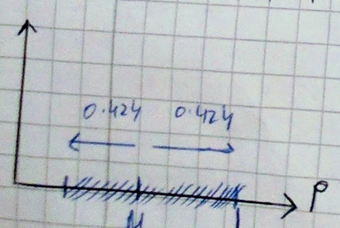
→ $1.96\sqrt{0.09/100}$

↳ CI shrinks with more samples i.e we have more confidence

↳ smaller boundary → less uncertainty

⇒ given a sample:

0 1 1 1



$M = \frac{0+1+1+1}{4} = \boxed{0.75}$ ← $\frac{\sum x_i}{n}$

$\sigma^2 = \frac{(0-0.75)^2 + 3(1-0.75)^2}{4} = \boxed{0.1875}$ ← $\frac{\sum (x_i - M)^2}{n}$ → empirical ~~Analytical~~ formula

$0.75 > 1.96\sqrt{\frac{\sigma^2}{N}}$ ← $CI = \boxed{0.424}$

Given Sample:

$\underbrace{000}_{T}\ \underbrace{1111111}_{H}$

$M = \frac{7}{10} = \boxed{0.7}$

$\sigma^2 = \frac{3(0-0.7)^2 + 7(1-0.7)^2}{10} = \boxed{0.21}$

$CI = 1.96\sqrt{\frac{0.21}{10}} = \boxed{0.284}$

But:

400 x Tails

600 x Heads.

$M = \frac{600}{1000} = \boxed{0.6}$

$\sigma^2 = \frac{400(0-0.6)^2 + 600(1-0.6)^2}{1000} = \boxed{0.24}$

$CI = 1.96\sqrt{\frac{0.24}{1000}} = \boxed{0.0304}$

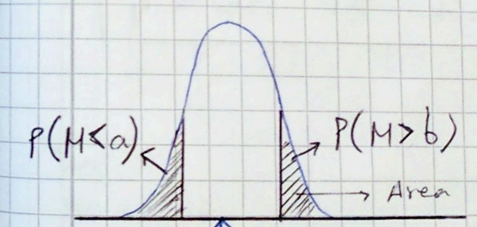
* with very large sample, we have very high confidence with narrow boundary.

* its b/c $1.96\sqrt{\frac{\sigma^2}{N}}$

⇒ Normal Quantiles

Confidence Interval = $\boxed{CI = \text{Mean} \pm 1.96\sqrt{\frac{\text{Variance}}{\text{Samples}}}}$

comes from Central Limit Theorem

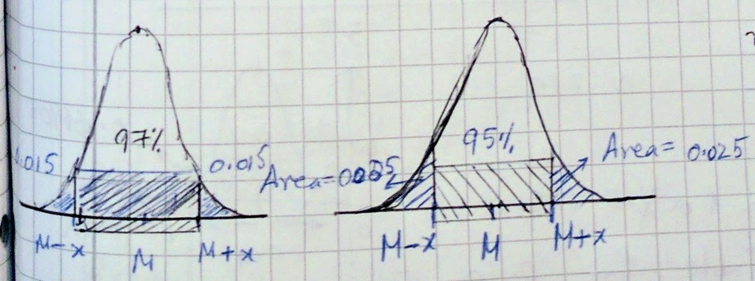


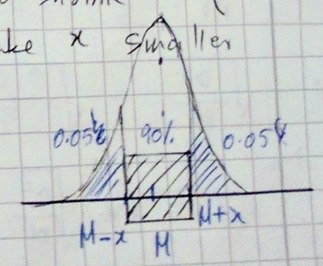
→ Area under the curve

Quantiles ← 1.96 → 95% Confidence Interval

P → N Samples → $M = \frac{1}{N}\sum X_i$

* To shrink CI (95% → 90%), make x smaller





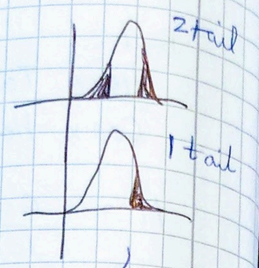
Higher x

Lower x

Central Limit

+/-	Quantiles
90%	~ 1.64
95%	~ 1.96
99%	~ 2.58

* These quantiles work when we have atleast 30 data samples.



* For fewer than 30 samples, we use t-distribution / table for statiscal estimates.

Used in Hypothesis

data samples = DOF + 1 ← T-table →

Degree of freedom (DOF)	~~1-0.99~~ 1-0.9	1-0.98	.....	.....
1				
2	quantiles values for 99% CI	quantiles values for 98% CI		
⋮				

→ Used to estimate normal distribution with < 30 samples.

* To see for 5 samples, Look the row = 4

** Question

estimate 90% sample 11000 → binary

find confidence interval = ?

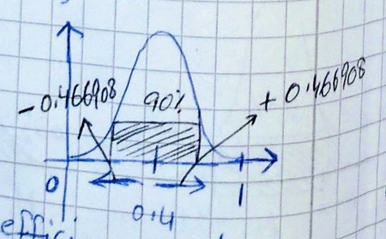
Sol.

$M = \frac{2}{5} = 0.4$; $\sigma^2 = \frac{2(1-0.4)^2 + 3(0-0.4)^2}{5} = 0.24$

$CI = \sqrt{\frac{\sigma^2}{N}} = \sqrt{\frac{0.24}{5}} = 0.219$

Now

$CI = (a)\sqrt{\frac{\sigma^2}{N}} = (a)(0.219)$



for a ⇒ see t-table for 0.10 efficiency in (5-1) row. So $a = 2.132$. Thus $CI = 0.466908$

$\boxed{0.4 \pm 0.466908}$ → confidence interval

⇒ Hypothesis Testing

Sample —statistics→ Information —statistics→ Decision

Null Hypothesis — H_0

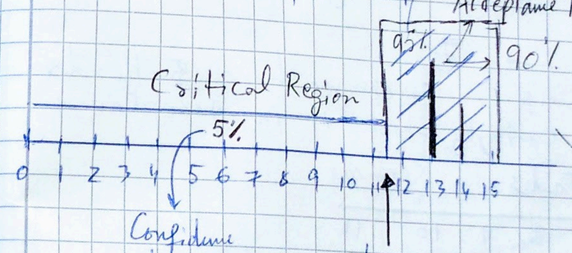
Alternate Hypothesis — H_1

Lose weight
Yes — 73%
No

So $H_0 \rightarrow p: 0.9$
$H_1 \rightarrow p: < 0.9$

yes 11 $H_0: p = 0.9$
No 4 $H_1: p < 0.9$

↳ Binomial
$N = 15$ $p = 0.9$



assume ⇒ H_0 is correct

doesn't work as advertised
$H_1 \neq 90\%$
doesn't work as well advertised
$H_1 < 90\%$
does work as advertised
$H_1 = 90\%$

low confidence on left
high confidence on right

* Critical region has many possible outcomes but in totality their probability ≤ 5%

$\frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$

Yes = 11
No = 4
$H_0: p = 0.9$

0	0
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	0.0002
9	0.001
10	0.01
11	0.04
12	0.12
13	0.27
14	0.34
15	0.20

1.12% < 5%

5% (Critical Region)

null hypothesis can't be in this region

95% Acceptance Region

→ $(15!/14!(1!))(0.9)^{14}(0.1)^1$

→ $\frac{15!}{15!(0!)}(0.9)^{15}(0.1)^0$

Question:

$H_0: p = 0.3$ → Claim: Head prob is 0.3

$H_1: p > 0.3$

So: $P(H) = 0.3$, $P(T) = 0.7$

* Also done using μ & σ^2 onwards

T H H T T T H T T H H → $p = \frac{5}{11} = 0.45$ → Mean $\mu = p$

probability of distribution

0	1	2	3	4	5	6	7	8	9	10	11
		0.12	0.25	0.22	0.13	0.06	0.02	0	0	0	0

most likely

smallest

0.02 0.09

→ outcome

→ probabilities of outcomes

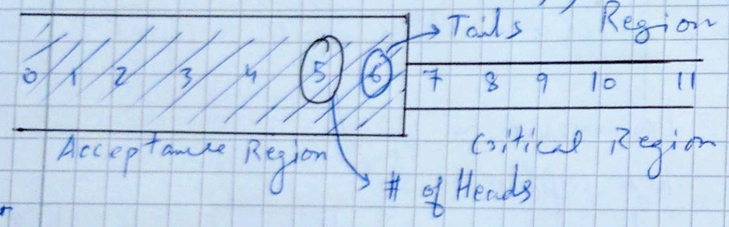
$\frac{n!}{k!(n-k)!}(p)^k(1-p)^{n-k}$

Select b/c

→ $H_1: > 0.3$

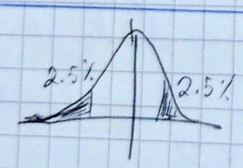
If $H_1: < 0.3$ we select 0

* Most likely outcome → 3 as its $p = 0.25$
* Critical Region (CR) → Right side of 3
* Smallest outcome in CR → 7
* Should we return the coin? since we have 5 Heads which is in Acceptance Region



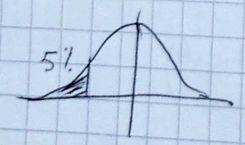
⇒ $H_0: p = 0.5$

Case-1: $H_1: p \neq 0.5$



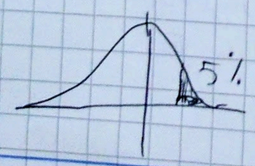
2 tailed

Case-2: $H_1: p < 0.5$



1 tail left

Case-3 $H_1: p > 0.5$



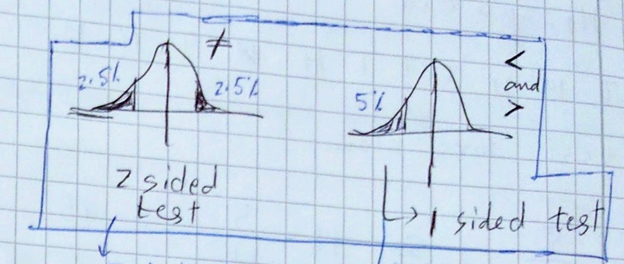
1 tail right

fair coin:

$H_0: p = 0.5$

$H_1: p \neq 0.5$ → either > 0.5 or < 0.5

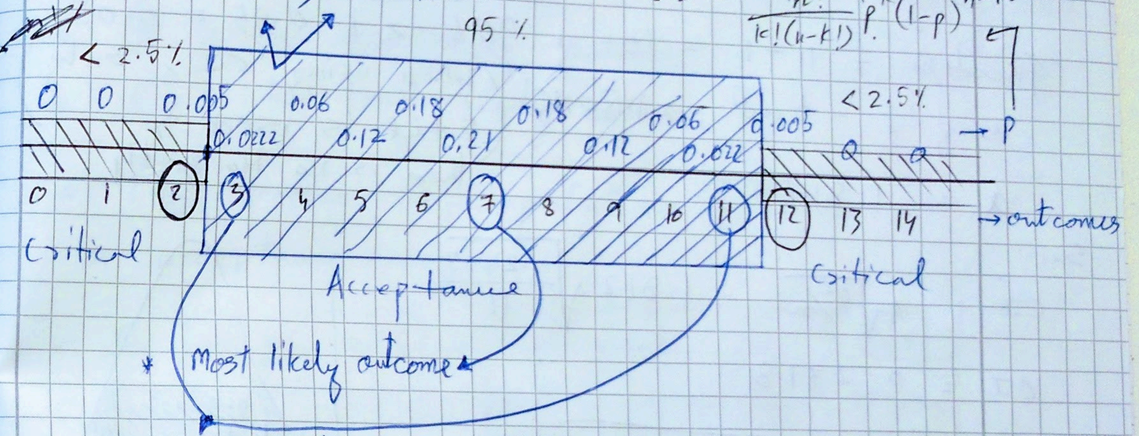
(18) (i)



T T T H H T H T T T T T T → $p = \frac{3}{14} = 0.21$

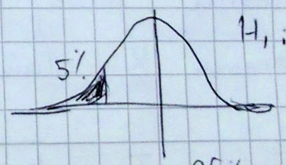
$N = 14$; $H = 3$; $T = 11$

0.5 ← $\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}$ → p

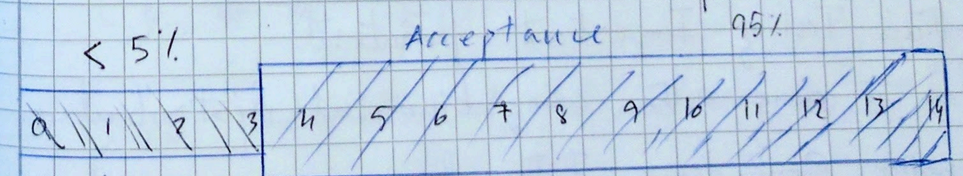


* Most likely outcome
* Smallest outcome
* Both 3 (H) & 11 (T) are in acceptance region.

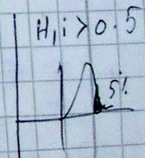
Case-2 ⇒ $H_1 < 0.5$



$H_1: < 0.5$



Selected b/c $H_1 < 0.5$



* Most likely → 7 with $p = 0.21$
* Smallest outcome → 11 with $p = 0.022$
* Test fails as 3 (H) is outside Acceptance Area

Question:

$H_0: p = 0.3$

$H_1: p > 0.3$

T H H T T T H T T H H

Test the claim.

Sol.

$p = \frac{5}{11}$; So: $M = \frac{5}{11} = 0.45$

$\sigma^2 = p(1-p) = 0.45(1-0.45) = 0.2475$

} Are these formulas to be used?

yes, b/c distribution is binary.

Now: Using t-table

Freedom $= 11 - 1 = 10$; $95\% \rightarrow 1 - 0.95 = 0.05$

Using 1-tail ..> WRONG (when using CI, only 2 tail is used & there is no H_0 & H_1)

$a = 1.812$

So:

$CI = a\sqrt{\sigma^2/n} = 1.812\sqrt{\frac{0.2475}{11}}$

$CI = 0.2718$

So: $\boxed{0.45 \pm 0.2718}$

* 0.3 is inside acceptance region, the test passes & claim ($p = 0.3$) is valid.

---

* Hypothesis Testing Important points:
  * Can be done either using binomial distribution or using Confidence Interval.
  * in case of binomial distr., a distribution is not required & we can generate it using # of Samples.
  * In binomial, $\mu = p$ & $\sigma^2 = p(1-p)$ ($H=1$, $T=0$) Same for binomial
  * In Confidence Interval, a distribution is required and $\mu = \frac{\sum x_i}{n}$ & $\sigma^2 = \frac{\sum (x_i - \mu)^2}{n}$ We don't use H_0 & H_1.

⟹ Cancer Treatment Hypothesis Test

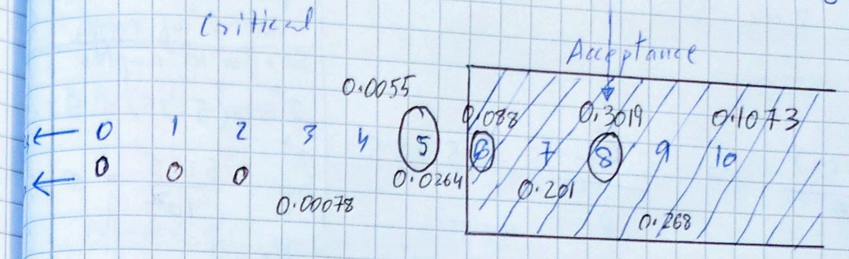
18(ii)

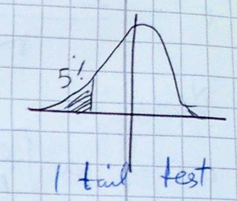
$H_0: p = 0.8$ → Claim

$H_1: p < 0.8$ → as well as advertised → Counter Claim

$N = 10$ → (expectation: 8 people healthy) → Sample

$\frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$





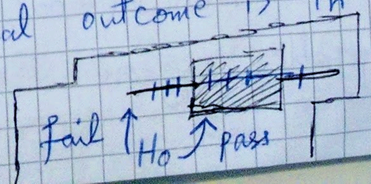
1 tail test

* 8 is the most likely outcome.
* 6 is the least Likely outcome
* Test passes as expectation 8 people is in acceptance region.
* For $\boxed{\text{Healthy} = 5}$ → $\boxed{\text{test fails as 5 is not in acceptance region.}}$

⟹ Summary of Hypothesis Test

given:
1- Null Hypothesis is given.
2- Alternative Hypothesis condition is given.
3- Actual outcome is given.
4- # of Sample.

find:
1- probabilities of outcomes using $\frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$
2- find Critical & Acceptance Region based on Alternate Hypothesis. ($<, >, \neq$)
3- Comment on most Likely, Least Likely outcome.
4- Test if actual outcome is in Acceptance region.



Question: Actual ⇒ 200. → [Method-1 Using CI]

199 200 201 202 203 204 205 206

find mean & confidence interval. ; $p = 95\%$

$$M = \frac{\Sigma X_i}{N} = 202.5$$

&

$$\sigma^2 = \frac{\Sigma (x_i - M)^2}{N} = \frac{42}{8} = 5.25$$

* Important: for CI, we always use σ^2 of avg.
So, for 10 samples
$\sigma^2 = 10(5.25) = 52.5$
$\sigma^2_{avg} = \frac{52.5}{10^2} = 0.525$

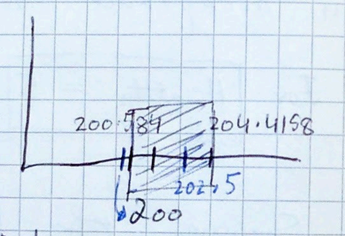
Now; Here $a = 2.365$ from table.

$$CI = a\sqrt{\frac{\sigma^2}{N}} = 2.365\sqrt{\frac{5.25}{8}}$$

$CI = 1.9158$

$$\boxed{202.5 \pm 1.9158}$$ → Confidence Interval

* Test is failed as 200 is not in acceptance region.



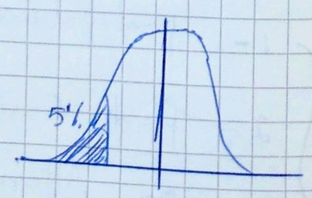
⇒ [Method-2 Using Hypothesis]

* Actual = 200

	199	200	201	202	203	204	205	206
p =	0	0	0	0	0.000356	0.00542	0.0546	0.2793

$H_0: p = 0.95$

$H_1: p < 0.95$



* 206 → Most likely
* 205 → Least Likely
* Test fails as Actual (200) is outside of ~~Accep~~ Acceptable region.

* For Variance:

$$\sigma^2 = \frac{\Sigma (x_i - M)^2}{n-1}$$

Can be used if sample size is small.

Question ⇒

AGE = 26

Distribution:

4 : 21, 6 : 24, 7 : 26, 11 : 29, 2 : 40

Number of samples (N) = 30

Test the Hypothesis using Confidence Interval:

1 – Mean:

$$M = \frac{\Sigma x_i}{n} = [4(21) + 6(24) + 7(26) + 11(29) + 2(40)]/30$$

$M = 26.9667$

2 – Variance:

$$\sigma^2 = \frac{\Sigma (M_i - X)^2}{n}$$

$\sigma^2 = 19.565$

3 – Let ~~Acc~~ probability = 90%

from table ⇒ freedom = 29, $p = 0.10$

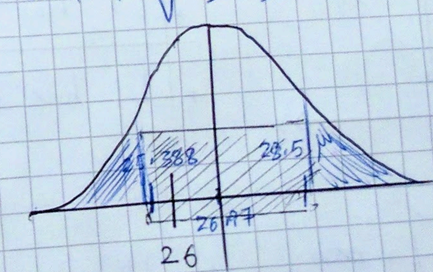
$a = 1.96$

4 – CI:

$$CI = a\sqrt{\sigma^2/n} = 1.96\sqrt{\frac{19.565}{30}} = 1.582$$

So:

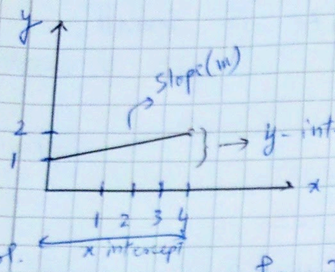
$$\boxed{26.97 \pm 1.582}$$



Since 26 is in Confidence Interval, the test passes → Means Avg. Age of person in Club is 26 with 90% accuracy.

⇒ Regression:

find m, b.

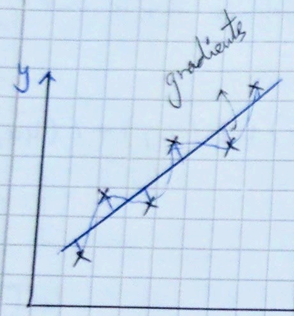


Sol.

$y = mx + b$

for $m = \frac{y_2 - y_1}{x_2 - x_1} = \frac{1-2}{0-4} \Rightarrow \boxed{m = \frac{1}{4}}$

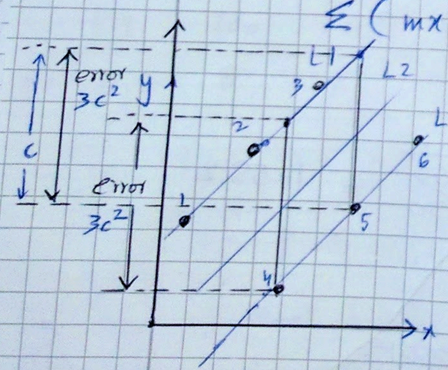
& for $x = 0 \Rightarrow 1 = m(0) + b \Rightarrow \boxed{b = 1}$



* non-Linear
* using Linear regression, we reduces y difference in Line → Line fitting.
* even if Linear, we can have variation b/c of noise.

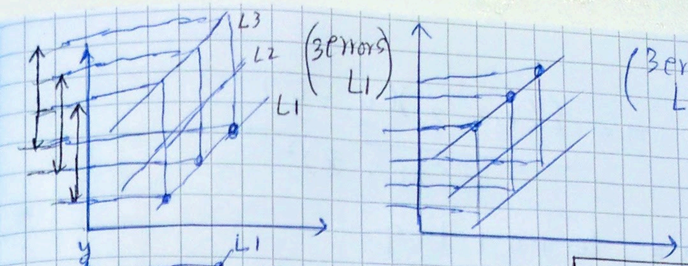
> If noise is Gaussian, then minimizing the quadratic deviation b/w data point & Line is a good method

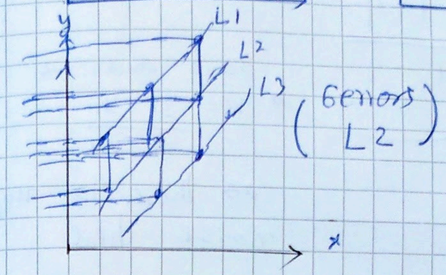
$\Sigma(mx + b - y)^2 \longrightarrow$ minimum



* Line 2 is best as it's min. error from all 6 points.
* L3 has errors for 1, 2, 3 points.
* L1 has errors for 4, 5, 6 points.

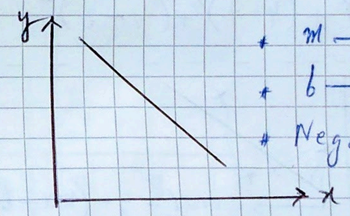
* error $= 3c^2$
  ↳ difference height





> * Line 1 errors $= 3c^2$
> * Line 3 errors $= 3c^2$
> * Line 2 errors $= 6\left(\frac{c}{2}\right)^2 = 1.5c^2$

> * Smaller but more errors are better than Larger & fewer error.
> This is due to quadratic nature of error.



* $m \rightarrow -ve$
* $b \rightarrow +ve$; $y = mx + b$
* Negative correlation as $x \propto \frac{1}{y}$

Data → $y = mx + b$ → m, b → Linear Regression

* $m = \frac{y_2 - y_1}{x_2 - x_1}$

$m = \frac{\Sigma(\bar{x} - x)(\bar{y} - y)}{\Sigma(\bar{x} - x)^2}$

* $y = mx + b$
  $b = y - mx$
  $b = \bar{y} - m\bar{x}$

⇒ Linear Regression:

$$\boxed{m = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2}}$$

$$\boxed{b = \bar{y} - b\bar{x}}$$

Question 1

x	y
6	7
2	3
1	2
-1	0

Find $m = \dfrac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2}$

$m = \dfrac{16+0+1+9}{16+0+1+9}$

$\boxed{m = 1}$

Same ans. as $m = \dfrac{y_2 - y_1}{x_2 - x_1}$, as there is no noise here

And $b = \bar{y} - m\bar{x} = 1$

$\boxed{b = 1}$

Question:

x	y
4	7
3	9
7	1
2	11

$m, b = ?$

Sol. $\bar{x} = 4;\ \bar{y} = 7$

And:

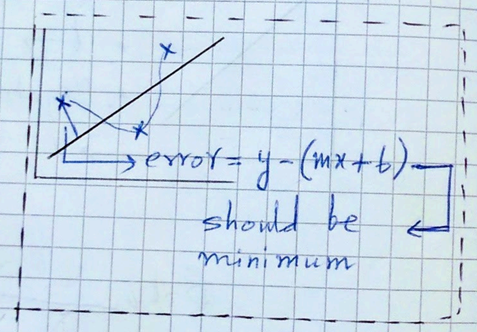
$m = \dfrac{(4-4)(7-7)+(4-3)(7-9)+(4-7)(7-1)+(4-2)(7-11)}{(4-4)^2+(4-3)^2+(4-7)^2+(4-2)^2}$

$\boxed{m = -2}$

And:

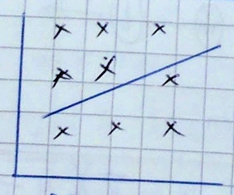
$b = 7 - (-2)(4)$

$\boxed{b = 15}$

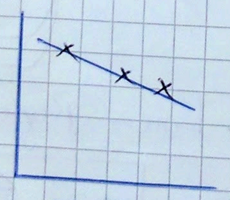


⇒ Correlation: $(-1 \le r \le +1)$

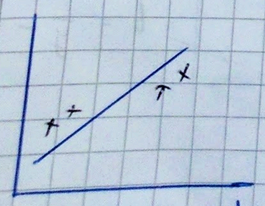
$r = 1$
Positive Corr.



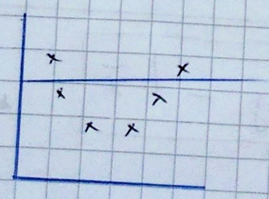
$r = 0$
No Corr.



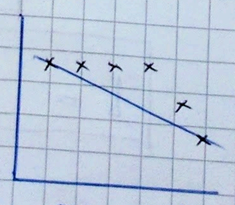
$r = -1$
Negative Corr.



$0 < r < 1$



$r = 0$



$r < 1$

$r \rightarrow$ Correlation Coeff.

* $\boxed{r = \dfrac{\Sigma(\bar{x} - x)(\bar{y} - y)}{\sqrt{\Sigma(\bar{x} - x)^2\ \Sigma(\bar{y} - y)^2}}}$ → Covariance (Variance of 2 variables)

  tells how covariance evolved

* $\boxed{m = \dfrac{\Sigma(\bar{x} - x)(\bar{y} - y)}{\Sigma(\bar{x} - x)^2}}$ → Mean Slope

* $\boxed{b = \bar{y} - m\bar{x}}$ → y - intercept

* $\boxed{m = \dfrac{y_2 - y_1}{x_2 - x_1}}$ → Slope only for Straight Lines

x	3	4	5
y	7	8	9

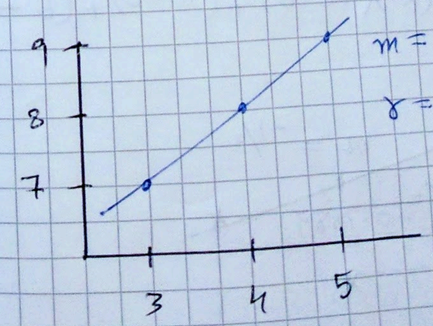
$m = ?$
$r = ?$

Sol.

$m = \dfrac{\Sigma(\bar{x} - x)(\bar{y} - y)}{\Sigma(x - \bar{x})^2} = \dfrac{2}{2} = 1$

$r = \dfrac{\Sigma(\bar{x} - x)(\bar{y} - y)}{\sqrt{\Sigma(\bar{x} - x)^2\ \Sigma(\bar{y} - y)^2}} = \dfrac{(4-3)(8-7)+(4-4)(8-8)+(4-5)(8-9)}{\sqrt{(1^2+0^2+1^2)(1^2+0^2+1^2)}}$

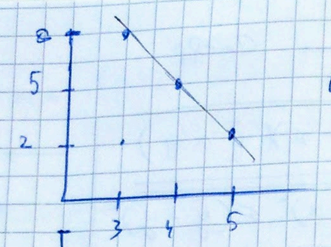
$r = \dfrac{2}{\sqrt{(2)(2)}} = 1$



$m = 1$
$r = 1$

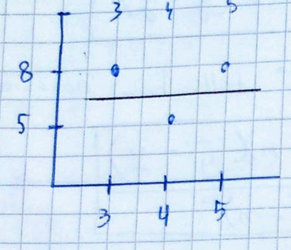
→ $\dfrac{y_2 - y_1}{x_2 - x_1}$ can be used as its straight line

x	3	4	5
y	8	5	3



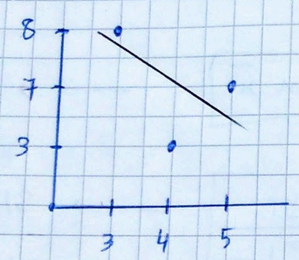
$\boxed{r < 0}$

x	3	4	5
y	8	5	8



$\boxed{r = 0}$

x	3	4	5
y	8	3	7



$\boxed{r < 0}$

x	3	4	5
y	8	3	7

$\bar{x} = 4$

$\bar{y} = 6$

$$m = \frac{\Sigma(\bar{x}-x)(\bar{y}-y)}{\Sigma(\bar{x}-x)^2} = \frac{(4-3)(6-8)+(4-4)(6-3)+(4-5)(6-7)}{(4-3)^2+(4-4)^2+(4-5)^2}$$

$$\boxed{m = \frac{-1}{2}}$$

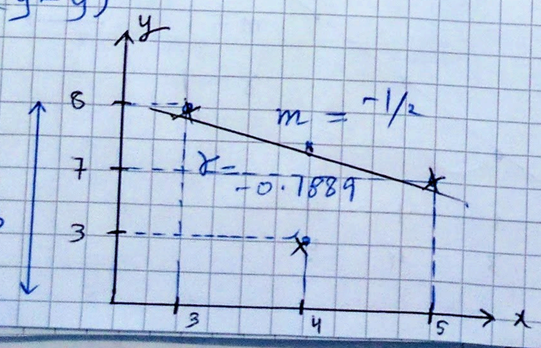
And:

$$b = \bar{y} - m\bar{x} = 6 - \left(\frac{-1}{2}\right)(4) = 8$$

And

$$r = \frac{\Sigma(\bar{x}-x)(\bar{y}-y)}{\sqrt{(\Sigma(\bar{x}-x)^2)\,\Sigma(\bar{y}-y)^2}} = \frac{-1}{\sqrt{(2)((6-8)^2+(6-3)^2+(6-7)^2}}$$

$$\boxed{r = -0.1889}$$



x	y
1	4
2	7
3	13

Find $m, r \rightarrow ?$

Sol.

$\bar{x} = 2$.

$\bar{y} = 8$.

Now:

$$m = \frac{9}{2} = 4.5$$

$$r = \frac{9}{\sqrt{(2)(45)}} = 0.9819$$

$\Rightarrow$ Case - 1: (y is twice).

So:

$$m = \frac{9(2)}{2} = 9$$

$$r = \frac{9(2)}{\sqrt{2^2(2) \times 45}} = 0.9819$$

$\sigma^2 \sigma^2$

* So $m \longrightarrow$ twice

$r \longrightarrow$ no change.

$\Rightarrow$ Case - 2 (x, y both twice):

$$m = \frac{(9)(2)(2)}{(2)^2(2)} = 4.5$$

$$r = \frac{(9)(2)(2)}{\sqrt{(2^2)(2)(2)^2(45)}} = 0.9819$$

So:

$m, r \longrightarrow$ no change

⇒ find $\frac{m}{r}$.

$$m = \frac{\Sigma(\bar{x}-x)(\bar{y}-y)}{\Sigma(\bar{x}-x)^2} \longrightarrow \frac{A \cdot B}{C}$$

$$b = \frac{\Sigma(\bar{x}-x)(\bar{y}-\bar{y})}{\sqrt{\Sigma(\bar{x}-x)^2\,\Sigma(\bar{y}-y)^2}}$$

So:

$$\frac{m}{r} = \frac{A \cdot B}{C} \times \frac{D}{A \cdot B} = \frac{D}{C}$$

$$\frac{m}{r} = \frac{\sqrt{\Sigma(\bar{x}-\bar{x})^2} \cdot \sqrt{\Sigma(\bar{y}-y)^2}}{\Sigma(\bar{x}-\bar{x})^2}$$

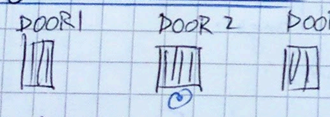
$$\frac{m}{r} = \frac{\sqrt{\Sigma(\bar{y}-y)^2}}{\left[\Sigma(x-\bar{x})^2\right]^{1-1/2}}$$

$$\frac{m}{r} = \sqrt{\frac{\Sigma(\bar{y}-y)^2}{\Sigma(\bar{x}-x)^2}}$$

$$\boxed{\frac{m}{r} = \frac{\sigma_y}{\sigma_x}} \longrightarrow \frac{\text{Variance } Y}{\text{Variance } x} \longrightarrow \text{Variance Ratio}$$

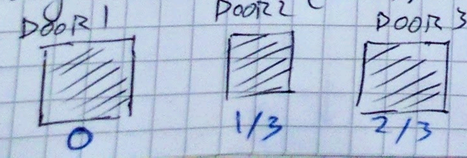
⇒ <u>Monty Hall Problem</u> (2 choices doesn't mean 50%/50%)

you selects = 2



	True Location	Monty Selects	Probability.
p = 1/3	1	1	0 → he wont show true
	1	2	0 → its yours
	1	3	1/3
p = 1/3	2	1	1/6 → (1/6) ÷ (1/2)
	2	2	0 → wont show true
	2	3	1/6 Normalizer ↵
p = 1/3	3	1	1/3 → (1/3) ÷ (1/2)
	3	2	0 → yours
	3	3	0 → wont show true

* Now say he selects (shows) = Door 1



* So door 3 actually has more chance than Door 2