RANDOM VARIABLE

Random Experiment: An experiment is said to be random if it's outcome can't be predicted with certainty. e.g. throw of dice, coin toss

Sample Space: The set of all possible outcomes of an experiment

e.g. throw of dice $S = \{1, 2, 3, 4, 5, 6\}$

coin toss $S = \{H, T\}$

Event: Subset of sample space

e.g. in a throw of dice, outcome being even

$E = \{2, 4, 6\}$

Random Variable: Set of all possible values from a random experiment

e.g. in a coin toss

$$\underbrace{X}_{\text{Random Var}} = \begin{cases} 0 \leftarrow \text{Head} \\ 1 \leftarrow \text{Tail} \end{cases}$$

Random Var — Possible values — Random Events

Imp. to note that assigning values to outcomes of random events/experiment and taking that set = Random Variable

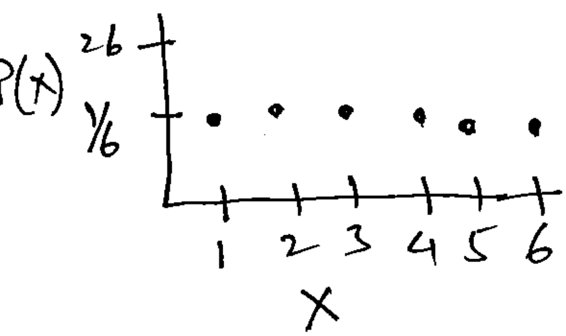
Random Variable (Alternate Def^n: Fn defined on sample space)

Discrete

PMF

(Prob. Mass Fn)

e.g. throw of dice



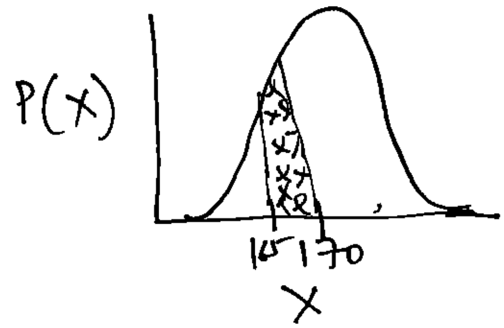
$P(x=3) = \frac{1}{6}$

Continuous

PDF

(Prob. Density Fn)

e.g. height of people



$P(X=170)$ does not make sense

Generally defined in some interval e.g. 165 to 170

$P(X>165, X<170)$ = Area shaded above

Statistical distribution or just "distribution" of a random var: describes freq with which values of random var. occur

Prob. distribution of a random var → describes how the probabilities are distributed over the values of random variable. Sum of prob for all values of random var = 1

EXPECTED VALUE OF A RANDOM VARIABLE

- Measure of central tendency of a random var
- mean value/outcome

$$E(X) = \sum_{i=1}^{n} x_i \, P(X_i) \quad \Rightarrow \text{Discrete}$$

x_i → outcome in numerical value

$P(X_i)$ → prob. of outcome

$$E(X) = \int_{-\infty}^{\infty} x \, P(X) \, dx \quad \Rightarrow \text{Continuous}$$

VARIANCE OF RANDOM VAR: $V(X) = E(X^2) - [E(X)]^2$

e.g. Expected value throw of dice

$$E(X) = 1\frac{1}{6} + 2\frac{1}{6} + 3\frac{1}{6} + 4\frac{1}{6} + 5\frac{1}{6} + 6\frac{1}{6}$$

$$= \frac{1}{6} + \frac{2}{6} + \frac{3}{6} + \frac{4}{6} + \frac{5}{6} + \frac{6}{6}$$

$$= \frac{21}{6} = 3.5$$

$$E(X^2) = 1^2\frac{1}{6} + 2^2\frac{1}{6} + 3^2\frac{1}{6} + 4^2\frac{1}{6} + 5^2\frac{1}{6} + 6^2\frac{1}{6}$$

$$= 1\cdot\frac{1}{6} + 4\cdot\frac{1}{6} + 9\cdot\frac{1}{6} + 16\frac{1}{6} + 25\frac{1}{6} + 36\frac{1}{6}$$

$$= \frac{91}{6}$$

$$V(X) = \frac{91}{6} - \left(\frac{21}{6}\right)^2 = \frac{105}{36}$$

EXPECTED VALUE PROBLEMS: (Mean value of experiment when we repeat experiment very large no. of times)

1. List "all" the outcomes
2. List the probabilities of each outcome
3. List the value (numerical) of the random variable (from the perspective of party whose expected value is asked)

Ques: A player throws a dice. If a prime number is obtained, he gains to win an amount equal to the number rolled times 100 dollars, but if a prime number is not obtained, he loses an amt equal to the number rolled times 100 dollars. Calculate the probability distribution and the expected value of the described game.

Sol^n

outcome	Probability	Value
1	1/6	$-1 \cdot 100$
2	1/6	$+2 \cdot 100$
3	1/6	$+3 \cdot 100$
4	1/6	$-4 \cdot 100$
5	1/6	$+5 \cdot 100$
6	1/6	$-6 \cdot 100$

$$E(x) = \frac{1}{6} \times -100 + \frac{1}{6} \times 200 + \frac{1}{6} 300 - \frac{1}{6} \cdot 400 + \frac{1}{6} \cdot 500 - \frac{1}{6} \cdot 600$$

$$= \frac{1}{6}\left[-100\right] = \frac{-100}{6}$$

Ques: A player tosses two coins into the air. He gains to win $1 times the number of heads that are obtained. However, he will lose $5 if neither coin is head. Calculate the expected value of this game and determine whether it is favorable for the player.

Sol^n

Outcome	Prob.	Value
HH	$\frac{1}{2}\times\frac{1}{2}$	2
HT	$\frac{1}{2}+\frac{1}{2}$	1
TH	$\frac{1}{2}+\frac{1}{2}$	1
TT	$\frac{1}{2}\times\frac{1}{2}$	−5

$$E(x) = \frac{1}{\cancel{4}_2}\cdot\cancel{2} + \frac{1}{4}\cdot 1\cdot + \frac{1}{4}\cdot 1 - 5\cdot\frac{1}{4}$$

$$= -1/4 \quad \text{(Unfavorable)}$$

Ques: An insurance company charges $150 for a policy that will pay for at most one accident. For a major accident, the policy pays $5000; for a minor accident, the policy pays $1000. The $150 premium is not returned.

P (major accident) = 0.005

P(minor accident) = 0.08

Expected value of policy to insurance company?

Solⁿ

Outcome	Probability	Cost to Company	Premium
major accident	0.005	−5000	\$150
minor accident	0.08	−1000	
no accident	$1-(0.005+0.08)$ $= 0.915$	0	

$$E(x) = 0.005(-5000) + 0.08(-1000) + 0.915(0) + 150$$

$$= 45 \$$$

Ques: Your Grade = # of correct answers − $\frac{1}{5}$ (# incorrect answers)

Every question has 5 options as answers

Suppose you guess at the answer to all 100 questions. What is the expected grade for the test?

Solⁿ:

Outcome	Prob.	Value
guess right	$\frac{1}{5}$	1×100
guess wrong	$\frac{4}{5}$	$-\frac{1}{5} \times 100$

$$E(x) = \frac{1}{5} \times 100 + \frac{4}{5}\left(-\frac{1}{5} \times 100\right)$$

$$= \frac{100}{5} - \frac{400}{25} = 20 - 16$$

$$= 4$$

Ques: FAIR GAME: EXPECTED VALUE = 0

Suppose for some game $p(\text{win}) = 2/6$ $p(\text{lose}) = 4/6$
If you lose you pay \$1, if you win other player pays you \$D. What should D be if the game is fair

$$E = \frac{2}{6} \times D + \frac{4}{6}(-1)$$

$$0 = \frac{2}{6} D - \frac{4}{6}$$

$$\therefore D = 2$$

Ques: Assume that it costs \$1 to play a state's daily number. The player chooses a three-digit number between 000 and 999, inclusive, and if the number is selected that day, then the player wins \$500

a) what is expected value of the game?

b) What should be the price of a ticket to make this game fair?

Solⁿ

Outcome	Prob	Value
Win	$1/1000$	$500 - 1 = 499$
Lose	$\frac{999}{1000}$	-1

$$E(x) = \frac{1}{1000} \times 499 + \frac{999}{1000}(-1) = -0.50$$

Let x be the fair price of ticket so that $E(x) = 0$

$$\frac{1}{1000}(500 - x) + \frac{999}{1000}(-x) = 0$$

$$\therefore x = 0.50$$

<u>Ques</u>: What is the number of heads we can expect when we flip four coins?

<u>Sol</u> Outcome relates to coming heads. In 4 coins heads can come 0, 1, 2, 3 or all 4 times

Outcome	Prob	Value is same as outcome
0	$1/16$ $\left(\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}\right)$	
1	$4/16$	
2	$6/16$	
3	$4/16$	
4	$1/16$	

$$= 0\cdot\frac{1}{16} + 1\cdot\frac{4}{16} + 2\cdot\frac{6}{16} + 3\cdot\frac{4}{16} + 4\cdot\frac{1}{16}$$

$$= \frac{32}{16} = 2$$

<u>GENERALIZATION</u>: $N/2$ where N = no. of coins

Ques: What is expected number of coin flips for getting a head?

Soln: Let the expected number of coin flips be x.

a) If the first flip is head, we are done. Prob. of this event is $1/2$ and no. of coin flips needed is 1

b) If the first flip is tail, then we have to start all over again. Prob of this event is $1/2$ and since we have wasted one flip, no. of coin flips now needed is $x+1$

$$\therefore x = \frac{1}{2}(1) + \frac{1}{2}(x+1)$$

$$x = \frac{1}{2} + \frac{x+1}{2}$$

$$2x = x+2$$

Ans: $x = 2$

Ques: What is expected number of coin flips for getting two consecutive heads?

Sol^n Let x be the no. of coin flips needed

1) first flip is tail. $P = \frac{1}{2}$ & no. of flip req $= x+1$ (since we have wasted one flip)

2) first flip is head & second flip is tail.

$P = \frac{1}{2}$ $\qquad$ $P = \frac{1}{2}$

total no. of flips req $= x+2$ (since we have wasted two flips)

3) Both flips are head. $P = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$ & no. of flips req $= 2$

$$x = \frac{1}{2}(x+1) + \frac{1}{2} \cdot \frac{1}{2}(x+2) + \frac{1}{4} \cdot 2$$

Ans: $x = 6$

Ques: Expected no. of flips for n consecutive heads

Sol^n $2^{N+1} - 2$ (Generalized formula)

Ques: Candidates are appearing for interview one after another. Probability of each cand. getting selected is 0.16. What is the expected no. of candidates that you will need to interview to make sure that you select somebody?

Let X be the no. of cand. that need to be interviewed

Soln 1) If first cand. is selected. $P = 0.16$

no. of cand. interviewed $= 1$

2) If the first cand is not selected

$P = 1 - 0.16$

no. of cand. to be interviewed now $= X+1$

$$X = (0.16) \times 1 + (1-0.16)(X+1)$$

Ans: $X = 6.25$

Ques: What is expected no. of dice throws to get a "four"?

Soln $P(4) = 1/6$ so if you throw dice 6 times one of them will be 4

$\therefore$ Ans: 6

LAW OF LARGE NUMBERS: If the same experiment is performed large number of times, then avg of results = Expected Value

where Expected Value = Σ Each possible outcome $\times$ it's prob

Eg. If a six sided dice is rolled large number of times then the avg. of their values or outcome $\Rightarrow \dfrac{1+2+3+4+5+6}{6} = 3.5$

The idea is if you want to know avg. outcome, you can use expected value

CHARACTERISTICS OF A DISTRIBUTION

→ A distribution is characterized by

— location (mean, median, mode

- scale (spread e.g. std. dev.)

If the above two are not sufficient to define a distribution

— shape (skew, kurtosis)

e.g. Normal distribution is characterized by mean/std.dev (location/scale)

Binomial distribution is characterized by

mean = np & variance = np(1-p)

where n = # of trials

p = prob. of success

→ E.g. of Continuous Distribution: Normal distribution
t-distribution

E.g. of Discrete Distribution: Binomial distribution
Poisson distribution

BINOMIAL DISTRIBUTION:

→ A binomial experiment has the following properties:

i) n identical trials

ii) two outcomes i.e. success or failure

iii) p : prob. of success does not change from trial to trial

iv) trials are independent

→ Binomial Prob. Mass Function (PMF) provides the prob. that k successes will occur in n trials

$$= \binom{n}{k} p^k (1-p)^{n-k}$$

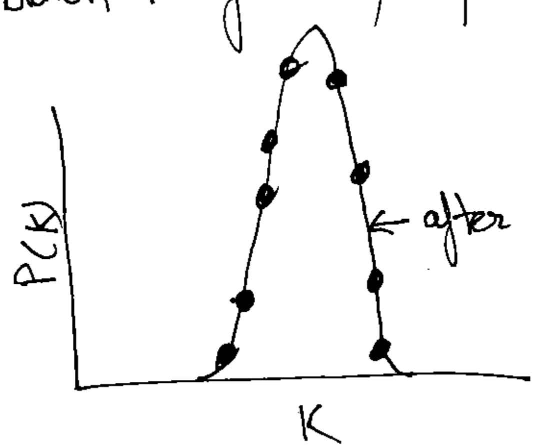
→ When $n=1$ ⇒ Bernoulli Distribution

→ The distribution has following properties:

$$\text{mean} = np$$
$$\text{Var} = np(1-p)$$

and the binomial distribution is generally represented as: $B(n,p)$

→ Sample distribution:
(remember → this is discrete dist.)



← after fitting a curve

POISSON DISTRIBUTION:

→ Models # of arrivals within a period of time

→ Characterized by λ = mean number of occurrence in interval
✓ e.g. $P(\lambda)$

→ mean $= \lambda$
✓ var $= \lambda$

→ PMF provides the probability of K arrivals within the same period of time for which λ is known
✓

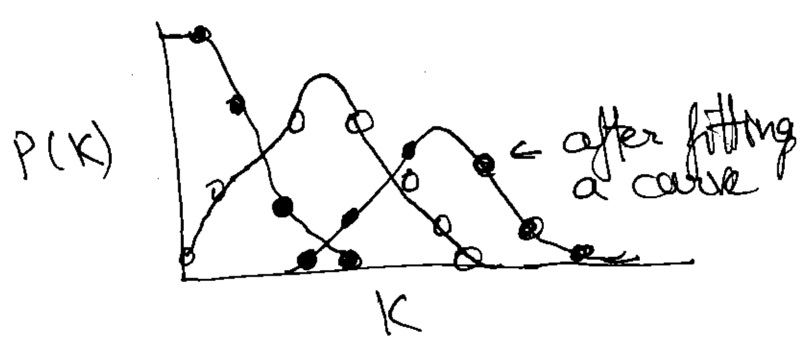
$$= \frac{\lambda^K e^{-\lambda}}{K!}$$

where e = Euler's const. $\simeq 2.71$
λ = mean # of occ. in interval

→ E.g. Mean # of calls coming in 15 min = 10
Prob. that 5 calls come in within next 15 min

ie. $\lambda = 10$, $K = 5$ $\Rightarrow \frac{10^5 e^{-10}}{5!} = 0.0378$

→ Sample dist.
(rem → discrete dist)



NORMAL DISTRIBUTION

→ mean = median = mode

symmetry about the center i.e. 50% values < mean
& 50% values > mean

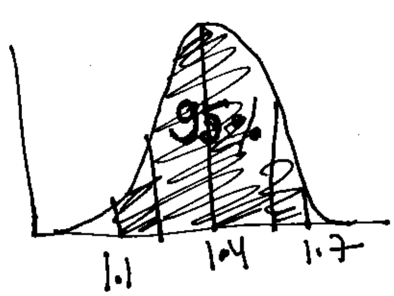
→ 68% of the values are within 1 std. dev. of mean
95% of the values are within 2 std. dev. of mean
99.7% of the values are within 3 std. dev. of mean

→ Example: 95% of students at school are b/w 1.1 m & 1.7 m tall. Assuming this data is normally distributed, can you calc. mean & std. dev

Mean = halfway b/w 1.1 & 1.7 $= \frac{1.1+1.7}{2} = 1.4$ m

95% → 2 std. dev. from mean on either side

$\therefore$ 1 std. dev. $= \frac{1.7\text{m} - 1.1\text{m}}{4} = 0.15$ m



→ Normal Dist. Vs. Standard Normal Distrib. (Z-score)

(Standard Normal Distrib. — also called z-distribution)

The value of the var. whose distribution needs to be investigated can be any number. To "standardize" the value of the var, it is transformed to values ~~between~~ [0, 1] ← Standard Normal Dist. with mean = 0 & std. dev = 1

⇓ since it is prob. dist, ∴ Area under curve = 1

✓ Now any one value of the var, when mapped into standard normal dist. is called → Z-score

Z-score = number of std. dev. from mean

(Z-score can be +ve or -ve: +ve (above mean), -ve (below mean))

Example: 95% of students at school are b/w 1.1 m & 1.7 m. One of the student has height = 1.85 m. What is his z-score?

$$\text{Mean} = \frac{1.1 + 1.7}{2} = 1.4 \text{ m}$$

How far is 1.85 m from mean $= 1.85 - 1.4$
$= 0.45$

$$\text{Std. dev} = \frac{1.7 - 1.1}{4} = 0.15$$

$$\therefore \frac{0.45}{0.15} = 3 \text{ std dev}$$

$\Rightarrow$ z-score $= 3$

Z-score formula

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

where μ = mean
σ = std. dev
x = value to be standardized
z = z-score

UNIFORM DISTRIBUTION

DISCRETE UNIFORM DIST:

Prob. of each outcome is same $= \dfrac{1}{\text{num. of outcomes}}$

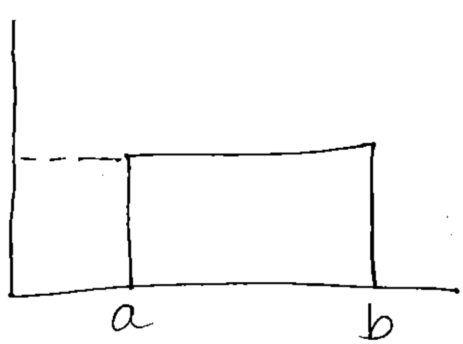
e.g. coin flip $\rightarrow \frac{1}{2}$, roll of dice $\rightarrow 1/6$

CONTINUOUS UNIFORM DIST:

If the data is temp, dist, income, mass etc., they can be measured very precisely to several decimal pts $\therefore$ Number of outcomes $= \infty$

In these cases, we use continuous uniform dist

$f(x)$ $\rightarrow$ Prob. density fn



Since, it is uniform dist $f(x)$ is constant over the possible values of x.

Since $f(x)$ is prob. density fn, area under curve $= 1$

Area $=$ base $\times$ ht.

$\cancel{f(x)}\ 1 = (b-a) \times f(x)$

$$\therefore f(x) = \frac{1}{b-a}$$

∴ pdf of uniform dist.

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{for } a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

Median = Mean $= \dfrac{a+b}{2}$

Variance $\sigma^2 = \dfrac{(b-a)^2}{12}$

Ques: What is the prob. that x is in between 550 & 1000 given x is Uniformly dist. b/w 400 & 800?

$P(550 < x < 1000) = ?$, given $P(400 < x < 800) =$ uniform

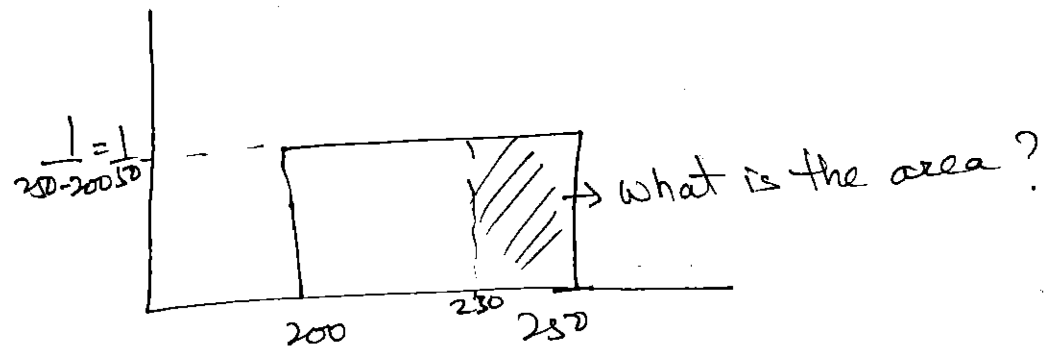
Sol^n

Area of region b/w 550 & 800

base × ht $= (800-550) \times \dfrac{1}{800-400} = \dfrac{800-550}{800-400} = \dfrac{250}{400} = 0.625$

Ques: What is the prob that random var $X > 230$ given X is uniform dist between 200 & 250?

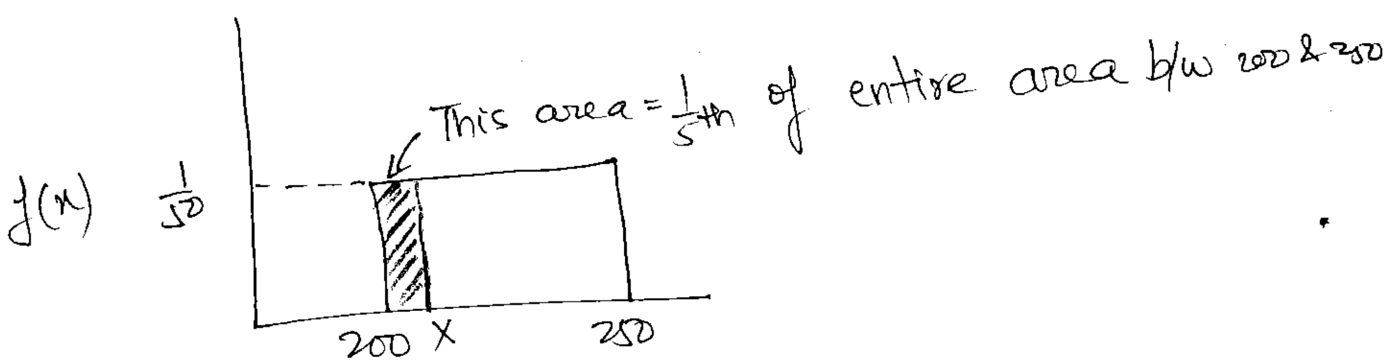
Sol^n



Area = base × ht

$= (200 - 230) \times \frac{1}{50} = \frac{2}{5} = 0.4$

Ques: What is 20th percentile of this uniform dist?



Sol^n: Area b/w 200 & 250 $= (250 - 200) \times \frac{1}{50} = 1$

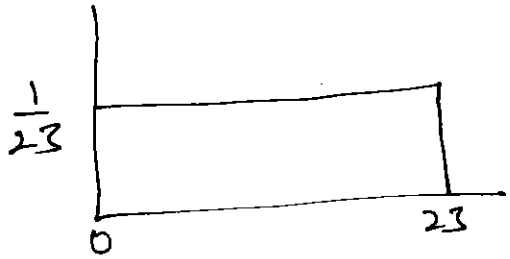
Area b/w 200 & $x \Rightarrow (x - 200) \times \frac{1}{50}$

$\therefore (x - 200) \times \frac{1}{50} = \frac{1}{5}(1)$

Solving for x: $x = 210$

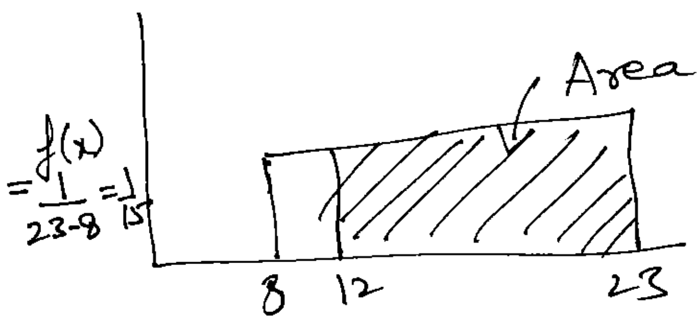
Ques: Given x is uniform dist. b/w 0 & 23, what is the prob that $X > 12$ given $X > 8$

Sol^n: This is conditional question



→ original space

↓ bcz of condition i.e. given $x > 8$



→ new space

Shaded Area : $(23-12) \times \frac{1}{15} = \frac{11}{15}$

ESTIMATION (STATISTICAL INFERENCE)

→ Suppose we want to find (estimate) mean ht. of all the people in world. Because of time, cost and other considerations data cannot be collected from every element of population. In such cases, a subset of population, called a sample, is used to provide the data. Data from the sample are then used to develop estimates of the characteristics of the larger population.

→ The process of using a sample to make inferences abt. a pop. ⇒ statistical inference

→ Parameters Vs sample statistics (or parameter estimates)

Parameters ↓ Characteristics of pop. e.g. pop. mean, pop. variance etc.

Sample statistics ↓ Characteristics of sample e.g. sample mean, sample var. etc

→ Types of estimates:

- Point estimate: value of sample statistics that is used as single estimate of pop. parameter
- Interval estimate (confidence interval): interval value of sample statistics that is used to estimate the pop. parameter

SAMPLING DISTRIBUTION:

→ Sampling dist : prob. dist. of a ~~sample~~ statistic — sample statistic e.g. mean / test statistic e.g. t-statistic, Z-statistic

→ Knowledge of sampling distribution is necessary for the construction of an interval estimate for a pop. parameter

→ This is why a prob. sample is needed; without a prob. sample, the sampling dist. cannot be determined and an interval estimate of a pop. parameter cannot be constructed

Note: a prob. sample is a sample in which each element of pop. has a "known" prob. of being included in the sample. If equal prob. → simple random sample

CENTRAL LIMIT THEOREM

→ Given a sufficiently large sample size the sampling dist. of mean for a var will approximate a normal distribution regardless of that var's dist in pop

→ Dist. of a var in a pop : can have any dist

→ Sampling dist. of Mean : Take n samples (of large size) with replacement, calc. their mean and graph them on histogram

gen. >30

Note: # of samples ≠ sample size

no. of elements in a sample

→ CLT links
- The dist. of var in pop
- Sampling dist. of mean

→ <u>Sample size & CLT:</u>

Normal dist is characterized by
- mean
- std. dev

As the sample size increases, the sampling dist. converges to a normal dist. where

mean of sampling dist = pop. mean ← point estimate not interval estimate

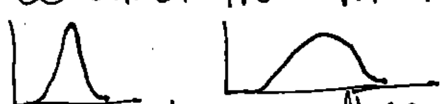
$$\text{stand. dev of sampling dist.} = \frac{\text{pop. std. dev}}{\sqrt{\text{sample size}}}$$

⇓ std. error of mean

If pop. std. dev is not known

$$= \frac{\text{sample std. dev}}{\sqrt{\text{sample size}}}$$

⇒ As ~~std.~~ sampling size increases, the std. dev of sampling dist decreases i.e. sampling dist. clusters more tightly around the mean



Large sample size — Small sample size

→ <u>Importance of CLT:</u>

✓ 1) Normality Assumption: Statistical Tests
- Parametric — e.g. t-test, F-test, z-test (normality assumption in pop. data) ⇒ more powerful (normal dist. has properties that can be used in statistical methods)
- Non-parametric (distribution-free) — e.g. chi-square test (normality is not assumed in pop. data) ⇒ less powerful

However, parametric tests of the mean are robust to departures from normality assumption when sample size is large ⇒ due to CLT

2) With larger sample size, the sampling dist. of mean clusters more tightly around pop mean ⇒ more precise estimates

Why CLT is overhyped?

(it only talks abt mean of sample & if sample size ≥ 30) → 1) To justify normal dist. is very common
2) In statistics, all parametric methods require normality assumption of pop - to justify it

If pop. is assumed normal, lot of tests & procedures can be used else not possible to calculate anything (or may be approximate). Hence most statistical methods assume that the pop. is normally distributed.

CLT justifies if large sample sizes → sampling dist. of mean = normal

---

Note: Standard Deviation - Population

$$\sigma = \sqrt{\frac{\Sigma (x_i - \mu)^2}{n}}$$

where x_i = each element in pop

μ = pop. mean

n = size of population

Std. Deviation - Sample

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

where

x_i = each element in sample

$\bar{x}$ = sample mean

n = size of sample

Why n-1 instead of n in denom.?

- Bessel's correction
- unbiased estimator

A random process generates data. The generated data has a distribution — Any distribution

Goodness of Fit Test : used to test if sample data fits a distribution from a certain pop. (pop. with poisson dist. or binomial dist.)

Now when pop. has diff. dist. than normal, parametric tests such as t-test, z-test, F-test cannot be applied. There are separate tests for these non-normal distributions.

CONFIDENCE INTERVAL

- It is the range of values, derived from sample statistics, which is likely to contain the value of unknown pop. parameter
- The wider the confidence interval, the ~~less~~ uncertainty abt. the value of pop. parameter (e.g. 90% CI is narrower than 95% CI and has smaller conf. of including pop. parameter)
- This uncertainty is bcz of the sampling method
  sampling error = | pop mean − sample mean |
- Suppose, we want to estimate pop. mean. ~~from~~ one sample with a given conf. level, we can estimate pop. mean

pop. mean = sample mean ± margin of error

where margin of error = critical value × std. dev. of sample statistic

& critical value is obtained from confidence level + sampling dist. of statistic

When the sampling dist. is nearly normal, the critical value can be expressed as t-score (sample size < 30) or z-score for a given confidence level

e.g.: for a conf. level of 95% $\xrightarrow{\text{z-score}}$ 1.96

for a conf. level of 90% $\xrightarrow{\text{z-score}}$ 1.645

→ Z-score = 1.96 can be interpreted as # of std. errors from the mean necessary to include 95% of the values in normal distribution

→ Interpretation of 95% conf. interval: 95% of the intervals constructed in this manner will contain pop. mean

Note that one sample is used to estimate pop. mean now if this experiment is repeated 20 times, 19 of these times the computed conf. interval will contain true pop. mean

Incorrect Interpretation: 95% prob. that pop mean will fall between computed conf. interval

Bayesian Interpretation of C.I. (params have prob, not constant)

In frequentist statistics, pop. parameters are always fixed (constant) and not a random var. It does not change. The prob. that a constant falls within any given range is always 0.0 or 1.0.

→ CI are used to bound mean, proportion, regression coeff, for the diff. b/w populations etc.

<u>EXAMPLE</u>: Find the avg/mean ht. of all men?? with 95% confidence

Given 40 random men have mean ht. of 175 cm. std. dev. of 20 cm

<u>Sol^n</u>

sample mean = 175

sample size = 40

sample std. dev = unknown

~~sample~~ std. dev = 20

Note: If pop. std. dev ~~is~~ not known, sample std. dev is calc $= \sqrt{\dfrac{(x_i - \bar{x})^2}{n-1}}$

If individual values in sample is not known, pop std. dev = sample std. dev.

95% CI $\xRightarrow{\text{Z-score}}$ 1.96 (95% of values in normal distrib.)

pop mean = sample mean ± margin of error

$= 175 \pm 1.96 \dfrac{20}{\sqrt{40}}$

Note: Margin of Error = Z-score $\times$ $\dfrac{\text{sample/pop std. dev}}{\sqrt{\text{sample size}}}$ std. dev. of sample statistic

Also, can be seen Margin of Error $\propto \dfrac{1}{\sqrt{\text{samplesize}}}$

GENERAL PROCEDURE TO CONSTRUCT CI

1. Identify sample statistic (sample mean, sample proportion etc)
2. Select conf. level (e.g. 90%, 95%)
3. Find Margin of Error

   Margin of Error = Critical Value × std. dev of sample statistic

   where critical value is derived from conf. level + sampling dist. of statistic

4. Compute CI: CI = sample statistic ± Margin of Error

<u>Example:</u> CI for regression coeff

1) regress. coeff
2) 95%, 90% etc
3) std error is generally given in regression O/P:

   Margin of error = std. error × crit. value

   For a conf level, using t-dist. table, get t-score

   e.g. crit. value = 2.63

   Note: t-dist. table also uses degree of freedom (+ conf level) to provide crit. value. For regression, degree of freedom = # of samples - 2
4) CI = reg. coeff ± 2.63 × std. error of the reg. coeff

HYPOTHESIS TESTING

- Hypothesis testing is a form of inferential statistics that allows ~~to~~ draw conclusions about an entire population based on a representative sample

- Steps in Hypothesis testing:
  1) Decide b/w one-tailed or two-tailed test
  2) Formulate Null Hypothesis & Alternate hypothesis
  3) Decide α (significance level)
  4) Find out p value
  5) Interpret the results

- Example: A researcher wants to find out if monthly expenditure on fuel for families has changed since last year. Last year's avg. exp on fuel = 260. The researcher draws a random sample of 25 families and calc. mean/avg. It comes out to be 330.6. Does that mean the avg. expenditure on fuel has changed?

- The diff (330.6 − 260) is based only on a sample.

Sampling error: Diff. b/w sample statistic and true pop. parameter.

True pop. parameter cannot be known as we can't collect data for an entire pop.

- We obtained a sample mean of 330.6. However, it's conceivable that, due to sampling error, the mean of the pop. is only 260. If the researcher drew another random sample, the next sample mean might be closer to 260.

- Hypothesis testing is used to solve such problems & to determine the likelihood of obtaining a sample mean of 260

- SAMPLING DISTRIBUTION DETERMINES WHETHER OUR SAMPLE MEAN IS UNLIKELY

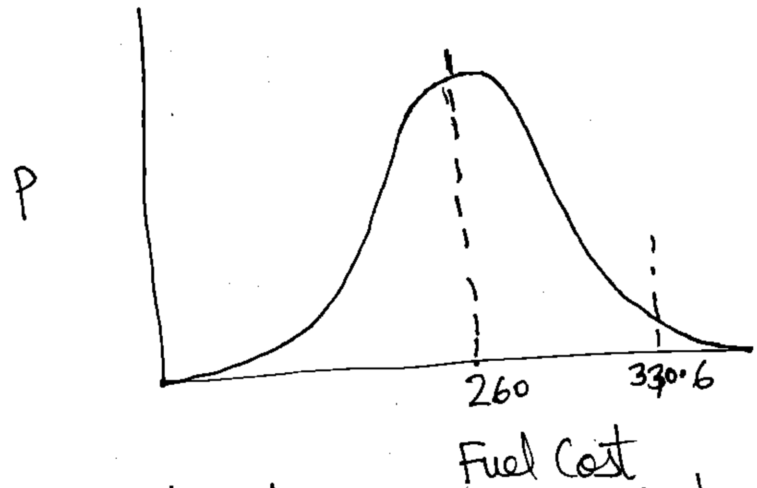
- It is very unlikely for any sample mean to equal the pop. mean bcz of sample error

- If we could obtain a substantial number of random samples and calc. sample mean of each sample and graph the distribution
  — Sampling dist. (of Mean)

- Sampling dist. allow to determine the likelihood of obtaining a sample statistic

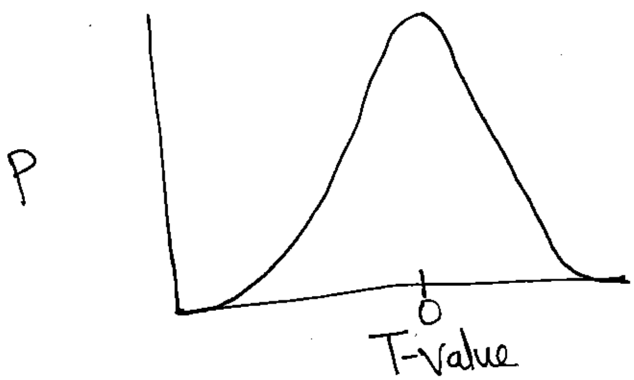
If estimating pop mean from sample mean:
- Sampling dist of mean (sample size < 30) = t-dist
  Sampling dist of mean (sample size > 30) = z-dist

If comparing pop. means from sample mean (in context of hypothesis testing)
- sampling dist of t-statistic/t-score/t-value (sample size < 30) → t-dist
- sampling dist. of z-statistic/z-score/z-value (sample size > 30) → z-dist



- Null distribution: Sampling dist. of test statistic when the null hypothesis is true. Ex. in an F-test, the null dist. is F-dist.
- Prob. distribution (e.g. t-dist, z-dist) display the prob. of obtaining a test statistic when the null hypothesis is true

t-dist



→ Null dist. is the dist. of two sets of data under null hypothesis.

- Hypothesis tests take all of the sample data and convert it to a single value → test statistic (e.g. t statistic, z statistic)
- A test statistic measures the degree of agreement b/w a sample of data and the null hypothesis
- For our example, the hypothesis

Null : Avg fuel cost did not change

$\mu_1 - \mu_2 = 0$ (this year's pop. mean equals null hypothesis mean i.e. 260)

or $\mu_1 = \mu_2$ (corresponds to point 0 on x-axis in t-dist.)

⇓

no effect

Alt : $\mu_1 \neq \mu_2$

or $\mu_1 - \mu_2 \neq 0$

→ Null hypothesis is generally opposite of researcher's hypothesis

→ In hypothesis tests, critical regions are ranges of distribution where the values represent statistically significant results. Defined by
- significance level (α)
- whether the test is one-tailed or two-tailed

→ Significance level (α) is a prob. value that the researcher sets before the study and is the prob. value below which null hypothesis is rejected (e.g. 0.05, 0.01)

→ Since the prob. dist (t-dist, z-dist etc) is the sampling dist. of test statistic assuming the null hypothesis is true

+

the significance level is set by researcher

⇓

The significance level is the prob. of rejecting a null hypothesis that is true

→ When the null hypothesis is rejected, the effect is said to be statistically significant

→ MAPPING CRITICAL REGIONS IN SAMPLING DIST (T-DIST, Z-DIST)

Decide
- Two Tailed Vs One-tailed
- Find out significance level

TWO-TAILED

1) Null (H_0): The effect equals zero

Alt (H_A): The effect does not equal zero

(Test for effect in both dir.)

ONE-TAILED

1.

a) Null: Effect is less than or equal to zero

Alt: Effect is greater than zero

OR

b) Null: Effect is greater than or equal to zero

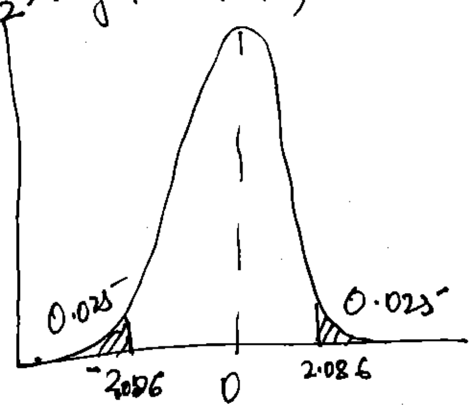
Alt: Effect is less than zero

(Test for effect in one dir)

Note: effect for our example: diff. of means

If point est. is taken into consideration: sample mean = pop mean

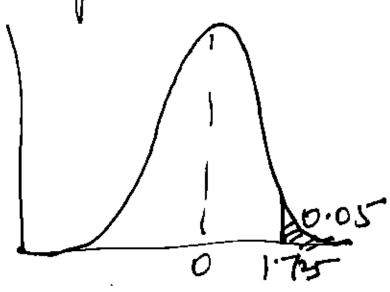
2) Split significance level % b/w both tails of dist ($\alpha/2$ % of the dist)



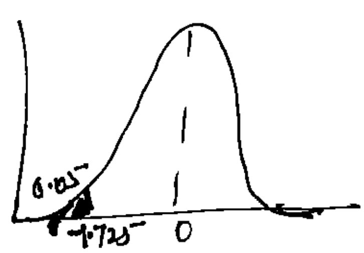
Note 1) Area Under Graph = 1 2) Shaded region = critical region

2) Do not Split significance level % (α% of the dist)

a.)



b)

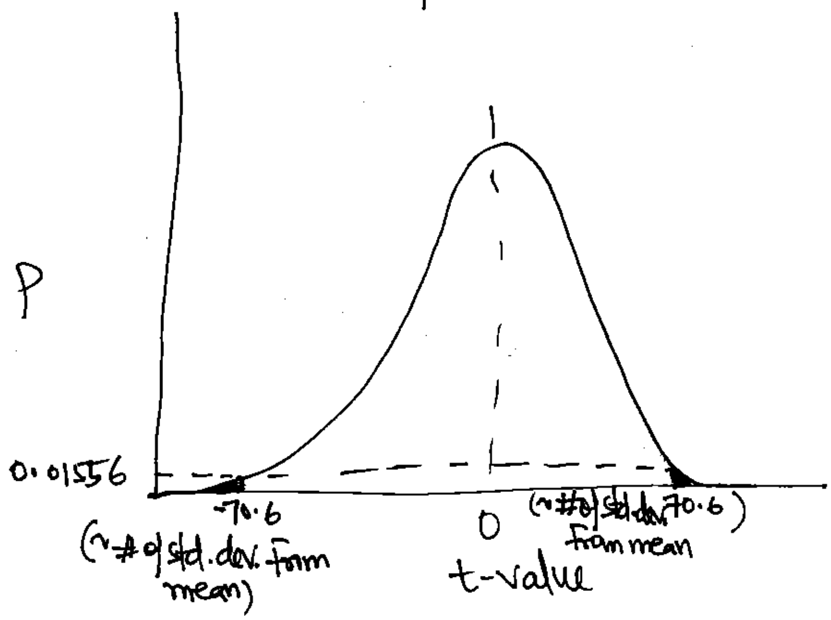


- P-VALUES:

Effect in our sample: 330.6 − 260 = 70.6

Let's shade the region on both sides of the dist. that are at least as far away as 70.6 from 0 (Two-tailed & sample size < 30 so t-dist)
& Find out the prob on y-axis

Note: 70.6 cannot be taken as is to plot on t-dist., need to convert to t-score

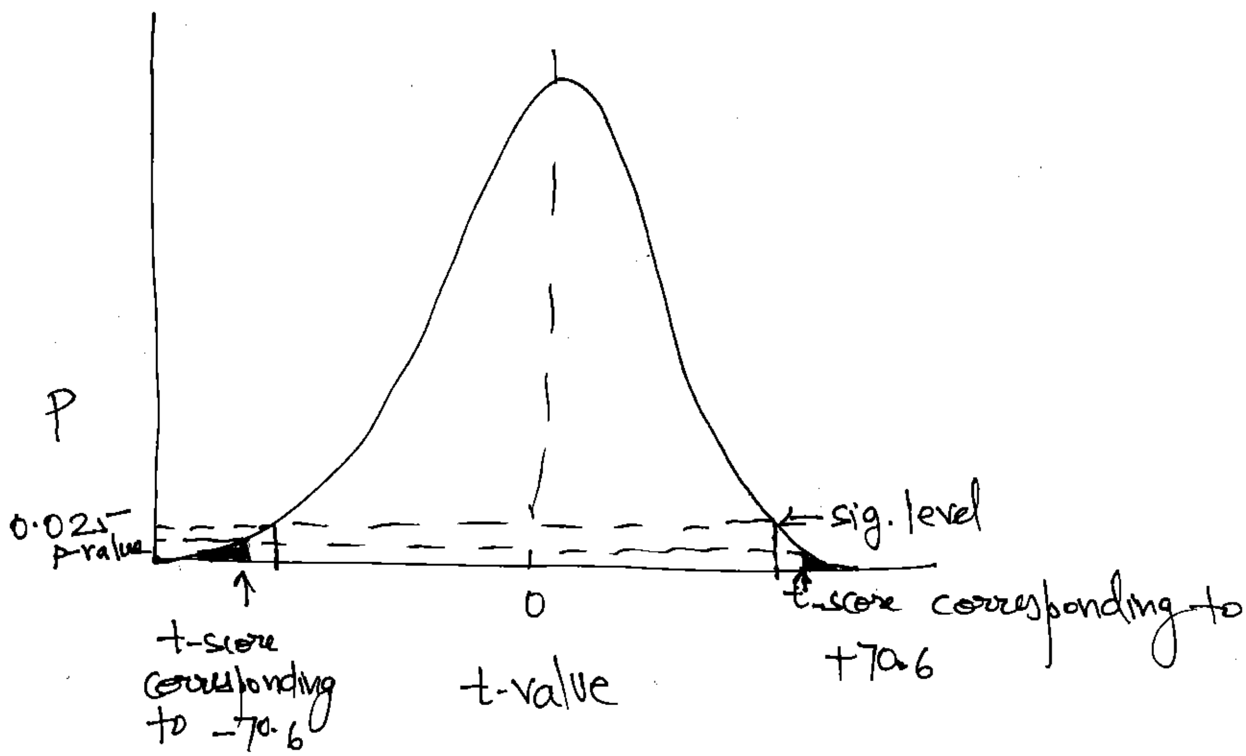


The total prob of two shaded regions = 0.01556 × 2
= 0.03112

If the null hypothesis is true and you drew many random samples, you'd expect sample means to fall in the shaded regions 3.1% of the time ⇒ P-value

P-value: prob. of an outcome, given the hypothesis (assuming true)
& not the prob of hypothesis given the outcome

USING P-VALUE & SIGNIFICANCE LEVEL TOGETHER TO FIND THE RESULT OF HYPOTHESIS TESTING :



✓ Now since P-value < Sig. level ⇒ reject null hypothesis

→ P-value is the prob. of an outcome as seen in Sampling dist. of the ~~test~~ stastistic, if the null hypothesis is true

MISCONCEPTIONS IN INTEPRETING P-VALUES:

× 1) p-value is the prob that null hypothesis is wrong/false
p-value in conjunction with significance level provides the result of hypothesis testing
In hypothesis testing, you either
a) reject null hypothesis → statistically sig
b) fail to reject hypothesis → inconclusive

× 2) Low p-value indicates large effect
Low p-value indicates that the sample outcome would be <u>very unlikely</u> if null hypothesis is true

× 3) High p-value indicates null hypothesis is true
High p-value indicates the data <u>do not conclusively</u> demonstrate that null hypothesis is false

DECISION	NULL HYPOTHESIS TRUE	FALSE
REJECT	α (Type I error) (significance level)	$1-\beta$ (POWER)
FAIL TO REJECT	$1-\alpha$	β (Type II error)

Consequently, the following definitions emerge:

1) Type I error: Rejecting a NULL Hypothesis, when in fact it is true
2) Type II error: Failing to reject a NULL Hypothesis, when in fact it is False
3) Impossible to make Type I error when NULL Hypothesis is False
4) Impossible to make Type II error when NULL Hypothesis is True

POWER: Prob. of correctly rejecting a false NULL Hypothesis
Also, prob that the test can detect an effect that truly exists

FACTORS AFFECTING POWER

$\propto$ sample size

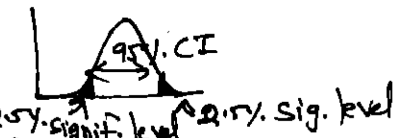
$\propto \frac{1}{\text{std. dev in sample}}$

$\propto$ significance level (e.g. power is lower for 0.01 level than 0.05 level)

$\propto$ effect size (larger delta $\mu_1 >> \mu_2$ ie. diff. b/w true value of parameter and value specified in null hypo.)

→ Confidence level (used in Conf Int.) $=$ $1-\alpha$ (significance level)

This means when significance level $= 0.05 \Leftrightarrow 95\%$ CI



→ DIFFERENT STATISTICAL ~~TESTS~~ METHODOLOGY:

Typically test method involves

i) test statistic (z-test, t-test etc) : computed from sample data e.g. sample mean, proportion, diff b/w means, diff b/w proportions etc.

ii) Sampling dist. of test statistic

Given a test statistic & its sampling dist, a researcher can assess prob. associated with the test statistic. If the test statistic prob. is less than significance level, the NULL Hypothesis is rejected

FREQUENTIST Vs BAYESIAN

F	B
1) Data is varying Parameters are fixed	Data is fixed Parameters have varying char. – prob
2) No prior concept	Prior (Choosing prior is an art based on observed data historically → belief & results can vary bcz of it)
3) Uncertainty Handling ∴ CI If this experiment is repeated multiple times, in 95% of those cases, the computed CI will contain the true parameter	Given our observed data, there is 95% prob. that the value of true parameter lies within credible region

※ Most common mistake: "conf. int." interpretation in Bayesian way

→ Results are similar for simple problems

Prior (may be uniform for simple problem)

$$P(\text{Hypothesis} \mid \text{Data}) = \frac{P(\text{Data} \mid \text{Hypothesis}) \; P(\text{Hypothesis})}{P(D) \; (\rightarrow \text{normalization})}$$

COVARIANCE AND CORRELATION COEFF

$$\rightarrow \mathrm{Cov}(X,Y) = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

where X = independent var
Y = dependent var
n = number of data points in the sample
$\bar{X}$ = mean of X
$\bar{Y}$ = mean of Y

→ Covariance does not standarize the strength of relationship between variables, hence correlation coeff is used → -1 to $+1$ → positive correlation

-1 ⇓ negative correlation; 0 ↓ no correlation

→ Correlation is a measure of linear dependence/association between two variables

$$\text{Correlation Coeff.}\ \gamma_{(X,Y)} = \frac{\mathrm{Cov}(X,Y)}{S_X \, S_Y}$$

Where
S_X = sample standard dev. of X
S_Y = sample standard dev of Y

and $$S_X = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}} \quad \& \quad S_Y = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}}$$

PROBABILITY

1. $0 \leq P \leq 1$

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

   If A & B are mutually exclusive i.e. $P(A \cap B) = 0$

   $P(A \cup B) = P(A) + P(B)$

3. $P(B|A) = \dfrac{P(A \cap B)}{P(A)}$ → joint probability, → marginal prob

   $P(B|A)$ ↓ conditional prob

4. If A & B are independent events then

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

   $\left(P(B|A) = P(B) \Rightarrow P(B) = \dfrac{P(A \cap B)}{P(A)} \Rightarrow P(A \cap B) = P(A) \cdot P(B) \right)$

5. If A & B are collectively exhaustive

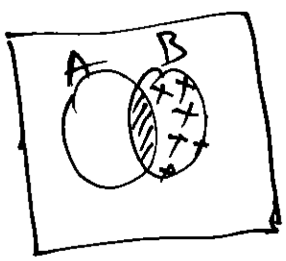
   $P(A) + P(B) = 1$

6. The joint probability of a set of events $A_1, A_2, A_3 \ldots A_n$ in terms of conditional probability

$$P(A_1 \cap A_2 \cap A_3 \ldots \cap A_n)$$
$$= P(A_1)\, P(A_2|A_1)\, P(A_3|A_1 \cap A_2) \ldots P(A_n|A_1 \cap A_2 \cdots \cap A_{n-1})$$

7. Total Probability Theorem:

$$P(B) = P(A \cap B) + P(\bar{A} \cap B)$$
$$= P(B|A)\, P(A) + P(B|\bar{A})\, P(\bar{A})$$



BAYES THEOREM:

$$P(A|B) = \frac{\overbrace{P(B|A)}^{\text{Likelihood Ratio}}\overbrace{P(A)}^{\text{prior prob}}}{P(B)}$$

$P(A|B)$: posterior prob

DERIVATION FROM CONDITIONAL PROBABILITY DEFINITION:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad (1) \Rightarrow P(A \cap B) = P(A|B)\,P(B)$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \Rightarrow P(A \cap B) = P(B|A)\,P(A) \quad (2)$$

Putting (2) in (1)

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

EXPANDED BAYES THEOREM FORMULA:

Since $P(B) = P(A \cap B) + P(\bar{A} \cap B)$

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

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

CONDITIONAL PROB. & BAYES THEOREM PROBLEMS

Ques: Imagine a test with true positive rate of 100%. and false positive rate of 5%. . Imagine a population with 1/1000 rate of having the condition the test identifies. Given a positive test, what is the probability of having that condition?

Sol^n:

$$P(+ve \mid Disease) = 1$$

$$P(+ve \mid \overline{Disease}) = 0.05$$

$$P(Disease) = 0.001 \qquad \therefore P(\overline{Disease}) = 1 - 0.001$$

$$P(Disease \mid +ve) = ?$$

$$P(Disease \mid +ve) = \frac{P(+ve \mid Disease)\, P(Disease)}{P(+ve)}$$

$$= \frac{P(+ve \mid Disease)\, P(Disease)}{P(+ve \mid Disease)\, P(Disease) + P(+ve \mid \overline{Disease})\, P(\overline{Disease})}$$

$$= \frac{1 \times 0.001}{1 \times 0.001 + 0.05 \times 0.999}$$

Ques: A disease test is advertised as being 99% accurate: if you have the disease, you will test positive 99% of the time, and if you don't have the disease, you will test negative 99% of the time. If 1% of all people have this disease and you test positive, what is the probability that you actually have the disease.

Soln:

$$P(+ve \mid Disease) = 0.99$$

$$P(+ve \mid \overline{Disease}) = 0.01$$

$$P(Disease) = 0.01 \qquad P(\overline{Disease}) = 0.99$$

$$P(Disease \mid +ve) = ?$$

$$P(Disease \mid +ve) = \frac{P(+ve \mid Disease)\, P(Disease)}{P(+ve)}$$

$$= \frac{0.99 \times 0.01}{0.99 \times 0.01 + 0.01 \times 0.99} = \frac{1}{2}$$

Ques:

A	B
100 gold coins	50 gold coins, 150 silver coins

You randomly choose a treasure chest to open and then randomly choose a coin from that treasure chest. If the coin you choose is gold, then what is the prob. that you chose chest A.

Sol^n

$$P(A \mid G) = ?$$

$$P(A|G) = \frac{P(G|A) \cdot P(A)}{P(G)}$$

$$= \frac{1 \cdot 1/2}{150/200} = \frac{1/2}{3/4} = \frac{1}{\cancel{2}} \times \frac{\cancel{4}^{2}}{3} = \frac{2}{3}$$

Ques: A diagnostic test has a probability 0.95 of giving a positive result when applied to a person suffering from certain disease and a probability 0.10 of giving (false) positive when applied to a non-sufferer. It is estimated that 0.5% of the population are sufferers. Suppose that the test is now administered to a person about whom we have no relevant information related to the disease. Calculate following probabilities

a) that the test result will be positive
b) that, given a positive result, the person is sufferer
c) that, given a negative result, the person is non-sufferer
d) that the person will be misclassified

Solⁿ: $P(+ve \mid D) = 0.95 \qquad P(+ve \mid \bar{D}) = 0.1$

$P(D) = 0.005 \quad \Rightarrow \quad P(\bar{D}) = 0.995$

a) $P(+ve) = P(+ve \mid D)\, P(D) + P(+ve \mid \bar{D})\, P(\bar{D})$

$= 0.95 \times 0.005 + 0.1 \times 0.995$

$= 0.10425$

b) $P(D \mid +ve) = \dfrac{P(+ve \mid D)\, P(D)}{P(+ve)}$

$= \dfrac{0.95 \times 0.005}{0.10425}$

c) $P(\bar{D} \mid \overline{+ve}) = \dfrac{P(\overline{+ve} \mid \bar{D})\, P(\bar{D})}{P(\overline{+ve})}$

$= \dfrac{(1 - P(+ve \mid \bar{D})) \times P(\bar{D})}{1 - P(+ve)}$

$= \dfrac{(1 - 0.1) \times 0.995}{1 - .10425} = \dfrac{0.9 \times 0.995}{1 - .10425} = 0.997$

d) Misclassified 1) Has disease but test results −ve

2) Does not have disease but test results +ve

$$= P(\overline{+ve} \mid D)\, P(D) + P(+ve \mid \bar{D})\, P(\bar{D})$$

$$= 0.05 \times 0.005 + 0.1 \times 0.995$$

$$= 0.09975$$

Ques: 1) A couple has two children, the older of which is a boy. What is the probability that they have two boys?

2) A couple has two children, one of which is a boy. What is the probability that they have two boys?

Soln 1)

Child Younger	Child Older
G	B
B	B

$= 1/2$

2)

Child 1	Child 2
B	G
G	B
B	B

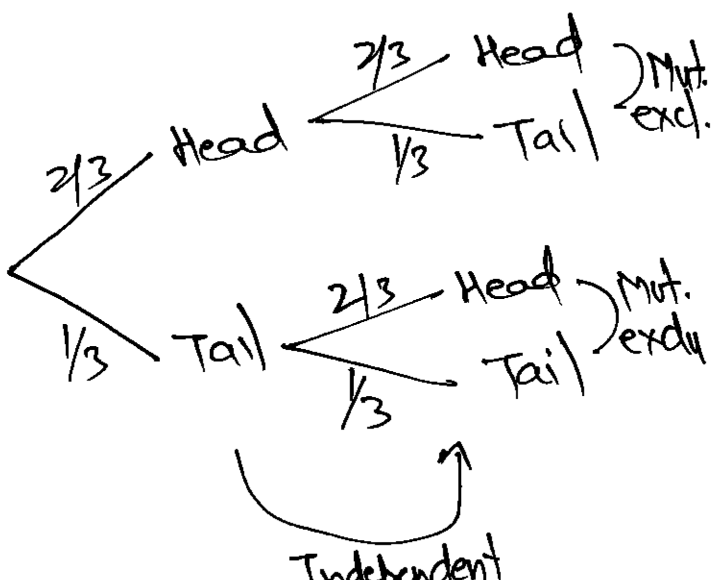
$= 1/3$

PROBABILITY TREE DIAGRAM

A coin is biased so that it is twice as likely to give heads as it is to give tails.

$\Rightarrow P(H) = 2/3 \qquad P(T) = 1/3$

The tree diagram below shows the possible outcomes when this coin is tossed twice



$P(H,H) = 2/3 \times 2/3 = 4/9$

$P(H,T) = 2/3 \times \frac{1}{3} = \frac{2}{9}$

$P(T,H) = 1/3 \times 2/3 = \frac{2}{9}$

$P(T,T) = 1/3 \times \frac{1}{3} = \frac{1}{9}$

Prob. that both tosses give the same result:

$P(H,H) + P(T,T) = 4/9 + \frac{1}{9} = \frac{5}{9}$

Prob that first toss = head & second is tail

$P(H,T) = 2/9$

Prob that both tosses give diff. result:

$P(H,T) + P(T,H) = \frac{2}{9} + \frac{2}{9} = \frac{4}{9}$

JOINT, MARGINAL AND CONDITIONAL PROBABILITY

→ Joint probability distribution of two random variables identifies the probability of any pair of outcomes occuring together

→ suppose random variable X_1 takes on values 1, 2 or 3
random variable X_2 takes on values 1, 2, 3 or 4

X_1 \ X_2	1	2	3	4
1	0.25	0.10	0.05	0.05
2	0.10	0.05	0.00	0.10
3	0.20	0.00	0.10	0.00

← joint prob. distribution

→ Numbers in cells sum to 1

$$\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} P(x_i, x_j) = 1$$

→ Marginal distributions tell you the probability that one random variable takes on any of its values regardless of the value of the other random variable

→

		X_2: 1	2	3	4
X_1	1	0.25	0.10	0.05	0.05
	2	0.10	0.05	0.00	0.10
	3	0.20	0.00	0.10	0.00
Marg. prob dist. of X_2		0.55	0.15	0.15	0.15

} → joint prob. dist

0.55 ↑ (0.25 + 0.10 + 0.20)

		X_2: 1	2	3	4	Marg prob dist of X_1
X_1	1	0.25	0.10	0.05	0.05	0.45 (0.25+0.10+0.05 +0.05)
	2	0.10	0.05	0.00	0.10	0.25
	3	0.20	0.00	0.10	0.00	0.30

→ Conditional Distributions:

Suppose we want to know the probability that X_1 takes on a specific value of 3, conditional on X_2 being equal to 1.

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

$$\therefore P(X_1 = 3 | X_2 = 1) = \frac{P(X_1 = 3 \text{ and } X_2 = 1)}{P(X_2 = 1)}$$

$$= \frac{0.20}{0.55}$$

Cond. dist of X_1 given $X_2 = 1$ ↓

X_1 \ X_2	1	2	3	4
1	0.25/0.55	0.10/0.15	0.05/0.15	0.05/0.15
2	0.10/0.55	0.05/0.15	0.00/0.15	0.10/0.15
3	0.20/0.55	0.00/0.15	0.10/0.15	0.00/0.15

Similarly, we can find cond. dist of X_2 associated with three different values for X_1.