Descriptive Statistics:

Term	Meaning	Population Formula	Sample Formula	Example {1,16,1,3,9}
Mean	Average	$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$	$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n}$	6
Median	The middle value – half are below and half are above			3
Mode	The value with the most appearances			1
Spread		1/n (⅀ x - μ)		
MAD	Mean Absolute Deviation	1/n (⅀ |x - μ|)		
Variance	The average of the squared deviations between the values and the mean	$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2$	$s^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$	$(1-6)^2 + (1-6)^2 + (3-6)^2 + (9-6)^2 + (16-6)^2$ divided by 5 values = 168/5 = 33.6
Standard Deviation	The square root of Variance, thought of as the "average" deviation from the mean.	$\sigma = \sqrt{\sigma^2}$	$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}$	Square root of 33.6 = 5.7966
Coefficient of Variation	The variation relative to the value of the mean		$CV = \frac{s}{\overline{X}}$	5.7966 divided by 6 = 0.9661
Range	Maximum minus Minimum			16 – 1 = 15
IQR	Interquartile Range	3rd quartile – 1st quartile(Dataset has 4 quartiles)	Outliers = Q1 - 1.5*IQR and Q3 + 1.5*IQR	

Linear Transformation:

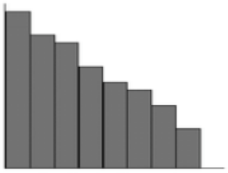
Var(a bX) b^2Var(X)
Average(a bX) a b[Average(X)]

Z-Score(Normalize data - Useful when two datasets with different metrics eg: GRE v/s GMAT):

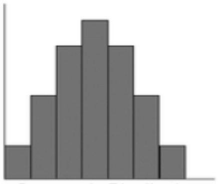
Z = (X-μ)/**σ**
By "z scoring" a data set, the new data set will have **mean 0 and variance 1**.
Different exams but Jason's Z-Score = -1, Mary's Z score = +2 → Jason is 1SD below mean, Mary is 2 SD above
Outlier Detection: Z-score ≥ 2

Dataset Distribution:

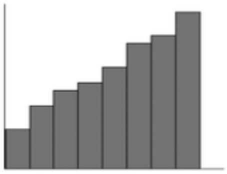
Empirical Rule - Mound Shaped/ Symmetric Data	**Chebyshev's Rule - Any set of data**
X ∓ **σ** → 68% of Data	At least 1 - $\frac{1}{K^2}$ data is in **K** SD
X ∓ 2**σ** → 95% of Data	1 - 1/4 i.e. 75% in 2 SD



Positively Skewed Histogram



Symmetric Distribution Histogram



Negatively Skewed Histogram

Negative Skewed - Skewed Left - Long tail to Left
Positive Skewed - Skewed Right - Long tail to right

Covariance, Correlation and not Causation:
Both measure existence of a **linear relationship**. Does not imply anything about another relationship.
Correlation does not imply causation.

Covariance	Correlation
$S_{xy} = \frac{1}{n-1} \sum (x-\mu_x)(y-\mu_y)$	Correlation coeff $r_{xy} = \sigma_{xy}/(\sigma_x * \sigma_y)$
possible: +ve, -ve, 0 | not possible: weak, strong, slightly Covariance gives you direction, not strength	Possible values: -1 → 1 Direction of relationship: +ve or -ve Strength of relationship: absolute value(-0.6 stronger than 0.4)
Cov(X,X) = Var(X)	$r_{xx} = r_{yy} = 1$
Cov(X,Y) != Cov(Y,X)	$r_{xy} = r_{yx}$

Regression:
One variable is considered the **explanatory variable(x-axis)** and the other the **dependent variable(y-axis)**.
Regression allows one to do predictions which cannot be done with correlation.

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = r\frac{s_y}{s_x}$$

$$b_0 = \bar{Y} - b_1\bar{X}$$

Fitted line: $Y^\wedge = b_0 + b_1X$ ($Y^\wedge$ ⟶ predicted value)
Fitting error: Y-Y^
Stock return: $\alpha + \beta$(Index Return)

Basic Probability:

P(Event Occurs) = $\frac{no.ofsuccessfulevents}{no.ofpossibleoutcomes}$

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
i.e. P(A or B) = P(A) + P(B) – P(A and B)
(Joint Probability) $P(A \cap B) = P(A|B) * P(B) = P(B|A) * P(A)$
If A,B **independent** → $P(A \cap B) = P(A)*P(B)$
→ $P(A|B) = P(A)$
If A,B **mutually exclusive** $P(A \cap B) = 0$

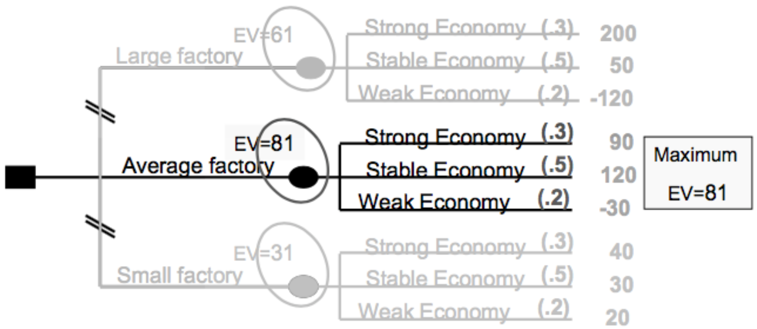
$P(A|B) > P(A)$ ⟶ Positively related
$P(A|B) < P(A)$ ⟶ Negatively related
$P(A|B) = \frac{P(AandB)}{P(B)}$

	B	$\bar{B}$
A	$P(A \text{ and } B)$ $= P(B)P(A \mid B)$	$P(A \text{ and } \bar{B})$ $= P(\bar{B})P(A \mid \bar{B})$
$\bar{A}$	$P(\bar{A} \text{ and } B)$ $= P(\bar{A})P(B \mid \bar{A})$	$P(\bar{A} \text{ and } \bar{B})$ $= P(\bar{A})P(\bar{B} \mid \bar{A})$

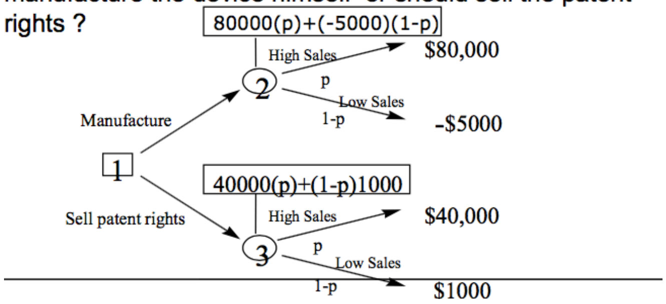
Decision Analysis:

Maximax solution	**Maximin Solution**	**Expected Value Solution**
Best possible scenario. Optimistic and aggressive. Highest payoff.	Worst case scenario.Pessimistic and conservative decision making. Guaranteed minimum payoff.	Probability estimate can be incorporated in search for optimal decision. Weighted average payoff. Never really the outcome
For each option, find the maximum payoff. Choose the option with greatest maximum payoff.	For each option, find the minimum payoff. Choose the option with the greatest maximum payoff.	EV of option = ∑(Prob. Of Outcome * Value of Outcome)

Decision Tree Analysis:



What is the value of the probability of high sales at which it makes no difference whether the engineer should manufacture the device himself or should sell the patent rights ?



The inventor should manufacture the device himself if

$$\begin{aligned}&80000(p)-5000(1-p) > 40000(p)+(1-p)1000\\ \Rightarrow\ &40000(p) > 6000(1-p)\\ \Rightarrow\ &\frac{p}{(1-p)} > 0.15\\ \Rightarrow\ &p > 0.15(1-p)\\ \Rightarrow\ &p > 0.13\end{aligned}$$

Conclusion: if p>.13 then he should manufacture the device himself.

Discrete Random Variables:

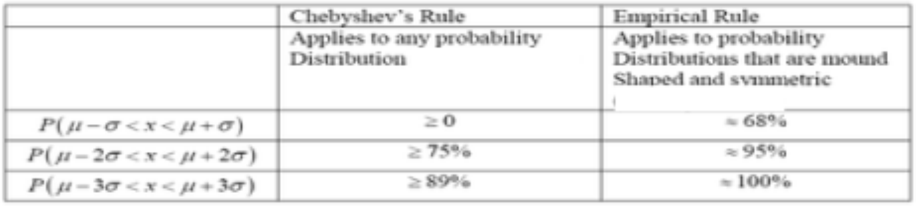
Probability mass function: For most, output will be zero. For specific inputs, some value between 0 and 1.

Cumulative Distribution Function: $F(x) = P(X \le x)$

$\mu_x = \sum x_i * P(x_i)$

$(\sigma_x)^2 = Var(X) = E[(x-\mu_x)^2] = \sum (x-\mu)^2 * P(X=x_i)$

$(\sigma_x)^2 = Var(X) = E[X^2] - (\mu_x)^2 \rightarrow E[X^2] = \sum x_i^2 * P(x_i)$



	Chebyshev's Rule Applies to any probability Distribution	Empirical Rule Applies to probability Distributions that are mound Shaped and symmetric
$P(\mu-\sigma<x<\mu+\sigma)$	≥0	≈68%
$P(\mu-2\sigma<x<\mu+2\sigma)$	≥75%	≈95%
$P(\mu-3\sigma<x<\mu+3\sigma)$	≥89%	≈100%

Binomial Distribution:

Criteria:

Each instance of a trial has only two outcomes, "success" or "failure".
The trials of the experiment are independent of each other.
The probability of a success, p, remains constant from trial to trial
We are interested in the total number of successes (out of n trials)

Not:

The college administration surveys students until they find one not on financial aid. Let the random variable *X* denote the number of students surveyed.

Learnings:

Probability of x successes in n trials =
where p = P(X=x) i.e. P(success)
All successes = p^n
All failures = $(1-p)^n$
$1-p^n$ = negating all successes = at least one failure
$1-(1-p)^n$ = negating all failures = at least one success

Mean: n*p

Var: n*p*q

Shape depends on p: p = 0.5 → symmetric
p = 0.1 → Skewed(tail) to the right

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

Number of ways to have x successes in n trials

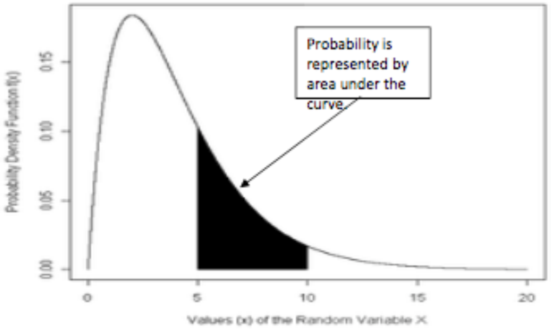
Probability of x successes

Probability of n-x failures

Continuous Random Variables:

P(X=x) = 0 for any x. There are just too many values out there.

CDF: Probability of being less than that value. Used to calculate probability of any interval.

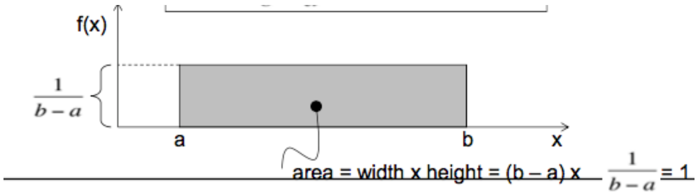


The density curve f(x) is such that f(x)>=0 for all x and the area under the curve=1

$F_X(X \le x) = P(X \le x) = \int_{-\infty}^{x} f(x)dx$ $\quad P(a \le X \le b) = F_X(b) - F_X(a)$

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x)dx \qquad \sigma^2 = Var(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx$$

Uniform Distribution(Example of Continuous RV):



$$f(x) = \begin{cases} \frac{1}{(b-a)} & \text{for x values between a and b} \\ 0 & \text{everywhere else} \end{cases}$$

Mean = E(X) = $\frac{(a+b)}{2}$ Variance = Var(X) = $\frac{(b-a)^2}{12}$

Normal Distribution(Example of Continuous RV):

σ governs height and width.Area has to be 1 so as you get higher, you must get skinnier.

For area under the curve, Z-score → Z = $\frac{(X - mean)}{SD}$

Once it is normalized, we can use the Z-score table that has cumulative probabilities for Z-values

$$P(a < X < b) = P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right)$$

$$P(110 \le X \le 175) = P\left(\frac{110-160}{25} \le \frac{X-160}{25} \le \frac{175-160}{25}\right)$$
$$= P(-2 \le Z \le 0.6)$$

Normal distribution is mound shaped: Empirical rules from Discrete RV apply
Reverse look up: Need to find top 10% of population? Find Z where the z-score is 0.9 i.e. z = 1.28 ➞ find x