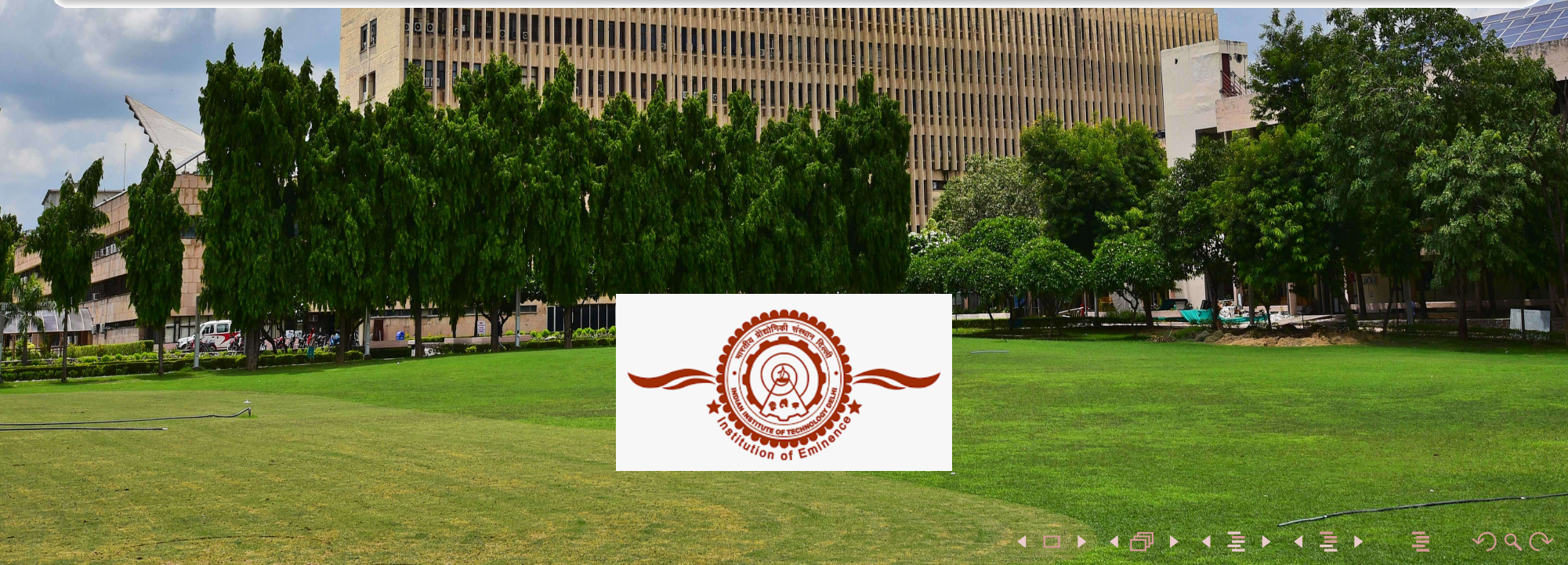
Certificate Course on Data Science and Machine Learning
Measures and descriptors of data

Hariprasad Kodamana, Manojkumar Ramteke, Agam Gupta
IIT DELHI

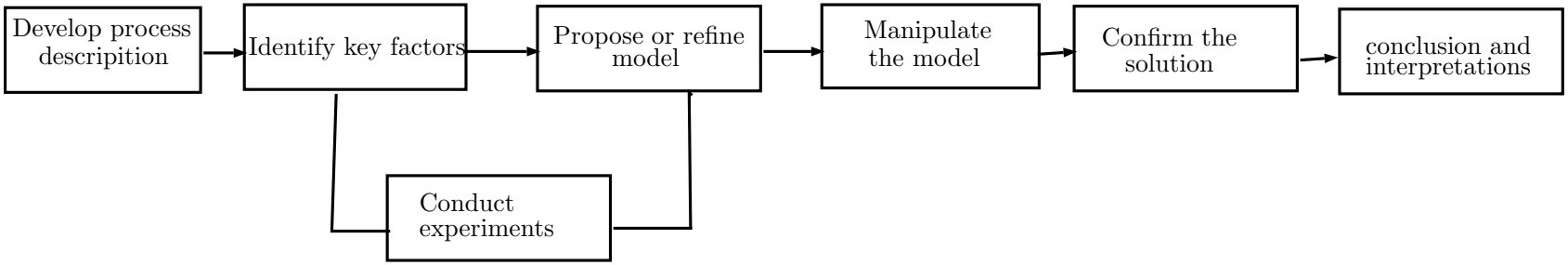


Overview of Presentation

1. Preamble
2. Data visualization
3. Description of data
4. Outlier detection
5. Normal distribution and its derivatives
6. backup for review

Preamble

- Data driven solution strategy

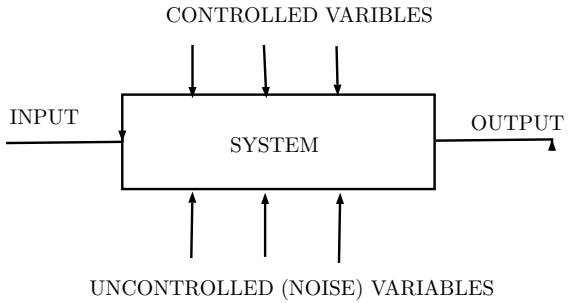


- Statistics-science of data $\rightarrow$ deals with collection, presentation, analysis and use of data
- Helps to describe/quantify the variability
- Eg: An engineer designed and fabricated a nozzle for a desired flow of 12.5 m^3/s. However, when tested, he observed that the flow rates are different : viz. 12.6, 12.9, 13.1, 12.1, 12.3, 12.4 m^3/s.
  Observation: there is a variability in flow rate, and it can be treated as a **random variable (RV)**

$$X = \mu + \epsilon \text{ here } \mu\text{=12.5, } \epsilon\text{-unceratinity,}$$

preamble(cont...)

- Every system is prone to be affected by uncertainties due to presence of uncontrolled variables



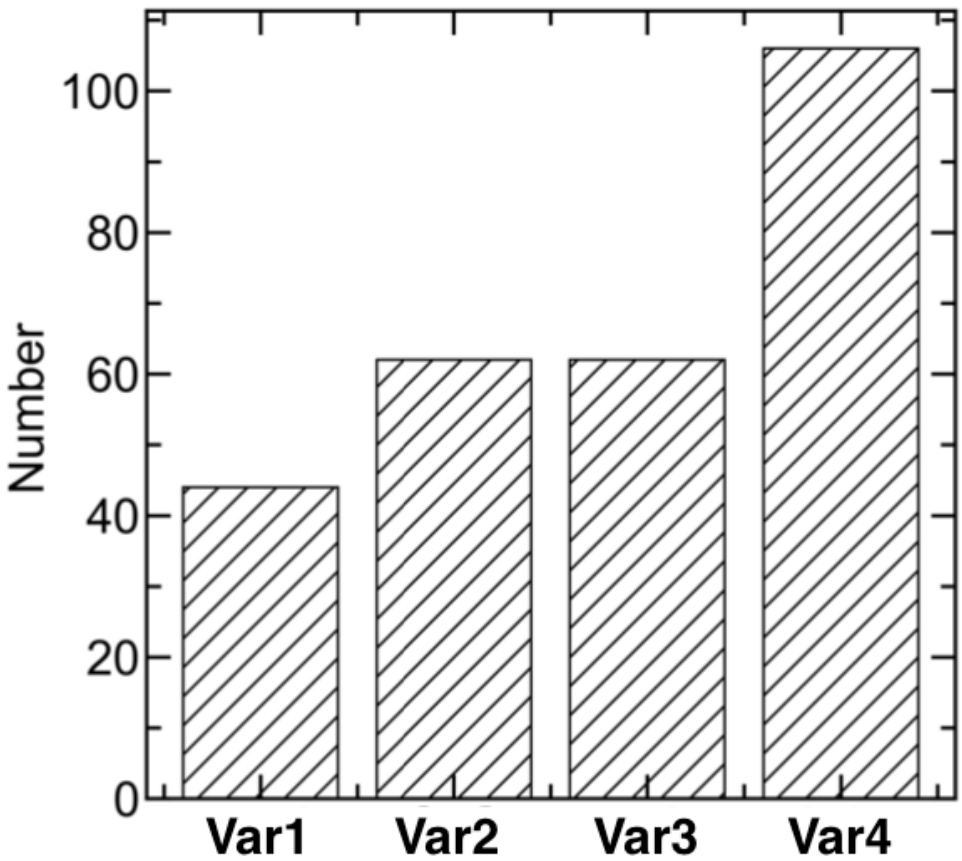
- **Random experiment**-an experiment that can result in different outcomes even though it is repeated in the same manner
- **Sample space**-the set of possible outcomes of a random experiment

  Eg: Time take to cover Delhi to Gurugram by car
- Discrete sample space-finite or countably infinite outcomes, continuous sample space- interval or real numbers
  1. $S = \mathcal{R}^+ = \{x | x > 0\}$
  2. $S = \{x | 0.5 < x < 2\}$
  3. $S = \{\text{less, medium, more}\}$

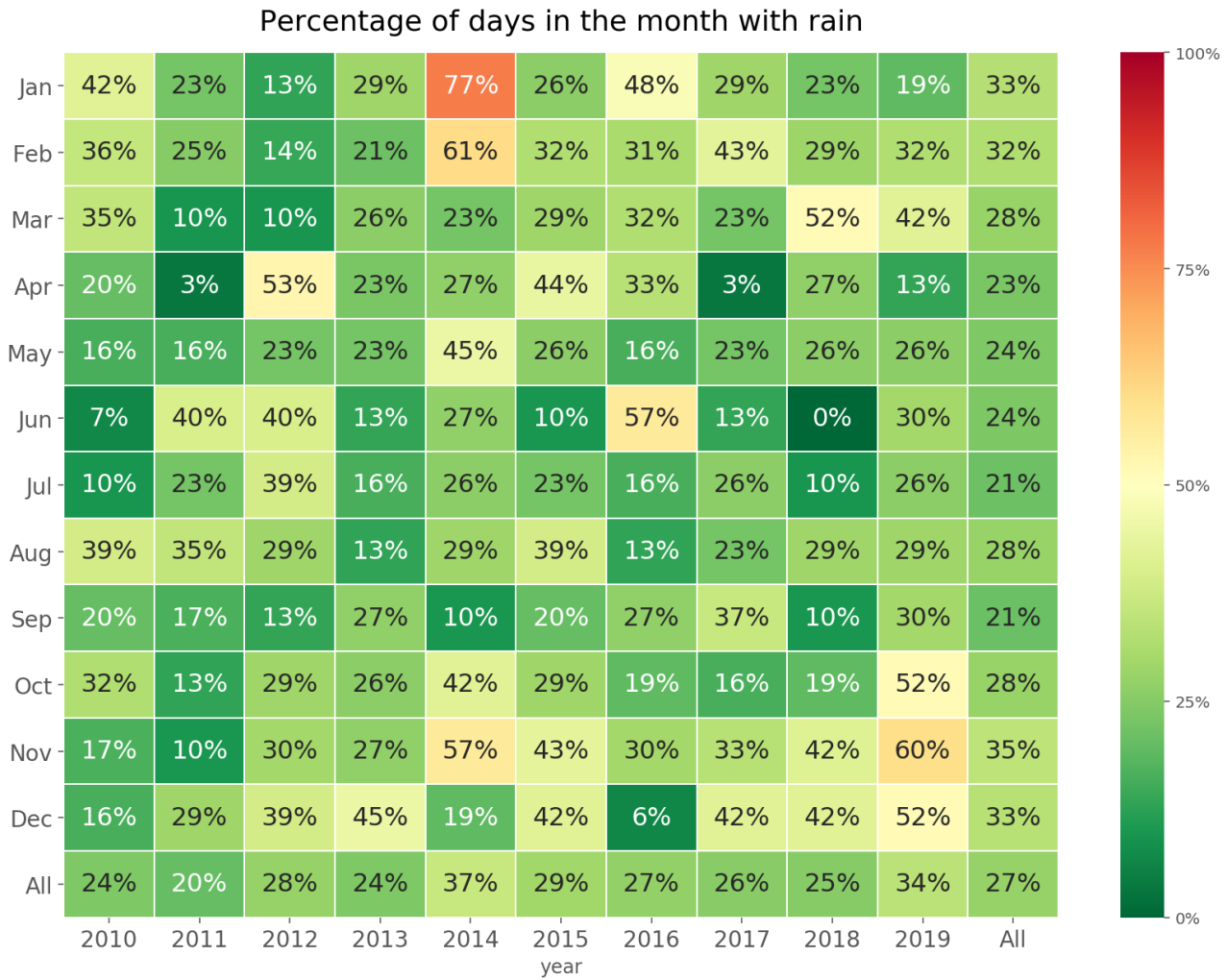
Overview of Presentation

1. Preamble
2. Data visualization
3. Description of data
4. Outlier detection
5. Normal distribution and its derivatives
6. backup for review

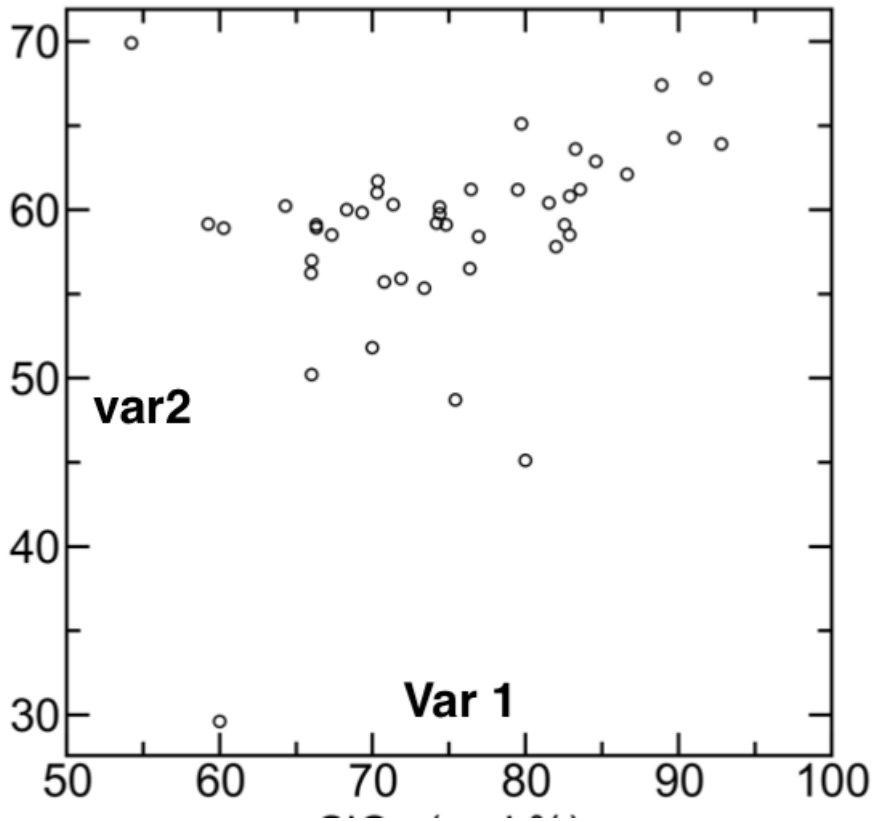
Bar plot



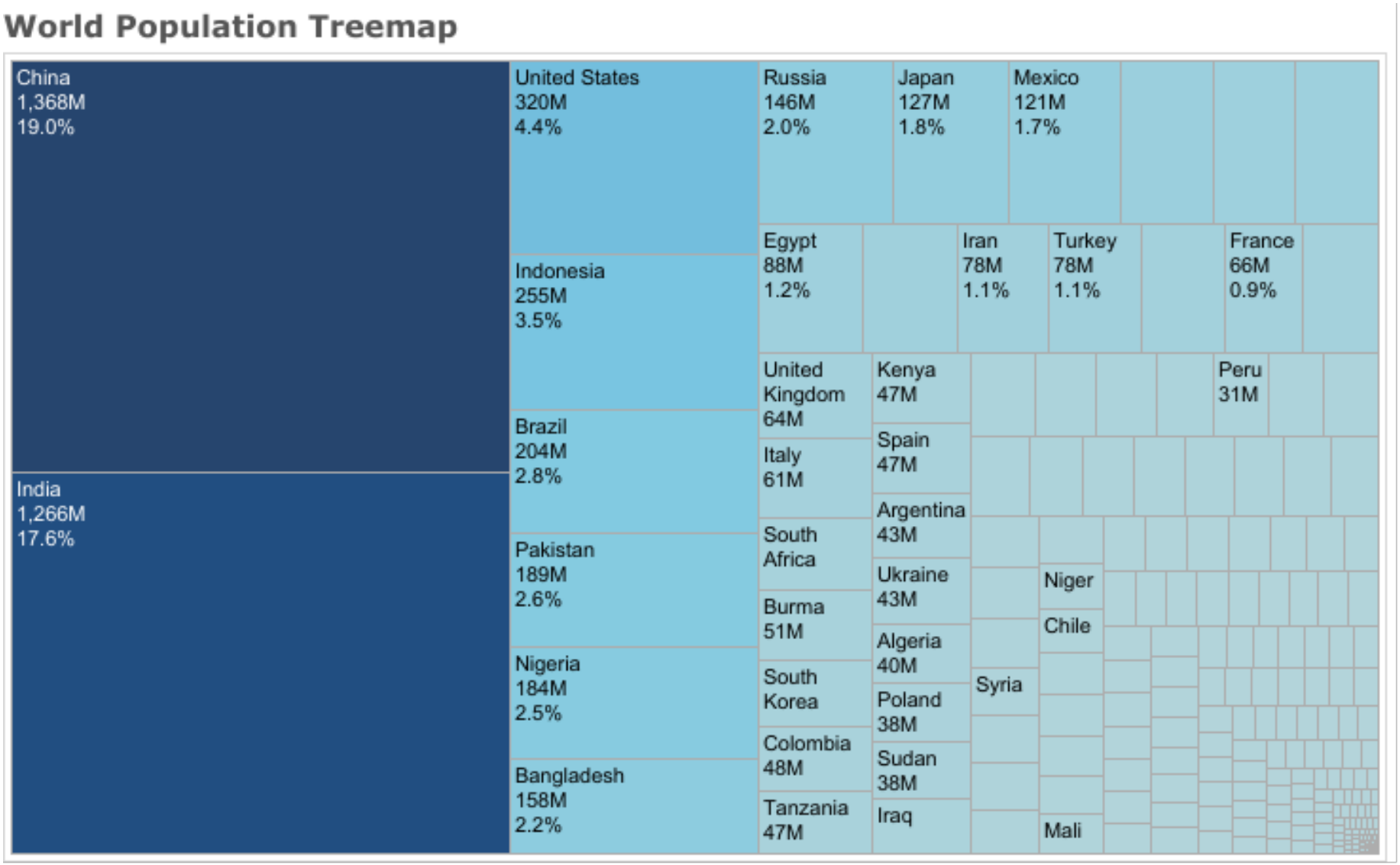
Heat map


Percentage of days in the month with rain
Jan 42% 23% 13% 29% 77% 26% 48% 29% 23% 19% 33%
Feb 36% 25% 14% 21% 61% 32% 31% 43% 29% 32% 32%
Mar 35% 10% 10% 26% 23% 29% 32% 23% 52% 42% 28%
Apr 20% 3% 53% 23% 27% 44% 33% 3% 27% 13% 23%
May 16% 16% 23% 23% 45% 26% 16% 23% 26% 26% 24%
Jun 7% 40% 40% 13% 27% 10% 57% 13% 0% 30% 24%
Jul 10% 23% 39% 16% 26% 23% 16% 26% 10% 26% 21%
Aug 39% 35% 29% 13% 29% 39% 13% 23% 29% 29% 28%
Sep 20% 17% 13% 27% 10% 20% 27% 37% 10% 30% 21%
Oct 32% 13% 29% 26% 42% 29% 19% 16% 19% 52% 28%
Nov 17% 10% 30% 27% 57% 43% 30% 33% 42% 60% 35%
Dec 16% 29% 39% 45% 19% 42% 6% 42% 42% 52% 33%
All 24% 20% 28% 24% 37% 29% 27% 26% 25% 34% 27%
2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 All
year
100%
75%
50%
25%
0%

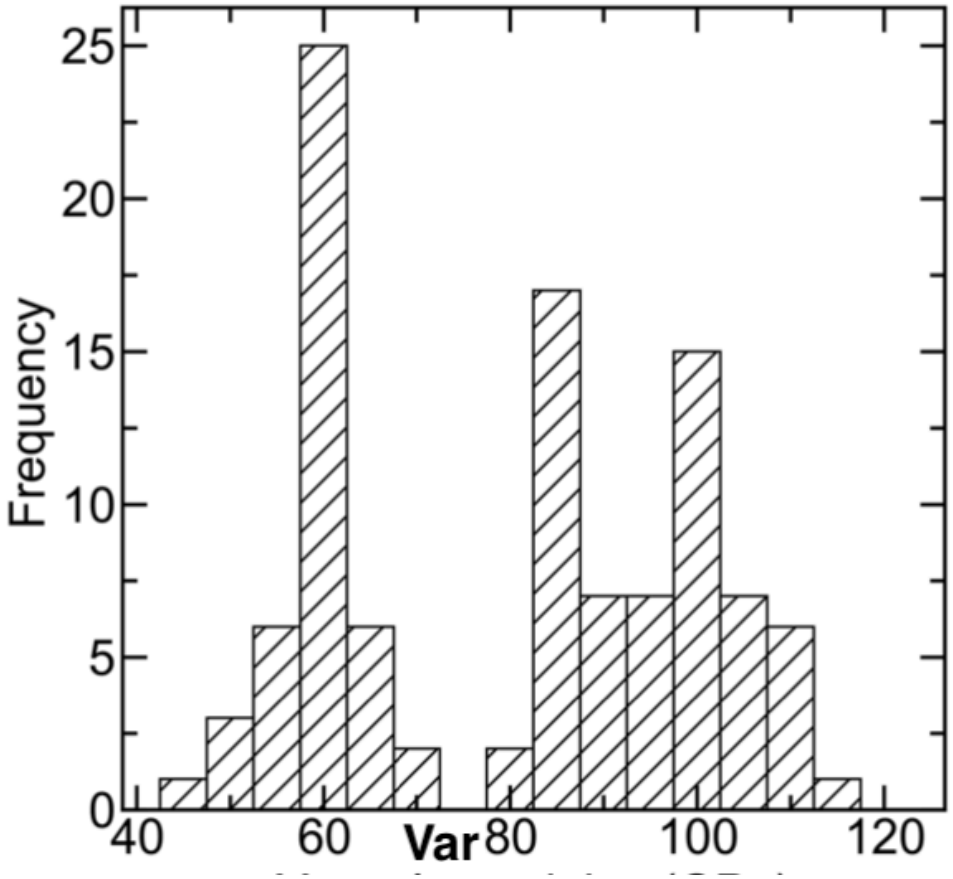
Scatter plot



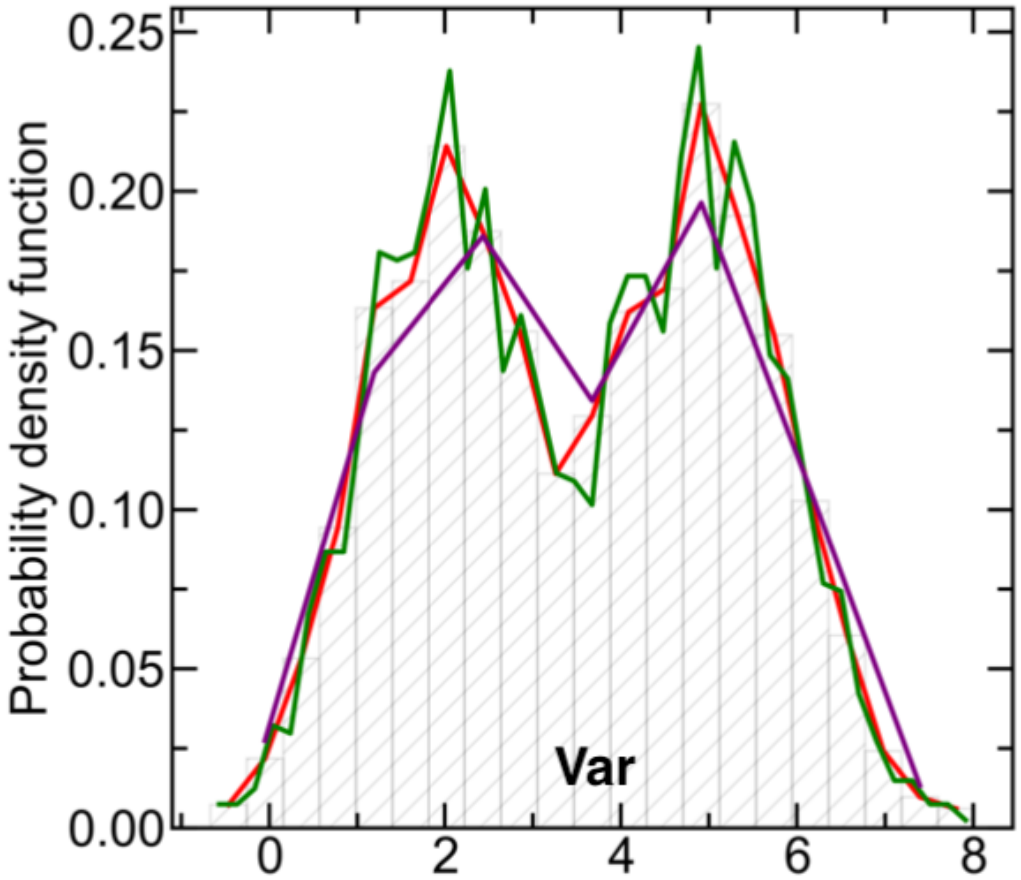
Tree map


World Population Treemap
China
1,368M
19.0%
India
1,266M
17.6%
United States
320M
4.4%
Indonesia
255M
3.5%
Brazil
204M
2.8%
Pakistan
189M
2.6%
Nigeria
184M
2.5%
Bangladesh
158M
2.2%
Russia
146M
2.0%
Japan
127M
1.8%
Mexico
121M
1.7%
Egypt
88M
1.2%
Iran
78M
1.1%
Turkey
78M
1.1%
France
66M
0.9%
United Kingdom
64M
Italy
61M
South Africa
Burma
51M
South Korea
Colombia
48M
Tanzania
47M
Kenya
47M
Spain
47M
Argentina
43M
Ukraine
43M
Algeria
40M
Poland
38M
Sudan
38M
Iraq
Peru
31M
Syria
Niger
Chile
Mali

Histogram



Distribution plots



Overview of Presentation

1. Preamble
2. Data visualization
3. Description of data
4. Outlier detection
5. Normal distribution and its derivatives
6. backup for review

Sample and Population

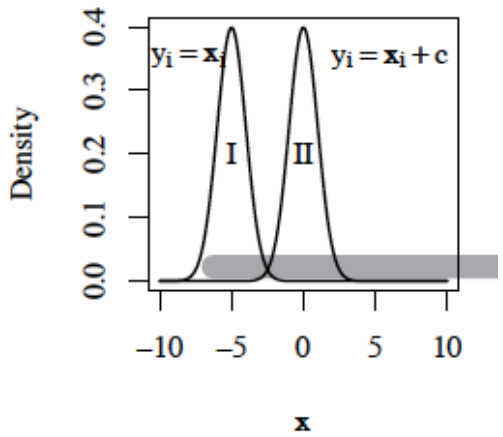
- Population: the set of all the observations of the characteristics under investigation: may be finite or infinite.
- Population: It can be thought as the set of all possible outcomes of the random variable. eg: all possible values of the speed recorded in the previous example, valid electorate in the country
- We wish to obtain parameters of the population model Eg. mean μ, variance σ (by convention Greek alphabets)
- Sample: A part or subset of population. eg: some experimental values of of speed as in the previous example, people who participate in exit polls

Statistic

- Statistic: A numerical quantity whose value depends on the collected data samples Eg. sample mean $\overline{x}$, sample variance S^2 (by convention English alphabets)
- Given a data set, we are interested in statistics which describe the central value, the spread in the data, and any other asymmetry in the data.
- Consider n samples from a population and calculate its statistic P_1. Now change the samples with the same sample size and calculate its statistic. It would be possibly different and let it be P_2. Now if we repeat this exercise, we may obtain different statistic P_3. Hence, the statistic P_i, $i = 1, \ldots, n$ is a RV
- We are always interested to obtain the population model parameter θ as an estimate $\hat{\theta}$ from the samples. Eg. True model for population: $y = \theta x + c$, Estimated model from samples: $\hat{y} = \hat{\theta} x + c$

Measures of centrality

- Population mean $\mu = \frac{1}{N}\sum_N x_i$, N-population size (usually unknown)
- Sample mean $\overline{x} = \frac{1}{n}\sum_n x_i$, n-sample size
- Mostly μ will not never be equal to $\overline{x}$
- Translation: if $y_i = x_i + c$, where c is a constant, then, $\overline{y} = \overline{x} + c$



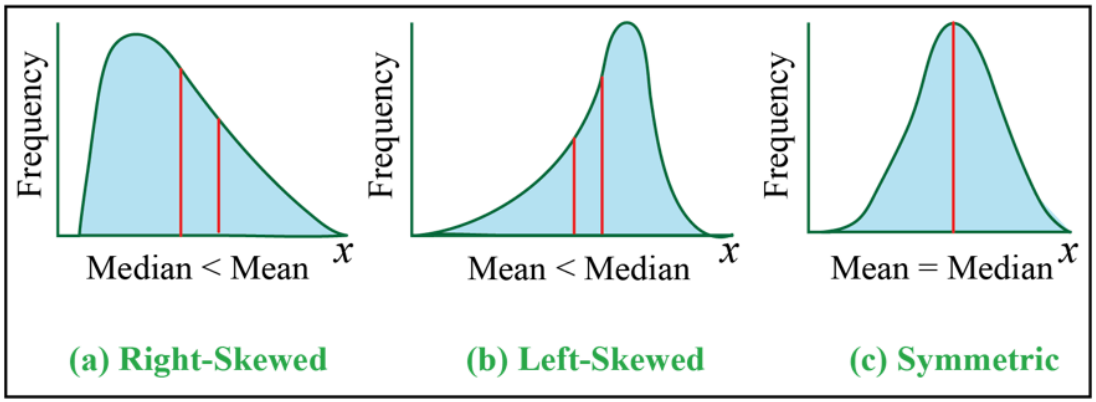
- Scaling: if $y_i = cx_i$ where c is a constant, then, $\overline{y} = c\overline{x}$

Measures of centrality (contd...)

- Median: For n observations, the sample median is $\frac{n+1}{2}$th largest observation for odd n, mean of $\frac{n}{2}$th and $\frac{n}{2}+1$th largest observations for even n.
- Hence, median is the value that splits the dataset into half.
- Due to this property, outliers present in the data may not affect the location of the median significantly.

Measures of centrality (contd...)

- Mode: the value which occurs with the greatest frequency. If each x_i is unique, it is difficult to represent.
- Typically, Mode is the best central measure while dealing with categorical or discrete data.



Measures of centrality (contd...)

Q A cricketer's scores in five ODI matches are as follows: 12, 34, 45, 50, 24

A 33

Q Find the median of number of wickets taken in by a bowler five ODI matches are as follows: 4, 4, 6, 3, 2.

A Let's arrange this data in ascending order: 2, 3, 4, 4, 6. Thus, median 4

Q For the data 6, 8, 9, 3, 4, 6, 7, 6, 3, compute the mode

A The value 6 appears the most number of times. It is the mode

Median and mode for continuous data

- When the data is continuous, mean/ median/ mode can be found using the following steps:
- mean $=\frac{f_1 m_1+\cdots+f_n m_n}{N}$; N total number of samples, m_i mid point of the range

1. Find median/modal class
2. Find median/mode using the following formulae:

- Mode$=l+\frac{(f_m-f_1)}{(2f_m-f_1-f_2)}h$
  Here, l=lower limit of modal class (i.e. the class with maximum frequency), f_m=frequency of modal class, f_1= frequency of class preceding modal class, f_2 frequency of class succeeding modal class, h class width
- Median$=l+\frac{(n/2)-c}{f}h$
  Here, l= lower limit of median class (the class where $n/2$ lies), c =cumulative frequency of the class preceding the median class, f= frequency of the median class, h=class size

Example: mean, mode and median calculation

Height	120-130	130-140	140-150	150-160	160-170	Total
frequency	2	8	12	20	8	50

- mean $= \frac{(\frac{120+130}{2} \times 2 + \cdots + \frac{160+170}{2} \times 8)}{2+\cdots+8}$
- Modal class =150-160 [as it has maximum frequency], l=150, h=10, f_m= 20, f_1=12, f_2=8, $\therefore$ Mode =154

Class Intervals	No.(f_i)	Cumulative frequency (c)
120-130	2	2
130-140	8	2+8=10
140-150	12	10+12=22 (c)
150-160	20 = f	22+20=42
160-170	8	42+8=50 (n)

- $n = 50$, $n/2 = 25$, Median class = 150-160, $c = 22$, $f = 20$, $h = 10$, $\therefore$ Median=151.5

Measures of spread

- Range: the difference between largest and smallest observations in a sample set. It is easy to compute and is sensitive to outliers.
- Percentile:The p^{th} percentile is that threshold such that $p\%$ of observations are at or below this value
- It is $(k+1)^{th}$ largest sample point if $np/100 \neq integer$, where k = largest integer less than $np/100$, or average of $(np/100)^{th}$ and $(np/100+1)^{th}$ values if $np/100 = integer$.

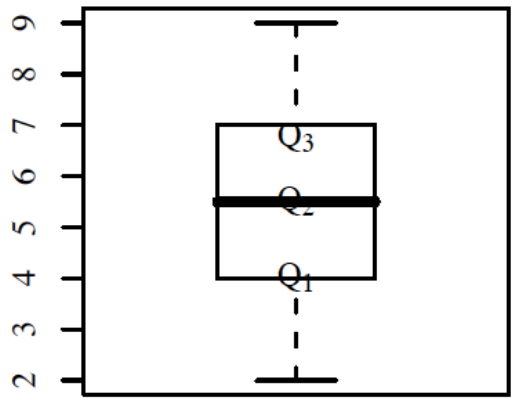
Q If $n = 25$; how (i) $p = 10$ and (ii) $p = 80$ is calculated

sol (i) $np/100 = 2.5$, i.e. the third smallest value. (ii) $n = 25$; $p = 80$, then $np/100 = 20$ i.e. average of the twentieth and twenty
first smallest values.

- First quartile ($Q1$)-25 p, Second quartile ($Q2$)-50 p (also the median), Third quartile ($Q3$)-75 p

Measures of spread (contd..)

- Box plot: A convenient graphic to represent range, median and quartiles



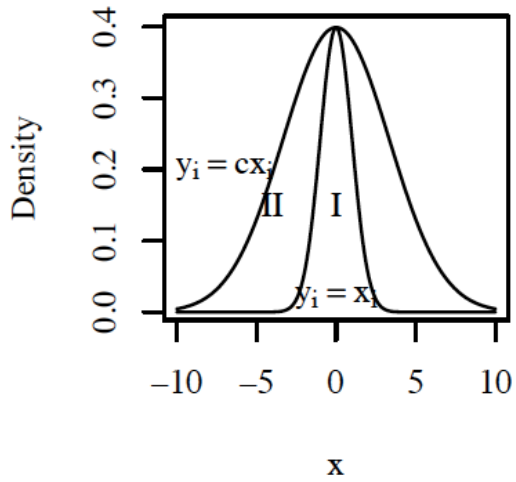
- A box is drawn from $Q1$ to $Q3$, $Q2$ (median) is drawn as a vertical line in the box, and outer lines are drawn either up to the outermost points, or at $1.5 \times (Q3 - Q1)$, and the line length represents the range

Measures of spread (contd..)

- Average deviation from mean $\sum_{i=1}^{n} \frac{(x_i - \overline{x})}{n} = 0$, hence not useful
- Mean absolute deviation: $\sum_{i=1}^{n} \frac{(|x_i - \overline{x}|)}{n}$, does capture spread, but do not reflect the bell shaped curve
- Population variance $\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}$, again N is unknown
- Sample variance $S^2 = \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n-1} = \sum_{i=1}^{n} \frac{(x_i - \frac{\sum_{i=1}^{n} x_i}{n})^2}{n-1}$
- A large S^2 implies large variability.
- $S^2 > 0$
- Standard deviation: Square root of variance
- For translated data: $y_i = x_i + c$, where c is a constant, $S_y^2 = S_x^2$

Measures of spread (contd..)

- For scaled data: $y_i = cx_i$, here c is a constant, $S_y^2 = c^2 S_x^2$



- Coefficient of variation: defined as $CV = 100\% \frac{S}{\bar{x}}$

  This metric is dimensionless, hence one can discuss S relative to the mean's magnitude and useful when comparing different sets of samples with different means, with larger means usually having higher variability.

Measures of spread (contd..)

Result

If a random sample of size n is taken from an infinite population with mean μ and variance σ^2, then probability distribution of sample mean $\overline{x}$ has mean μ and has variance of $\frac{\sigma^2}{n}$

Steps

Consider the samples $x_i (i = 1, ..., n)$. By definition, $E(x_i) = \mu$. Then, $E(\overline{x}) = E(\frac{1}{n}\sum_{i=1}^{n} x_i) = \frac{1}{n}E(x_1) + \cdots + \frac{1}{n}E(x_n) = \frac{1}{n}\mu + \cdots + \frac{1}{n}\mu = n \times \frac{1}{n}\mu = \mu.$

Consider the samples $x_i (i = 1, ..., n)$. By definition, $E(x_i) = \mu, V(x_i) = \sigma^2.$ $V(\overline{x}) = V(\frac{1}{n}\sum_{i=1}^{N} x_i) =$ $V(\frac{1}{n}x_1) + \cdots + V(\frac{1}{n}x_n) = \frac{1}{n^2}\sigma^2 + \cdots + \frac{1}{n^2}\sigma^2 = n \times \frac{1}{n^2}\sigma^2 = \frac{\sigma^2}{n}$ *(Hint.* $V(cX) = c^2 V(X)$*)*

Measures of distortion

- Two higher-order measures of data representation are skewness and kurtosis. While skewness is a measure of the distortion, kurtosis is a measure heavy-tailed nature of the data relative to a Normal distribution.
- Skewness measure the degree of distortion of the data from the normal distribution. A symmetrical distribution will have a skewness of 0. Skewness is calculated by

$$G_1 = \frac{\sqrt{n(n-1)}}{n-2} \sum_{i=1}^{n} \frac{(x_i - \overline{x})^3}{nS^3} \tag{1}$$

- If the $-0.5 \leq G_1 \leq 0.5$ then data are fairly symmetrical. If the $G_1 < -0.5$ it is called negatively skewed while if the $G_1 > 0.5$ it is called positively skewed.

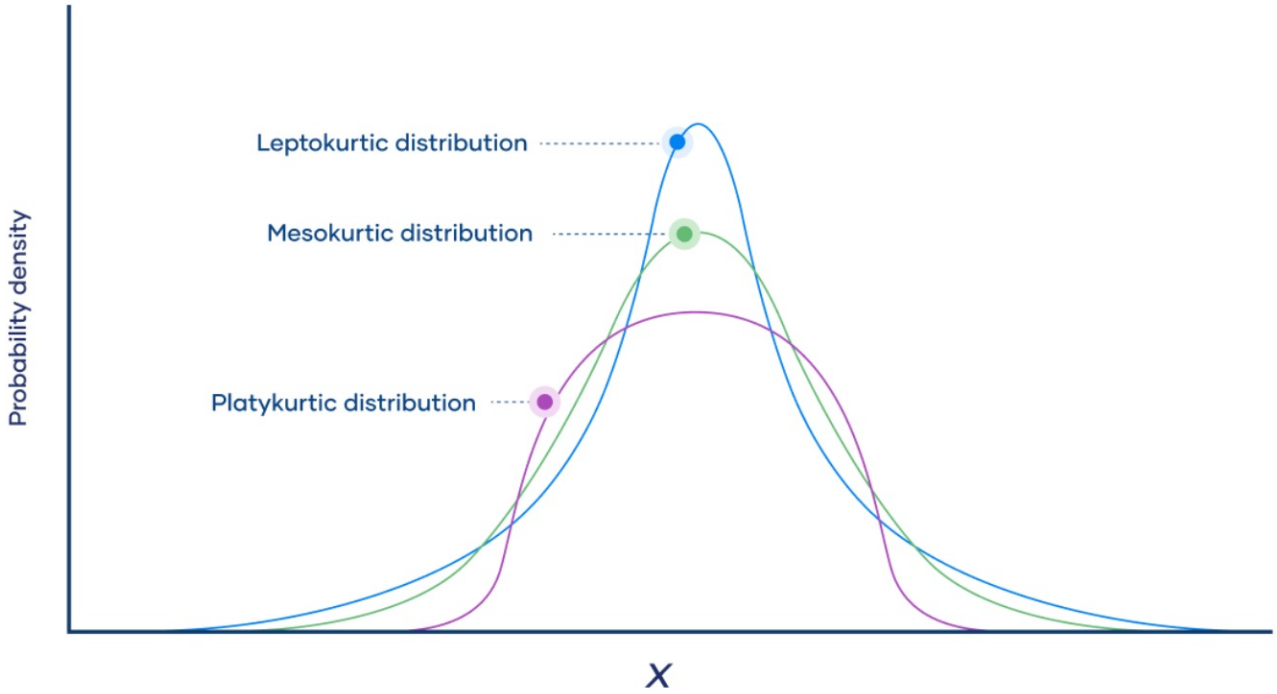
Measures of distortion

- Kurtosis is used to describe the extreme values in one versus the other tail and therefore, it is a measure of outliers present in the distribution. Kurtosis is calculated as:

$$(Excess)\ Kurtosis = \sum_{i=1}^{n} \frac{(x_i - \overline{x})^4}{nS^4} - 3 \tag{2}$$

- High value of kurtosis in a data set is an indicator that data has heavy tails or outlier and vice versa.
- Mesokurtic - medium tailed - Normal
- Platykurtic - thin tailed - Uniform
- Leptokurtic - fat hailed - Laplace

Measures of distortion



	Mesokurtic	Platykurtic	Leptokurtic
Tailedness	Medium-tailed	Thin-tailed	Fat-tailed
Outlier frequency	Medium	Low	High
Kurtosis	Moderate (3)	Low (< 3)	High (> 3)

Overview of Presentation

1. Preamble
2. Data visualization
3. Description of data
4. Outlier detection
5. Normal distribution and its derivatives
6. backup for review

Outlier detection : standard deviation approach

- We first calculate the mean and standard deviation of the data. A data point is identified as an outlier, if it is away from the mean by a pre-specified threshold in terms of the the standard deviations.
- That is, if a data point x_i satisfies calculated the Z score,

$$\frac{|x_i - \overline{x}|}{S} > k, \text{where, } k = 1, 2, \text{ or } 3 \tag{3}$$

  then it is detected as an outlier.
- In other words, a data point is an outlier if it beyond $\overline{x} + kS$. For a normally distributed dataset, $1S$, $2S$, and $3S$ rerpresents, 68.27%, 95.45%, and 99.73%, respectively, of the dataset.
- However, this method can fail to detect outliers if the S is large.

Outlier detection : Median Absolute Deviation (MAD) approach

- Median is a central of measure of data that is less susceptible to outliers.
- Median Absolute Deviation (MAD) is calculated as the median of absolute difference between each point and the median as:

$$MAD = median(|x_i - M|),\ 1, \ldots, n \tag{4}$$

- Then the modified Z_M- score is calculated using MAD values as:

$$Z_M = \frac{\overbrace{0.6745}^{75p\ of\ \mathcal{N}}\,(x_i - M)}{MAD} \tag{5}$$

- As a rule of thumb, if Z_M is greater than 3, an outlier is detected.

Outlier detection : interquartile approach

- The interquartile range (IQR) is calculated the same way as the range.
- IQR is computed by subtracting the first quartile from the third quartile:

$$IQR = Q_3 - Q_1 \tag{6}$$

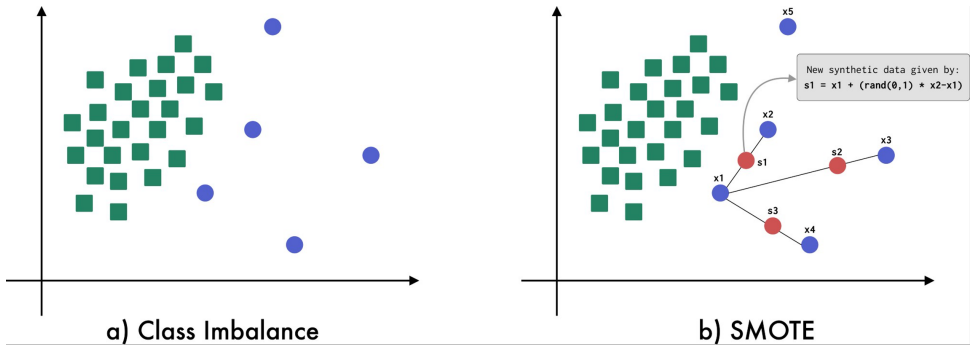
- IQR can be used to detect outliers as follows: if any data point x_i if lies outside

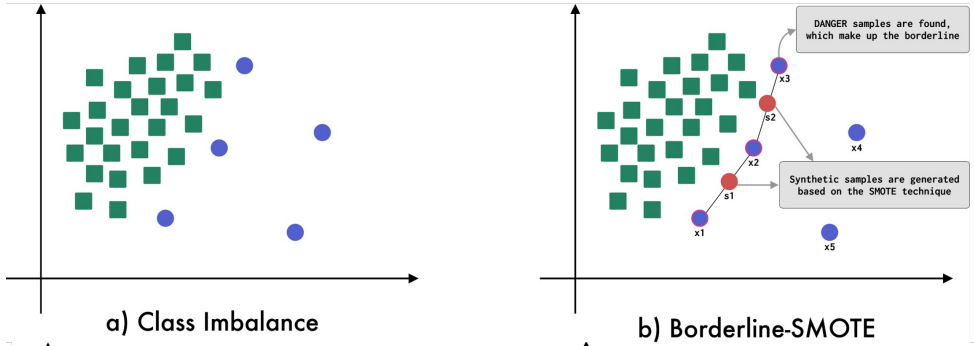
$$Q_3 + 1.5IQR < x_i < Q_1 - 1.5IQR \tag{7}$$

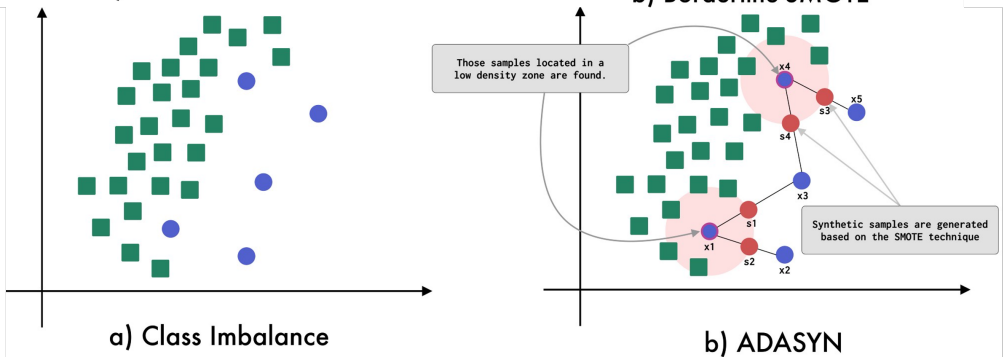
can be a potential outlier.

Imbalanced data

- Under sampling
- Over sampling
  - SMOTE
  - Borderline SMOTE
  - ADASYN







Covariance

- Covariance of two RVs X, Y defined as: $Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$, where $\mu_X,\ \mu_Y$ are the means of $X,\ Y$, respectively
- $Cov(X, Y) = E(XY) - E(\mu_X Y) - E(\mu_Y X) + \mu_X \mu_Y = E(XY) - \mu_X \mu_Y$
- $Cov(X, Y) = Cov(Y, X)$
- $Cov(X, X) = Var(X)$
- If X, Y are independent RVs, then $Cov(X, Y) = 0$

Correlation

- $+$ covariance $\implies$ $+$ relationship: X tends to increase with increasing Y and vice-versa.
- Similarly for negative covariance
- Magnitude of covariance is sensitive to units (scaling) of X, Y
- Define correlation between X, Y as: $Corr(X, Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$
- $-1 \leq Corr(X, Y) \leq 1$
- $Corr(X, Y)$ is dimensionless

Moments

- n^{th} moment (about the origin) of a random variable X defined as: $E[X^n]$
- $\mu = E(X)$ is the first moment about origin and indicates the mean value X can take (It is the centre of gravity of the density function.)
- $Var(X) = E[(X - \mu)^2]$ is the second moment about the mean and indicates the spread (variation) in the value that X will take around the mean.
- Third moment - Skewness
- Fourth moment - Kurtosis
- Fifth moment - Hyperskewness
- Sixth moment - Hypertailedness

Overview of Presentation

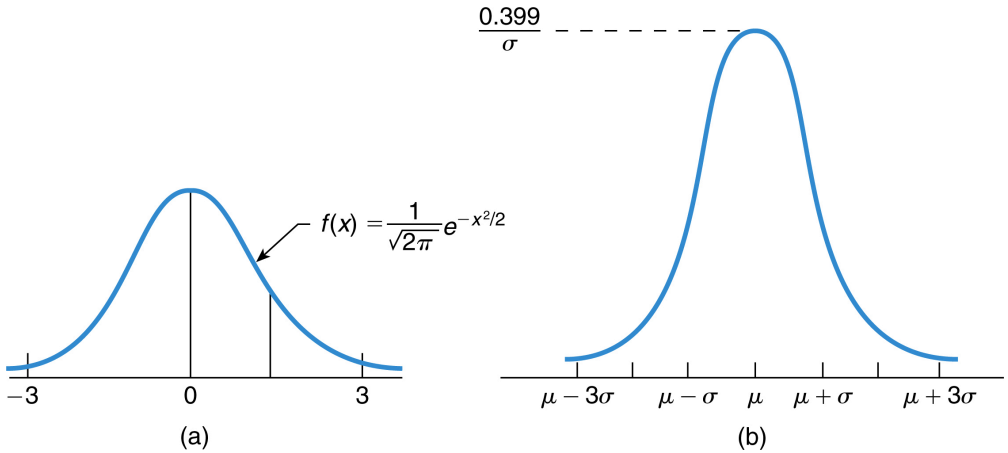
1. Preamble
2. Data visualization
3. Description of data
4. Outlier detection
5. Normal distribution and its derivatives
6. backup for review

Normal RV and derivatives

- Normal distribution: most useful distribution, used approximate other distributions.
- χ^2-distribution: sum of squares of independent standard normal RVs
- t-distribution: ratio of standard Normal RV and chi-squared RV
- F-distribution: ratio of two independent chi-square RVs

Gaussian (Normal) RV

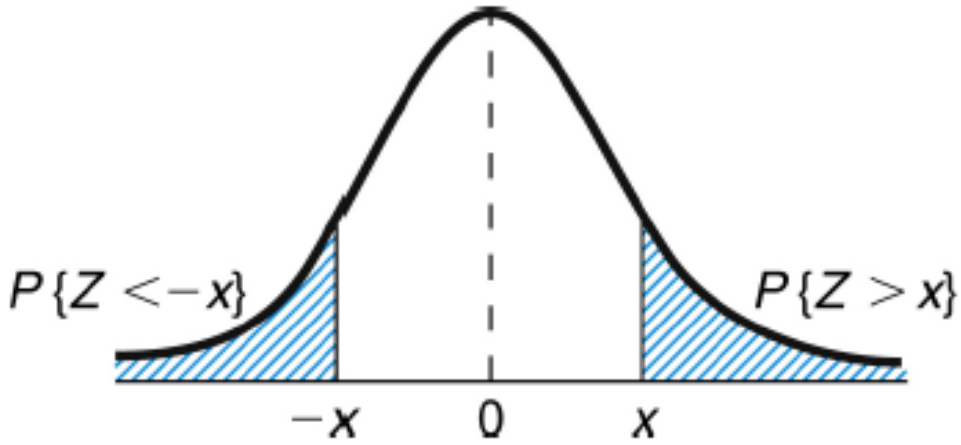
- PDF: $f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$



- Represented as $\mathcal{N}(\mu, \sigma^2)$, where μ, and σ^2 are mean and variance, respectively.
- For normal, mean=median=mode.
- The maximum height of $\mathcal{N}(\mu, \sigma^2)$ is $\frac{1}{\sqrt{2\pi}\sigma}$ and hence maximum height $\propto \frac{1}{\sigma}$
- If $X \sim \mathcal{N}(\mu, \sigma^2)$, then for any constants a and $b \neq 0$, the random variable $Y = a + bX$ is also a normal random variable with $E(Y) = E(a + bX) = a + b\mu$, $Var(Y) = b^2\sigma^2$

Standard Normal RV

- A Standard Normal RV Z is $Z \sim \mathcal{N}(0, 1)$
- CDF of Z : $\phi(Z) = P(Z \leq z) = \frac{1}{\sqrt{2\pi}} \int_{\infty}^{z} e^{-t^2/2} dt$
- $\phi(Z)$ is readily available for reference
- If the table contains values of $\phi(z)$ for $z \geq 0$ only; use symmetry



- Given a Normal random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ we can use the following transformation to make it standard Normal RV $Z = \frac{X-\mu}{\sigma}$
- Verify that: $E(Z) = \frac{E(X)-\mu}{\sigma} = 0$ and $Var(Z) = \frac{Var(X)}{\sigma} = 1$

Standard Normal RV: examples

Q Average rain fall at IIT Delhi campus follows normal distribution, with mean 60 cm/year with a variance of 20 cm/year. What is the probability that we get at least 80 cm/yr?

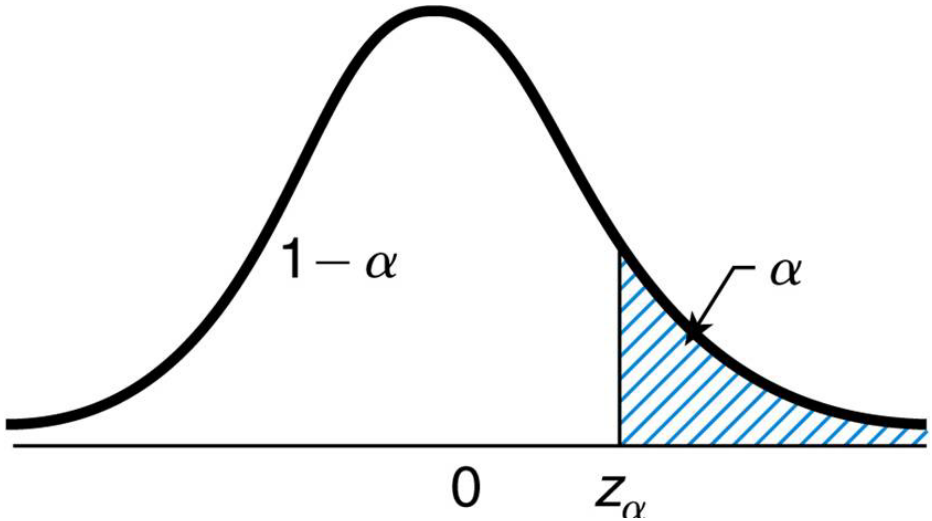
Sol We need to calculate $P(X \geq 80)$

$P(X \geq 80) \sim \mathcal{N}(60, 20)$. Then $P(\frac{X-\mu}{\sigma} = Z \geq \frac{80-60}{20}) \sim \mathcal{N}(0, 1)$

$P(Z \geq 1) = 1 - P(Z \leq 1) = 1 - \phi(1) = 1 - 0.8413 = 0.1587$

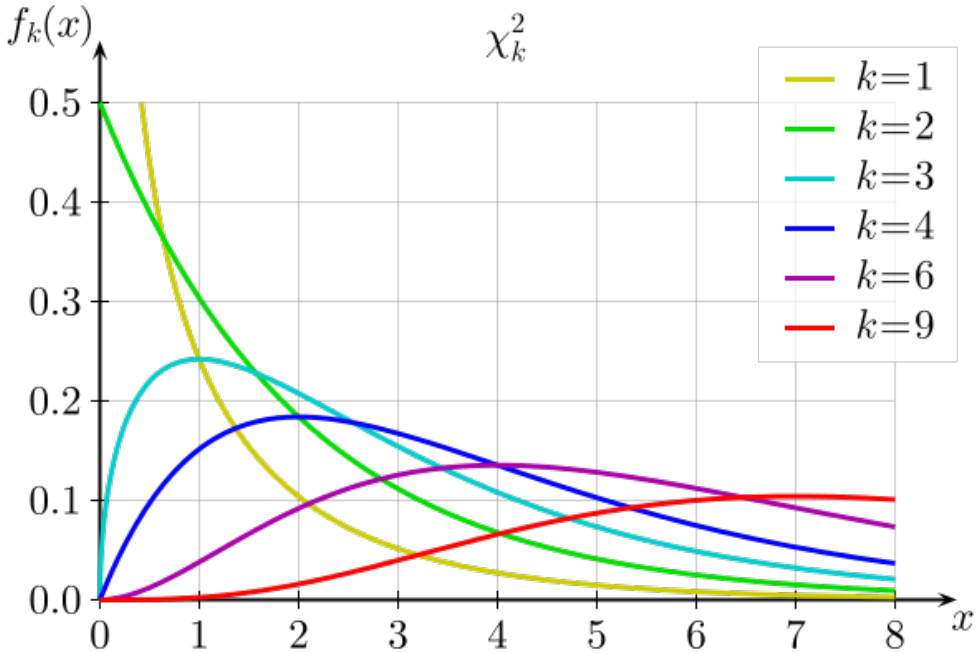
Standard Normal RV

- $P(X \leq x) = P(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}) = P(Z \leq \frac{x-\mu}{\sigma}) = \phi(\frac{x-\mu}{\sigma})$
- $100 \times (1-\alpha)^{th}$ percentile $\mathcal{N}(0,1) = z_\alpha$ where z_α is such that $P(Z > z_\alpha) = \alpha$, or $(1-\alpha) \times 100$ percent of the time a standard normal random variable will be less than z_α
- $P(Z \leq z_\alpha) = 1 - \alpha$



The Chi-squared distribution

- If $X_1, X_2, \ldots, X_n$ are independent standard normal random variables, then the RV X, defined as:
  $X = \sum \frac{(X_1-\overline{X})^2}{\sigma_1} + \cdots + \frac{(X_n-\overline{X})^2}{\sigma_n}$ follows χ^2 distribution with $n-1$ degrees of freedom, $X \in \chi^2_{n-1}$

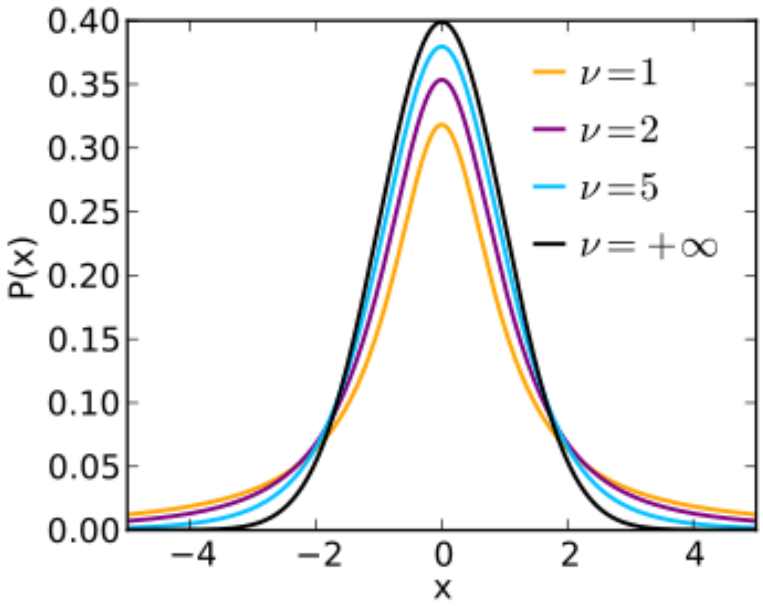


The Chi-squared distribution

- Let X_1, X_2 be independent chi-square random variables with n_1 and n_2 degrees of freedom, respectively. Then $X_1 + X_2$ is a chi-square with $n_1 + n_2$ degrees of freedom.
- For a chi-squared RV with n degrees of freedom: $E[X] = n$, $Var(X) = 2n$
- For a χ^2 RV X with n degrees of freedom $P(X > \chi^2_{\alpha,n}) = \alpha$
- Not symmetric
- χ^2 statistics - $\sum \frac{(Expected - Observed)^2}{Expected}$

The t distribution

- If Z and χ^2_n are random variables, with Z having a standard normal distribution and χ^2_n having a chi-squared distribution with n degrees of freedom, then the random variable T_n defined by $T_n = \frac{Z}{\sqrt{\chi^2/n}}$ is is said to have a t-distribution with n degrees of freedom.



ν-degrees of freedom (n)

The t distribution

- t density is symmetric about 0 like standard normal density.
- As n becomes larger ($\rightarrow \infty$), t density tends to a standard normal density.
-

$$E(T_n) = \begin{cases} 0 \ for \ n > 1 \\ undefined \ otherwise \end{cases}$$

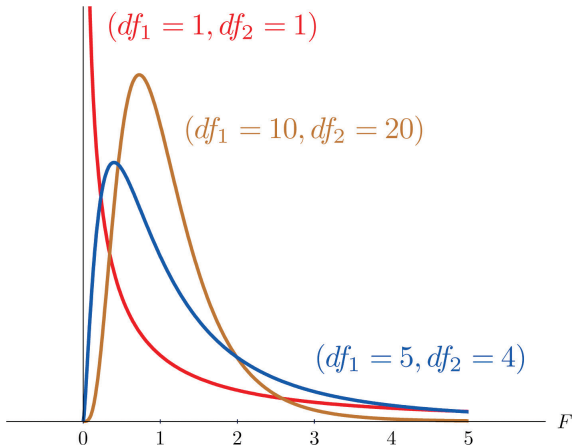
-

$$Var(T_n) = \begin{cases} \frac{n}{n-2} \ for \ n > 2 \\ \infty \ 1 < n \leq 2 \\ undefined \ otherwise \end{cases}$$

- From symmetry of the density $t_{1-\alpha,n} = -t_{\alpha,n}$
- Let $t_{\alpha,n}$ such that $P(T_n > t_{\alpha,n}) = \alpha, \ 0 < \alpha < 1$

The F distribution

- Ratio of two indpendent chi-square variables
- If χ^2_n and χ^2_m are independent chi-square RVs with n and m degrees of freedom respectively, then the RV $F_{n,m}$ defined by $F_{n,m} = \frac{\chi^2_n/n}{\chi^2_m/m}$ is said to have an F-distribution with n and m degrees of freedom.



- For any $\alpha \in (0, 1)$, let let $F_{\alpha,n,m}$ be such that $P(F_{n,m} > F_{\alpha,n,m}) = \alpha$
- For $\alpha > 0.5$, $F_{1-\alpha,m,n} = \frac{1}{F_{\alpha,n,m}}$
- $F_{0.9,5,7} = 1/F_{0.1,7,5}$

Central Limit Theorem (CLT)

Result (Central Limit Theorem)

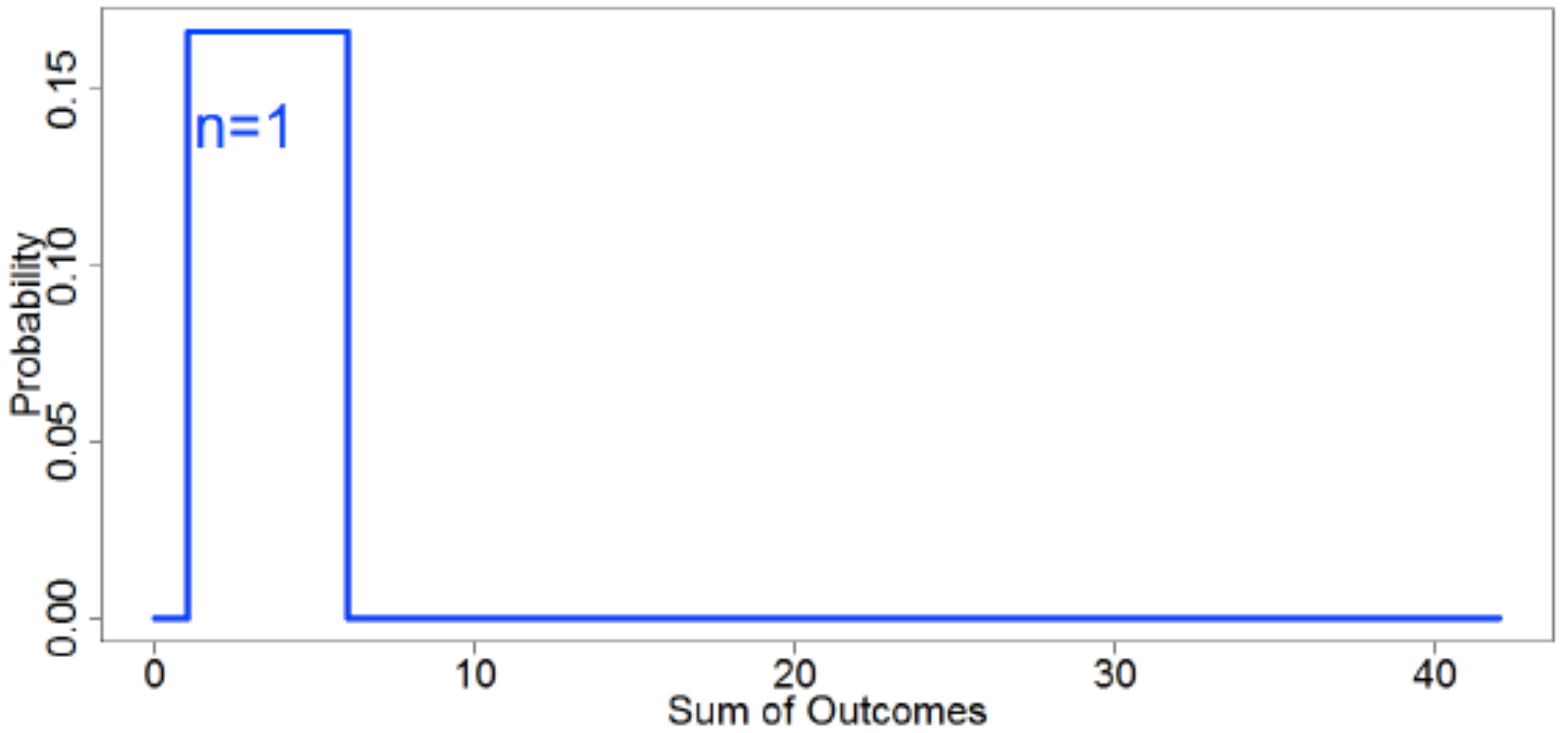
Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables each having a mean μ and variance σ^2. Then for large n, the distribution of $X_1 + \cdots + X_n$ is approximately normal with mean $n\mu$ and variance $n\sigma^2$.

i.e. for large n, $X_1 + X_2 + \cdots + X_n \sim \mathcal{N}(n\mu, n\sigma^2)$ or $\frac{(X_1+X_2+\cdots+X_n)}{n} = \overline{X} \sim \mathcal{N}(\mu, \sigma^2)$

- One of the most powerful results in probability Proposed in 1733 by French mathematician A. deMoivre. Forgotten until Laplace published it in 1812: used normal to approximate binomial. In 1901 Lyapunov defined it in general terms and proved it formally

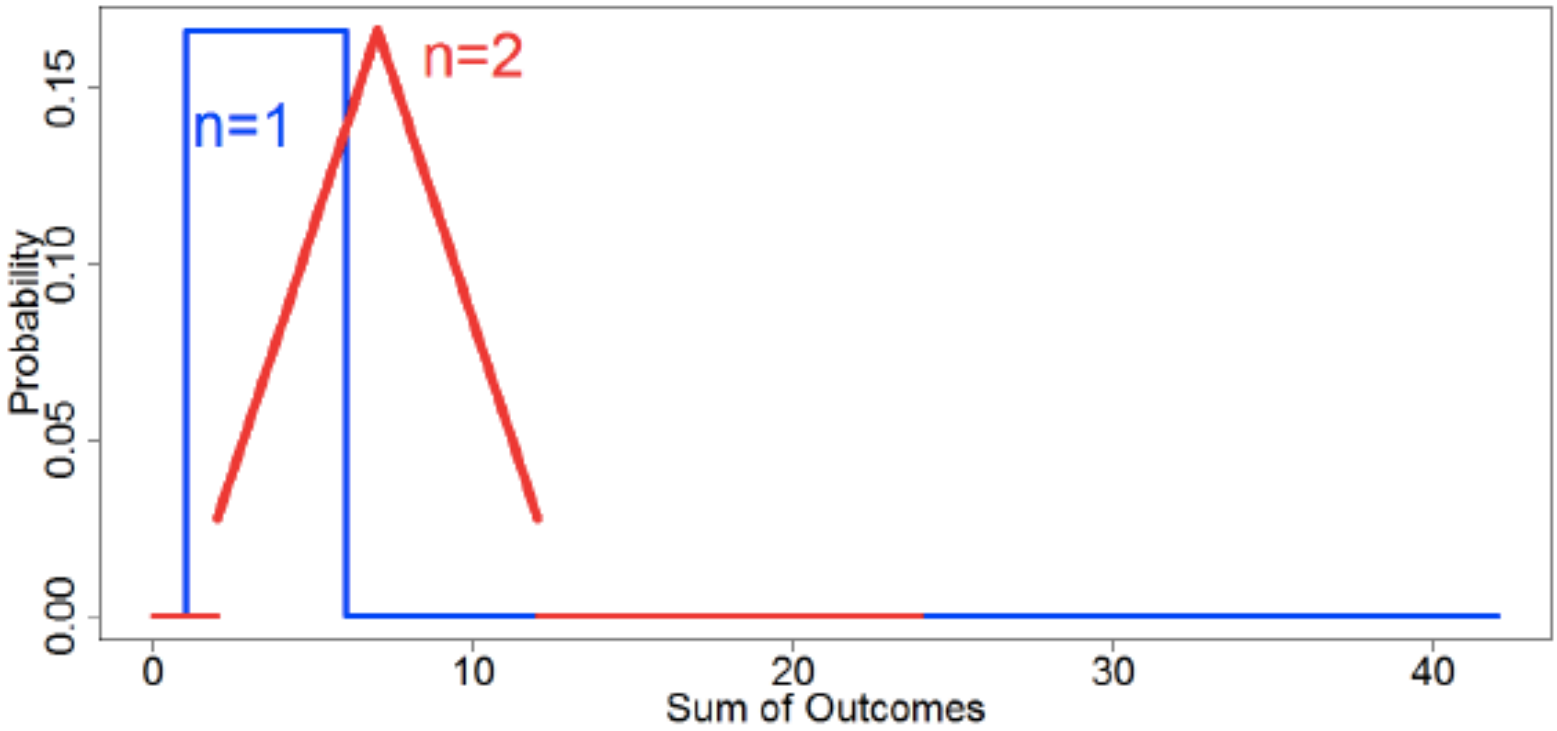
CLT: illustration

Consider the total obtained on rolling several dice:



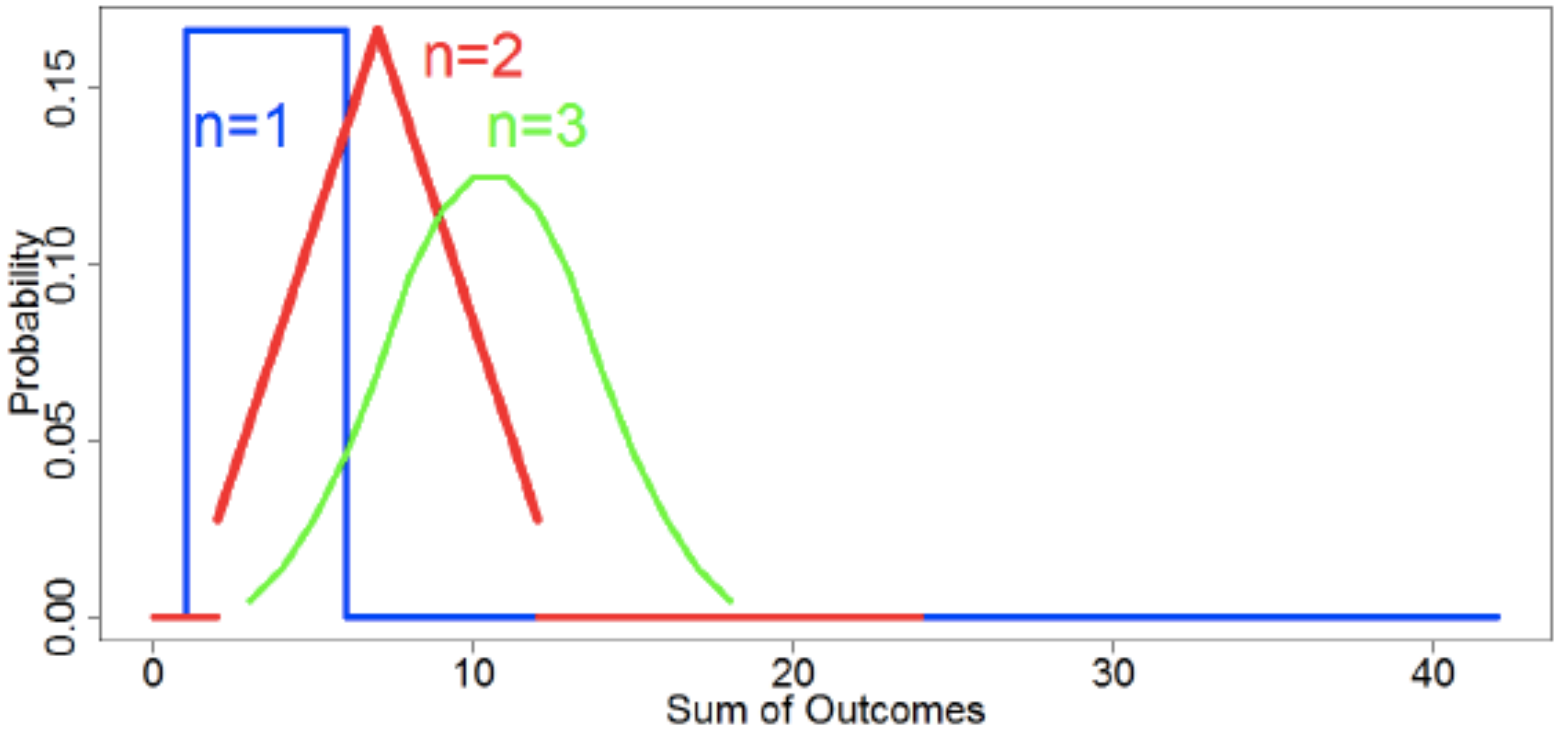
CLT: illustration

Consider the total obtained on rolling several dice:



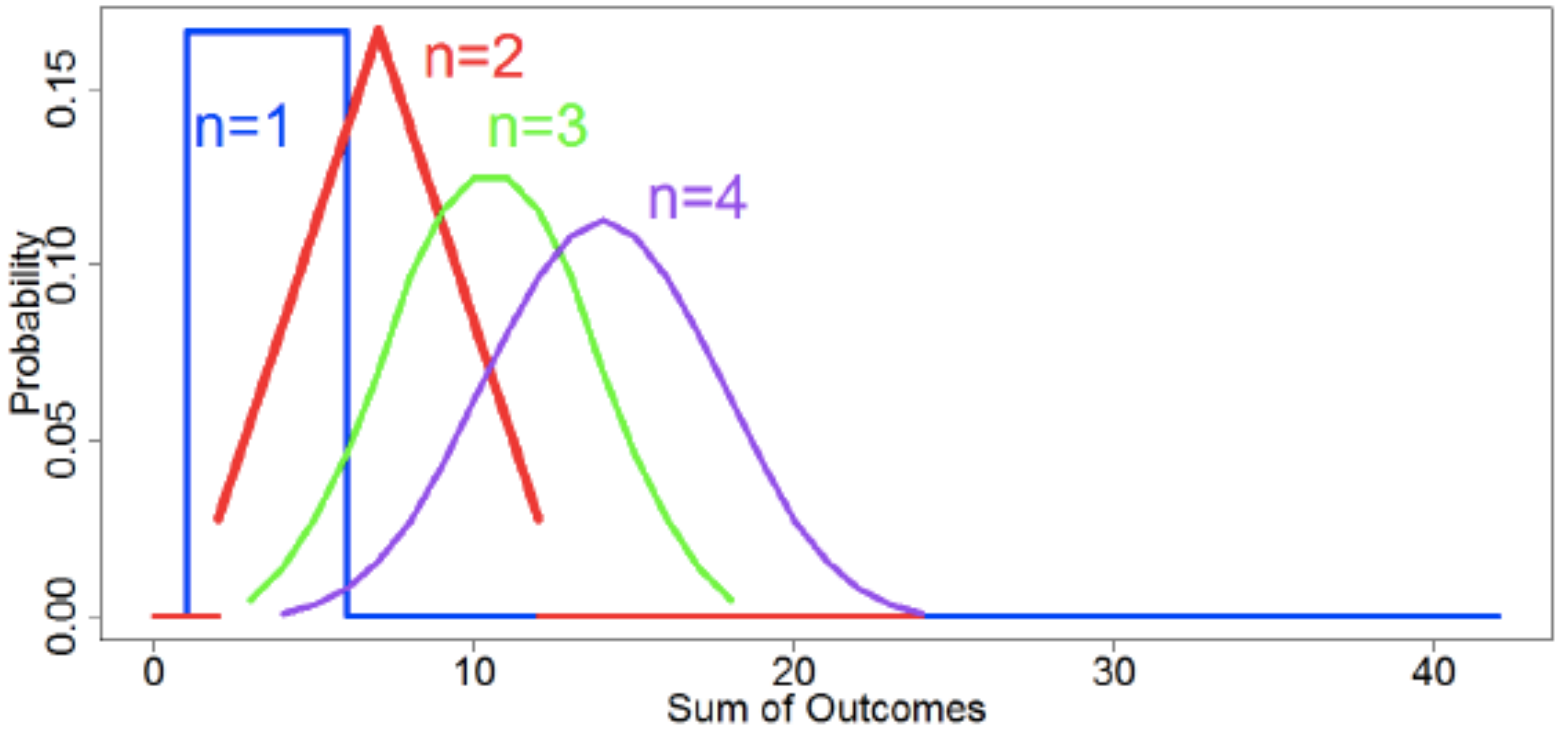
CLT: illustration

Consider the total obtained on rolling several dice:



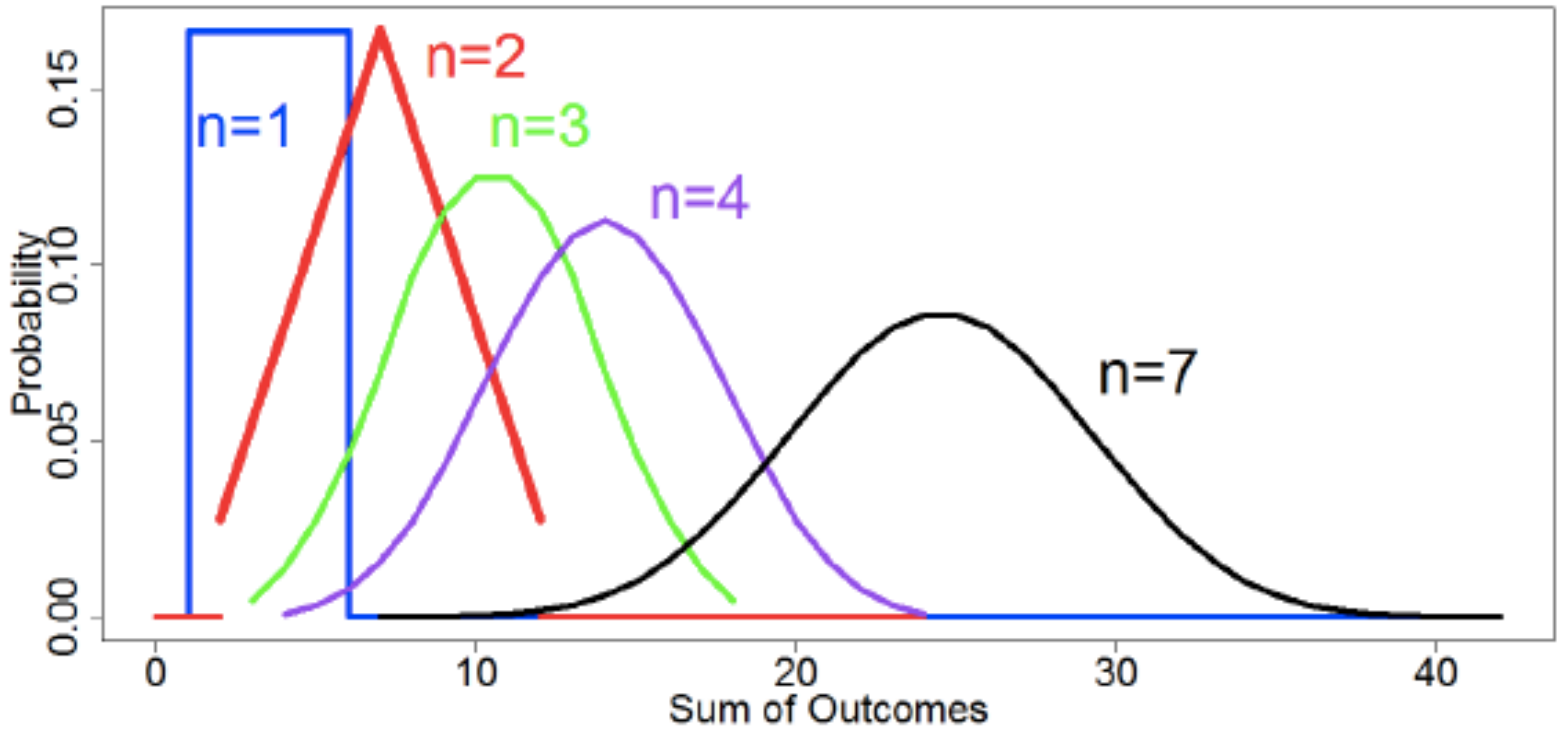
CLT: illustration

Consider the total obtained on rolling several dice:



CLT: illustration

Consider the total obtained on rolling several dice:



CLT and sample size

- CLT does not tell us how large n should be for the normal approximation of the X to be valid
- As a general rule of thumb,the normal approximation may be applied if $n > 30$
- In most cases, normal approximation will be valid for much smaller sample sizes
- Sample mean $\overline{X} = \frac{\sum_i X_i}{n}$, tend to follow Normal distribution, no matter how non-normal the underlying population

Samples and Population: recap

- Population $\supset$ Sample
- Population parameters are typically in Greek.
  μ= population mean, σ^2= population variance
- The corresponding sample statistic is mostly represented in English
  $\overline{X}$= sample mean, S^2 = sample variance.
- Statistic is a RV, that is why, we need to learn about RVs and distributions
- Our aim is to calculate Population parameters from Sample parameters (Estimation) and also verify its significance (Hypothesis testing)
- $\overline{X}$ is an estimator of μ. There can be other estimators as well, like median, mode etc.
- Similarly S^2 is an estimate σ^2

Thank you!

Overview of Presentation

1. Preamble
2. Data visualization
3. Description of data
4. Outlier detection
5. Normal distribution and its derivatives
6. backup for review

Probability

- Probability: used to quantify the likelihood, or chance, that an outcome of a random experiment will occur.
- Subjective belief: the likelihood of an outcome is quantified by assigning a number from the interval [0, 1] to the outcome
- For a discrete sample space, the probability of an event E, denoted as $P(E)$, equals the sum of the probabilities of the outcomes in E.

Axioms of probability

If S is the sample space and E is any event in a random experiment,

1. $P(S) = 1$
2. $0 \leq P(E) \leq 1$
3. For two events E_1 and E_2 with $E_1 \cap E_2 = \emptyset$, $P(E_1 \cup E_2) = P(E_1) + P(E_2)$
4. Bayes theorem: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$, *for* $P(B) > 0$

Random variables: more formal

- A RV is a function that assigns a real number to each outcome in the sample space of a random experiment, i.e, if X is RV, then, $X : S \to \mathbb{R}$
- The sequence $HHTH$ of coin tosses is not an RV. It must have a numerical mapping.
- A RV is denoted by an uppercase letter such as X, the measured value of the RV is denoted by a lowercase letter ($x = 2$ *heads*)
- Two coin toss: $S = \{(H, H), (H, T), (T, H), (T, T)\}$, let X : no heads turning up in one experiment. Then, $X(H, H) = 2$, $X(H, T) = 1$, etc.
- A discrete RV is has a finite (or countably infinite) range. Eg: number of transmitted bits, result of coin toss, etc.
- A continuous RV has an interval (either finite or infinite) of real numbers for its range. Eg: electrical current, length, pressure, temperature, time, voltage, weight, etc.

Probability functions

- The assignment of a probability to the value x that a random variable X takes is represented as $P_X(X = x)$ or simply $P(X = x)$

Q Consider two coin toss. If X = number of heads, calculate $P(X = x)$

sol $S = \{(H, H), (H, T), (T, H), (T, T)\}$

$$P(X = x) = \begin{cases} \frac{1}{2} if \ x = 1 \\ \frac{1}{4} if \ x = 2, 0 \\ 0 \ otherwise \end{cases}$$

- $0 \leq P(X = x) \leq 1$
- $\sum_x P(X = x) = 1$

Probability mass function (PMF) $p(x)$

The probability distribution of a random variable X is a description of the probabilities associated with the possible values of X.

Probability mass function $p(x)$

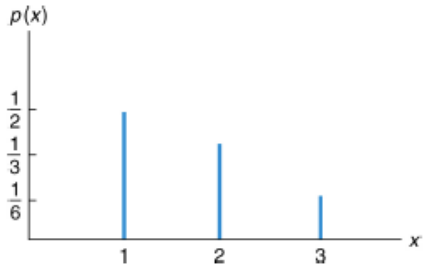
For a discrete random variable X with possible values $x_1, x_2, \ldots, x_n$ a probability mass function is a function such that,

- $p(x_i) = P(X = x_i)$
- $p(x_i) \geq 0$
- $\sum_{i=1}^{n} p(x_i) = 1$

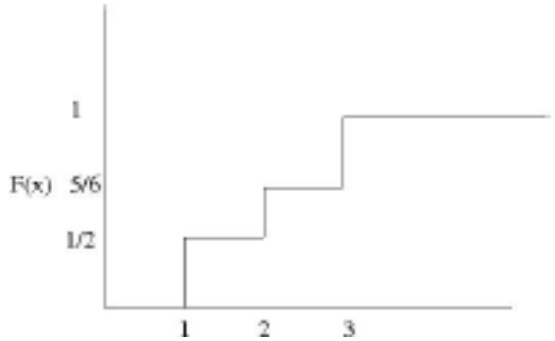
Cumulative distribution function (CDF) ($F(x)$)

- $F(x) = P(X \leq x) = \sum_{x_{\leq} x} p(x_i)$
- $0 \leq F(x) \leq 1$
- $P(a < X \leq b) = F(b) - F(a)$

PMF and CDF



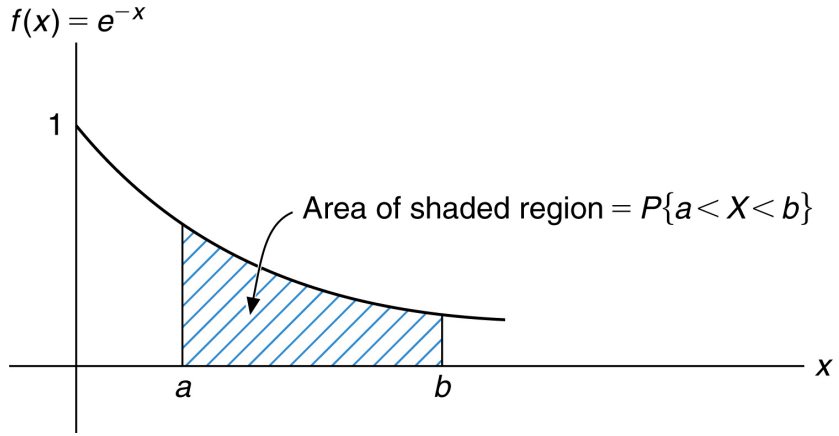
$$p(0) = 0, p(1) = 1/2, p(2) = 1/3, p(3) = 1/6$$



$$F(X \leq a) = \begin{cases} 0 \ if \ x < 1 \\ 1/2 \ if \ 1 \leq x \leq 2 \\ 5/6 \ if \ 2 \leq x \leq 3 \\ 1 \ if \ 3 \leq a \end{cases}$$

Probability density function (PDF)

- For a continuous random variable X: probability density function $f(x)$ is a nonnegative function, defined for all $x \in \mathbb{R}$, $P(X \in B) = \int_B f(x)dx$, where B, any set of real numbers
- For $B \in [a, b]$, $P(a \leq X \leq B) = \int_a^b f(x)dx$



- If $a = -\infty, b = \infty$, $\int_a^b f(x)dx = 1$
- If $a = b$, $\int_a^b f(x)dx = 0$, that is, probability that a continuous random variable will assume any particular value is 0.
- $f(x)$ is not a probability value

Cumulative density function

- $F(a) = P(X \in (-\infty, a)) = \int_{-\infty}^{a} f(x)dx$
- Differentiating both the sides (Hint: Lebnitz rule: $\frac{d}{dx}\left(\int_{f_1(x)}^{f_2(x)} g(t)\,dt\right) = g[f_2(x)]f_2'(x) - g[f_1(x)]f_1'(x).$)
  $\frac{d}{da}F(a) = f(a)$
- Density function is derivative of cumulative distribution function

Expectation operator

- Discrete random variable (RV): The expectation of X is denoted as $E(x) = \mu = \sum_i x_i p(x_i)$
- Continuous RV: For a continuous RV, $E(x) = \mu = \int_{-\infty}^{\infty} xf(x)dx$
- It is the average value of X in a large number of repetitions of the experiment and not necessarily the value that we expect X to have.
- It is the centre of gravity of the probability mass (density) function of discrete (continuous) random variable.

Expectation operator: discrete rv

Q What is the expected value when you roll a die?

sol $S = [1, 2, 3, 4, 5, 6]$ each having probability of $1/6$
$\mu = 1 \times (1/6) + \cdots + 6 \times (1/6) = 3.5$

Q2 What is the probability that a drug works on $\{0, 1, 2, 3, 4\}$out of four patients:

r	$P(X = r)$
0	0.008
1	0.076
2	0.265
3	0.411
4	0.240
total	1

sol $E[X] = 0(0.008) + 1(0.076) + 2(0.265) + 3(0.411) + 4(0.240) = 2.80$

Expectation operator: continuous rv

Q You are waiting for a phone call from your girl (boy) friend after dinner. Based on the past history, it has been observed that, number of hours (X) after dinner until the call arrives is a random variable with the following probability density function:

$$f(x) = \begin{cases} 1/1.5, \ \textit{if} \ 0 < x < 1.5 \\ 0 \ \textit{otherwise} \end{cases}$$

If the time is now $9PM$, when can you expect his(her) call?

sol $E[X] = \int_0^{1.5}(1/1.5)xdx = 0.75$. He (she) will get call around $9:45PM$

Variance operator

- Variance operator: $Var(X) = \sigma^2 = E(X - \mu)^2$
- Variance of a Discrete RV:

$$
\begin{aligned}
&= \sum_i (x_i - E(X))^2 p(x_i) \\
&= \sum_i (x_i^2 - 2x_i E(X) + (E(X))^2) p(x_i) \\
&= \sum_i x_i^2 p(x_i) - 2E(X) \sum_i x_i p(x_i) + \sum_i (E(X))^2 \\
&= E(X^2) - 2E(X)E(X) + (E(X))^2 \\
&= E(X^2) - (E(X))^2 \\
&= E(X^2) - \mu^2
\end{aligned}
$$

- continuous RV: $\int_{-\infty}^{\infty} (X - \mu)^2 f(x) dx$
- Standard deviation of RV: $\sigma = \sqrt{Var(X)}$