Elements of Statistics

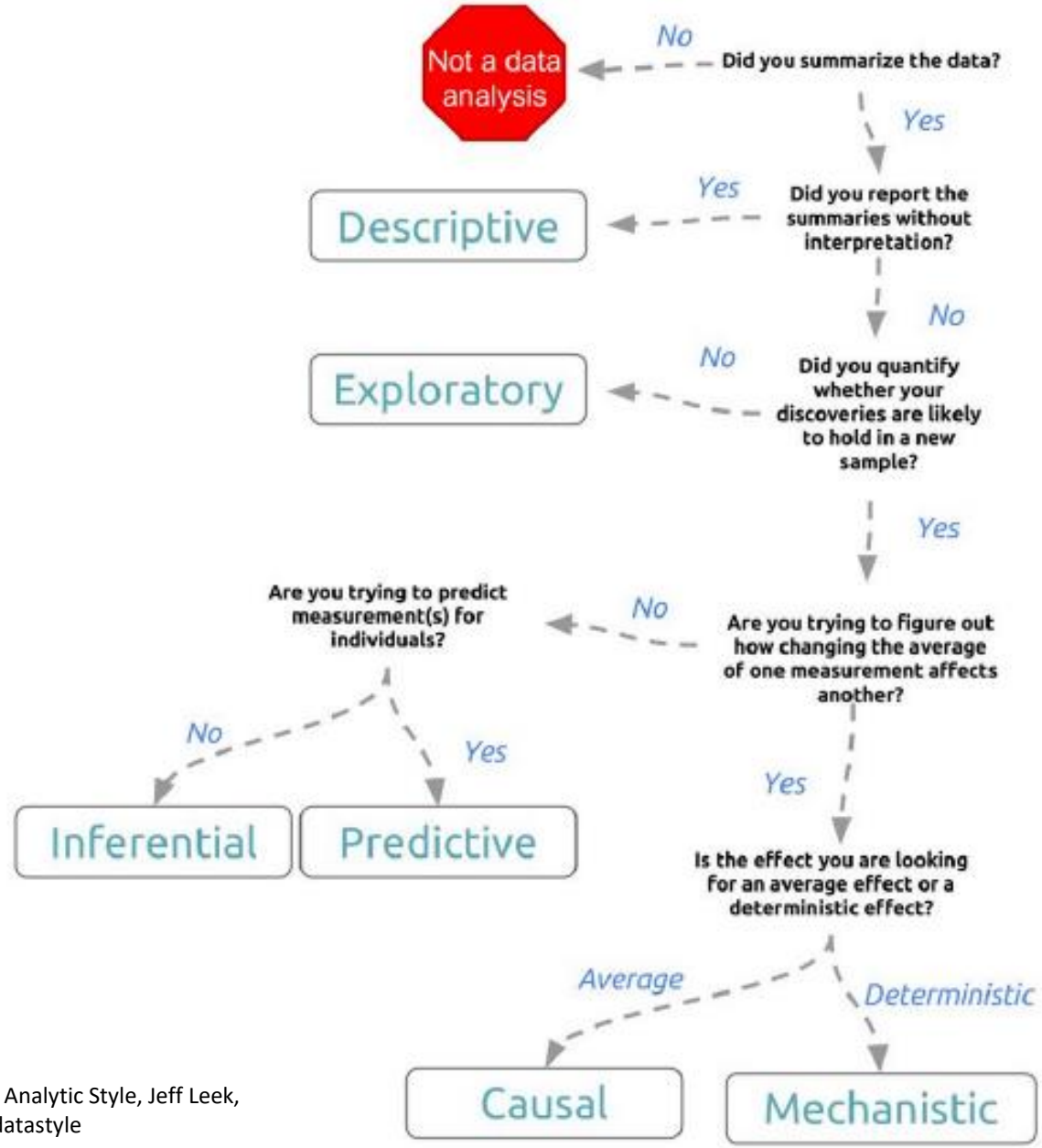
Vinay Kulkarni

What is ‘Statistics’

STATISTICS

- The branch of science that deals with
  - Collecting data
  - Organizing and summarizing data
  - Analysis of data
  - Inferring / Predicting / Deciding based on the data and its analysis

Statistics - Branches



Ref: Elements of Data Analytic Style, Jeff Leek, http://leanpub.com/datastyle

More About Statistics

- Descriptive Statistics: Introduction
- Central Tendency
  - Measures of Central Tendency
- Measures of Position
- Measures of Dispersion
- Measures of Quality and Outliers

Summarizing Quantitative Data

Measures used to summarize quantitative data

- **Measures of Central Tendency**
  - The Mean
  - The Mode
  - The Median
- Measures of Variation / Dispersion
- Measures of Position
- Measures of Quality and Outliers

The Mean

- For the Population

$$\mu = (x_1 + x_2 + \ldots + x_N) / N = \sum x_i / N$$

- For the Sample

$$\overline{x} = (x_1 + x_2 + \ldots + x_n) / n = \sum x_i / n$$

Summarizing Quantitative Data

Measures used to summarize quantitative data

- Measures of Central Tendency
- **Measures of Variation / Dispersion**
  - The Range
  - The Variance
  - The Standard Deviation
- Measures of Position
- Measures of Quality and Outliers

The Range

- The Range

$$\text{Range} = R = \text{Largest data value} - \text{smallest data value}$$
$$= \text{Maximum} - \text{minimum}.$$

- Range: Measure of distance between the extremes in the data
- It does not tell us how the observations are distributed between the smallest and the largest data values

The Variance and Standard Deviation (Population)

- Variance (Population)

$$\sigma^2 = \frac{\sum_{i=1}^{N} (Y_i - \mu)^2}{N}$$

- Standard Deviation (Population)

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (Y_i - \mu)^2}{N}}$$

The Variance and Standard Deviation (Sample)

- Variance (Sample)

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

- Standard Deviation (Sample)

$$S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

Rule: Sample v/s Population

Any set of data should be considered as a **Sample** until it is clearly specified that data is the whole **Population**

Summarizing Quantitative Data

Measures used to summarize quantitative data

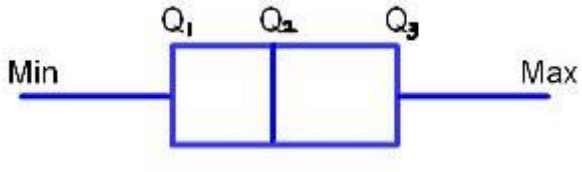
- Measures of Central Tendency
- Measures of Variation / Dispersion
- **Measures of Relative Position**
- **Measures of Quality and Outliers**
  - The Percentiles
  - The Deciles
  - The Quartiles
  - The z-score

The Percentiles, Deciles and Quartiles

- Percentiles
  - Divide the data set, in order of magnitude, into 100 parts
  - Hence 99 percentiles can be determined
- Deciles
  - Divide the data set, in order of magnitude, into 10 parts
- Quartiles
  - Divide the data set, in order of magnitude, into 4 equal parts, each a quartile

Characteristics of Quartiles

- Quartiles help us to identify the following
  - Min, 25th Percentile, Median, 75th Percentile, Max



- Inter Quartile Range : Q3 – Q1
  - Range of the middle 50% of the data set
- IQR is resistant to extreme values
  - Variance and Standard Deviation are not
- **Quartiles can help identify 'outliers' by defining the 'fences'**
  - Lower Fence = Q1 – 1.5 * IQR
  - Upper Fence = Q3 + 1.5 * IQR

The z-score

- **The z-score**
  - Distance of a data point, in terms of the Standard Deviation, from the mean of all observations

Population Z- Score: $$Z = \frac{x - \mu}{\sigma}$$

Sample Z- score: $$Z = \frac{\left(Y - \bar{Y}\right)}{S}$$

- What is the ‘mean’ of z-scores?
- What is its standard deviation?
- Uses of z-scores?

Normal Distribution Table

STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.9	.00005	.00005	.00004	.00004	.00004	.00004	.00004	.00004	.00003	.00003
-3.8	.00007	.00007	.00007	.00006	.00006	.00006	.00006	.00005	.00005	.00005
-3.7	.00011	.00010	.00010	.00010	.00009	.00009	.00008	.00008	.00008	.00008
-3.6	.00016	.00015	.00015	.00014	.00014	.00013	.00013	.00012	.00012	.00011
-3.5	.00023	.00022	.00022	.00021	.00020	.00019	.00019	.00018	.00017	.00017
-3.4	.00034	.00032	.00031	.00030	.00029	.00028	.00027	.00026	.00025	.00024
-3.3	.00048	.00047	.00045	.00043	.00042	.00040	.00039	.00038	.00036	.00035
-3.2	.00069	.00066	.00064	.00062	.00060	.00058	.00056	.00054	.00052	.00050
-3.1	.00097	.00094	.00090	.00087	.00084	.00082	.00079	.00076	.00074	.00071
-3.0	.00135	.00131	.00126	.00122	.00118	.00114	.00111	.00107	.00104	.00100
-2.9	.00187	.00181	.00175	.00169	.00164	.00159	.00154	.00149	.00144	.00139
-2.8	.00256	.00248	.00240	.00233	.00226	.00219	.00212	.00205	.00199	.00193
-2.7	.00347	.00336	.00326	.00317	.00307	.00298	.00289	.00280	.00272	.00264
-2.6	.00466	.00453	.00440	.00427	.00415	.00402	.00391	.00379	.00368	.00357
-2.5	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
-2.4	.00820	.00798	.00776	.00755	.00734	.00714	.00695	.00676	.00657	.00639
-2.3	.01072	.01044	.01017	.00990	.00964	.00939	.00914	.00889	.00866	.00842
-2.2	.01390	.01355	.01321	.01287	.01255	.01222	.01191	.01160	.01130	.01101
-2.1	.01786	.01743	.01700	.01659	.01618	.01578	.01539	.01500	.01463	.01426
-2.0	.02275	.02222	.02169	.02118	.02068	.02018	.01970	.01923	.01876	.01831
-1.9	.02872	.02807	.02743	.02680	.02619	.02559	.02500	.02442	.02385	.02330
-1.8	.03593	.03515	.03438	.03362	.03288	.03216	.03144	.03074	.03005	.02938
-1.7	.04457	.04363	.04272	.04182	.04093	.04006	.03920	.03836	.03754	.03673
-1.6	.05480	.05370	.05262	.05155	.05050	.04947	.04846	.04746	.04648	.04551
-1.5	.06681	.06552	.06426	.06301	.06178	.06057	.05938	.05821	.05705	.05592
-1.4	.08076	.07927	.07780	.07636	.07493	.07353	.07215	.07078	.06944	.06811
-1.3	.09680	.09510	.09342	.09176	.09012	.08851	.08691	.08534	.08379	.08226
-1.2	.11507	.11314	.11123	.10935	.10749	.10565	.10383	.10204	.10027	.09853
-1.1	.13567	.13350	.13136	.12924	.12714	.12507	.12302	.12100	.11900	.11702
-1.0	.15866	.15625	.15386	.15151	.14917	.14686	.14457	.14231	.14007	.13786
-0.9	.18406	.18141	.17879	.17619	.17361	.17106	.16853	.16602	.16354	.16109
-0.8	.21186	.20897	.20611	.20327	.20045	.19766	.19489	.19215	.18943	.18673
-0.7	.24196	.23885	.23576	.23270	.22965	.22663	.22363	.22065	.21770	.21476
-0.6	.27425	.27093	.26763	.26435	.26109	.25785	.25463	.25143	.24825	.24510
-0.5	.30854	.30503	.30153	.29806	.29460	.29116	.28774	.28434	.28096	.27760
-0.4	.34458	.34090	.33724	.33360	.32997	.32636	.32276	.31918	.31561	.31207
-0.3	.38209	.37828	.37448	.37070	.36693	.36317	.35942	.35569	.35197	.34827
-0.2	.42074	.41683	.41294	.40905	.40517	.40129	.39743	.39358	.38974	.38591
-0.1	.46017	.45620	.45224	.44828	.44433	.44038	.43644	.43251	.42858	.42465
-0.0	.50000	.49601	.49202	.48803	.48405	.48006	.47608	.47210	.46812	.46414

Normal Distribution Table

STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158
2.4	.99180	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361
2.5	.99379	.99396	.99413	.99430	.99446	.99461	.99477	.99492	.99506	.99520
2.6	.99534	.99547	.99560	.99573	.99585	.99598	.99609	.99621	.99632	.99643
2.7	.99653	.99664	.99674	.99683	.99693	.99702	.99711	.99720	.99728	.99736
2.8	.99744	.99752	.99760	.99767	.99774	.99781	.99788	.99795	.99801	.99807
2.9	.99813	.99819	.99825	.99831	.99836	.99841	.99846	.99851	.99856	.99861
3.0	.99865	.99869	.99874	.99878	.99882	.99886	.99889	.99893	.99896	.99900
3.1	.99903	.99906	.99910	.99913	.99916	.99918	.99921	.99924	.99926	.99929
3.2	.99931	.99934	.99936	.99938	.99940	.99942	.99944	.99946	.99948	.99950
3.3	.99952	.99953	.99955	.99957	.99958	.99960	.99961	.99962	.99964	.99965
3.4	.99966	.99968	.99969	.99970	.99971	.99972	.99973	.99974	.99975	.99976
3.5	.99977	.99978	.99978	.99979	.99980	.99981	.99981	.99982	.99983	.99983
3.6	.99984	.99985	.99985	.99986	.99986	.99987	.99987	.99988	.99988	.99989
3.7	.99989	.99990	.99990	.99990	.99991	.99991	.99992	.99992	.99992	.99992
3.8	.99993	.99993	.99993	.99994	.99994	.99994	.99994	.99995	.99995	.99995
3.9	.99995	.99995	.99996	.99996	.99996	.99996	.99996	.99996	.99997	.99997

Example

- The manager of a manufacturing unit is trying to quantify the variation in the weekly demand for coolant in his factory.
- Based on historical data, it appears that this demand is normally distributed.
- He knows that on an average 100 litres are required every week and that 90 % of the time weekly demand is below 115

- Help him calculate the Standard Deviation of the weekly coolant demand

Random Variables

Random Variables

- **Random experiment**
  - Process of measurement or observation in which the outcome **cannot** be completely determined in advance
- **Sample space**
  - All possible outcomes of a random experiment
- **Random Variable**
  - A real-valued quantity, or numerical measure, whose value depends on the outcomes of a random experiment
  - Can be **Discrete** or **Continuous**

Random Variable

- The probability that a Random Variable may assume a particular value is governed by a Probability Function:
  - For Discrete variables
    - **Probability Mass Function (PMF)**
  - For Continuous variables
    - **Probability Density Function (PDF)**
  - For Discrete / Continuous variables
    - **Cumulative Distribution Function**

Random Variable and Probabilities

- Let X be the Random Variable
- Let x be one of its possible values
- Let P(x) be the probability that X=x
- Then

$$0 \leq P(x) \leq 1$$

$$\sum_{\text{all values of } x} P(x) = 1$$

Random Variable: Expected Value

- Expected value : E(X)
  - Weighted average, of all possible values, considering their probabilities

- For Discrete random variable

$$\mathrm{E(X)} = \mu_X = \sum_{all\ values\ of\ x} [x.p(x)]$$

- For Continuous random variable

$$\mathrm{E(X)} = \mu_X = \int_{-\infty}^{\infty} x.f(x)dx$$

Random Variable: Variance and Standard Dev

- Variance of a Random Variable

$$\sigma_X^2 = E[(X - \mu_X)^2] = E(X^2) - \mu_X^2$$

- Standard Deviation of a Random Variable

$$\sigma_x = +\sqrt{\sigma_X^2}$$

- Where

$$E(X^2) = \sum_{\text{all values of } x} [x^2 . p(x)], \text{ in the discrete case, and}$$

$$E(X^2) = \int_{-\infty}^{\infty} x^2 . f(x) dx, \text{ in the continuous case}$$

- Variance and Standard Deviation reflects the extent to which the Random variable is close to its mean

Exercise

- Which of the following tables represent valid discrete probability distributions?

(a)

X	p(x)
1	0.2
2	0.35
3	0.12
4	0.40
5	-0.07

(b)

x	p(x)
1	0.2
2	0.25
3	0.10
4	0.14
5	0.49

(c)

x	p(x)
1	0.2
2	0.25
3	0.10
4	0.15
5	0.30

- For valid distributions, calculate:
  - The expected value , variance and standard deviation

Sample – to – Population

- Goal of all these definitions:
  - Given the nature of the phenomena
    - Probability distributions
  - And a sample with certain deductions
    - Measured observations
  - Predict additional properties and confidence levels
- We therefore need to
  - Understand generic phenomena
  - And the probabilistic nature of their events

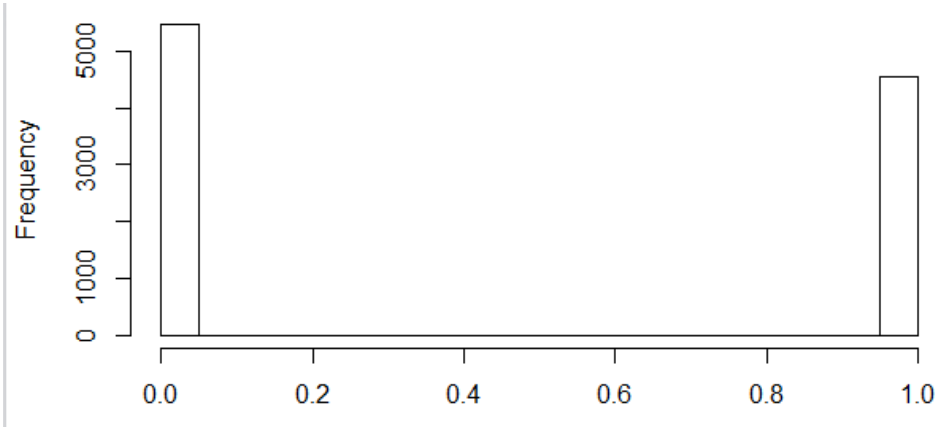
Probability Distributions

- Binomial Distribution
- Poisson Distribution
- Students T-Distribution
- Chi-Square Distribution
- F-Distribution
- Normal Distribution
- Log normal Distribution
- Other Distributions
  - Bernoulli
  - Geometric
  - Hypergeometric
  - Multinomial
  - Exponential
  - Beta
  - Gamma

Probability Distributions

Bernoulli Trials and Bernoulli Distribution

- If an experiment has **only two outcomes**, it is known as a **Bernoulli Trial**
- p = Probability of success
- q = Probability of failure



Bernoulli Dist: p = 0.45

Bernoulli

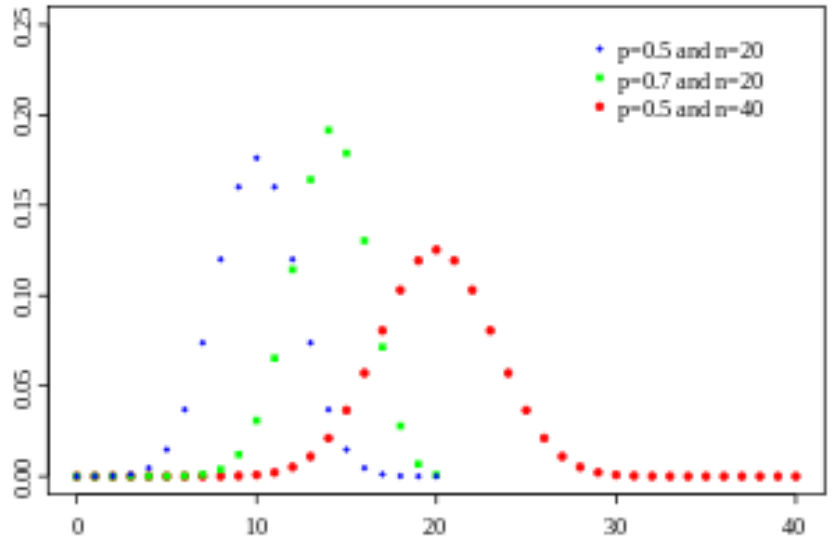
Parameters	$0 < p < 1, p \in \mathbb{R}$
Support	$k \in \{0, 1\}$
pmf	$\begin{cases} q = (1-p) & \text{for } k = 0 \\ p & \text{for } k = 1 \end{cases}$
CDF	$\begin{cases} 0 & \text{for } k < 0 \\ 1-p & \text{for } 0 \leq k < 1 \\ 1 & \text{for } k \geq 1 \end{cases}$
Mean	p
Median	$\begin{cases} 0 & \text{if } q > p \\ 0.5 & \text{if } q = p \\ 1 & \text{if } q < p \end{cases}$
Mode	$\begin{cases} 0 & \text{if } q > p \\ 0,1 & \text{if } q = p \\ 1 & \text{if } q < p \end{cases}$
Variance	$p(1-p)(=pq)$

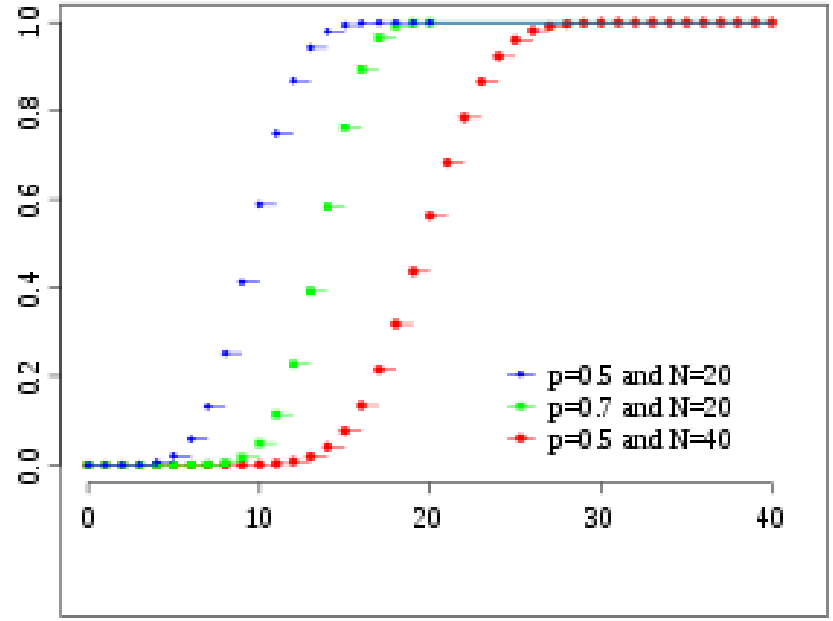
Binomial Distribution

- If an experiment has **only two outcomes**, it is known as a **Bernoulli Trial**
- Such an experiment is said to have Binomial Probability Distribution, if:
  - There are finite, **independent**, trial
  - Probability of success / failure is constant throughout the experiment
  - We are interested in the **number** of successes ‘x’, regardless of how they occur
- The number of successes is given by:

$$P(X = x) = P(x) = {}_nC_x\, p^x(1-p)^{n-x} = \binom{n}{x} p^x(1-p)^{n-x},\ x = 0, 1, 2, \ldots, n$$

Binomial Distribution





Notation	$B(n, p)$
Parameters	$n \in \mathbb{N}_0$ — number of trials $p \in [0,1]$ — success probability in each trial
Support	$k \in \{0, \ldots, n\}$ — number of successes
pmf	$\binom{n}{k} p^k (1-p)^{n-k}$
CDF	$I_{1-p}(n-k, 1+k)$
Mean	np
Median	$\lfloor np \rfloor$ or $\lceil np \rceil$
Mode	$\lfloor (n+1)p \rfloor$ or $\lfloor (n+1)p \rfloor - 1$
Variance	$np(1-p)$
Skewness	$\frac{1-2p}{\sqrt{np(1-p)}}$

Src: Wikipedia

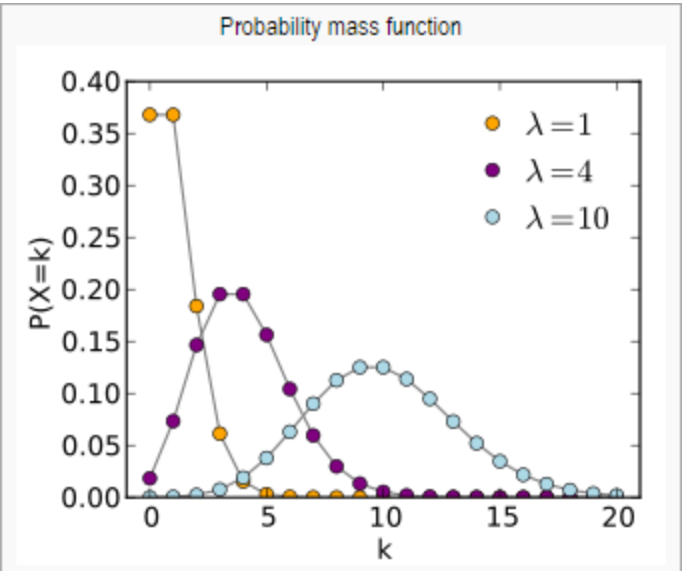
Exercise: Binomial Distribution

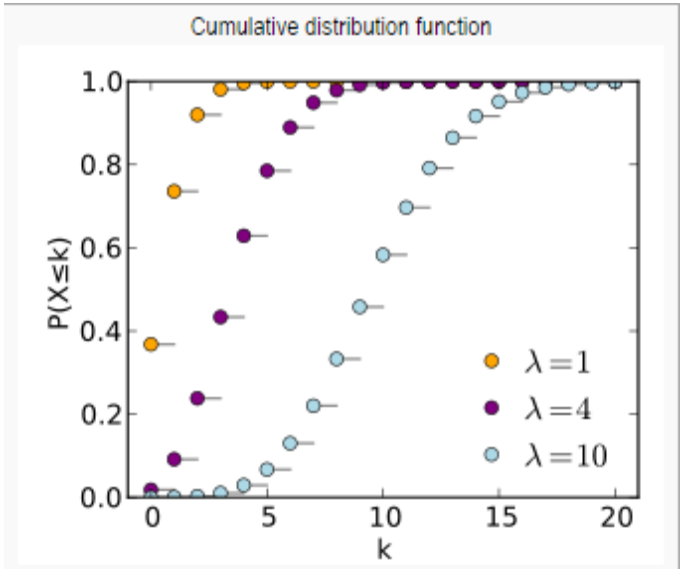
- Given:
  - From the tests run on a tool / workpiece combination, while removing a certain volume of material, it is observed that on an average 2 in 10 trials result in tool breakage
- Now find, if 15 such operations are carried out
  - What is the probability that at most 3 tools will break
  - What is the probability that exactly 5 tools will break

Poisson Distribution

- In case of random phenomena
- Where events are continuous
  - Calls arriving at a switchboard during lunch
  - Accidents at an intersection between 10 and noon
  - Misprints per page in a book
- The probability that a continuous measure X will take on value x, in a given unit of measurement, is governed by Poisson Distribution

Poisson Distribution





Notation	$\mathrm{Pois}(\lambda)$
Parameters	$\lambda > 0$ (real)
Support	$k \in \{0, 1, 2, 3, \ldots\}$
pmf	$\frac{\lambda^k}{k!}e^{-\lambda}$
CDF	$\frac{\Gamma(\lfloor k+1 \rfloor, \lambda)}{\lfloor k \rfloor !}$, or $e^{-\lambda}\sum_{i=0}^{\lfloor k \rfloor}\frac{\lambda^i}{i!}$, or $Q(\lfloor k+1 \rfloor, \lambda)$ (for $k \geq 0$, where $\Gamma(x, y)$ is the incomplete gamma function, $\lfloor k \rfloor$ is the floor function, and Q is the regularized gamma function)
Mean	λ
Median	$\approx \lfloor \lambda + 1/3 - 0.02/\lambda \rfloor$
Mode	$\lceil \lambda \rceil - 1, \lfloor \lambda \rfloor$
Variance	λ
Skewness	$\lambda^{-1/2}$

Src: Wikipedia

Poisson Distribution

- Probability Mass Function (PMF)

$$P(X = x) = p(x) = \lambda^x e^{-\lambda} / x!, \qquad x = 0, 1, 2, \ldots$$
$$= \text{zero, otherwise.}$$

  - Where:
    - λ = average number per unit measurement
    - X = Specific value of the measure
- Expected value / Mean
    - λ
- Variance
    - λ

Exercise

Consider the number of accidents between 8 and 9 am on an intersection on Saturday. From data recorded let the mean of accidents on that intersection have a mean of 4. Hence this follows what we call a Poisson distribution with $\lambda = 4$. Find the probability that on a given Saturday, between 8 and 9 am, there will be:

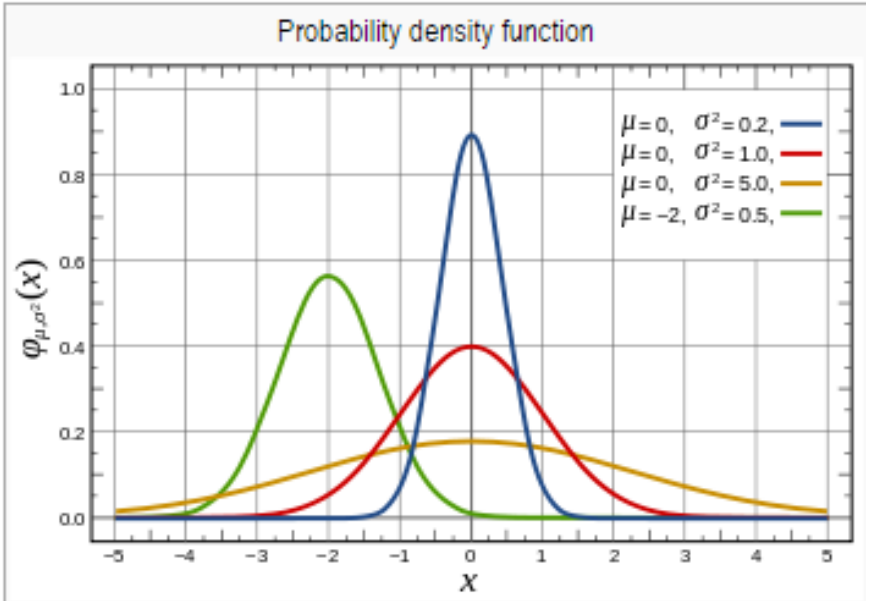
a) No accident,
b) At least one accident,
c) Exactly 4 accidents.

Normal Distribution

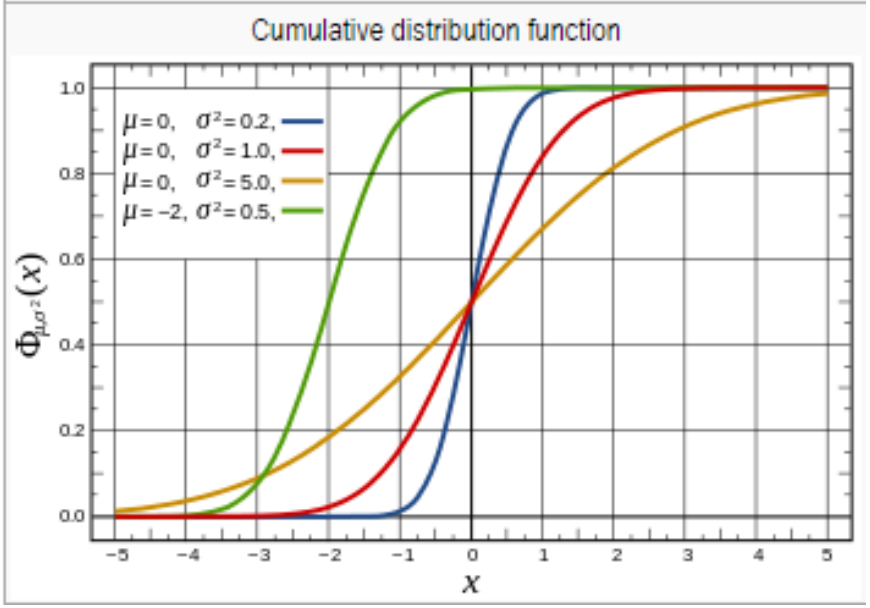
- A very well known Continuous Probability Distribution Function (PDF)
  - Applies to many phenomena
    - Human characteristics, physical quantities and processes, errors in physical and econometric measurements
  - Provides accurate approximation to a large number of probability laws
  - Important role in statistics and inferences
- PDF

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

Normal Distribution



The red curve is the *standard normal distribution*



Notation	$\mathcal{N}(\mu, \sigma^2)$
Parameters	$\mu \in \mathbf{R}$ — mean (location) $\sigma^2 > 0$ — variance (squared scale)
Support	$x \in \mathbf{R}$
pdf	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
CDF	$\frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$
Quantile	$\mu + \sigma\sqrt{2}\,\operatorname{erf}^{-1}(2F-1)$
Mean	μ
Median	μ
Mode	μ
Variance	σ^2
Skewness	0
Ex. kurtosis	0

Src: Wikipedia

Normal Distribution

- The Cumulative Distribution Function for Normal Distribution is available
  - In the form of a Standard Normal Distribution Table
  - Based on $Z = (x-\mu)/\sigma$
- The Standard Normal Distribution table is used in solving problems

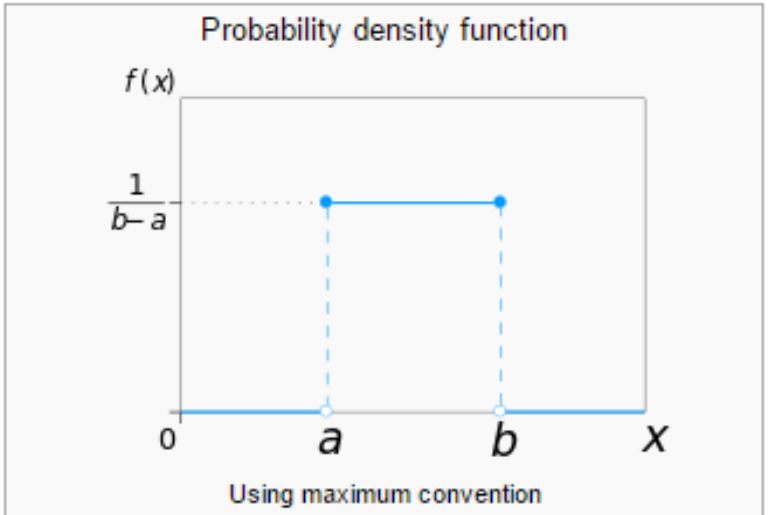
Example

- The weight of a certain category of watermelon follows normal distribution with mean 1.0 kg and standard deviation of 0.20 kg. Find:
  1. The probability that a watermelon weighs less than 1.5kg
  2. Probability that it weighs between 0.9kg and 1.2kg
  3. Probability that it weighs more than 1.6kg
  4. Percentage of watermelons that weigh between 0.8kg and 1.50kg
  5. Among a group of 300 watermelons how many will weigh between 0.8kg and 1.5kg?

Answers

1. 0.9938
2. 0.5328
3. 0.0013
4. 83.51%
5. About 251

Uniform Distribution



Notation	$\mathcal{U}(a,b)$ or $\operatorname{unif}(a,b)$
Parameters	$-\infty < a < b < \infty$
Support	$x \in [a,b]$
PDF	$\begin{cases} \frac{1}{b-a} & \text{for } x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$
CDF	$\begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } x \in [a,b) \\ 1 & \text{for } x \geq b \end{cases}$
Mean	$\frac{1}{2}(a+b)$
Median	$\frac{1}{2}(a+b)$
Mode	any value in (a,b)
Variance	$\frac{1}{12}(b-a)^2$

Estimation and Intervals

Estimation related Topics

- Sampling
- Point Estimation
- Sample Mean and Sample Variance
- Interval Estimation
- Confidence interval: one parameter
- Sample Size

Samples, Parameters & Statistics

- Sampling
  - Allows us to make inferences about a population based on a sample of that population
- Parameters
  - Numerical characteristics about the population that are of interest
- Statistics
  - Parameters cannot be exactly determined
  - They can only be estimated from samples
  - These estimates or summaries, based on the sample, are known as **Statistics**
- Major aspects of samples and statistics:
  - How accurate are the estimators (statistics)?
  - Is the sample truly representative of the population?

Statistics as Estimates for Parameters

- We use statistics to estimate parameters
  - Proportions
  - Arithmetic averages
  - Ranges
  - Quartiles
  - Deciles
  - Percentiles
  - Variances
  - Standard deviations

Sampling Methods

- Non-probability sampling
  - Convenience sampling
    - Randomly pick-up the easily accessible apples
  - Judgment or subjective sampling
  - Volunteer sampling
    - Especially used in clinical trials and research
- Probability sampling methods
  - This involves the planned use of **chance**
  - There is no selection bias

Point Estimation

- Estimation
  - First step of Inferential Statistics
    - (Second step is Hypothesis Testing)
  - Two types:
    - Point estimation
    - Interval estimation
- Point Estimate
  - Value of a statistic
  - Calculated from a sample
  - Estimates the parameter of the population

Discussion

- Different possible samples can be drawn from the same population
- Each of those samples can yield a different value of a statistic
  - Example: Mean and Variance
- It becomes important to investigate the sampling distributions of estimators
- **Sampling Distribution** for a given sample size, n:
  - Collection of all the estimators of that parameter
  - Of all possible samples of size ‘n’ from the population

Central Limit Theorem

The **central limit theorem** states that given a distribution with a mean μ and variance σ^2, the sampling distribution of the mean approaches a normal distribution with a mean (μ) and a variance σ^2/N as N, the sample size, increases

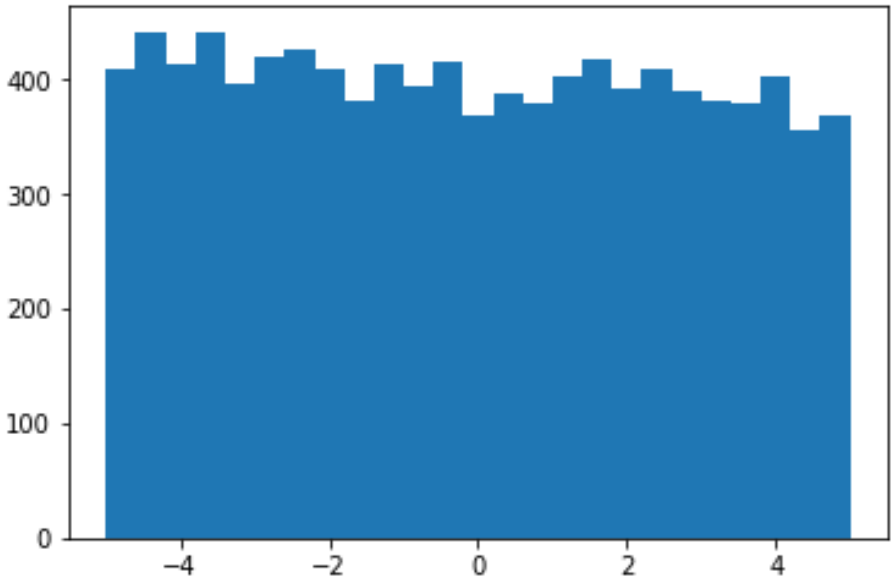
Sampling Distribution: Mean

- The sampling distribution of the Mean
  - Becomes approximately normal as the size of the sample 'n' increases
  - **Regardless of the shape of the population**
- Standard deviation of the sampling distribution of the mean: Also known as the **standard error**
- Where σ is the standard deviation of the population

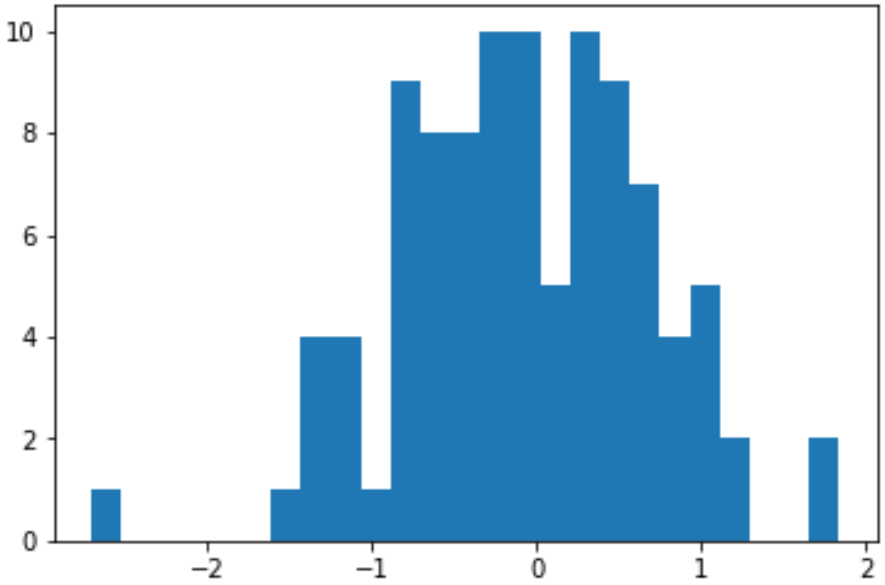
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

- http://onlinestatbook.com/

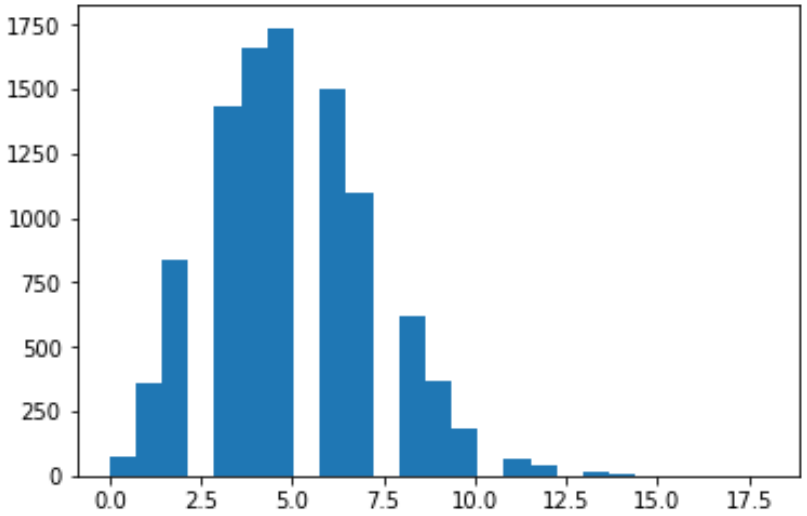
Sampling distribution of the mean



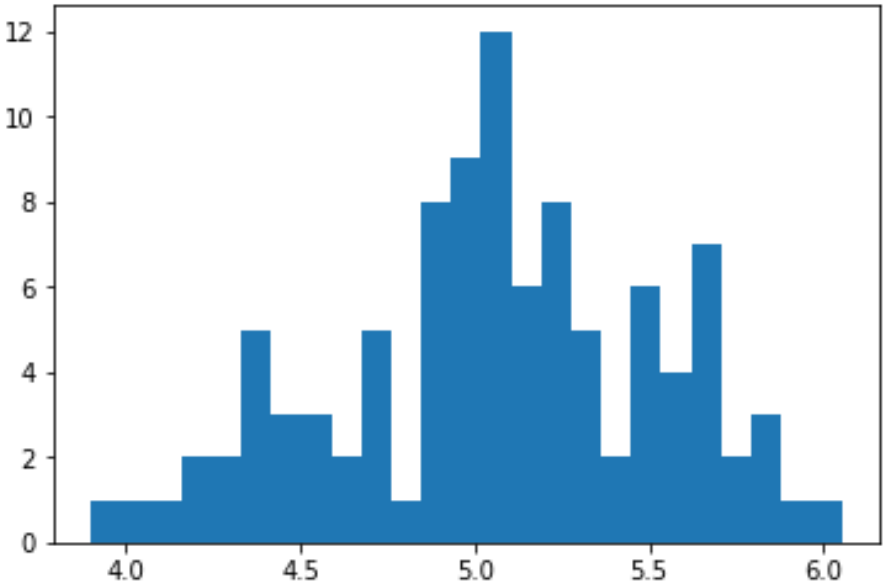
UNIFORM DISTTRIBUTION



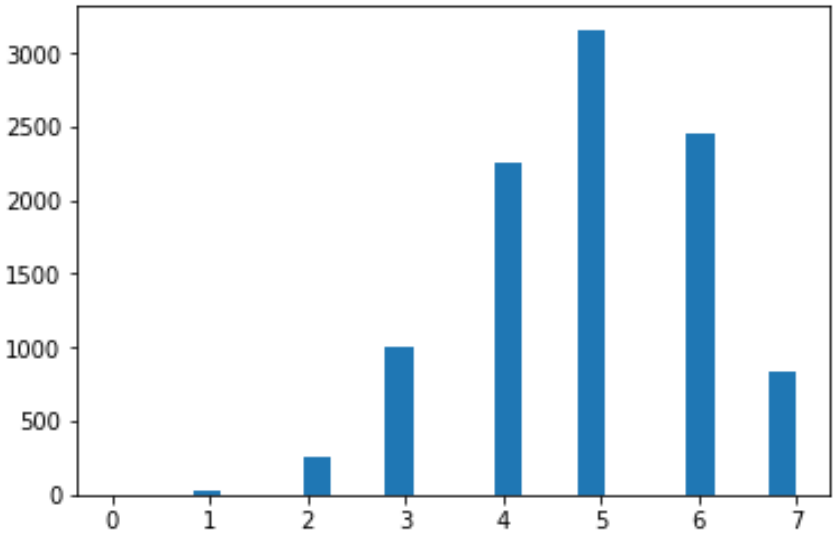
Sampling distribution of the mean



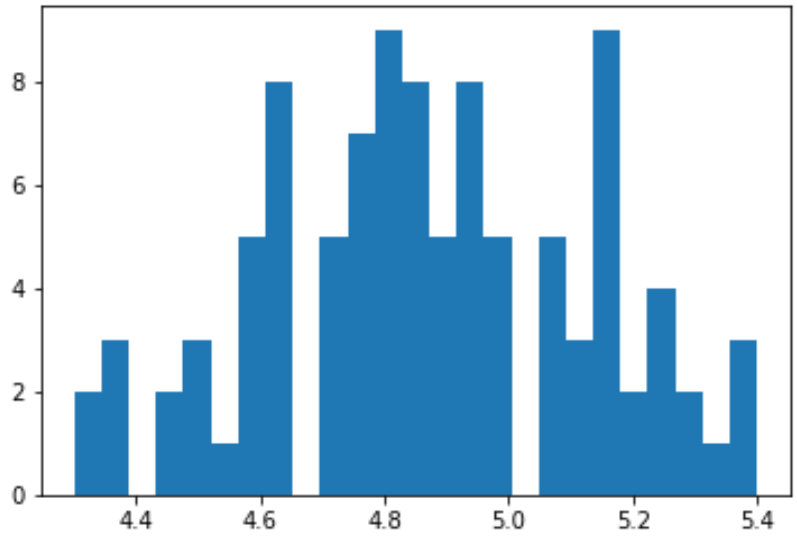
POISSON DISTRIBUTION



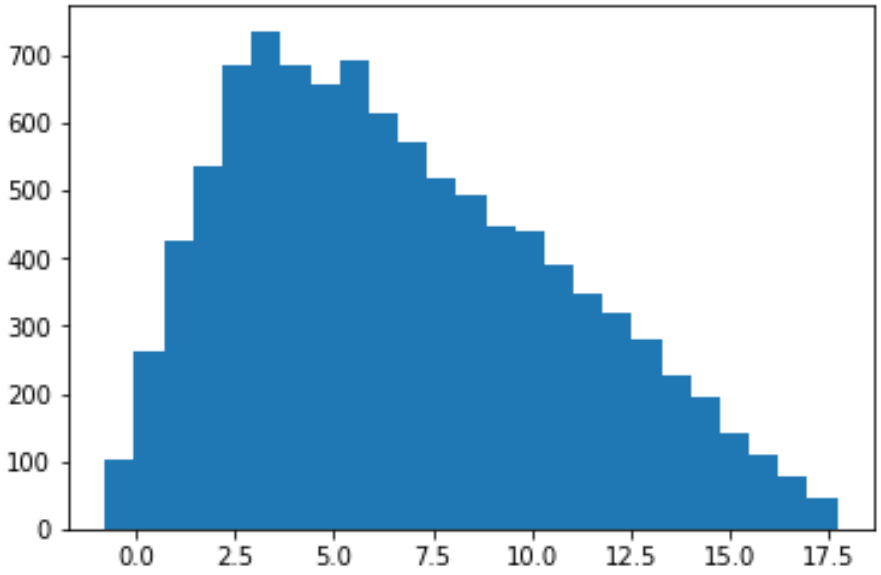
Sampling distribution of the mean



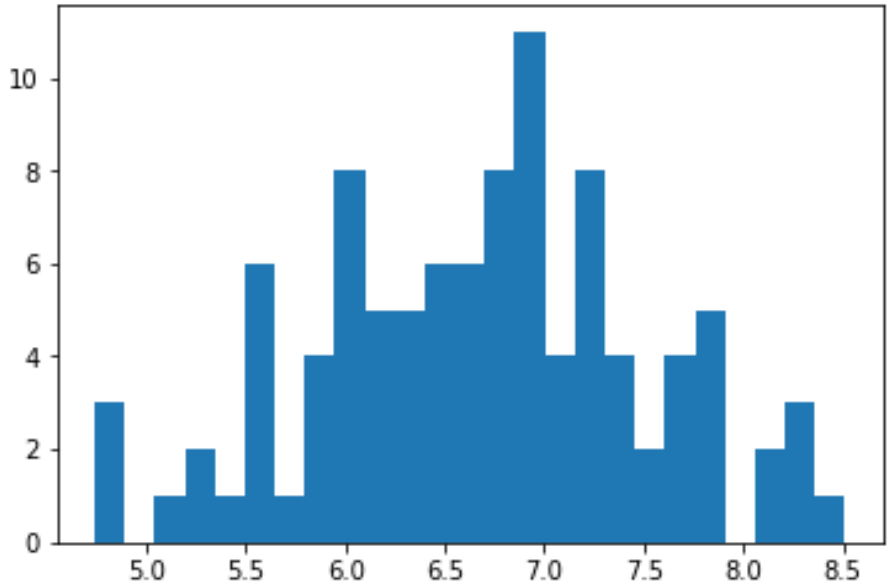
BINOMIAL DISTRIBUTION



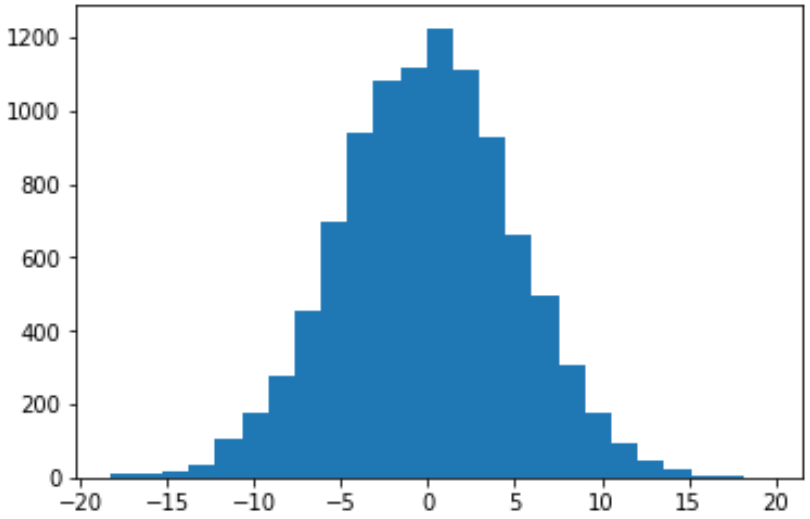
Sampling distribution of the mean



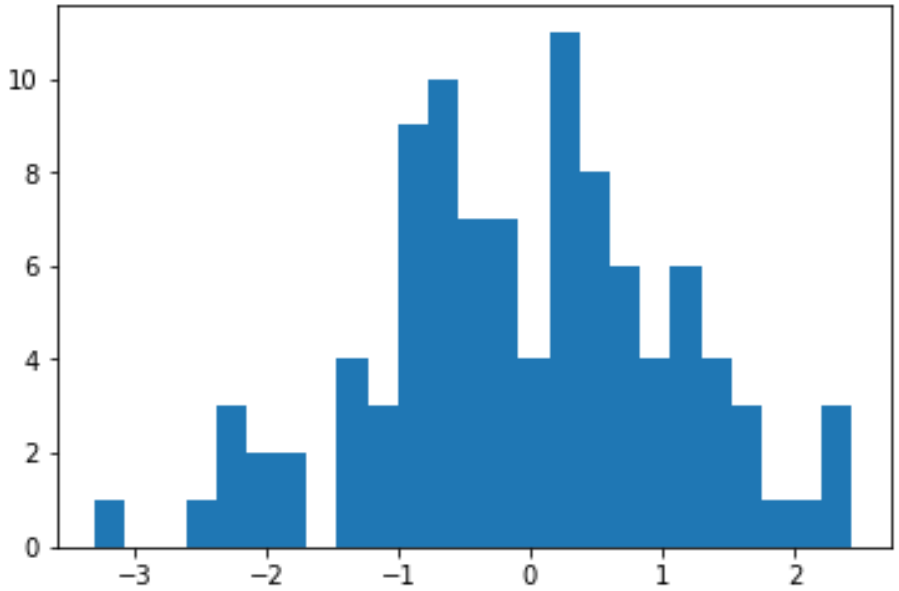
TRIANGULAR DISTRIBUTION



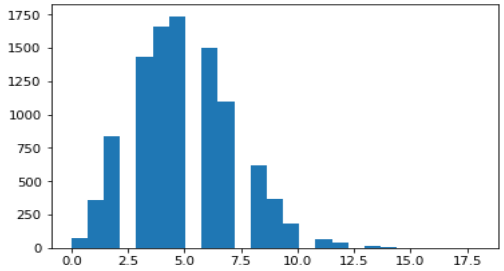
Sampling distribution of the mean

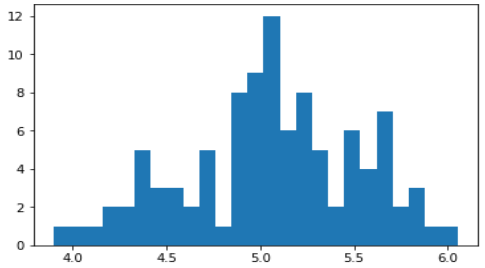


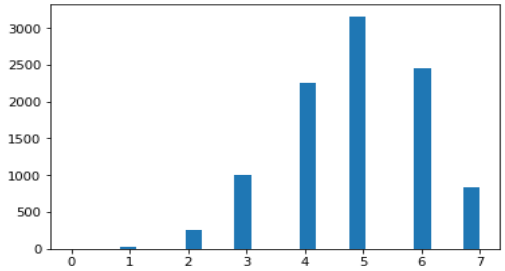
NORMAL DISTRIBUTION

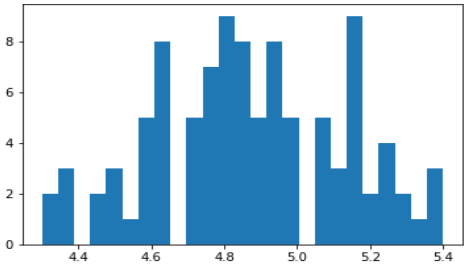


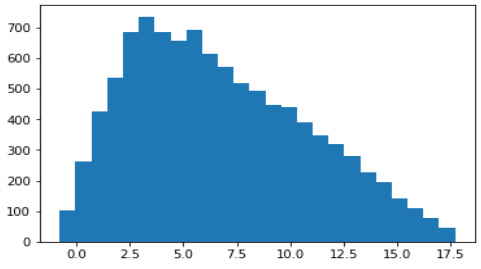
Sampling distribution of the mean

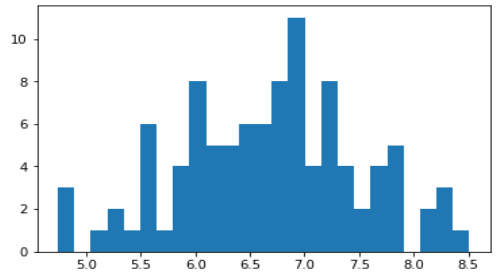


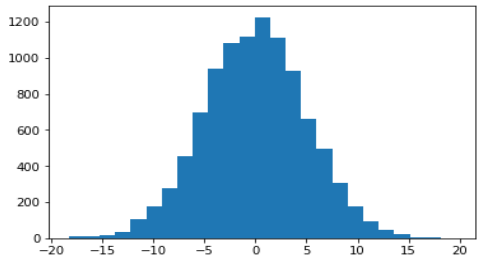


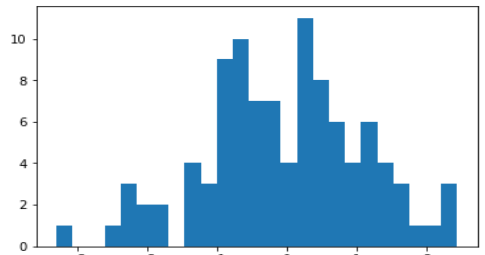












Interval Estimation

- Definition:
  - It is a range of values related to a parameter
  - Calculated based on the sample
  - Such that
    - The parameter will be within that range
    - With some degree of confidence
- Use
  - A statistic, such as the mean, can be presented
    - As a Point Estimate, X
    - As an interval, X $\pm E$ where ***E* is the margin of error**

Point v/s Interval Estimates

- Point estimate is often insufficient
  - It is either right or wrong!
  - It also does not indicate the confidence level in that estimate

- Interval estimate
  - Better option : report an interval estimate
  - It provides the range, as well as the degree of confidence

Confidence Interval : Mean

- Margin of error for the Mean
  - Where the population SD is known

$$\bar{x} - Z_{\alpha/2}\sigma / \sqrt{n} < \mu < \bar{x} + Z_{\alpha/2}\sigma / \sqrt{n}$$

  - Where sample size is large, and population SD is not known

$$\bar{x} - Z_{\alpha/2}S / \sqrt{n} < \mu < \bar{x} + Z_{\alpha/2}S / \sqrt{n}$$

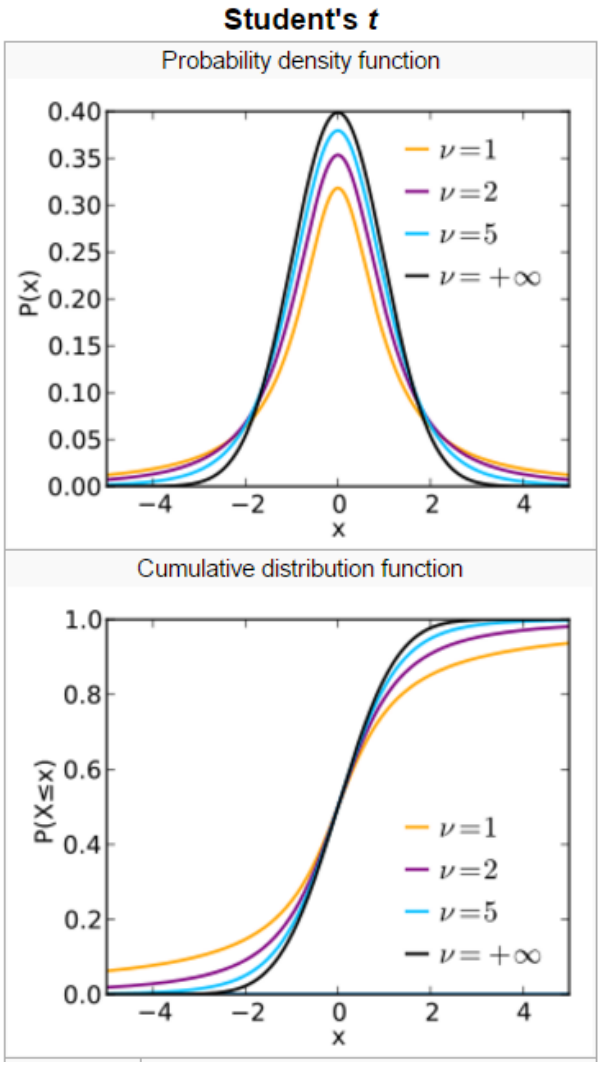
  - Where sample size is < 30, student t-distribution is used (**degrees of freedom = n-1**)

$$\bar{x} - t_{\alpha/2}S / \sqrt{n} < \mu < \bar{x} + t_{\alpha/2}S / \sqrt{n}$$

Students t-distribution

- Used to estimate the mean of a parameter that follows normal distribution
  - But, where the sample size is small
  - Typically < 30
- t-distribution varies with sample size
- It approaches normal distribution with large, very large sample size
- It is based on the ‘degrees of freedom’
  - Where degree of freedom = number of samples - 1

Students t-distribution



Parameters	$\nu > 0$ degrees of freedom (real)
Support	$x \in (-\infty; +\infty)$
PDF	$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$
CDF	$\frac{1}{2}+x\Gamma\left(\frac{\nu+1}{2}\right)\times \frac{{}_2F_1\left(\frac{1}{2},\frac{\nu+1}{2};\frac{3}{2};-\frac{x^2}{\nu}\right)}{\sqrt{\pi\nu}\,\Gamma\left(\frac{\nu}{2}\right)}$ where ${}_2F_1$ is the hypergeometric function
Mean	0 for $\nu > 1$, otherwise undefined
Median	0
Mode	0
Variance	$\frac{\nu}{\nu-2}$ for $\nu > 2$, ∞ for $1 < \nu \leq 2$, otherwise undefined
Skewness	0 for $\nu > 3$, otherwise undefined
Ex. kurtosis	$\frac{6}{\nu-4}$ for $\nu > 4$, ∞ for $2 < \nu \leq 4$, otherwise undefined

Src: Wikipedia

Exercise

- Mercury needs to be estimated in the water of a certain area. 16 samples are collected and their ppm values of mercury are as given below

409, 400, 406, 399, 402, 406, 401, 403, 401, 403, 398, 403, 407, 402, 410, and 399

- Manually calculate the mean, variance and standard deviation
- Estimate the ppm levels of mercury as an interval for 95% confidence level

Answer

- Sample size small
  - n <= 30
  - Hence students t-distribution will be used
- n = 16
- Sample mean = 403.063
- Variance = 12.996
- Standard deviation = 3.605
- Confidence level 95%, DOF = 16-1 = 15
- From t charts $t_{.025} = 2.131$
- Based on this the interval (401.143, 404.983)

Problem

A random sample of 100 families from a large city is chosen to estimate the current average annual demand for milk in that city. The mean family demand from the sample is 150 gallons with a standard deviation of 40 gallons.

a) Construct a 95% C.I. for the mean annual demand of milk by all families in the city.

Exercise

- In a factory that used coal as fuel, the consumption was observed for 10 consecutive weeks. It was found that an average value of 11400 tons of coal was consumed per day with a standard deviation of 700 tons.
- From this data, the plan manager wants to estimate an interval for the mean consumption such that he can be 95% confident about the coal requirement. Can you help him out?

Sample Size Determination

- Why determine sample size?
  - To ensure that the error in estimating a population parameter is less than a desired threshold
- When sample is too small:
  - Required precision is not achieved
- When sample size is too large
  - Wastage of resources required to estimate the parameter

Sample Size Determination

- In the case of sampling distribution of the mean, the margin of error is:

$$E = Z_{\alpha/2}.\sigma / \sqrt{n}$$

- Therefore

$$n = (Z_{\alpha/2}.\sigma / E)^2$$