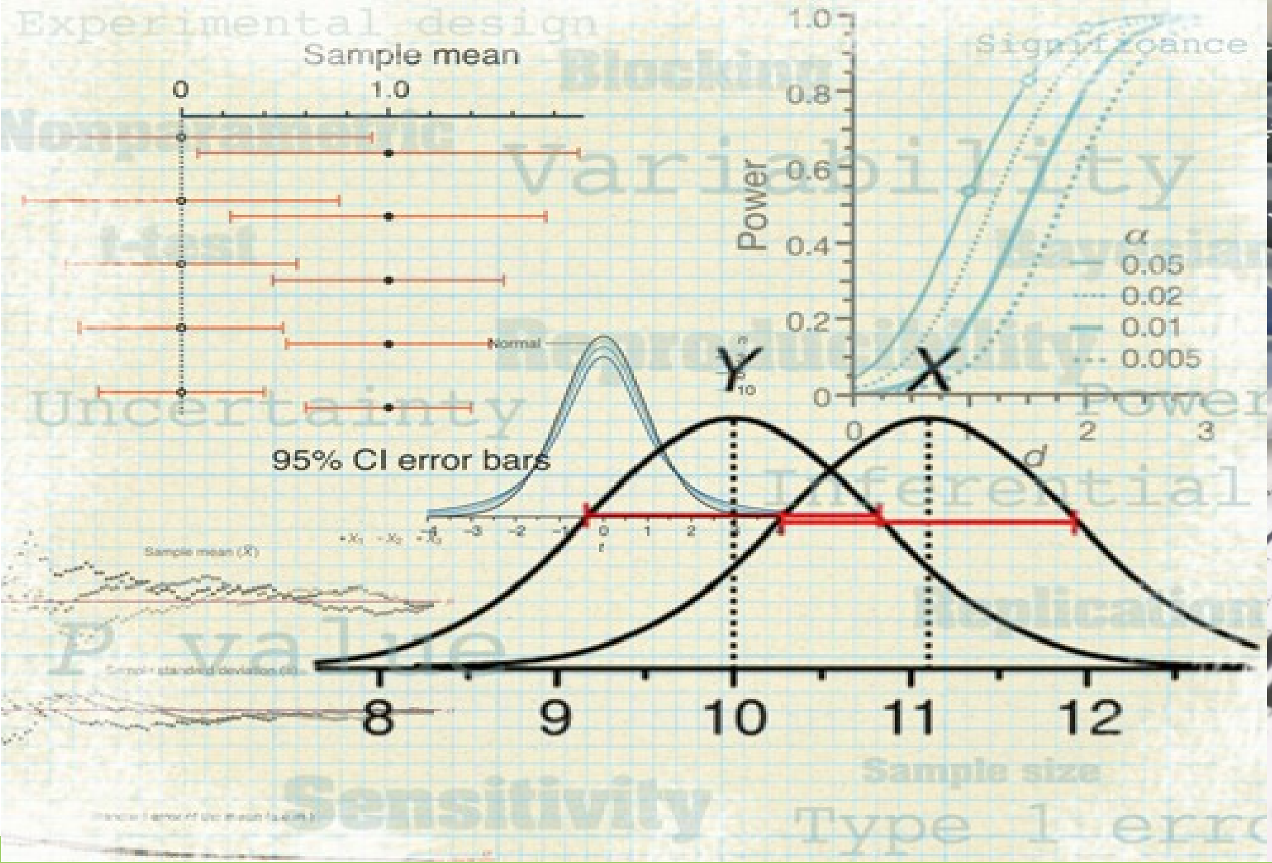
Machine Learning – Basics of Statistics

May 23, 2018

SRIKRISHNA S

Overview

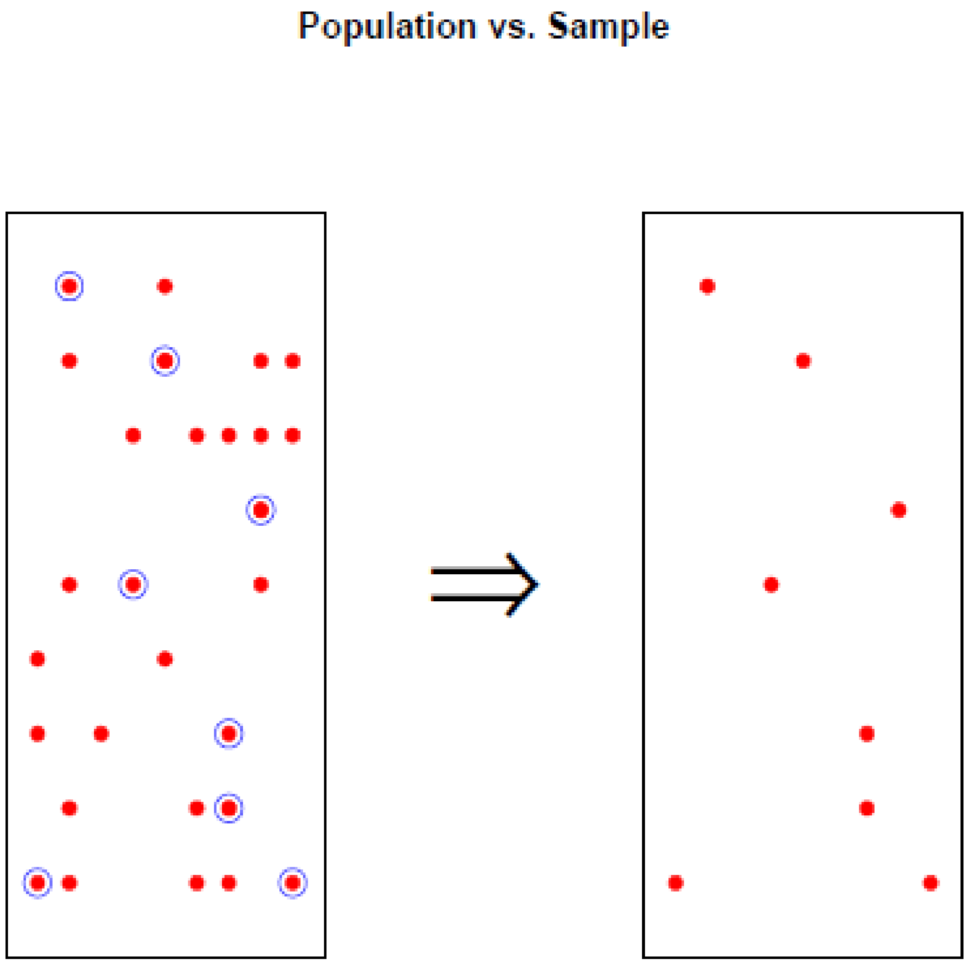
- Introduction
- Process Flow
- Variables & Organization of data
- Plottings
- Measures of Centre
- Measures of Variation
- Probability Distribution function(s)
- Sampling Distribution
- Estimation
- Hypothesis testing
- Linear regression & R-Studio
- ANOVA


Experimental design
Sample mean
0
1.0
Blocking
Nonparametric
Variability
t-test
Power
1.0
0.8
0.6
0.4
0.2
0
Significance
α
0.05
0.02
0.01
0.005
Bayesian
Reproducibility
Normal
Y
X
Uncertainty
95% CI error bars
0
1
2
3
d
Inferential
P value
Replication
8
9
10
11
12
Sample size
Sensitivity
Type 1 error

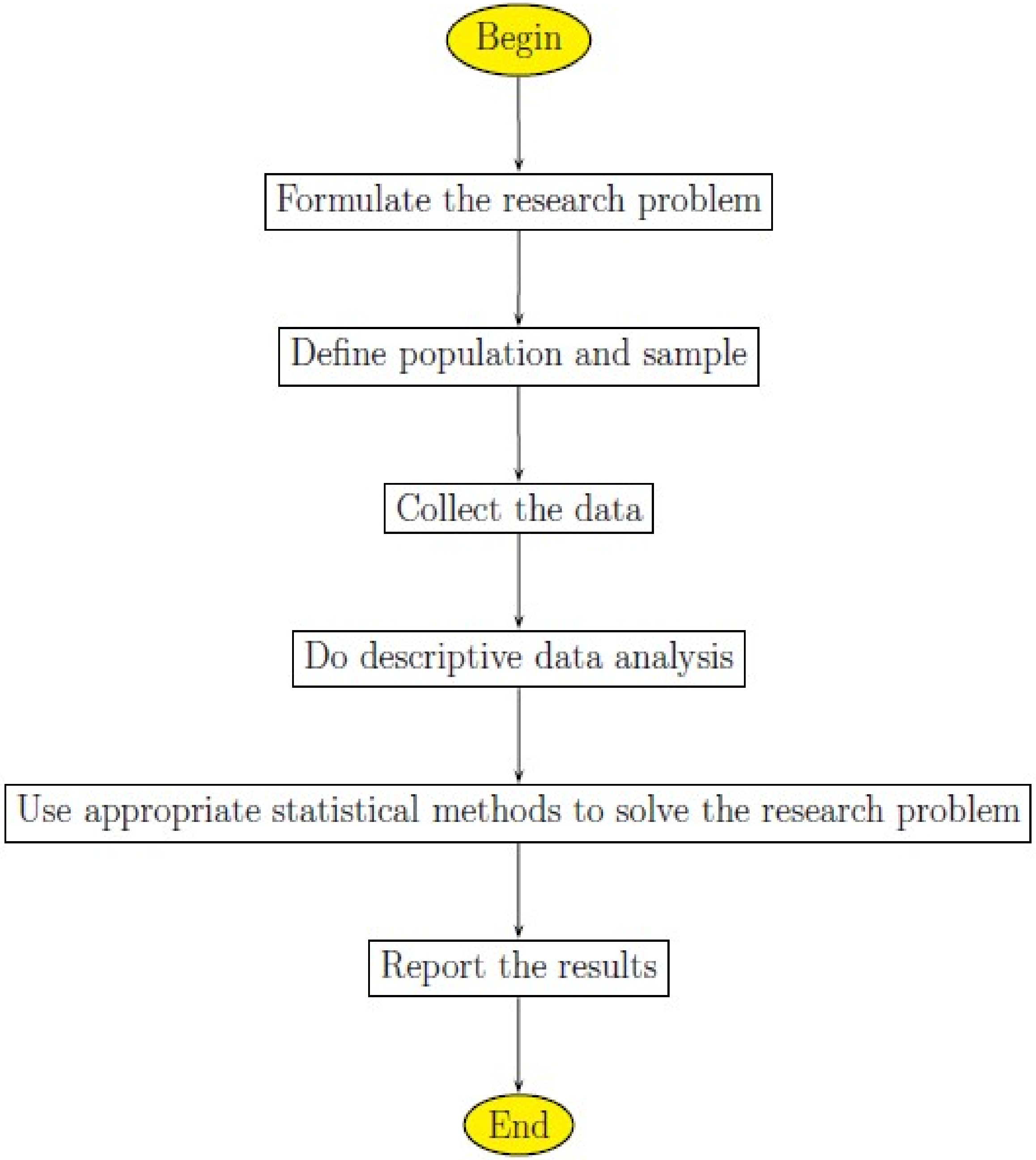
13-02
27-02
13-03
27-03
10-04
24-04
08-05
22-0
13-03
27-03

Introduction

- Statistics consists of body of methods for collecting and analyzing data
- **Design:** Planning and carrying out research studies
- **Description:** Summarizing and exploring data
- **Inference:** Making predictions and generalizing about phenomena represented by data
- **Population** is set of measurements corresponding to entire collection of units for which inferences are to be made
- **sample:** Set of measurements collected from statistical population that are actually collected in the course of an investigation
- Finite population vs Hypothetical population
- Descriptive statistics vs Inferential statistics
- Construction of graphs, charts, tables and calculation of measures like **centre**, **variance** etc.
- Point **estimation**, interval estimation, hypothesis testing based on probability theory



Process Flow



Begin

Formulate the research problem

Define population and sample

Collect the data

Do descriptive data analysis

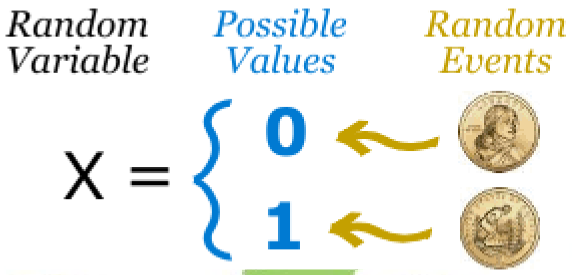
Use appropriate statistical methods to solve the research problem

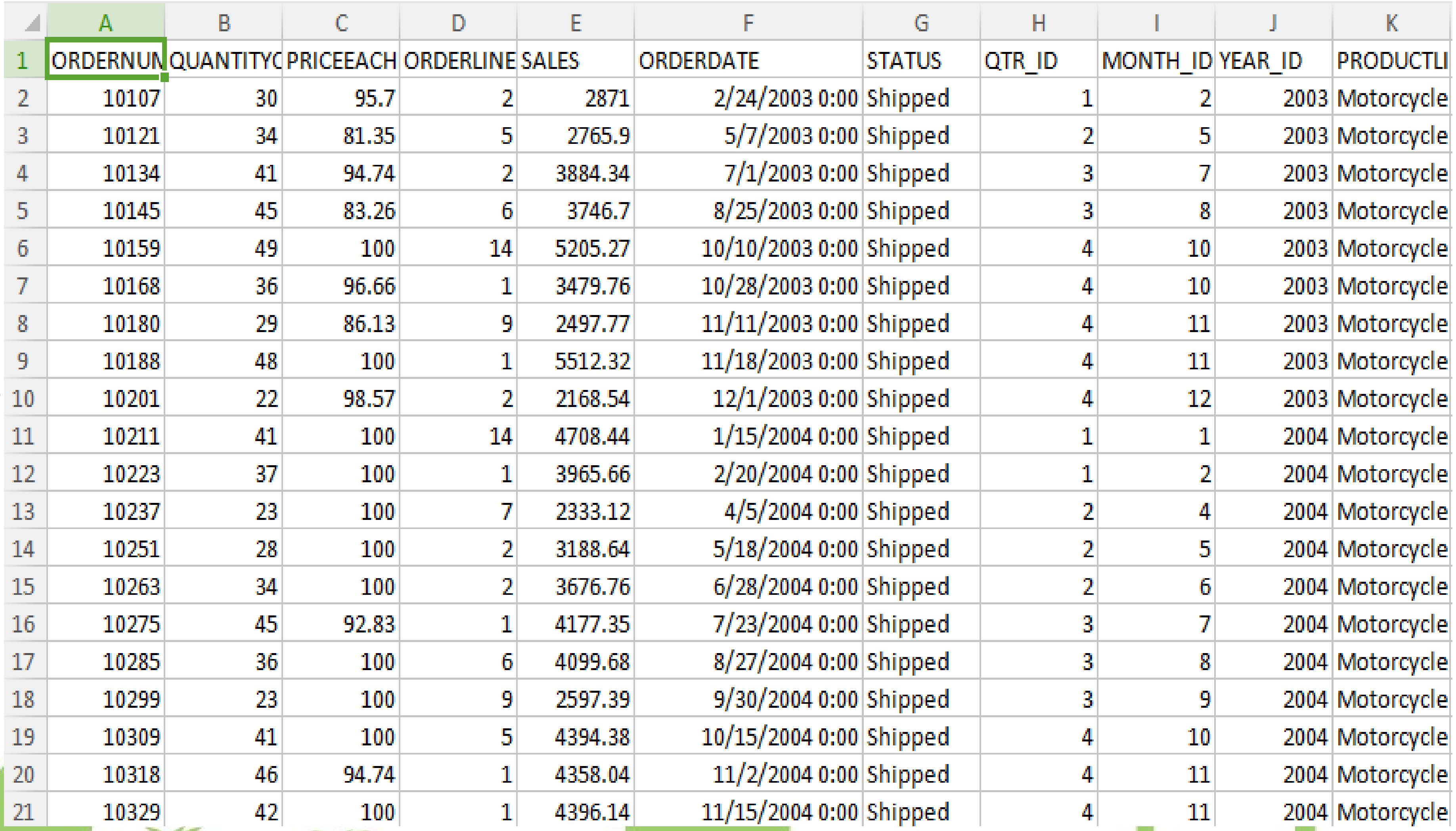
Report the results

End

Variables & organization of data

- Quantitative or numerical
- Qualitative or categorical
- Discrete and continuous
- Interval and ratio scaling
- Nominal and ordinal
- Discrete random variable
- Continuos random variable
- Frequency distribution

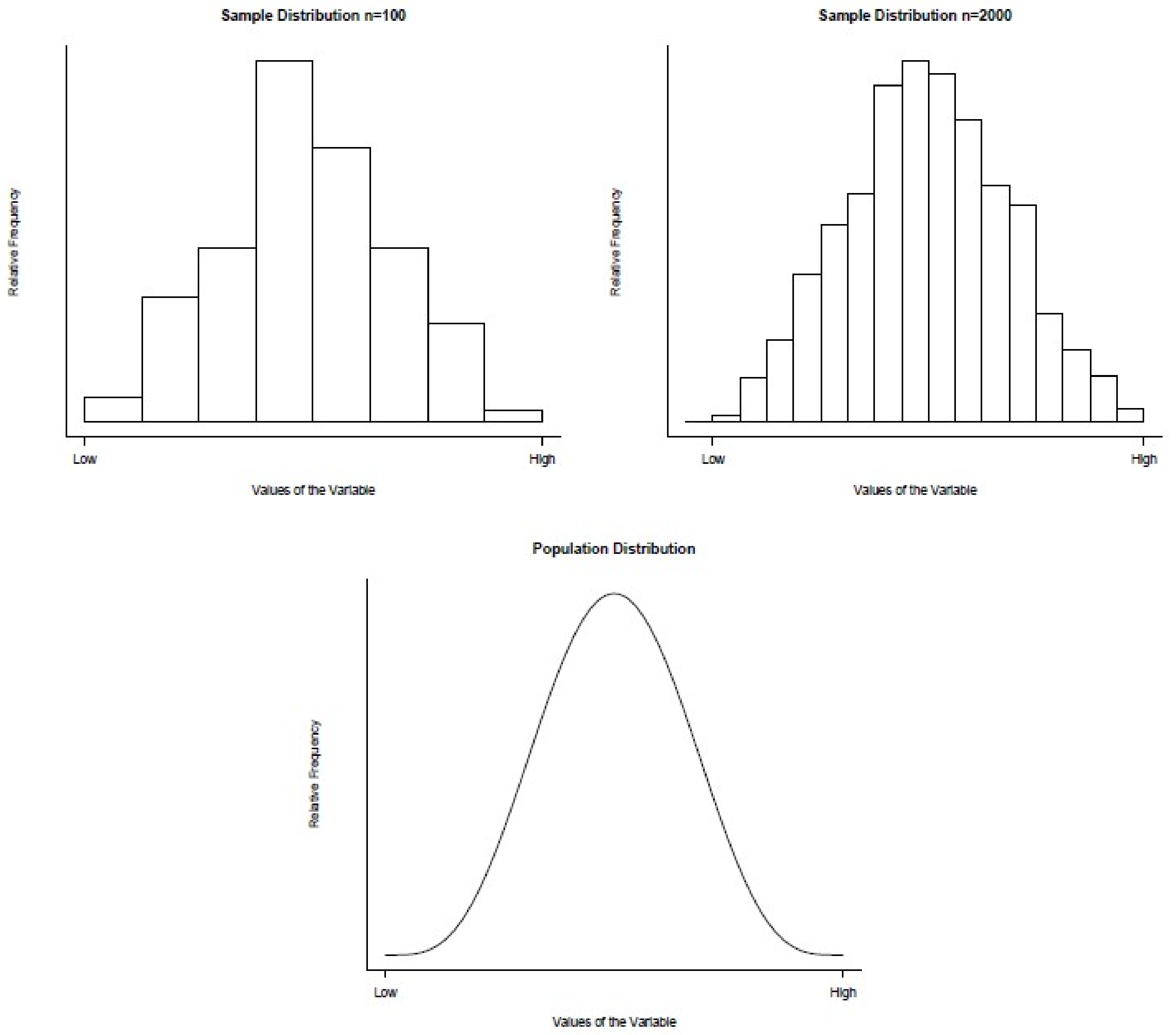




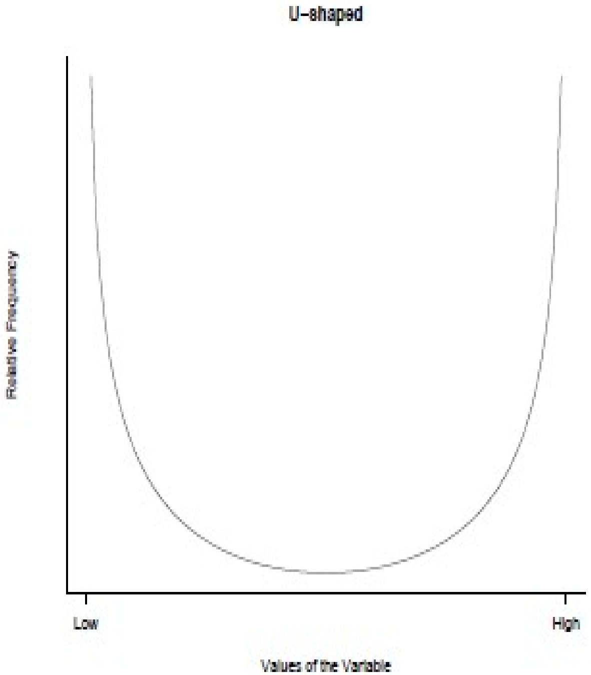
	A	B	C	D	E	F	G	H	I	J	K
1	ORDERNUN	QUANTITYC	PRICEEACH	ORDERLINE	SALES	ORDERDATE	STATUS	QTR_ID	MONTH_ID	YEAR_ID	PRODUCTLI
2	10107	30	95.7	2	2871	2/24/2003 0:00	Shipped	1	2	2003	Motorcycle
3	10121	34	81.35	5	2765.9	5/7/2003 0:00	Shipped	2	5	2003	Motorcycle
4	10134	41	94.74	2	3884.34	7/1/2003 0:00	Shipped	3	7	2003	Motorcycle
5	10145	45	83.26	6	3746.7	8/25/2003 0:00	Shipped	3	8	2003	Motorcycle
6	10159	49	100	14	5205.27	10/10/2003 0:00	Shipped	4	10	2003	Motorcycle
7	10168	36	96.66	1	3479.76	10/28/2003 0:00	Shipped	4	10	2003	Motorcycle
8	10180	29	86.13	9	2497.77	11/11/2003 0:00	Shipped	4	11	2003	Motorcycle
9	10188	48	100	1	5512.32	11/18/2003 0:00	Shipped	4	11	2003	Motorcycle
10	10201	22	98.57	2	2168.54	12/1/2003 0:00	Shipped	4	12	2003	Motorcycle
11	10211	41	100	14	4708.44	1/15/2004 0:00	Shipped	1	1	2004	Motorcycle
12	10223	37	100	1	3965.66	2/20/2004 0:00	Shipped	1	2	2004	Motorcycle
13	10237	23	100	7	2333.12	4/5/2004 0:00	Shipped	2	4	2004	Motorcycle
14	10251	28	100	2	3188.64	5/18/2004 0:00	Shipped	2	5	2004	Motorcycle
15	10263	34	100	2	3676.76	6/28/2004 0:00	Shipped	2	6	2004	Motorcycle
16	10275	45	92.83	1	4177.35	7/23/2004 0:00	Shipped	3	7	2004	Motorcycle
17	10285	36	100	6	4099.68	8/27/2004 0:00	Shipped	3	8	2004	Motorcycle
18	10299	23	100	9	2597.39	9/30/2004 0:00	Shipped	3	9	2004	Motorcycle
19	10309	41	100	5	4394.38	10/15/2004 0:00	Shipped	4	10	2004	Motorcycle
20	10318	46	94.74	1	4358.04	11/2/2004 0:00	Shipped	4	11	2004	Motorcycle
21	10329	42	100	1	4396.14	11/15/2004 0:00	Shipped	4	11	2004	Motorcycle

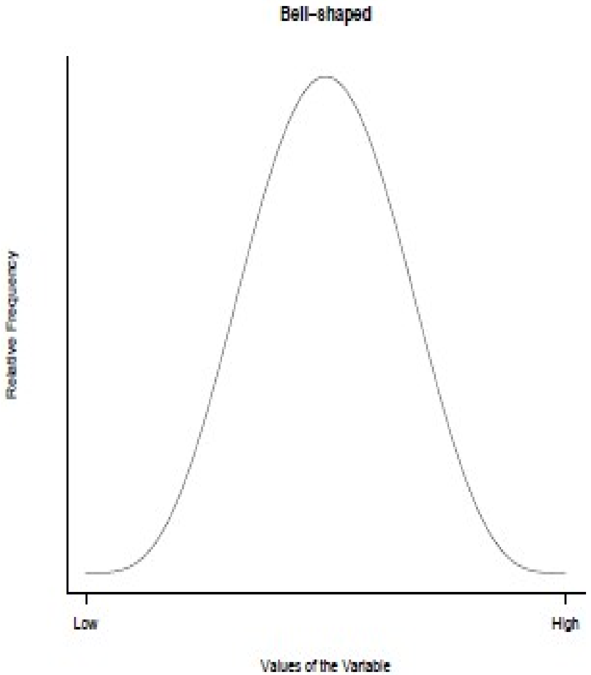
Plottings

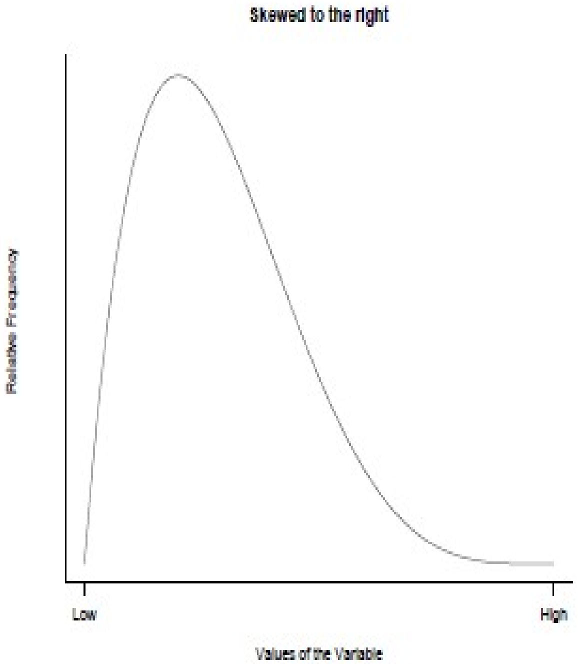
- Histograms
- Boxplots
- Scatter Plots
- Pie charts
- dot plot
- stem plot

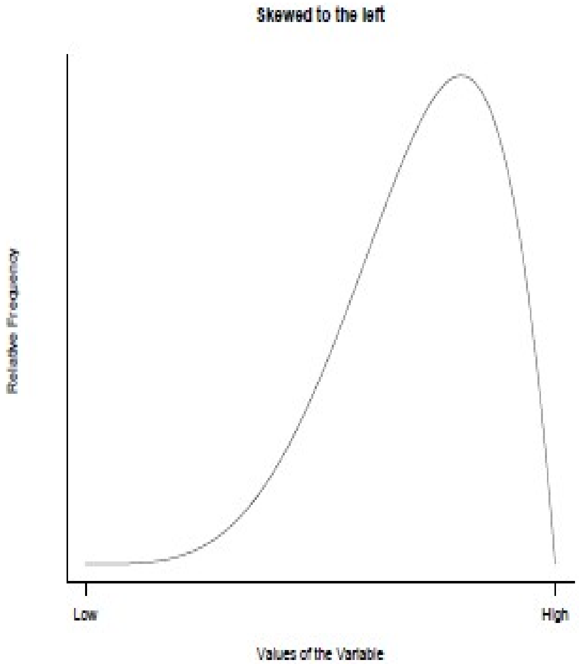


Observations after plotting



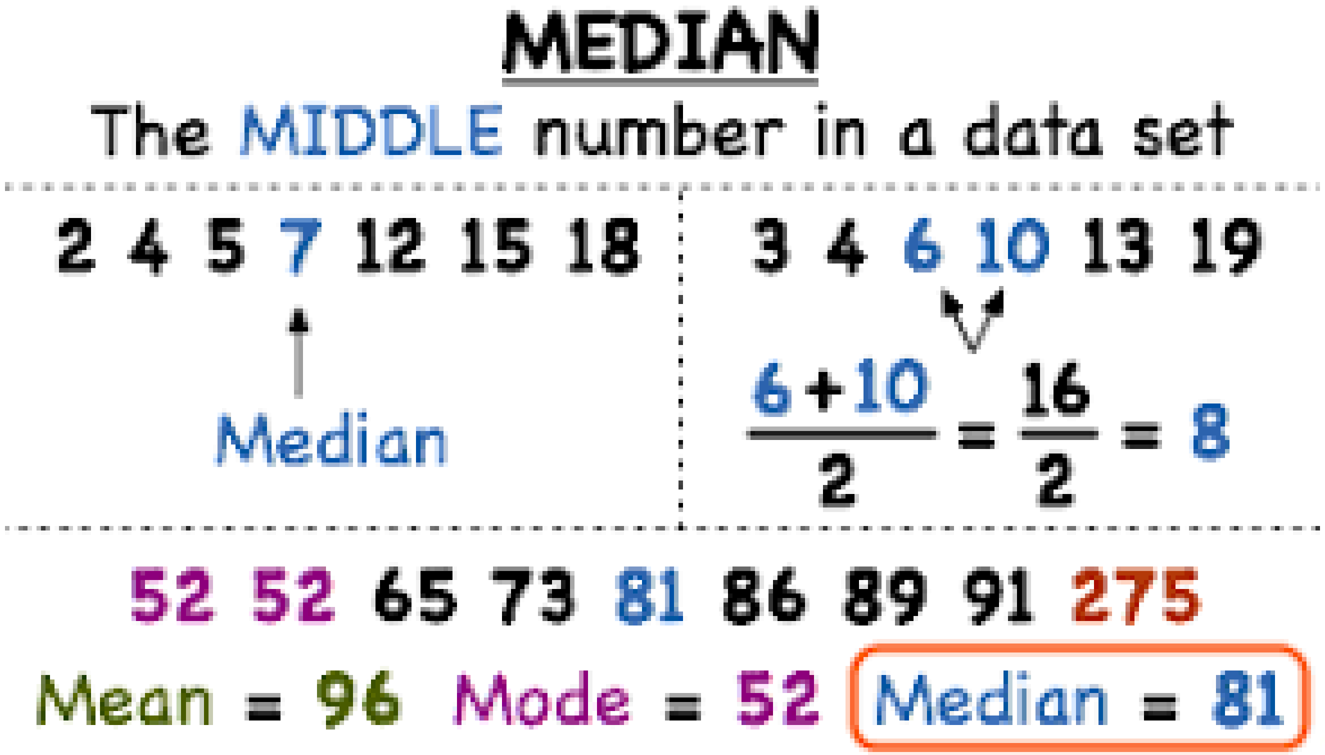






Measures of centre

- Mode
- Median
- Mean
- Which measure to chooose ?



MEDIAN

The MIDDLE number in a data set

2 4 5 7 12 15 18

Median

3 4 6 10 13 19

$$\frac{6+10}{2} = \frac{16}{2} = 8$$

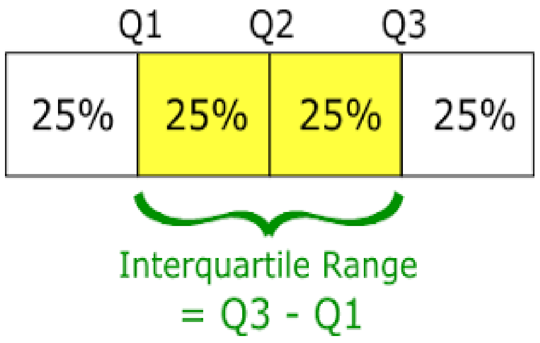
52 52 65 73 81 86 89 91 275

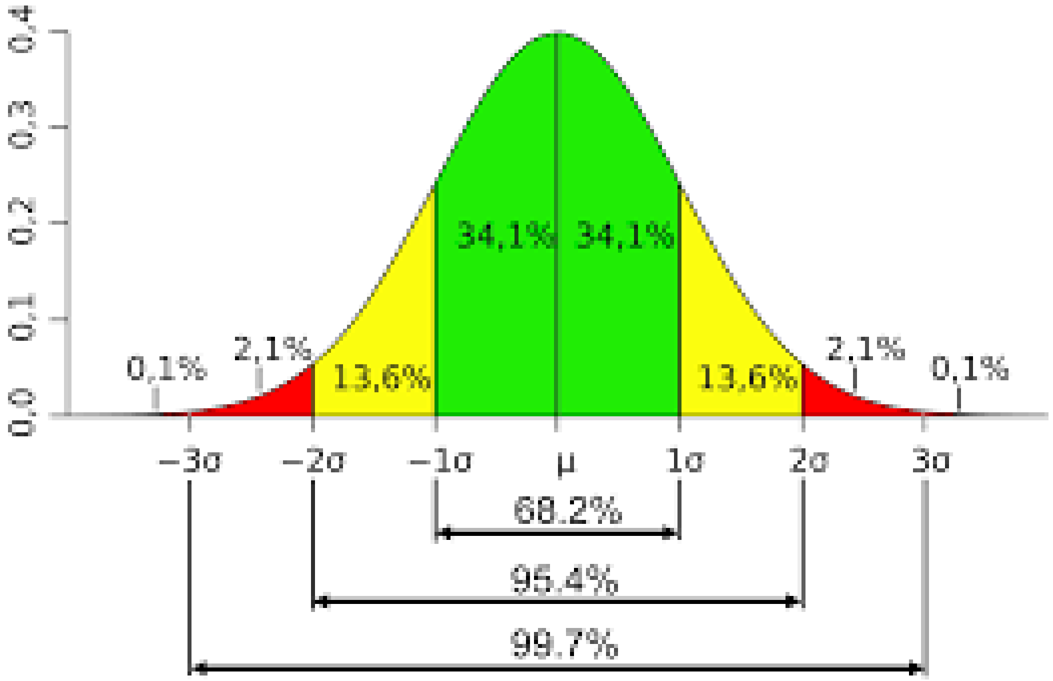
Mean = 96 Mode = 52 Median = 81

Measures of variation

- Range
- Inter quartile range
- Five number summary box-plot {min, Q1, Q2, Q3, max}
- standard deviation

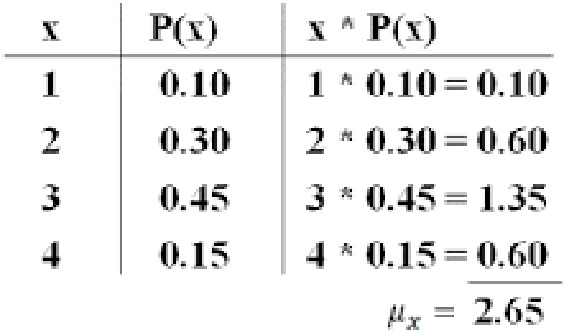
$$\mathrm{SD} = \sqrt{\frac{\sum |x - \bar{x}|^2}{n}}$$





Probability Distribution Function

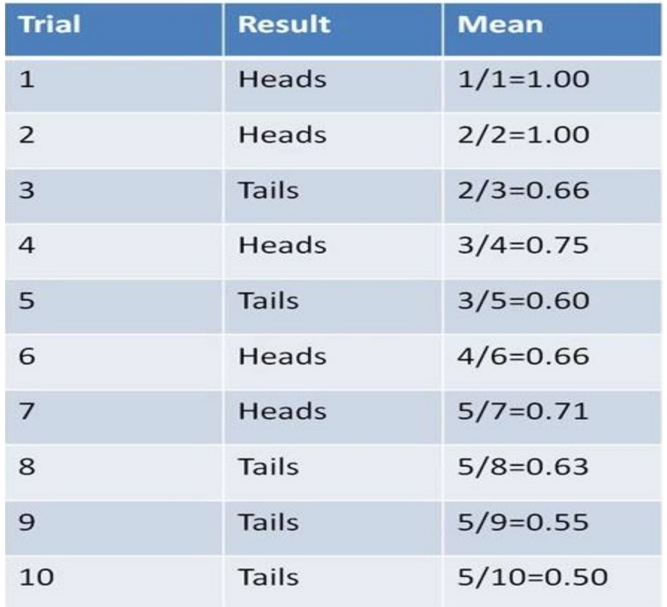
- Law of large numbers
- random variable
- Continous random variable
- Discrete random variable
- Mean and standard deviation of random variable
- Variance of discrete random variable
- Mean, SD, Variance of continuous random variable
- Normal Distribution
- Binomial, Bernouli, Poisson etc. distributions



x	P(x)	x * P(x)
1	0.10	1 * 0.10 = 0.10
2	0.30	2 * 0.30 = 0.60
3	0.45	3 * 0.45 = 1.35
4	0.15	4 * 0.15 = 0.60
		$\mu_x =$ 2.65

Mean Formula:

$$\mu_x = \sum[x * P(x)]$$



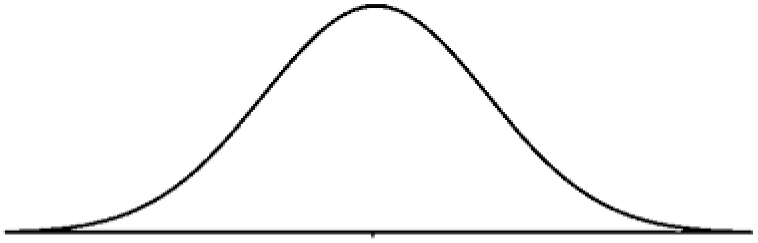
Trial	Result	Mean
1	Heads	1/1=1.00
2	Heads	2/2=1.00
3	Tails	2/3=0.66
4	Heads	3/4=0.75
5	Tails	3/5=0.60
6	Heads	4/6=0.66
7	Heads	5/7=0.71
8	Tails	5/8=0.63
9	Tails	5/9=0.55
10	Tails	5/10=0.50

- The variance of a discrete random variable is:

$$\boxed{\sigma_X^2 = \sum_{All\,x} (x - \mu_X)^2 \, p(x)}$$

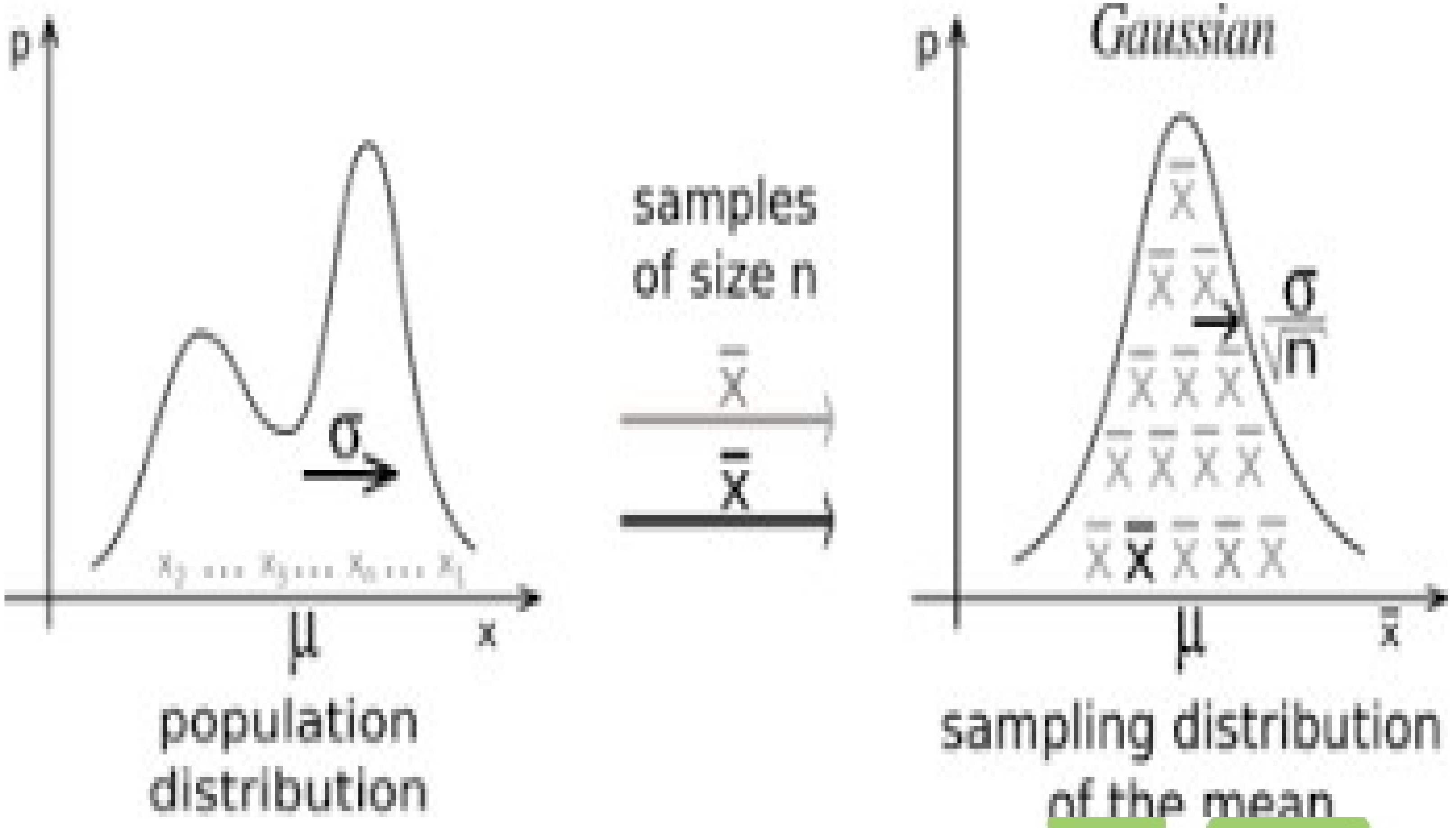
- The standard deviation is the square root of the variance.

$$\boxed{\sigma_X = \sqrt{\sigma_X^2}}$$



Sampling Distributions

- Sample distribution {
- Sample means, samples standard deviation etc. -> standard error
- Central limit theoram {whatever the distributions, but sample means distribution is normal}

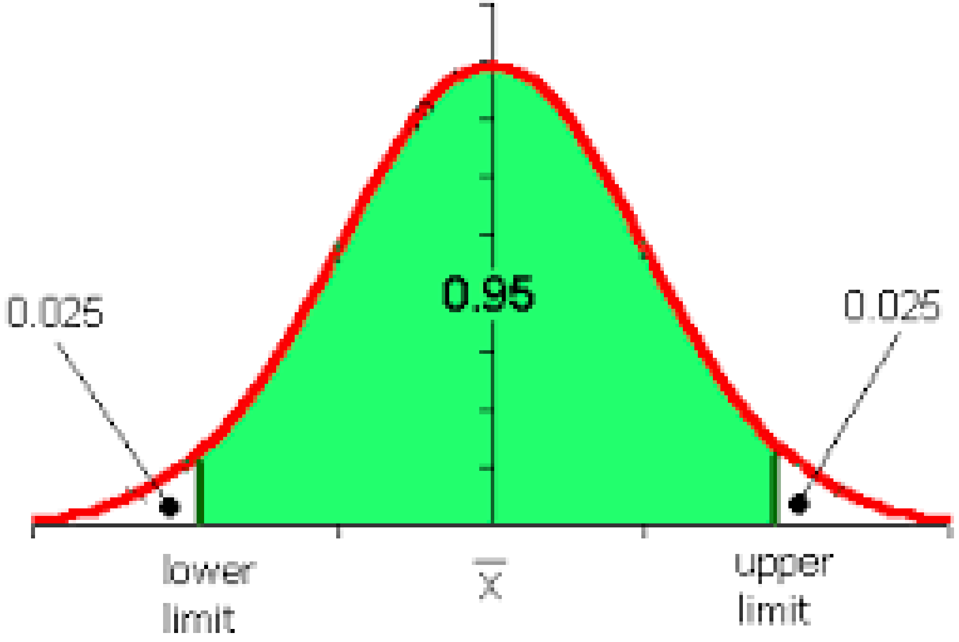


Estimation

- Point estimation - Estimating population data point using sample data
- Interval estimation -
- Confidence intervals
- Large sample confidence interval
- Small sample confidence interval
- Degrees of freedom

[N-> N-1]

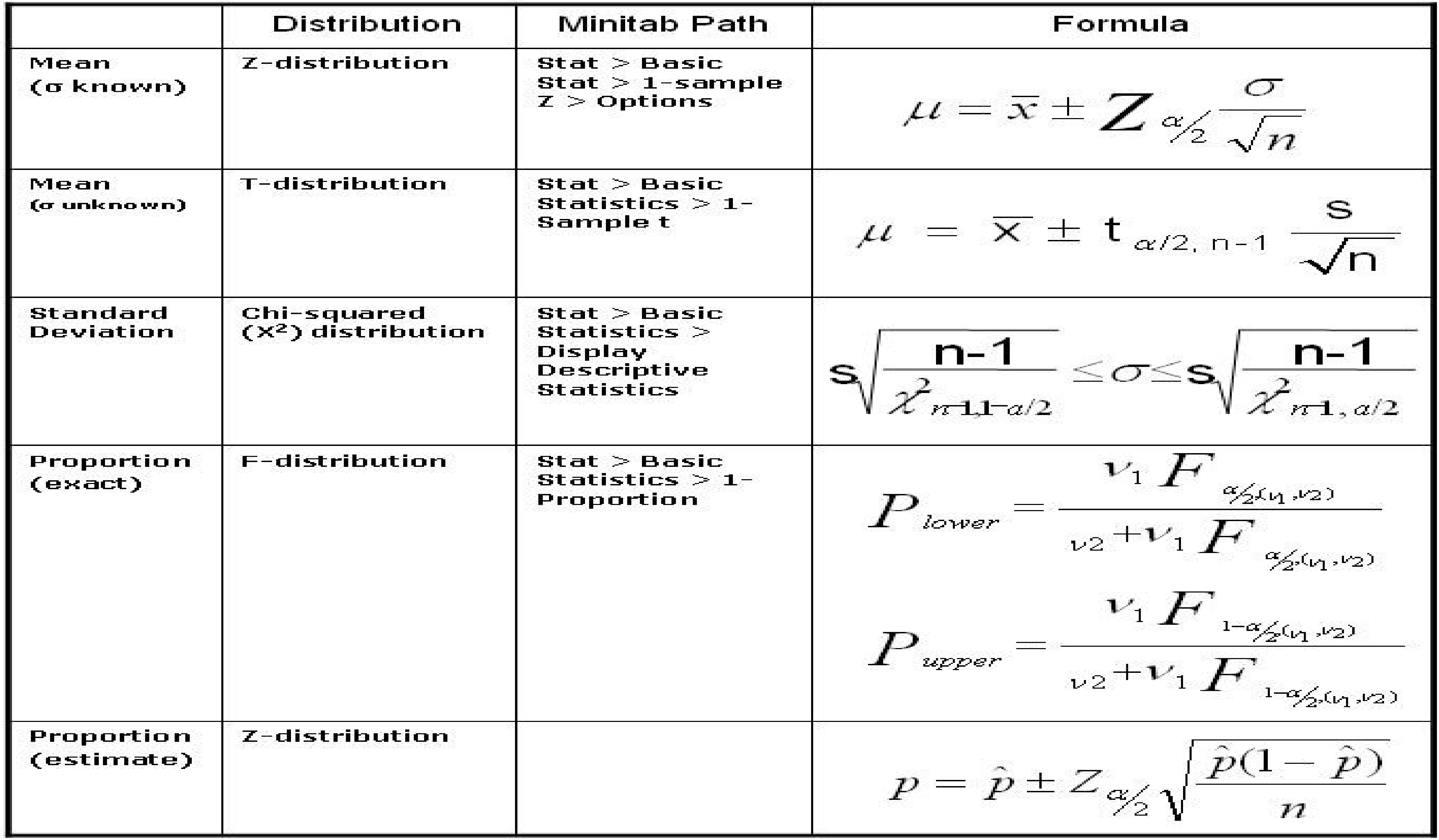
[N1, N2 -> N1+N2-1]



Hypothesis testing

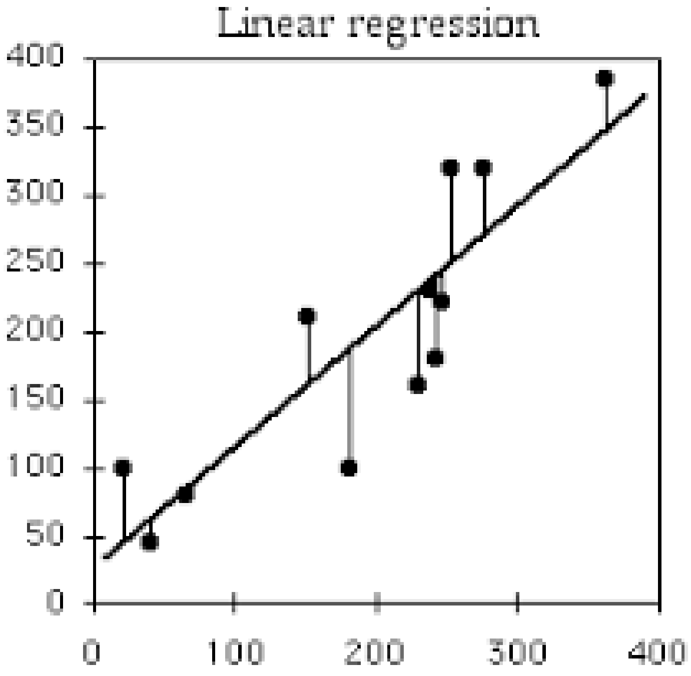
- hypothesis
- assumptions
- hypothesis {Null hypothesis & alternate hypothesis}
- test stastic
- p-value
- conclusion

ANOVA



	Distribution	Minitab Path	Formula
Mean (σ known)	Z-distribution	Stat > Basic Stat > 1-sample Z > Options	$\mu = \bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$
Mean (σ unknown)	T-distribution	Stat > Basic Statistics > 1- Sample t	$\mu = \bar{x} \pm t_{\alpha/2,\, n-1} \frac{s}{\sqrt{n}}$
Standard Deviation	Chi-squared (X^2) distribution	Stat > Basic Statistics > Display Descriptive Statistics	$s\sqrt{\frac{n-1}{\chi^2_{n-1,1-\alpha/2}}} \leq \sigma \leq s\sqrt{\frac{n-1}{\chi^2_{n-1,\alpha/2}}}$
Proportion (exact)	F-distribution	Stat > Basic Statistics > 1- Proportion	$P_{lower} = \frac{\nu_1 F_{\alpha/2(\nu_1,\nu_2)}}{\nu_2 + \nu_1 F_{\alpha/2,(\nu_1,\nu_2)}}$ $P_{upper} = \frac{\nu_1 F_{1-\alpha/2(\nu_1,\nu_2)}}{\nu_2 + \nu_1 F_{1-\alpha/2,(\nu_1,\nu_2)}}$
Proportion (estimate)	Z-distribution		$p = \hat{p} \pm Z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Linear Regression



Dependent Variable

Population Y intercept

Population Slope Coefficient

Independent Variable

Random Error term

$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}$$

The World Lies Within !!