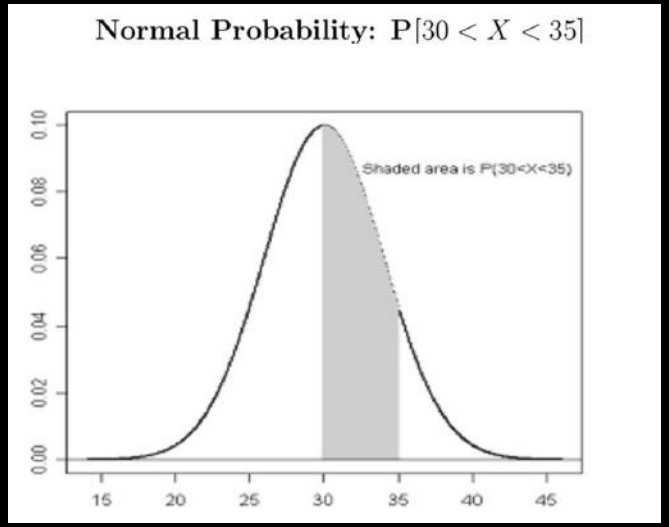
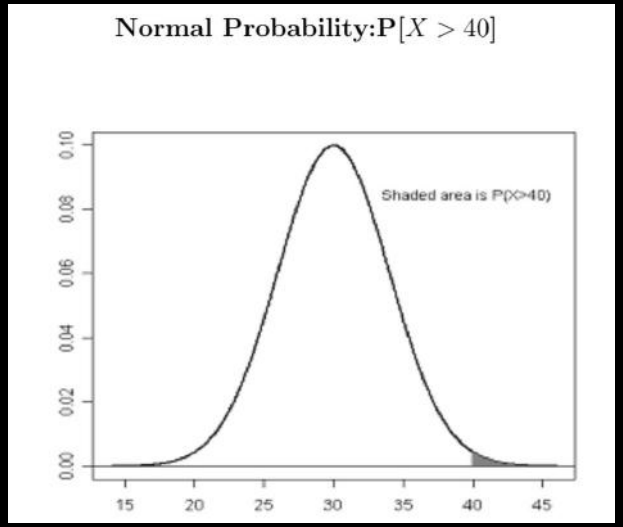
Computing for Medicine

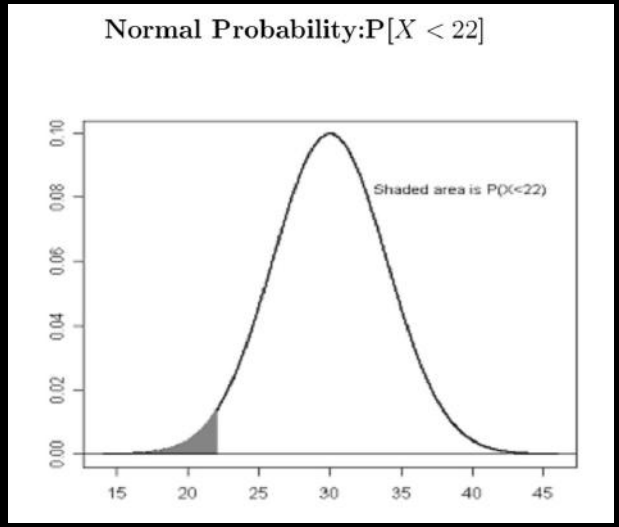
Data Science: Data Science 101 Contd

Tavpritesh Sethi

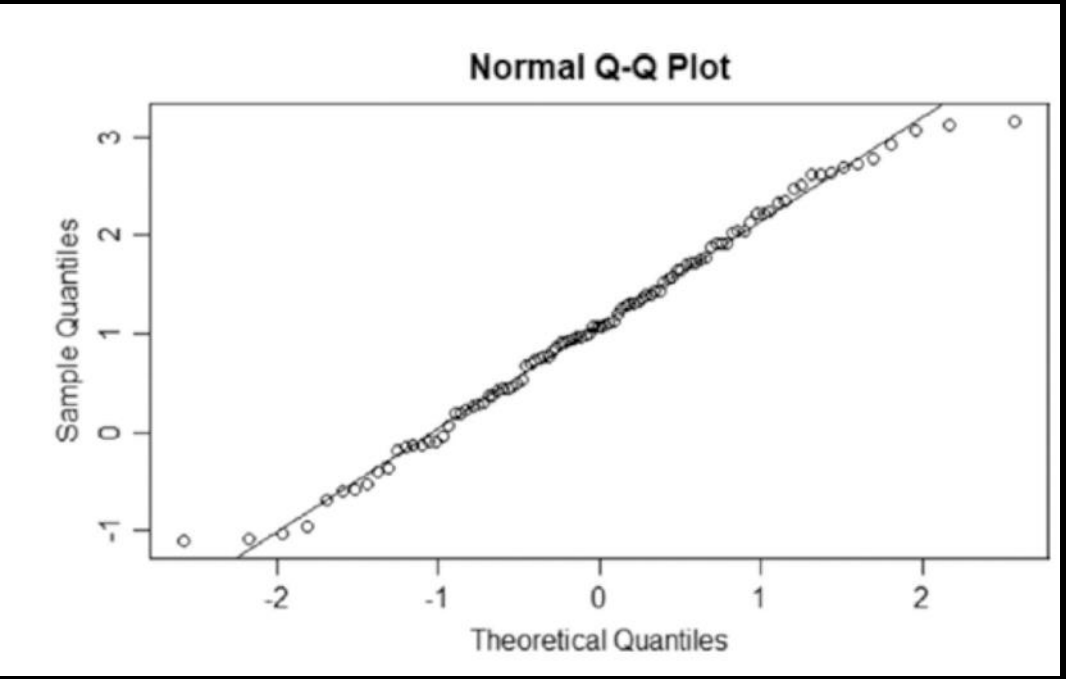
Quantiles (Rank Statistics)







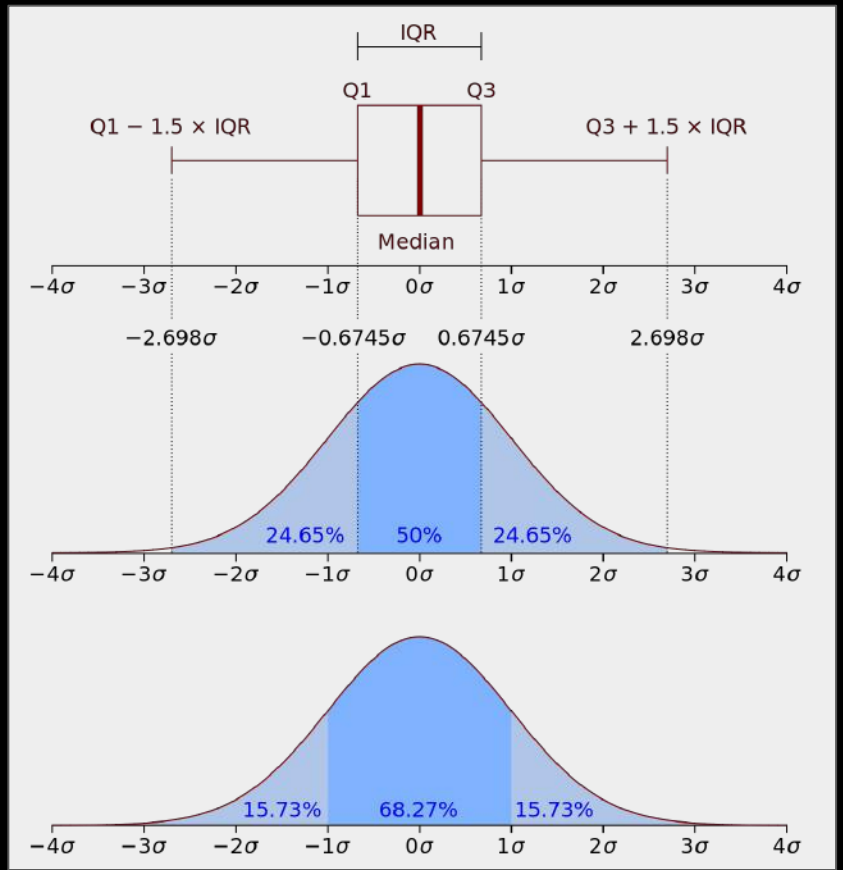
Quantile Plots



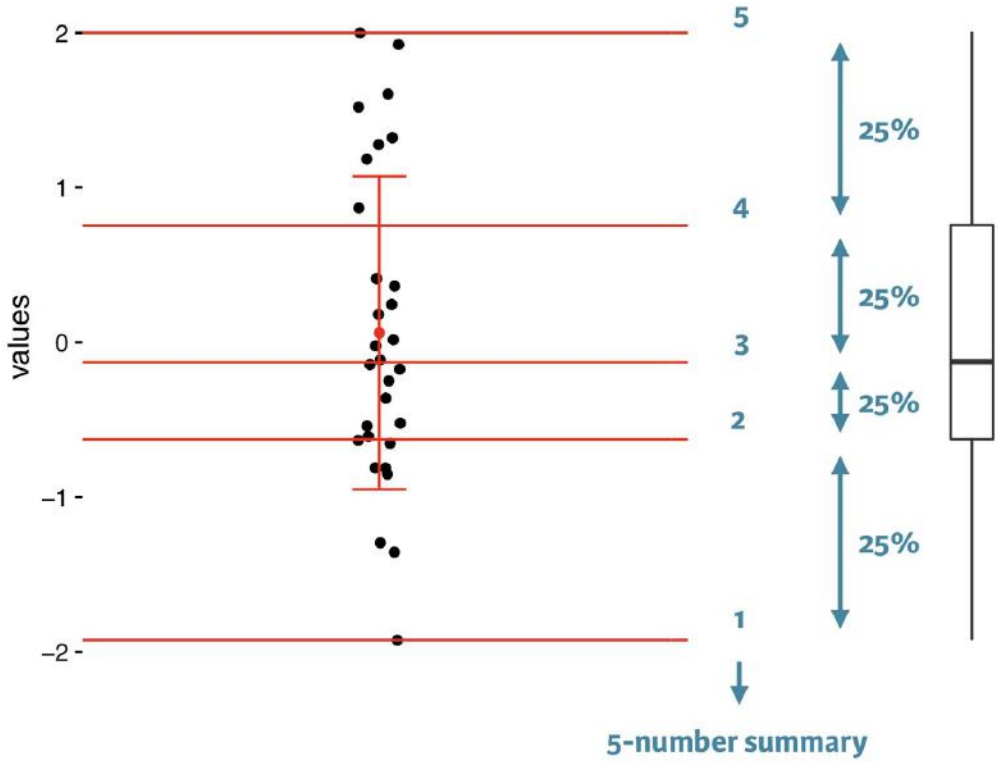

```
xuni = runif(100); qqnorm(xuni)
xnorm = rnorm(100); qqnorm(rnorm)
shapiro.test(xuni)
shapiro.test(xnorm)
```

Interquartile Range

- Midspread, H-spread
- Rank-order statistic



Box and Whisker Plots




```
x <- rnorm(1000)

mean(x)
sd(x)

summary(x)

boxplot(x)
```

```
hist(x)
rug(x)
```

Population and Sample Statistics: Sense of Spread

range(wts) # minimum and maximum values

diff(range(wts)) # the distance between values

QUIZ: What is the problem with the range as a measure of spread?

$$sample\ variance = s^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2.$$

Population Variance vs Sample Variance

$$\sigma^2 = \frac{\sum(X-u)^2}{N}$$

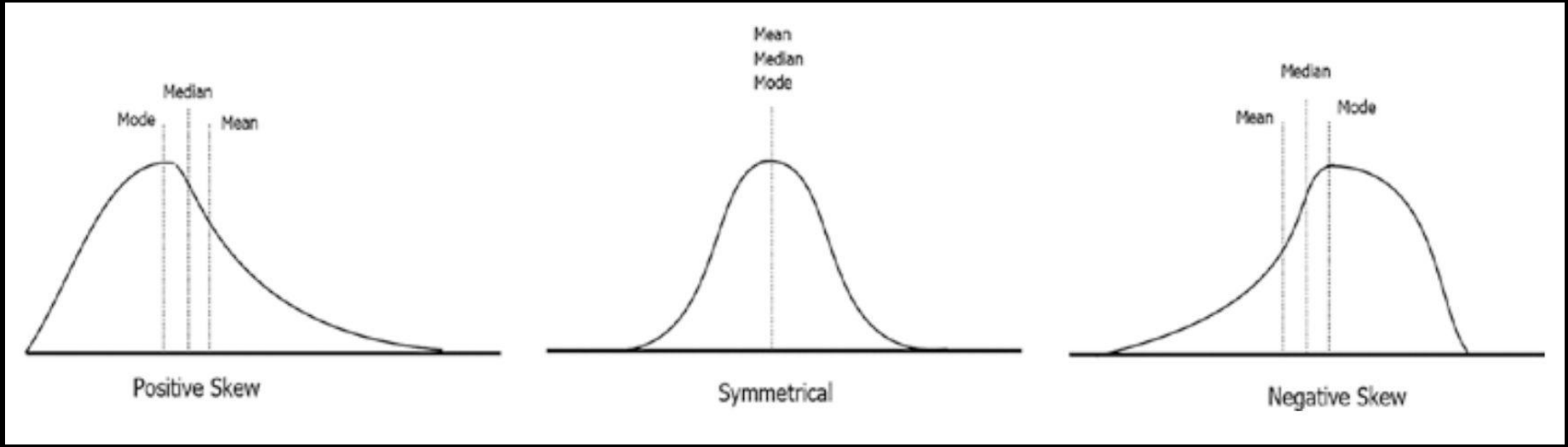
$$S^2 = \frac{\sum(X-\bar{X})^2}{n-1}$$

Population SD and Sample SD

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

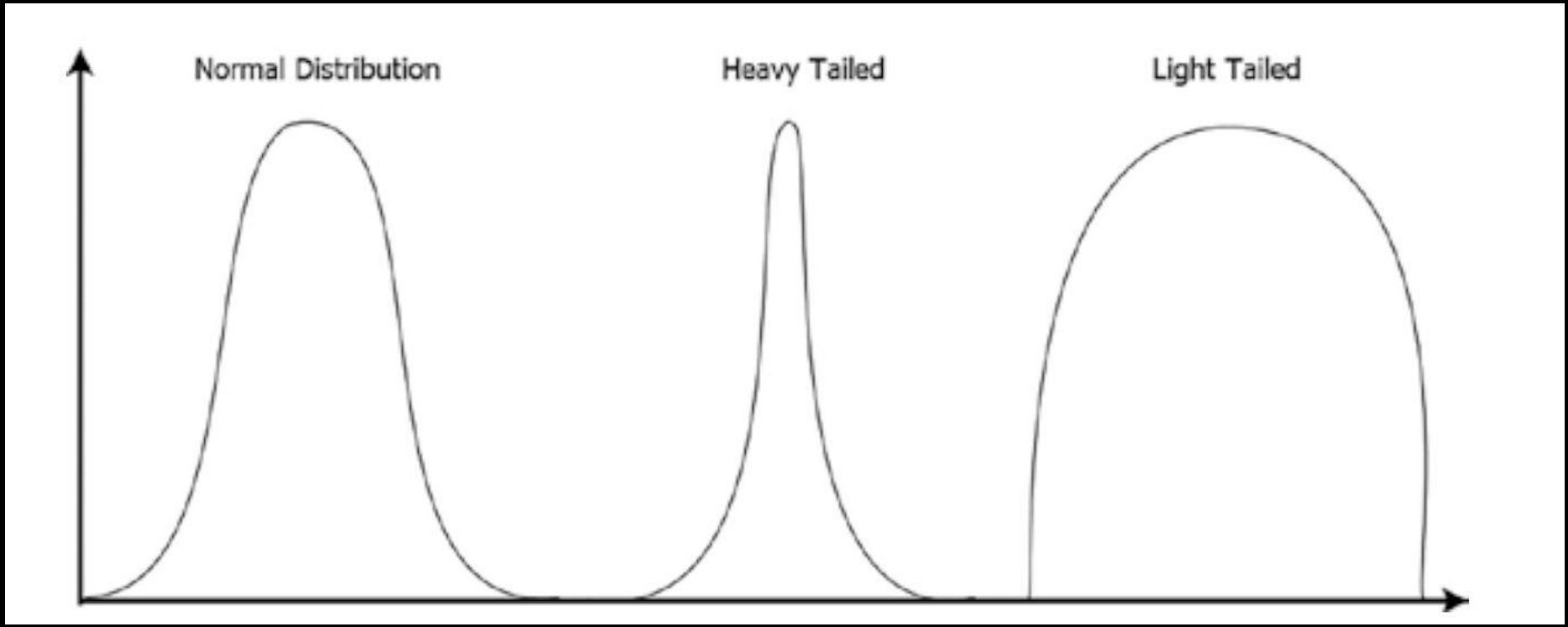
Skewness


Mode
Median
Mean
Positive Skew
Mean
Median
Mode
Symmetrical
Mean
Median
Mode
Negative Skew

Skewness

$$\textit{sample skewness} = \sqrt{n}\,\frac{\sum(x_i - \bar{x})^3}{(\sum(x_i - \bar{x})^2)^{3/2}} = \frac{1}{n}\sum z_i^3.$$

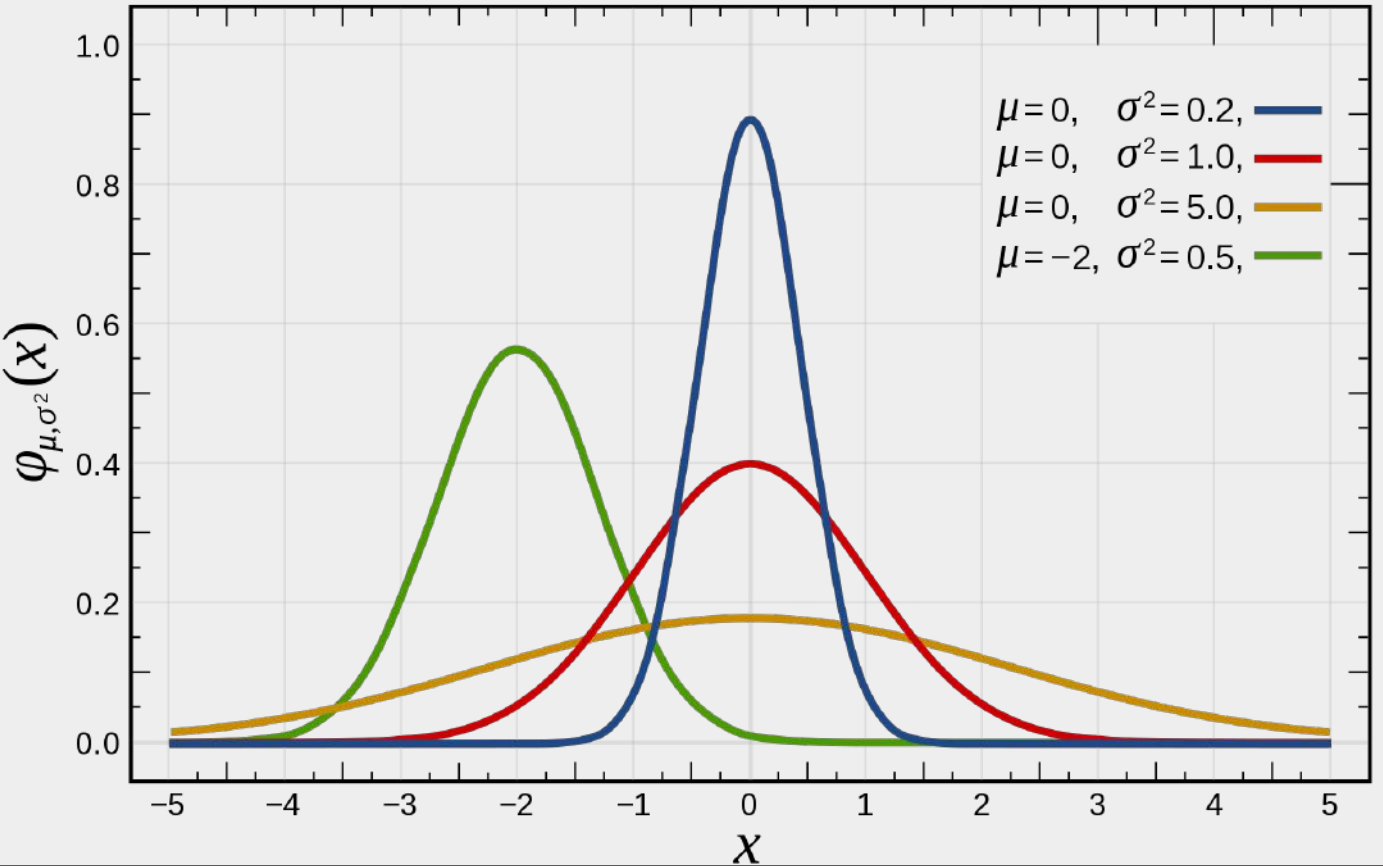
Kurtosis



Kurtosis

$$\textit{sample excess kurtosis} = n\frac{\sum(x_i - \bar{x})^4}{(\sum(x_i - \bar{x})^2)^2} - 3 = \frac{1}{n}\sum z_i^4 - 3.$$

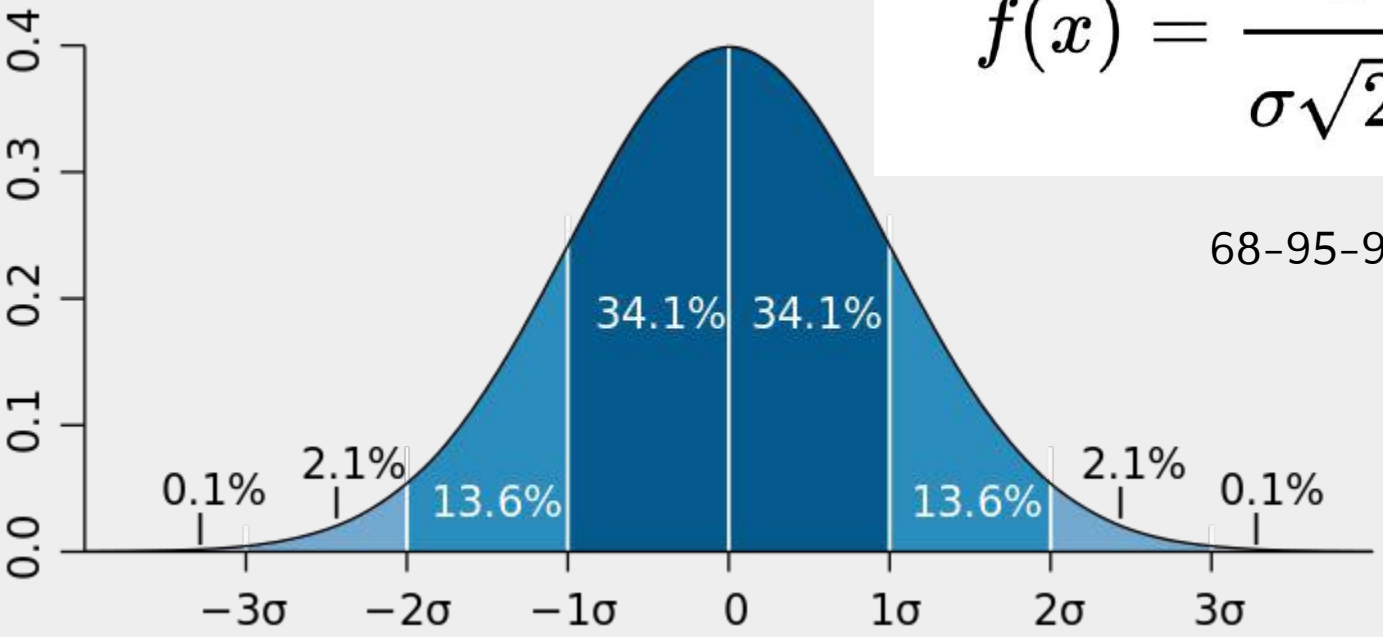
Normal Distribution



What characterizes “Normal”?

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

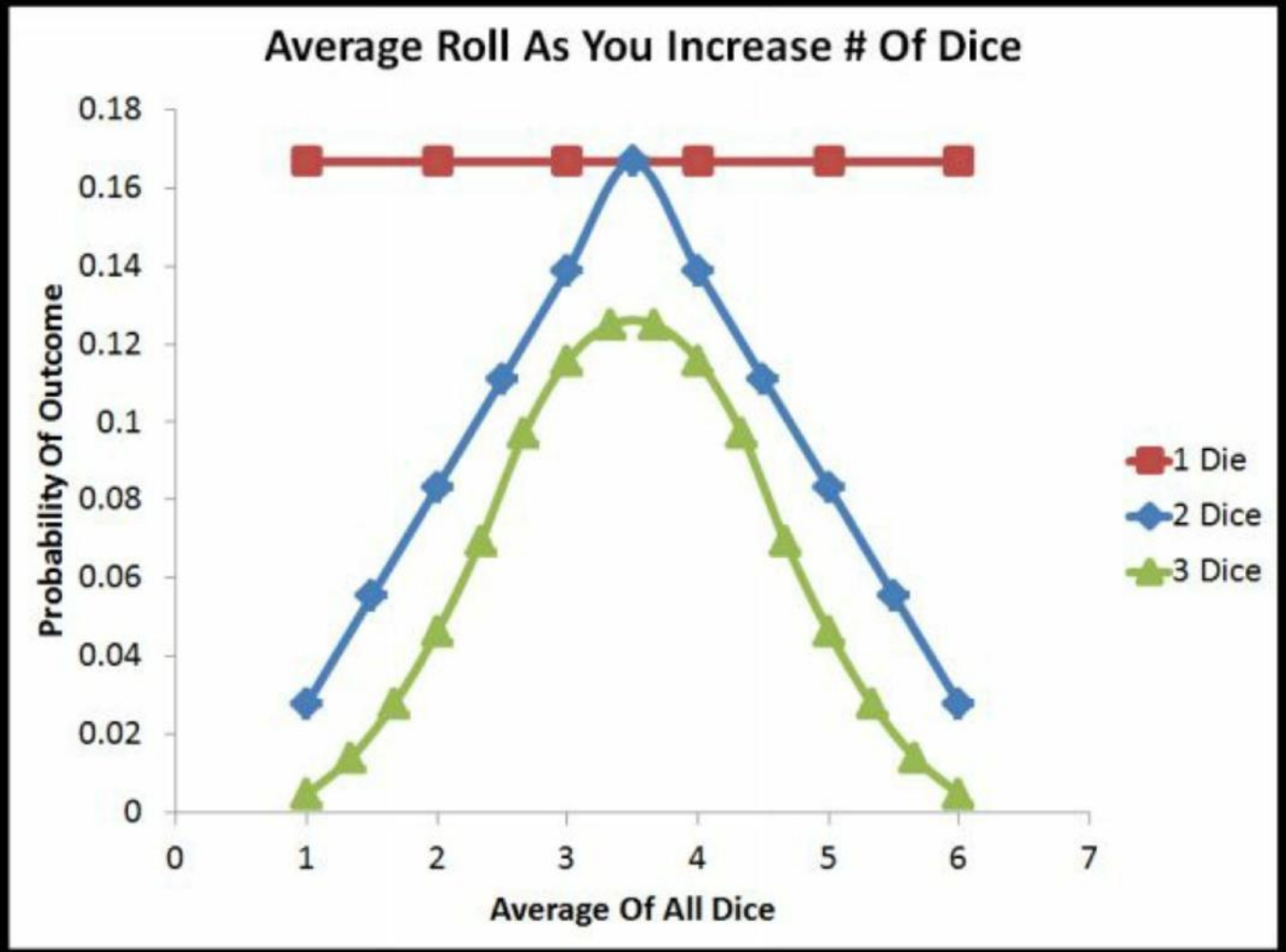
68–95–99: Three Sigma Rule

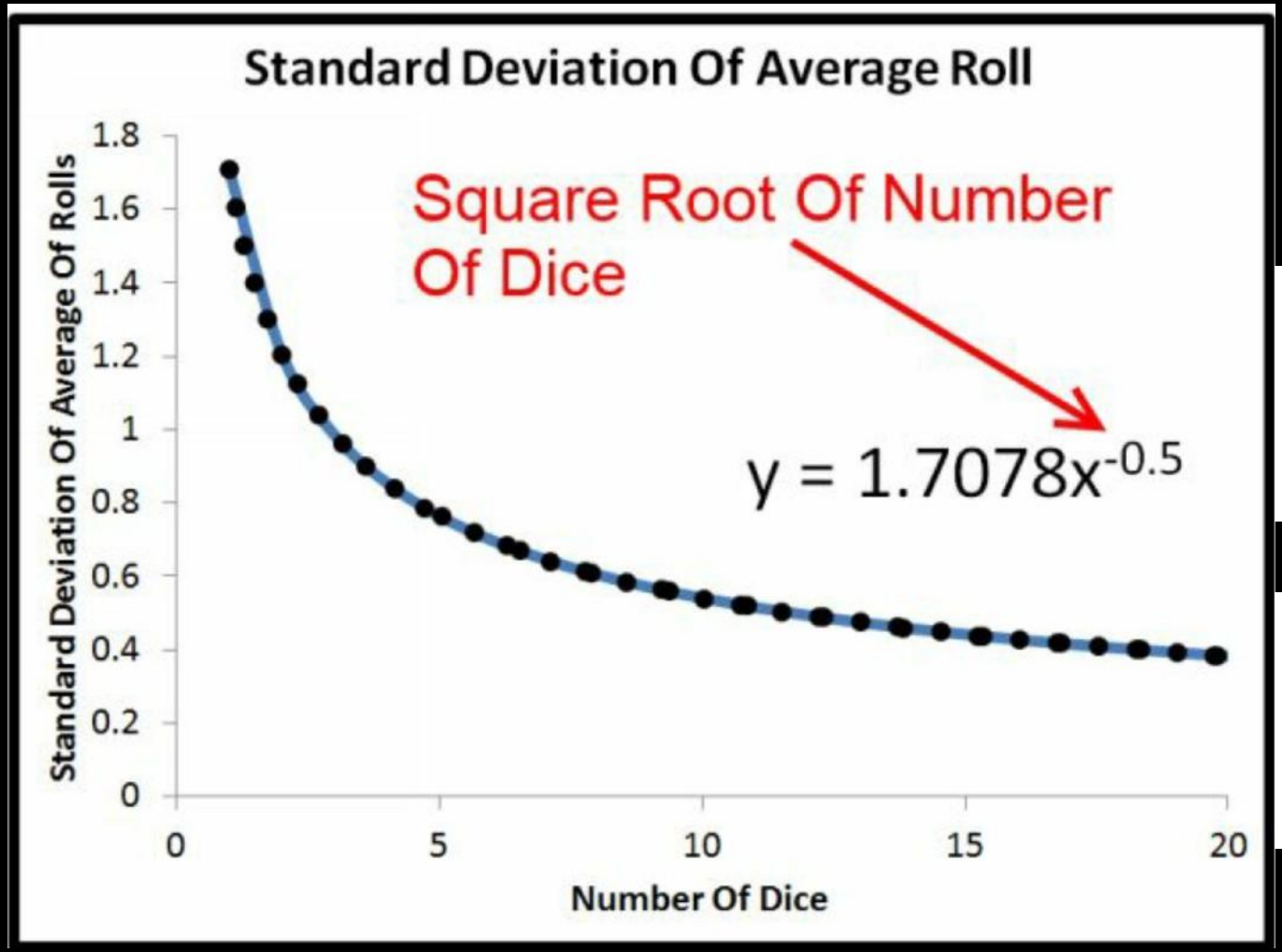


Central Limit Theorem

For ANY distribution

“Given a large enough sample size, the distribution of mean is approximately normal **regardless** of the form of the population distribution.”

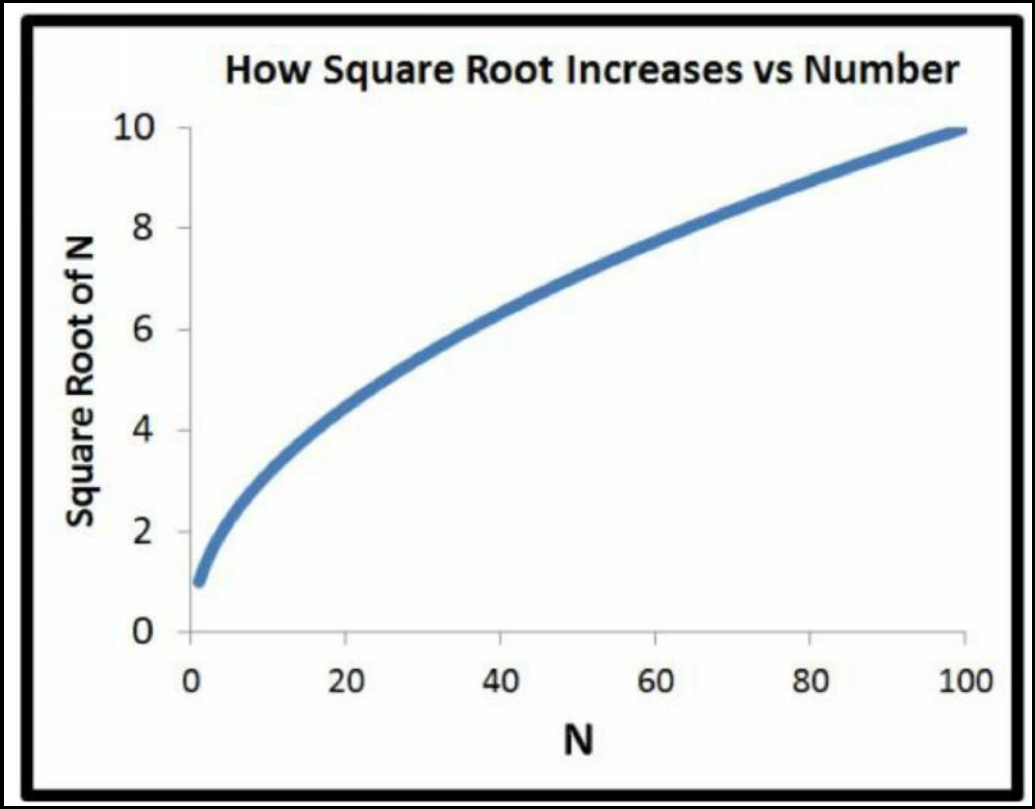

Average Roll As You Increase # Of Dice
Probability Of Outcome
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0
1
2
3
4
5
6
7
Average Of All Dice
1 Die
2 Dice
3 Dice



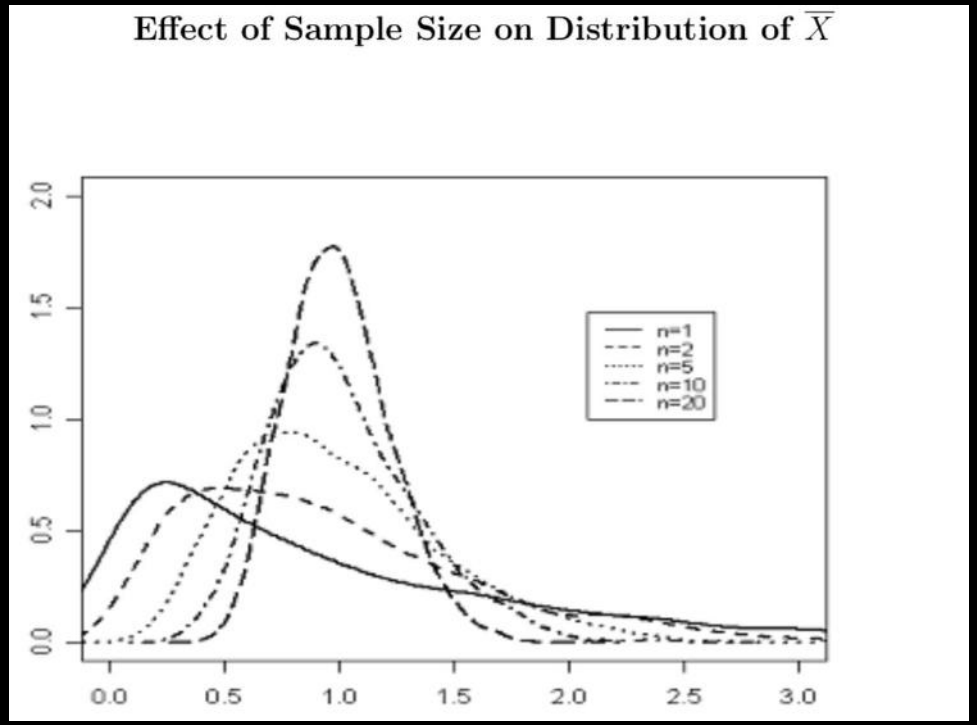
$$y = \frac{1.7078}{\sqrt{x}}$$

$$y = \frac{\sigma}{\sqrt{n}}$$

Diminishing Returns



How many samples do I need?

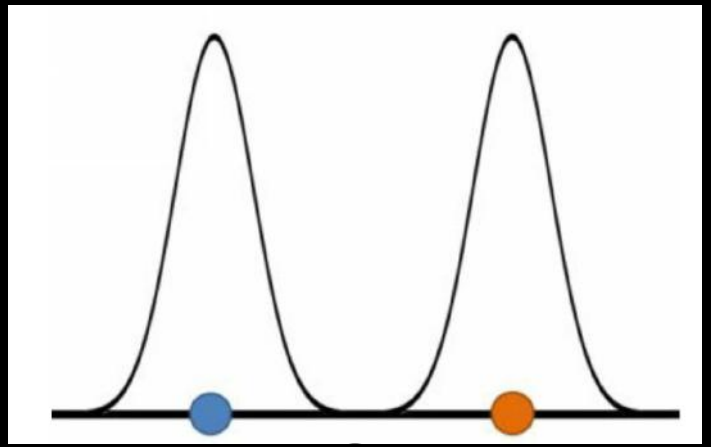


Statistical Inference

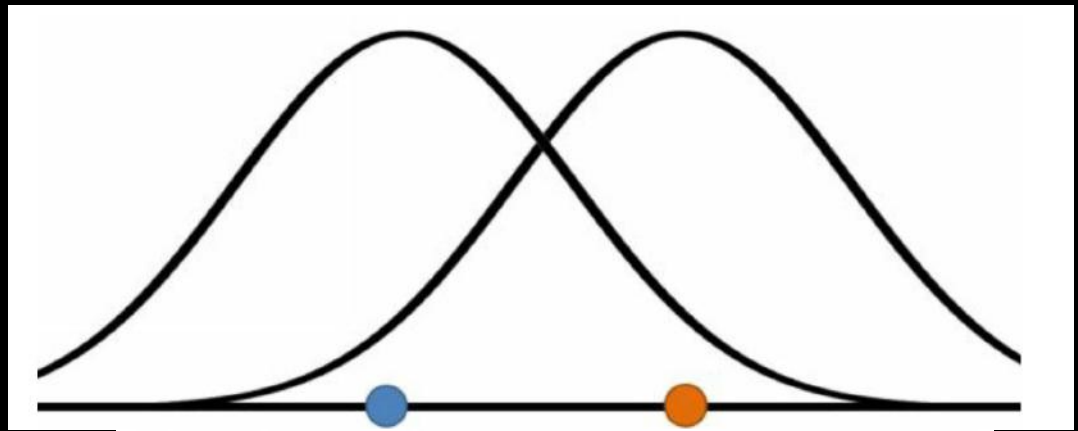
- Hypothesis Testing
  - T test, ANOVA, Wilcoxon Rank sum test etc

- Parameter Estimation
  - Simple Linear Models, Generalized Linear Models etc

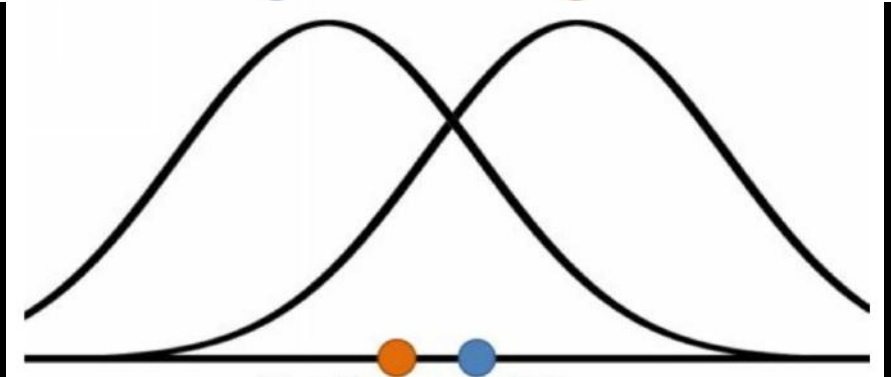
Example: Use of Normal distribution in Inference



1

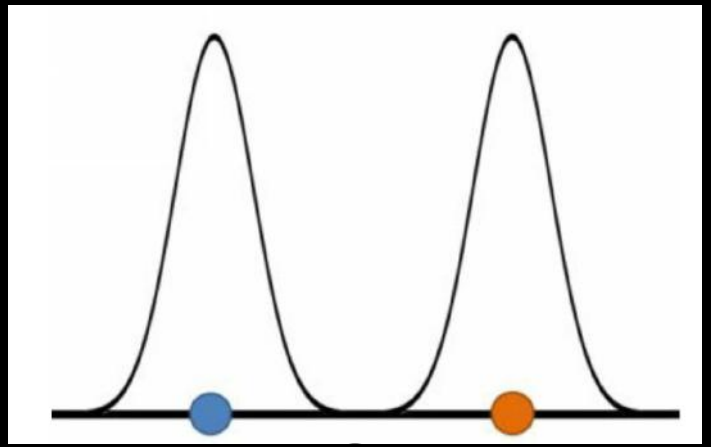


2

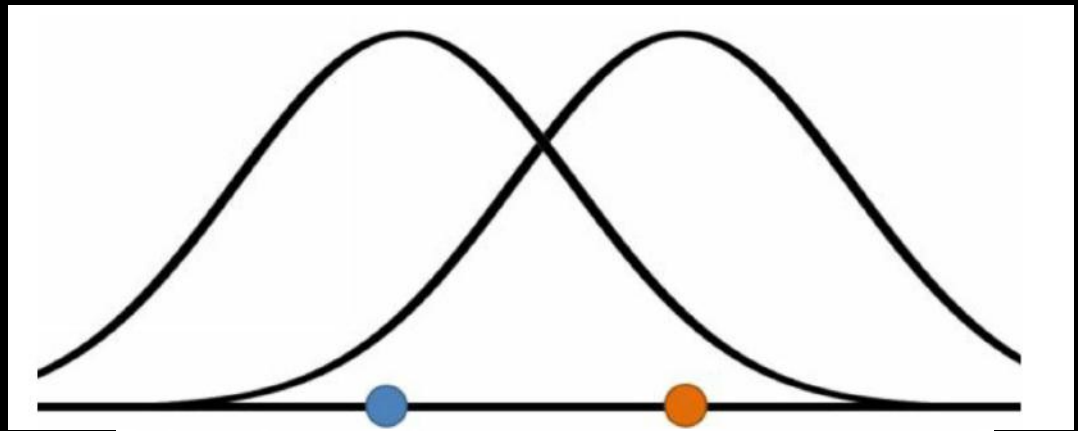


3

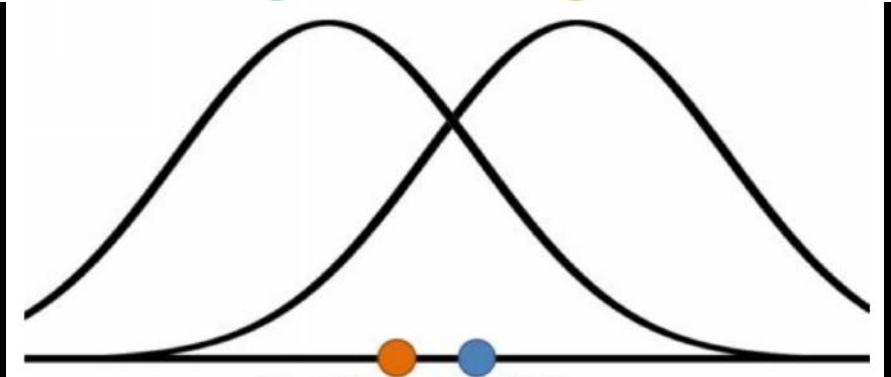
Hypothesis testing



1



2



3