Computing for Medicine

Data Science: Data Science 101

Tavpritesh Sethi

Statistical Features and Distributions

Descriptive Statistics

Common Terms

Population: Total set of observations. A population is the whole, and it comprises every member or observation.

Sample: Portion of a population. A sample can be extracted from the population using **random sampling**.

Observations: Something you measure or count during a study or experiment. Often called “samples”

Variables: Feature being measured. Numeric, Categorical, Factor variables.

Moments of a Distribution: Define Shape. Total probability, Mean, Variance, Skewness, Kurtosis

Random Variable

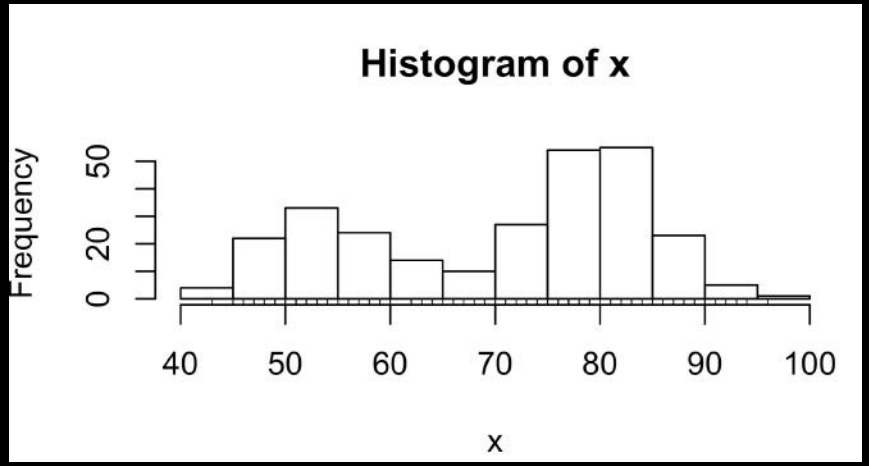
Tossing an unbiased coin twice. The sample space is

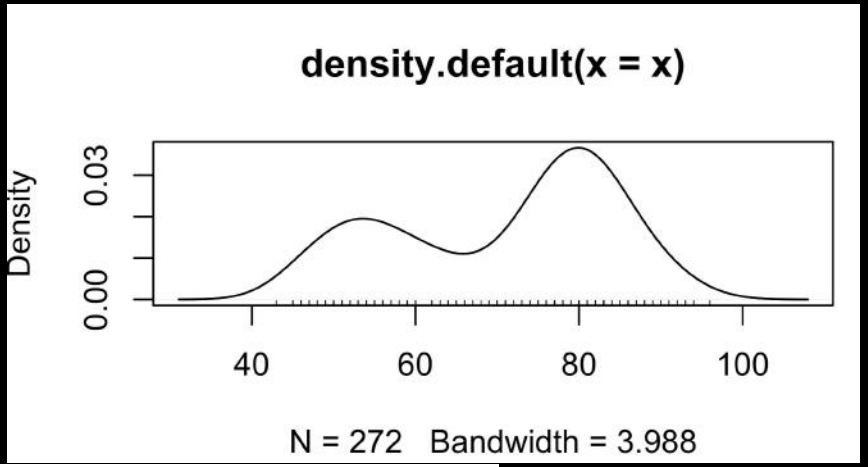
$$S = \{(H,T),(T,H),(H,H),(T,T)\}.$$

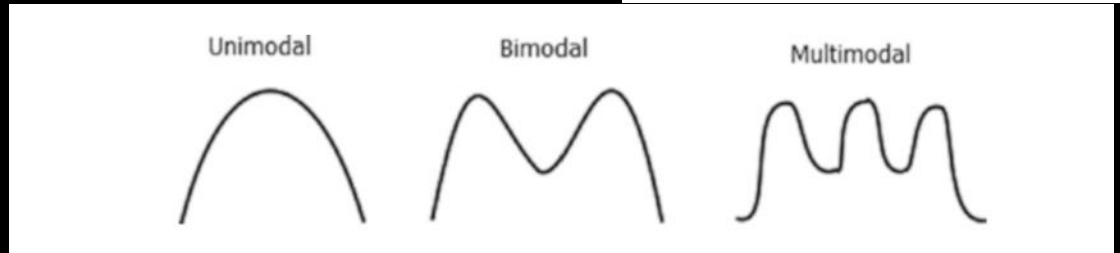
$$f(x) = \begin{cases} 0.25, & \text{for } x = 0 \\ 0.50, & \text{for } x = 1 \\ 0.25, & \text{for } x = 2 \\ 0, & \text{elsewhere.} \end{cases}$$

Distributions and their Visualization

x <- faithful$waiting; hist(x); rug(x); plot(density(x));



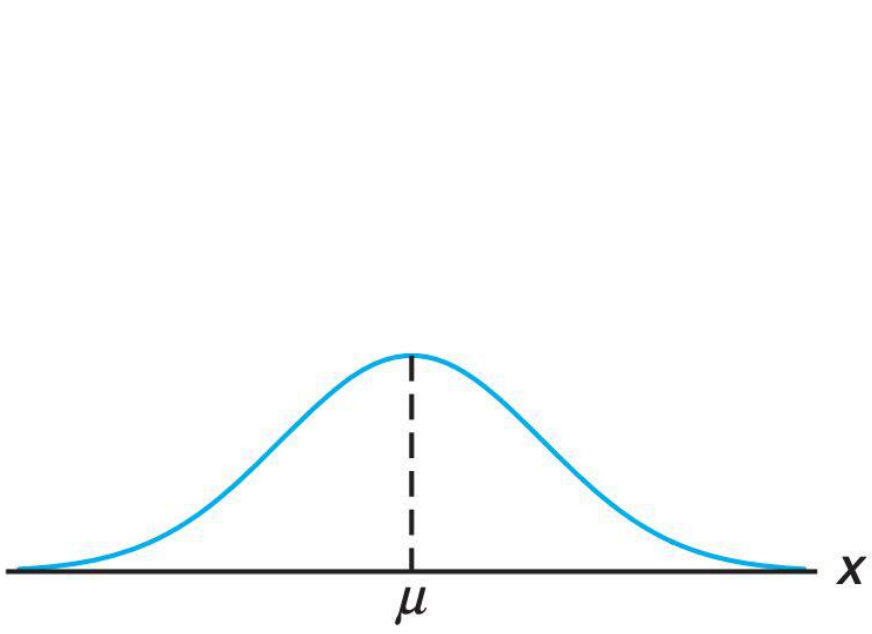




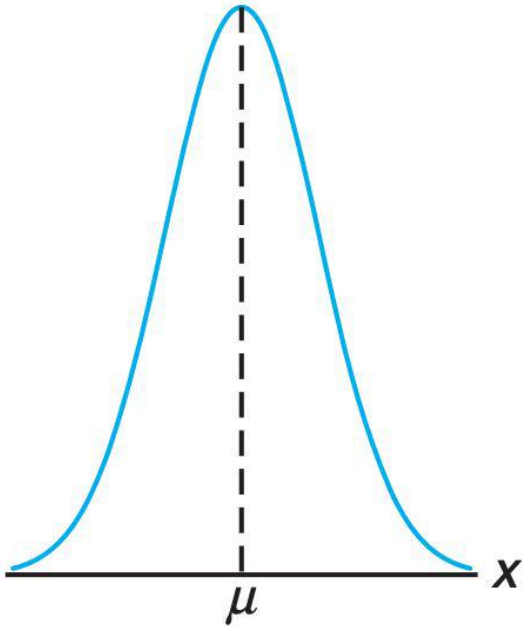
Dummy Data

Patient ID	Age (Years)	BMI (kg/m²)	HbA1c (Baseline, %)	HbA1c (6 Months, %)	Daily Insulin Dose (Units)	Outcome (Target HbA1c < 7%: 1=Yes, 0=No)
1	45	28.5	9.2	7.8	30	1
2	32	24.8	8.5	7.4	25	1
3	56	31.2	10.1	8.1	40	0
...	...	...	...	...	...	...
20	41	26.9	9.8	7.6	32	1

Spread in Distributions

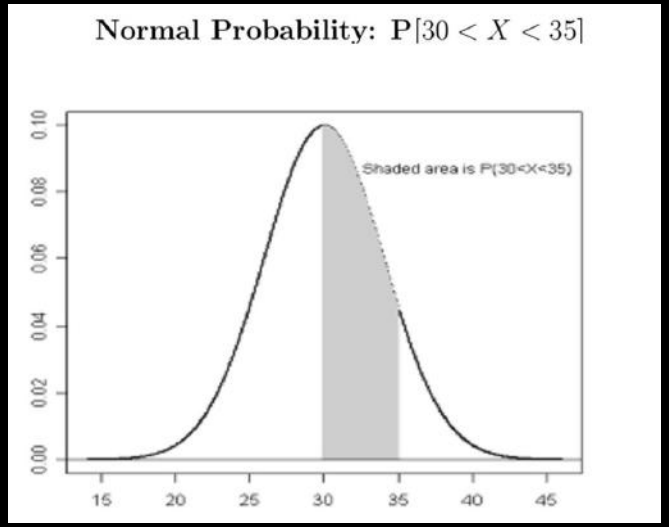


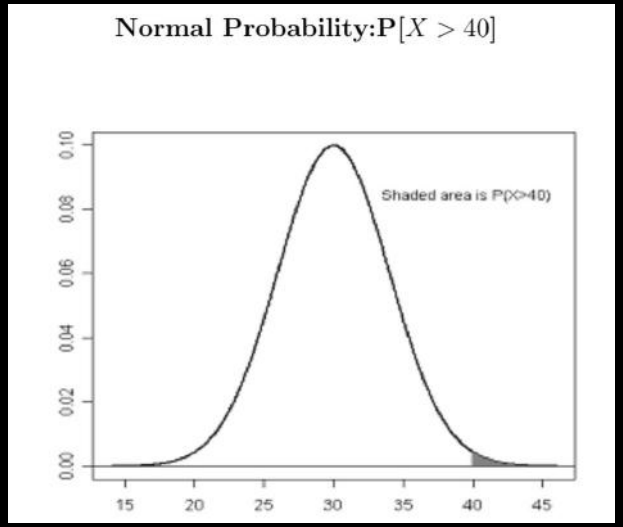
(a)

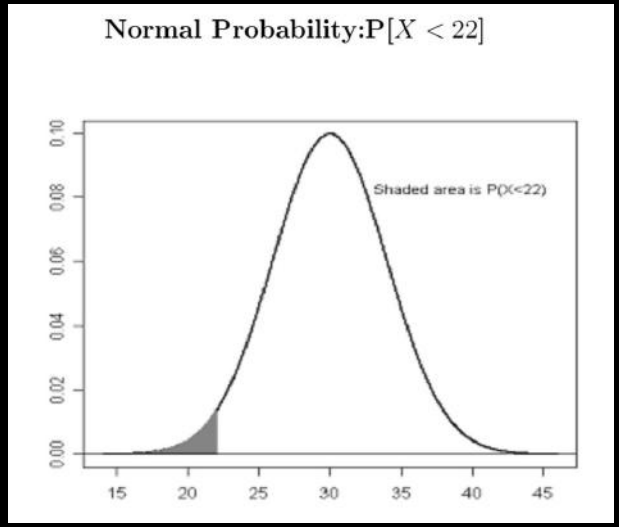


(b)

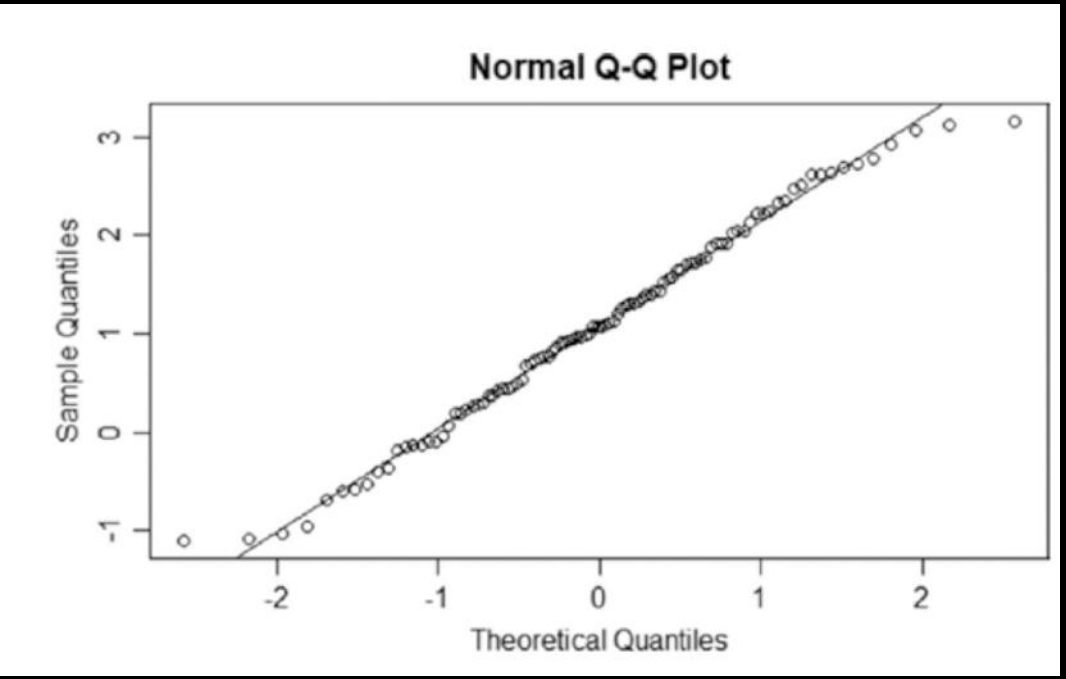
Quantiles (Rank Statistics)







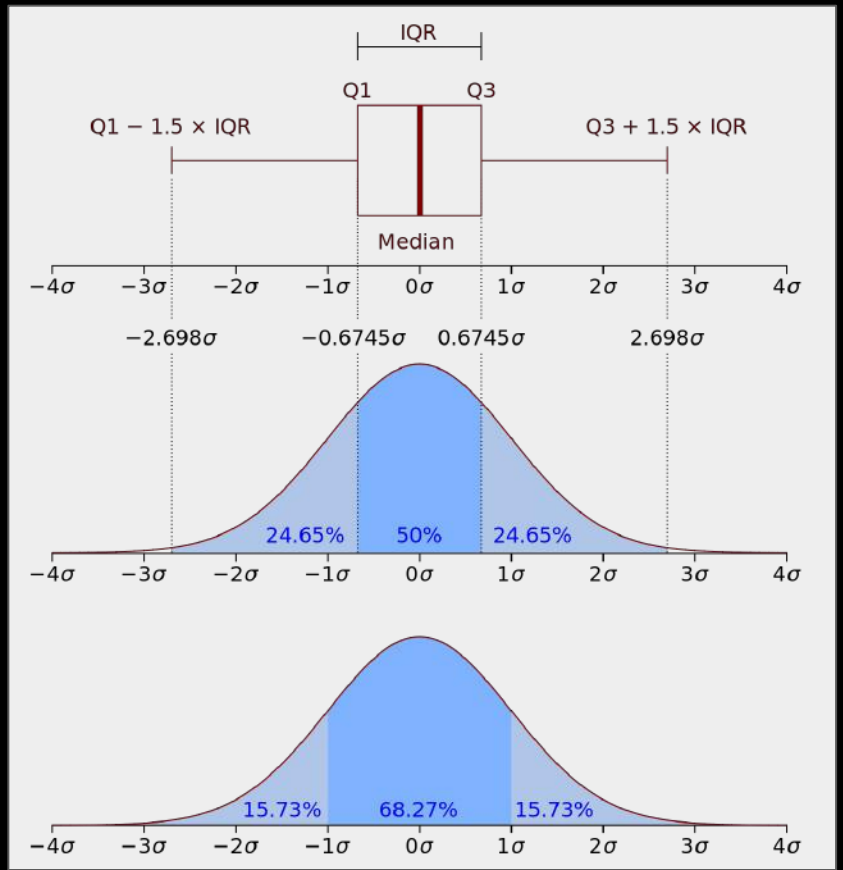
Quantile Plots



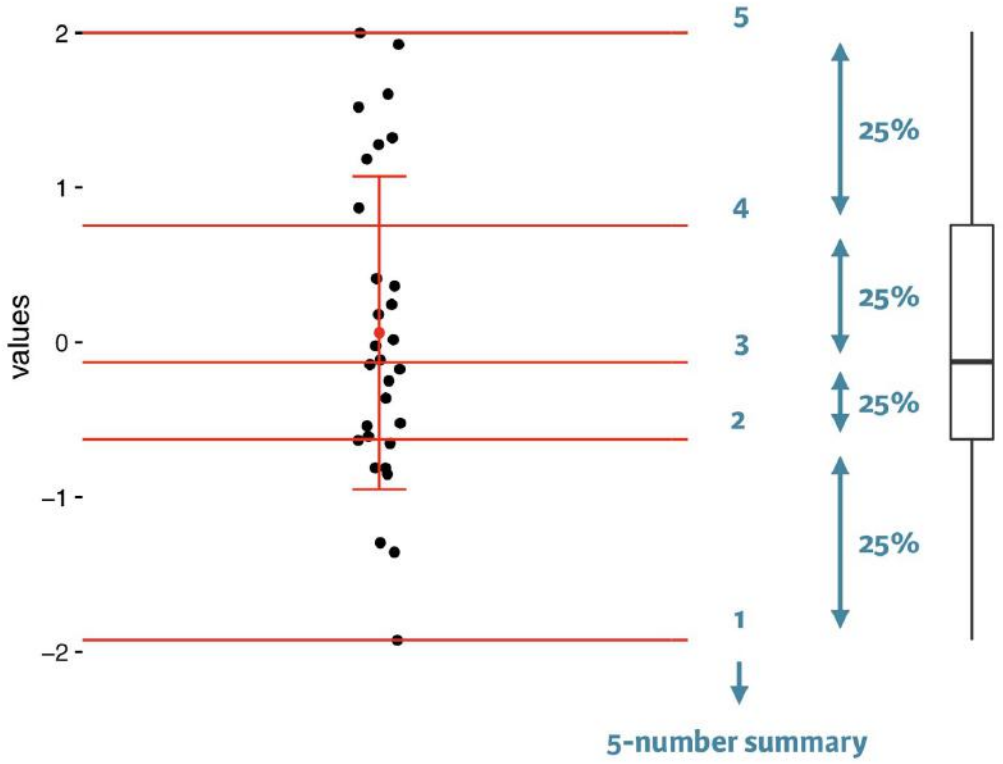

```
xuni = runif(100); qqnorm(xuni)
xnorm = rnorm(100); qqnorm(rnorm)
shapiro.test(xuni)
shapiro.test(xnorm)
```

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

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

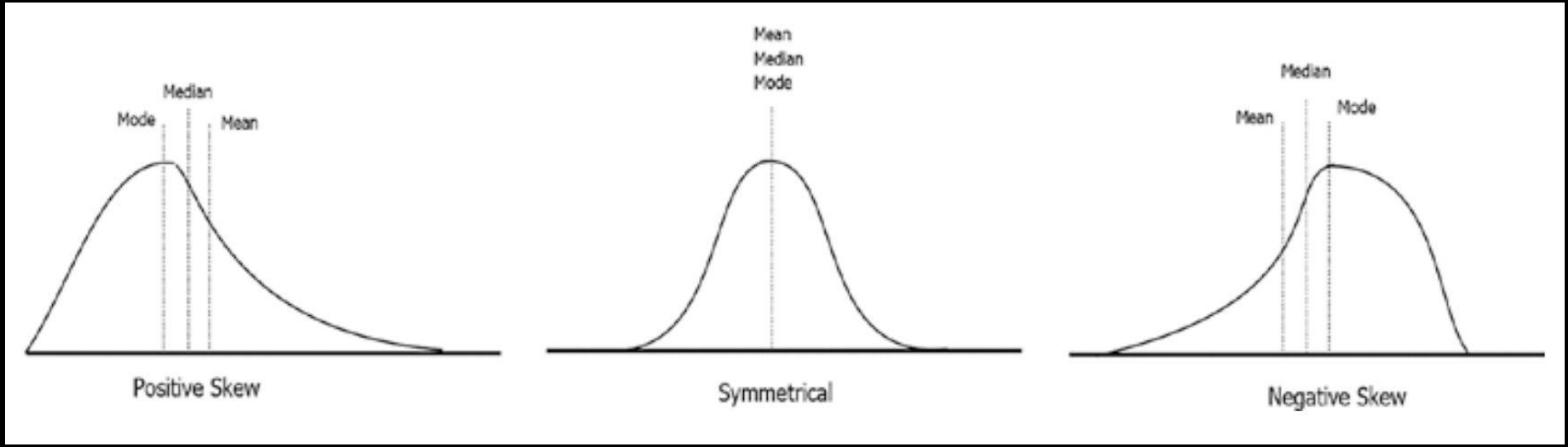
$$S^2 = \frac{\sum(X-\bar{X})^2}{n-1}$$

Population SD and Sample SD

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

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

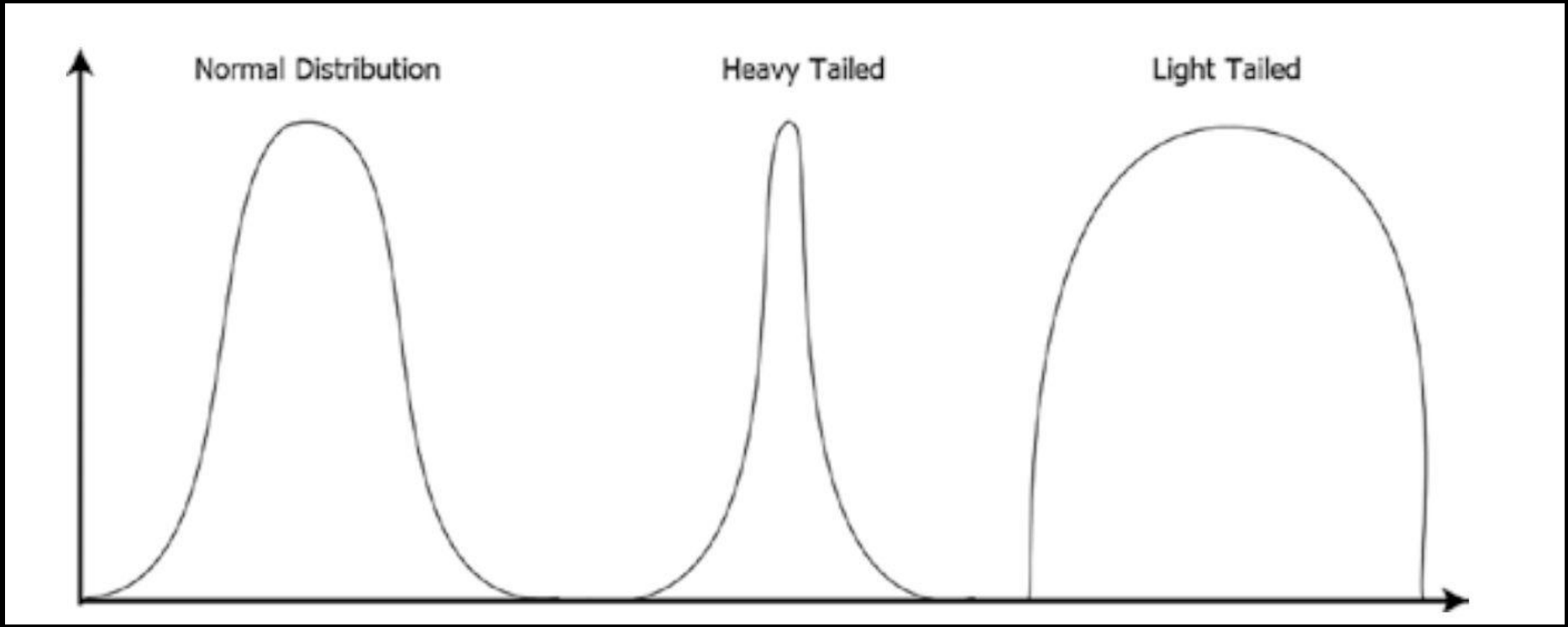
Skewness


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

Skewness

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

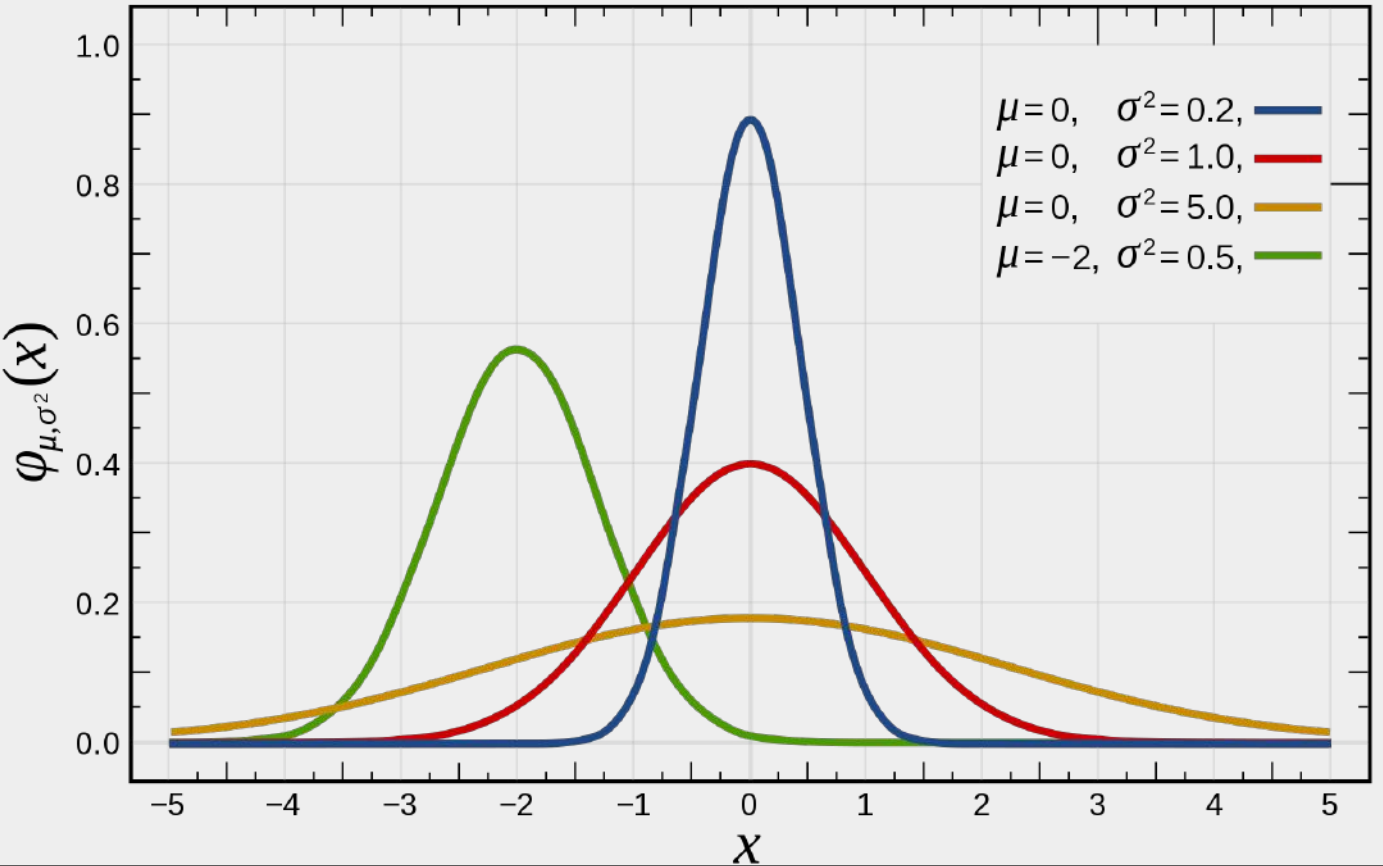
Kurtosis



Kurtosis

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

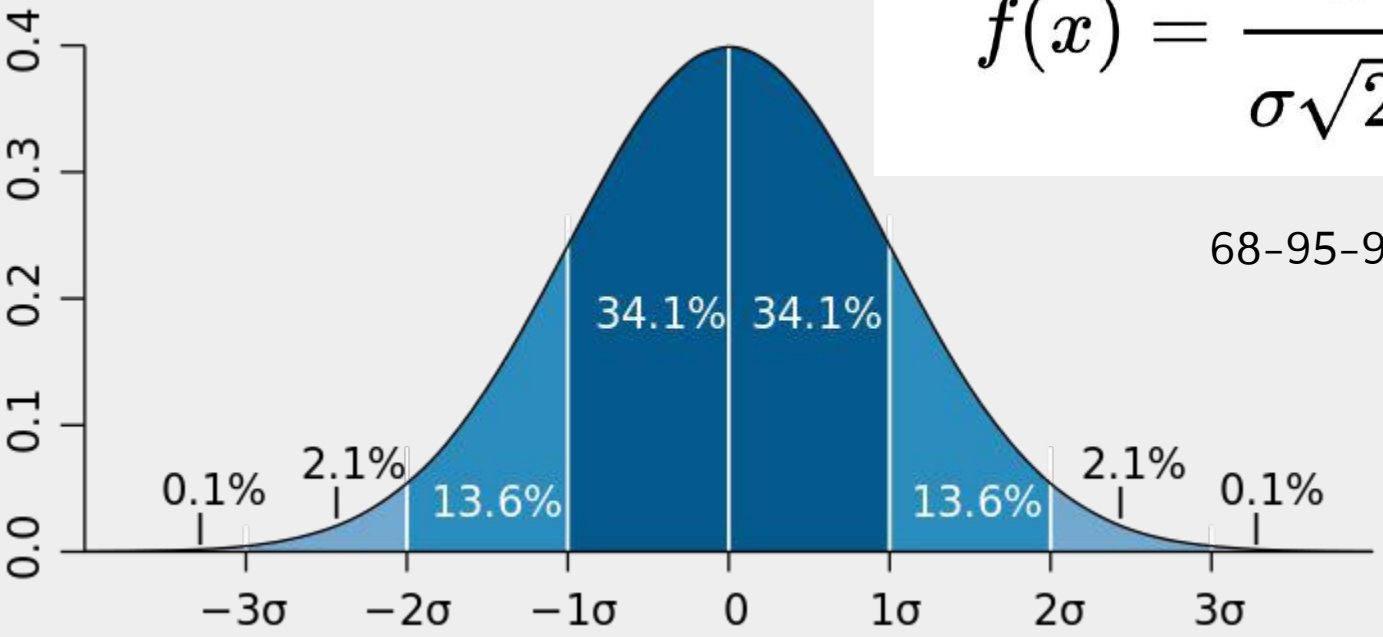
Normal Distribution



What characterizes “Normal”?

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

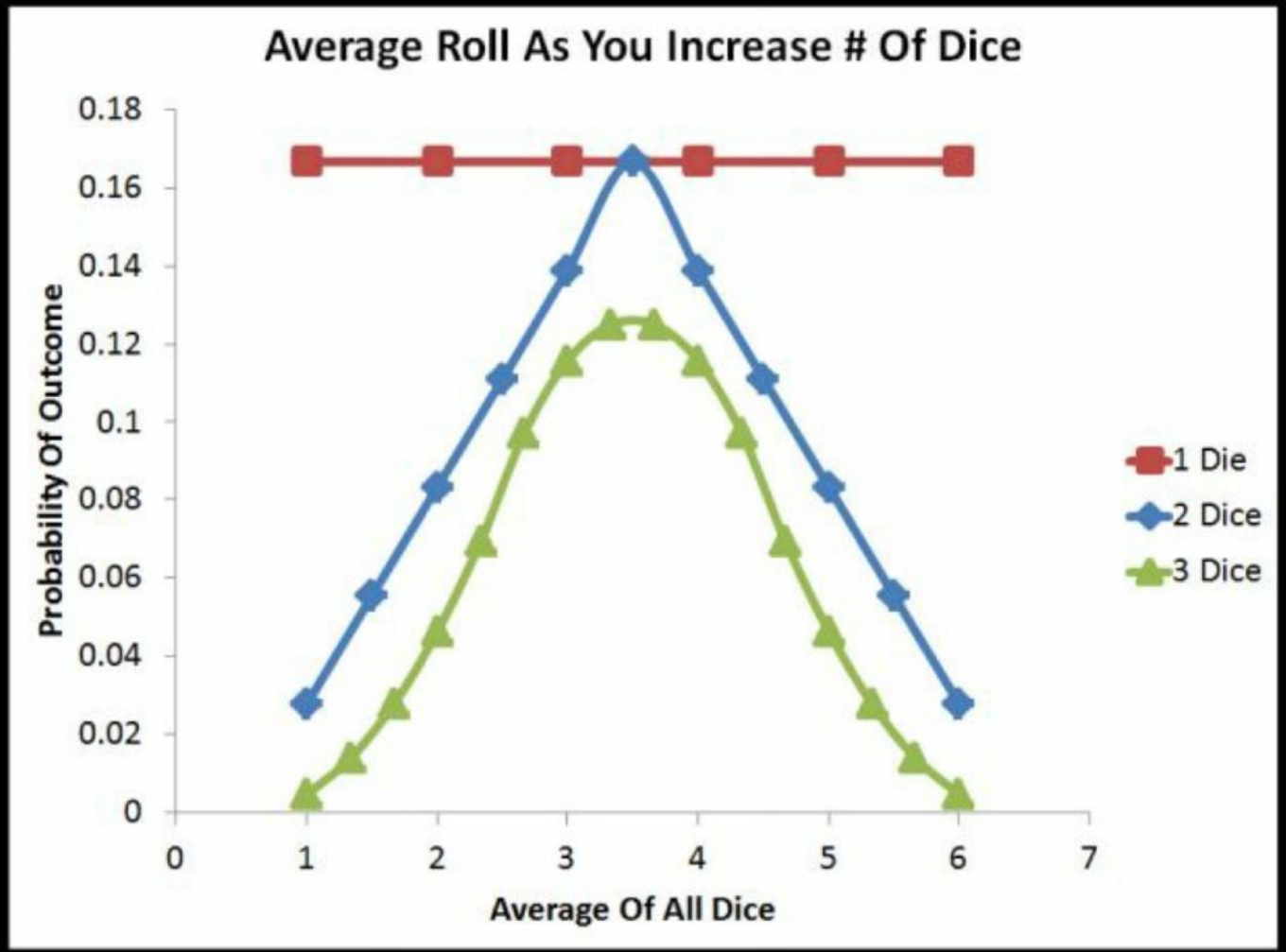
68–95–99: Three Sigma Rule

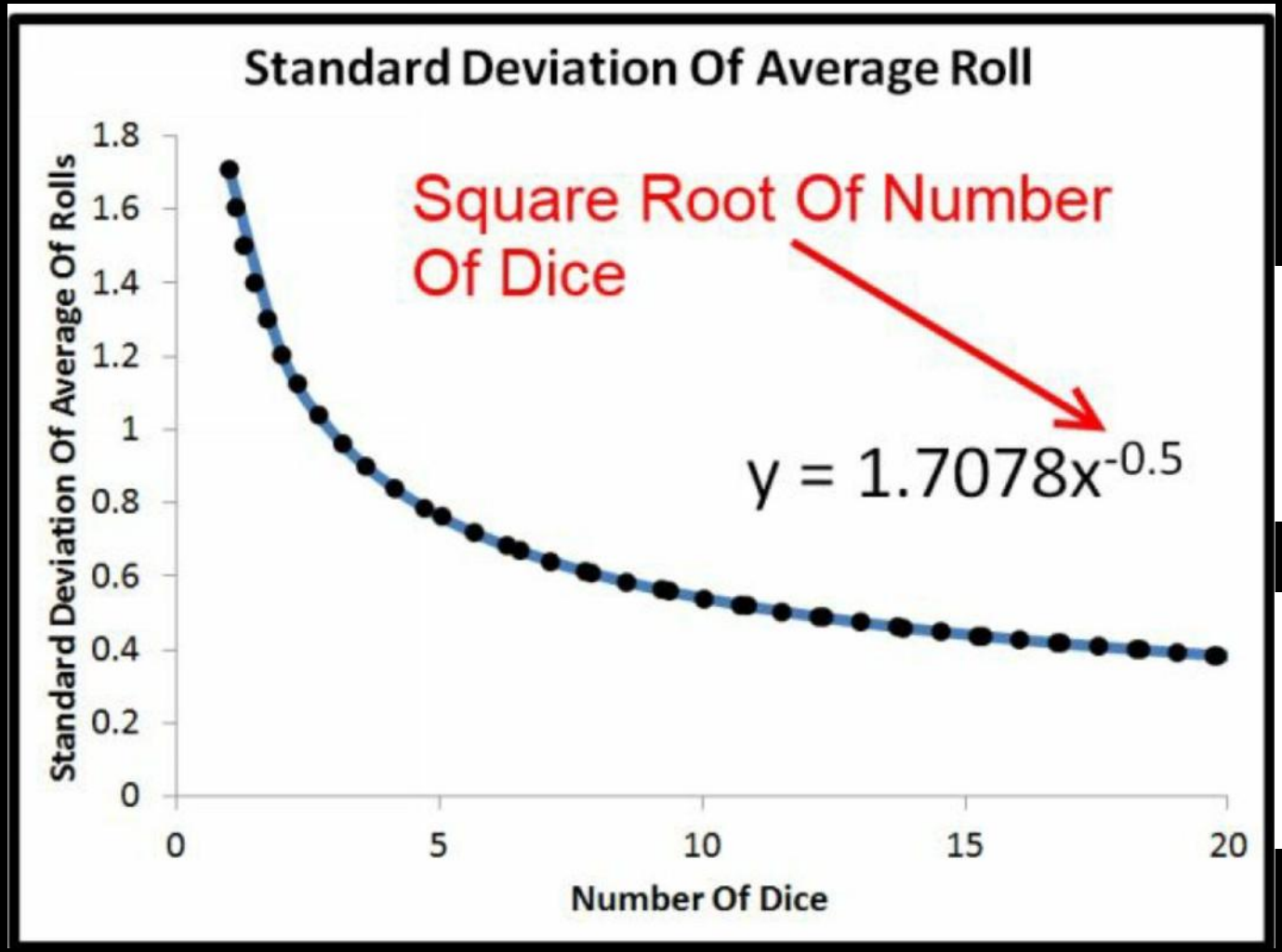


Central Limit Theorem

For ANY distribution

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

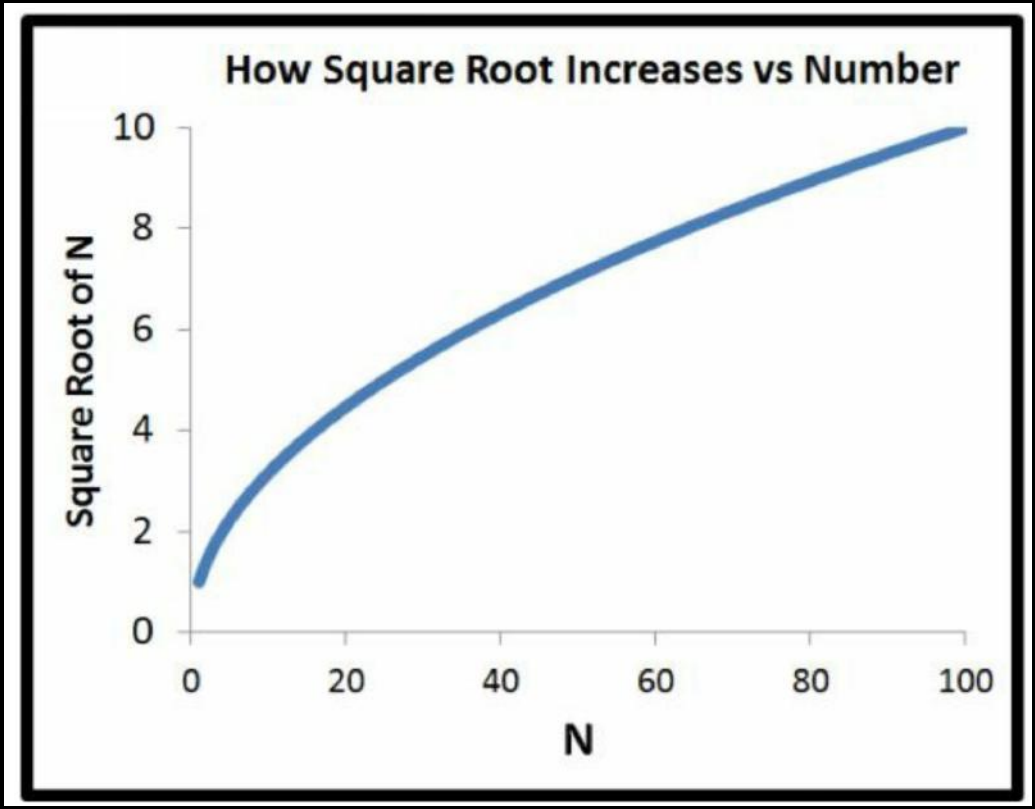

Average Roll As You Increase # Of Dice
Probability Of Outcome
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0
1
2
3
4
5
6
7
Average Of All Dice
1 Die
2 Dice
3 Dice



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

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

Diminishing Returns



How many samples do I need?