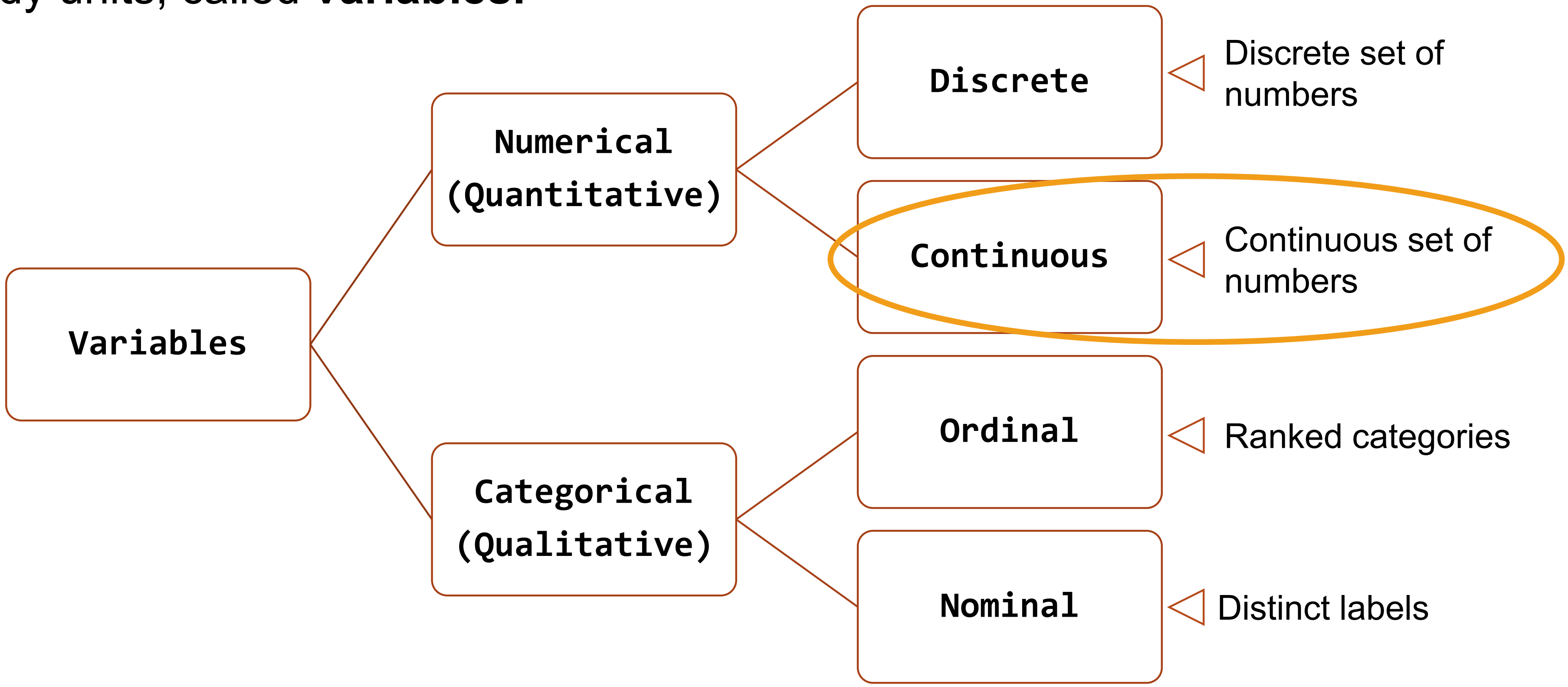
Basic Probability and Statistics: **Exploring and Summarizing Data**

Flood Frequency Analysis

Greg Karlovits, PE, PH, CFM

Hydrologic Engineering Center, May 2022

Data are the result of observing or measuring selected characteristics of the study units, called **variables.**



Variables

Numerical (Quantitative)

Categorical (Qualitative)

Discrete

Continuous

Ordinal

Nominal

Discrete set of numbers

Continuous set of numbers

Ranked categories

Distinct labels

See Tamhane and Dunlop (2000), chapter 4

USGS Flow Measurements

Measuring Agency	USGS USACE Other	**Nominal**
Measure Rating	Excellent Good Fair Poor Unknown Unspecified	**Ordinal**
Measure Duration	[<blank>, 0.0, 0.1, 0.2, …] hours	**Discrete**
Streamflow	in $ft^3 s^{-1}$	**Continuous**

Numerical Variables

Interval vs. Ratio

Comparable by difference, but not ratio

Example: Temperature

80°F is **not** 4 times hotter than 20°F.

Comparable by both, has "natural zero"

STRONGER SCALE

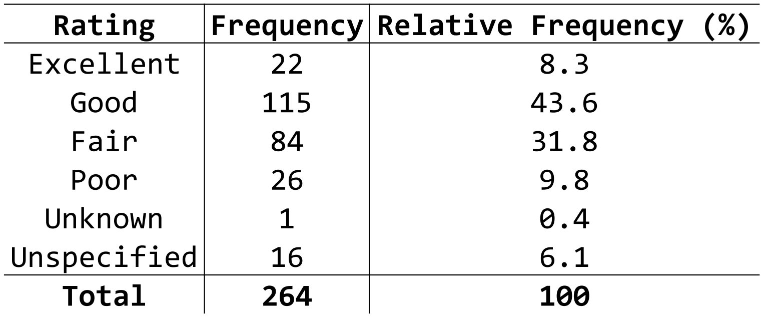
Example: Distance

50 km **is** 10 times farther than 5 km.

Categorical Data Summaries

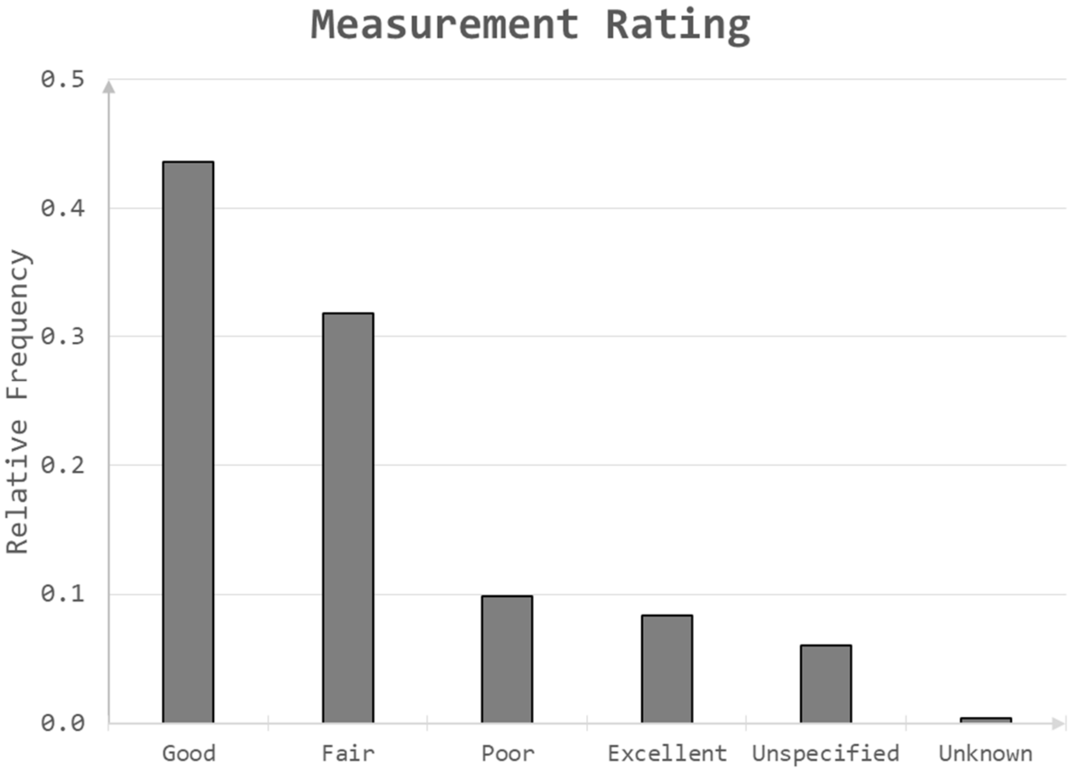
Arithmetical operations are not meaningful for categorical data.

Summary statistic:
Count



Rating	Frequency	Relative Frequency (%)
Excellent	22	8.3
Good	115	43.6
Fair	84	31.8
Poor	26	9.8
Unknown	1	0.4
Unspecified	16	6.1
Total	**264**	**100**

Frequency Table



Pareto Chart

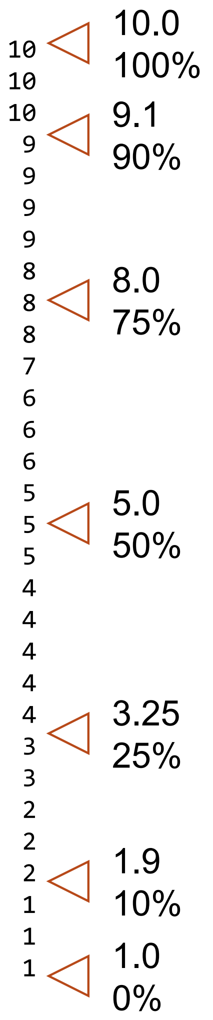
Numerical Data Summaries: Percentiles

The *α*-percentile of a dataset is the data value where ***α*% of the data are below it.**

Values shown at right have been interpolated.

Excel:
`=PERCENTILE.INC(x, k)`

[R]:
`quantile(x, probs)`



10 ◁ 10.0 100%
10
10 ◁ 9.1 90%
9
9
9
9
8
8 ◁ 8.0 75%
8
7
6
6
6
5
5 ◁ 5.0 50%
5
4
4
4
4
4 ◁ 3.25 25%
3
3
2
2
2 ◁ 1.9 10%
1
1
1 ◁ 1.0 0%

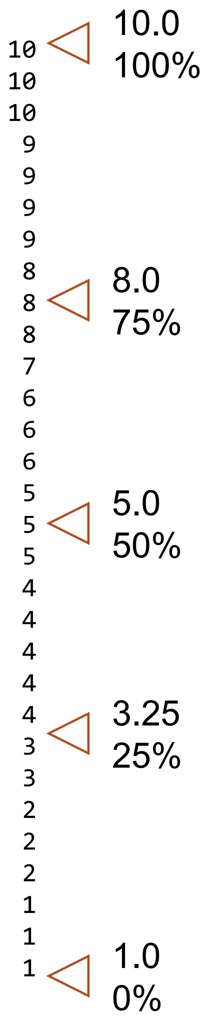
Numerical Data Summaries: Five-Number Summary

A quick, standard way to represent a dataset.

Other measures can be derived from it.

- Minimum
- 25th percentile (first quartile)
- 50th percentile (median/second quartile)
- 75th percentile (third quartile)
- Maximum

```
[R]:
fivenum(x)
```



```
10
10
10
 9
 9
 9
 9
 8
 8
 8
 7
 6
 6
 6
 5
 5
 5
 4
 4
 4
 4
 4
 3
 3
 2
 2
 2
 1
 1
 1
```

10.0
100%

8.0
75%

5.0
50%

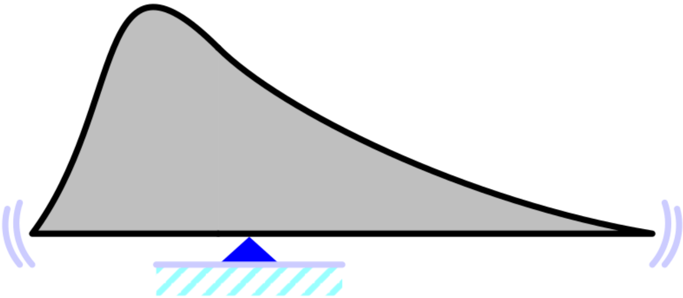
3.25
25%

1.0
0%

Numerical Data Summaries: Central Tendency

Mean

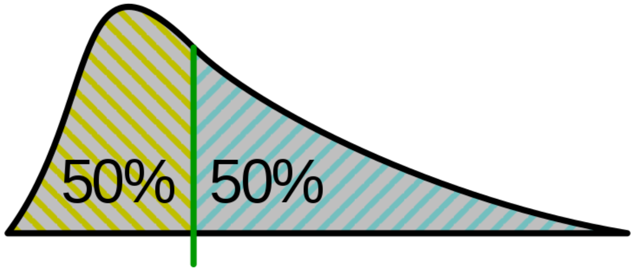
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$



Median

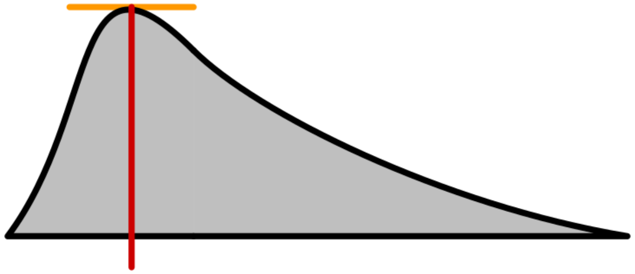
$$x_{min} = x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)} = x_{max}$$

$$\tilde{x} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & n\ odd \\ \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n+1}{2}\right)}}{2} & n\ even \end{cases}$$



Mode

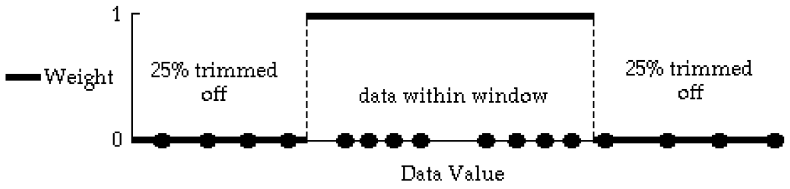
Most frequently-occurring value



Numerical Data Summaries: Central Tendency (Robust)

Weighted averaging schemes

Trimmed Mean



Weighted average of many values

[R]:

```
mean(x, trim = 0.25)
```

Tukey's Trimean

$$TM = \frac{Q_1 + 2Q_2 + Q_3}{4}$$

Q_1 – first quartile (25th percentile)
Q_2 – median (50th percentile)
Q_3 – third quartile (75th percentile)

Weighted average of 3 values

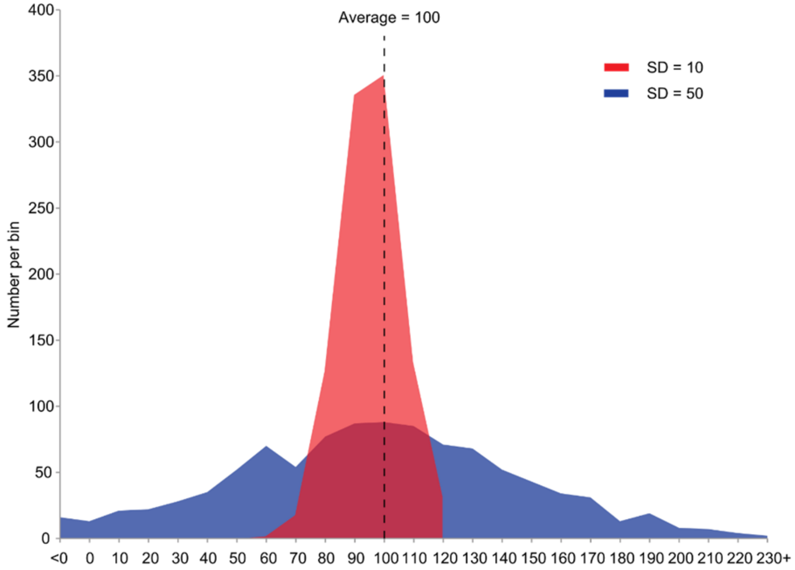
Numerical Data Summaries: Dispersion

Variance

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Standard Deviation

$$s_x = \sqrt{s_x^2}$$



Coefficient of Variation

$$CV = \frac{s_x}{\bar{x}}$$

Numerical Data Summaries: Dispersion (Robust)

Inter-Quartile Range

$IQR = Q_3 - Q_1$

Q1 – first quartile (25th percentile)
Q3 – third quartile (75th percentile)

Quartile Coeff. of Dispersion

$CQV = \frac{Q_3 - Q_1}{Q_3 + Q_1}$

Scale-invariant

Median Absolute Deviation

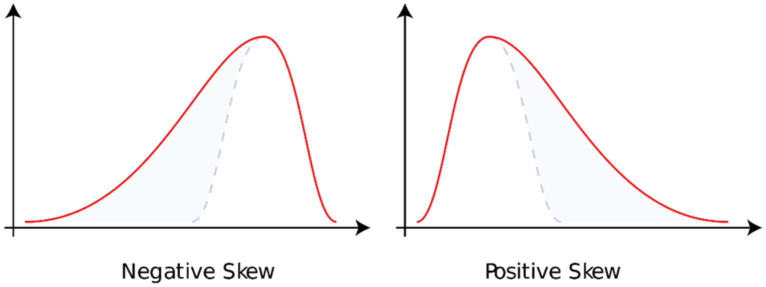
$MAD = \mathrm{median}(|x_i - \tilde{x}|)$

median distance between each data point and the sample median

Numerical Data Summaries: Asymmetry (Skew)

Coeff. of skewness

$$g = \frac{n}{(n-1)(n-2)} \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{s_x^3}$$



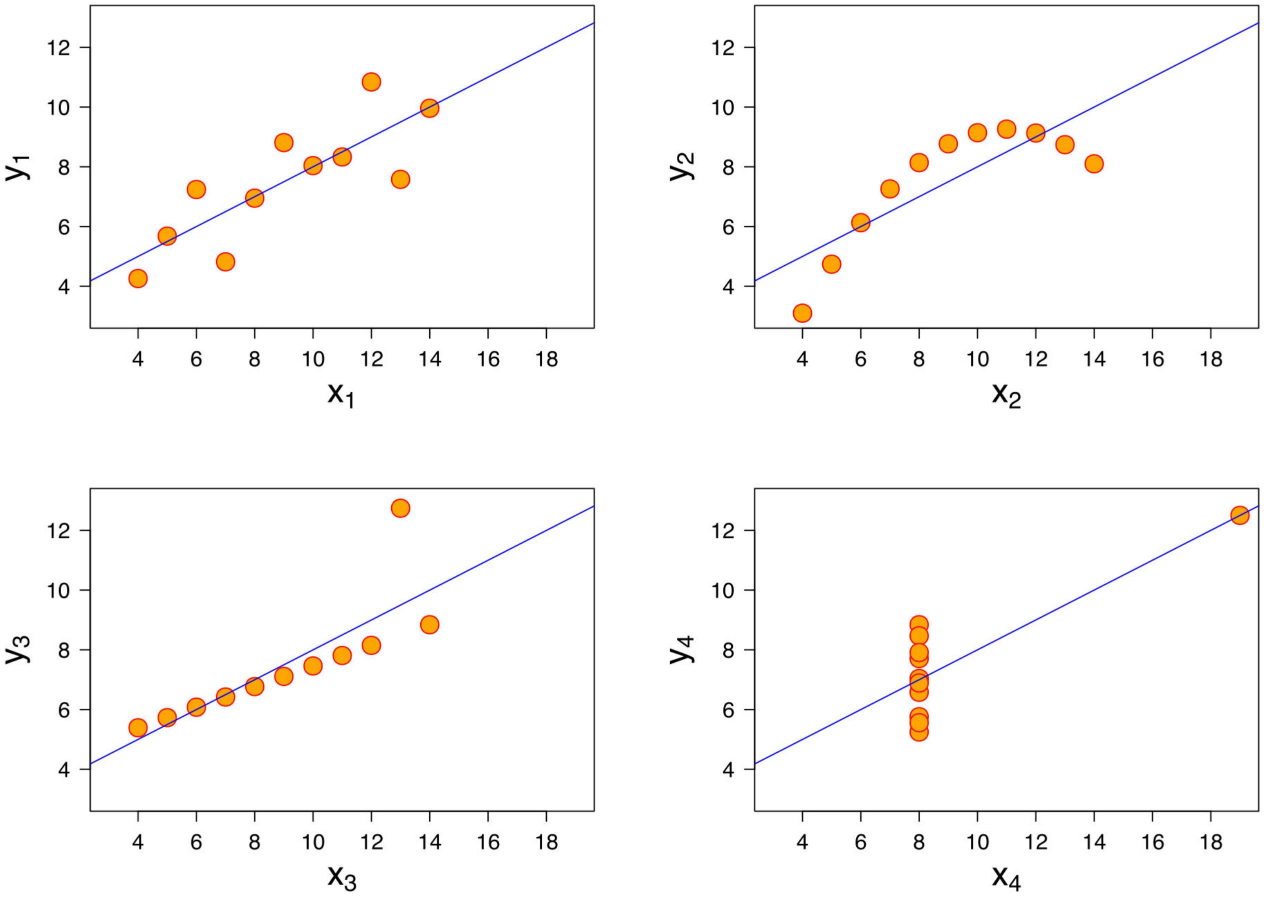
Yule's Coeff.

$$\frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1} = \frac{\frac{Q_3 + Q_1}{2} - Q_2}{\frac{Q_3 - Q_1}{2}}$$

L-Moments

- A formulation of moment measure less susceptible to outliers
- Mainly used in precipitation-frequency analysis
- Central tendency – “L-Mean”
- Dispersion – “Coefficient of L-Variation”
- Asymmetry – “Coefficient of L-Skewness”

Why should you look at your data?

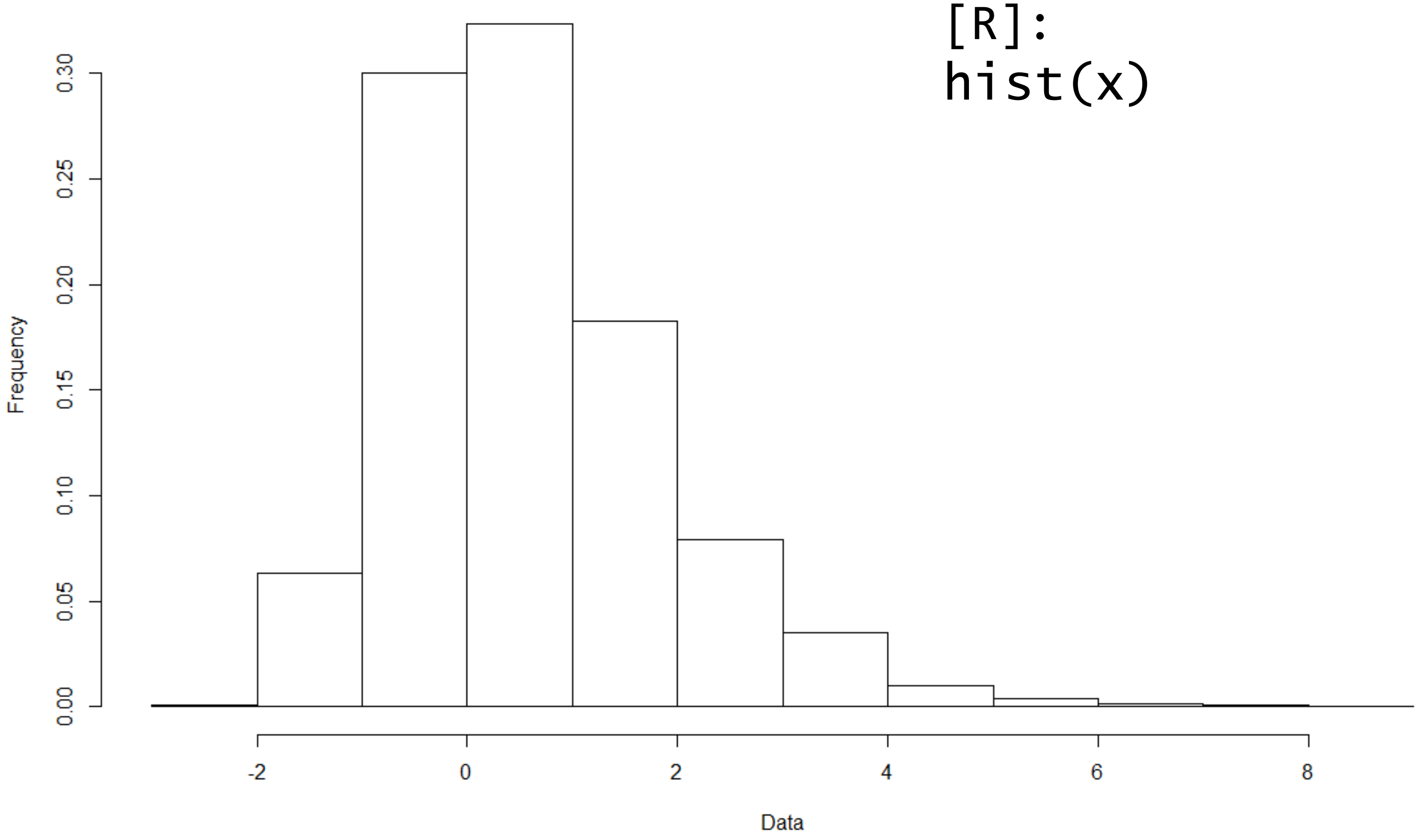


Property	***Value***
Mean of x	9
Sample variance of x	11
Mean of y	7.50
Sample variance of y	4.125
Correlation between x and y	0.816
Linear regression line	y = 3.00 + 0.500x
Coefficient of determination of the linear regression	0.67

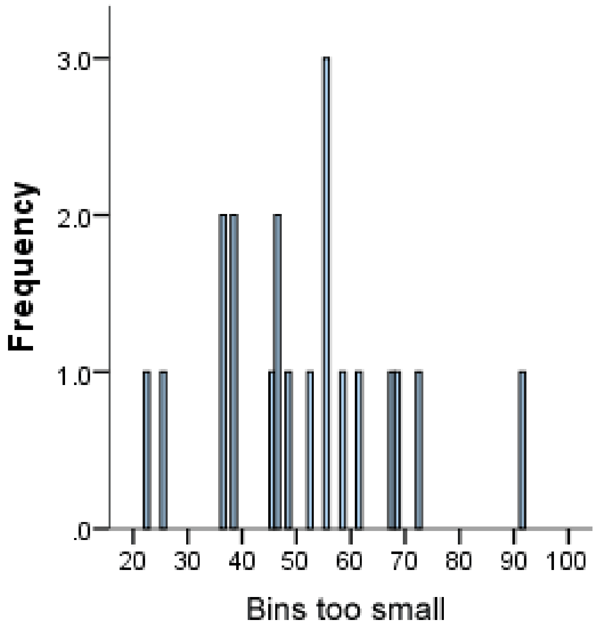
Histogram

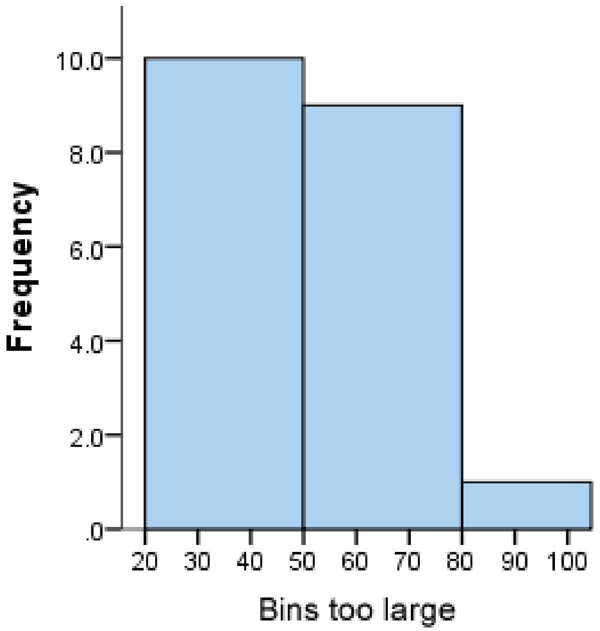
Excel:
=FREQUENCY(data, bins)

[R]:
hist(x)



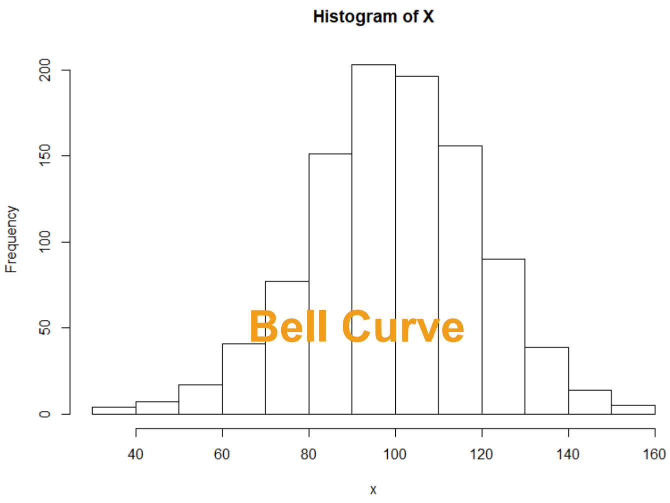
Histogram

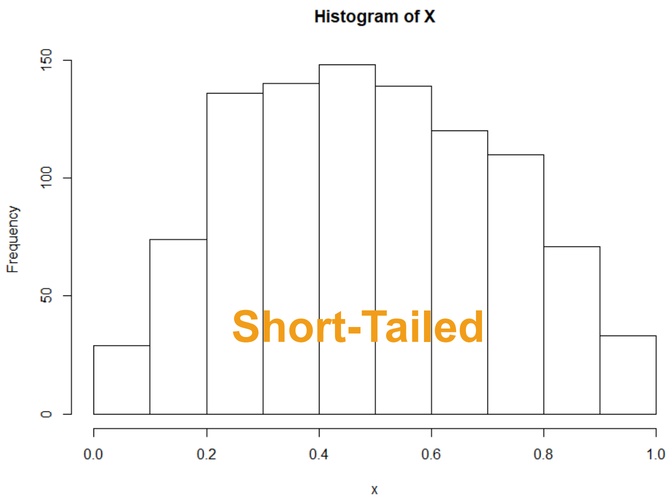


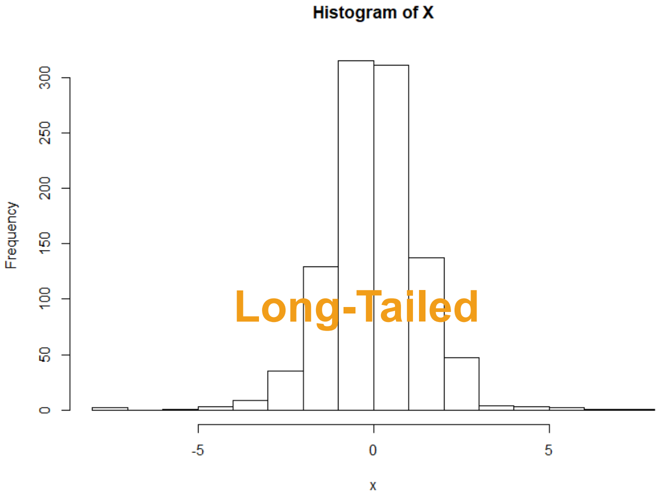


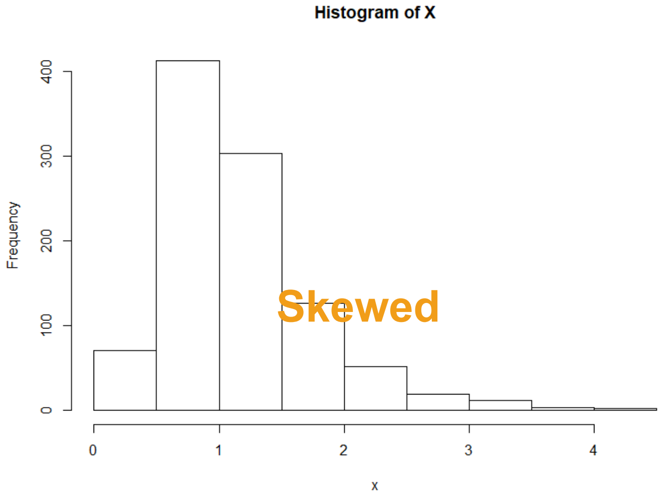
https://statistics.laerd.com/statistical-guides/understanding-histograms.php

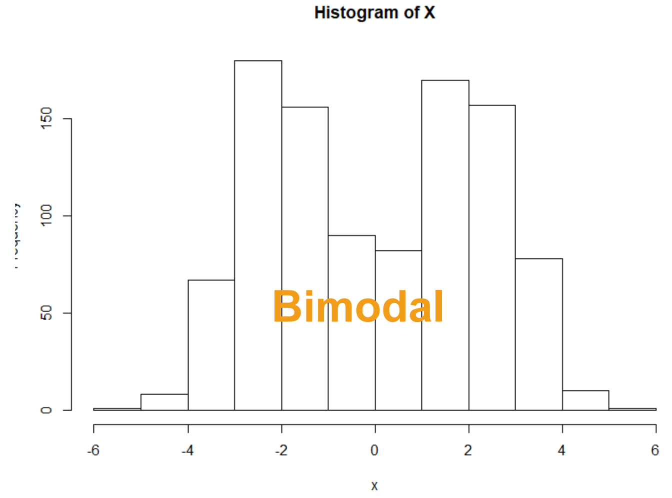
Histogram Diagnostics





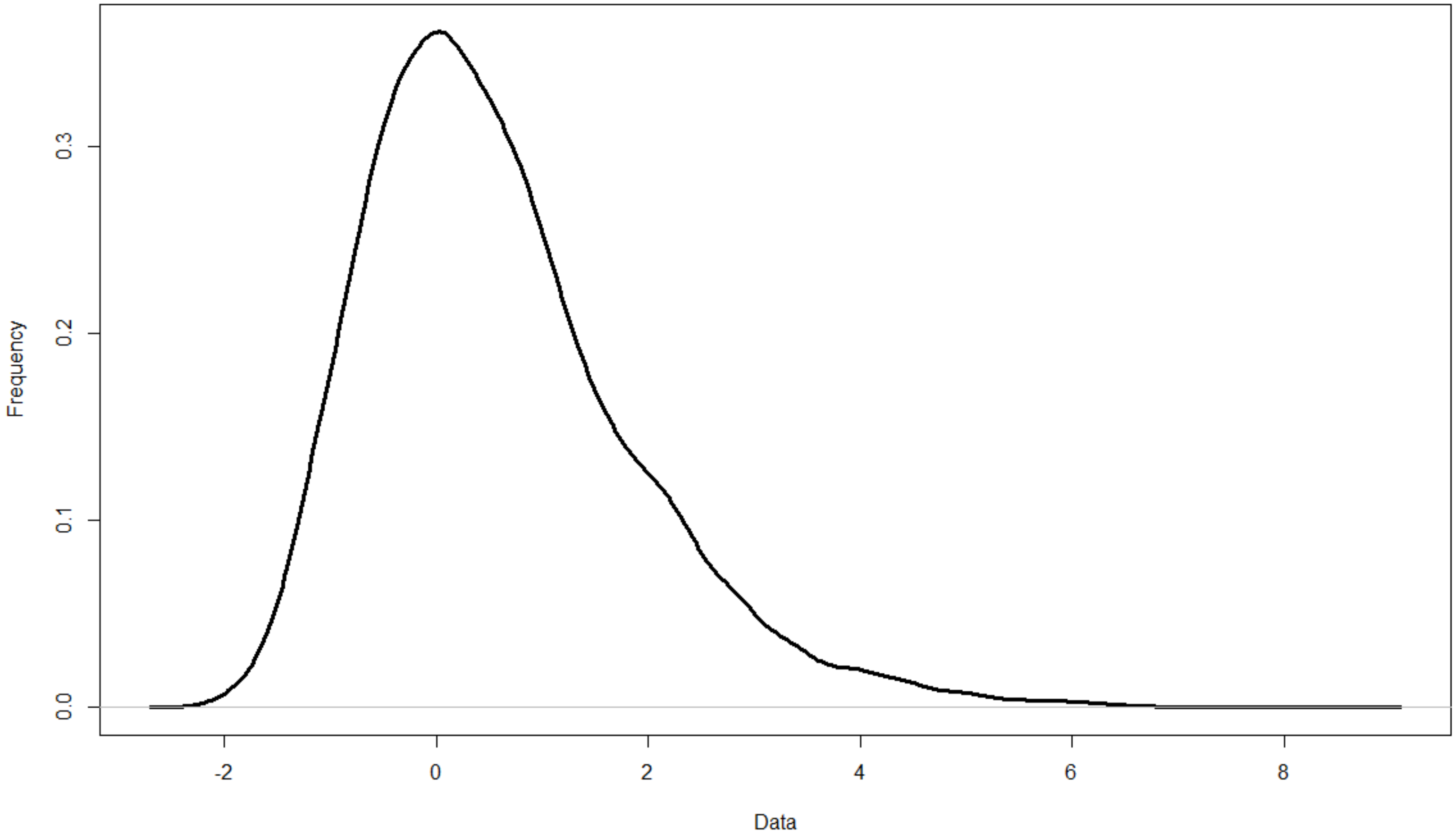




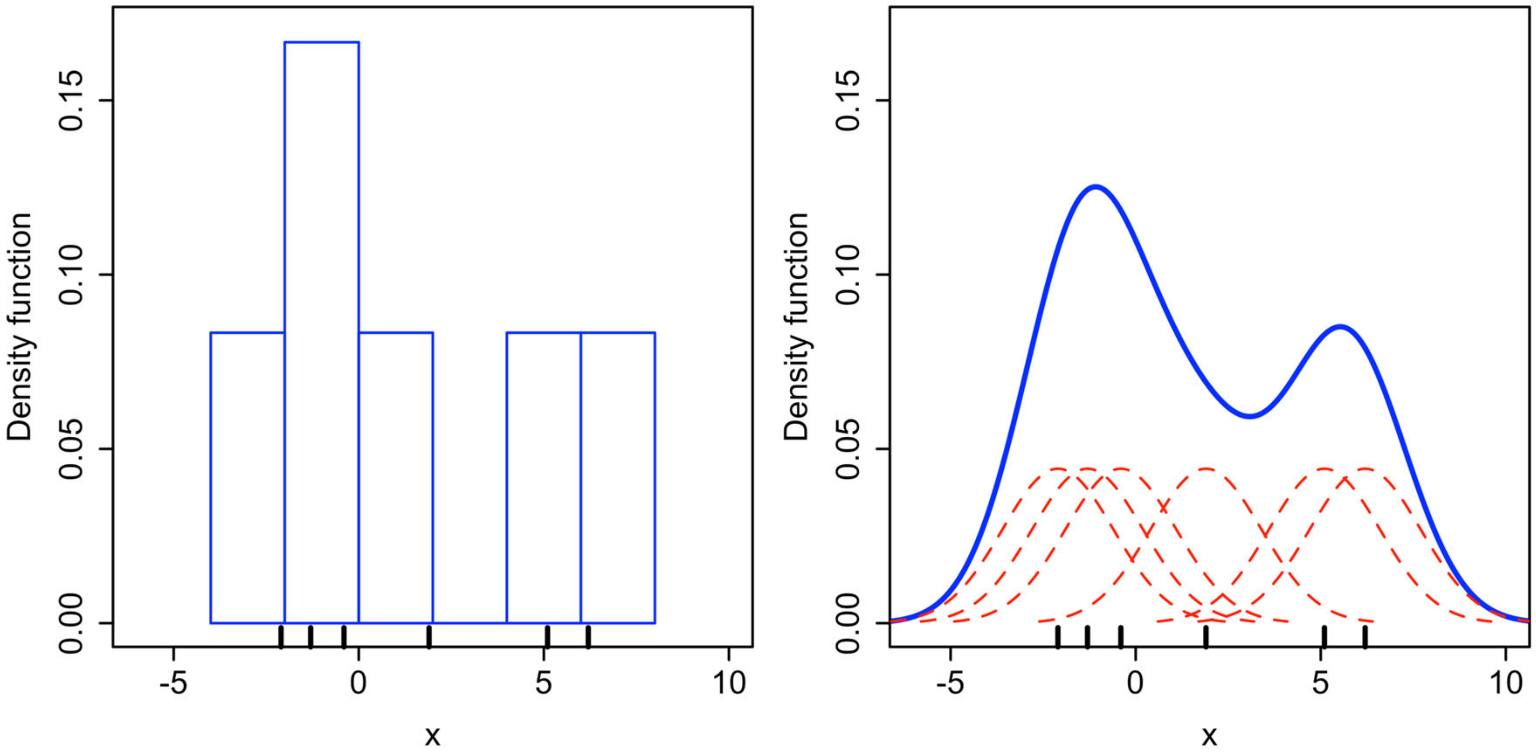


Kernel Density Estimation

[R]:
`density(x)`

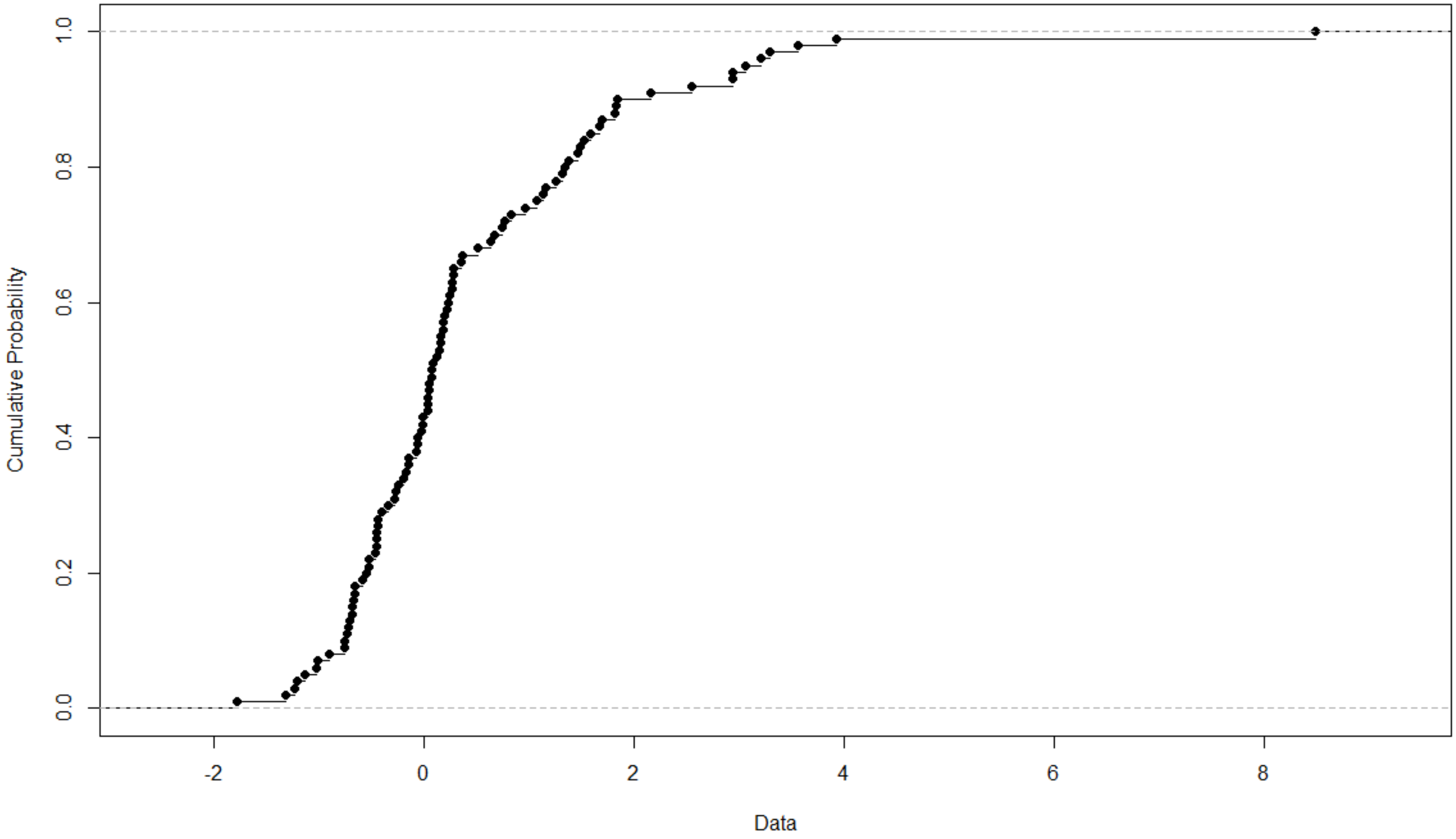


Kernel Density Estimation

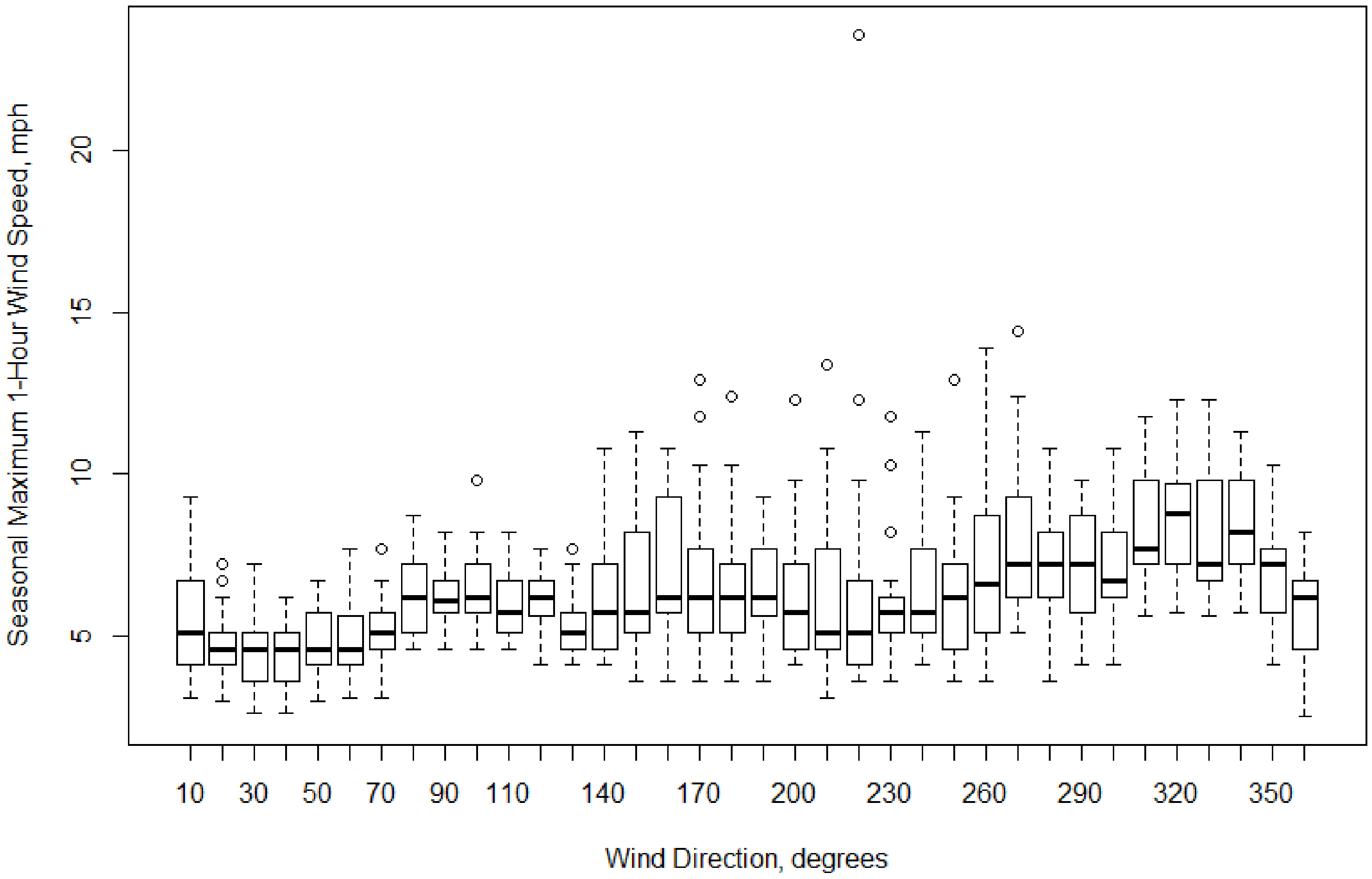


Empirical CDF (eCDF)

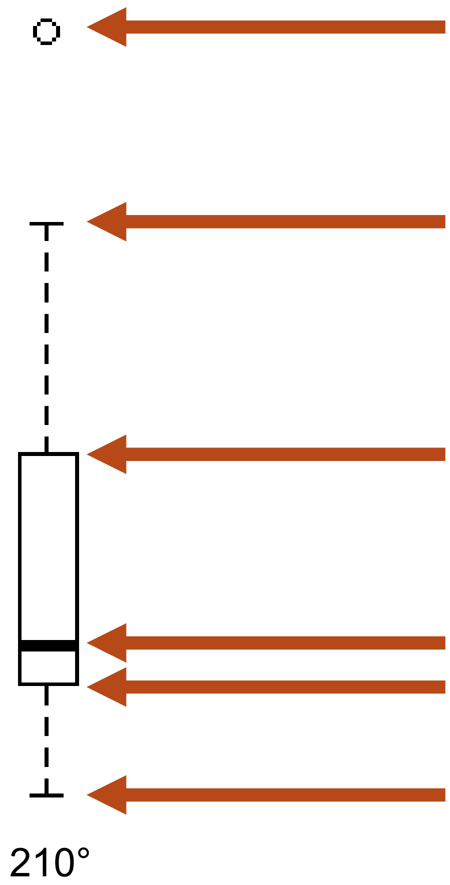
[R]:
`ecdf(x)`



Box Plots


Seasonal Maximum 1-Hour Wind Speed, mph
5
10
15
20
10
30
50
70
90
110
140
170
200
230
260
290
320
350
Wind Direction, degrees

Box Plots



Outlier

Q_3 + 1.5 * IQR

75th percentile "Q_3"

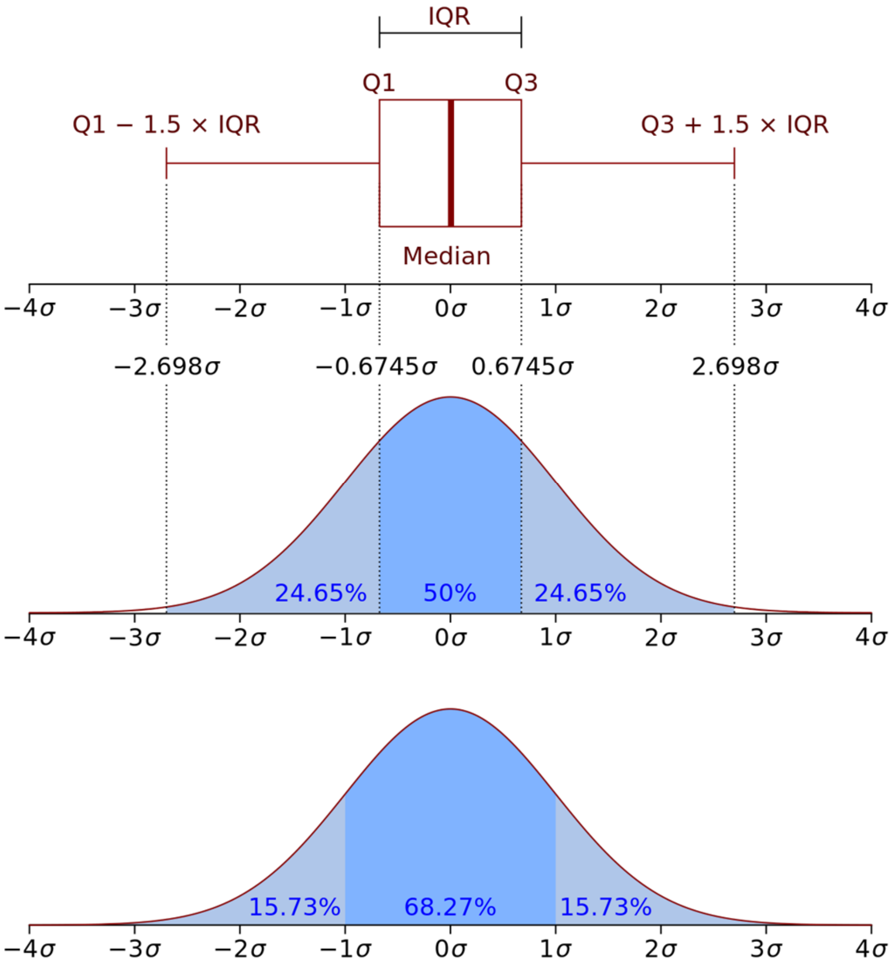
Median "Q_2"

25th percentile "Q_1"

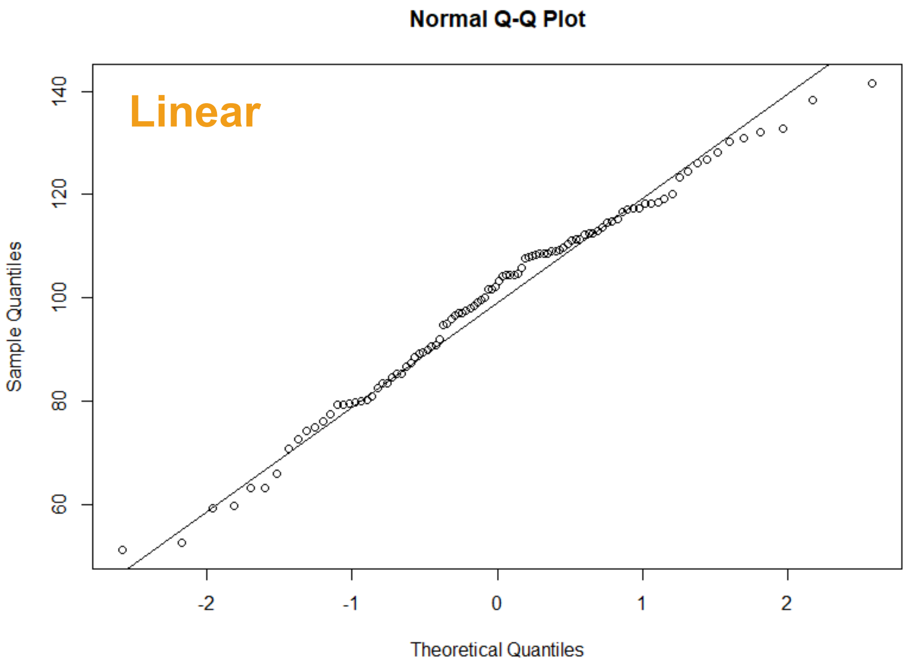
Q_1 - 1.5 * IQR

Inter-quartile range
(IQR = Q_3 – Q_1)

Box Plots


IQR
Q1
Q3
Q1 − 1.5 × IQR
Q3 + 1.5 × IQR
Median
−4σ −3σ −2σ −1σ 0σ 1σ 2σ 3σ 4σ
−2.698σ
−0.6745σ
0.6745σ
2.698σ
24.65%
50%
24.65%
−4σ −3σ −2σ −1σ 0σ 1σ 2σ 3σ 4σ
15.73%
68.27%
15.73%
−4σ −3σ −2σ −1σ 0σ 1σ 2σ 3σ 4σ

Normal Q-Q Plot



Compute z-scores for data

$$z_i = \frac{x_i - \bar{x}}{s_x}$$

Plot against sorted data

Plot line through Q_1 and Q_3

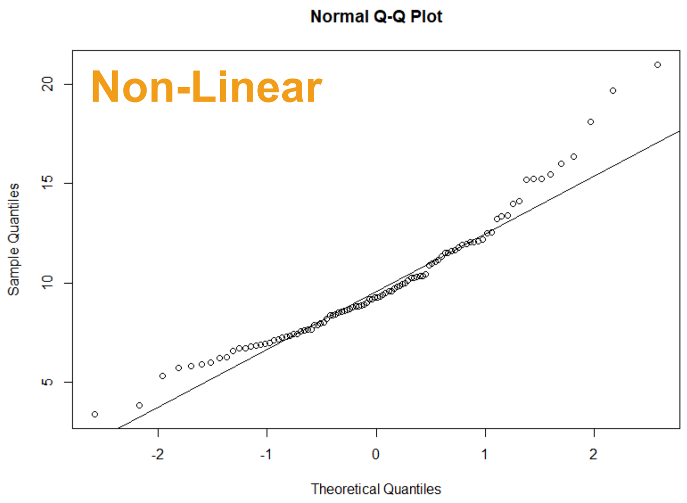
Used to test:

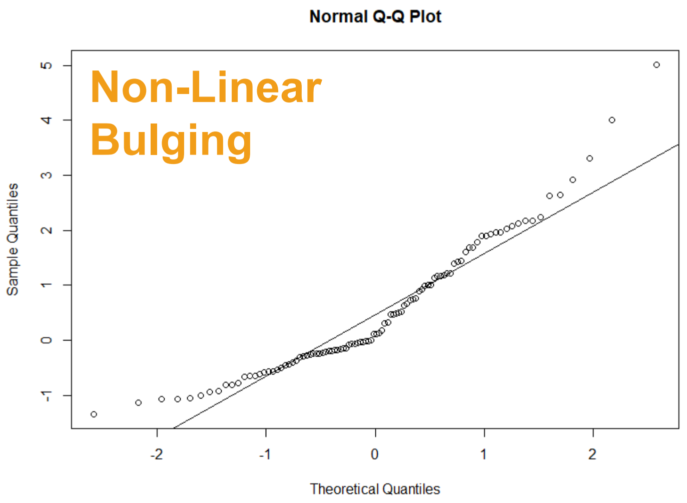
- Normality

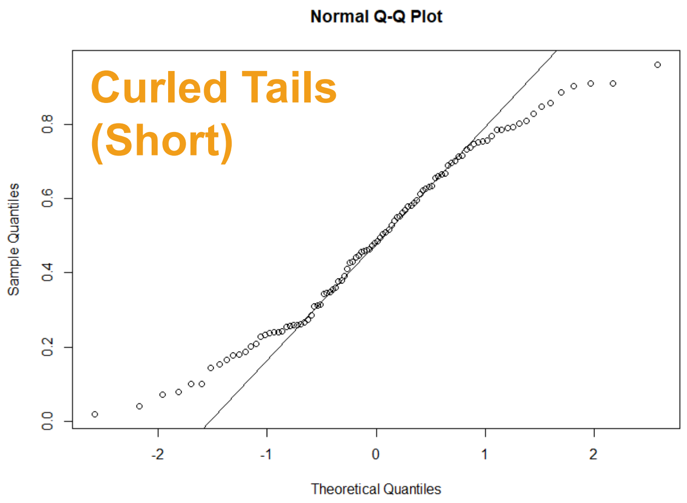
[R]:

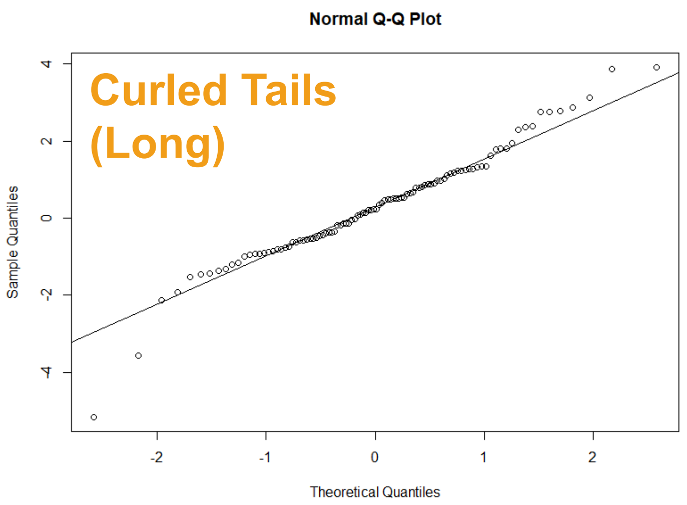
```
qqnorm(x)
qqline(x)
```

Normal Q-Q Plot Diagnostics

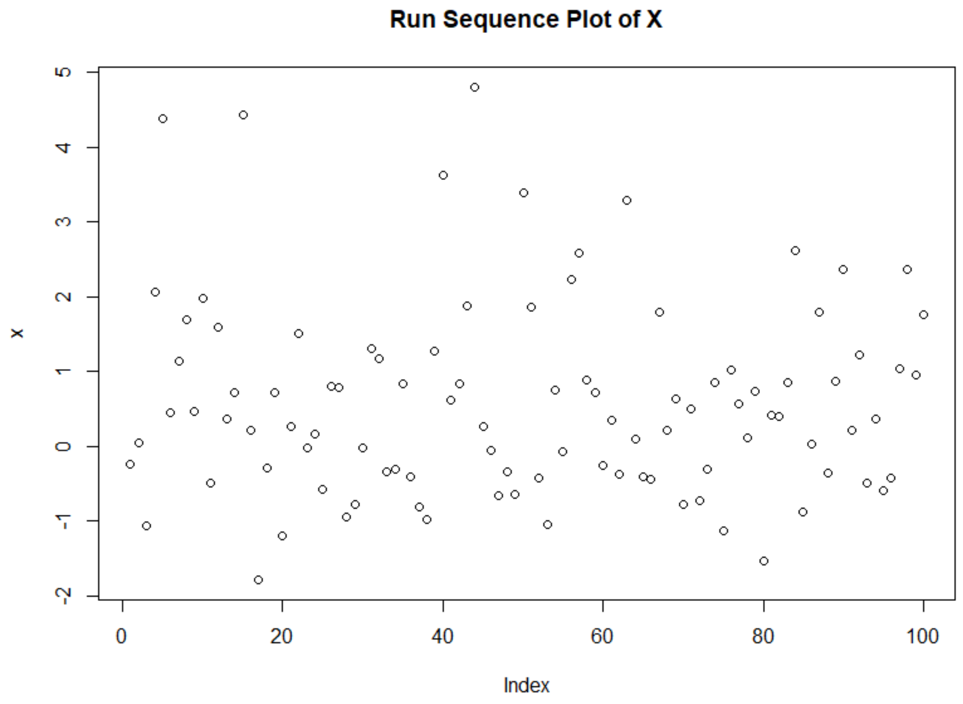

Non-Linear


Non-Linear Bulging


Curled Tails (Short)


Curled Tails (Long)

Run Sequence/Time Series Plot



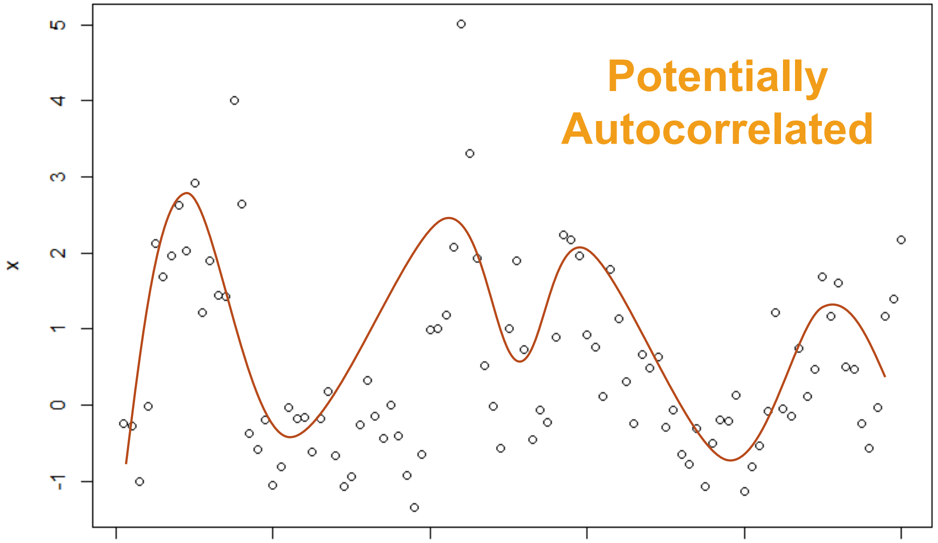
Plot the data in the order they were observed

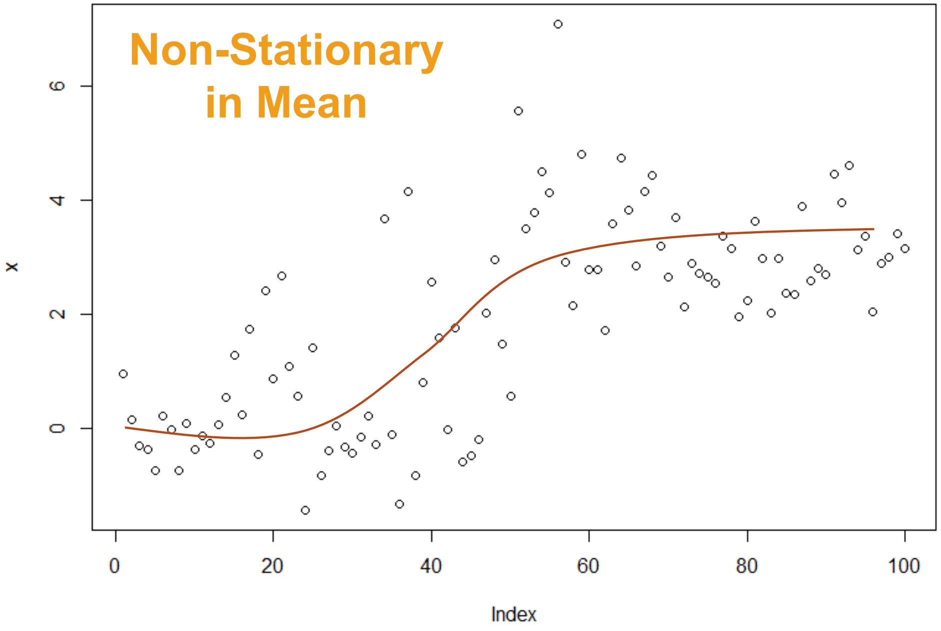
Use the order (index) or time as the x-axis variable

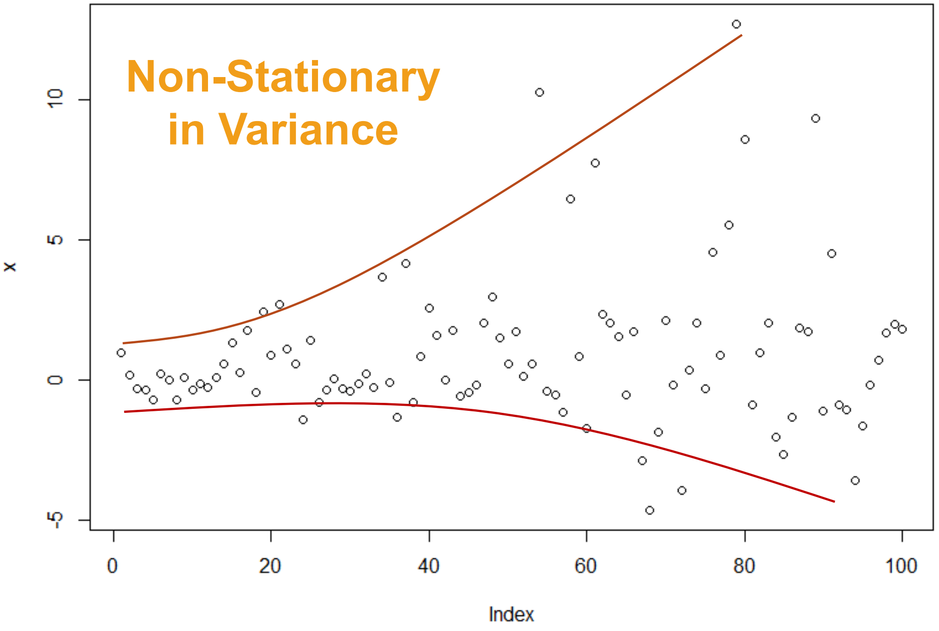
Used to test:

- Randomness
- Fixed location
- Fixed variation

Run Sequence and Time Series Plot Diagnostics





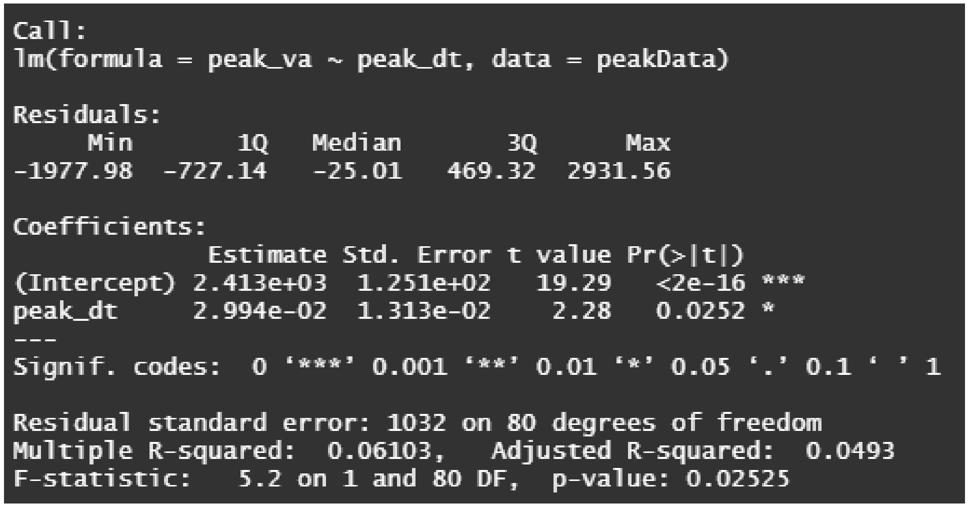


Non-Stationarity

- Properties of the time series are changing with respect to time
- Can be attributed to physical causes
  - Land use change/urbanization
  - Climate change
- Manifests as changes in mean or variance
- Often can be identified visually

Detecting Non-Stationarity

- Run sequence/time series plot
- Check data flags
- Split sample testing
- Simple regression
- Nonstationarity Detection Tool



```
Call:
lm(formula = peak_va ~ peak_dt, data = peakData)

Residuals:
     Min       1Q   Median       3Q      Max 
-1977.98  -727.14   -25.01   469.32  2931.56 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2.413e+03  1.251e+02   19.29   <2e-16 ***
peak_dt     2.994e-02  1.313e-02    2.28   0.0252 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1032 on 80 degrees of freedom
Multiple R-squared:  0.06103,	Adjusted R-squared:  0.0493 
F-statistic:   5.2 on 1 and 80 DF,  p-value: 0.02525
```

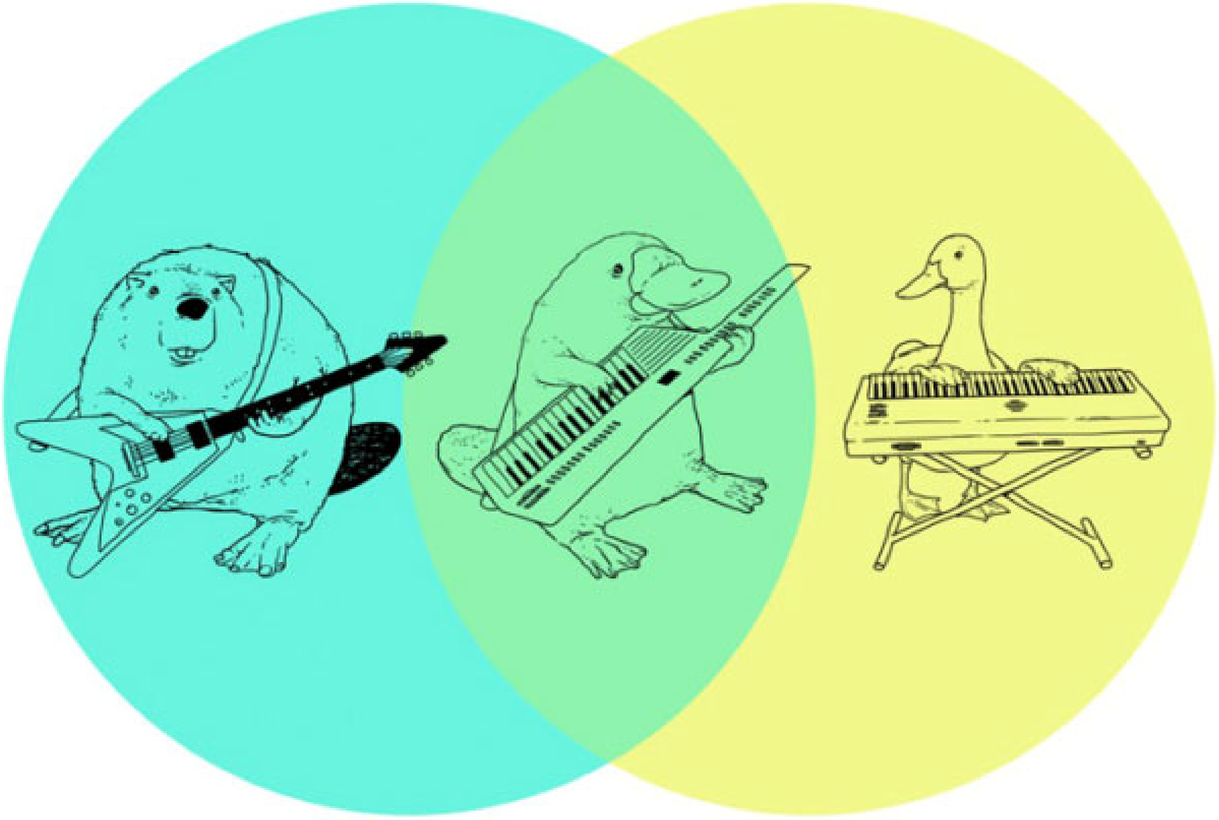
Basic Probability and Statistics: **Events and Relationships – Venn Diagrams**

Flood Frequency Analysis

Greg Karlovits, PE, PH, CFM

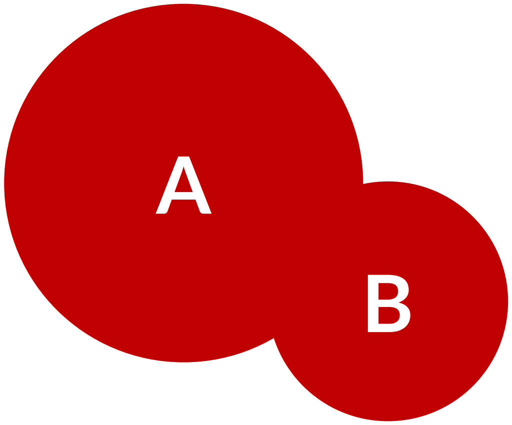
Hydrologic Engineering Center, May 2022

Venn Diagrams

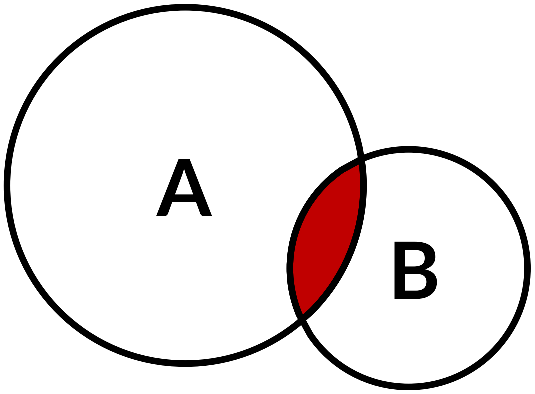


Union and Intersection

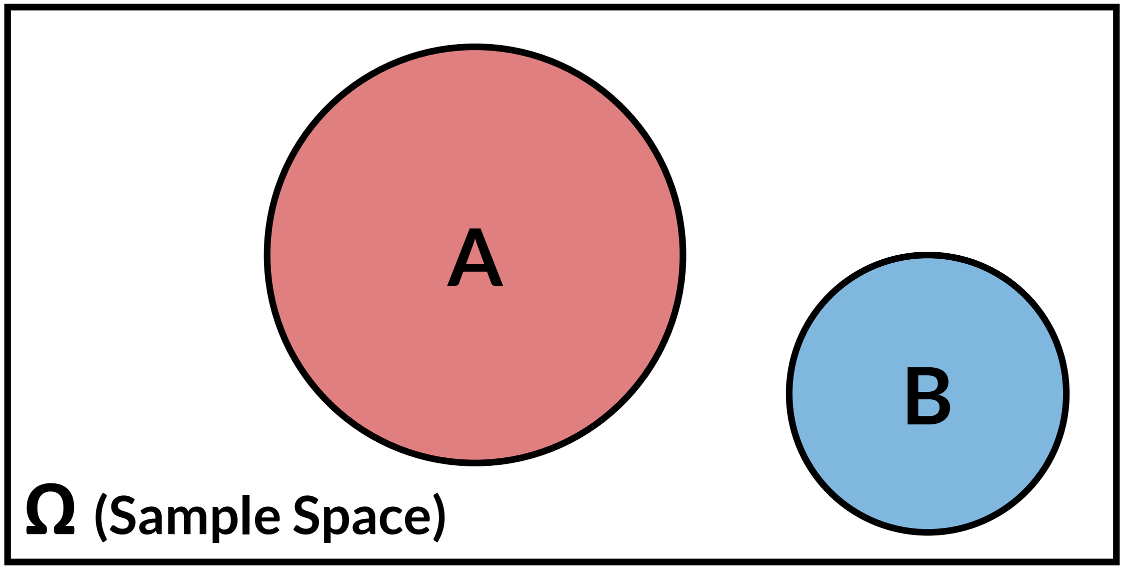
Union
A OR B



Intersection
A AND B



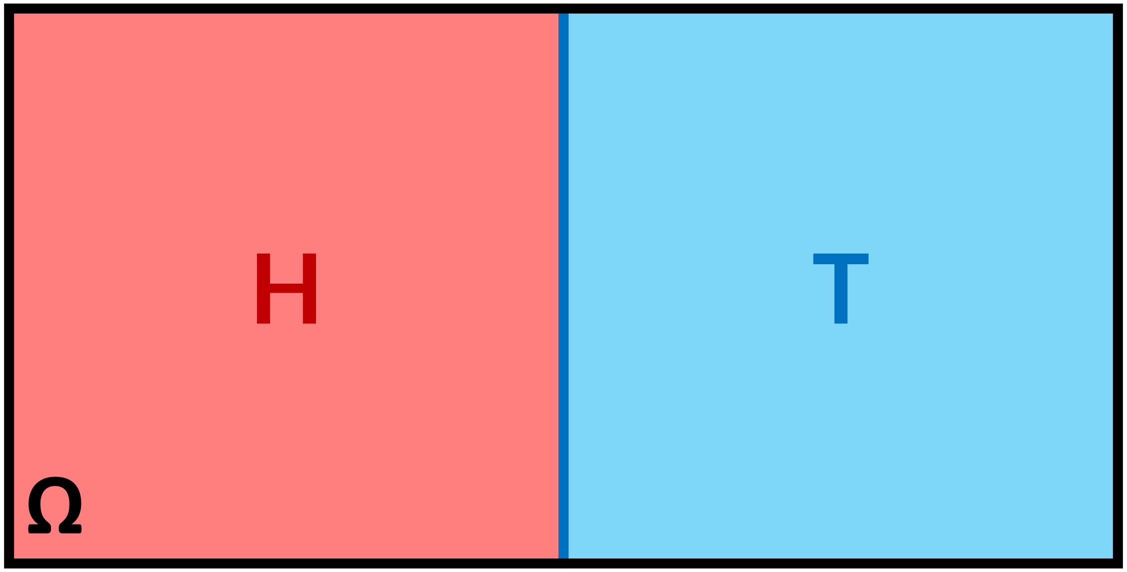
Venn Diagrams



$$p(\Omega) = 1$$
$$p(A\ or\ B) = p(A) + p(B)$$

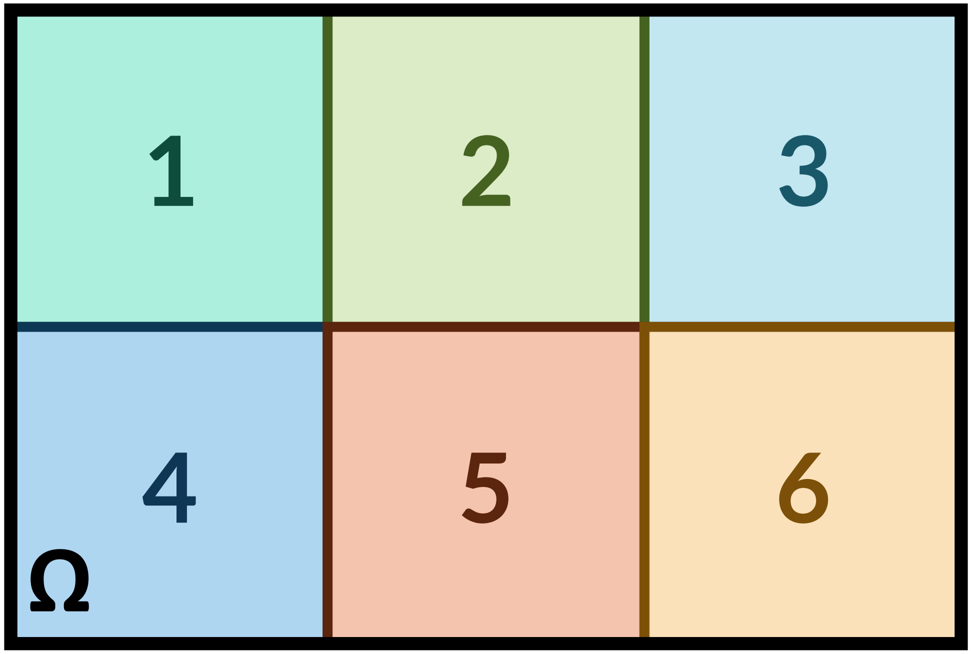
Coin Flip

- Mutually exclusive and exhaustive



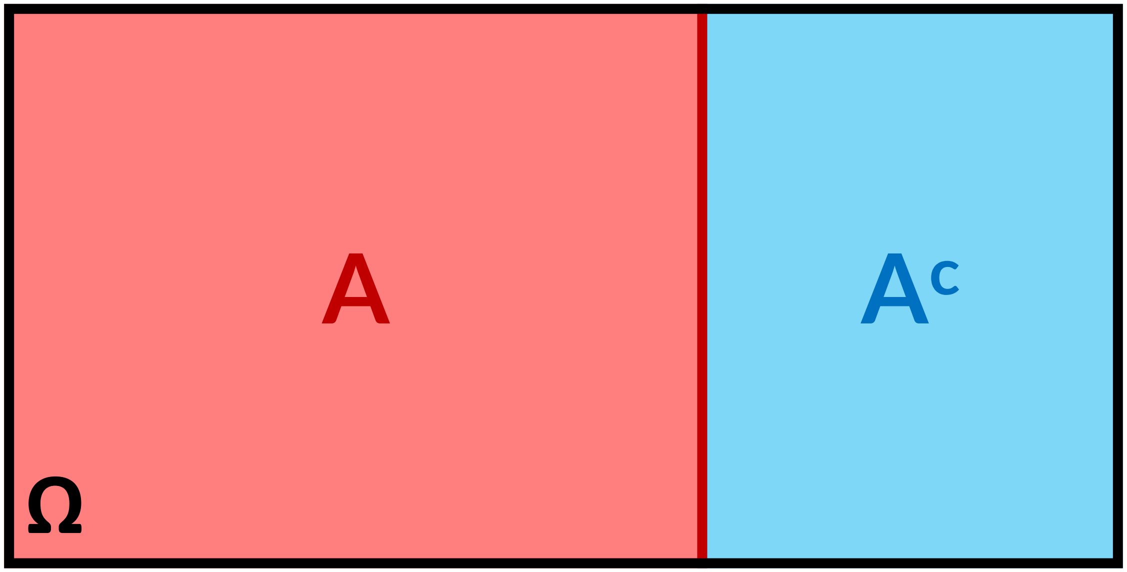
$$p(A\ or\ B) = p(A) + p(B) = 1$$

Die Roll



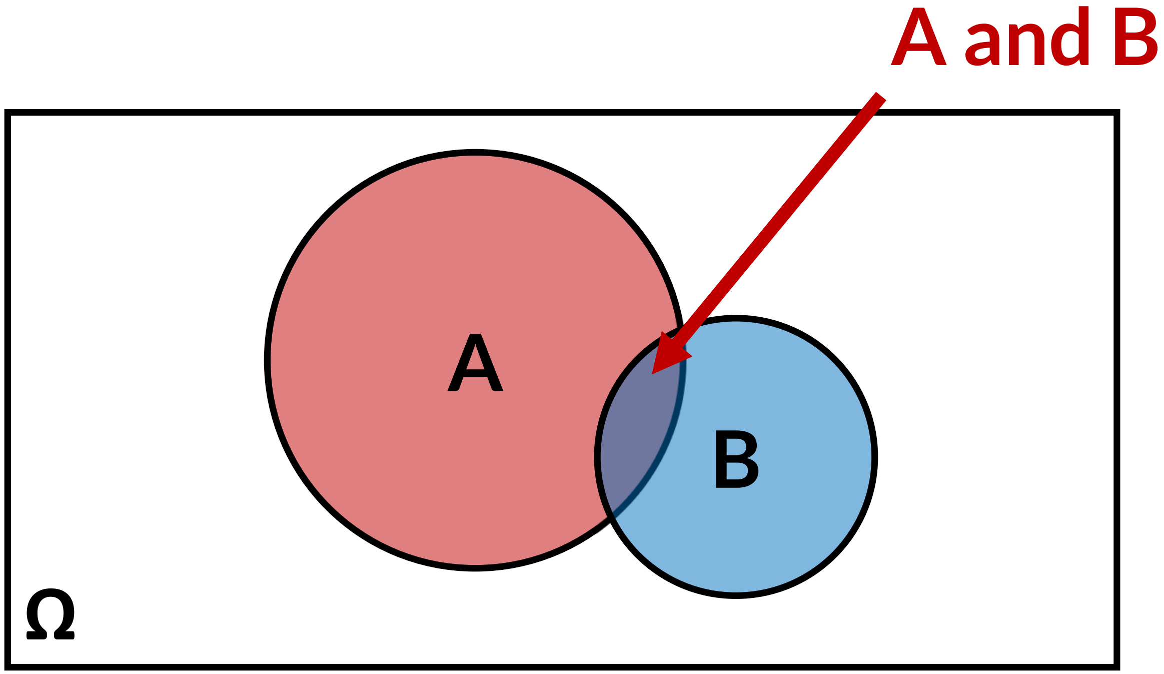
Complements

- All the space in “not A”



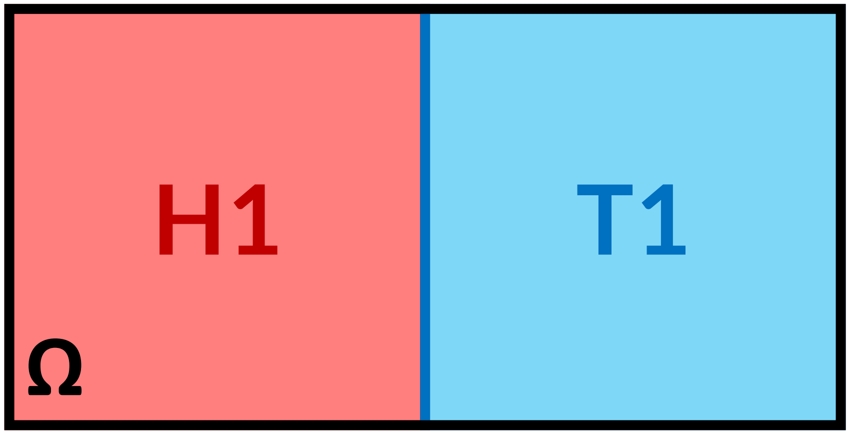
$$p(A^c) = p(\Omega) - p(A) = 1 - p(A)$$

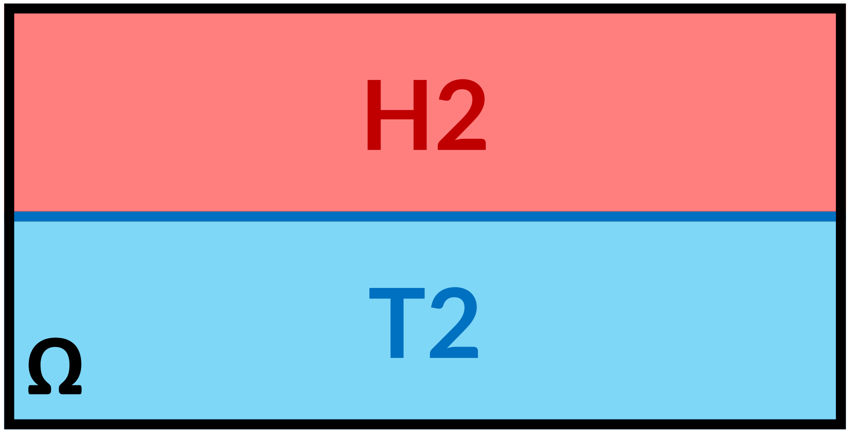
When Events Collide – General Additivity



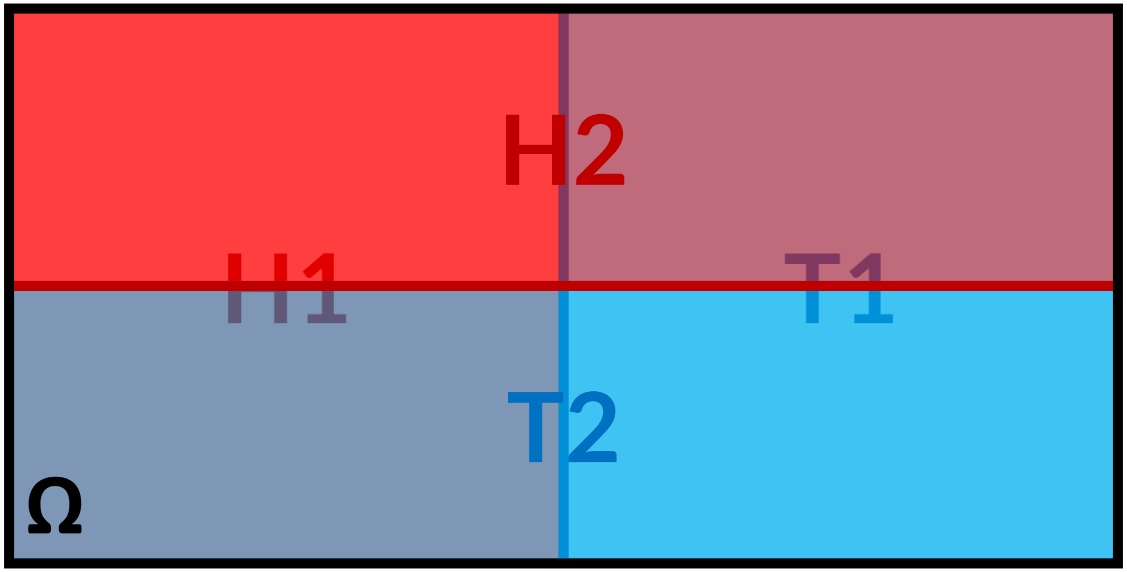
$$p(A\ or\ B) = p(A) + p(B) - p(A\ and\ B)$$

Two Coins



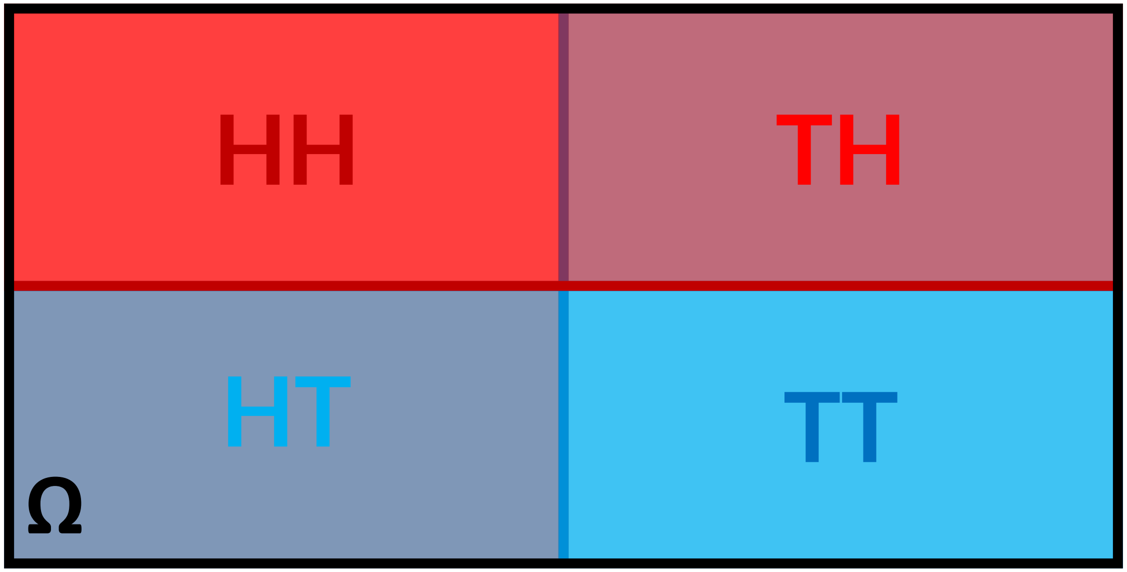


Two Coins



$$p(H1\ or\ H2) = p(H1) + p(H2) - p(H1\ and\ H2)$$

Two Coins



$$p(H1 \text{ } and \text{ } H2) = p(H1) * p(H2)$$

$$p(H1 \text{ } or \text{ } H2) = p(H1) + p(H2) - p(H1) * p(H2)$$

Only because H1 and H2 are independent!

Independence

Joint probability of A and B

Marginal probability of B

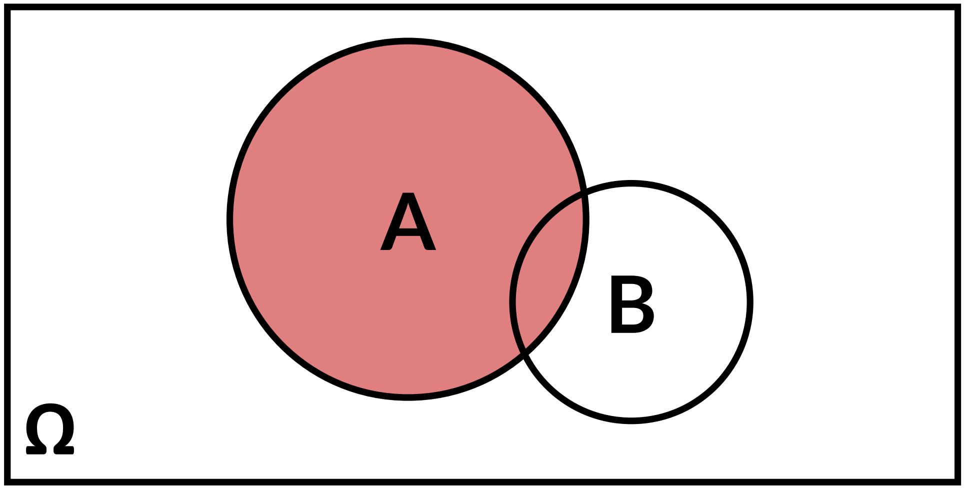
$$p(A\ and\ B) = p(A) * p(B)$$

iff $A \perp B$

Marginal probability of A

Marginal Probability

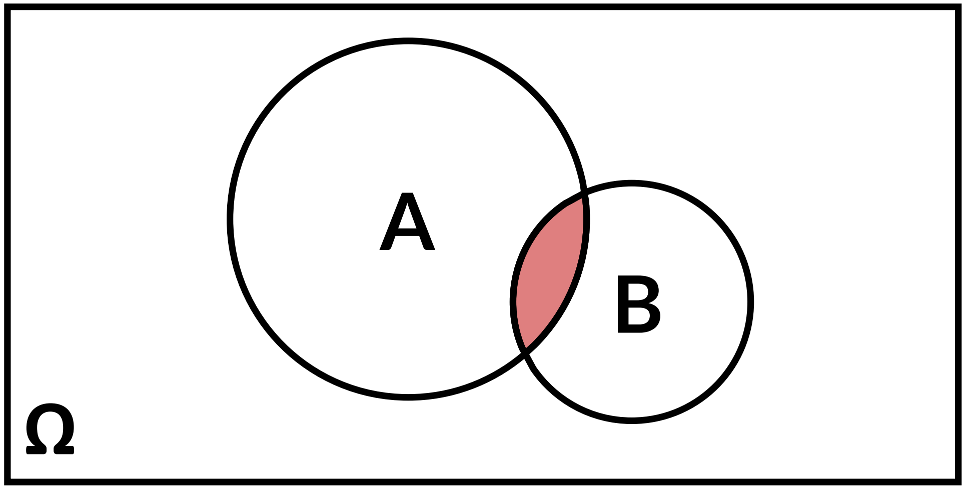
- *What is the probability of A occurring irrespective of what happens with B?*

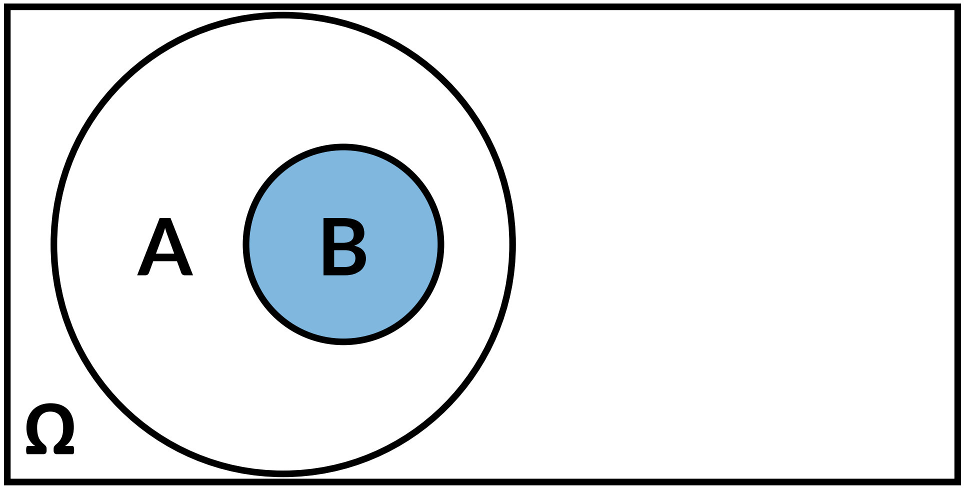


Joint Probability

- *What is the probability of A and B occurring together?*

$$p(A\ and\ B)$$

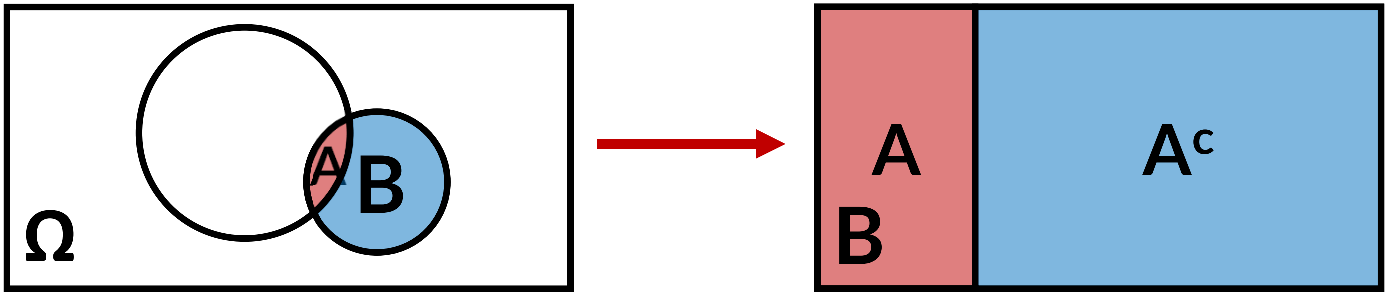




Conditional Probability

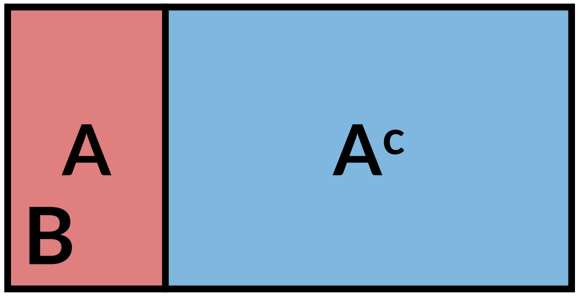
- *Given that B has occurred, what is the probability that A occurs?*
- Once we have observed that B has occurred, it becomes our "universe"

$$p(A|B)$$



Conditional and Joint Probability

$$p(A|B) = \frac{p(A\ and\ B)}{p(B)}$$



if A $\perp$ B,

$$p(A|B) = \frac{p(A) * p(B)}{p(B)} = p(A)$$