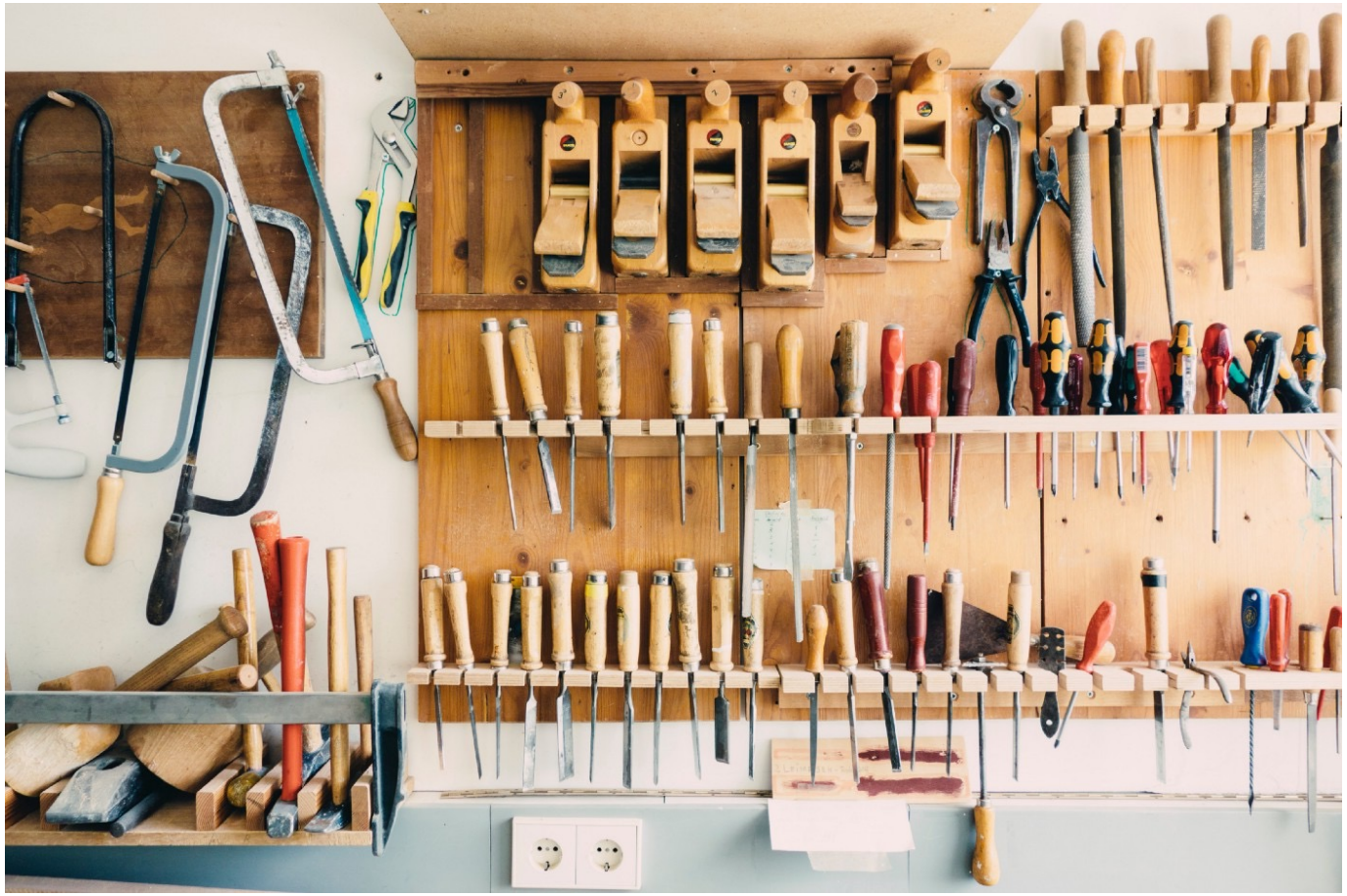
Statistical Methods

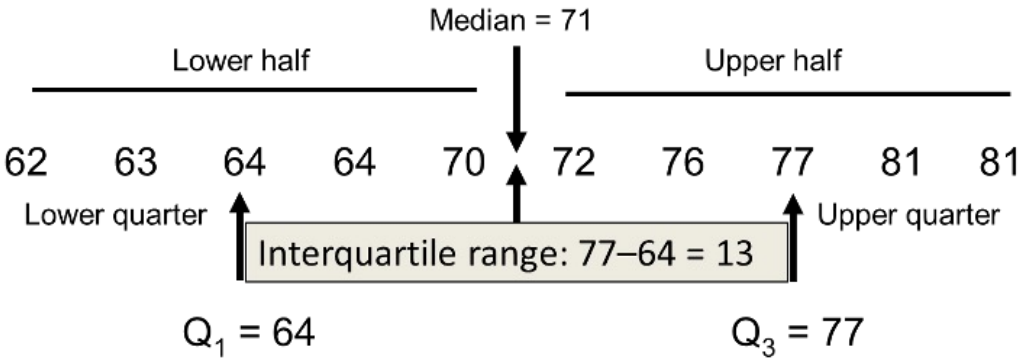


Thinking of statistical methods as **tools,** instead of **tests**

Philip Sweet Summer 2021

What can we do with statistics?

1. Estimate parameters
   - Mean, Median, SD, IQR



2. Make and test predictions
   - Use a sample to predict the populations (t-tests)

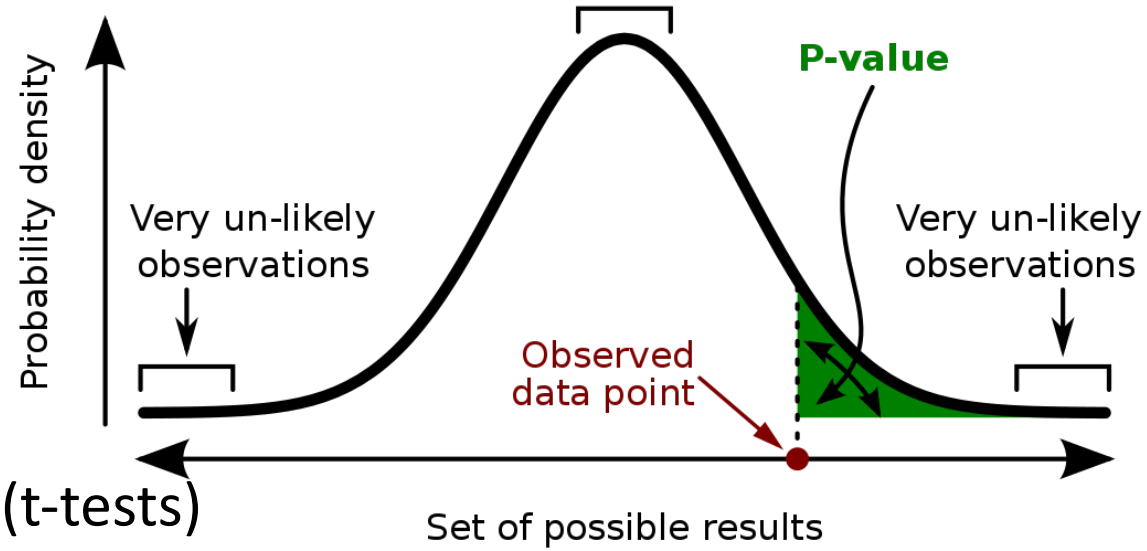
3. Build and compare models
   - Regressions, classifiers (ie ”big data” stuff)

What can we do with statistics?

1. Estimate parameters
   - Mean, Median, SD, IQR

2. Make and test predictions
   - Use a sample to predict the population (t-tests)

3. Build and compare models
   - Regressions, classifiers (ie ”big data” stuff)

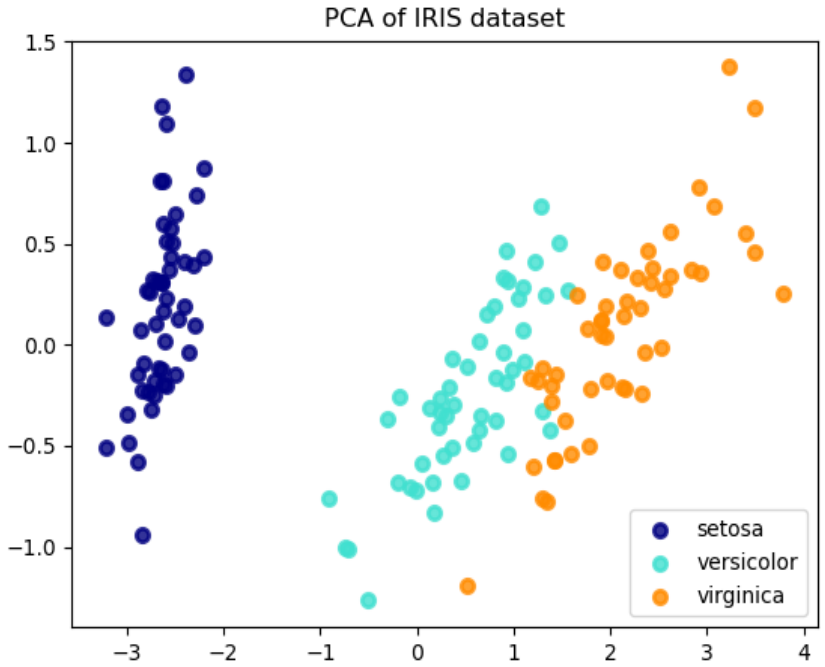


What can we do with statistics?

1. Estimate parameters
   - Mean, Median, SD, IQR

2. Make and test predictions
   - Use a sample to predict the populations (t-tests)

3. Build and compare models
   - Regressions, ANOVA, classifiers (ie ”big data” stuff)



Conducting meaningful statistical analysis requires that you understand your data

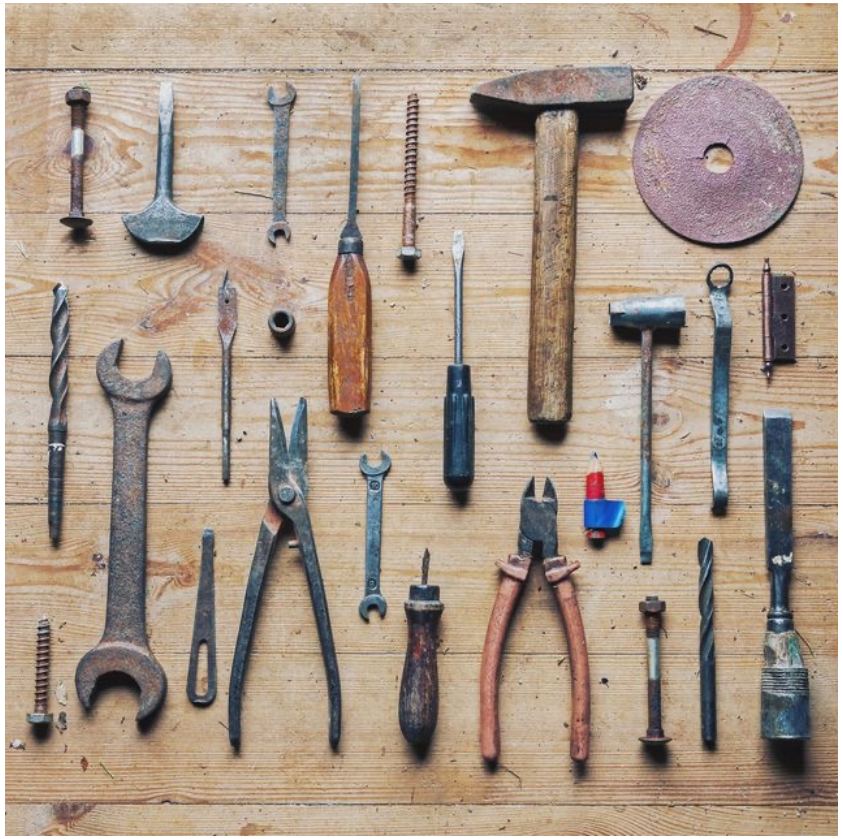
Understanding your data

- What kind of data are you working with?
- How much data will you have?
- Do you have a sense of how it will be distributed?
- Are you trying to be descriptive or predictive?

Conducting meaningful statistical analysis requires that you know your tools

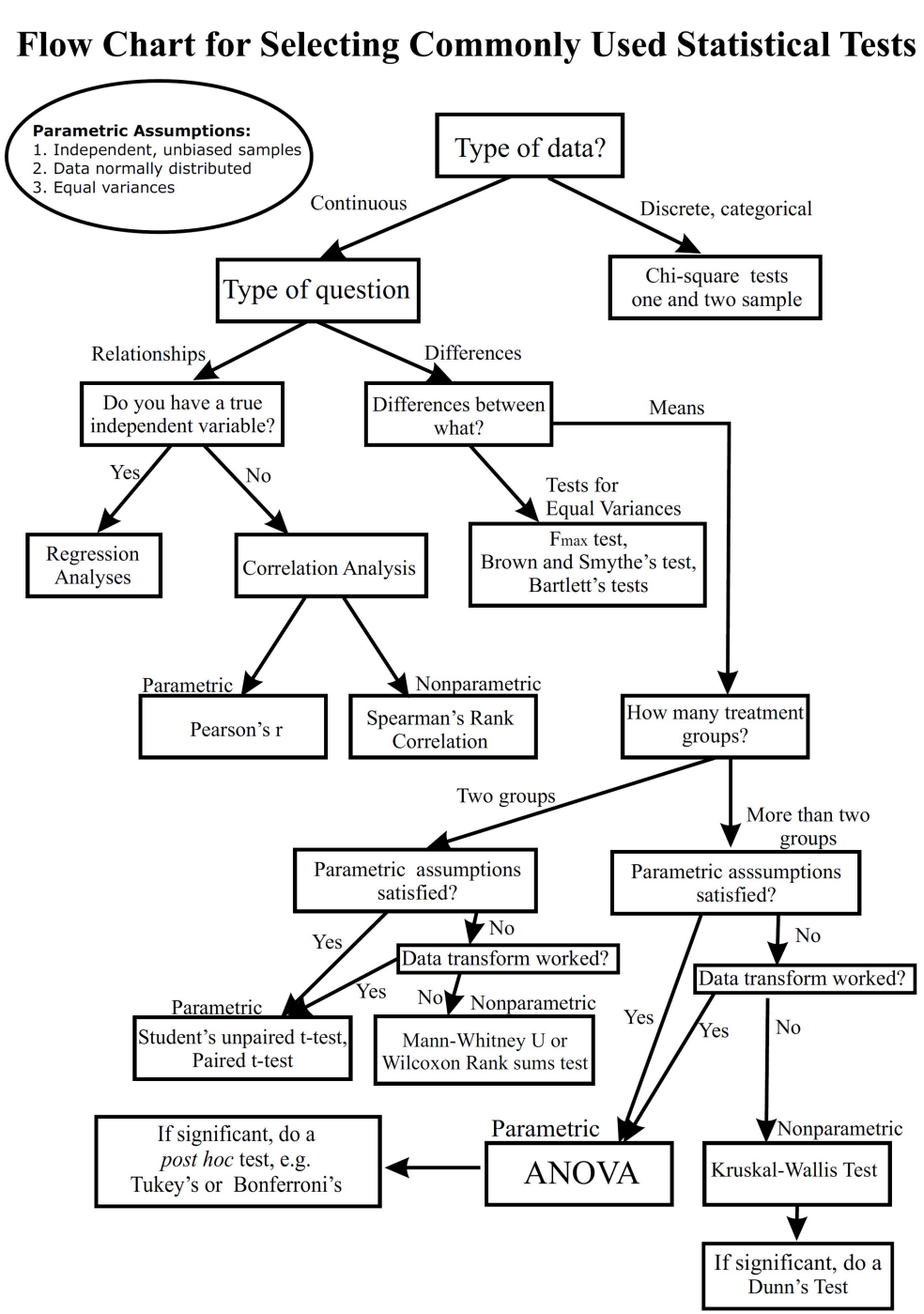
Classes of Statistical Methods

- *Frequentist*
  - **Parametric**
    - T-tests
    - ANOVA/F-test
    - Regression
    - Chi-squared
  - **Non-parametric**
    - Mann–Whitney/Kolmogorov–Smirnov
    - Bootstrapping
- *Bayesian*
  - ***Bayes' Theorem***
    - Naïve Bayes Classifier
    - Markov Chain Monte Carlo
- *~~Machine Learning~~*



Methods Crash Course

- Introduction to Classes of Statistics
  - Frequentist vs Bayesian
  - Parametric vs Non-parametric
  - Bayesian Statistics
- Biology Examples
  - **Parametric:**
    - ANOVA comparing mean of qPCR data
  - **Non-parametric:**
    - KS test of distribution of cell sizes
  - **Bayesian:**
    - Assign bird gender using weight
    - Model development



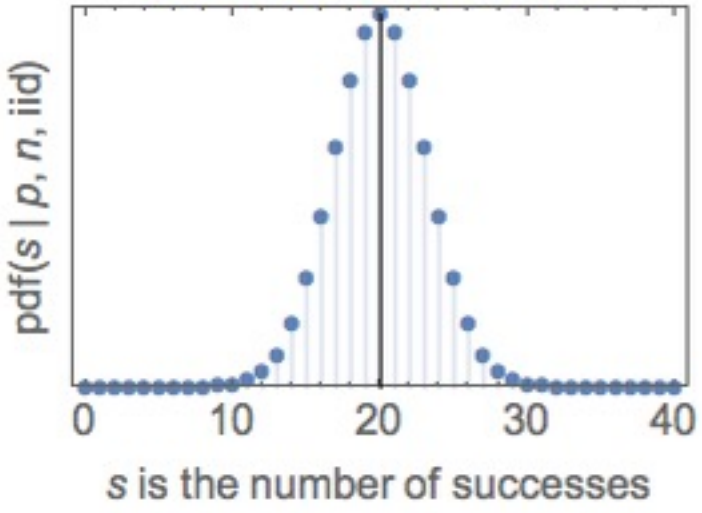
Frequentist vs Bayesian Statistics

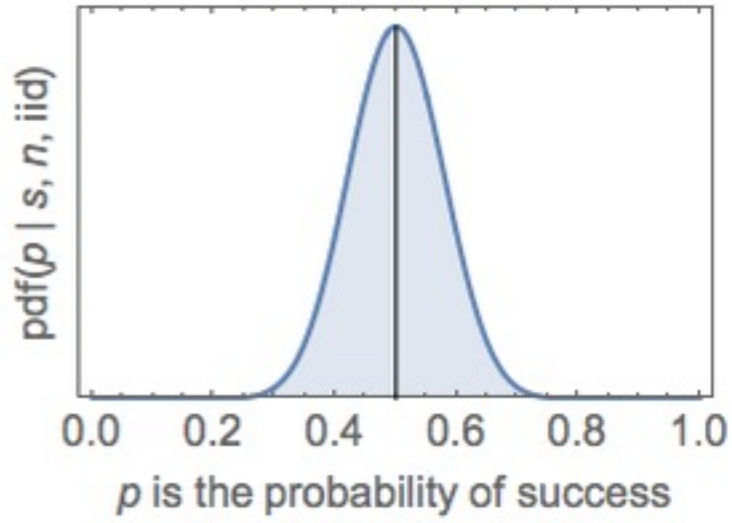
Frequentist

- Based on the probability that the experiment would have the same outcomes if it were to replicated
- Only takes into account the data collected
- Tests against randomness, ie a ”positive” result is that the data/outcome isn’t random

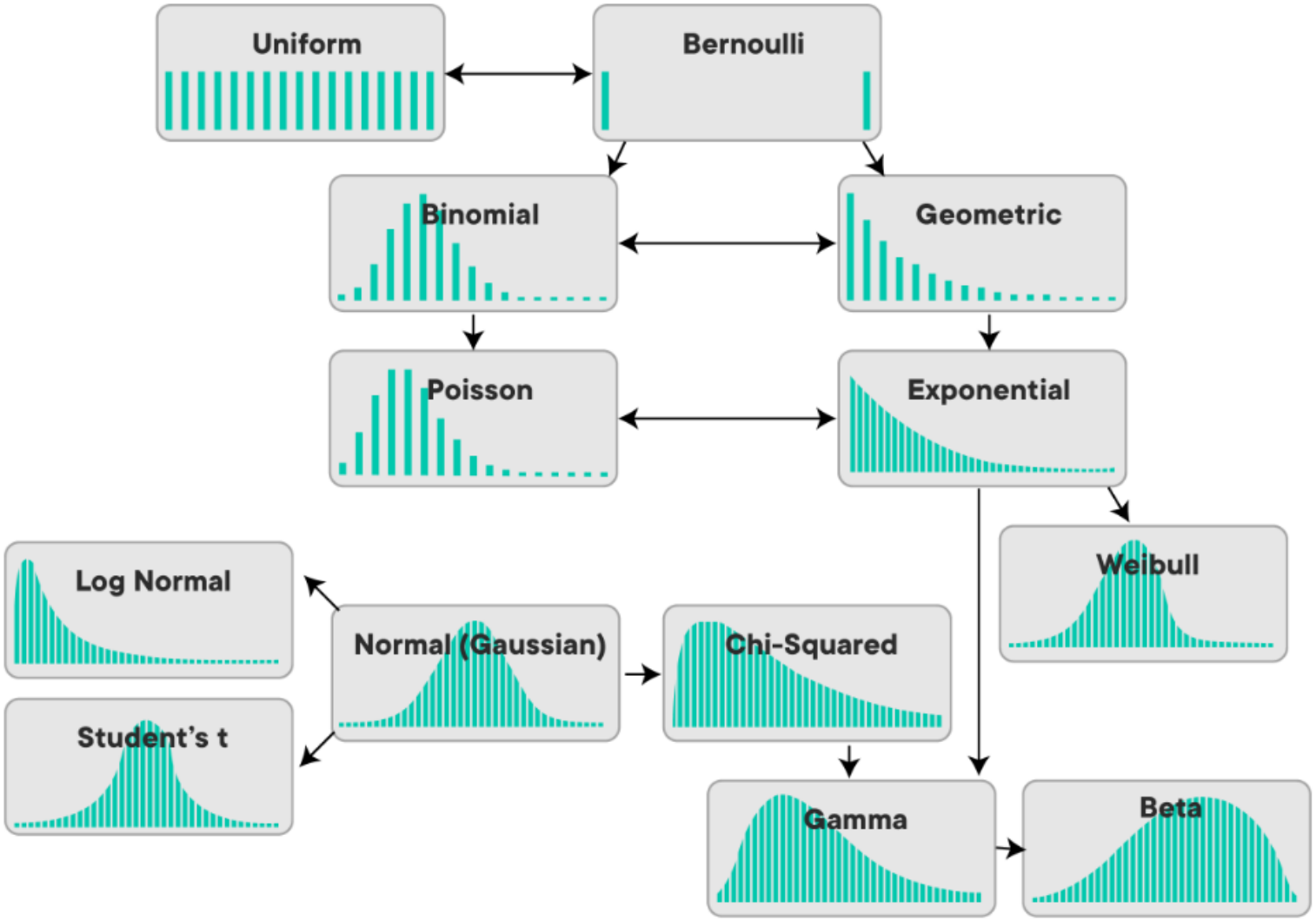
Bayesian

- Uses probability to expresses a degree of belief in an experimental outcome
- Takes into account existing data and assumptions as well as the current experiment data
- Tests which model is more predictive (more true)



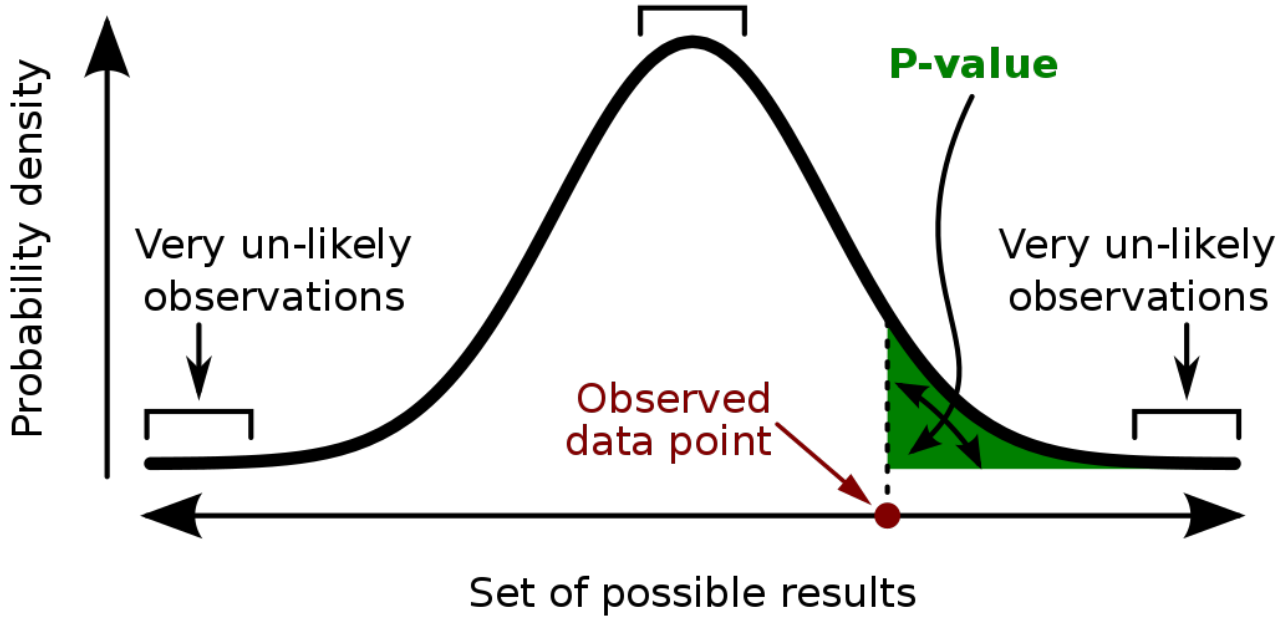


Parametric Methods Rely on a Distribution



Parametric Methods Rely on a Distribution

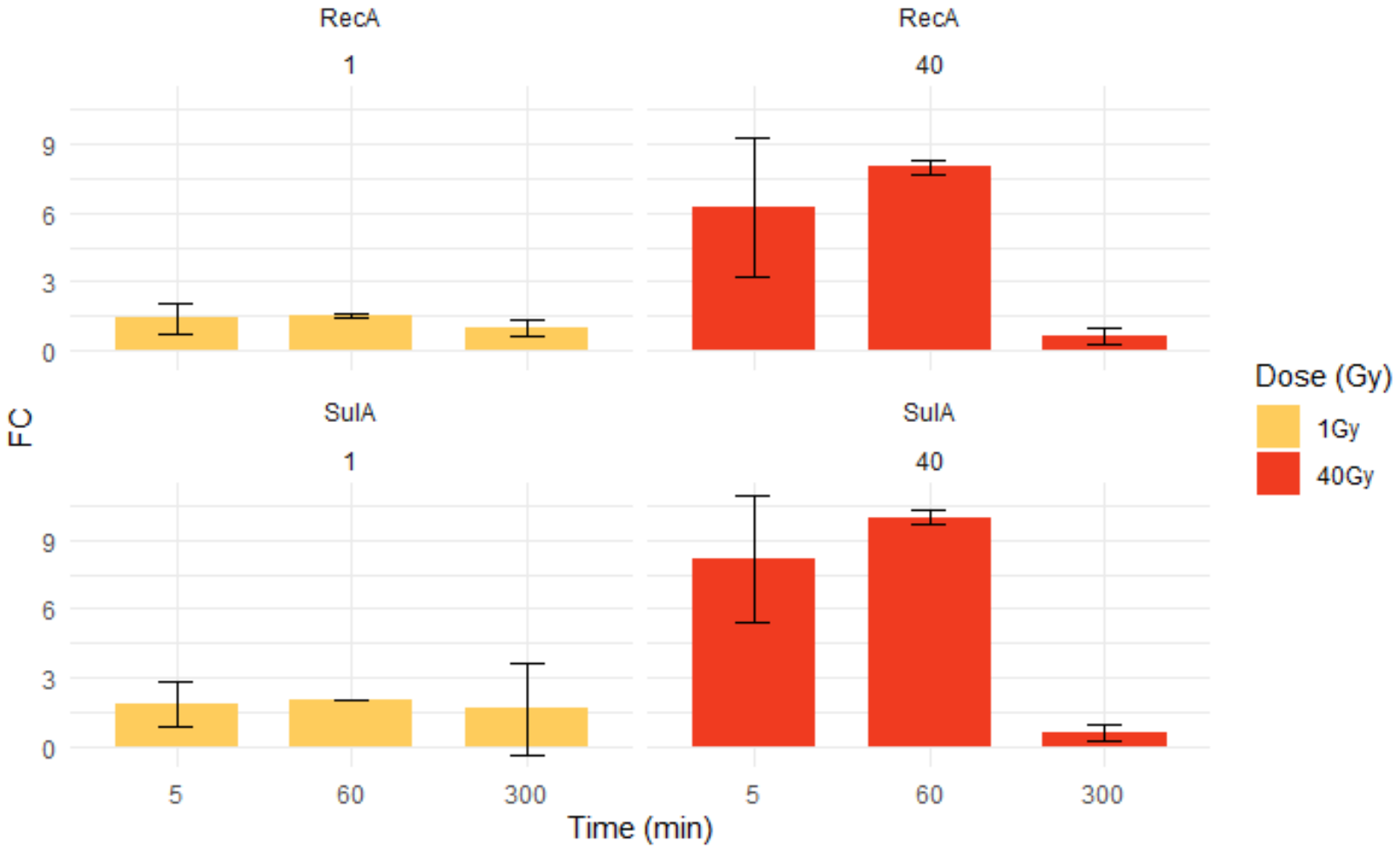
- Most of the commonly employed statistical tests in the life sciences are parametric.
- Tests rely on the data coming from an assumed distribution against which hypothesis can be tested
- Null Hypothesis = Assumption
- Alt Hypothesis = Prediction
- Parametric test allow you to **REJECT** the null, but they do not **PROVE** the alternative.



Example: Analyzing qPCR data with ANOVA

DeltaDeltaCT Method

1. **CT** = abundance of gene in the sample
2. **DeltaCT** = internally normalized abundance of the gene
3. **DeltaDeltaCT** = relative abundance to the control sample
4. From this we can calculate a foldchange by raising the 2^DeltaDeltaCT

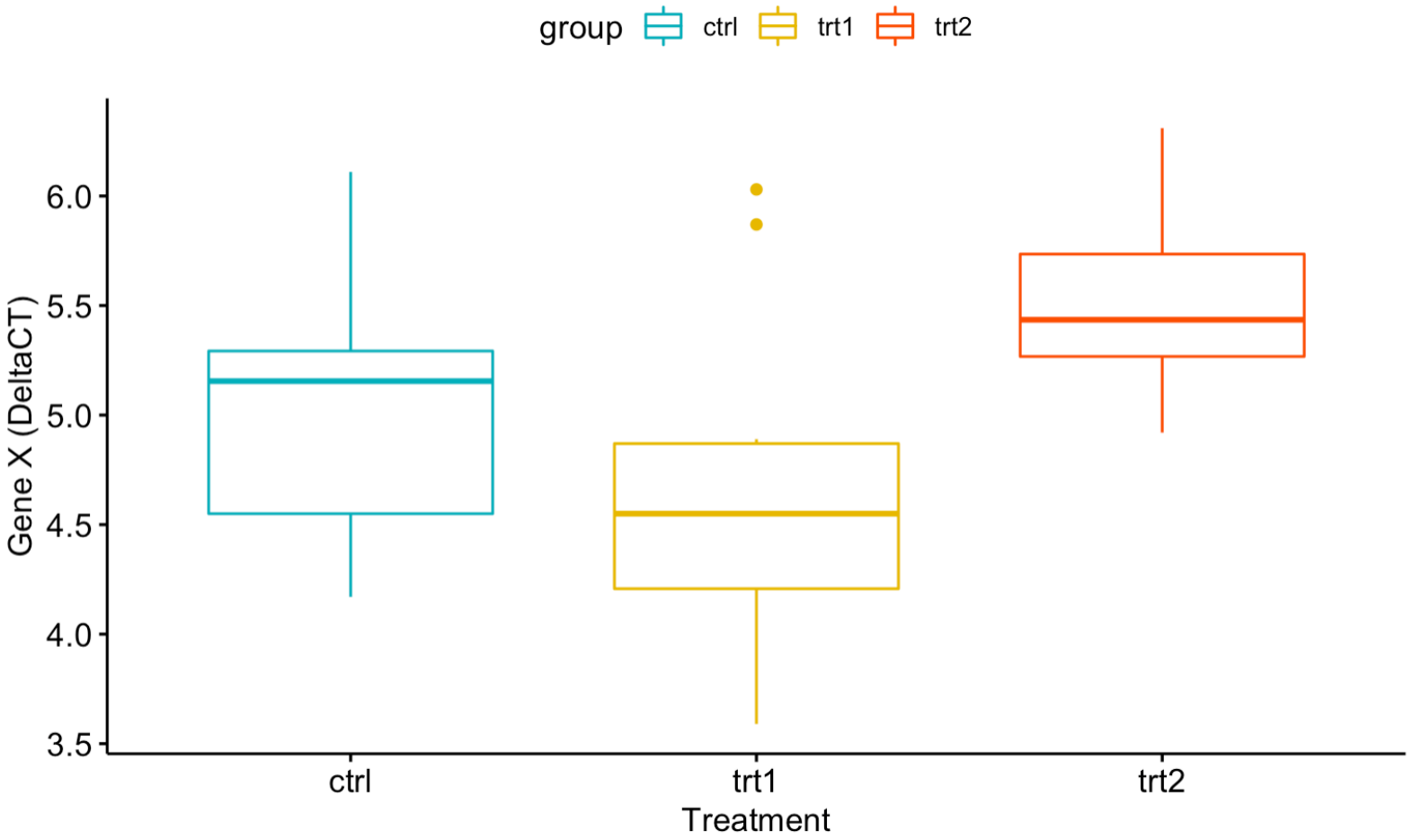


Example: Analyzing qPCR data with ANOVA

DeltaDeltaCT Method

1. **CT** = abundance of gene in the sample
2. **DeltaCT** = internally normalized abundance of the gene
3. **DeltaDeltaCT** = relative abundance to the control sample
4. From this we can calculate a foldchange by raising the 2^DeltaDeltaCT

For determining significance of the effect, it is best to use the DeltaCT since it is the least modified



Example: Analyzing qPCR data with ANOVA

```r
# Compute the analysis of variance
res.aov <- aov(weight ~ group, data = my_data)
# Summary of the analysis
summary(res.aov)
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
group        2  3.766  1.8832   4.846 0.0159 *
Residuals   27 10.492  0.3886
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

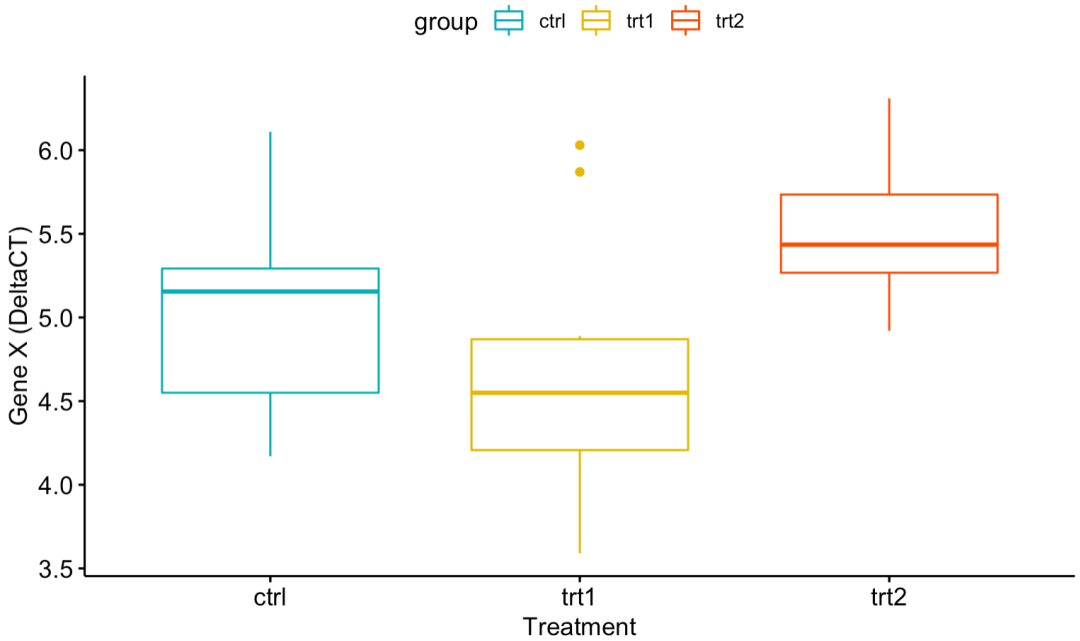
✔ The output includes the columns *F value* and *Pr(>F)* corresponding to the p-value of the test.

http://www.sthda.com/english/wiki/one-way-anova-test-in-r

Example: Analyzing qPCR data with ANOVA

```
TukeyHSD(res.aov)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level
Fit: aov(formula = weight ~ group, data = my_data)
$group
             diff        lwr       upr     p adj
trt1-ctrl -0.371 -1.0622161 0.3202161 0.3908711
trt2-ctrl  0.494 -0.1972161 1.1852161 0.1979960
trt2-trt1  0.865  0.1737839 1.5562161 0.0120064
```



- **diff**: difference between means of the two groups
- **lwr**, **upr**: the lower and the upper end point of the confidence interval at 95% (default)
- **p adj**: p-value after adjustment for the multiple comparisons.

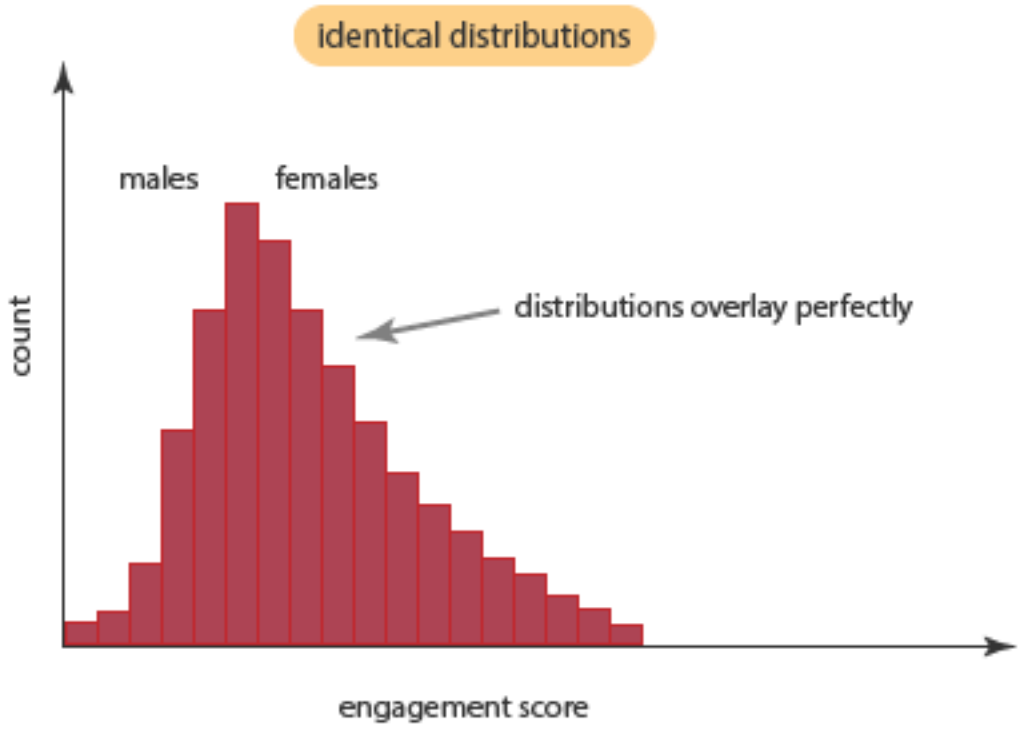
✔ It can be seen from the output, that only the difference between trt2 and trt1 is significant with an adjusted p-value of 0.012.

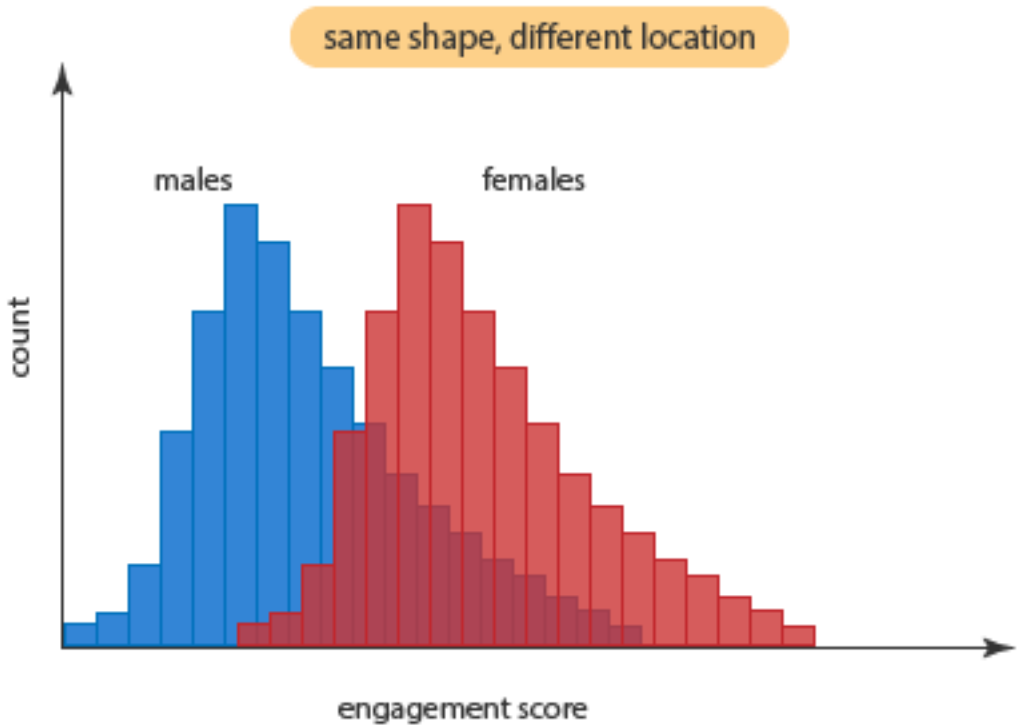
http://www.sthda.com/english/wiki/one-way-anova-test-in-r

Non-Parametric Methods Only Need Data

- For when your data looks “weird” or you want to compare distributions of data
- A very broad category of methods defined by what they aren’t, i.e. **they use no assumptions about the shape of the data**
- Each method builds a distribution using the data to be tested
- Often rely on ranking the data
- Good for low occurrence data sets in which you want to establish a difference
- Still based around Null and Alt hypothesis

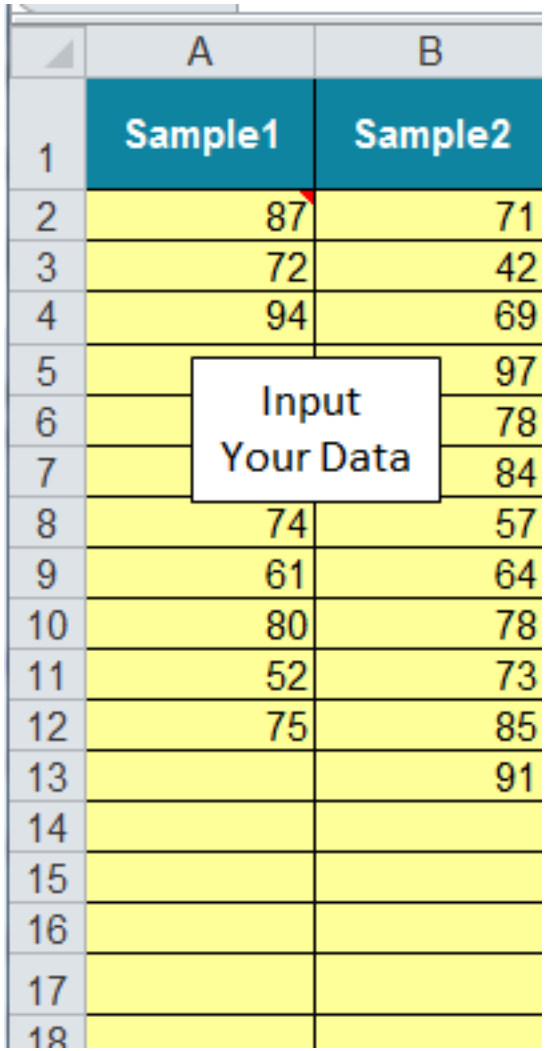
Mann-Whitney U Test of Independence



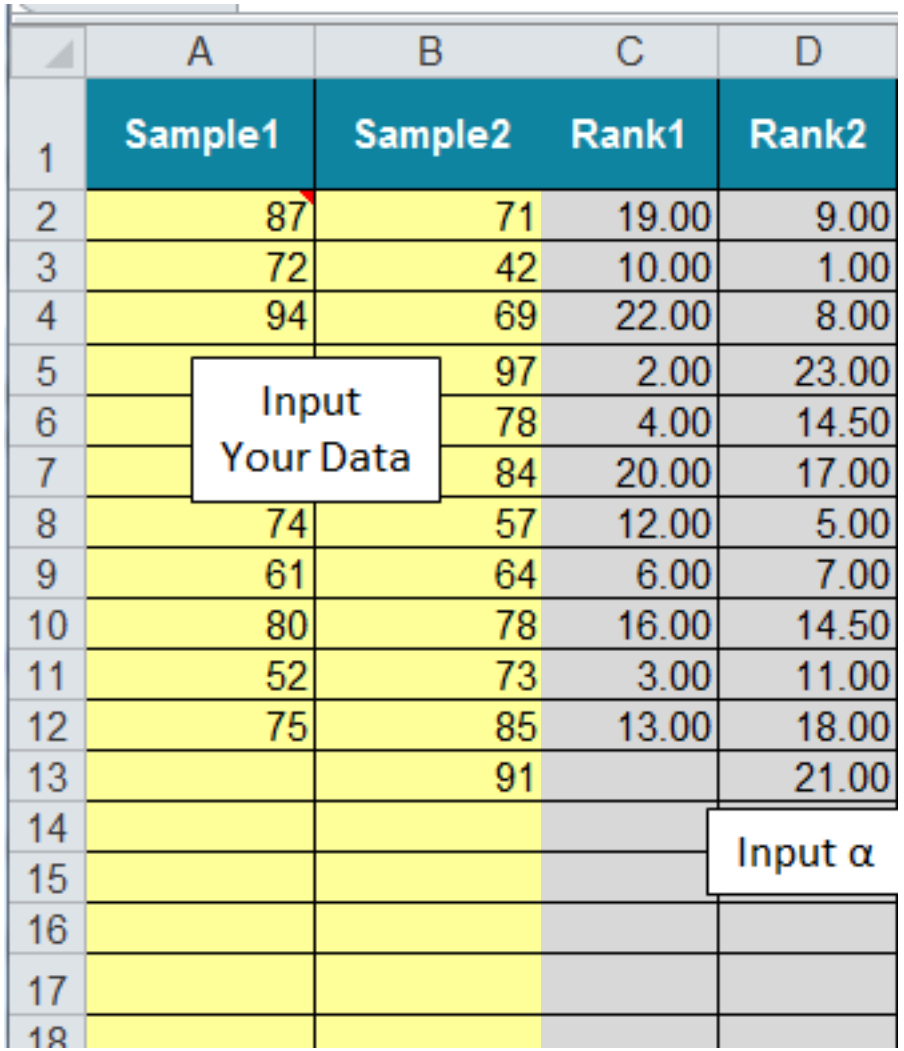


https://statistics.laerd.com/spss-tutorials/mann-whitney-u-test-using-spss-statistics.php

Mann-Whitney U Test of Independence

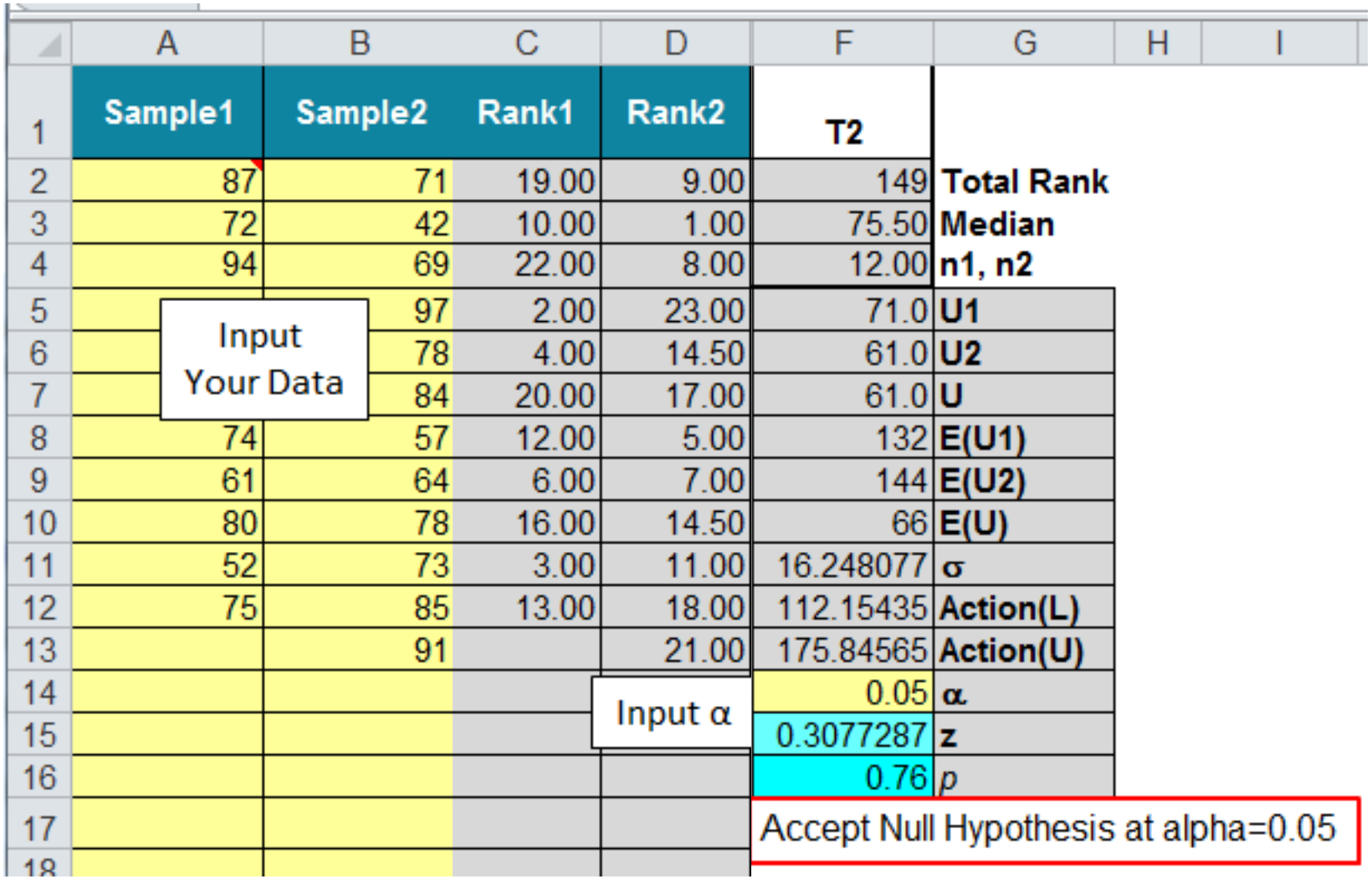


Mann-Whitney U Test of Independence



	A	B	C	D
1	Sample1	Sample2	Rank1	Rank2
2	87	71	19.00	9.00
3	72	42	10.00	1.00
4	94	69	22.00	8.00
5		97	2.00	23.00
6		78	4.00	14.50
7		84	20.00	17.00
8	74	57	12.00	5.00
9	61	64	6.00	7.00
10	80	78	16.00	14.50
11	52	73	3.00	11.00
12	75	85	13.00	18.00
13		91		21.00
14				
15				
16				
17				
18				

Mann-Whitney U Test of Independence



	A	B	C	D	F	G	H	I
1	Sample1	Sample2	Rank1	Rank2	T2			
2	87	71	19.00	9.00	149	Total Rank		
3	72	42	10.00	1.00	75.50	Median		
4	94	69	22.00	8.00	12.00	n1, n2		
5		97	2.00	23.00	71.0	U1		
6		78	4.00	14.50	61.0	U2		
7		84	20.00	17.00	61.0	U		
8	74	57	12.00	5.00	132	E(U1)		
9	61	64	6.00	7.00	144	E(U2)		
10	80	78	16.00	14.50	66	E(U)		
11	52	73	3.00	11.00	16.248077	σ		
12	75	85	13.00	18.00	112.15435	Action(L)		
13		91		21.00	175.84565	Action(U)		
14					0.05	α		
15					0.3077287	z		
16					0.76	*p*		
17					Accept Null Hypothesis at alpha=0.05			
18								

Input Your Data

Input α

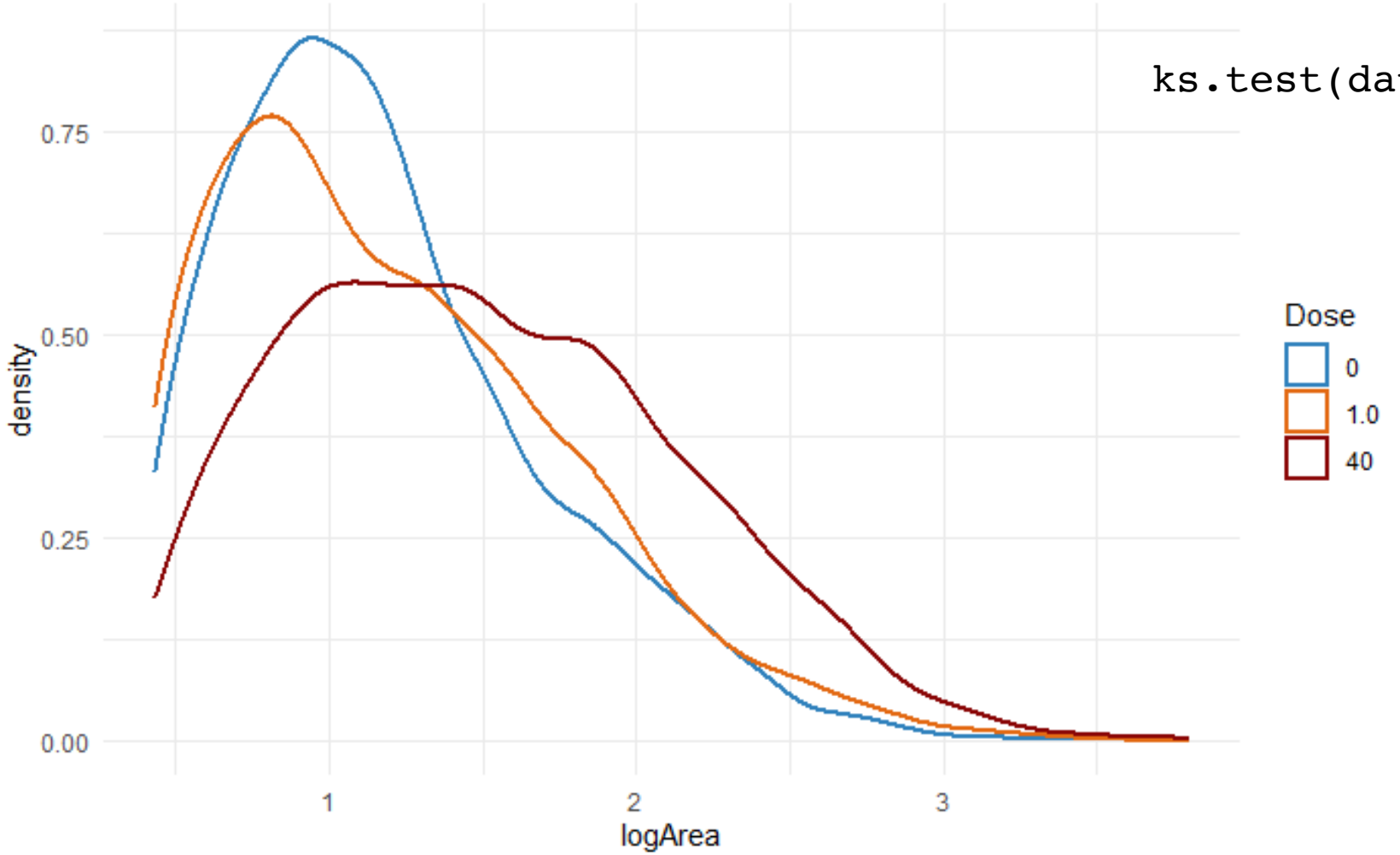
Mann-Whitney in R

```
#create a data frame with two columns, one for each group
drug_data <- data.frame(attacks = c(3, 5, 1, 4, 3, 5, 4, 8, 6, 2, 1, 9),
                        drug_group = c(rep("old", 6), rep("placebo", 6)))

#perform the Mann Whitney U test
wilcox.test(attacks~drug_group, data = drug_data)

#output
data:  attacks by drug_group
W = 13, p-value = 0.468
alternative hypothesis: true location shift is not equal to 0
```

Example: Comparing the distribution of cell sizes using the Kolmogorov–Smirnov test

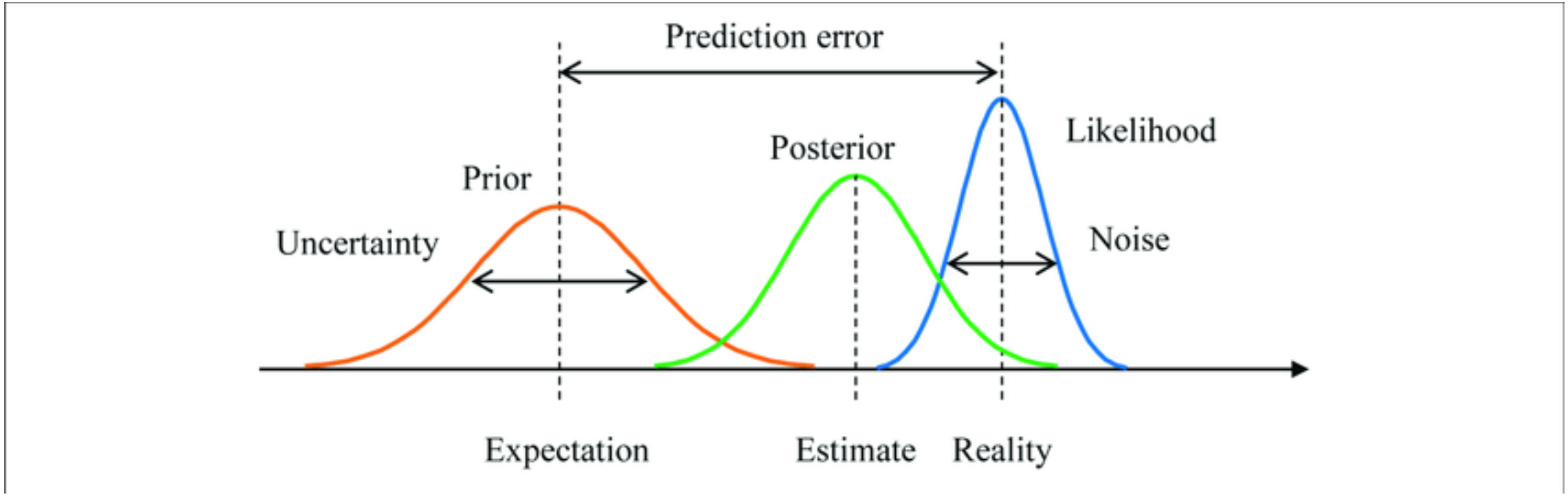



```
ks.test(data$SampleA, data$sampleB)
```

Dose (Gy)	KS Test of Distribution
1	< 0.001
40	< 0.001

Bayesian Statistics

- Based on probability i.e Bayes' Therom
- Useful for assigning classifications
- Let's you build models using **assumptions** and **collected data**



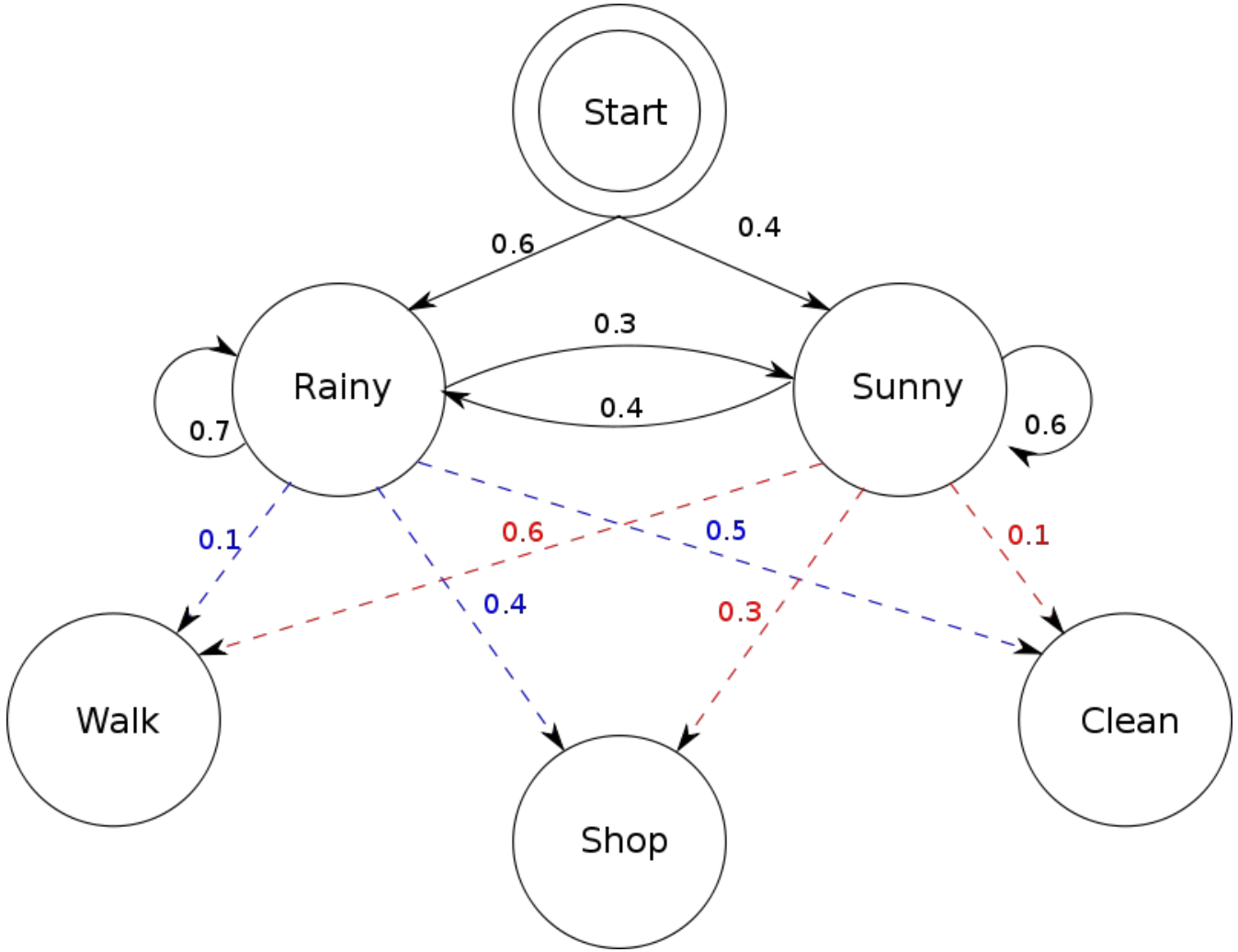
Example: Determining falcon gender using weight

We are ecologists with access to large number of falcon weights and we want to determine the gender of the measured birds.

- We know half of birds are male and half are female thus our **Likelihood** is 0.5
- A previous dataset tells us a weight ranges of male and female birds this is our **Prior**
- Combining the *likelihood* and *prior* produces the **Posterior**
- Each bird weight of unknown can then be feed into this model and a probable gender assignment can be produced

Posterior $P(A|B)$ = Likelihood $P(B|A)$ * Prior $P(A)$ / Evidence $P(B)$

Example: Sally Determines the Weather Using John's Actions Hidden Markov Model



Bayesian statistics is great for big data sets and building complex models

Example applications

- Determining ORFs in unannotated genomes
- Determining Introns and Exon boundaries
- Assigning cell type in mixed population
- Determining diseased states using biometrics
- Testing drug treatments

But what about Machine Learning??

"A computer program is said to **learn** from experience E with respect to some class of tasks T and performance measure P, if its performance at the tasks improves with the experiences." --- (Mitchell 1997)

- **Supervised:** You provide the labels on the data and the program tries to determine which data predict the label by building a model (Regression, Naïve Bayes Classifier, Random Forest)
- **Unsupervised:** You provide the data, the algorithm provides the clusters, no model is produced (PCA and K-means)
- **Deep Learning:** ~Neural Networks~ Huge amounts of data to train a black box classifier

Basic machine learning functions for R can be found in the caret package

Caret Resources

- https://topepo.github.io/caret/
- https://cran.r-project.org/web/packages/caret/vignettes/caret.html

Resources abound!

- Portfolio in applied statistical modeling
  - Like a minor for your PhD
  - https://stat.utexas.edu/graduate/portfolio-in-applied-statistical-modeling
- UT Summer Statistics Institute
  - Week-long courses in specific topics
  - https://stat.utexas.edu/training/ssi
- TACC Summer Institute
  - Computationally focused short courses
  - https://www.tacc.utexas.edu/education/institutes/intro-advanced-computing
- LinkedIn Learning
  - Free w/UTID
  - Go at your own pace
  - https://www.linkedin.com/learning/topics/statistics?u=36306084