Perform appropriate statistical tests

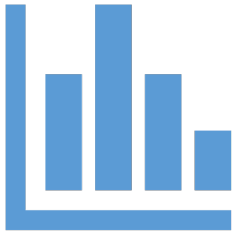
Serghei Mangul, Ph.D
Assistant Professor of Clinical Pharmacy and Biological Sciences,
University of Southern California

Learning Objectives

By the end of this module you should be able to:

- Relate some basic statistical foundations to real data
- Use NumPy and SciPy to apply statistical functions to data
- Use various tests of significance, such as:
- Apply statistical methods to analyze results

Statistics



What is statistics?

It represents how we can mathematically model our world

Why is this important?

- Defines how we construct experiments
- Lets us describe the accuracy of our measures
- Helps us put a quantitative measure to significance

Sampling

Statistics would be simple if we could measure an entire population

Not feasible in most situations

Instead we need to take samples from a larger population

Sampling (a bit more specifically)

Population: The entire pool of all subjects of interest

That could be humans living in North America, or all the cells of one individual

Sample: A subset of the population

The subjects we can actually measure

Parameter: The true measurement from the population

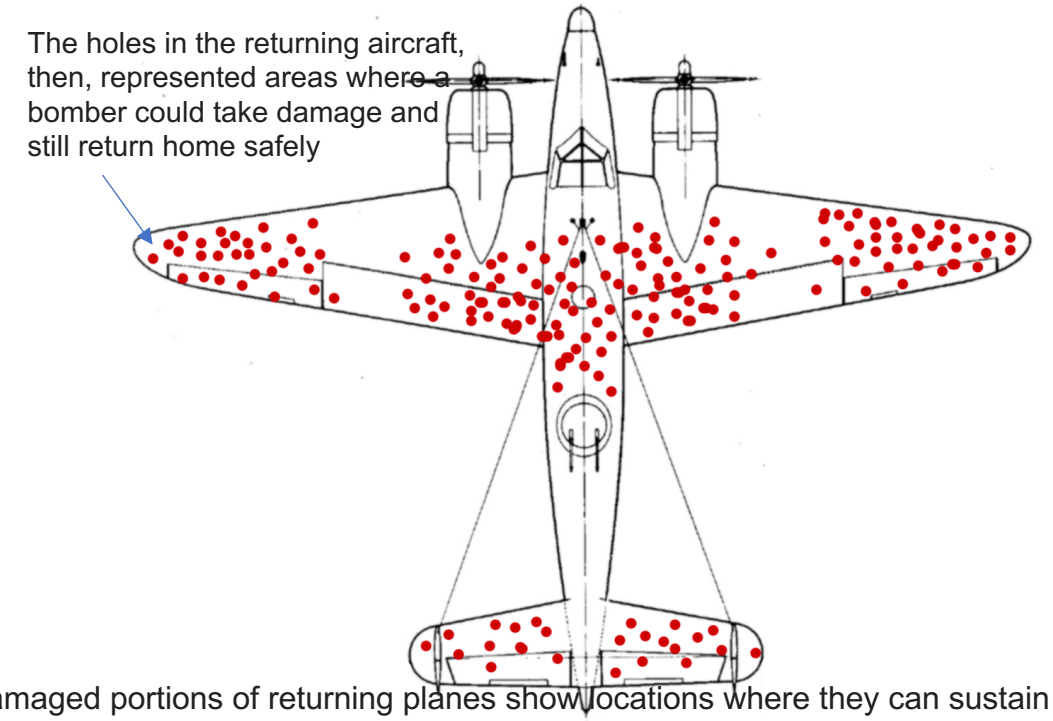
Typically we want to know what this is

Statistic: The measurement from the sample

This is how we will estimate the parameter

Sampling bias

- Populations are typically unified by at least one thing (e.g. location or disease)
  - Many other factors can still vary, and might create **subpopulations**
- Sampling bias comes from imbalanced sampling from these subpopulations
  - Typically due to the nature of how the population was sampled
- For example
  - Assume you are studying COVID-19
  - Unfortunately, you can only measure individuals who recovered



The damaged portions of returning planes show locations where they can sustain damage and still return home; those hit in other places do not survive

Distributions

- A description of the probability of every outcome possible for a random variable (a numerical description of the outcome of a statistical experiment)
- Two categories of distributions
  - Discrete: Measurements have distinct and separate outcomes
    - A die can only roll 1, 2, 3, 4, 5, or 6
    - A person can only make or miss a free throw
  - Continuous: Measurements have essentially infinite outcomes
    - Height can be any value within the range of human heights
    - Glucose levels vary across a range of measurements

Law of Large Numbers

- The **law of large numbers (LLN)** states that as a **sample** size grows, its mean gets closer to the average of the whole population
- What does this mean?
  - The larger your sample, the more likely your sample mean is close to the population mean
  - Somewhat intuitive, as more data means outliers make a smaller difference

Central Limit Theorem

- In the study of probability theory, the **central limit theorem** (CLT) states that the distribution of sample means approximates a normal distribution (also known as a “bell curve”)

Normal Approximation

- The Law of Large Numbers and Central Limit Theorem together make a very important point
- If you have a very large enough sample from a population
  - Then the sample mean follows a normal distribution centered at the population mean
- This is why we use the Normal distribution so much

More Python libraries

- NumPy
  - Includes many useful functions for computing mathematical operations
  - Often required for data manipulation
  - Typically named np
    - import numpy as np
- SciPy
  - Has many useful features
  - We will focus on the stats module
    - from scipy import stats
  - This includes plenty of statistical tests and operations already implemented for us

Obtaining basic distribution stats

- When drawing a boxplot or other distribution visualization
  - Useful to know exact values defining this distribution
- In fact, Nature requires this for boxplots (as mentioned before)
  - Minimum
  - 25th percentile
  - Median (50th percentile)
  - 75th percentile
  - Maximum

Results are based on pairs of tissues that are represented by at least 10 individual donors. **a** Box plots depicting the Sørensen–Dice similarity indexes for Ig clonotype sequences shared across samples from the same individual (orange color), IGH (n = 1542, min = 0.0006, Q_1 = 0.0080, median = 0.0158, Q_3 = 0.0317, max = 0.2731), IGK (n = 30498, min = 0.0005, Q_1 = 0.0131, median = 0.0297, Q_3 = 0.0597, max = 0.4545), and IGL (n = 5124, min = 0.0006, Q_1 = 0.0142, median = 0.0299, Q_3 = 0.0548, max = 0.3871); shared across samples from different individuals (blue color), IGH (n = 2988, min = 0.0004, Q_1 = 0.0016, median = 0.0025, Q_3 = 0.0041 max = 0.1463), IGK (n = 10213560, min = 0.0002, Q_1 = 0.0091, median = 0.0208, Q_3 = 0.0424, max = 0.4615), and IGL (n = 1484735, min = 0.0004, Q_1 = 0.0090, median = 0.0188, Q_3 = 0.0336, max = 0.2500). Only samples with at least ten reported clonotype sequences were used to compute Sørensen–Dice similarity indexes. Each boxplot represents the median and interquartile range, with whiskers extending to 1.5 times the interquartile range. The p-values were generated using a two-

Obtaining basic distribution stats

- Pandas provides a function to do this

```
[10] categorical_data.describe()
```

	total_bill	tip	size
count	244.000000	244.000000	244.000000
mean	19.785943	2.998279	2.569672
std	8.902412	1.383638	0.951100
min	3.070000	1.000000	1.000000
25%	13.347500	2.000000	2.000000
50%	17.795000	2.900000	2.000000
75%	24.127500	3.562500	3.000000
max	50.810000	10.000000	6.000000

Obtaining basic distribution stats

- This can further be applied to single columns or grouped dataframes

```
[11] categorical_data['tip'].describe()
```

```
count    244.000000
mean       2.998279
std        1.383638
min        1.000000
25%        2.000000
50%        2.900000
75%        3.562500
max       10.000000
Name: tip, dtype: float64
```

```
[12] categorical_data.groupby('sex').describe()
```

	total_bill								tip								size							
	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
sex																								
Male	157.0	20.744076	9.246469	7.25	14.00	18.35	24.71	50.81	157.0	3.089618	1.489102	1.0	2.0	3.00	3.76	10.0	157.0	2.630573	0.955997	1.0	2.0	2.0	3.0	6.0
Female	87.0	18.056897	8.009209	3.07	12.75	16.40	21.52	44.30	87.0	2.833448	1.159495	1.0	2.0	2.75	3.50	6.5	87.0	2.459770	0.937644	1.0	2.0	2.0	3.0	6.0

Obtaining basic distribution stats

- Using the unstack function can make this vertical if it becomes difficult to read

```
[15] categorical_data.groupby('sex')['tip'].describe().unstack()
```

```
          sex
count     Male      157.000000
          Female     87.000000
mean      Male        3.089618
          Female      2.833448
std       Male        1.489102
          Female      1.159495
min       Male        1.000000
          Female      1.000000
25%       Male        2.000000
          Female      2.000000
50%       Male        3.000000
          Female      2.750000
75%       Male        3.760000
          Female      3.500000
max       Male       10.000000
          Female      6.500000
dtype: float64
```

Statistical Tests

MANN-WHITNEY U-TEST

STUDENT’S T-TEST

Hypothesis Tests Refresher

- Null Hypothesis
  - The hypothesis opposite to what we are investigating
    - I.e. no significant difference between two groups
- Alternate Hypothesis
  - The hypothesis we are investigating
    - I.e. a significant difference between two groups
- p-value
  - The probability of an event occurring given the null hypothesis
    - I.e. the probability of a difference between two groups due to random chance
- If the probability of an event occurring is small enough that we think it is not merely random chance (classically <0.05), we call it significant and rule the null hypothesis false

Mann-Whitney U Test

- Sometimes referred to as the Wilcoxon Rank Sum Test
- Non-parametric test of significance
  - Does not rely on the assumption of normality
    - Thus useful for smaller samples if we cannot satisfy Law of Large Numbers/Central Limit Theorem
  - Does not rely on the mean
    - Thus it's not really affected by large outliers
- Null Hypothesis
  - Samples come from equal populations
- Alternate Hypothesis
  - Samples come from populations that are not equal

Mann-Whitney U Test

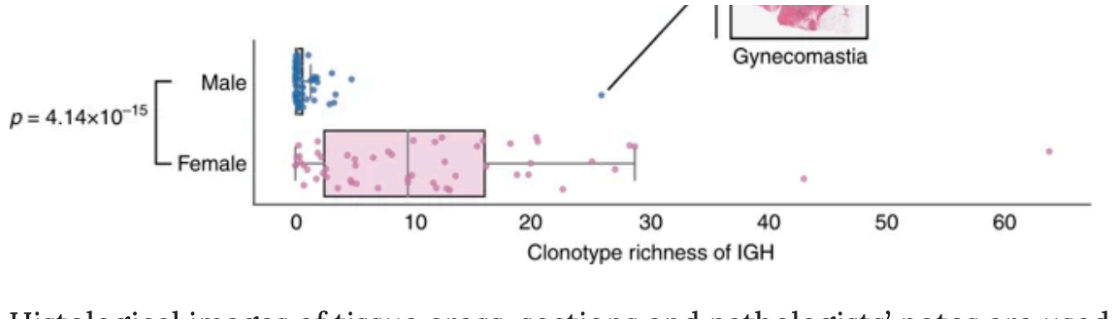
- Takes 2 arrays x and y
  - Here x = tips to males, and y = tips to females
- The alternative parameter defines the type of test we run
  - greater: test whether x > y
  - less: test whether x < y

```
[21] male_tips = categorical_data.loc[categorical_data['sex'] == 'Male', 'tip']
     female_tips = categorical_data.loc[categorical_data['sex'] == 'Female', 'tip']
     mwu_results = stats.mannwhitneyu(male_tips, female_tips, alternative='greater')
     print(mwu_results)

     MannwhitneyuResult(statistic=7289.5, pvalue=0.19167724679681963)
```

Mann-Whitney U Test

- How to report the test?
- Because the test is non-parametric, but still comparing distributions
  - Make sure to report the medians of both samples, at the least
- Additionally, you should report the statistic and p-value that are generated with the test
- Lastly, make sure you mention the alternative hypothesis you used



We detect a significant difference between the clonotypic richness of IGH in the breast tissue of males and females (two-sided Mann–Whitney *U*-test: $U = 376$, *p*-value = 4.14×10^{-15}).

https://www.nature.com/articles/s41467-020-16857-7

Mann-Whitney U Test

Further things to consider:

Recommended minimum sample size

- Each sample should have at least 20 data points
- Based on recommendation in SciPy documentation

Independence requirement

- The test relies on the assumption that the two samples are independent

Student's t-test

2-sample t-test

- Test comparing two independent samples

Is a parametric tests

- Thus it assumes normal distribution
  - For t-test, it is enough that it is approximately normal
- Will apply to the means
  - Thus they are affected by outliers
- Parametric tests are typically stricter in assumptions, but more powerful as a result
  - If you meet requirements, it is recommended to try and use them

Two-sample t-test

- Null Hypothesis
  - Sample means are equal
- Alternate Hypothesis
  - Sample means are not equal
- Using scipy.stats, this will always be a two-sided test
- Assumes that the two samples have identical variances
  - Standardizing each sample with respect to sample mean can make this true

Two-sample t-test

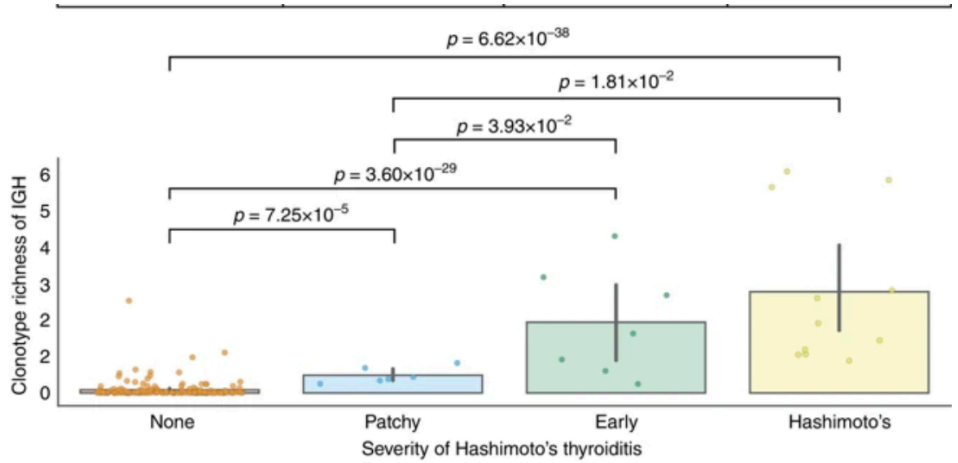
- Takes 2 arrays x and y
  - Here x = tips to males, and y = tips to females

```
[7] male_tips = categorical_data.loc[categorical_data['sex'] == 'Male', 'tip']
    female_tips = categorical_data.loc[categorical_data['sex'] == 'Female', 'tip']
    tt2_results = stats.ttest_ind(male_tips, female_tips)
    print(tt2_results)
```

```
Ttest_indResult(statistic=1.3878597054212687, pvalue=0.16645623503456763)
```

Student's t-test

- How to report the test?
- Because the test is parametric, and comparing distributions
  - At the least report the means of samples and populations
- Additionally, you should report the statistic and p-value that are generated with the test
- Lastly, make sure you mention that it is two-sided



per one million RNA-seq reads (CPM). The p-values were generated using a two-sided, two-sample t-test for each pair of severities (None vs. Patchy: $t = -4.06$ and p-value = 7.25×10^{-5}; None vs. Early: $t = -13.44$ and p-value = 3.60×10^{-29}; None vs. Hashimoto's: $t = -16.31$ and p-value = 6.62×10^{-38}; Patchy vs. Early: $t = -2.34$ and p-value = 3.93×10^{-2}; Patchy vs. Hashimoto's: $t = -2.65$ and p-value = 1.81×10^{-2}; Early vs. Hashimoto's: $t = -0.92$ and p-value = 0.37). Error

Correlation

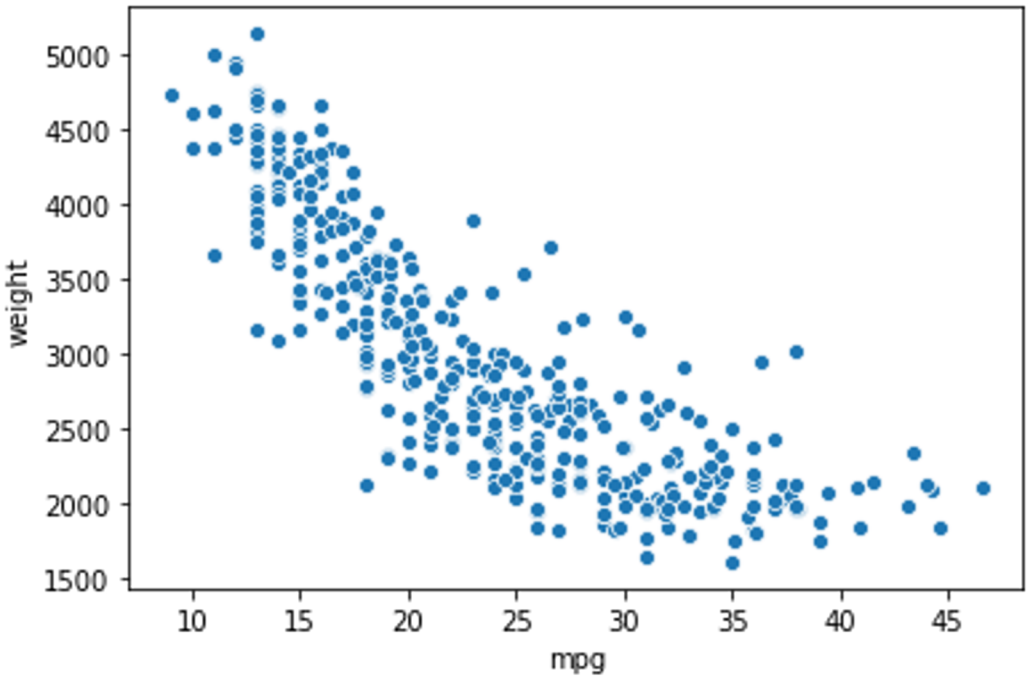
- Correlation measures the linear relationship between two datasets
- This is useful for measuring whether two variables appear to be related
  - Remember: Correlation does not equal causation
- Many ways to measure, but we will focus on Pearson Correlation Coefficient
  - Technically assumes normality
- +1 correlation coefficient means perfect linear relationship
  - As x increases y increases
- -1 correlation coefficient means perfect inverse linear relationship
  - As x increases y decreases
- 0 correlation coefficient means no relationship
  - As x increases y can do anything

Correlation

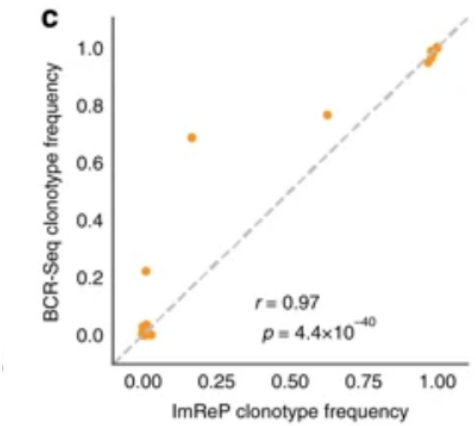
- The scipy pearson calculation also includes a two-tailed p-value
- Both arrays passed in should be the same length

```
[10] mpg = relational_data['mpg']
     weight = relational_data['weight']
     p_corr = stats.pearsonr(mpg, weight)
     print(p_corr)
```

```
(-0.831740933244335, 2.9727995640500577e-103)
```



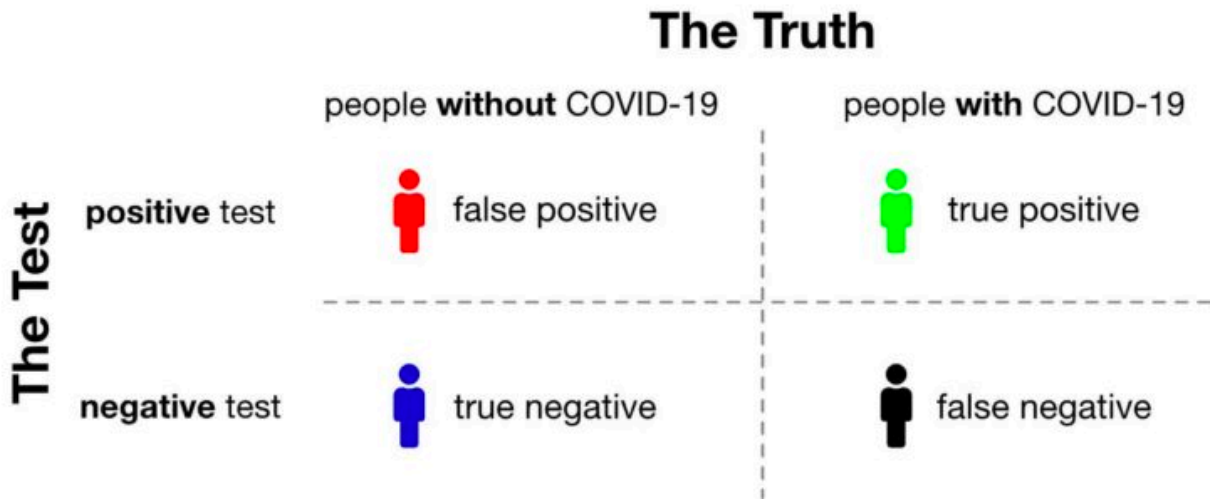
Visualize correlation in a manuscript



the BCR-seq-confirmed clonotypes abundances. **c** Pearson correlation of IGH clonotype frequencies estimated based on the BCR-Seq data (y-axis) and the RNA-seq data (x-axis) across all the samples for ImReP ($n = 63$, $r = 0.97$, p-value $= 4.4 \times 10^{-40}$). **d** Pearson correlation of IGH clonotype

Confusion Matrices

- Fancy term for the table indicating false positives, false negatives, true positives, and true negatives



Confusion Matrices

- Precision
  - Also referred to as positive predictive value (PPV)
  - A measure of well our selections represented correct predictions
- Recall
  - Also referred to as sensitivity
  - A measure of many correct predictions we made, given how many were possible
- F1-Score
  - The harmonic mean of precision and recall
  - Balances false positives and negatives

$$\frac{TP}{(TP + FP)}$$

$$\frac{TP}{(TP + FN)}$$

$$\frac{2 * (precision * recall)}{(precision + recall)}$$