Machine learning workshops for material scientists

Lecture 1: Statistics review and warm up

October 5, 2022

Sira Sriswasdi, PhD

- Research Affairs
- Center of Excellence in Computational Molecular Biology (CMB)
- Center for Artificial Intelligence in Medicine (CU-AIM)

Today's content

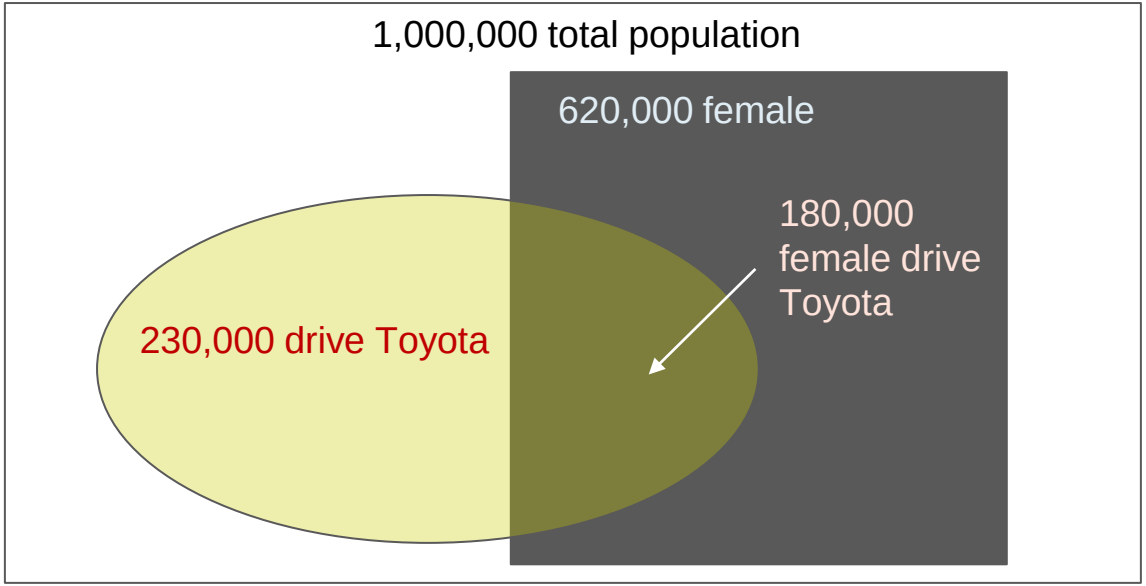
- Quick overview of the roles of probability and statistics
- Maximum likelihood principle
- P-value and test statistics
- Permutation test
- Correction for multiple testing

Probability is the basis of statistics

- P-value = probability of observing the same or more extreme result, given that the null hypothesis is true
  - Probability of seeing >2-fold up-regulation of gene A in drug treated patient by chance, given that the drug does not affect gene A

- Likelihood ratio = $\frac{\text{P(observed data | model 1)}}{\text{P(observed data | model 2)}}$
  - If LR is high, reject model 2. If LR is low, reject model 1
  - $\frac{\text{P(observed monkey pox genome diversity | mutation rate=0.3)}}{\text{P(observed monkey pox genome diversity | mutation rate=0.01)}}$

Probability tells us what to expect



- How likely or unlikely is this observation?
- P(data | no relation between female drivers and Toyota) $= \frac{\binom{620,000}{180,000}\binom{380,000}{50,000}}{\binom{1,000,000}{230,000}}$

Fisher's Exact Test

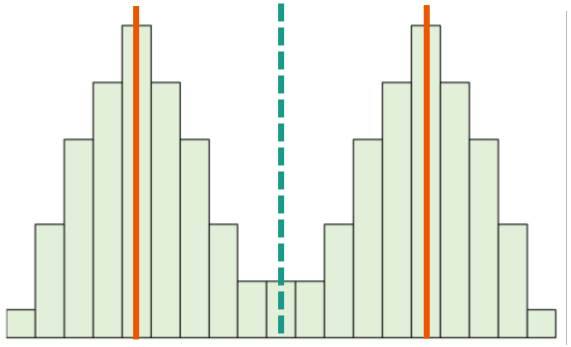
Group	Female	Male	Total
Drive Toyota	$k \geq 180{,}000$	$230{,}000 - k$	230,000
Don't drive Toyota	$620{,}000 - k$	$150{,}000 + k$	770,000
Total	620,000	380,000	1,000,000

- P-value for this observation = P(Female & Toyota ≥ 180,000)
  - Summation of Hypergeometric probability for $k \geq 180{,}000$

Probability helps us suggest hypothesis

Group I: 50 samples
Mean = 10, Variance = 2

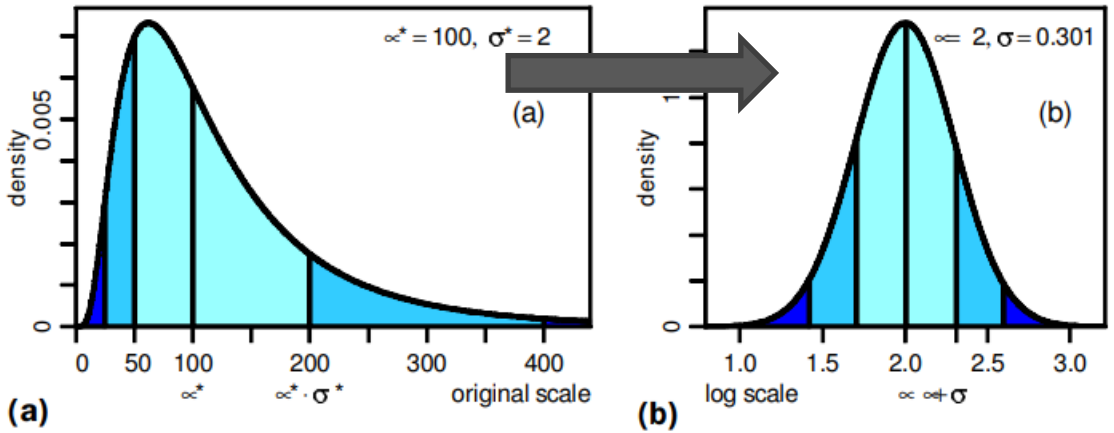
Group II: 50 samples
Mean = 26, Variance = 5



Combined: Mean = 18, Variance = 137

- Which assumption better explain the observed data?
- P(data | bimodal) versus P(data | unimodal)

Log-normal distribution



Limpert, Stahel, and Abbt. BioScience 2001.

- Some data are not normally distributed but their log-values are
  - Intensity and count data
- Visualize histogram to check

Statistics helps us learn from the data

- 1,000 untreated pancreatic cancer patients survived for 1, 5, 3, …, 5 years. What is the estimated yearly survival rate?
  - Estimate the value of model parameters

- 1,000 pancreatic cancer patients treated with drug X survived for 6, 9, 10, …, 8 years. Does the drug significantly increase the survival rate?
  - Assess how data support a certain conclusion

Maximum likelihood

Maximum likelihood for parameter estimation

- Likelihood = P(data | model)
- Find model that maximize the likelihood
  - Why?

- True objective = find model that maximize P(model | data)
- Bayes' Rule:
  - P(model | data) = P(data | model) x P(model) / P(data)
  - P(data) is a constant

  - If all models are equally likely to be true, P(model | data) $\propto$ P(data | model)

Maximum likelihood is all around

- 10 tosses of a coin yielded T,T,T,H,T,H,T,T,T,H. How much do you think this coin is weighted on the tail side?
  - Set P(Tail) = p
  - Likelihood = P(getting 7 tails out of 10 tosses | p) = $\binom{10}{7}p^7(1-p)^3$
  - Which p maximize the likelihood?
    - Solve the equation $\frac{d\text{Likelihood}}{dp} = 0$
    - $0 = -3p^7(1-p)^2 + 7p^6(1-p)^3$
    - $0 = p^6(1-p)^2[-3p + 7(1-p)]$
    - $7 = -10p$
    - $P = 0.7$

Maximum likelihood is all around

- 5 pancreatic cancer patients passed away after 1, 5, 3, 4, and 5 years, respectively. What is the maximum likelihood estimate for the yearly survival rate?
  - Set the yearly survival rate = r
  - P(survive exactly k years) = $r^k(1 - r)$
  - Likelihood = $r^1(1 - r)\, r^5(1 - r)\, r^3(1 - r)\, r^4(1 - r)\, r^5(1 - r) = r^{18}(1 - r)^5$
  - Which r maximize the likelihood?
    - Solve the equation $\frac{dLikelihood}{dr} = 0$ → $r_{MLE} = 18/23$

Likelihood ratio test for selecting models

- Likelihood ratio = $\frac{\mathrm{P}(\text{data} \mid \text{model 1})}{\mathrm{P}(\text{data} \mid \text{model 2})}$
  - If LR is high, reject model 2. If LR is low, reject model 1

- Example:
  - Test for impact of treatment: $\frac{\mathrm{P}(\text{survival} \mid \text{treated and untreated differ})}{\mathrm{P}(\text{survival} \mid \text{all patients are the same})}$

- This is theoretically the most powerful test (Neyman-Pearson Lemma)

Balancing between complexity and likelihood

- **Simple model**: Omicron and Delta have the same spreading rate
  - One parameter
- **Complex model**: Omicron and Delta have different spreading rates
  - Two parameters

- Complex model always achieve higher likelihood

- But is the additional complexity worth it?
  - Akaike information criterion (AIC): 2 x # parameters – log likelihood
  - Bayesian information criterion (BIC): log(# samples) x # parameters – log likelihood

P-value

The rise of P-value

- Probability of observing the same or more extreme result, given that the null hypothesis is true

- Using the null hypothesis as reference point
  - Is age an important indicator of cancer risk?
  - Build several logistic regression models with age as an input
  - Test whether the coefficients of age is zero
    - Null: Coefficients of age is normally distributed with mean $= 0$
    - Alternative: Coefficients of age is normally distributed with mean $\neq 0$

- Working with null hypothesis is convenient and practical

P-value example 1

- Before BA.5, a study estimated the rate of daily increase in COVID-19 infections with a normal distribution $N(1.3, 0.01)$

- After BA.5, data show that the rate of daily increase in COVID-19 infections is 1.5

- P-value = P(daily rate ≥1.5 | BA.5 has the same spread rate as prior strains)
  = P(getting value ≥1.5 from $N(1.3, 0.01)$)
  = P-value of Z-score of 2 = 0.02275

- Reject null hypothesis

P-value example 2

- Before BA.5, a study estimated the rate of daily increase in COVID-19 infections with a normal distribution $N(1.3, 0.01)$

- After BA.5, data show that the rate of daily increase in COVID-19 infections is 1.4

- P-value = P(daily rate ≥1.4 | BA.5 has the same spread rate as prior strains)
  = P(getting value ≥1.4 from $N(1.3, 0.01)$)
  = P-value of Z-score of 1 = 0.158655

- Do we accept null hypothesis?

The rise of P-value

- Probability of observing the same or more extreme result, given that the null hypothesis is true

- We use ranking scores called test statistics

Test statistics

Test statistics

- A score for ranking observations

- In one-sample t-test of whether the mean $\bar{x}$ of data $\{x_1, x_2, \ldots, x_n\}$ is equal to β, the test statistics is $t = \frac{\bar{x} - \beta}{\mathrm{SE}}$
  - Measure how close is $\bar{x}$ to β, subject to the variability of the data
  - Correspond to the null hypothesis that the mean of the data is β
  - The more t deviates from zero, the more extreme the result

- P-value = $P(t \geq t_{\text{observed}} \mid \text{the data is normally distributed with mean } \beta)$
  - t follows $N(0, 1)$ by Central Limit Theorem

Ranking score behind popular tests

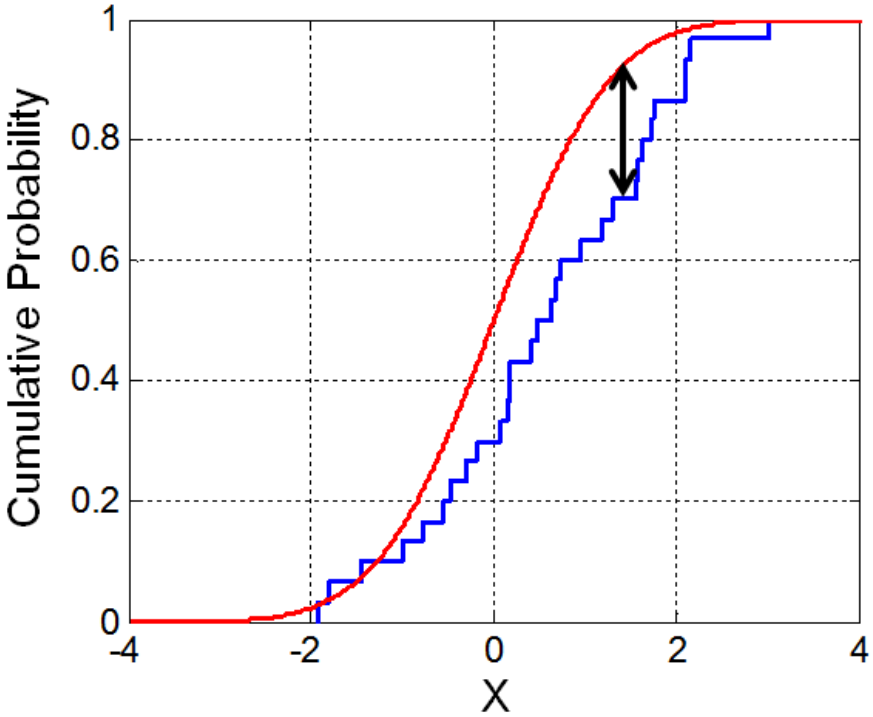
- Mann-Whitney U test: $$U = \sum_{i=1}^{n} \sum_{j=1}^{m} S(X_i, Y_j), \quad S(X,Y) = \begin{cases} 1, & \text{if } X > Y, \\ \frac{1}{2}, & \text{if } Y = X, \\ 0, & \text{if } X < Y. \end{cases}$$

- Wilcoxon rank-sum test:

1. Compute $|X_1|, \ldots, |X_n|$.
2. Sort $|X_1|, \ldots, |X_n|$, and use this sorted list to assign ranks $R_1, \ldots, R_n$

$$T = \sum_{i=1}^{N} \operatorname{sgn}(X_i) R_i.$$

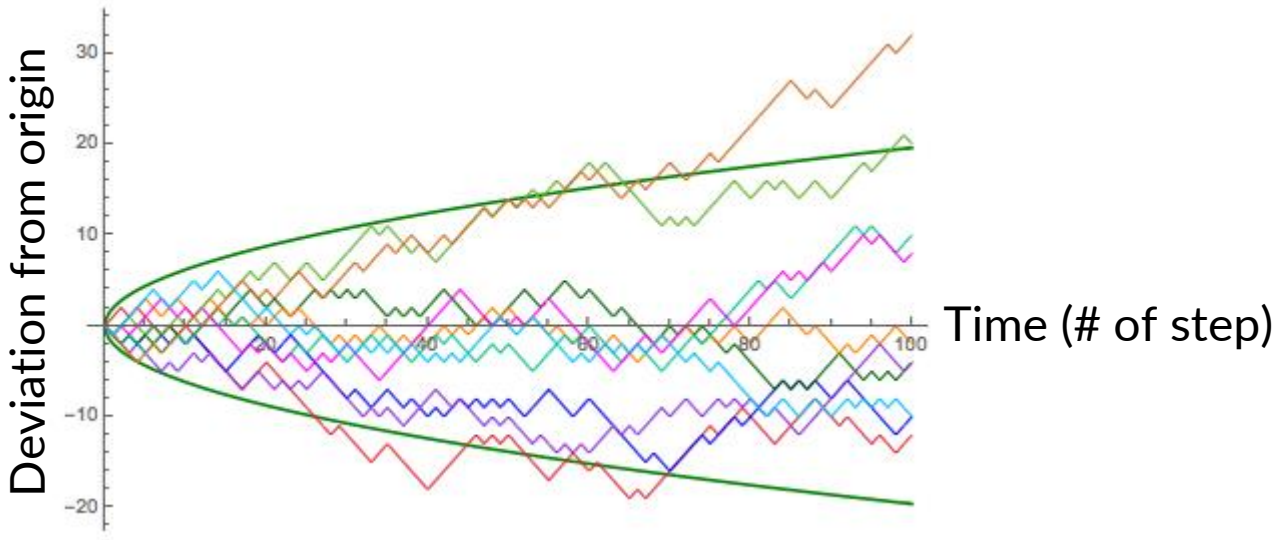
Kolmogorov-Smirnov test



https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test

- Test whether two probability distribution are equal
- Compare cumulative density (red and blue trends)
- If they are equal, the two curves should stay close to each other
- Test statistics = maximal deviation

Maximal deviation of random walks



https://demonstrations.wolfram.com/SimulatingTheSimpleRandomWalk/

- P(maximal deviation > d) $\approx 2\sum_{k=1}^{\infty}(-1)^{k-1}\, e^{-2(kd)^2}$

Chi-squared test

Group	Asian	Black	White
Head & neck cancer	200	120	70
Healthy	110	150	130

- Test whether the two categorical distributions are the same
- Test statistics $\sum_i \frac{(O_i - E_i)^2}{{E_i}^2}$ follows Chi-squared distribution
  - Larger difference between observed and expected → more extreme
  - Large difference on small group → more extreme

Choosing the right test

AUROC	Split 1	Split 2	Split 3	Split 4	Split 5	Split 6	Split 7
M1	0.701	0.503	0.991	0.827	0.623	0.728	0.596
M2	0.691	0.478	0.905	0.739	0.589	0.719	0.508

- Comparing AUROCs between models M1 and M2
  - Unpaired Student's t-test p-value = 0.5687
  - Mann-Whitney U test p-value = 0.6101
  - Paired Student's t-test p-value = 0.0137
  - Wilcoxon signed rank test p-value = 0.0156

Hypothesis testing framework

- Propose null hypothesis

- Design the test statistic t
  - This score should reflect the extreme aspect of the data
- Derive the distribution of test statistic under the null hypothesis

- Specify the significance level α to reject null hypothesis (e.g., 0.05)
- Calculate p-value: $P(t \geq t_{\text{observed}} \mid \text{null hypothesis})$

- By following this framework, new tests can be created!

Permutation test

Correlation example

	Obs 1	Obs 2	Obs 3	Obs 4	Obs 5	Obs 6	Obs 7
Feature 1	0.701	0.503	0.991	0.827	0.623	0.728	0.596
Feature 2	0.691	0.478	0.905	0.739	0.589	0.719	0.508

- Correlation of between Feature 1 and Feature 2 = 0.9746
- Can we assess the significance of this correlation?

- **Null hypothesis**
  - Feature 1 and Feature 2 are uncorrelated
  - Feature 2 values can be shuffled and still lead to the same correlation

Permutation test

- **Alternative hypothesis**: The observed property of the data, such as high correlation, is due to some structure in the data

- **Null hypothesis**: That structure in the data does not contribute to the property of interest

- **P-value** = Probability that shuffled data still yield the same or more extreme property than the original data

- Shuffle data in such a way that the structure of interest is disrupted
- Re-calculate the property of interest and compared to the original value

Permutation test for correlation

	Obs 1	Obs 2	Obs 3	Obs 4	Obs 5	Obs 6	Obs 7
Feature 1	0.701	0.503	0.991	0.827	0.623	0.728	0.596
Feature 2	0.691	0.478	0.905	0.739	0.589	0.719	0.508

Correlation = 0.97

1,000 times

Correlation = 0.24

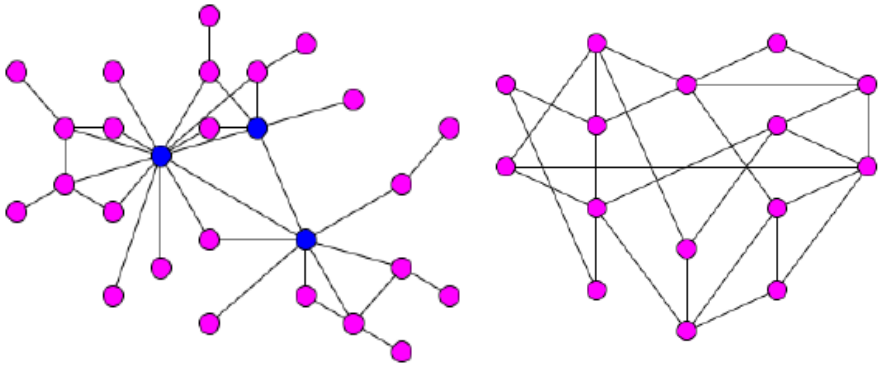
Correlation = -0.30

Feature 2	0.719	0.691	0.739	0.589	0.508	0.905	0.478

Correlation = 0.32

Permutation test for network data

Real-world network has hubs that serve as shortcut between other nodes



Random network is not well-connected

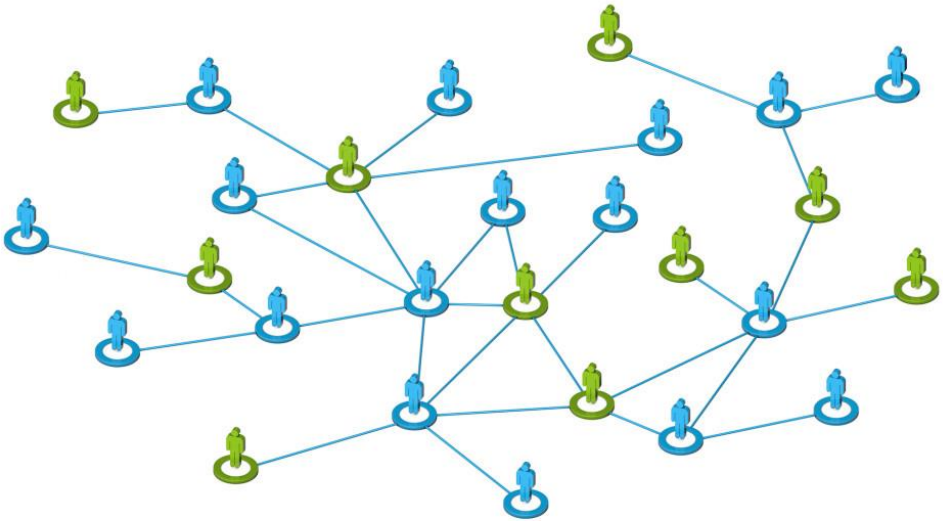
Source: Segura-Cebrera *et al*. Analysis of Protein Interaction Networks to Prioritize Drug Targets of Neglected-Diseases Pathogens

- **Null hypothesis**: The small diameter of real-world network appear by chance in any network with the same number of nodes and edges

- **Permutation test**: Generate 1,000 random networks with the same number of nodes and edges and compute the diameter of these networks

Permutation test for network data

Same-gender Facebook friendships occur more often than cross-gender ones



- **Null hypothesis**: The high number of same-gender relationship of Facebook network occur by chance in any random network with the same number of people and relationship

Correction for multiple testings

Correction for multiple testings

- P-value cutoff of 0.05 means that under the null hypothesis, there is only 5% chance of observing the same or more extreme result

- Applying the same test 1,000 times will result in 50 tests on average with smaller p-value than 0.05 just by chance
  - Differential expression analysis tests thousands of gene at once

- This in unacceptable if a conclusion relies on multiple tests
  - Functional enrichment analysis assumes that all input DEGs are true

Bonferroni method

- Divide the p-value cutoff by the number of test
- Adjusted p-value cutoff = 0.05 / 1000 = 0.00005

- Applying the same test 1,000 times will result in 0.05 tests on average with smaller p-value than 0.00005 just by chance

- Easy but lose power

False discovery rate (FDR)

- P-value operates under the null hypothesis
- But in practice, we want to control the number of errors in the output
  - The number of DEGs that were incorrectly proposed
- FDR = Probability of getting a false positive
  = # false positive / # all predicted positives
- But FDR involves alternative hypothesis, which is difficult to calculate
- We can control FDR somewhat through p-value!

Benjamini-Hochberg procedure

- Valid under broad situations (independent tests, some dependency, etc.)
- Control false discovery rate (FDR)

- Target FDR = 0.05
- Given a series of tests with p-values, p_1, p_2, ..., p_n
  - Sort p-values from low to high, p'_1, p'_2, ..., p'_n
  - Find largest k such that $p'_k \leq 0.05 \times k / n$
    - For the smallest p-value, this is equivalent to Bonferroni
    - For other p-values, this technique gradually loosens the cutoff
  - Reject null hypothesis for tests corresponding to p'_1, p'_2, ..., p'_k

Impact of correction method

P-value	Bonferroni	B-H	B-Y
Smallest	0.0005	0.0005	0.0005
2nd smallest	0.0005	0.001	0.000667
3rd smallest	0.0005	0.0015	0.000818
4th smallest	0.0005	0.002	0.00096
5th smallest	0.0005	0.0025	0.001095

- **Assumptions**: There are 100 tests. Target *p*-value or FDR cutoff = 0.05
- More powerful tests can yield more significant results

Summary

- Probability and statistics are intertwined
- **Yet another important ML**: Maximum likelihood principle
- P-value and test statistics
- **Weapon of mass statistical testing**: Permutation test
- Correction for multiple testing

Any question?