3050571 Practical Clin Data Sci

Session 4: Statistics revisited

February 6, 2024

Sira Sriswasdi, PhD

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

Two main goals of statistical analysis

- **Inference**
  - Estimate a parameter from the data
  - Ex: Yearly survival probability of a stage III lung cancer patient
  - Ex: Effect size of “age” on fall risk
  - Confidence interval

- **Hypothesis testing**
  - Is radiotherapy dose associated with normal tissue complication?
  - Is drug A better than drug B at suppressing symptom?

Inference

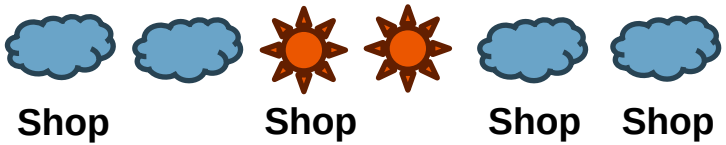
How do we infer a value of a parameter?

- A group of patients developed AD at **65, 82, 71, and 59 years old**. What is your estimated age of AD onset?

- A group of stage III pancreatic cancer patients survived for **1, 5, 3, 4, and 5 years**. What is your estimated yearly survival rate?

- Gene A has two alleles, *A* and *a*. A study of 1,000 Thai individuals found **700 with genotype *AA*, 200 with genotype *Aa*, and 100 with genotype *aa***. What is the estimated allele frequency of *a*?

Let's review probability



- P(A) = probability that A occurs

Joint probability

- P(A, B) = probability that A and B occurs

Conditional probability

- P(A | B) = probability that A occurs given that B already occurred
- P(A | B) = P(A, B) / P(B)
- P(A, B) = P(A | B) P(B) = P(B | A) P(A)

- Let's calculate P(Shopping | Sunny) and P(Sunny | Shopping)

Which expression do you prefer?

- **Objective**: Find the model that best fit to the data

- **Option 1**: Find the model that maximize P(model | data)
  - P(yearly survival = *p* | patients survived for 1, 5, 3, 4, and 5 years)
  - P(allele frequency of *a* = *f* | observed 700, 200, and 100 patients)

- **Option 2**: Find the model that maximize P(data | model)
  - P(patients survived for 1, 5, 3, 4, and 5 years | yearly survival = *p*)
  - P(observed 700, 200, and 100 patients | allele frequency of *a* = *f*)

Bayes' rule links the two options

$$P(\text{model} \mid \text{data}) = P(\text{data} \mid \text{model}) \times P(\text{model}) / P(\text{data})$$

- We want the model that maximize P(model | data)
  - Data is already collected. But the model is yet to be determined.
  - But, how to calculate?

- **Using Bayes' rule**:
  - We can calculate P(data | model)
  - P(data) is just a constant
  - Define prior belief, P(model), from domain knowledge
    - If all model are equally likely, P(data | model) = P(model | data)

Maximum likelihood principle

- Likelihood = P(data | model) ← Find the model that maximize this

- Gene A has two alleles, *A* and *a*. A study of 1,000 Thai individuals found 700 with genotype *AA*, 200 with genotype *Aa*, and 100 with genotype *aa*. What is the estimated allele frequency of *a*?
  - Set the allele frequencies $f_A = p$ and $f_a = 1 - p$
  - $P(AA) = p^2$, $P(Aa) = 2p(1 - p)$, and $P(aa) = (1 - p)^2$
  - $P(\text{data} \mid p) = P(AA)^{700}\, P(Aa)^{200}\, P(aa)^{100} = p^{1400} 2^{200} p^{200} (1 - p)^{200} (1 - p)^{200}$
    $= 2^{200} p^{1600} (1 - p)^{400}$
  - Which p maximize the likelihood?
    - Solve the equation $\frac{dLikelihood}{dp} = 0$ → $p_{\text{MLE}} = 0.8$

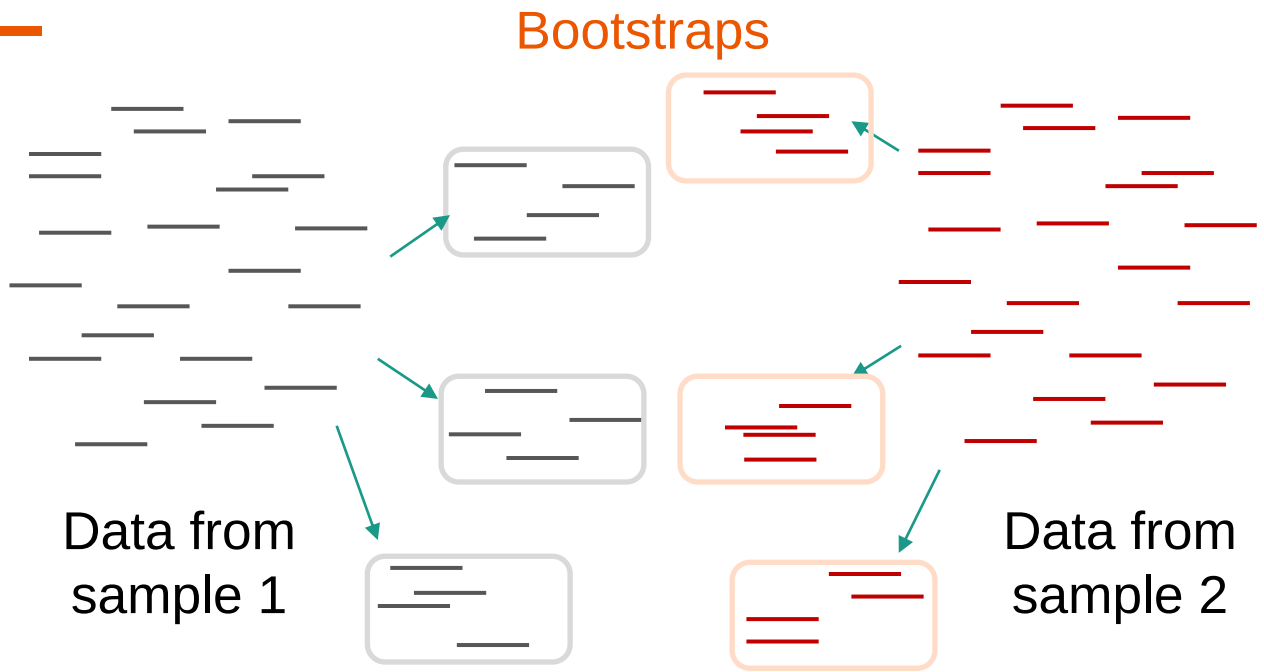
Optimization Problem!

Another maximum likelihood example

- In a study of 5 pancreatic cancer patients, they passed away after 1, 5, 3, 4, and 5 years, respectively. What is the estimated yearly survival rate?
  - Let's set the yearly survival rate = r
  - P(a patient survives exactly k years) = $r^k(1 - r)$
  - P(data | r) = $r^1(1 - r)\, r^5(1 - r)\, r^3(1 - r)\, r^4(1 - r)\, r^5(1 - r) = r^{18}(1 - r)^5$
  - Which r maximize the likelihood?
    - Solve the equation $\frac{dLikelihood}{dr} = 0$ → r_{MLE} = 18/23

Bootstrapping

Bootstrapping



- Instead of using the whole data to perform one calculation
- Sampling from data to perform multiple calculations → Estimate variance

Bootstrapping

- Out of 1,000 treated patients, 731 fully recovered
- **Rate of recovery = 73.1%**

- How do you estimate the uncertainty of this result?
  - Sample 300 patients at random 20 times
  - Each time, calculate the rate of recovery
  - Calculate the variance of these 20 estimated rate of recovery

- Bootstrapping = synthetic data generation

Hypothesis testing

Testing for nested models

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

- The more complex model always achieve higher likelihood

- But is the additional complexity worth it?
  - $\frac{\mathrm{P(data \mid complex\ model)}}{\mathrm{P(data \mid simple\ model)}}$ must be >>1 to reject the simpler model
  - Akaike information criterion (AIC): 2 x # parameters – log likelihood

Likelihood ratio test

- 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

- Examples:
  - Test for viral spreading rate: $\frac{\mathrm{P}(\text{COVID}-19 \text{ infection} \mid \text{spreading rate}=4.0)}{\mathrm{P}(\text{COVID}-19 \text{ infection} \mid \text{spreading rate}=1.5)}$
  - Test for impact of treatment: $\frac{\mathrm{P}(\text{recovery rate} \mid \text{drug has some effect})}{\mathrm{P}(\text{recovery rate} \mid \text{drug has no effect})}$

- We do not necessarily know the right hypothesis!

Alternative vs null hypothesis

Alternative Hypothesis

This is the knowledge that we want to prove via the data.

But we may not be able to express it quantitatively.

Ex: Omicron spreads **faster** than other variants.

Null Hypothesis

This is the “default” assumption if nothing special is going on.

Ex: Omicron spreads **at the same rate** as other variants.

The likelihood can be calculated

Likelihood under the null hypothesis

- **Case 1**: Omicron spreads at the same rate as other variants
  - Estimate spread rate **before** Omicron
  - Calculate P(case counts **after** Omicron | spread rate)

- **Case 2**: Drug A has no effect on patient recovery
  - Estimate recovery rate from control group
  - Calculate P(patient **in treatment group** | recovery rate)

- What is your interpretation if the likelihood is **low**? How about **high**?

P-value

From likelihood to P-value

- Likelihood = Probability of observing a data

- P-value = Probability of observing the same or more extreme data, given that the **null hypothesis** is true
  - **Low p-value** lets us reject the **null hypothesis**
  - **High p-value** means we cannot reject the **null hypothesis**

- How to quantify whether a data is the same or more extreme?
  - If drug A has no effect, increased drug A usage shouldn't increase the fraction of recovered patients
  - The fraction of recovered patients is a score!

Test statistics rank the extremeness of the data

- Fraction of recovered patients
- Percentage of additional daily COVID-19 case count
- 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 $t = \frac{\bar{x} - \beta}{\frac{SD}{\sqrt{n}}}$
- P-value = P(score ≥ $\text{score}_{\text{observed}}$ | null hypothesis)

Test statistics behind well-known 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 $R_1, \ldots, R_n$

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

- **Sign test**: Each observation is equally likely to be positive or negative
  - Probability of *k* positive values out of *N* observations = Binomial(*N*, *k*, *p* = 0.5)

Chi-squared test

Genotype	*A/A*	*A/a*	*a/a*
Expected frequency	200	120	70
Observed frequency	90	210	80

- $\sum_i \frac{(O_i - E_i)^2}{{E_i}^2}$ follows Chi-squared distribution

The ingredients of a statistical test

One-sample *t*-test example

- Define null hypothesis: data are normally distributed with mean = β
- Design the test statistic $t = \frac{\bar{x} - \beta}{\frac{SD}{\sqrt{n}}}$
- 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($|t| \geq |t_{observed}|$ under the null hypothesis) based on the derived distribution
- By following this framework, new tests can be created!

Mann-Whitney U test example

- Define null hypothesis: two samples have the same mean
- Design the test statistic: $U = \sum_{i=1}^{n} \sum_{j=1}^{m} S(X_i, Y_j), \quad S(X,Y) = \begin{cases} 1, & \text{if } X > Y, \\ \frac{1}{2}, & \text{if } Y = X, \\ 0, & \text{if } X < Y. \end{cases}$
- 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($U \geq U_{\text{observed}}$ | the null hypothesis)

Significance of a correlation

Patient	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7
P1	0.701	0.503	0.991	0.827	0.623	0.728	0.596
P2	0.691	0.478	0.905	0.739	0.589	0.719	0.508

- Correlation of test results between the two patients = 0.9746
- Can we say anything about the significance of this correlation?
- **Null hypothesis**
  - Test results are uncorrelated across patient
  - P2 data can be shuffled and still give the same correlation

Permutation test

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

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

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

- Shuffle data in such a way that the structure of interest is disrupted
- Calculate the property of interest and compared to the original score

Permutation test

Patient	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7
P1	0.701	0.503	0.991	0.827	0.623	0.728	0.596
P2	0.691	0.478	0.905	0.739	0.589	0.719	0.508

Correlation = 0.97

1,000 times

Correlation = 0.24

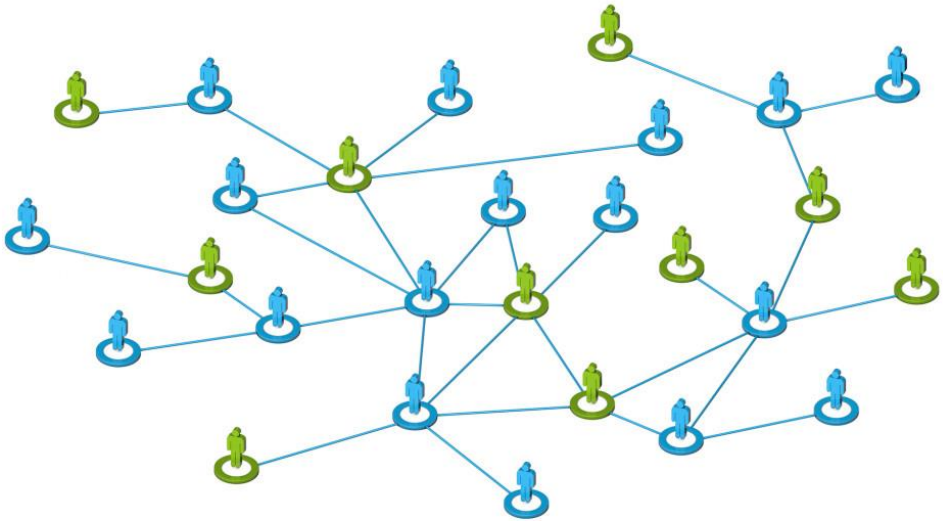
Correlation = -0.30

P2	0.719	0.691	0.739	0.589	0.508	0.905	0.478

Correlation = 0.32

Permutation test for network data

Same-gender FB friendship occur more easily than different-gender ones



- **Null hypothesis**: The high number of same-gender edge of Facebook friendship network can be achieved by chance in any random network with the same number of male/female nodes and the same number of edges

Any questions?

See you on February 8th