Statistical Testing: BMED 6517

Saurabh Sinha

saurabh.sinha@bme.gatech.edu
http://sinhalab.net/

BME & ISyE
Georgia Tech

January 4, 2023

Statistical Testing

- Toss a coin 50 times, get heads 50 times. I ask you "Is the coin a fair coin?"
- You do not want to believe the "hypothesis" that "the coin is unbiased". i.e., you reject this hypothesis.
- Same experiment, but 40 heads and 10 tails. Do you still want to reject the hypothesis?
- Maybe? Maybe not? Need a systematic procedure.
- That's what a statistical test does.

Statistical Testing: a not-so-good solution

- Step 1: State the null hypothesis. "The coin is unbiased". This allows probability calculations. For instance, coin tosses are Bernoulli trials with prob. of heads $p = 0.5$.
- Step 2: Decide upon the "test statistic", a random variable whose value depends on the data. For instance, "X = number of heads in 50 coin tosses".
- Step 3. Calculate the probability of the observed value of test statistic (X). For instance, $P(X = 40) = 0.00001$. (Use Binomial distribution.)
- Step 4. If the probability from Step 3 is "small" (say less than 5%), then reject the null hypothesis – the data do not seem likely under that hypothesis, so it is likely to be wrong.

Where's the problem?

- Say you saw 30 heads out of 50 coin tosses. What does your intuition say? Is it very unlikely?
- No. You probably are not surprised to see 30 heads in 50 coin tosses, when using a regular (fair) coin.
- Yet, Step 3 will calculate the probability $P(X = 30) = 0.042$ and in Step 4 you will reject the null hypothesis.
- This problem gets exacerbated as you have more data.
- Say you saw 250 heads out of 500 coin tosses. Surely this is not unlikely under the fair coin hypothesis? It's the most likely outcome after all! Yet, Step 3 calculates $P(X = 250) = 0.036$ and you will reject the null hypothesis!
- Our test is not so good!

Statistical Testing: the common solution

- Step 1: State the null hypothesis. "The coin is unbiased". This allows probability calculations. For instance, coin tosses are Bernoulli trials with prob. of heads $p = 0.5$.

- Step 2: Decide upon the "test statistic", a random variable whose value depends on the data. For instance, "X = number of heads in 50 coin tosses".

- Step 3. Calculate the probability of the observed value of test statistic (X) *being equal to or more "extreme" than the observed value*. For instance, $P(X \geq 40) = 0.00001$. (Use Binomial distribution: $P(X \geq 40) = \sum_{k=40}^{50} B(50, 0.5, k)$.

- Step 4. If the probability from Step 3 is "small" (say less than 5%), then reject the null hypothesis – the data do not seem likely under that hypothesis, so it is likely to be wrong.

- Note: The "5%" or "0.05" threshold defining "small" in Step 4 is called the "significance level" of the test.

- Note: The probability calculated in Step 3 is called the "p-value".

Discussion points

- Let's see if the "problem" got fixed. First, 30 heads out of 50 tosses gives us $P(X \geq 30) = 0.10$, which is not that small. Second, 250 heads out of 500 coin tosses gives us $P(X \geq 250) = 0.52$, clearly not a small number.
- Another point: what does "more extreme" mean, in the language of Step 3?
- When you see 40 heads out of 50 tosses, you're really testing if it's "too many", so "as extreme" means $X \geq 40$.
- You might also be testing 40 out of 50 is "too far from what you expected (25)", in which case you should calculate $P(X \geq 40) + P(X \leq 10)$, since both $X = 40$ and $X = 10$ are as far from your expectation and thus "as extreme".
- In the former scenario, we say "Null hypothesis: $p = 0.5$, Alternative hypothesis: $p > 0.5$. A "one-sided test".
- In the latter scenario, we say "Null hypothesis: $p = 0.5$, Alternative hypothesis: $p \neq 0.5$. A "two-sided test".

Null distribution of p-value

- Let's say that the null hypothesis (unbiased coin) is true. Do 50 coin tosses, count heads. This count is a random variable, call it X.
- Calculate the p-value of X, e.g., with a one-sided test. Denote the p-value by p. Note that p is determined by X, and is thus a random variable itself.
- What is the probability distribution of p?
- $\Pr(p \leq \pi) = \pi$.
- In other words, the p-value follows a uniform distribution between 0 and 1.

Parametric tests

- Parametric test: null hypothesis involves the *value of parameter(s)* of probability distribution(s). Test above was parametric. We knew that the distribution is Binomial, the null hypothesis was a statement about the parameter p of the distribution.
- Another example: consider two groups of hypertension patients, each of size K. The first group is given a medication, while the second was given placebo. Measure blood pressure in each individual.
- Assume that blood pressure in both groups is a normally distributed variable (X_1 and X_2). Null hypothesis: X_1 and X_2 have the same mean and variance. Alternative hypothesis: Means are not equal.

Parametric tests (cont'd)

- Calculate averages $\overline{X_1}$ and $\overline{X_2}$ respectively and the standard deviations s_1 and s_2 respectively for both groups.
- Calculate the "statistic"

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{1}{K}(s_1^2 + s_2^2)}}$$

- The statistic t follows a t-distribution with $2(K-1)$ degrees of freedom. This allows determining p-value of an observed value of t.
- This is the "T-test".

Non-parametric tests

- In the above, we assumed normal distribution of the "blood pressure" variable. In many cases, we cannot assume such a distributional form about the observations. Therefore, the null hypothesis cannot be of the form "assume the distribution is *bla* and that the parameters are equal/not equal".
- Example: In the previous example, suppose we had reason to believe that the distribution of blood pressure is not Normal, and is in fact unknown. We still have the same data, and we want to compare the two groups to see if the medication made a difference.

- Sort all $2K$ observations in increasing order: $D_1, D_2, \ldots D_{2K}$, and assign them ranks $r(D_i)$. Then calculate the "rank-sum" statistic:

$$T^+ = \sum_{i \in I^+} r(D_i)$$

, where I^+ is the set of indices of the first group of individuals.

- The null hypothesis is that the medicine makes no difference to the blood pressure. The distribution of T^+ can be computed under the null hypothesis, and we can prove that this distribution *does not depend* on the distribution of the blood pressure measurements.

- This is the Wilcoxon rank-sum test.

Enrichment tests

- A study is looking a "population" of N genes
- A subset of n genes have been identified as being turned on in cancer
- Suspiciously many of these n cancer genes are known to be involved in "cell division"
- Can we demonstrate a connection?

- Collect the set of all genes involved in cell division, say this is of size m

- Find k genes to be in the intersection of the cancer set and the cell division set

- Is this (k) significantly large, given N, m, n?

- The Hypergeometric test:
- Let us keep the n cancer genes to be a fixed set.
- If we had picked a random sample of m genes, how likely is an intersection equal to k?
- $f(k; N, n, m) = \frac{\binom{n}{k}\binom{N-n}{m-k}}{\binom{N}{m}}$
- How likely is an intersection equal to or greater than k?
- $P = \sum_{j \geq k} f(j; N, n, m)$

Enrichment tests

- If P calculated this way is below some threshold α, e.g., 0.05, we say that the association between the cancer set and the cell division set is statistically significant.
- In other words, we have just discovered a link between cancer and cell division, which is probably worthy of further investigation

An unusual observation?

- You are working with ten DNA sequences, each of length 1000 bp. These are the "promoter" regions of ten related genes
- You were searching for the words "TCACGT" and "AATTGA", both of which would implicate a specific transcription factor in the regulation of these genes
- You found that each of the ten sequences has at least one match to each of these two words.
- Could this be statistically significant, and perhaps suggestive of a specific biochemical mechanism?

Statistical testing through permutations

- Let θ be a boolean random variable that is true iff a set of random DNA sequences has the property you observed, i.e., each of them has at least one occurrence of "TCACGT" and at least one occurrence of "AATTGA".
- Generate ten random sequences to "mimic" the input sequences. Each of length 1000 bp. Base composition similar to original sequences
- Sample one nucleotide at a time, from the appropriate probability distribution on (A,C,G,T). Sampling a nucleotide based on "pseudo-random number" generators
- After generating one complete sample (i.e., ten sequences), check if the observation θ is true or not
- Repeat the entire procedure (simulate and check) 1000000 times, and count how frequently θ was true.
- If this count (rather, the corresponding probability) is very low, say ≤ 0.05, you have discovered a significant pattern in the sequence data

Permutation tests

- In the previous example, instead of generating sequences randomly, you could have "shuffled" or "permuted" each sequence.
- For each of the ten given sequences, do a random permutation.
- Check if observation θ is true in this permuted version or not.
- Repeat 1000000 times and estimate probability of θ being true by chance.
- How to permute a sequence randomly?

Testing a gene for differential expression

- Suppose a gene's expression was measured in 100 different samples from cancer patients and 100 samples from healthy individuals
- Test whether gene is 'differentially expressed' between the two groups: T-test or Wilcoxon rank sum test
- Test produces a p-value, and if this p-value is $\leq \alpha$ (say $\alpha = 0.05$), we can proclaim this gene to be differentially expressed.

False positives?

- Let's look at this again, a little more closely
- Let's denote the null hypothesis of the test (e.g., "the two groups have the same mean gene expression level") as H_0.
- The test amounts to asking $Pr(X \geq \tau | H_0) \leq \alpha$, where X is the test statistic (e.g., t-statistic) and τ is the observed value of this test statistic.
- This test has a margin of error: α. That is, even if see $Pr(X \geq \tau | H_0) \leq \alpha$ and thus consider this gene to be differentially expressed ("non-random" difference between groups), this might in fact have happened by chance.
- In other words, even if the null hypothesis is true and the gene is not significantly deviating from random, the test may call it interesting and non-random. The probability of such a "false positive" prediction is α.

False positives?

- False Positive: "Positive" because rejecting null hypothesis usually incriminates the gene as being interesting in some way. "False because H_0 being true means that the prediction is false.
- With our testing procedure we are able to control "false positive" rate

Testing multiple genes

- Now consider repeating the above test on 10000 genes, one by one.
- In each test, the probability of false positive is α.
- In other words, if I do 10000 tests, I might make false positive errors $10000 \times \alpha$ times. (For $\alpha = 0.05$, this amounts to 500 false predictions!) This is the multiple hypothesis testing problem. A significance level (α) that looks convincing on a single test no longer looks so convincing when doing many many tests.
- We'd like to predict a set of genes as being interesting, i.e., as violating null hypothesis, but with 'control' over the false positive rate.

False Discovery Rate

- Is one such procedure
- The final outcome will be a set of genes predicted to be differentially expressed
- We will have some control on the proportion of false positives among these predicted genes
- The theory talks about 'tests' and not 'genes', of course. Here, we are using it in the context of tests involving genes.
- Say there are 10000 genes and 100 are truly differentially expressed. It is probably OK then to predict some set of 100 genes as being differentially expressed, with the disclaimer that (say) 10 of these may be false positives.
- Our testing is based on p-values. So we need a way to go from a p-value (e.g., "probability of a false positive call on this gene is 0.05") to an overall false positive proportion (e.g., "of all genes found significant by us, we expect 10% to be false positives").

An FDR Procedure

- Proposed by Benjamini and Hochberg in 1995. Many other procedures since then, but we'll only see this original one.
- Begin with a per-gene p-value, i.e., $Pr(X \geq \tau | H_0)$, for every one of the g genes being studied.
- Let the g p-values be denoted by $p_{(i)}$
- Consider these g p-values to be sorted in ascending order, i.e., $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(g)}$
- Let $H_0^{(i)}$ be the null hypothesis corresponding to $p(i)$
- Let $q_i = \frac{i\alpha}{g}$ for $i = 1, 2, 3 \ldots g$ where α is the desired FDR
- Let k be the max i such that $p_{(i)} \leq q_i$
- Procedure: Reject null hypothesis $H_0^{(1)}, H_0^{(2)}, \ldots H_0^{(k)}$ and accept all others.
- Theorem: This controls the FDR at level α. *What does that mean?*

Formal definition of FDR

- Consider the genes for which null hypothesis is rejected. (Say R in number.)
- Let V be number of genes (from these R) for which null hypothesis is true (i.e., falsely rejected, or false positive)
- Let $S(= R - V)$ be the number of genes (from these R) for which null hypothesis is false (i.e., correctly rejected)
- Define $Q = 0$ if $R = 0$ and $Q = V/R$ if $R > 0$.
- Define FDR = expectation of random variable Q.
- When using the FDR procedure we saw, the theorem says $E(Q) \leq \alpha$.

A note on FDRs vs p-values

- FDR is fundamentally different from a p-value.
- P-value assesses significance of data. If we publish some data that we claim to be significant, we should present a small p-value for the data (e.g., ≤ 0.05)
- FDR is generally used as a "culling tool"; the investigator wants to predict a set of genes to test experimentally, and an FDR of 0.1 or even 0.5 may be acceptable to them (they will do twice as much experimental work, which may be fine)