Stanford Stats 200

Self-study
Riley Wilson

1 Two sample t-test

This section of Rice is concerned with comparing samples from two distributions that may, or may not, be identical.

Suppose we're running an experiment and have the control group and the treatment group. We model the observations from both as a normal random variable with means μ_X and μ_y for the two distributions.In order to see if the treatment had any affect on the control, we'd look for the difference of means, or $\mu_x - \mu_y$. A natural estimate for this is $\bar{X} - \bar{Y}$. Since $\bar{X} - \bar{Y}$ is a difference of normal random variables, we know it's distributed as $N(\mu_x = \mu_y, \sigma(1/n+1/m))$.

Unfortunately, we often don't know the σ for X and Y, so we have to estimate it with the **pooled sample variance**.

$$s_p^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{m+n-2}$$

Where s_x^2 and s_y^2 are the sample variances of the observations from X and Y. This means S_p^2 is the weighted average of the sample variances we see from the two groups.

Theorem 1.1. Suppose $X_1 \dots X_n$ are independent and normally distributed random variables with mean μ_x and variance σ_x^2, and that $Y_1 \dots Y_m$ are iid random variables with mean μ_y and variance σ^2, and that Y_i are independent of the X_i. Then, the statistic:

$$t = \frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{s_p\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

follows a t distribution with $m+n-2$ degrees of freedom. Often, we refer to the denominator of the prior expression as $s_{\bar{x}-\bar{y}}$. Note also that we're asuuming a common σ between X and Y. This is a necessary part of the theorem that's, according to Rice, often problematic in practice. [1]

This part isn't motivated well in the lecture notes, and Rice's derivation is tedious, but it's also possible to apply a test between two distributions assuming unequal variances.

[1] the proof to this relies on a chi-squared random variable. I need to learn this to understand what else is happening in here.

Theorem 1.2 (**Welche's t-test**)**.** Suppose $X_q, \ldots, X_n \sim_{\text{iid}} N(\mu_x, \sigma_x)$ and $Y_1, \ldots, Y_n \sim_{\text{iid}} N(\mu_y, \sigma_y)$. For possibly different σ_x and σ_y. Then we can use the following test statistic:

$$T_{Welch} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{1}{n}S_X^2 + \frac{1}{m}S_Y^2}}$$

Apparently we don't even know the exact distribution this follows. What Welch did figure out is it's really close to a t distribution with

$$\frac{(S_X^2/n + S_Y^2/m)^2}{(S_X^2/n)^2/(n-1) + (S_Y^2/m)^2/(m-1)}$$

degrees of freedom. It's nasty, yes, but I foresee this being very helpful. Unfortunately, I can't tell you anything about when what is given behaves poorly. I'd feel uncomfortable using a tool unless I understand when it can go wrong. For more (but not much more) information, visit the wikipedia page for Welch's t-test.

What's important It's straightforward to conduct a significance test or construct a confidence interval.

Definition 1.1 (Confidence intervals)**.** A confidence interval of level α for a parameter μ is an interval $[a, b]$ such that

$$P(a < \mu < b) \geq \alpha$$

In words, it's an interval where we are α-percent sure the mean lies within it. Although frequentists are bound to disagree.

[2]

2 Signed rank test (Wilcoxian signed rank test)

One way to test whether the mean of two distributions is the same is to use the Signed Rank Test. It's an odd test, in that it doesn't utilize any parameters and requires pairing off observations between your two samples. In fact, what I'm about to describe is a more specific instance of the general signed rank test. [3]

[2] Rice does something frustrating in the text here. He talks about how some people manipulate data so it's more symmetric before applying statistical tests. He says there are mixed views about the practice, but doesn't give any arguments as to why.

[3] In the general case and our case, what we're actually doing is figuring out whether a distribution is symmetric about its mean.

We apply this test when we want to know whether two distributions share the same mean. Suppose we have n observations from two samples consisting of $x_1, \ldots, x_n$ and $y_1 \ldots y_n$. The steps are as follows:

1. Pair off all the observations and calculate $w_i = x_i - y_i$.
2. Rank all the w_i by the value of $|w_i|$ from smallest to largest. Let $R(w_i)$ be the rank of item w_i.
3. Now, we get to the namesake of the test. Throw away all of the $w_i < 0$. For the remaining $w_i's$, add up their ranks. Call this sum W. Formally, $W = \sum_{w_i > 0} R(w_i)$.
4. W is now our test statistic. For extreme values of W we can reject the null hypothesis.

Calculating our rejection zone for a given α is done like so. The PMF of W is:

$$P(W = w) = (\text{all the ways numbers} < w \text{ can sum to } w) * \frac{1}{2^n}$$

Calculating a CDF is computationally difficult for all but the smallest values of n, but we can do it.[4]. When n is large then we can rely on the fact W is asymptotically normal with $E[W] = \frac{n(n+1)}{4}$ and $Var(W) = \frac{n(n+1)(2n+1)}{24}$

The signed rank test is nice for small values of n, and when we don't want to assume normality.

Lastly, if $W_i = 0$, then we're going to toss the data point. If there are ties, but not too many, then we can let $R(w_i)$ be the average of the ranks it is tied with. Suppose w_i is tied with two other values, such that if you ranked them they would occupy the ranks 4,5, and 6. Then all of the ranks would be 5.

3 Rank-sum test

This test is quite similar to the signed-rank test, although we'll use the rank-sum test under different circumstances. Suppose we have a group of $n + m$ that we're going to run an experiment on. n of the things will be put in the control group, and m will be in the experiment group. Our null hypothesis is our experiment does nothing, and any differences between the two groups are due to chance. The alternative is some difference exists. After creating the control and experiment groups, we proceed as follows:

1. Run the experiment.

[4] A naive algorithm works in $O(2^n)$ time by trying out every subset

2. Take the measurements from the control and experimental groups.
3. Rank these measurements from smallest to largest.
4. Sum the ranks of the control (or experimental) group.

The sum of the ranks — call it W — is the test statistic. Two facts help us determine the null distribution. First, we're working with the ranks of the measurements, not the measurements themselves. This means we don't need to speculate about the distributions they came from. Second, under the null hypothesis, the ranks are distributed randomly (by us). As a result, every collection of ranks we see in the control group is equally likely, and we can compute a probability for it, and hence a PMF.

For instance, suppose $n = m = 3$, so we have three things in our control and three in our experimental group. Suppose also that the three lowest ranks, 1,2,3, are all in the control group for a rank-sum statistic of 6. We can calculate the probability of this occurring under the null hypothesis. There are $\binom{6}{3} = 20$ possible ranks that could have landed in the control group, all with an equal probability if the experiment did nothing. Since we can only achieve the sum of 6 one way, the probability we observe the test statistic take on this value is $\frac{1}{20} = .5$. If this was a one-sided hypothesis test with $\alpha = .05$, then we would (barely) reject the null hypothesis. Rice sets up the rank-sum test as discriminating between whether nothing happened or there is any difference — regardless of the sign — between the two groups, but in principle I see no reason why we can't have a single-sided test.

In practice, tables for the null distribution of the rank-sum test are common.

4 Permutation test

The permutation test is incredibly flexible. We can use it with any test statistic where it's difficult to understand the null distribution, though for expository purposes we're going to consider the statistic $T = \bar{X} - \bar{Y}$ for a sample from X and a sample from Y. Imagine also we're testing whether X and Y share the same distribution.

To conduct the permutation test, we apply the following steps. It helps to think about T as a function of all your observations, $T(X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_m)$.

1. Calculate your test statistic with the data you have.
2. Now permute all your observations. In our case, we can imagine $m + n$ data points arranged in a line. The first n of them are samples from X while the next m are samples from Y. Switch them all up in the line, but still compute $\bar{X}$ as the first n observations in the line, and $\bar{Y}$ as the second m observations.
3. Compute T again under this permutation. Repeat for all $(m+n)!$ possible permutations of your data.

4. Now we have a bunch of $T's$ based on the permutations. In order to calculate the p-value of your first T — the one with the original data — we need to figure out the band in which α of the $T's$ lie in.

5. If our original statistic lies outside the band, then we can reject with significance α. If not, then we're stuck with the null hypothesis.

[add in assumptions about exchangeability and why one would want to use this test.]

5 A couple notes on experimental design

At this point in the course, Stats 200 takes a departure from a traditional math class. We begin to talk about how we can design experiments to get the most from the statistical techniques we've developed. You might think this is a discussion about assuming normality when we're not confident, or supposing two distributions share a variance when they don't (which are good things to look out for) but it's more basic than that.

He wants to talk about randomization, controls, and blinds. Randomization should be an intuitive concept. The units we perform an experiment on should be randomized in order for our results to be generalizable. If our analysis is on a subset of the population, then it's unclear whether changes are due to preexisting characteristics or our intervention. Controls are another crucial part of running a successful experiment. If we don't have a baseline, how do we know our results are good? Is the effect size large or small? These are basic questions that are best answered by keeping a control group. Blinds are the last part Rice mentions. I find this surprising, since we're crossing a bit into the psychological domain here. Humans are quite effective at believing what's expedient, so preventing all involved from subconsciously influencing the results of a study is important.

There are also ways we can effect study design that clearly influence the mathematics of our analysis. Suppose we're designing an experiment. We want to know whether popcorn increases fan enjoyment of viewing Puss in Boots: The Last Wish. You propose the following experiment: you separate the movie theater into two groups. The first group is given popcorn at the beginning of the film. The second is given popcorn at the halfway point. At the halfway point, and at the end of the film, you administer surveys to the audience to measure their enjoyment. To avoid bias, you randomly pick who gets in the first group as the second.

Statistically, how is this going to work? One way to do the analysis is with a t-test. Model each of observation of the first group's satisfaction as a sample from $X\ N(\mu_x, \sigma)$ and each observation from the second group as a sample from $Y\ N(\mu_y, \sigma)$. (We're assuming both distributions share σ and we know it).[5] A good statistic to see if the means of these two distributions are different is $\bar{X} - \bar{Y}$.

[5] normally when we don't know the variance, we would use s_p as an estimate, which is the pooled sample variance.

From what we know about normal distributions, $T = \bar{X} - \bar{Y} \sim N(\mu_x + \mu_y, \sigma^2(\frac{1}{n} + \frac{1}{m}))$. Under the null hypothesis, both of the means are the same, so we can normalize:

$$Z = \frac{\bar{X} - \bar{Y}}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

This allows us to compute the p-value of our statistic.

How powerful is our experiment? Given the alternate hypothesis is true, what's the probability we will reject the null? In order to get a bearing on this, we need to start thinking about the effect size. Most of the time we think of effect size as the following: $\frac{\mu_x - \mu_y}{\sigma}$. In words, this is the difference between the two distributions as a fraction of the shared variance. Let $d = \frac{\mu_x - \mu_y}{\sigma} \frac{1}{\sqrt{\frac{1}{n} + \frac{1}{m}}}$

The larger this quantity is, the larger the effect size under the alternative distribution. To calculate the power, we can say:

$$P_{H_1}[Z > z(\alpha)] = P_{H_1}[Z - d > z(\alpha) - d] = 1 - \Phi(z(\alpha) - d)$$

Which is tractable for a given α. We want to maximize this quantity, as it's the power. To do so, make $m = n = 150$. There are two things missing in order to get a number out of the expression for power we have above. The first is to get a value for effect size.. In our case, assume an effect size of .11 for the popcorn. This step is feasible in practice because we know the effect size of other, similar phenomena. There are tables you can look up for different disciplines and whatnot. The second thing we need to do is solve for d. Then we can calculate $1 - \Phi(z(\alpha) - d)$

Now we need to solve for n in the equation:

$$d = \frac{\mu_x - \mu_y}{\sigma}\sqrt{\frac{n}{2}} = .11\sqrt{\frac{150}{2}} = .95$$

Do the rest of the algebra, and it turns out the power is .244 for $\alpha = .05$. This means we would only have a .244 chance of rejecting the null if the effect was really there! Hopefully you can see why this is bad. Imagine if your alternative hypothesis is correct, you've spent all this time and money on an experiment, but your effort is only going to detect the effect a quarter of the time!

Suppose now we wanted to boost the power of our test. What kind of sample size would we need to achieve a power of .9 for the given effect size? I'll spare you the algebra, but it turns out to be ≈ 1400 samples in each group. This means instead of doing this at a large movie theater, you'd have to do it at a massive one. Maybe the premier of Puss In Boots.

5.1 improving power

One way to improve power is to run bigger experiments. This is all well and good if you have money and time to burn. Most of us don't. Instead, we can change the design of our experiment to increase the power with the same number of samples.

In our example, the statistic was the difference of sample means (in effect, comparing the average control score to the average intervention score). What if instead we compared each person's sans-popcorn satisfaction score to their popcorn satisfaction score? A random half of the audience would be given popcorn during the first half of the movie, and the others will be given popcorn during the latter half.

Formally, suppose we have n people and $X_1, \dots, X_n$ are their satisfaction scores during the non-popcorn section of the movie. Likewise, $Y_1 \dots Y_n$ are their scores during the popcorn portion.

In this scenario we're going to use a paired two sample t-test.

Let $D_i = X_i - Y_i$ and reject the null hypothesis H_0, for large values of the t statistic: [6]

$$T = \frac{\sqrt{n}\bar{D}}{S}$$

$\bar{D}$ and S are the sample mean and variance of the D_i's.

We have a test, but what the feck is its power? since $X \sim N(\mu_x, \sigma^2)$ and $Y \sim N(\mu_y, \sigma^2)$, (X_i, Y_i) corresponds to a bivariate normal distribution with some amount of correlation.

$$(X_i, Y_i) \sim N(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{pmatrix})$$

To that end, we know each D_i is normally distributed with $E[D_i] = \mu_x - \mu_y$ and variance:

$$\begin{aligned} Var(D_i) &= Cov[X_i - Y_i, X_i - Y_i] \\ &= Cov[X_i, X_i] - Cov[X_i, Y_i] - Cov[Y_i, X_i], Cov[Y_i, Y_i] \\ &= 2\sigma^2(1 - \rho) \end{aligned}$$

If n is large, then S^2 should be very close to $Var[D_i]$. Suppose for simplicity we know $Var[D_i]$ and our statistic becomes:

[6] how is the average difference distinct from the difference of the averages?

$$Z = \frac{\sqrt{n}\bar{D}}{\sqrt{2\sigma^2(1-\rho)}}$$

We're going to play a similar game as we did earlier to figure out the power of this statistic. In order to figure out the power, we need to find the value of:

$$P_{H_1}[Z > z(\alpha)] = P_{H_1}[Z - d > z(\alpha) - d] = 1 - \Phi(z(\alpha) - d) \tag{1}$$

and in particular, d. The useful thing we know about d is it's the thing that normalizes our test statistic, which we're calling Z. Under an alternate hypothesis, which is $\mu_x \neq \mu_y$, we know $\bar{D}$ is not distributed according to a standard normal, but in order to standardize it we need to subtract the means of the sample distributions with the correct denominators:

$$\frac{\sqrt{n}\bar{D}}{\sqrt{2\sigma^2(1-\rho)}} - \frac{\mu_x - \mu_y}{\sigma}\sqrt{\frac{n}{2(1-\rho)}} \tag{2}$$

Is a normal distribution. This means $\frac{\mu_x-\mu_y}{\sigma}\sqrt{\frac{n}{2(1-\rho}}$ is our d value. Examining equation 1, we want our d to be as large as possible to maximize power. As it demonstrates, we can do this either by increasing n, or increasing ρ, the correlation between X and Y. When we compare satisfaction scores between individuals, we're effectively introducing some correlation between the what we measure in the control and what we measure after the intervention.

At first, this was confusing to me. Running a good experiment looked all about randomizing. Having regularities sneak in was experimental sin, but the value of correlated observations becomes clearer when we consider extremes. Imagine our Xs and Ys were perfectly correlated. Any fluctuation in one would perfectly determine moves in another. For our experimental purposes, this is perfect. Since we know the value of X given Y, or vice versa, any deviations from the expected value would be caused by our experiment.

Another cool observation is what we can know is asymmetric. If $\rho = 1$, then we're running an infinitely-powered experiment, which is cool. However, if our observations in the control and intervention groups are uncorrelated, $\rho = 0$, then we can still learn something. It's just going to take a big sample to get the power we want.

5.2 bonus proof: S^2 to σ^2

In this section, we used the fact $S^2 \to \sigma^2$ as $n \to \infty$. We're going to prove it, since we're good mathematicians.

Let X_i's be iid with mean 0 and variance σ^2.

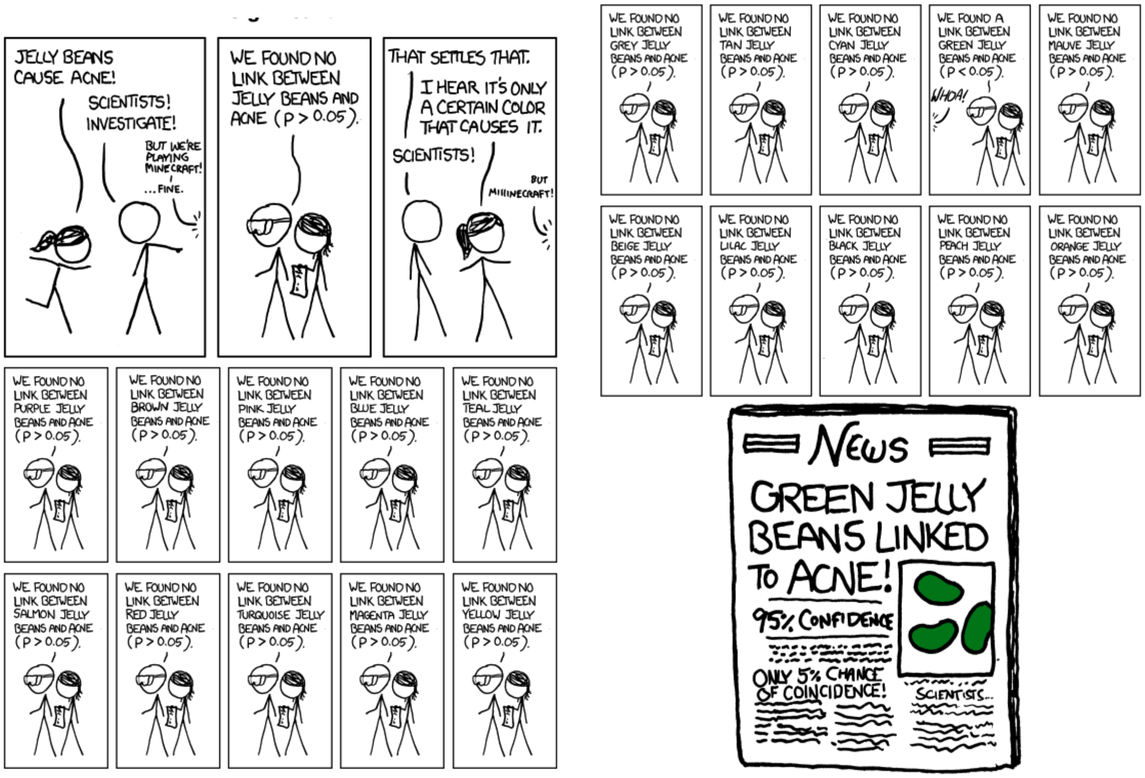
$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}\sum(X_i - n\bar{X}^2) \tag{3}$$

$$= \frac{n}{n-1}\underbrace{\frac{1}{n}\sum X_i^2}_{\to\sigma^2\text{by LLN}} - \frac{n}{n-1}\underbrace{\bar{X}^2}_{\to 0\text{by LLN}} \tag{4}$$

The law of large numbers works here since we're working with random variables with mean 0. Note we need to use the continuous mapping theorem in order to say that the product of these quantities that limit to certain values in probability is the product of their limits.

6 Multiple testing problem

Maybe you've heard of p-hacking, or data dredging. If you have, you might understand it vaguely as a dark art of science that produces unreplicable papers. That characterization is roughly true, but now is your opportunity to understand the statistical basis of it. Xkcd can give us a rough primer:


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

Source: here.

Suppose you have a big dataset. To use Xkcd's example, it includes information on jelly bean consumption and prevalence of acne for many individuals. You run a statistical test (like those discussed above) in order to see whether consuming jelly beans is linked to acne. That test might not reject the null hypothesis, as in the comic. However, you run the test again, on the same data, for the different colors of jelly bean, looking for a relationship. Eventually, you find a statistically significant link between green jelly bean consumption and acne.

What went wrong here? I like to think of statistical tests like unreliable metal detectors. Sometimes they'll (correctly) say nothing is underfoot. Sometimes they'll give a false positive. Eventually, if you use one of these metal detectors long enough, it'll malfunction and you'll get a false positive, even if, in actuality, there's only sand beneath you. The scientists in the comic applied a metal detector again, and again, and again. Run the process enough times and you'll get a false positive and a publicity-worthy result. This is called p-hacking.

We can also discuss the phenomenon formally. The α value of a test is the probability of falsely rejecting the null hypothesis for an alternative. Even if α is really small, if we run the test enough times, we're bound to get a false positive.[7] In fact, if we test n times we should expect $n\alpha$ false positives. This is bad if we have n in the thousands.

A lot has been written about p-hacking in science. A famous paper in the field is Ionnadis' "why most published research findings are false." For a popular science introduction to p-hacking, bias, fraud, and all the other ugly parts of science, see Stuart Ritchie's "Science Fictions."

6.1 Bonferroni's correction

Here's a simple way to control for p-hacking, or random false positives: lower the p-value. In fact, lower it enough so if each of n tests has significance α_t, the collection of tests has significance α. In other words, change α_t for each test such that the probability of falsely rejecting any one of the n tests you'll carry out is α. Bonferroni's correction is a straightforward way to do this. Suppose $H_1, \dots, H_n$ are a bunch of true null hypotheses:

$$\begin{aligned} P(\text{falsely rejecting any hypothesis}) &= P(\text{reject } H_1, \dots, \text{ reject } H_n) && (5)\\ &= \leq P(\text{reject } H_1, \vee \text{ reject } H_2) + \cdots + P(\text{reject } H_n) && (6)\\ &= \alpha_t n && (7) \end{aligned}$$

And if we want $\alpha_t n = \alpha$ we set $\alpha_t = \alpha / n$ Now, the probability of falsely rejecting any null hypothesis is α. The advantage of Bonferroni's correction is it's conceptually straightforward. We think of all of the tests we're going to do as one big significance test and then control the

[7] I want to include the infinite monkey theorem. It gives good justification as to why this is true

α of that big test. The only issue is sometimes Bonferroni's correction is quite conservative.[8]

6.2 The Benjamini-Hochberg procedure

So you're no fan of the Bonferroni method. No worries. There's another way we can attenuate the difficulties of the multiple hypothesis problem. The Bonferroni method controls the α of the ensemble of tests, but perhaps we'd want to control something different. Suppose we have a n null hypotheses, $H_0^0, H_0^1 \dots H_0^n$, and we perform tests to see which ones to reject. An omniscient figure (say, God) may use the following table summarize our results.

	H_0 is true	H_0 is false	Total
Reject H_0	V	S	R
Accept H_0	U	T	$n - R$
Total	n_0	$n - n_0$	n

V is the number of hypotheses we rejected when they were actually true. T is the number of hypotheses we accept when they're actually false. Why God chose to the letters he did, I do not know. Do they make an acronym? TUVS? SVUT? STUV? They all sound like slurs. Maybe God's playful side comes out in statistics.

Regardless, one thing we may be interested in controlling is the so-called False Disovery Rate (FDR). The FDR tells us how many of the ”discoveries”, or rejections of the null hypothesis, we should expect to be erroneous, or false.[9]

In order to do this, we're going to implement the Benjamini-Hochberg procedure. It will allow us to control the FDR as we see fit. The procedure itself is quite simple, if initially inscrutable. Suppose we want $FDR < q$.

The Procedure

1. Do all your tests. Calculate n P-values and order them $P_{(1)}, P_{(2)}, \dots P_{(n)}$.
2. Find the largest R such that $P_{(R)} < \frac{qR}{n}$
3. Reject all the null hypotheses $P_{(1)}, P_{(2)} \dots P_{(R)} < \frac{qR}{n}$
4. Done.

It's straightforward to implement, but I was agog the first time I saw this. In my textbook I wrote ”This seems absolutely occult. How is this supposed to work?” For your curiosity and mine, we'll go through the proof.

[8] Are there additional critiques of Bonferroni's correction?

[9] Why is this a better statistic that α? I would want a better explanation here

Proof. The first thing to recognize is the FDR is defined as the expectation of the False Discovery Proportion(FDP).[10] The FDP is $\frac{V}{R}$. The strategy is to get an expression for the FDR and then show how it will always be less than or equal to q if we follow the BH procedure. Let's begin. Suppose we have n null hypotheses, and the first n_0 of them are true.

$$FDR = E[FDP] \tag{8}$$

$$= E[\sum_{r=1}^{n} \frac{V}{r} \mathbb{1}_{R=r}] \tag{9}$$

All we've done so far is rewrite the FDP. We only count the $\frac{V}{r}$ term when the r takes on the correct value, that is, R. Notice also $V = \sum_{j=1}^{n_0} \mathbb{1}_{\text{reject } H_o^{(j)}}$

$$= E[\frac{1}{r} \sum_{r=1}^{n} \sum_{j=1}^{n_0} \mathbb{1}_{\text{reject } H_o^{(j)}} \mathbb{1}_{R=r}] \tag{10}$$

$$= \frac{1}{r} \sum_{r=1}^{n} \sum_{j=1}^{n_0} P(\text{reject } H_o^{(j)} \text{ and } r = R) \tag{11}$$

Suppose we're now working with the BH procedure. Under what circumstances will we reject $H_0^{(j)}$? Only when its associated P-value, $P_{(j)}$, is less than $\frac{qR}{n}$. We also need the BH to reject $R-1$ hypotheses other hypotheses, not including our $H_0^{(j)}$. Call this event:

$$\epsilon^{(r)} := \{P_{(1)}, \dots P_{(r-1)} \leq \frac{qr}{n}, P_{(r)} > \frac{q(r+1)}{n} \dots P_{(n-1)} > q\} \tag{12}$$

Two more facts are relevant. First, we assume the P-values are independent. Second, that $P_{(j)} \sim U(0,1)$, since $H_0^{(j)}$ is a true null hypothesis.

$$FDR = \frac{1}{r} \sum_{r=1}^{n} \sum_{j=1}^{n_0} P(\text{reject } H_o^{(j)}) P(\epsilon^{(r)} \text{is true}) \tag{13}$$

$$= \frac{1}{r} \sum_{r=1}^{n} \sum_{j=1}^{n_0} P(P_j < \frac{qr}{n}) P(\epsilon^{(r)} \text{is true}) \tag{14}$$

$$= \sum_{j=1}^{n_0} \sum_{r=1}^{n} \frac{q}{n} P(\epsilon^{(r)} \text{is true}) \tag{15}$$

$$\tag{16}$$

[10] why is this? I do not know.

Just consider the expression $\sum_{r=1}^{n} P(\epsilon^{(r)}\text{is true})$ In words, it's asking for the probability we reject 1 hypothesis, or 2 hypotheses, or 3 hypotheses, all the way up to n. Since we have to reject at least one hypothesis, $\sum_{r=1}^{n} P(\epsilon^{(r)}\text{is true}) = 1$.

$$FDR = \frac{q}{n}\sum_{j=1}^{n_0} 1 = \frac{qn_0}{n} \leq q \tag{17}$$

Finishing the proof. □

7 Parameter estimation

I don't think statistics is taught well. Every course starts with the basics of probability and then covers statistical tests. I think statistical tests are necessary to learn, but boring. What's more exciting — and ambitious — is creating models. Ideally, we want to extract the guiding principles, or forces, from a phenomenon and use those to create predictions. We might not need infinite computing power or huge brains in order to adequately understand or forecast the world. It might be enough to understand the principles at play and simulate those in our tiny heads.

Parameter estimation is the first step towards building models. Here's the scene: we have data, and we conjecture it comes from a random variable distributed in a certain way. Suppose we (somehow) reason it's normally distributed. But what are the parameters? And how confident can we be our estimation of these parameters is correct? Being certain our data is governed by a standard normal versus weakly confident it is a $N(1.43, 10^{99})$ are two vastly different epistemic situations.