Class 20 in-class problems, 18.05, Spring 2022

Concept questions

Concept question 1. Significance tests

Three different tests are run, all with significance level $\alpha = 0.05$.

Experiment 1: finds $p = 0.003$ and rejects its null hypothesis H_0.

Experiment 2: finds $p = 0.049$ and rejects its null hypothesis.

Experiment 3: finds $p = 0.15$ and fails to rejects its null hypothesis.

Which result has the highest probability of being correct?

1. Experiment 1 *2. Experiment 2*

3. Experiment 3 *4. Impossible to say.*

Solution: Impossible to say. You can't compute probabilities of hypotheses from p values.

Concept question 2. Multiple testing

(a) *Suppose we have 6 treatments and want to know if the average recovery time is the same for all of them. If we compare two at a time, how many two-sample t-tests do we need to run?*

(i) *1* **(ii)** *2* **(iii)** *6* **(iv)** *15* **(v)** *30*

(b) *Suppose we use the significance level 0.05 for each of the 15 tests. Assuming the null hypothesis, what is the best estimate of the probability that we reject at least one of the 15 null hypotheses?*

(i) < 0.05 **(ii)** *0.05* **(iii)** *0.10* **(iv)** > 0.25

Solution: (a) (iv) 6 choose 2 = 15.

(b) (iv) Greater than 0.25.

Under H_0 the probability of rejecting for any given pair is 0.05. Because the tests aren't independent, i.e. if the group1-group2 and group2-group3 comparisons fail to reject H_0, then the probability increases that the group1-group3 comparison will also fail to reject.

We can say that the following 3 comparisons: group1-group2, group3-group4, group5-group6 are independent. The number of rejections among these three follows a binom$(3, 0.05)$ distribution. The probability the number is greater than 0 is $1 - (0.95)^3 \approx 0.14$.

Even though the other pairwise tests are not independent, they do increase the probability of rejection. In simulations of this with normal data, the false rejection rate was about 0.36.

Board questions

Problem 1. Stop!

Experiments are run to test a coin that is suspected of being biased towards heads. The significance level is set to $\alpha = 0.1$

Experiment 1: *Toss a coin 5 times. Report the sequence of tosses.*

Experiment 2: *Toss a coin until the first tails. Report the sequence of tosses.*

(a) *Give the test statistic, null distribution and rejection region for each experiment. List all sequences of tosses that produce a test statistic in the rejection region for each experiment.*

(b) *Suppose the data is $HHHHT$.*

(i) Do the significance test for both types of experiment.

(ii) Do a Bayesian update starting from a flat prior: Beta(1,1).

Draw some conclusions about the fairness of coin from your posterior. (Use R: `pbeta` *for computation in part (b).)*

Solution: (a) Experiment 1: The test statistic is the number of heads x out of 5 tosses. The null distribution is binomial(5,0.5). The rejection region is $\{x = 5\}$.

The sequence of tosses $HHHHH$ is the only one that leads to rejection.

Experiment 2: The test statistic is the number of heads x until the first tails. The null distribution is geom(0.5), the rejection region $\{x \geq 4\}$.

The sequences of tosses that lead to rejection are $\{HHHHT, HHHHH**T\}$, where '**' means an arbitrary length string of heads.

(b) (i) For experiment 1 and the given data, 'as or more extreme' means 4 or 5 heads. So for experiment 1 the p-value is $P(4 \text{ or } 5 \text{ heads} \mid \text{fair coin}) = 6/32 \approx 0.20$.

For experiment 2 and the given data 'as or more extreme' means at least 4 heads at the start. So $p =$ `1 - pgeom(3,0.5)` $= 0.0625$.

(ii) Since the likelihood functions are proportional, the Bayesian posterior will be the same for both experiments. Let θ be the probability of heads, Since we have a conjugate pair (beta-binomial or beta-geometric), Bayesian updating is just updating the parameters in the Beta distribution. Four heads and a tail updates the prior Beta(1,1) to the posterior Beta(5,2). Using R we can compute

$$P(\text{Coin is biased to heads}|\text{data}) = P(\theta > 0.5|\text{data}) = \texttt{1 -pbeta(0.5, 5,2)} = 0.89.$$

If the prior is reasonable then the probability the coin is biased towards heads is fairly high.

Problem 2. Stop!

For each of the following experiments (all done with $\alpha = 0.05$)

(a) *Comment on the validity of the claims.*

(b) *Find the true probability of a type I error in each experimental setup.*

1. *Experiment 1. By design Alessandre did 50 trials and computed $p = 0.04$.*
   They report $p = 0.04$ with $n = 50$ and declare it significant.

2. *Experiment 2. Sara did 50 trials and computed $p = 0.06$.*
   Since this was not significant, she then did 50 more trials and computed $p = 0.04$ based on all 100 trials.
   She reports $p = 0.04$ with $n = 100$ and declares it significant.

3. *Experiment 3. Gabriel did 50 trials and computed $p = 0.06$.*
   Since this was not significant, he started over and computed $p = 0.04$ based on the

next 50 trials.
He reports $p = 0.04$ with $n = 50$ and declares it statistically significant.

Solution: Experiment 1. (a) This is a reasonable NHST experiment.

(b) The probability of a type I error is 0.05.

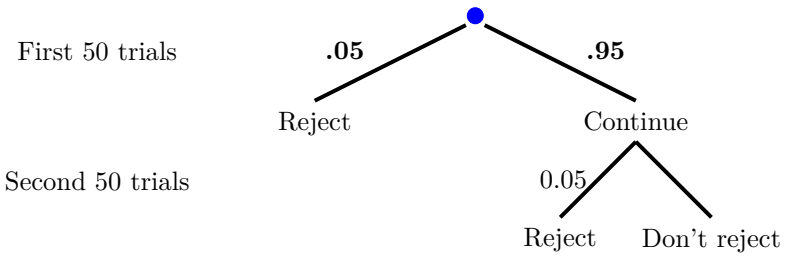
Experiment 2. (a) The actual experiment run:
(i) Do 50 trials.
(ii) If $p < 0.05$ then stop.
(iii) If not run another 50 trials.
(iv) Compute p again, pretending that all 100 trials were run without any possibility of stopping.

This is not a reasonable NHST experimental setup because the second p-values are computed using the wrong null distribution.

(b) If H_0 is true then the probability of rejecting is already 0.05 by step (ii). It can only increase by allowing steps (iii) and (iv). So the probability of rejecting given H_0 is more than 0.05. We can't say how much more without doing a more complicated computation.

Experiment 3. (a) See answer to (2a).

(b) The total probability of a type I error is more than 0.05. We can compute it using a probability tree. Since we are looking at type I errors all probabilities are computed assuming H_0 is true.



The total probability of falsely rejecting H_0 is $0.05 + 0.05 \times 0.95 = 0.0975$.

Problem 3. From Class 19: Chi-square for independence
(From Rice, Mathematical Statistics and Data Analysis*, 2nd ed. p.489)*

Consider the following contingency table of counts

Education	*Married once*	*Married multiple times*	*Total*
College	*550*	*61*	*611*
No college	*681*	*144*	*825*
Total	*1231*	*205*	*1436*

Use a chi-square test with significance level 0.01 to test the hypothesis that the number of marriages and education level are independent.

Solution: The null hypothesis is that the cell probabilities are the product of the marginal probabilities. Assuming the null hypothesis we estimate the marginal probabilities in orange and multiply them to get the cell probabilities in blue.

Education	Married once	Married multiple times	Total
College	0.365	0.061	611/1436
No college	0.492	0.082	825/1436
Total	1231/1436	205/1436	1

We then get expected counts by multiplying the cell probabilities by the total number of women surveyed (1436). The table shows the observed, expected counts:

Education	Married once	Married multiple times
College	550, 523.8	61, 87.2
No college	681, 707.2	144, 117.8

We then have

$$G = 16.55 \qquad \text{and} \qquad X^2 = 16.01$$

The number of degrees of freedom is $(2-1)(2-1) = 1$. (We can count this: we needed the marginal counts to compute the expected counts. Now setting any one of the cell counts determines all the rest because they need to be consistent with the marginal counts from the data.) So, we get

$$p = \texttt{1-pchisq(16.55,1)} = 0.000047$$

Therefore we reject the null hypothesis in favor of the alternate hypothesis that number of marriages and education level are not independent

Discussion questions

1. From Class 18: Type I errors Q1

Suppose a journal will only publish results that are statistically significant at the 0.05 level. What percentage of the papers it publishes contain type I errors?

Solution: This is asking for $P(H_0|\text{rejected } H_0)$. This is the probabilitiy of a hypothesis. Since we are not given a prior (baserates), we can't know this. **The percentage could be anywhere from 0 to 100!**

Remember: signifcance is the false positive rate, i.e $P(\text{rejection}|H_0)$. You need the base rate (prior) to know how often the test as a whole is wrong

2. From Class 18: Type I errors Q2

Jerry desperately wants to cure diseases but he is terrible at designing effective treatments. He is however a careful scientist and statistician, so he randomly divides his patients into control and treatment groups. The control group gets a placebo and the treatment group gets the experimental treatment. His null hypothesis H_0 is that the treatment is no better than the placebo. He uses a significance level of $\alpha = 0.05$. If his p-value is less than α he publishes a paper claiming the treatment is significantly better than a placebo.

(a) *Since his treatments are never, in fact, effective what percentage of his experiments result in published papers?*

(b) *What percentage of his published papers contain type I errors, i.e. describe treatments that are no better than placebo?*

Solution: (a) Since in all of his experiments H_0 is true, roughly 5%, i.e. the significance level, of his experiments will have $p < 0.05$ and be published.

(b) This is asking for $P(H_0|\text{rejected } H_0)$. This is the probabilitiy of a hypothesis. Since we are given the prior (baserate), that is, since all his treatments are no better than placebo, we can answer this: All of his published papers contain type I errors.

3. From Class 18: Type I errors Q3

Jen is a genius at designing treatments, so all of her proposed treatments are effective. She is also a careful scientist and statistician, so she too runs double-blind, placebo controlled, randomized studies. Her null hypothesis is always that the new treatment is no better than the placebo. She also uses a significance level of $\alpha = 0.05$ and publishes a paper if $p < \alpha$.

(a) *How could you determine what percentage of her experiments result in publications?*

(b) *What percentage of her published papers contain type I errors, i.e. describe treatments that are, in fact, no better than placebo?*

Solution: (a) The percentage that get published depends on the power of her treatments. If they are only a tiny bit more effective than placebo then roughly 5% of her experiments will yield a publication. If they are a lot more effective than placebo then as many as 100% could be published.

(b) This is asking for $P(H_0|\text{rejected})$. Since we are given the prior (baserate), that is, since all her treatments are better than placebo, we can answer this: None of her published papers contain type I errors.

18.05 Introduction to Probability and Statistics
Spring 2022