

# Comparison of Bayesian and Frequentist Inference

18.05 Spring 2014

Jeremy Orloff and Jonathan Bloom

## Concept question

Three different tests are run all with significance level  $\alpha = .05$ .

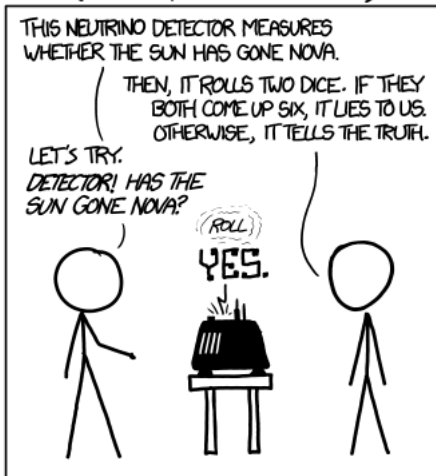
1. Experiment 1: finds  $p = .03$  and rejects its null hypothesis  $H_0$ .
2. Experiment 2: finds  $p = .049$  and rejects its null hypothesis.
3. Experiment 3: finds  $p = .15$  and fails to reject its null hypothesis.

Which result is most likely to be correct?

(Click **4** if you don't know.)

## Solution

DID THE SUN JUST EXPLODE?  
(IT'S NIGHT, SO WE'RE NOT SURE.)



Courtesy of xkcd. CC-BY-NC.

<http://xkcd.com/1132/>

## Board question: 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

Are the number of marriages and education level independent?

Test this using a chi-squared test with significance level 0.01.

## 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 red and multiply them to get the cell probabilities in blue.

Education	Married once	Married multiple times	Total
College	.365	.061	611/1436
No college	.492	.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

## Solution continued

We then have

$$G = 16.55 \quad \text{and} \quad \chi^2 = 16.01$$

The number of degrees of freedom is 1. This is because we are given the marginal counts and now any one of the cell counts determines all the rest. We get

$$p = 1 - \text{pchisq}(16.55, 1) = .000047$$

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

## Concept question: multiple-testing

1. Suppose we use two-sample  $t$ -tests at  $\alpha = .05$  level to determine whether 6 treatments all have the same recovery time. How many  $t$ -tests might we need to run?

- 1) 1      2) 2      3) 6      4) 15      5) 30

## Concept question: multiple-testing

1. Suppose we use two-sample  $t$ -tests at  $\alpha = .05$  level to determine whether 6 treatments all have the same recovery time. How many  $t$ -tests might we need to run?

- 1) 1      2) 2      3) 6      4) 15      5) 30

2. In the situation above, assuming all 6 means are the same, what is the probability that we reject at least one of the 15 null hypotheses?

- 1) Less than .05    2) .05    3) .10    4) Greater than .50



## Concept question: multiple-testing

1. Suppose we use two-sample  $t$ -tests at  $\alpha = .05$  level to determine whether 6 treatments all have the same recovery time. How many  $t$ -tests might we need to run?

- 1) 1      2) 2      3) 6      4) 15      5) 30

2. In the situation above, assuming all 6 means are the same, what is the probability that we reject at least one of the 15 null hypotheses?

- 1) Less than .05    2) .05    3) .10    4) Greater than .50

**Discussion:** Recall that there is an  $F$ -test that tests if all the means are the same. What is an advantage of using the  $F$ -test rather than many two-sample  $t$ -tests?

## Board question: Stop!

Experiments are run to test a coin that is suspected of being biased towards heads. The significance level is set to  $\alpha = .1$

**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.

1. 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.

2. Suppose the data is *HHHHT*.

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

(b) Do a Bayesian update starting from a flat prior:  $\text{Beta}(1,1)$ .

Draw some conclusions about the fairness of coin from your posterior.

(Use R: `pbeta` for computation)

## Solution

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

**3a.** For experiment 1 and the given data, 'as or more extreme' means 4 or 5 heads. So for experiment 1 the  $p$ -value is  $P(4 \text{ or } 5 \text{ heads} \mid \text{fair coin}) = 6/32 \approx .20$ .

For experiment 2 and the given data 'as or more extreme' means at least 4 heads at the start. So  $p = 1 - \text{pgeom}(3, .5) = .0625$ .

**3b.** Let  $\theta$  be the probability of heads, Four heads and a tail updates the prior on  $\theta$ ,  $\text{Beta}(1,1)$  to the posterior  $\text{Beta}(5,2)$ . Using R we can compute

$$P(\text{Coin is biased to heads}) = P(\theta > .5) = 1 - \text{pbeta}(.5, 5, 2) = .89.$$

## Board question: Stop II

For each of the following experiments (all done with  $\alpha = .05$ )

(a) Comment on the validity of the claims.

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

- 1 By design Peter did 50 trials and computed  $p = .04$ .  
He reports  $p = .04$  with  $n = 50$  and declares it significant.
- 2 Erika did 50 trials and computed  $p = .06$ .  
Since this was not significant, she then did 50 more trials and computed  $p = .04$  based on all 100 trials.  
She reports  $p = .04$  with  $n = 100$  and declares it significant.
- 3 Jerry did 50 trials and computed  $p = .06$ .  
Since this was not significant, he started over and computed  $p = .04$  based on the next 50 trials.  
He reports  $p = .04$  with  $n = 50$  and declares it statistically significant.

## Solution

- (a) This is a reasonable NHST experiment.

(b) The probability of a type I error is .05.
- (a) This is a reasonable NHST experiment.

(b) The probability of a type I error is .05.
- (a) The actual experiment run:

  - Do 50 trials.
  - If  $p < .05$  then stop.
  - If not run another 50 trials.
  - Compute  $p$  again, pretending that all 100 trials were run without any possibility of stopping.

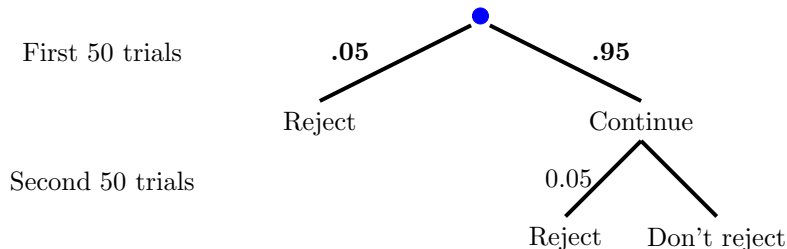
This is not a reasonable NHST experimental setup because the second  $p$ -values are computed using the wrong null distribution.

(b) If  $H_0$  is true then the probability of rejecting is already .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 .05. We can't say how much more without more details.

## Solution continued

4. (a) See answer to (3a).

(b) The total probability of a type I error is more than .05. We can compute it using a probability tree. Since we are looking at type I errors all probabilities are computed assume  $H_0$  is true.



The total probability of falsely rejecting  $H_0$  is  $.05 + .05 \times .95 = .0975$

MIT OpenCourseWare  
<http://ocw.mit.edu>

## 18.05 Introduction to Probability and Statistics

Spring 2014

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.