18.05 Problem Set 9, Spring 2022 Solutions

Problem 1. (15: 10,5 pts.)
We perform a t-test for the null hypothesis $H_0 : \mu = 10$ at significance level $\alpha = 0.05$ by means of a dataset consisting of $n = 16$ elements with sample mean 11 and sample variance 4.

(a) *Should we reject the null hypothesis in favor of $H_A : \mu \neq 10$?*

Solution: This is a two-sided alternative. The t-statistic is

$$\frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{1}{2/4} = 2.$$

Since we have $n = 16$ our t statistic has 15 degrees of freedom.

We have the two-sided p-value

$$p = P(|t| > 2|H_0) = \texttt{2*(1-pt(2,15))} = 0.063945.$$

Since $p > \alpha = 0.05$ we don't reject the null hypothesis.

Alternatively we could have done the problem in terms of rejection regions. We are given $\bar{x} = 11$, $s^2 = 4$, and $n = 16$. The null hypothesis is $\mu = 10$. Using $\bar{x}$ as our test statistic the rejection region is

$$(-\infty, 10 - t_{15,0.025}\frac{s}{\sqrt{n}}] \ \cup \ [10 + t_{15,0.025}\frac{s}{\sqrt{n}}, \infty) \ = \ (-\infty, 8.93] \ \cup \ [11.07, \infty)$$

Here $t_{15,0.025}$ means a *critical value*, i.e. the value with right tail probability 0.025: for $T \sim t(15)$ we have $P(t > t_{15,0.025}) = 0.025$.

Since 11 lies outside the rejection region, we should *not* reject the null-hypothesis.

(b) *What if we test against $H_A' : \mu > 10$?*

Solution: This is a one-sided alternative. The t-statistic is the same

$$\frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{1}{2/4} = 2.$$

So we have the one-sided p-value

$$p = P(t > 2|H_0) = \texttt{1-pt(2,15)} = 0.031973.$$

Since $p < \alpha = 0.05$ we reject the null hypothesis in favor of the alternative.

Again looking at rejection regions. We use the critical value $t_{15,0.05} \approx 1.753$. The rejection region for $\bar{x}$ is

$$[10 + t_{15,0.05}\frac{s}{\sqrt{n}}, \infty) \ = \ [10.876, \infty).$$

Since 11 lies inside the rejection region, we should reject the null-hypothesis in favor of $H_1 : \mu > 10$.

Problem 2. (40: 10,10,5,10,5 pts.)
Jerry took a JP Licks token and asked Jon to perform a test at significance level $\alpha = 0.05$

to investigate whether the coin is fair or biased toward tails (the side that says 'Token'). Jon recorded the following data

$$THTTHTTTTTTH$$

showing 3 heads and 9 tails.

Before Jon could compute the one-sided p-value for $H_0 : \theta = 0.5$ versus $H_A : \theta < 0.5$, he needed to take Aviva to the playground.

(a) *Erika believes that Jon's intention was to count the number of heads in twelve flips. Let's call this Experiment 1. Compute the rejection region and p-value. Sketch the null-distribution and rejection region. What does Erika conclude?*

Solution: Let x = number of heads

Model: $x \sim \text{binomial}(12, \theta)$.

Null distribution binomial$(12, 0.5)$.

Data: 3 heads in 12 tosses.

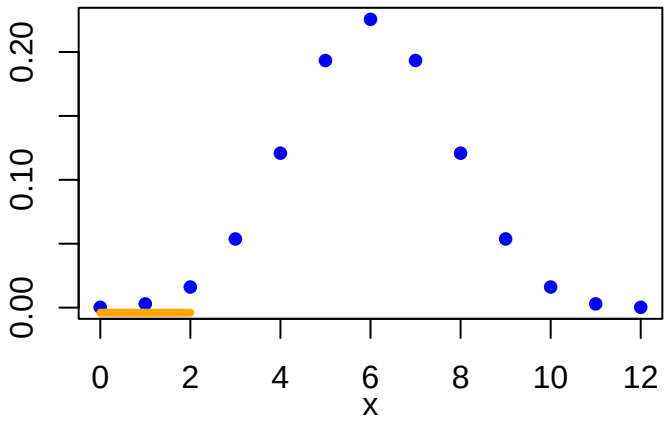
Since H_A is one-sided the rejection region is one-sided. Since H_A says that θ is small it predicts a small number of heads in 12 tosses. That is, we reject H_0 on a small number of heads.

So, rejection region = left tail of null distribution.

$$c_{0.95} = \texttt{qbinom(0.05, 12, 0.5) - 1} = 2$$

Rejection region is $0 \le x \le 2$.

$$p = \texttt{pbinom(3, 12, 0.5)} = 0.072998$$



Binomial(12,15) null distribution and rejection region $x \le 2$.

Erika concludes there is not enough evidence to reject the null hypothesis at the significance level 0.05.

(b) *Ruthi believes that Jon's intention was to stop after the third heads and report the number of tails, e.g., in the data the third head came on flip 12 so the number of tails is 9. Let's call this Experiment 2. Compute and sketch the corresponding null-distribution,*

rejection region, and p-value. What does Ruthi conclude? Hint: if a counting argument eludes you, google "negative binomial distribution".

The R functions `dnbinom, pnbinom, etc.` *might be helpful. For example,* `dnbinom(5, 3, 0.2)` *gives the probability of seeing 5 tails before the third head in a sequence of tosses of a coin with probability 0.2 of heads.*

Solution: Let n = number of tosses that were tails before the third that is heads
Probability model: Choose two tosses in the first n+2 for heads; the $n+3$rd toss must be heads:

$$p(n) = \binom{n+2}{2}(1-\theta)^n\theta^3.$$

This is called the negative binomial distribution with parameters 3 and θ.

Our data is: 9 tails to get 3 heads.

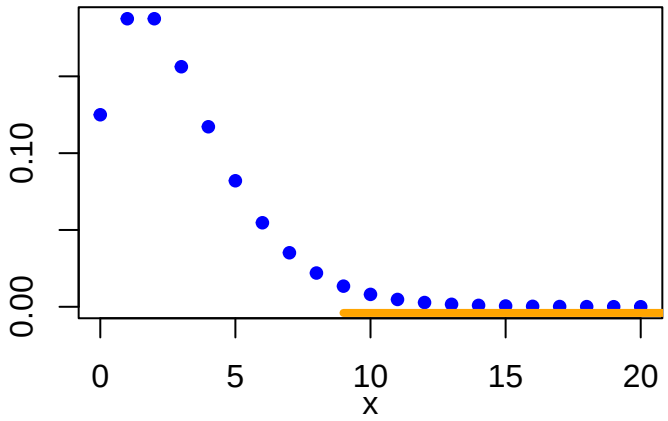
Since H_A is one-sided the rejection region is one-sided. Since H_A has θ small (< 0.5), it predicts a large number of tails before 3 heads. So we reject on a large number of tails.

Rejection region = right tail of null distribution.

$$c_{0.05} = \texttt{qnbinom(0.95, 3, 0.5) + 1} = 9$$

Rejection region is $n \geq 9$.

$$p = \texttt{1 - pnbinom(8, 3, 0.5)} = 0.032715$$



Negative binomial null distribution and rejection region

Ruthi rejects the null hypothesis in favor of H_A at significance level 0.05.

(c) *Jerry actually told Jon to count the number of heads in 100 flips (Experiment 3), so Jerry figures that Jon must have gotten bored and quit right after the 12^{th} flip. Strictly speaking, can Jerry compute a p-value from Jon's partial data? Why or why not?*

Solution: No. Computing a p-value requires that the experiment be fully specified ahead of time so that the definition of 'data at least as extreme' is clear.

Having said that, it's a shame to waste good data. You can still analyze the data for suggestive results. As long as you report everything honestly people can reach their own conclusions.

(d) *Let's reexamine the same data from the Bayesian perspective. What is the likelihood function in Experiment 1? What is the likelihood function in Experiment 2? How are these likelihood functions related? Given the prior Beta(n,m), find the posterior in each case. How are they related?*

Solution: Prior: Beta(n, m) has pdf $c\,\theta^{n-1}(1-\theta)^{m-1}$

Likelihood experiment 1: $\binom{12}{3}\theta^3(1-\theta)^9$

Likelihood experiment 2: $\binom{11}{2}\theta^3(1-\theta)^9$

Since the likelihoods are the same up to a constant factor the posterior has the same form

$$c\,\theta^{n+3-1}(1-\theta)^{m+9-1}$$

which is the pdf of a Beta$(n+3, m+9)$ distribution.

The two posteriors are identical. In the Bayesian framework the same data produces the same posterior.

(e) *Read* ***https://en.wikipedia.org/wiki/Likelihood_principle****, appreciate the voltmeter story, and summarize the main points we are getting at via the earlier parts of this problem regarding frequentist and Bayesian experiments.*

Solution: The main point is that in the frequentist framework the decision to reject or accept H_0 depends on the exact experimental design because it uses the probabilities of unseen data as well as those of the actually observed data.

Problem 3. (10 pts.) **(Chi-square for variance)**

The following data comes from a normal distribution which you suspect has variance equal to 1. You want to test this against the alternative that the variance is greater than 1.

```
 1.76,  -2.28,  -0.56,   1.46,  0.59,  1.26,
-1.94,  -0.79,  -0.86,  -1.41,  2.07,  1.30
```

There is a chi-square test for this. Look at
https://www.itl.nist.gov/div898/handbook/eda/section3/eda358.htm
and run the test with significance level 0.05. You can use R for the computations, but explain what you are doing and give the value of the test statistic, and the p-value.

Solution: We use the χ^2 statistic with hypotheses H_0: $\sigma^2 = 1$, H_A: $\sigma^2 > 1$.

So, we have $\sigma_0^2 = 1$ and s^2 = sample variance = `var(data)` = 2.34.

$n = 12$ = number of data points.

χ^2-statistic: $X^2 = (n-1)s^2/\sigma_0^2 = 25.74$

Since the alternative hypothesis is $\sigma > \sigma_0$, this is a right-sided test. The right-sided p-value is `p = 1 - pchisq(`X^2`, n-1) = 0.0071.`

Since $p < 0.05$ we reject the null hypothesis H_0 in favor of the alternative that $\sigma^2 > 1$.

Problem 4. (10 pts.) **(Chi-squared for categorical data)**

Jon and Jerry spent a fortune on dice and and bent coins for 18.05, so they decide to submit an invoice to the math department for reimbursement. The math department suspects that

their six figure expense report is made up, so they call you to test the data for fraud. You do some research and learn that accounting data should follow something called Benford's law. *This states that the relative frequency of the first digits of each entry should have the following distribution:*

First digit k	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
probability p(k)	*0.301*	*0.176*	*0.125*	*0.097*	*0.079*	*0.067*	*0.058*	*0.051*	*0.046*

You go back to the math department and tell them that the only data you need is the counts of all the first digits in their invoice. They are skeptical, but they know you have taken 18.05 and, so, must know what you are doing. They give you the following counts.

First digit k	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
count	*7*	*13*	*12*	*9*	*9*	*13*	*11*	*10*	*16*

The math department doesn't want to unjustly accuse Jon and Jerry, so they ask you to test at 0.001 significance level. Run a significance test to see how well this data fits Benford's distribution and make a recommendation to the math department.

See https://en.wikipedia.org/wiki/Benford%27s_law

Solution: We do a χ^2 test of goodness of fit comparing the observed counts with the counts expected from Benford's distribution.

You can use either test statistic

$$G = 2\sum O_i \ln\left(\frac{O_i}{E_i}\right).$$

or

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where O_i are the observed counts and E_i are the expected counts from Benford's distribution. The total count = 100.

First digit k	1	2	3	4	5	6	7	8	9
observed	7	13	12	9	9	13	11	10	16
expected	30.103	17.609	12.494	9.691	7.918	6.695	5.7992	5.1153	4.5757
X^2 components	17.731	1.206	0.200	0.049	0.148	5.939	4.664	4.665	28.523

The χ^2-statistics are $G = 56.3919$ and $X^2 = 62.6998$.
There are 9 cells that must sum to 100 so the degrees of freedom = 8.
The p-value using G is

$$p = P(\text{G test stat} > 56.3919 \,|\, H_0) = \texttt{1-pchisq(56.3919, 8)} = 2.4 \times 10^{-9}$$

The p-value using X^2

$$p = P(\text{X}^2 \text{ test stat} > 62.6998 \,|\, H_0) = \texttt{1-pchisq(62.6998, 8)} = 1.4 \times 10^{-10}$$

Since $p < \alpha$ we reject H_0 in favor of the notion that Jon and Jerry were trying to embezzle money.

Problem 5. (20: 10,10 pts.) **(Two-sample F test for equal variances)**
In this problem, we want you to become comfortable using the web to learn about new

tests and R commands. There are dozens of statistical software packages used by labs and industry, so being able to learn new commands is probably more important than trying to memorize them all. Look at the help file in R for ***`var.test`***

We used R to secretly create a vector x of 20 random samples from an $N(0,1)$ distribution and a vector y of 20 random samples from a $N(10,1)$ distribution. These are listed below. With a tiny bit of editing you should be able to copy and paste these into R.

```
-0.802,  0.457,  0.972,  0.044,  0.318, -1.380,  0.111,
-0.023, -0.700, -1.977, -0.497,  1.471, -1.314, -0.078,
-0.505,  0.583,  1.363, -1.863, -2.105,  0.488


 9.019,  9.852,  7.947,  9.465, 10.060, 10.508, 9.506,
 9.540, 10.218,  9.407, 11.455, 11.422,  7.698, 9.972,
10.928, 11.577, 10.376,  8.605,  9.347, 10.715
```

(a) *Temporarily erase from your mind the variances used in generating the data. We want to test the null hypothesis that the normal distributions which generated x and y have equal variances. Now use* ***`var.test`*** *to find the p-value for the test for equal variances. (You may need to go back and look at the documentation for* ***`var.test`*** *to see how to do this.)*

Solution: We let $x =$ the first set of 20 numbers and y the second. R makes it almost too easy. We give the command

`var.test(x,y)`.

R then prints out

```
      F test to compare two variances

data:  x and y
F = 0.97034, num df = 19, denom df = 19, p-value = 0.9484
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  0.3840737 2.4515249
sample estimates:
ratio of variances
     0.9703434
```

So, the p-value is 0.9484 with F-statistic 0.9703. With this p-value we do not reject the null hypothesis that the variances are the same.

(b) *You should look up the F-test for equal variances. Give the formula and compute the F statistic and p-value directly. You can use R to compute sample means, sample variances and do arithmetic as needed, but you can't use* `var.test` *to do all the work. You will need the R function* `pf` *to compute probabilities for the F distribution*

Should you reject the null-hypothesis at $\alpha = 0.05$?

Note: here is how you get a two-sided p-value for this test.
If the F-statistic is greater than 1 then

$$p = 2 \times \text{ the right tail probability}$$

If the F-statistic is less than 1 then

$$p = 2 \times \text{ the left tail probability}$$

Solution: We found the formula for the F statistic for this test at
https://en.wikipedia.org/wiki/F-test_of_equality_of_variances

$$s_x^2 = \texttt{var(x)} = 1.1302$$
$$s_y^2 = \texttt{var(y)} = 1.1647$$

Our F-statistic is

$$\texttt{fstat} = \frac{s_x^2}{s_y^2} = 0.9703$$

The degrees are freedom are both 19. We are running a two-sided test. So,

$$p = \texttt{2*min(pf(fstat, 19, 19), 1-pf(fstat, 19, 19))} = 0.9484$$

which matches our result in part (a).

Problem 6. (20: 10,10 pts.) **(One-way ANOVA)**
Read the abstract of the following paper:

https://www.sciencedirect.com/science/article/pii/S1090513813000226

Barnaby J. Dixson, Robert C. Brooks. The role of facial hair in women's perceptions of men's attractiveness, health, masculinity and parenting abilities. *Evolution and Human Behavior, Volume 34, Issue 3, May 2013, Pages 236-241*

Note that one of the authors may have a personal bias:

https://www.researchgate.net/profile/Barnaby_Dixson

For this problem you will need the F-test for equal means from the reading for class 19. For the purposes of the problem, we made a slight simplification to the experimental protocol and data.

(a) *The table below records the mean attractiveness rating for 351 heterosexual or bisexual women who rated the attractiveness of one male face of each type from 0 (very low) to 5 (very high). So we have four groups with 351 samples per group.*

facial hair state	*clean*	*5-day*	*10-day*	*full*
sample mean	*1.32*	*1.26*	*1.53*	*1.39*
sample variance	0.56	0.80	0.93	0.82

Run a one-way ANOVA (F-test) at $\alpha = 0.01$ for equal means "by hand". That is, compute the F-statistic and corresponding p value using the data in the table, and decide whether to reject the null hypothesis.

Solution: Let's specify the assumptions and hypotheses for this test.
We have 4 groups of data: Clean, 5-day, 10-day, full

Assumptions: Each group of data is drawn from a normal distribution with the same variance σ^2; all data is drawn independently.
H_0: the means of all the normal distributions are equal.
H_A: not all the means are equal.

The test compares the between group variance with the within group variance. Under the null hypothesis both are estimates of σ^2, so their ratio should be about 1. We'll reject H_0 if this ratio is far from 1.

We used R to do the computation. Here's the code.

```
mns = c(1.32, 1.26, 1.53, 1.39)
v = c(0.56, 0.80, 0.93, 0.82)
m = 351 # number of samples per group
n = length(mns) # number of groups
msb = m*var(mns) # between group variance
msw = mean(v) # within group variance
fstat = msb/msw
df1 = n-1;
df2 = n*(m-1)
p = 1 - pf(fstat, df1,df2)

print(fstat)
print(p)
```

This produced an F-statistic of 6.09 and $p = 0.00041$. Since the p-value is much smaller than 0.05 we reject H_0.

(b) *Following up, use 3 two-sample t-tests to investigate the hypothesis that the mean for 10-day is higher than the each of the others. For each test use significance level $\alpha = 0.01$. (The two-sample t-test is described in the reading for class 19.)*

If exactly one of these tests resulted in a rejection, would it be appropriate to conclude that we have rejected the null hypothesis in part (a) at significance level $\alpha = 0.01$?

Solution: We compare 10-day beards with each of the others. In each case we have: H_0: the means are the same
H_A: the 10-day mean is greater than the other mean.

Note carefully that this is a one-sided test while the F-test in part (b) is a two-sided test.

From the class 19 reading we have the t-statistic for two samples. Since both samples have the same size $m = 351$ the formula looks a little simpler.

$$t = \frac{\bar{x} - \bar{y}}{s_P},$$

where the pooled sample variance is

$$s_P^2 = \frac{s_x^2 + s_y^2}{m}$$

Note: the test assumes equal variances which we should verify in each case. This raises the issue of multiple tests from the same data, but it is legitimate to do this as exploratory analysis which merely suggests directions for further study.

The following table shows the one-sided, 2-sample t-test comparing the mean of the 10-day growth against the other three states.

	t-stat	one-sided p-value	F-stat
clean	3.22	0.00066	10.39
5-day	3.85	0.00007	14.79
full	1.98	0.0239	3.93

We also give the F-statistic for the two samples. You can check that the F-statistic for two-samples is just the square of the t-statistic.

If one of the tests has significance level less than 0.01, it is not proper to reject the null hypothesis. This is because the significance level of the entire experiment is greater than 0.01. That is, since we ran 3 tests each with probability 0.01 of a type I error the total probability of type I error is greater than 0.01 –it will be close to 0.03.

We reiterate that with multiple testing the true significance level of the test is larger than the significance level for each individual test.

18.05 Introduction to Probability and Statistics
Spring 2022