Math 490 HW #13

Maxwell Levin

March 16, 2018

Question 1.

Test whether the means of the two random samples of the 2001 and 2002 real estate sales data given by tables 16.1 and 16.5 are significantly different.

a. State the null and alternative hypotheses.

Our null hypothesis is

$$H_0 : \mu_0 = \mu_1,$$

where μ_0 is our sample mean for the 2001 data and μ_1 is our sample mean for the 2002 data.

Our alternative hypothesis is

$$H_1 : \mu_0 \neq \mu_1.$$

b. Perform a two-sample t-test. What is the P-value?

In order to compute the test statistic,

$$\text{t-stat} = \frac{\overline{Y}_{2001} - \overline{Y}_{2002}}{\sqrt{\frac{S_{2001}^2}{n_{2001}} + \frac{S_{2002}^2}{n_{2002}}}},$$

we first need to compute $\overline{Y}$, S, and n for the 2001 and 2002 samples.

To do this we run the following code in R after reading in our tables:

```r
y2001 = mean(price2001)
y2002 = mean(price2002)

s2001 = sd(price2001)
s2002 = sd(price2002)

n2001 = length(price2001)
n2002 = length(price2002)
```

Running this we see that

$$\overline{Y}_{2001} = 288.9265,$$
$$\overline{Y}_{2002} = 329.2571,$$
$$S_{2001} = 157.7778,$$
$$S_{2002} = 316.83,$$
$$n_{2001} = n_{2002} = 50.$$

We plug these numbers into our test-statistic formula to get

$$t - stat = \frac{288.9265 - 329.2571}{\sqrt{\frac{(157.7778)^2}{50} + \frac{(316.83)^2}{50}}} \approx -0.80573.$$

We use the liberal degree of freedom equation to get $df = 50 + 50 - 2 = 98$. We then use R to calculate our P-value:

```
pt(-0.80573, 98) + 1 - pt(0.80573, 98)
```

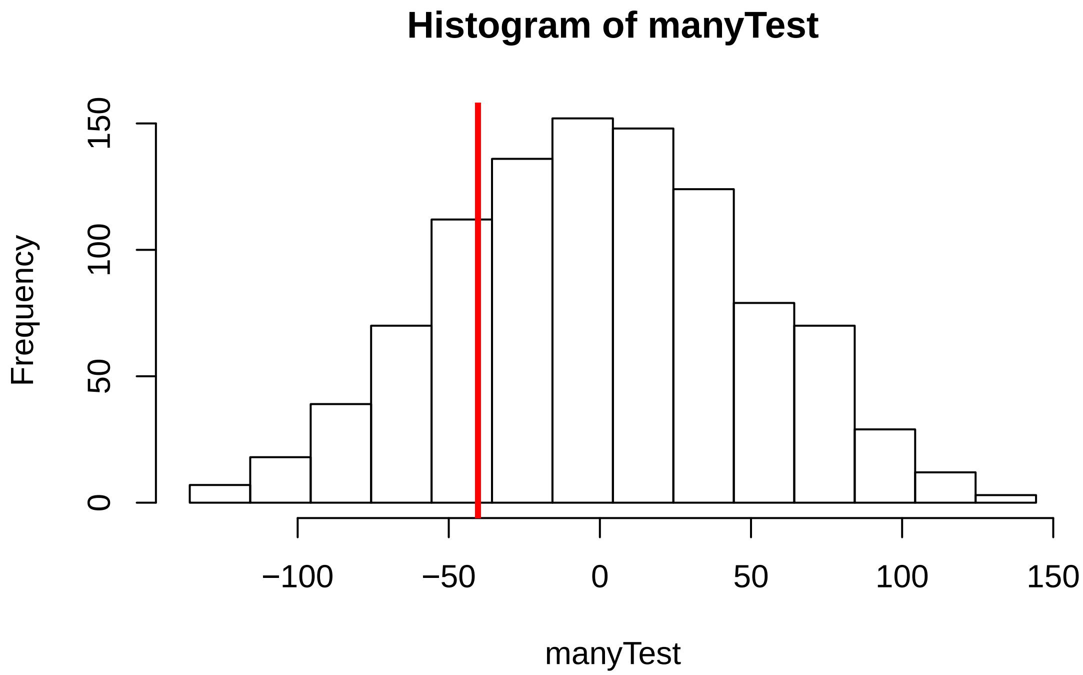
```
[1] 0.4223493
```

Thus we see that our P-value is about 0.4223493, which is more than a 5% α level, so we do not have enough evidence to reject the null hypothesis.

c. Perform a permutation test on the difference in means. What is the P-value? Compare it with the P-value you found in part (b). What do you conclude based on the tests?

We run the following code in R;

```
price = c(price2001, price2002)
year = rep(c(2001, 2002), c(50, 50))
yearAndPrice = data.frame(year, price)

obsDiff = mean(price[year == 2001]) - mean(price[year == 2002])

oneTest = function(x, y) {# x is the category label like year
                           # y is the response variable like price
    xPerm = sample(x);     # a random permuation of label x
    mean(y[xPerm == 2001]) - mean(y[xPerm == 2002]);
                           # categories are 2001 and 2002 in our data
}

manyTest = replicate(999, oneTest(year, price))

hist(manyTest, breaks=seq(min(manyTest), max(manyTest) + 20, 20))

abline(v = obsDiff, lwd = 3, col="red")
```



We get the P-value from this simulation by

```
print(mean(abs(manyTest) > abs(obsDiff)))
```

```
[1] 0.4474474
```

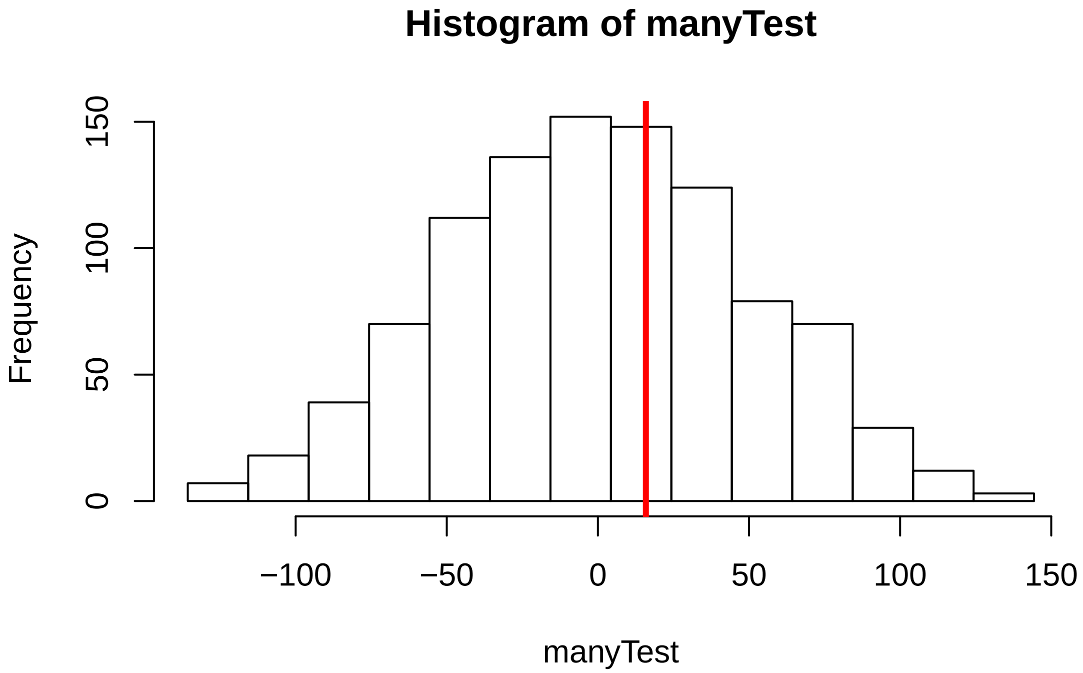
Thus we see that our P-value is about 0.43. This is way more than 0.05, which is our default α level, so this suggests that cannot reject the null hypothesis. The P-value from our permutation sample is slightly larger than our P-value from part (b), but both are well above the threshold of 5%, so our tests suggest that the means of the two random samples of the 2001 and 2002 real estate sales data given by tables 16.1 and 16.5 are not significantly different.

Question 2.

Because distributions of real estate prices are typically strongly skewed, we often prefer the median to the mean as a measure of center. We would like to test the null hypothesis that Seattle real estate prices in 2001 and 2002 have equal medians. Carry out a permutation test for the difference in medians, find the P-value, and explain what the P-value tells us.

To run a permutation test for the difference in medians we can run the following code in R:

```
price = c(price2001, price2002)
year = rep(c(2001, 2002), c(50, 50))
yearAndPrice = data.frame(year, price)

obsDiff = median(price[year == 2001]) - median(price[year == 2002])

oneTest = function(x, y) {# x is the category label like year
                          # y is the response variable like price
    xPerm = sample(x);    # a random permuation of label x
    median(y[xPerm == 2001]) - median(y[xPerm == 2002]);
                                    # categories are 2001 and 2002 in our data
}

medianTest = replicate(999, oneTest(year, price))

hist(manyTest, breaks=seq(min(manyTest), max(manyTest) + 20, 20))

abline(v = obsDiff, lwd = 3, col="red")
```



We can get the P-value of our simulation by;

```
[1] 0.6136136
```

Thus we see that our P-value is around 0.6, which is way more than our specified level of 0.05, and still a good portion larger than our P-values in Question 1. Such a high P-value tells us that we do not have sufficient evidence to reject the null hypothesis that medians of the two random samples of the 2001 and 2002 real estate sales data given by tables 16.1 and 16.5 are not significantly different.