

# STAT20 Homework #8

Seth Metcalf

## Table of Contents

Introduction.....	1
Chapter 26 .....	1
1. Loosely based on Ch 26 F4:.....	1
2. Ch 26 Rev 8: .....	3
3. Ch 26 Rev 11:.....	3
Chapter 27 .....	4
4. Ch 27 A6: .....	4
5. Ch 27 B3: .....	5
6. Ch 27 C2:.....	5
7. Ch 27 D4: .....	6
8. Ch 27 Rev 3: .....	7
9. Ch 27 Rev 4: .....	8
10. Ch 27 Rev 7:.....	9

## Introduction

This is Homework #8, which contains questions from Chapters 26 & 27. *Due 9 April 2021.*

## Chapter 26

### 1. Loosely based on Ch 26 F4:

A scale is calibrated using a weight that is known to be exactly 1 kg. Each data set below represents repeated measurements of the weight. Assume the Gauss model (measurement = true weight of 1 kg + measurement error + bias), with measurement errors following the normal curve. Since the weight actually weighs 1 kg, the null hypothesis says that the expected value should be 1 kg (no bias). For each dataset below, make a t-test to see whether the scale is properly calibrated or not. In one case, this is impossible. Which one, and why? The data is listed as the amount above or below 1kg, and the units are micrograms.

a) 1, -2, 9

```
aWeight = c(1, -2, 9)
t.test(aWeight, mu = 1)

##
##  One Sample t-test
##
## data:  aWeight
## t = 0.50767, df = 2, p-value = 0.6621
## alternative hypothesis: true mean is not equal to 1
## 95 percent confidence interval:
##  -11.45874 16.79207
## sample estimates:
## mean of x
## 2.666667
```

In this scenario, it is still possible for the scale to be properly calibrated. 95% confidence interval includes the mean value that it should have.

b) 1, -2, 9, 14, 8, 15, -1

```
bWeight = c(1, -2, 9, 14, 8, 15, -1)
t.test(bWeight, mu = 1)

##
##  One Sample t-test
##
## data:  bWeight
## t = 1.993, df = 6, p-value = 0.09333
## alternative hypothesis: true mean is not equal to 1
## 95 percent confidence interval:
##  -0.2039156 12.7753441
## sample estimates:
## mean of x
## 6.285714
```

In this scenario, it is still possible for the scale to be properly calibrated, however it is much less likely. 95% confidence interval includes the mean value that it should have.

c) 1

It is impossible to commit a t-test with this data as there is no way to construct a 95% interval with only 1 data value.

d) 1, 14

```
dWeight = c(1, 14)
t.test(dWeight, mu = 1)

##
##  One Sample t-test
##
## data:  dWeight
```

```
## t = 1, df = 1, p-value = 0.5
## alternative hypothesis: true mean is not equal to 1
## 95 percent confidence interval:
## -75.09033 90.09033
## sample estimates:
## mean of x
## 7.5
```

In this scenario, it is still possible for the scale to be properly calibrated, but there should be more samples taken. 95% confidence interval includes the mean value that it should have.

## 2. Ch 26 Rev 8:

Bookstores like education, one reason being that educated people are more likely to spend money on books. National data show the nationwide average educational level to be 13 years of schooling completed, with an SD of about 3 years, for persons age 18 and over. A bookstore is doing a market survey in a certain county, and takes a simple random sample of 1,000 people age 18 and over. They find the average educational level to be 14 years, and the SD is 5 years. Can the difference in average educational level between the sample and the nation be explained by chance variation? If not, what other explanation can you give?

### Explanation

```
educationSample = 14
nationSD = 3
educationDraws = 1000
educationSE_sum = sqrt(educationDraws) * nationSD # = 94.86833
educationSE_avg = educationSE_sum/educationDraws # = 0.09486833

nationMean = 13
educationZ = (educationSample - nationMean)/educationSE_avg # 10.54093

(1 - pnorm(10.54093)) * 100

## [1] 0
```

In conclusion, this result is so close to 0 that it can not be due to chance and in this scenario the null hypothesis is rejected. However an alternate explanation would be that the area that is chosen has a higher average education than the rest of the country and there are other areas with lower average education to even it out.

## 3. Ch 26 Rev 11:

According to the census, the median household income in Atlanta (1.5 million households) was \$52,000 in 1999. In June 2003, a market research organization takes a simple random sample of 750 households in Atlanta; 56% of the sample households had incomes over \$52,000. Did median household income in Atlanta increase over the period 1999 to 2003?

a) Formulate null and alternative hypotheses in terms of a box model. (It's a little tricky, think about what percent you should expect the 56% to be if the median household income didn't change.)

null = The percentage drawn is due to chance and median household income has not changed since 1999.

alt = Median household income has increased to over \$52,000 since 1999.

b) Calculate the appropriate test statistic and P.

```
atlantaObserved = 750*.56 # = 420
atlantaDraws = 750

atlantaEV_sum = 750*.50 # = 375
atlantaSD = (1-0) * sqrt(0.5 * 0.5) # = 0.5
atlantaSE_sum = sqrt(atlantaDraws) * atlantaSD # = 13.69306

atlantaZ = (atlantaObserved - atlantaEV_sum)/atlantaSE_sum # = 3.286335

(1 - pnorm(atlantaZ)) * 100

## [1] 0.05075005
```

c) Did median family income go up?

Since the p-value is quite small this number is highly significant, meaning that we reject the null hypothesis and it is most likely not due to chance. Meaning that the average value income in a home has likely gone up from \$52,000 since 1999.

## Chapter 27

### 4. Ch 27 A6:

One hundred draws are made at random with replacement from box F: the average of these draws is 51 and their SD is 3. Independently, 400 draws are made at random with replacement from box G: the average of these draws is 48 and their SD is 8. Someone claims that both boxes have the same average. What do you think? You don't need to do a formal hypothesis test, but think about it in that way.

a) What do you think?

Given the number of draws and the average and SD gotten from these draws I find it unlikely that the means would be the same. Without doing a formal hypothesis test we can find that just by looking at the SEs between the two

```
boxfSE_sum = sqrt(100) * 3 # = 30
boxfSE_avg = boxfSE_sum/100 # = 0.3
```

```
boxgSE_sum = sqrt(400) * 8 # = 160
boxgSE_avg = boxgSE_sum/400 # = 0.4
```

That it is quite unlikely to have a difference in the average draw sum of 3 between the two boxes meaning that it is quite unlikely that they have the same mean.

### 5. Ch 27 B3:

In 1970, 59% of college freshmen thought that capital punishment should be abolished; by 2005, the percentage had dropped to 35%. Is the difference real, or can it be explained by chance? You may assume that the percentages are based on two independent simple random samples, each of size 1,000. For this one, do a formal hypothesis test.

#### Hypothesis test

```
percent1970 = 0.59
percent2005 = 0.35
collegeSample = 1000

SD_1970 = (1 - 0) * sqrt(0.59 * 0.41) # = 0.4918333
SE_1970_sum = sqrt(collegeSample) * SD_1970 # = 15.55313
SE_1970_avg = (SE_1970_sum/collegeSample) * 100 # = 1.555313%

SD_2005 = (1 - 0) * sqrt(0.35 * 0.65) # = 0.4769696
SE_2005_sum = sqrt(collegeSample) * SD_2005 # = 15.0831
SE_2005_avg = (SE_2005_sum/collegeSample) * 100 # = 1.50831%

SE_1970_2005_diff = sqrt(SE_1970_avg^2 * SE_2005_avg^2) # = 2.345895

collegeMean = 0
collegeDiff = (percent1970 - percent2005) * 100 # = 24%
collegeZ = (collegeDiff - collegeMean)/SE_1970_2005_diff # = 10.23064

(1 - pnorm(10.23064)) * 100

## [1] 0
```

Since the p-value is about 0 it is enough evidence to reject the null hypothesis. This difference in results is not due to chance. The number of college students that want to abolish capital punishment has decreased since 1970.

### 6. Ch 27 C2:

(Hypothetical.) Is Wheaties a power breakfast? A study is done in an elementary statistics class; 499 students agree to participate. After the midterm, 250 are randomized to the treatment group, and 249 to the control group. The treatment group is fed Wheaties for breakfast 7 days a week. The control group gets Sugar Pops.

a) Final scores averaged 66 for the treatment group; the SD was 21. For the control group, the figures were 59 and 20. What do you conclude?

```
wheaties_avg = 66
wheaties_SD = 21
wheaties_size = 250
wheaties_SE_sum = sqrt(wheaties_size) * wheaties_SD # = 332.0392
wheaties_SE_avg = wheaties_SE_sum/wheaties_size # = 1.328157

sugarpops_avg = 59
sugarpops_SD = 20
sugarpops_size = 249
sugarpops_SE_sum = sqrt(sugarpops_size) * sugarpops_SD # = 315.5947
sugarpops_SE_avg = sugarpops_SE_sum/sugarpops_size # = 1.267449

SE_wheaties_sugarpops_diff = sqrt(wheaties_SE_avg^2 + sugarpops_SE_avg^2) # =
1.835872
elementaryDiff = (wheaties_avg - sugarpops_avg) # 7
elementaryZ = (elementaryDiff - 0)/SE_wheaties_sugarpops_diff # = 3.812902

(1 - pnorm(elementaryZ)) * 100

## [1] 0.006867225
```

b) What aspects of the study could have been done “blind?”

The blindness could have been done by not letting the patients know which group they were in (treatment/control). But, this would be extremely difficult as you can not control which of the students recognized the specific cereal. However, the person that was in charge of receiving this data could have been blind to affect it even less.

## 7. Ch 27 D4:

Many observational studies conclude that low-fat diets protect against cancer and cardiovascular “events” (heart attacks, stroke, and so forth). Experimental results, however, are generally negative. In 2006, the Women’s Health Initiative (WHI) published its results. This was a large-scale randomized trial on women who had reached menopause. As one part of the study, 48,835 women were randomized: 19,541 were assigned to the treatment group and put on a low-fat diet. The other 29,294 women were assigned to the control group and ate as they normally would. Subjects were followed for 8 years. Among other things, the investigators found that 1,357 women on the low-fat diet experienced at least one cardiovascular event, compared to 2,088 in the control group. Can the difference between the two groups be explained by chance? What do you conclude about the effect of the low-fat diet?

### Conclusion

```
menopauseT_sum = 1357
menopauseT_size = 19541
menopauseT_percent = menopauseT_sum/menopauseT_size # 0.06944373
menopauseT_SD = (1 - 0) * sqrt(menopauseT_percent * (1 - menopauseT_percent))
```

```
# 0.2542072
menopauseT_SE_sum = sqrt(menopauseT_size) * menopauseT_SD # = 35.5354
menopauseT_SE_avg = menopauseT_SE_sum/menopauseT_size # = 0.001818505

menopauseC_sum = 2088
menopauseC_size = 29294
menopauseC_percent = menopauseC_sum/menopauseC_size # 0.07127739
menopauseC_SD = (1 - 0) * sqrt(menopauseC_percent * (1 - menopauseC_percent))
# 0.2572876
menopauseC_SE_sum = sqrt(menopauseC_size) * menopauseC_SD # = 44.03604
menopauseC_SE_avg = menopauseC_SE_sum/menopauseC_size # = 0.001503244

SE_menopauseT_menopauseC_diff = sqrt(menopauseT_SE_avg^2 +
menopauseC_SE_avg^2) * 100 # = 0.2359386
womenDiff = (menopauseT_percent - menopauseC_percent) * 100 # -0.1833661
womenZ = (womenDiff - 0)/SE_menopauseT_menopauseC_diff # = -0.007771771

pnorm(womenZ) * 100

## [1] 21.85271
```

Yes, it is most likely to be due to just chance. In this scenario you do not reject the null hypothesis and having a low-fat diet does not have a significant effect on cardiovascular diseases.

### 8. Ch 27 Rev 3:

The Gallup poll asks respondents how they would rate the honesty and ethical standards of people in different fields-very high, high, average, low, or very low. The percentage who rated clergy “very high or high” dropped from 60% in 2000 to 54% in 2005. This may have been due to scandals involving sex abuse; or it may have been chance variation. (You may assume that in each year, the results are based on independent simple random samples of 1,000 persons in each year.)

a) Should you make a one-sample z-test or a two-sample z-test? Why?

Two-sample z test to compare the different percentage rates depending on the years especially because it is given that they are independent.

b) Formulate the null and alternative hypotheses in terms of a box model. Do you need one box or two? Why? How many tickets go into each box? How many draws? What do the tickets show? What do the null and alternative hypotheses say about the box(es)?

null = The number of people who rate clergy as very high or high has not changed from 2000 to 2005.

alt = The number of people who rate clergy as very high or high has dropped from 2000 to 2005.

size = The total number of respondents, unknown.

draws = 1000 for each year

c) Can the difference between 60% and 54% be explained as a chance variation? Or was it the scandals? Or something else?

```
clergy2000_avg = 0.60
clergy2000_size = 1000
clergy2000_SD = (1 - 0) * sqrt(clergy2000_avg * (1 - clergy2000_avg)) #
0.4898979
clergy2000_SE_sum = sqrt(clergy2000_size) * clergy2000_SD # = 15.49193
clergy2000_SE_avg = clergy2000_SE_sum/clergy2000_size # = 0.01549193

clergy2005_avg = 0.54
clergy2005_size = 1000
clergy2005_SD = (1 - 0) * sqrt(clergy2005_avg * (1 - clergy2005_avg)) #
0.4983974
clergy2005_SE_sum = sqrt(clergy2005_size) * clergy2005_SD # = 15.76071
clergy2005_SE_avg = clergy2005_SE_sum/clergy2005_size # = 0.01576071

SE_clergy2000_clergy2005_diff = sqrt(clergy2000_SE_avg^2 +
clergy2005_SE_avg^2) * 100 # = 2.209977
clergyDiff = (clergy2000_avg - clergy2005_avg) * 100 # 6
clergyZ = (clergyDiff - 0)/SE_clergy2000_clergy2005_diff # = 2.71496

(1 - pnorm(clergyZ)) * 100

## [1] 0.331419
```

It is unlikely to be due to chance variation. The number of people that think of clergys as very high or high has dropped since the year 2000. In this scenario we should reject the null hypothesis.

#### 9. Ch 27 Rev 4:

This continues the previous exercise. In 2005, 65% of the respondents gave medical doctors a rating of “very high or high,” compared to a 67% rating for druggists. Is the difference real, or chance variation? Or do you need more information to decide? If the difference is real, how would you explain it? Discuss briefly. You may assume that the results are based on a simple random sample of 1,000 persons taken in 2005; each respondent rated clergy, medical doctors, druggists, and many other professions.

#### Discuss Briefly

```
doctor2005_avg = 0.65
doctor2005_size = 1000
doctor2005_SD = (1 - 0) * sqrt(doctor2005_avg * (1 - doctor2005_avg)) #
0.4769696
doctor2005_SE_sum = sqrt(doctor2005_size) * doctor2005_SD # = 15.0831
doctor2005_SE_avg = doctor2005_SE_sum/doctor2005_size # = 0.0150831

druggist2005_avg = 0.67
```



```

druggist2005_size = 1000
druggist2005_SD = (1 - 0) * sqrt(druggist2005_avg * (1 - druggist2005_avg)) #
0.4702127
druggist2005_SE_sum = sqrt(druggist2005_size) * druggist2005_SD # = 14.86943
druggist2005_SE_avg = druggist2005_SE_sum/druggist2005_size # = 0.01486943

SE_doctor2005_druggist2005_diff = sqrt(doctor2005_SE_avg^2 +
druggist2005_SE_avg^2) * 100 # = 0.02242772
dDiff = (doctor2005_avg - druggist2005_avg) * 100 # -2
dZ = (dDiff - 0)/SE_doctor2005_druggist2005_diff # = -0.8917537

pnorm(dZ) * 100

## [1] 17.25135

```

This is most certainly due to chance variation. There is no reason to reject the null hypothesis in this scenario as it is quite a high probability to occur. The rates between the two are most likely the same to one another but it was just a bad sample.

## 10. Ch 27 Rev 7:

When convicts are released from prison, they often return to crime and are arrested again (recidivism). The Department of Labor ran a randomized controlled experiment to find out if providing income support to ex-convicts during the first months after their release reduces recidivism. The experiment was done on a group of convicts being released from prisons in Georgia. Income support was provided for the treatment group, like unemployment insurance, and the control group received no payment.

a) 592 prisoners were assigned to the treatment group, and of them 48.3% were rearrested within a year of release. 154 were assigned to the control group, and of them 49.4% were rearrested within a year of release. Did income support reduce recidivism? Answer yes or no, and explain briefly.

```

prisonerT_avg = 0.483
prisonerT_size = 592
prisonerT_SD = (1 - 0) * sqrt(prisonerT_avg * (1 - prisonerT_avg)) #
0.4997109
prisonerT_SE_sum = sqrt(prisonerT_size) * prisonerT_SD # = 12.15849
prisonerT_SE_avg = (prisonerT_SE_sum/prisonerT_size) * 100 # = 2.053799

prisonerC_avg = 0.494
prisonerC_size = 154
prisonerC_SD = (1 - 0) * sqrt(prisonerC_avg * (1 - prisonerC_avg)) # 0.499964
prisonerC_SE_sum = sqrt(prisonerC_size) * prisonerC_SD # = 6.20439
prisonerC_SE_avg = (prisonerC_SE_sum/prisonerC_size) * 100 # = 4.028825

SE_prisonerT_prisonerC_diff = sqrt((2.053799 * 2.053799) + (4.028825 *
4.028825)) # = 4.522115
prisonerDiff = (prisonerT_avg - prisonerC_avg) * 100 # -1.1
prisonerZ = (prisonerDiff - 0)/SE_prisonerT_prisonerC_diff # = -0.243249

```

```
(1 - pnorm(dZ)) * 100
```

```
## [1] 82.74865
```

In this scenario the p-value is not small and allows for the assumption for it to be due to chance error. There is no reason to reject the null hypothesis. It also implies that income support did not reduce recidivism.

b) In the first year after their release from prison, those assigned to the treatment group averaged 16.8 weeks of paid work; the SD was 15.9 weeks. For those assigned to the control group, the average was 24.3 weeks; the SD was 17.3 weeks. Did income support reduce the amount that the ex-convicts worked? Answer yes or no, and explain briefly.

```
workprisonerT_avg = 16.8
```

```
workprisonerT_size = 592
```

```
workprisonerT_SD = 15.9
```

```
workprisonerT_SE_sum = sqrt(workprisonerT_size) * workprisonerT_SD # =  
386.8637
```

```
workprisonerT_SE_avg = (workprisonerT_SE_sum/workprisonerT_size) # =  
0.653486%
```

```
workprisonerC_avg = 24.3
```

```
workprisonerC_size = 154
```

```
workprisonerC_SD = 17.3
```

```
workprisonerC_SE_sum = sqrt(workprisonerC_size) * workprisonerC_SD # =  
214.6874
```

```
workprisonerC_SE_avg = (workprisonerC_SE_sum/workprisonerC_size) # = 0.653486
```

```
SE_workprisonerT_workprisonerC_diff = sqrt((workprisonerT_SE_avg^2) +  
(workprisonerC_SE_avg^2)) # = 1.539638
```

```
workprisonerDiff = (workprisonerT_avg - workprisonerC_avg) # -7.5
```

```
workprisonerZ = (workprisonerDiff - 0)/SE_workprisonerT_workprisonerC_diff #  
= -4.871275
```

```
pnorm(workprisonerZ) * 100
```

```
## [1] 5.54403e-05
```

In this scenario the p-value is very small (about 0). This implies that we reject the null hypothesis and that it is not due to chance error meaning that this supports that income support reduced the amount that the ex-convicts worked.