Name: SOLUTIONS **Midterm (take home version)**

To help you budget your time, questions are marked with *s. One * indicates a straight forward question testing foundational knowledge. Two ** indicate a more challenging question requiring use of several key concepts. Three *** indicate a challenging question requiring a clever use of key concepts and/or techniques not covered in class yet.

If you can't solve a questions completely, solve it as far as you can. If you require a laptop to get your final solution, describe in detail what you would do to solve the problem.

***1)** A twenty-sided die is a icosahedron numbered such that the numbers 1, 2, 3, …, 20 show up with equal probability. Suppose you roll ONE regular six-sided die and TEN twenty-sided dice. Assuming the outcomes of the dice are independent, what is the probability that all eleven dice land on the number 6? (Solve for the number as a decimal or fraction and show all your work.)

Solution

P(all land on 6) = P(ten d20 land on 6) * P(one d6 lands on 6)

= P(one d20 land on 6)^{10} * P(one d6 lands on 6)

= (1/20)^{10} * (1/6)

= 1.63 * 10^{-14}

***2)** Suppose you roll ONE regular six-sided die and TEN twenty-sided dice. Let X = the sum of the eleven dice. Assuming the outcomes of the dice are independent, what is the expectation of X, i.e. E[X].? (Solve for the number as a decimal or fraction and show all your work.)

Solution

E[sum of all eleven dice] = E[sum of ten d20] + E[one d6]
= 10 * E[one d20] + E[one d6]

Notice:
E[one d6] = 1*(1/6) + 2*(1/6) + 3*(1/6) + 4*(1/6) + 5*(1/6) + 6*(1/6)
= (1/6) * (1+2+3+4+5+6)
= (1/6) * (7*6/2) by sum of sequence from 1 to n = (n+1)*n/2. This shortcut step isn't essential.
= (1/6) * (21)
= 3.5
E[one d20] = 1*(1/20) + 2*(1/20) + 3*(1/20) + … + 20*(1/20)
= (1/20) * (1+2+3+…+20)
= (1/20) * (21*20/2)
= (1/20) * 210
= 10.5

Thus,
E[sum of all eleven dice] = 10 * 10.5 + 3.5
= 108.5.

***3)** Suppose you roll ONE regular six-sided die and TEN twenty-sided dice. You randomly pick out one of the dice and observe it had landed on a five. What is the probability you had selected the regular six-sided die? (Solve for the number as a decimal or fraction and show all your work.)

Solution

We want P(picked d6 | landed on a 5).

We know:

P(picked d6) = 1/11

P(landed on 5 | d6) = 1/6

P(picked d20) = 10/11

P(landed on 5 | d20) = 1/20.

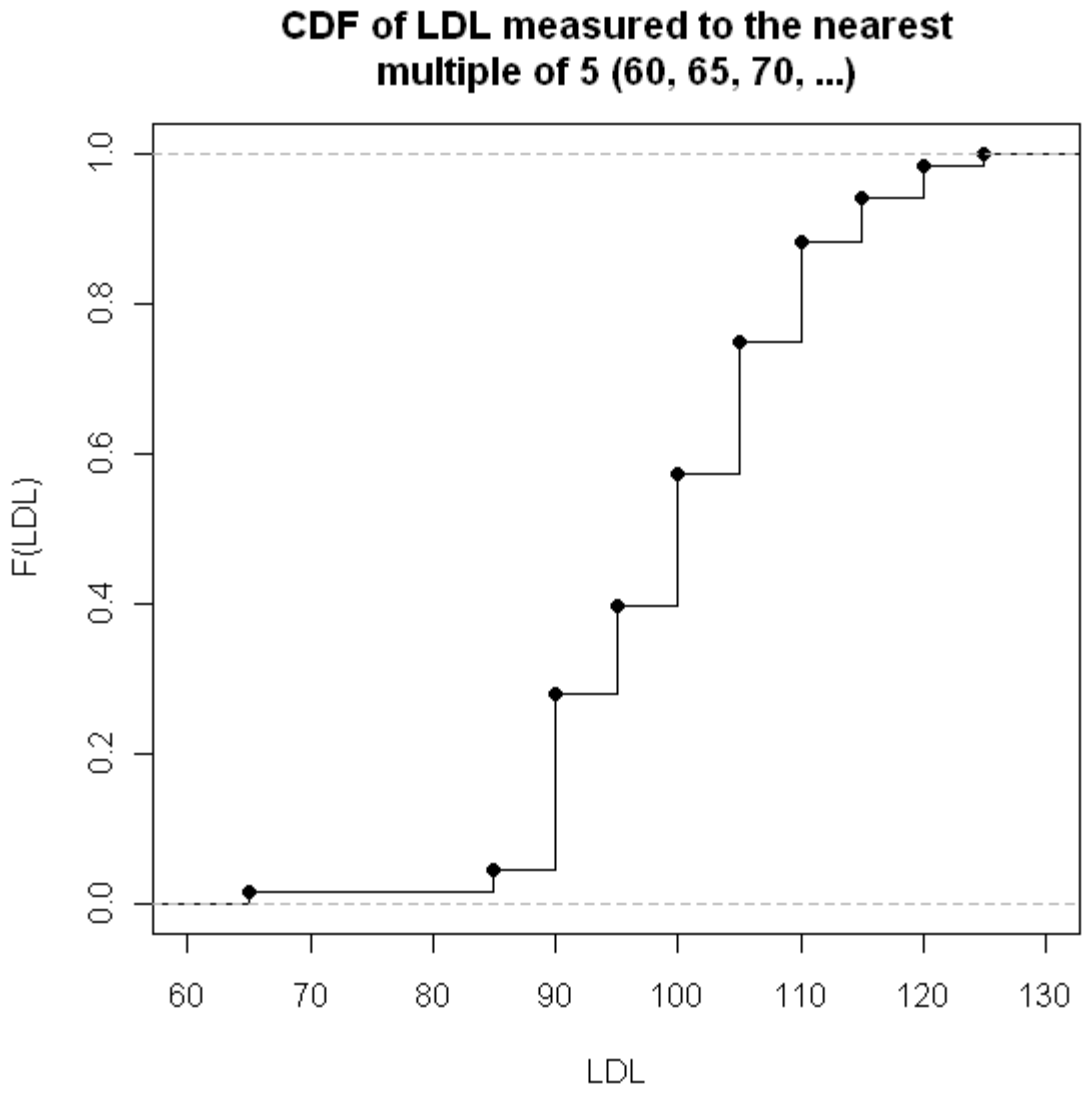
This is a perfect setup to use Bayes Theorem.

$$P(\text{picked d6} \mid \text{landed on a 5}) = \frac{P(\text{landed on 5} \mid \text{d6}) * P(\text{picked d6})}{P(\text{landed on 5} \mid \text{d6}) * P(\text{picked d6}) + P(\text{landed on 5} \mid \text{d20}) * P(\text{picked d20})}$$

$= (1/6)*(1/11) / [(1/6)*(1/11) + (1/20)*(10/11)]$

$= 1/4$

$= 0.25.$



***4)** Above is a plot of the cumulative distribution function, CDF, of low density lipoprotein, LDL, for a sample of healthy adults. LDL was measured to the nearest multiple of 5, e.g. 60, 65, 70, 75, …, 130. Based on this graph, roughly what proportion of the sample had an LDL > 110? (Provide the number or explain why the problem can't be solved from this graph.)

Solution
From the CDF we can read P(LDL $\leq$ 110) = roughly 0.90. (Anywhere from 0.85 to 0.90 is close enough.)
Thus, P(LDL > 110) = 1- 0.90 = 0.10. (Anywhere from 0.10 - 0.15 is close enough.)

***5)** What was the mode of this sample, i.e. what was the most common value observed? (Provide the number or explain why the problem can't be solved from this graph.)

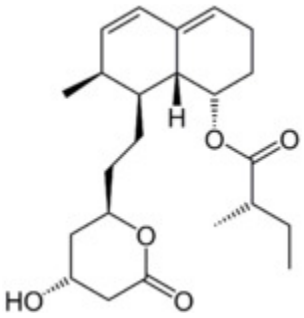
Solution
Spikes on the CDF show where there are lots of data. The biggest spike is at 90.

***6)** Researchers conducted a study of the impact of eating Cheerios for breakfast on LDL. 78 of the 96 participants were observed to have lower LDL after three months of the Cheerios breakfast diet. Based on this study, provide a 95% confidence interval for the probability that the Cheerios breakfast diet will reduce LDL. (Solve for the upper and lower bounds as decimals or fractions and show all your work.)

Solution

The 95% Wilson interval is formed by adding two successes and two failures to the data set and using the Wald interval formula on the updated data.

$(80/100) \pm 1.96 * \text{sqrt}((80/100)*(20/100)/100)$
$0.8 \pm 1.96 * \text{sqrt}(0.0016)$
$0.8 \pm 1.96 * 0.04$
0.8 ± 0.0784
$(0.7216, 0.8784)$



Mevastatin

***7)** Suppose that in spite of its side effects, Mevastatin is known to reduce LDL in 88.5% of those taking it. **Write the conclusion** for a two-sided, 5% significance level hypothesis test of whether the Cheerios for breakfast diet in problem 6 performs differently from Mevastatin. (Write your answer in plain English and include an appropriate interpretation of what it means to reject or not reject the null hypothesis in this case. You may choose to use your solution in problem 6 in this answer if it is appropriate to do so.)

Solution 1

The Wilson interval corresponds with the asymptotically normal one-sample proportions test; therefore, I can use the Wilson interval solution for problem 6.

88.5% is not contained in the 95% Wilson interval. Thus, at a 5% significance level, we **reject** the null hypothesis that the true rate of the Cheerios breakfast diet reducing the LDL is 88.5%. The observed rate of 81.25% provides evidence that the true rate for the Cheerios diet **is less than** 88.5%.

Solution 2

Ho: $\theta = 0.885$. Ha: $\theta \neq 0.885$. TS = (0.8125-0.885)/sqrt(0.885*(1-0.885)/96) = -0.0725/0.03256 = -2.227. Conclusion: |-2.227| > 1.96; therefore, at a 5% significance level, we **reject** the null hypothesis that the true rate of the Cheerios breakfast diet reducing the LDL is 88.5%. The observed rate of 81.25% provides evidence that the true rate for the Cheerios diet **is less than** 88.5%.

***8)** Researchers conducted a follow-up study of the impact of eating Cheerios **for dinner** on LDL. LDL was observed to be lowered in 35 of the 50 participants. Consider the hypothesis test of whether the effectiveness of Cheerios in lowering LDL depends on whether they are eaten for breakfast or for dinner, i.e. letting θ be the true probability of lowering LDL, Ho: $\theta_{breakfast} = \theta_{dinner}$ and Ha: $\theta_{breakfast} \neq \theta_{dinner}$.

Calculate an appropriate test statistic for this hypothesis test. (Show all work.)

Solution 1
Breakfast: pb = 78/96. nb = 96.
Dinner: pd = 35/50. nd = 50.
Pooled: pp = (78+35)/(96+50) = 113/146

$$TS_{b-d} = \frac{(pb - pd) - 0}{\text{sqrt}(pp\ (1-pp)\ (1/nb + 1/nd))}$$

$$= \frac{(78/96 - 35/50)}{\text{sqrt}(113/146 * (1-113/146) * (1/96 + 1/50))}$$

$$= \frac{0.1125}{0.07294561}$$

$$= 1.542245$$

$$= 1.54$$

Solution 2

$$TS_{b-d} = \frac{(pb - pd) - (1/(2nb) + 1/(2nd))}{\text{sqrt}(pp\ (1-pp)\ (1/nb + 1/nd))}$$, using continuity correction

= 1.3338.

Solution 3
A Chi-square test may also be used here. This solution is not shown here.

***9)** What critical value would you use if you were interpreting your test statistic in #8 at a 5% significance level?

Solution 1 or 2
I used the Z-test, as opposed to the χ^2 test, so the appropriate critical value is 1.96.

****10)** A study of the effect of eating Cheerios for lunch measured the LDL in mg/dl. The researchers assumed the distribution of the sample mean was sufficiently normal to calculate the standard normal 95% confidence interval using 1.96 as the critical value. They reported the 95% CI for the mean LDL as (93.06, 98.94). Calculate the test statistic for the hypothesis test that the true mean is 100 mg/dl, i.e. Ho: $\mu = 100$. (Show all work.)

Solution
95% CI = (point estimate $\pm$ 1.96 * SE)
TS = (point estimate - 100) / SE

point estimate = (93.06 + 98.94) / 2 = 96

1.96 * SE = (98.94 - 96) = 2.94
SE = 1.5

TS = (96-100)/1.5
= -4/1.5
= -2.67

****11)** American and British researchers separately conducted independent studies in their native countries on the effect of the drug Cooleeez on reducing fevers. The U.S. Study with 10 subjects reported a 1°F average reduction with a standard deviation of the participants' change in temperatures of s_{US} = 2°F. The U.K. Study with 12 subjects reported a 1°C average reduction with a standard deviation of the participants' change in temperatures of s_{UK} = 2°C. **Calculate the test statistic** for the hypothesis that the two studies agree with each other, i.e. let μ_i = the true mean reduction in temperature for country i, Ho: $\mu_{US} = \mu_{UK}$, and Ha: $\mu_{US} \neq \mu_{UK}$. (Show all work and recall that y°C = 5/9*(x°F - 32). Use the rule of thumb $s_{bigger}^2/s_{smaller}^2 < 2$ to decide if the equal variances assumption is appropriate.)

Hint: Notice they are reporting "reductions", i.e. Xpost - Xpre.
Reduction°C = 5/9*(Xpost°F - 32) - 5/9*(Xpre°F - 32)
= 5/9 * Reduction°F.

US in °F	US in °C	UK in °C	Check equal variance assumption:
n = 10	n = 10	n = 12	suk^2/sus^2 = 324/100 = 3.24 > 2
mean = 1	mean = 5/9	mean = 1	By the rule of thumb, the equal
s = 2	s = 10/9	s = 2	variances assumption is not
			appropriate for this dataset.

$TS_{us\text{-}uk}$ = (5/9 - 1 - 0) / sqrt((10/9)^2 / 10 + 2^2 / 12)
= -0.4444 / 0.6759
= -0.657596

****12a)** Provide a p-value to four decimal places for your test in question 11.

Solution using R Method 1

```
TS <-  (5/9 - 1) / sqrt( (10/9)^2 / 10 + 2^2 / 12 )
#Welch-Satterthwaite approximation for degrees of freedom :
DFapprox <- ( (10/9)^2/10 + 2^2/12 )^2 / ( ((10/9)^2/10)^2/(10-1) + (2^2/12)^2/(12-1) )
# df = 17.69
2*pt(TS, df= DFapprox)
= 0.5192631
round( 2*pt(TS, df= DFapprox), 4)
```

= 0.5193

Solution using R Method 2

```
# Marquitta's Solution
UK <-rnorm(12)
UK <-2*(UK-mean(UK))/sd(UK)+1
US <-rnorm(10)
US <-(10/9)*(US-mean(US))/sd(US)+(5/9)
t.test(US, UK, var.equal = FALSE)
round( t.test(US, UK, var.equal = FALSE)$p.value, 4)
```

= 0.5193

Solution using Stata Method 3

```
ttesti 12 1 2 10 0.5555556 1.1111111, unequal
```

= 0.5193

***12b)** Is your conclusion to the test in question 11 robust to your assumptions about the variances?

Solution

Yes, in fact the p-values are both around 0.5 leaving no ambiguity in a conclusion made at the 0.05 level. We would not reject the null hypothesis regardless of whether we assumed equal variances in our test.
Support: Solve equal variance test by hand or as follows.

```
round(  t.test(US, UK, var.equal = TRUE)$p.value,  4)  = 0.5388
ttesti 12 1 2 10 0.5555556 1.1111111  = 0.5388
```

*****13)** Referring to the studies in question 11, notice that neither of the studies was able to reject the null hypothesis that Cooleeez doesn't reduce fevers. The researchers meet and ask if they combined the data from their two studies, could they show Cooleeez is effective. **Calculate the test statistic** for an appropriate test of Cooleeez reducing fevers, i.e. defining μ as the true mean reduction in fevers for the US and the UK, Ho: $\mu = 0$.

Solution 1

If we think about how the mean is the weighted average of the US and UK means, i.e.

Xbar_pooled = (10/22)*(Xbar_us) + (12/22)*(Xbar_uk)

= (10/22)*(5/9) + (12/22)*(1),

then we can solve for the standard error of Xbar_pooled by using what we know about linear functions of variances and the variance of sums of RVs.

SE(Xbar_pooled) = sqrt(VAR(Xbar_pooled)).

VAR(Xbar_pooled) = VAR((10/22)*Xbar_us) + VAR((12/22)*Xbar_uk)

= (10/22)^2 * VAR(Xbar_us) + (12/22)^2 * VAR(Xbar_uk)

= (10/22)^2 * VAR(X_us)/10 + (12/22)^2 * VAR(X_uk)/12

= (10/22)^2 * (10/9)^2 / 10 + (12/22)^2 * 2^2 / 12

= 0.124

SE(Xbar_pooled) = sqrt((10/22)^2 * (10/9)^2 / 10 + (12/22)^2 * 2^2 / 12)

= 0.353

TS = (79/99) / sqrt((10/22)^2 * (10/9)^2 / 10 + (12/22)^2 * 2^2 / 12)

= 2.259912

Solution 2

$s_{uk}^2 = 4 = ((x_{uk,1} - 1)^2 + (x_{uk,2} - 1)^2 + \dots + (x_{uk,12} - 1)^2) / 11 = SS_{uk} / 11$

$s_{us}^2 = (10/9)^2 = ((x_{us,1} - 5/9)^2 + (x_{us,2} - 5/9)^2 + \dots + (x_{us,10} - 5/9)^2) / 9 = SS_{us} / 9$

$xbar_p^2 = (10*(5/9) + 12*(1)) / 22 = 158/198 = 79/99$

$s_p^2 = ((x_{uk,1} - 79/99)^2 + \dots + (x_{uk,12} - 79/99)^2 + (x_{us,1} - 79/99)^2 + \dots + (x_{us,10} - 79/99)^2) / 21$

$= ((x_{uk,1} - 1 + 1 - 79/99)^2 + \dots + (x_{uk,12} - 1 + 1 - 79/99)^2$
$+ (x_{us,1} - 5/9 + 5/9 - 79/99)^2 + \dots + (x_{us,10} - 5/9 + 5/9 - 79/99)^2) / 21$ by adding and subtracting contants

$= ((x_{uk,1} - 1)^2 + \dots + (x_{uk,12} - 1)^2 + 2(x_{uk,1} - 1)(1 - 79/99) + \dots + 2(x_{uk,12} - 1)(1 - 79/99) + 12*(1 - 79/99)^2$
$+ (x_{us,1} - 5/9)^2 + \dots + (x_{us,10} - 5/9)^2 + 2(x_{us,1} - 5/9)(5/9 - 79/99) + \dots + 2(x_{us,12} - 5/9)(5/9 - 79/99)$
$+ 10*(5/9 - 79/99)^2) / 21$ by $(a+b)^2 = a^2 + 2ab + b^2$

$= (SS_{uk} + SS_{us} + 12*(1 - 79/99)^2 + 10*(5/9 - 79/99)^2) / 21$

by combining terms and noticing that the two sums of deviations from the mean = 0

$= (11*s_{uk}^2 + 9*s_{us}^2 + 12*(1 - 79/99)^2 + 10*(5/9 - 79/99)^2) / 21$

$= (11*2^2 + 9*(10/9)^2 + 12*(1 - 79/99)^2 + 10*(5/9 - 79/99)^2) / 21$

$= 1.63574$

TS = (79/99 - 0) / (1.63574 / sqrt(22))

= 2.288174

***13) Continued

Solution 3

```
# we can adapt Marquitta's trick for problem 12 to get the solution here

UK <-rnorm(12)
UK <-2*(UK-mean(UK))/sd(UK)+1
US <-rnorm(10)
US <-(10/9)*(US-mean(US))/sd(US)+(5/9)
Pooled <- c(US, UK)
sdPooledSimulated <- sd(Pooled)
sePooledSimulated <- sdPooledSimulated / sqrt(22)
TS <- (79/99) / sePooledSimulated
TS
```

= 2.288174

*****14a)** Based on a normal approximation for the test statistic you came up with in problem 13, can you reject the null hypothesis at a 5% level of significance?

Solution
Yes, 2.3 > 1.96.

*****14b)** The normal approximation for your test is likely to be anti-conservative given the small sample sizes; however, you don't have the theory worked out for what the appropriate degrees of freedom for the t distribution. Argue for a conservative estimate of the number of degrees of freedom and state whether or not your conclusions are robust to whether or not you used the normal approximation.

Solution 1
Here we took the weighted average of the two sample means. The degrees of freedom are certainly at least 10, the smaller of the two merged sample sizes. Because qt(1-0.025, df=10) = 2.228 < 2.26, we will reject the null hypothesis for any value of df $\geq$ 10.

Solution 2 or 3
Here we really did merge the samples, so we do know the theory for the number of degrees of freedom. Because qt(1-0.025, df=21) = 2.080 < 2.288, we reject the null hypothesis at a 5% significance level.

*****15)**
Let X1, X2, …, X13 ~iid Poisson(3).
Let Y1, Y2, …, Y17 ~iid Poisson(5).
All Xs and Ys are independent.

Let sum(X) = X1 + X2 + X3 + … + X13
Let sum(Y) = Y1 + Y2 + Y3 + … + Y17
Let the point estimate for λx - λy = sum(X)/13 - sum(Y)/17

What is the standard error of the point estimate, i.e.
what is the StandardDeviation(sum(X)/13 - sum(Y)/17)?
(solve for the number, showing all work)

Solution 1
$\text{Var}(\text{sum}(X)/13 - \text{sum}(Y)/17) = \text{Var}(\text{sum}(X)) / 13^2 + \text{Var}(\text{sum}(Y)) / 17^2$
$= 13*\text{Var}(X) / 13^2 + 17*\text{Var}(Y) / 17^2$
$= \text{Var}(X) / 13 + \text{Var}(Y) / 17$
$= 3/13 + 5/17$
$= 0.5248869.$

$\text{SD}(\text{sum}(X)/13 - \text{sum}(Y)/17) = \text{sqrt}(3/13 + 5/17)$
= 0.7244908

Solution 2 -- Emily's simulation

```
#Here I am doing 10,000 simulations of a poisson distribution with n=13 and lambda=3
simulations <- 50000
n1<-13
TrueMean1<-3
Xbars1<-c()
for (i in 1:simulations){
      Sample1 <-rpois(n1,TrueMean1)
      Xbars1<-c(Xbars1, mean(Sample1))
      }

#Here I am doing 10,000 simulations of a poisson distribution with n=17 and lambda=5
simulations <- 50000
n2<-17
TrueMean2<-5
Xbars2<-c()
for (i in 1:simulations){
      Sample2 <-rpois(n2,TrueMean2)
      Xbars2<-c(Xbars2, mean(Sample2))
      }

#Here is the sd for the point estimate. It will be different for each run, but it's always close to 0.72.
sd(Xbars1-Xbars2)
```

= 0.725

***16)** Assume the under the null hypothesis Ho: λx = λy, that the following test statistic is roughly normally distributed. That is you can use the critical value 1.96 to interpret the test again a two-sided alternative.

$$TS = \frac{(\text{point estimate for } \lambda x - \lambda y) - 0}{\text{StandardDeviation}(\text{point estimate for } \lambda x - \lambda y)}$$

What is the power of this test to detect a significant difference given the true values of λx and λy given in question 15?

Solution 1 -- per formula on page 333 of Rosner
Power = Phi(-1.96 + 2/sqrt(3/13 + 5/17))
= pnorm(qnorm(0.025) + 2/sqrt(3/13 + 5/17))
= 0.7883

Solution 2 -- expanding on Emily's solution to problem 15

```
# use the following to calculate power
diffs <- (Xbars1-Xbars2)
SeExact <- sqrt(3/13 + 5/17)
TSes <- diffs/SeExact

rejected <- sum( abs(TSes)>1.96 )
power <- rejected / simulations
power
```

= 0.7870

Solution 3 -- Emily's t-test solution

```
#simulation to calculate power if using a two-sample unequal variances t-test
simulations<-50000
pSamples<-c()
for (i in 1:simulations){
 Sample1 <- rpois(13,3)
 Sample2<- rpois(17,5)
 pSamples<-c(pSamples, t.test(Sample1,Sample2)$p.value)
 }
 powerSamples<-sum(pSamples <0.05)/simulations
 powerSamples
```

= 0.7607