HW 22 - Solutionsc

For Problems 1-2, consider the following table represents sales of beer in Pennsylvania where it must be sold only by the case. A friend suggests that this means beer is likely less expensive in PA than elsewhere. To test this theory, consider the price per case for $n = 12$ popular beers at *Rite Buy* in PA and at *Total Wine* in DE, where beer is also sold in 6-packs and 12-packs and isn't taxed.

```
mydata <- read.csv("https://raw.githubusercontent.com/ProfSuzy/Stat61/main/beer_data_2005.csv",
                   header=T)
mydata <- mydata[,1:4]
mydata
```

```
##             Beer DE.price..Total.Wine. PA.price..Rite.Buy. Difference..PA...DE.
## 1      Yuengling                 13.99               18.59                 4.60
## 2   Rolling Rock                 13.99               19.19                 5.20
## 3          Coors                 16.38               18.99                 2.61
## 4      Budweiser                 18.99               18.59                -0.40
## 5        Fosters                 19.98               23.99                 4.01
## 6         Pete's                 21.98               22.19                 0.21
## 7    Dock Streed                 22.99               21.99                -1.00
## 8      Dos Equis                 23.98               26.59                 2.61
## 9      Newcastle                 21.48               29.99                 8.51
## 10 Sierra Nevada                 22.98               29.99                 7.01
## 11        Corona                 23.98               30.19                 6.21
## 12       Victory                 27.99               28.29                 0.30
```

We are interested in estimating μ_{diff}, the mean difference in price for a case of beer in PA vs DE. The matched pairs t-test is useful since it seems reasonable to treat the price differences as a set of $n = 12$ independent values with mean μ_{diff} and some standard deviation σ_{diff}.

Problem 1

Suppose the distribution of differences in price is close enough to a Normal distribution to invoke the CLT even though $n = 12$.

(a) Compare the standard error for the difference in averages you would get if you treated these as two independent samples compared to the (more appropriate) standard error for matched pairs. Explain how this shows the value of matching when it is reasonable to do so.

```
## First I am going to rename the columns so the data is easier for me to work with:
colnames(mydata) <- c("Beer", "DE_TotalWine", "PA_RiteBuy", "Difference")
mydata
```

```
##             Beer DE_TotalWine PA_RiteBuy Difference
## 1      Yuengling        13.99      18.59       4.60
## 2   Rolling Rock        13.99      19.19       5.20
## 3          Coors        16.38      18.99       2.61
## 4      Budweiser        18.99      18.59      -0.40
## 5        Fosters        19.98      23.99       4.01
## 6         Pete's        21.98      22.19       0.21
## 7    Dock Streed        22.99      21.99      -1.00
```

```
## 8      Dos Equis        23.98      26.59       2.61
## 9      Newcastle        21.48      29.99       8.51
## 10 Sierra Nevada        22.98      29.99       7.01
## 11        Corona        23.98      30.19       6.21
## 12       Victory        27.99      28.29       0.30
```

```
## The standard error if we treat DE and PA prices as
##     coming from independent populations is s_p*sqrt(1/12 + 1/12)
## Note: you have to be really careful with parentheses in the following code
s_p = ( sum((mydata$DE_TotalWine-mean(mydata$DE_TotalWine))^2) +
          sum((mydata$PA_RiteBuy-mean(mydata$PA_RiteBuy))^2) ) / 12+12-2
SE_ind = s_p*sqrt((1/12) + (1/12))
SE_ind
```

```
## [1] 19.34908
```

```
## The standard error if we treat DE and PA prices as
##  linearly dependent is s_diff * sqrt(1/12)
s_diff = (1/11) * sum((mydata$Difference -
                         mean(mydata$Difference))^2)
SE_dep = s_diff * sqrt(1/12)
SE_dep
```

```
## [1] 2.804207
```

As expected, the SE for the paired data is smaller than the SE for the un-paired data. A smaller SE corresponds to a tighter CI for the difference in population means which corresponds to an increase in the power to detect a difference if one does in fact exist.

(b) Find an approximate 99% CI for μ_{diff} and interpret what this interval suggests about the price differences.

As the problem specifies that we can assume the CLT applies (even though we only have a sample of $n = 12$ differences), we know that

$$\bar{D} \overset{n\to\infty}{\sim} N(\mu_{diff}, \sigma^2/12).$$

But since we don't know σ^2 we must estimate it and thus we will instead consider

$$\frac{\bar{D}}{SE(\bar{D})} = \frac{\bar{D}}{s_{diff}/\sqrt{12}} \overset{n\to\infty}{\sim} t_{(11)}.$$

Thus a 99% CI for μ_{diff} is found to be $[0.518, 6.117]$ by:

```
mean(mydata$Difference) -
  ( qt(0.01/2, df=11) * sd(mydata$Difference)/sqrt(12)  )
```

```
## [1] 6.116871
```

```
mean(mydata$Difference) +
  ( qt(0.01/2, df=11) * sd(mydata$Difference)/sqrt(12)  )
```

```
## [1] 0.5281288
```

Problem 2

Now let's not suppose we can justify the use of the CLT. The sign test is a non-parametric alternative that only assumes the data of $n = 12$ differences are IID with some constant probability θ of being positive. Thus we can treat the number of positive differences, Y, as a Binomial(n, θ) RV and test $H_0 : \theta = 0.5$ vs $H_1 : \theta \neq 0.5$.

(a) Interpret the null and alternative hypotheses in the context of this problem.

Here H_0 implies that whether the price of a DE beer is higher or lower than a PA beer is arbitrary and each option is equally likely. H_1 on the other hand implies that this is not so and that systematically, we can expect one of the states to have consistently higher prices for the same beer.

(b) Use the binomial probability function (*dbinom*() in R) to find the exact p-value for this test and explain its meaning in the context of this problem.

Using the information from the problem, we can use Y as our test statistic and the rejection region will be $\{(PA_{diff,i}, DE_{diff,i}) :| \ (PA_{diff,i} - DE_{diff,i} \ |\geq c_\alpha\}$ where c_α solves: $Pr(| \ (PA_{diff,i} - DE_{diff,i} \ |\geq c_\alpha) = 0.01$. Since the problem asks us to find a p-value and we have a significance level, we don't actually need to calculate c_α and can instead directly calculate the p-value which is

$$
\begin{aligned}
&Pr(| \ (PA_{diff,i} - DE_{diff,i} \ |\geq 10 \mid H_0 : \theta = 1/2) \\
&\quad = Pr(| \ d_i \ |\geq 10 \mid H_0 : \theta = 1/2) \\
&\quad = Pr(Y \leq 2 \mid H_0 : \theta = 1/2) + Pr(Y \geq 10 \mid H_0 : \theta = 1/2) \\
&\quad = Pr(Y = 0 \mid H_0 : \theta = 1/2) + Pr(Y = 1 \mid H_0 : \theta = 1/2) + Pr(Y = 2 \mid H_0 : \theta = 1/2) + \\
&\quad\quad Pr(Y = 10 \mid H_0 : \theta = 1/2) + Pr(Y = 11 \mid H_0 : \theta = 1/2) + Pr(Y = 12 \mid H_0 : \theta = 1/2)
\end{aligned}
$$

```
dbinom(0, size=12, prob=1/2) +
dbinom(1, size=12, prob=1/2) +
dbinom(2, size=12, prob=1/2) +
dbinom(10, size=12, prob=1/2) +
dbinom(11, size=12, prob=1/2) +
dbinom(12, size=12, prob=1/2)
```

```
## [1] 0.03857422
```

This p-value is larger than $\alpha = 0.01$ so based on this data set, we do not have enough evidence to indicate a systematic difference between the prices of DE beers and PA beers.

Problem 3

A 3-year study with 72 chronic cocaine user considered an antidepressant drug called desipramine as a possible treatment for addiction. A clinical trial compared outcomes for subjects randomly assigned to take either desipramine, lithium, or a placebo. The counts of subjects who relapsed within the 3 years are reported in the following table.

```
coc_dat = data.frame( drug = c("Desipramine", "Lithium", "Placebo"),
                      n = c(24, 24, 24),
                      relapse = c(10, 18, 20))
coc_dat
```

```
##          drug  n relapse
## 1 Desipramine 24      10
## 2     Lithium 24      18
## 3     Placebo 24      20
```

(a) Construct a 95% CI for the difference in relapse rates for the placebo compared to desipramine.

Note that we are asked to find a CI for the difference in proportions. Since proportions are actually means of a bunch of zeros and ones, this procedure is mathematically equivalent to a (un-paired) t-test/CI for the difference in population means.

```
desi = c(rep(0, 24-10), rep(1, 10))
plac = c(rep(0, 24-20), rep(1, 20))
t.test(desi, plac, alternative="two.sided",
       paired=TRUE, conf.level = 0.95)
```

```
##
##  Paired t-test
##
## data:  desi and plac
## t = -4.0532, df = 23, p-value = 0.0004929
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.6293226 -0.2040107
## sample estimates:
## mean of the differences
##              -0.4166667
```

Thus a 95% CI for the difference in mean relapse rates $\mu_{diff} = \mu_{Desi} - \mu_{plac}$ is $[-0.629, -0.204]$.

(b) The lithium group was included because this had been show previously to be an effective treatment for addiction. To argue for desipramine, it should be shown to be at least as effective as lithium. Define parameters, state hypotheses and carry out the test using Fisher's exact test. Give the exact p-value and explain your conclusion at an $\alpha = 0.01$ significance level.

```
## Desi is first row, Lithium is second row,
## relapse=TRUE is the first column, relapse=FALSE is the second column
coc_fisher_table = matrix(c(10,24-10,18,24-18), nrow=2, byrow=TRUE)
chisq.test(coc_fisher_table, simulate.p.value = TRUE)
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  coc_fisher_table
## X-squared = 5.4857, df = NA, p-value = 0.04348
```

The p-value of a Fisher's exact test for an association between the drug treatment type and whether or not a patient relapses is 0.039, which is larger than the significance level. Therefore, from this data set, we do not see any statistical indication that there is a difference in relapse rates between these two drugs.

(c) State hypotheses and carry out a Chi-square test of independence for all three groups. Give an approximate p-value for the test and explain what it represents in the context of this problem.

First we need to get the data into the proper format where each row is an observational unit (a person) and each column corresponds to a `factor` variable, one for drug type the other for relapse status. Then,

```
coc_data_formatted = data.frame(drug = c(rep("Desipramine", 24),
                                         rep("Lithium", 24),
                                         rep("Placebo", 24)),
                                relapse = factor( c(rep(0, 24-10), rep(1, 10),
##Note: I'm making sure R recognizes these 0s and 1s as levels/factors
                                                    rep(0, 24-18), rep(1, 18),
                                                    rep(0, 24-20), rep(1, 20))) )
table(coc_data_formatted)
```

```
##              relapse
## drug           0  1
##   Desipramine 14 10
##   Lithium      6 18
##   Placebo      4 20
```

Now we can conduct a chi-square test for independence between the treatment (drug) levels and the incidence of a relapse with:

```
chisq.test(table(coc_data_formatted))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(coc_data_formatted)
## X-squared = 10.5, df = 2, p-value = 0.005248
```

Since the p-value is 0.0052, which is much smaller than the significance level, there is strong statistical evidence that relapse and drug treatment are not independent. However, we should keep in mind that by now we've conducted a total of three hypothesis tests on the same data so it isn't appropriate for us to conclude that all tests were/were not significant at an overall level of $\alpha = 0.01$. It would be more transparent (although probably overly conservative), to report these results as significant/not significant based on individual significance levels of $\alpha = 0.01/3 = 0.0033$ since we've conducted three separate tests.