Homework 2 Solutions

Statistics 411: Spring 2018, 25 points

1. Regarding the article *More pregnant women are using pot*:

   (a) (3 pts) The scientific method is:

      i. Observe some phenomenon
      ii. State a hypothesis explaining the phenomenon
      iii. Collect data
      iv. Analyze the data
      v. Test: Does the data support the hypothesis?
      vi. Conclusion. If the test fails, go back to step ii.

   The scientists in the article followed the scientific method by:

      i. Many states are legalizing marijuana use and hence it is becoming more normalized in society.
      ii. The study's lead author, Kelly Young-Wolff, licensed clinical psychologist and research scientist at the Kaiser Permanente Northern California Division of Research, must have been curious how pot use in pregnant women may be changing. Is it different than in previous years? Does the change depend on the age of the expecting mother? Maybe she even thought that it would be higher because "California legalized medical marijuana use in 1996, so they have grown up with the idea of it not only not being illegal but being a medical therapy."
      iii. $n = 279,457$ pregnant women answered questionnaires and took a cannabis toxicology test during their standard prenatal care visits from 2009 through 2016.
      iv. CNN reported that "The prevalence of marijuana use among a sample of moms-to-be in California climbed from 4.2% to 7.1% ... Among pregnant teens younger than 18, marijuana use climbed from 12.5% to 21.8%, and among women 18 to 24, marijuana use climbed from 9.8% to 19%, the researchers found."
      v. and vi. "We were concerned to find that the prevalence of marijuana use in pregnancy is increasing more quickly among younger females, aged 24 and younger, and to see the high prevalence of use in this age group," concluded Kelly Young-Wolff.

   (b) (1 pt) The population of interest to the study is pregnant females. The actual population studied was women in Kaiser Permanente in Northern California health care system.

   (c) (1 pt) The pregnant women in the study are likely from a convenience sample (women easily contacted by Kelly Young-Wolff or one of her collaborators). Hence, the sample may suffer from voluntary response bias (all women undergoing prenatal care in Kaiser Permanente in Northern California health care system from 2009 - 2016 were contacted and asked to be in the study).

   (d) (2 pts) This is an observational study of marijuana use in pregnant women. Neither pregnancy nor pot-use were randomly assigned to some women in some years and then marijuana use studied.

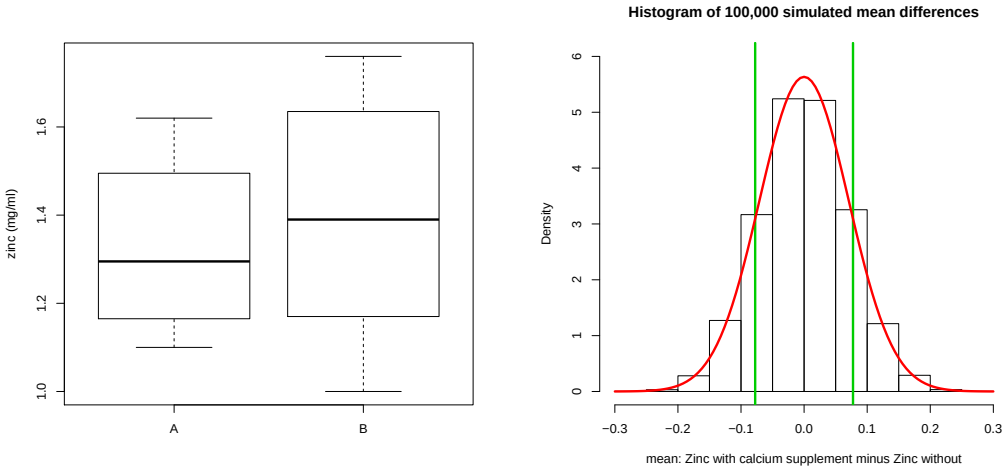
   (e) (1 pt) "The prevalence of marijuana use among a sample of moms-to-be in California climbed from 4.2% to 7.1% ... Among pregnant teens younger than 18, marijuana use climbed from 12.5% to 21.8%, and among women 18 to 24, marijuana use climbed from 9.8% to 19%, the researchers found."

   (f) (4 pts) **Scope of Inference**: This was not a randomized experiment, so it is possible that other behaviors besides pregnancy could have contributed to marijuana use in these women. Due to the lack of random sampling from a large population, it would be tenuous to predict marijuana use of pregnant women outside of Kaiser Permanente in northern CA. In conclusion, this study shows that the $279,457$ pregnant women who participated in the study were associated with brain damage.

2. The statistical report for Exercise 25 on page 25 is on the next page.

Introduction[1]

Dieticians are concerned that taking calcium supplements can affect zinc levels. To test this hypothesis, an experiment was performed on 39 rats. Typically, rats for experimental purposes are bred to be genetically be very similar and purchased from specialized companies. Twenty rats received a calcium supplement, the others received no supplement. We will assume that the calcium treatment was randomly assigned. The zinc levels in all rats were then measured - see the side-by-side boxplots in the left pane of the figure below.



Statistical Procedures Used

A two-sample two-sided randomization test was used to test for any difference in mean zinc levels between the group of rats that received the calcium treatment (Group A) and the group of rats that did not (Group B), so the hypotheses are $H_0 : \mu_A = \mu_B$ vs. $H_a : \mu_A \neq \mu_B$. 10^5 simulations were performed (see the Appendix for the R code). A histogram of the 10^5 simulated mean differences is shown in the right pane to the figure above. A Welch two-sample two-sided t-test was also performed to test these same hypotheses.

Summary of Statistical Findings

The 20 rats in the treatment group that received a calcium supplement (Group A) had slightly lower zinc levels, of 0.078 mg/ml on the average, compared to the control group (Group B). Results for each group are shown in the following table.

	Group	
	A	B
Mean	1.323500	1.4010526
SD	0.174997	0.2572481
n	20	19

The evidence failed to suggest an effect on the mean zinc levels (randomization test statistic 0.76, $p = 0.272$)[2]. With 95% confidence, the effect of calcium was between a 0.22 mg/ml decrease or a 6mg/ml increase in mean zinc levels. Similar results were obtained by a Welch two-sample t-test ($t = -1.1$, df $= 31$, p =0.282, 95% CI for Group A mean - Group B mean: [-0.22, .06]).

[1] Give an overview of the study's goal, study design, and sampling plan in the Introduction.

[2] Or, I could have stated that *The difference in means was not statistically significant.* I would not say there was *no* evidence of a difference. You could say that there was *no* evidence of an *increase* in zinc levels due to the calcium.

Scope of Inference[3]

Because these rats were a convenience sample bought from a company, then any conclusions from these data may only hold for the population of rats bred by the company. Because a randomized experiment was performed, the evidence fails to suggest that the calcium supplement caused any change in the mean zinc levels of the rats.

Appendix

In the left pane below, side-by-side boxplots compare the zinc levels (mg/ml) between the group of rats that received the calcium supplement (Group A) and the group of rats that did not receive calcium (Group B) are presented next. The right pane summarizes the results of the randomization test as a histogram of the simulated differences in mean zinc levels (mg/nml) between the group of rats that received the calcium supplement (Group A) and the group of rats that did not receive calcium (Group B). The test statistic value of $\overline{Y}_A - \overline{Y}_B = -0.078$ is shown by a green vertical line. There is a second green vertical line at $+0.078$ to indicate how the p-value was calculated. The area in the histogram in the lower tail (below -0.078) plus the area in the upper tail (above $+0.078$) is the two-sided p-value $= 0.272$ for the test of $H_0 : \mu_A = \mu_B$ vs. $H_a : \mu_A \neq \mu_B$. The normal approximation, $N\left(0, \sqrt{0.175^2/20 + 0.257^2/19}\right)$, to the histogram representation of the randomization distribution is given by the red bell-shaped curve.

R-code and R-output

```
# Get the data
library(Sleuth3)
rats = ex0125

summary(rats)
##  Group       Zinc
##  A:20  Min.   :1.000
##  B:19  1st Qu.:1.165
##        Median :1.330
##        Mean   :1.361
##        3rd Qu.:1.525
##        Max.   :1.760

# Graph the data
boxplot(rats$Zinc~rats$Group,ylab="zinc (mg/ml)")

# Some simple stats into a table
Mean = tapply(rats$Zinc,rats$Group,mean)
Mean
##       A        B
##1.323500 1.401053

SD = tapply(rats$Zinc,rats$Group,sd)
SD
##        A         B
##0.1749970 0.2572481

n = tapply(rats$Zinc,rats$Group,length)
n
## A  B
## 20 19

rbind(Mean,SD,n)
##             A          B
##Mean  1.323500  1.4010526
```

[3]Explain how study design affects conclusions of cause-and-effect vs. association; and how the sampling plan affects conclusions regarding the population to whom results can be generalized.

```
##SD    0.174997  0.2572481
##n    20.000000 19.0000000

# observed difference is 1.323500 - 1.401053 = -0.07755
test.stat = as.numeric(Mean[1] - Mean[2] )    # Group A mean - Group B mean
test.stat
## [1] -0.07755263

# How many possible randomizations are there?
choose(39,20)
## [1] 68923264410    # 6.9e10

# Perform 1e5 simulations, a tiny fraction of the total number of randomizations 1e5/6.9e10
num_sim=1e5
diff.mean <- numeric(num_sim)       # storage vector

# generate random assignments and calculate difference in means
for(i in 1:num_sim)
{
  grp<-sample(rats$Group,39,replace=F)
  diff.mean[i]<- mean(rats$Zinc[grp=="A"])- mean(rats$Zinc[grp=="B"])
}

# Graph the approximate randomization distribution
hist(diff.mean,freq=FALSE,ylim=c(0,6),main="Histogram of 100,000 simulated mean differences",
  xlab="mean: Zinc with calcium supplement minus Zinc without") # a density histogram

# Visualize the test statistic and the p-value
abline(v=test.stat,col=3,lwd=3) # puts a green vertical line at the test.stat = observed difference
abline(v=-test.stat,col=3,lwd=3)

# Visualize the normal approximation
SE_Unpooled =  sqrt(SD[1]^2/n[1] + SD[2]^2/n[2])
curve(dnorm(x,0,SE_Unpooled),add=TRUE,col=2, lwd=3)

# Get the two-sided p-value
sum(abs(diff.mean)>=abs(test.stat))/num_sim # two-sided p-value
## [1] 0.27237

# 95% Confidence interval
quantile(diff.mean + test.stat, p=c(0.025,0.975))
##       2.5%       97.5%
##-0.21463158  0.05939474

# Confirmatory analysis by a t-test
t.test(Zinc ~ Group,data=rats)
##
##        Welch Two Sample t-test
##
##data:  Zinc by Group
##t = -1.0952, df = 31.532, p-value = 0.2817
##alternative hypothesis: true difference in means is not equal to 0
##95 percent confidence interval:
## -0.22187364  0.06676838
##sample estimates:
##mean in group A mean in group B
##       1.323500        1.401053
```