STAT22000 Summer 2020 Homework 12

All page, section, and exercise numbers below refer to the course text (*OpenIntro Statistics*, 3rd edition, by Diez, Barr, and Cetinkaya-Rundel.).

Reading: Section 5.3, 5.2 (Skip 5.4 and 5.5)

Problems for Self-Study (Do Not Turn In): Exercise 5.25, 5.35 on p.260-267 and Exercise 5.17, 5.19, 5.21 on p.260-261 The answers can be found at the end of the book.

Problems to Turn In: due **midnight of Tuesday, July 21, on Canvas**.

1. This problem is about the data set in Lab 8:

   `http://www.stat.uchicago.edu/~yibi/s220/labs/lab08.html`

   which is a random sample of 1000 birth records in the state of North Carolina. We are interested in comparing the average weights of babies born to smoking and non-smoking mothers. In the data, the variable `weight` stores the birth weights of babies, and the variable `habit` indicator whether the mother is a smoker or a nonsmoker.

   (a) First lets load the data set.

   ```
   nc = read.csv("https://www.openintro.org/stat/data/csv/ncbirths.csv")
   ```

   The original data set contains both full term babies and premature babies. Here let's just focus on *full term* babies only.

   ```
   nc.full = subset(nc, premie=="full term")
   ```

   Make a side-by-side histogram and a side-by-side boxplot comparing the weights of *full term* babies born to smoking and non-smoking mothers. Comment on the appropriateness of using the two-sample t-tests and t-intervals.
   Hint: Try the following R codes.

   ```
   library(mosaic)
   histogram(~weight | habit, data=nc.full, layout=c(1,2), width=0.5,
            xlab="Birth Weight of Full Term Babies (lbs)")
   bwplot(weight ~ habit, data=nc.full,
          ylab="Birth Weight of Full Term Babies(lbs)", xlab="Mother")
   ```

   (b) Test whether if the mean birth weights of full term babies born to smoking and non-smoking mothers are different. Write down the hypotheses. Report the test statistics, the degrees of freedom, and the P-value (without assuming equal population SDs). What is your conclusion at significance level $\alpha = 0.05$? Please show how the test statistic is calculated using the sample means and the sample SDs. Do NOT use the `t.test()` command.
   Hint: The summary statistic (mean, SD, sample size) can be obtain using the R command

   ```
   favstats(weight ~ habit, data=nc.full)
   ```

   (c) Give an estimate of the difference in the mean birth weights of full term babies born to smoking and those born to non-smoking mothers and construct a 95% confidence interval for the difference. Please show how the confidence interval is calculated using the sample means and the sample SDs. Do NOT use the `t.test()` command.

   (d) Check your computation in (c) and (d) with the `t.test()` function in R as follows:

   ```
   t.test(weight ~ habit, data=nc.full)
   ```

2. This problem is a continuation of Problem #3 in HW11. It has been hypothesized that allergies result from a lack of early childhood exposure to antigens. If this hypothesis were true, then we would expect allergies to be more common in very hygienic households with low levels of bacteria and other infectious agents. To test this theory, researchers at the University of Colorado sampled the houses of 61 children 9-24 months old and recorded two variables: (1) whether the child tested positive for allergies and (2) the concentration of bacterial endotoxin in the house dust (endotoxin units per ml, EU/ml)[1]. The data file `allergy.txt` is posted along with HW12 on Canvas.

   (a) Load the data set to R and make a side-by-side boxplot of bacterial endotoxin concentration by the commands below.

   ```
   allergy = read.table("allergy.txt", h=T)
   library(mosaic)
   bwplot(Endotoxin ~ Allergic, data=allergy)
   ```

   Comment on whether it is appropriate to conduct two-sample t test on the equality of the mean endotoxin levels between the "sensitive" and "normal" groups.

   Review Section 3 in Lab #1 `http://www.stat.uchicago.edu/~yibi/s220/labs/lab01.html` *about changing the working directory if you have trouble loading the data file to R.*

   (b) Make a side-by-side boxplot of the log of bacterial endotoxin concentration by the commands below.

   ```
   bwplot(log(Endotoxin) ~ Allergic, data=allergy)
   ```

   Comment on whether it is appropriate to conduct a two-sample t test on the equality of the mean of the log concentration of bacterial endotoxin in the house dust of the two groups.

   (c) Test if the mean of the log endotoxin levels of the normal group μ_n is higher than that of the sensitive group μ_s, i.e., $H_0 : \mu_n = \mu_s$ v.s. $H_a : \mu_n > \mu_s$, WITHOUT assuming the equality of the two population SDs. Report the t-statistic, degrees of freedom, and give a range of the p-value. The summary statistic of the data can be obtained by the following command.

   ```
   favstats(log(Endotoxin) ~ Allergic, data=allergy)
   ```

   Please show how the test statistic is calculated using the sample means and the sample SDs. Do NOT use the `t.test()` command.

   (d) Construct a 90% confidence interval for the difference in the mean log endotoxin levels in the "normal" group and the "sensitive" group (normal − sensitive). Please show how the confidence interval is calculated using the sample means and the sample SDs. Do NOT use the `t.test()` command. Remark: Recall that $\log(a) - \log(b) = \log(a/b)$. If a 90% CI for the mean difference in the mean log of the endotoxin levels is (L, U) (normal − sensitive), then (e^L, e^U) is a 90% CI for the ratio of the two means without the log transformation. The confidence interval can be described as: with 90% confidence, the mean endotoxin levels in the houses of 9-24 month children without allergy was e^L to e^U times as large as those with allergy.

   (e) Check your computation in (c) and (d) with the `t.test()` function in R as follows:

   ```
   t.test(log(Endotoxin) ~ Allergic, data=allergy, alternative = "greater")
   t.test(log(Endotoxin) ~ Allergic, data=allergy, alternative = "two.sided", conf.level=0.9)
   ```

[1] Gereda JE, Leung DYM, Thatayatikom A, Streib JE, Price MR, Klinnert MD, and Liu AH. (2000). Relation between house-dust endotoxin exposure, type 1 T-cell development, and allergen sensitisation in infants at high risk of asthma. *The Lancet*, **355**: 1680-1683.

Are any physiological indicators associated with schizophrenia? Early studies, based largely on postmortem analysis, suggest that the sizes of certain areas of the brain may be different in persons afflicted with schizophrenia than in others. Confounding variables in these studies, however, clouded the issue considerably. In a 1990 article, researchers reported the results of a study that controlled for genetic and socioeconomic differences by examining 15 pairs of identical twins, where one of the twins was schizophrenic and the other was not. The twins were located through an intensive search throughout Canada and the United States[a]. The researchers used magnetic resonance imaging (MRI) to measure the volumes (in cm^3) of several regions and subregions inside the twins' brains. The table presents data based on the reported summary statistics from one subregion, the left hippocampus.

[a]Data from R. L. Suddath et al., 'Anatomical Abnormalities in the Brains of Monozygotic Twins Discordant for Schizophrenia,' *New England Journal of Medicine* 322(12) (1990): 789-93.

	Unaffected	Affected	diff
	1.94	1.27	0.67
	1.44	1.63	−0.19
	1.56	1.47	0.09
	1.58	1.39	0.19
	2.06	1.93	0.13
	1.66	1.26	0.40
	1.75	1.71	0.04
	1.77	1.67	0.10
	1.78	1.28	0.50
	1.92	1.85	0.07
	1.25	1.02	0.23
	1.93	1.34	0.59
	2.04	2.02	0.02
	1.62	1.59	0.03
	2.08	1.97	0.11
Mean	1.759	1.560	0.199
SD	0.242	0.301	0.238

3. (a) Test the hypothesis H_0: $\mu = 0$ versus H_a: $\mu \neq 0$, where μ is the mean difference between the left hippocampus volumes of twins discordant on schizophrenia. Please report the test statistic, the degrees of freedom, and the P-value. What do you conclude at 0.05 significance level?

   (b) Construct a 95% confidence interval for the mean difference in volumes of the left hippocampus between the unaffected and the affected individuals.

   (c) Check your computation in (a) and (b) with the R commands below.

```
unaffected = c(1.94,1.44,1.56,1.58,2.06,1.66,1.75,1.77,1.78,1.92,1.25,1.93,2.04,1.62,2.08)
affected = c(1.27,1.63,1.47,1.39,1.93,1.26,1.71,1.67,1.28,1.85,1.02,1.34,2.02,1.59,1.97)
t.test(unaffected, affected, paired=T)
```