STAT22000 Summer 2020 Homework 12 Solutions

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.

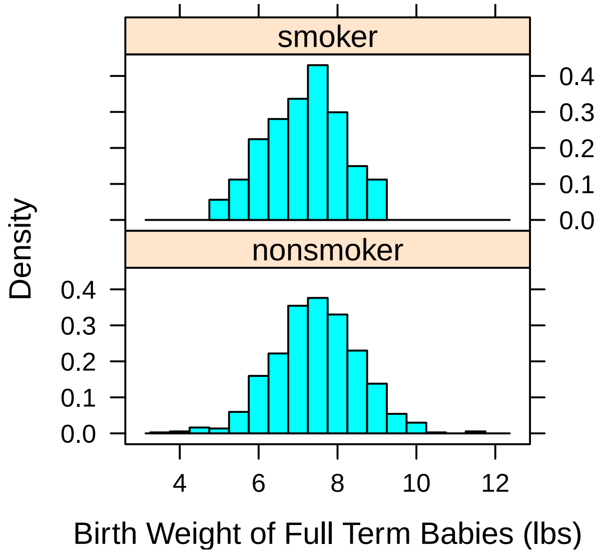
   ```
   nc.full = subset(nc, premie=="full term")
   ```

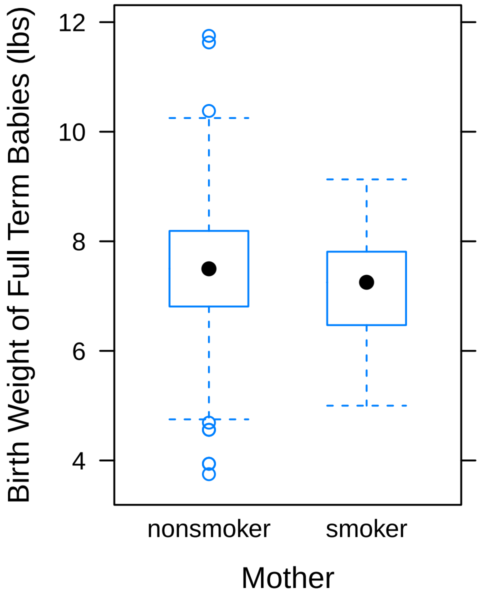
   Make a side-by-side histogram and a side-by-side boxplot comparing the weights of *full term* babies born to smoking and non-smoking mothers. Comment on the appropriateness of using the two-sample t-tests and t-intervals.

   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")
   ```

   Answer: *[1pts = 0.5pt for the plots + 0.5pt for the comment.]*





There remain a few outliers in the non-smoking group but they are not as extreme. It is safer to use the t-procedures given the large sample size.

(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)
```

Answer: *[6pts = 1pt for the hypotheses + 2pts for the SE and t-statistic + 1pt for the df + 1pt for the P-value + 1pt for the conclusion.]*

```
> favstats(weight ~ habit, data=nc.full)
      habit  min    Q1 median   Q3   max     mean        sd   n missing
1 nonsmoker 3.75 6.81   7.50 8.19 11.75 7.501123 1.0832873 739       0
2    smoker 5.00 6.47   7.25 7.81  9.13 7.171308 0.9723654 107       0
```

The mean, SD, and size of the the two groups of babies are as follows:

	n	$\overline{x}$	s
nonsmoker	739	7.5011	1.0829
smoker	107	7.1713	0.9724

The hypotheses are $\mathrm{H}_0 : \mu_s = \mu_{ns}$ and $\mathrm{H}_a : \mu_s \neq \mu_{ns}$, where μ_s and μ_{ns} are respectively the population mean birth weights of full-term babies born to smoking and non-smoking mothers. The t-statistic is

$$t\text{-statistic} = \frac{\overline{x}_{ns} - \overline{x}_s}{\sqrt{\frac{s_{ns}^2}{n_s} + \frac{s_s^2}{n_s}}} \approx \frac{7.5011 - 7.1713}{\sqrt{\frac{(1.0829)^2}{739} + \frac{(0.9724)^2}{107}}} \approx \frac{0.3298}{0.1021} \approx 3.2302.$$

with $\mathrm{df} = \min(739 - 1, 107 - 1) = 106$. From the R code below, we can fine the two-sided P-value to be around 0.00165.

```
> 2*pt(3.2302, df=106, lower.tail=F)
[1] 0.001647791
```

Conclusion: As the two-sided p-value $\approx 0.00165 < 0.01$, we can conclude that the mean birth weight of full term babies born to smoking mothers is significantly lower than that of smoking mothers.

(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.

Answer: *[3pts in total = 1pt for the estimate + 1pt for t^* + 1pt for the SE]* Full term babies born to smoking mothers had a lower mean birth weight, by an estimate of $7.501 - 7.171 = 0.330$ lbs. With $\mathrm{df} = 106$, the critical value for 95% CI is $t^* \approx 1.98$

```
> qt(0.05/2, df=106, lower.tail=F)
[1] 1.982597
```

The 95% CI for the mean difference $\mu_{ns} - \mu_s$ is

$$\overline{x}_s - \overline{x}_p \pm t^* \sqrt{\frac{s_s^2}{n_s} + \frac{s_p^2}{n_p}} \approx 7.5011 - 7.1713 \pm 1.9826 \sqrt{\frac{(1.083)^2}{739} + \frac{(0.972)^2}{107}}$$
$$\approx 0.3298 \pm 0.20235 = (0.12745, 0.53215)$$

Conclusion: The mean birth weight of full term babies born to smoking mothers is 0.128 lbs to 0.532 lbs lower than that of those born to non-smoking mothers, with 95% confidence.

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

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

Answer: *[0pt]* The t-statistic obtained is 3.2303, with a larger df 146.84 because R use the software formula to compute the df. The p-value 0.001526 is slightly lower than the one 0.00165 we obtained. The 95% CI $(0.1280399, 0.5315896)$ is slightly narrower than our CI $(0.12745, 0.53215)$ because R uses the software formula to compute the df.

```
Welch Two Sample t-test

data:  weight by habit
t = 3.2303, df = 146.84, p-value = 0.001526
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1280399 0.5315896
sample estimates:
mean in group nonsmoker    mean in group smoker
               7.501123                7.171308
```

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.

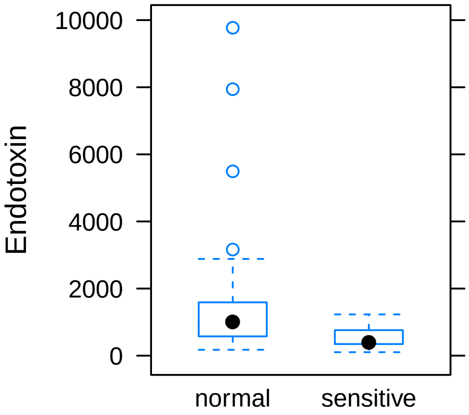
   (a) Load the data set to R and make a side-by-side boxplot of bacterial endotoxin concentration by the commands below.

   ```
   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.*

[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.

Answer:



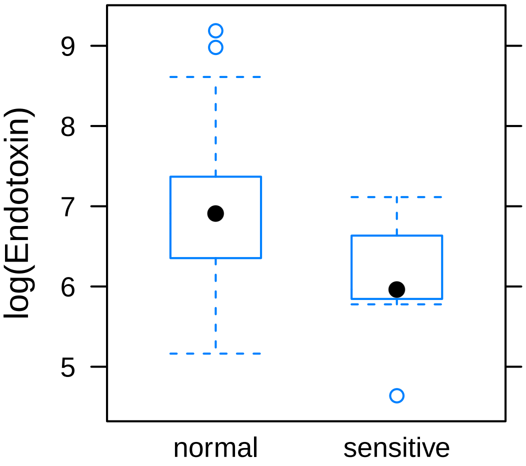
[1pts = 0.5pt for the plot + 0.5 pt for the comment]
The distributions of endotoxin levels are severely right-skewed. There are outliers more than 3 or 4 IQRs above Q3 in the normal group. The two-sample t-test is not reliable when there are extreme outliers. So it's not appropriate to use a two-sample t-test.

(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.

Answer:



[1pts = 0.5pt for the plot + 0.5 pt for the comment]
After log-transformation, the distributions of endotoxin levels become a lot more symmetric. The three outliers tagged by the 1.5 IQR rule are far less extreme than in the untransformed data. It's more appropriate to use use the two-sample t-test on log endotoxin levels than on endotoxin levels.

(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., $\text{H}_0 : \mu_n = \mu_s$ v.s. $\text{H}_a : \mu_n > \mu_s$, 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.

Answer: *[4 pts = 2pts for the t-statistics + 1pt for df + 1pt for p-value]*
The summary statistics of log endotoxin levels for the normal and the sensitive group are as follows.

```
> favstats(log(Endotoxin) ~ Allergic, data=allergy)
   Allergic      min       Q1   median       Q3      max     mean        sd  n missing
1    normal 5.163242 6.353943 6.908465 7.368487 9.187285 6.916581 0.8584405 51       0
2 sensitive 4.638605 5.846055 5.965161 6.507390 7.114241 6.077740 0.6910635 10       0
```

The two-sample t-statistic is

$$T = \frac{\bar{x}_{\text{norm}} - \bar{x}_{\text{sen}}}{\sqrt{\frac{s^2_{\text{norm}}}{n_{\text{norm}}} + \frac{s^2_{\text{sen}}}{n_{\text{sen}}}}} = \frac{6.916581 - 6.077740}{\sqrt{\frac{0.8584405^2}{51} + \frac{0.6910635^2}{10}}} = \frac{0.838841}{0.2494119} \approx 3.3633$$

with $\min(51-1, 10-1) = 9$ degrees of freedom. The upper one-sided p-value is about 0.00417.

```
> pt(3.3633, df=9, lower.tail=F)
[1] 0.00417265
```

Since p-value $\approx 0.00417 < 0.05$, we reject H_0. The data provide strong evidence that the mean log endotoxin levels in the "normal" group is significantly higher than the "sensitive" group.

(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.

Answer: *[3 pts, 1pt for t^*]* With df = 9, the critical value for 90% CI is $t^* = 1.8331$.

```
> qt(0.1/2, df=9, lower.tail=F)
[1] 1.833113
```

The 90% CI for the mean difference $\mu_n - \mu_s$ is

$$\begin{aligned}\bar{x}_n - \bar{x}_s \pm t^* \sqrt{\frac{s_n^2}{n_s} + \frac{s_s^2}{n_s}} &= 6.916581 - 6.077740 \pm 1.8331 \sqrt{\frac{0.8584405^2}{51} + \frac{0.6910635^2}{10}} \\ &\approx 0.838841 \pm 1.8331 \times 0.2494119 = 0.838841 \pm 0.457197 \approx (0.381644, 1.296038).\end{aligned}$$

(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)
```

Answer: *[0pt]*
For your reference only, one can perform the two-sample t-test in R:

```
> t.test(log(Endotoxin) ~ Allergic, data=allergy, alternative = "greater")

Welch Two Sample t-test

data:  log(Endotoxin) by Allergic
```

```
t = 3.3633, df = 15.022, p-value = 0.00213
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.4016521       Inf
sample estimates:
   mean in group normal mean in group sensitive
               6.916581                6.077740
```

The t-statistic obtained is still 3.633, but the degrees of freedom is 15.022 because R use the software formula. The p-value 0.00213 is lower than the one 0.00417 we obtained in (b) since R uses a t-distribution w/ a higher df which has slimmer tails.

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

Welch Two Sample t-test

data:  log(Endotoxin) by Allergic
t = 3.3633, df = 15.022, p-value = 0.00426
alternative hypothesis: true difference in means is not equal to 0
90 percent confidence interval:
 0.4016521 1.2760295
sample estimates:
   mean in group normal mean in group sensitive
               6.916581                6.077740
```

The 90% CI given by R (0.4016521 1.2760295) is narrower than the one (0.381644, 1.296038) we obtained in (c) as R uses a t-distribution w/ a higher df which has slimmer tails.

---

3. 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.

	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	1.7587	1.5600	0.1987
SD	0.2424	0.3013	0.2383

(a) Test the hypothesis H$_0$: $\mu = 0$ versus H$_a$: $\mu \neq 0$, where μ is the mean difference between the left

[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.

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?

Answer: *[3pts = 1pt for the t-statistic + 1pt for the df + 1pt for the P-value. Give 0 pt if using a two-sample t test.]*
First we take the difference between the left hippocampus volumes of the unaffected and the affected twins, and then compute the sample mean and sample SD of the 15 differences to be $\overline{d} \approx 0.199$ and $s \approx 0.238$.

The paired t-statistic is $t = \dfrac{\overline{d} - 0}{s_d/\sqrt{n}} = \dfrac{0.1987 - 0}{0.2383/\sqrt{15}} \approx 3.2294$ with df $= 15 - 1 = 14$. The two-sided P-value is around 0.006.

```
> 2*pt(3.2294, df=14, lower.tail=F)
[1] 0.006055856
```

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

Answer: *[3pts in total, 1pt for t^*. Give 0 pt if using a two-sample t-CI.]*
With df $= 14$, the critical value for 95% CI is $t^* \approx 2.1448$.

```
> qt(0.05/2, df=14, lower.tail=F)
[1] 2.144787
```

The 95% CI for μ is

$$\overline{d} \pm t^* s_d/\sqrt{n} = 0.1987 \pm 2.1448 \times 0.2383/\sqrt{15} \approx 0.1987 \pm 0.1320 = (0.0667, 0.3307).$$

(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)
```

Answer: *[0pt]* The R output for the paired t test is as follows. We see the t-statistic 3.2289, df = 14, P-value 0.006062 and the 95% CI $(0.0667041, 0.3306292)$ are slightly off from our computation in (a) and (b) because of rounding errors.

```
        Paired t-test

data:  unaffected and affected
t = 3.2289, df = 14, p-value = 0.006062
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.0667041 0.3306292
sample estimates:
mean of the differences
              0.1986667
```
