Section 6 Solutions

Statistics 104

Fall 2019

Topics

- Two-sample t-test for paired data
- Two-sample t-test for independent group data
- Statistical power and sample size
- ANOVA

1. **The Paleo Diet.** The Paleo diet allows only for foods that humans typically consumed prior to the development of agriculture about 10,000 years ago. To study the efficacy of the Paleo diet, researchers recruited 500 adult volunteers interested in losing weight from around the Boston area. Volunteers were randomly assigned to either of two equally sized treatment groups. One group spent six months following the Paleo diet, while the other group received a weekly email about healthy eating practices such as reducing portion size. At the beginning of the study, the average difference in weights between the two groups was about 0. The file `diets.Rdata` contain the change in weight for each volunteer, measured in pounds and calculated as (weight at beginning of study - weight at end of study).

   a) Calculate the 90% confidence interval for the difference between the mean weight loss of the two groups (Paleo - control). Interpret this interval in the context of the data.

   The 90% confidence interval is (-0.28, 4.49) pounds. With 90% confidence, the mean amount of pounds that people on the Paleo diet lose relative to people receiving a weekly email about healthy eating lies between 0.28 pounds fewer to 4.49 pounds more.

```
#load the data
load("diets.Rdata")

#calculate the confidence interval
t.test(diets$paleo.group, diets$control.group, mu = 0,
   paired = FALSE, conf.level = 0.90)$conf.int
```

```
## [1] -0.2762898  4.4894930
## attr(,"conf.level")
## [1] 0.9
```

   b) Based on the confidence interval, do the data provide convincing evidence that the Paleo diet is more effective for weight loss than the weekly email newsletter?

   The data do not provide convincing evidence of the effectiveness of the Paleo diet, since the value representing no difference (0) is within the confidence interval. The interval represents plausible observed values for a difference in means if mean weight loss is equal between the two treatment regimens.

c) Would it be reasonable to believe that the study results generalize to the population of American adults? Explain your answer.

The study participants were a group of volunteers interested in losing weight from around the Boston area. Since the participants do not represent a random sample of American adults, the study results do not generalize to the population of America adults.

As volunteers interested in losing weight, the participants of the study are probably more inclined to follow a healthy lifestyle in general, even outside of keeping to a healthy diet. Additionally, the weekly email about healthy eating may appear more effective in a population of people interested in losing weight than in the population at large, since these volunteers have a specific interest in following the emailed advice or may even already be watching their diet. It may be that in the general population, specific instructions to follow the Paleo diet (or any diet) will be more effective for weight loss than receiving an email newsletter.

Furthermore, generalizing the study results to American adults would require assuming that adults in Boston are representative of adults in the United States. It could be more plausible to generalize to, say, the population of American adults living in large metropolitan areas; note that this would require assuming that adults in Boston are reasonably representative of adults living in cities like NYC, Chicago, Seattle, etc.

d) Without explicitly performing the hypothesis test, determine whether the results would have indicated a significant difference in population means if the Paleo group had lost 8.25 pounds on average instead of 7.25. Assume that all other numbers remained constant.

This would have shifted the confidence interval one unit to the left, since $\overline{x}_1 - \overline{x}_2$ would have equaled 3.11 instead of 2.11. The confidence interval of (0.724, 5.489) pounds does not include 0; there would be sufficient evidence to reject H_0 at $\alpha = 0.10$ and conclude that mean weight loss in volunteers on the Paleo diet is greater than in volunteers receiving a weekly newsletter about healthy eating practices.

e) Suppose a new study will be conducted that specifically recruits twins; of a pair of twins, one individual will be assigned to the Paleo diet group, while the other will be assigned to the email newsletter group. The study team is interested in detecting any average difference of at least 5 pounds, and anticipate that the standard deviation of weight will be about 15 pounds. Calculate the number of twin pairs that should be recruited to achieve a power level of 80%, if the study data will be analyzed at $\alpha = 0.10$.

The study should recruit 58 pairs of twins to achieve a power level of 80%. Note how the paired design requires far fewer participants than the independent group design.

```
#paired design
power.t.test(n = NULL, delta = 5, sd = 15,
         sig.level = 0.10, power = 0.80,
         type = "paired",
         alternative = "two.sided")$n
```

```
## [1] 57.02048
```

```r
#suppose the data were not paired...
power.t.test(n = NULL, delta = 5, sd = 15,
         sig.level = 0.10, power = 0.80,
         type = "two.sample",
         alternative = "two.sided")$n
```

```
## [1] 111.9686
```

2. **Indoor Air Quality.** A study was conducted in 1980 to compare the indoor air quality in offices where smoking was permitted with that in offices where smoking was not permitted. Measurements were made of carbon monoxide (CO) at 1:20 p.m. in 40 work areas where smoking was permitted and in 40 work areas where smoking was not permitted; CO levels were measured in parts per million (ppm). The data are in `air_quality.Rdata`. [1]

   Analyze the data using a hypothesis test and confidence interval. Summarize your conclusions in language that a non-statistician would understand.

   Test & Calculations

   H_0: $\mu_{NS} = \mu_S$, H_A: $\mu_{NS} \neq \mu_S$. Let $\alpha = 0.05$.

   The p-value is 0.00071. Since $p < \alpha$, there is sufficient evidence to reject the null hypothesis that there is no difference in mean CO levels between work areas where smoking is permitted and where smoking is not permitted.

   The 95% confidence interval is (2.19, 7.67) ppm.

   Summary

   An analysis was done to assess whether average carbon dioxide levels in areas where smoking is permitted differs from average carbon dioxide levels where smoking is not permitted. When 40 smoking areas were compared to 40 non-smoking areas, the average CO level in smoking areas was almost 5 ppm higher than in non-smoking areas—there is a less than 0.01% chance of observing such a large difference if the CO levels are actually the same in smoking and nonsmoking areas. Instead, it is much more likely that CO levels in areas where smoking is permitted is higher than in areas where smoking is not permitted. Based on the observed data, it is plausible that the amount by which mean CO level in smoking areas is higher than in non-smoking areas is in the range (2.19, 7.67) ppm, for offices in general.

```r
#load the data
load("air_quality.Rdata")

#conduct the test and calculate the interval
t.test(air.quality$smoking, air.quality$nonsmoking, paired = FALSE, mu = 0)
```

```
##
##  Welch Two Sample t-test
##
## data:  air.quality$smoking and air.quality$nonsmoking
## t = 3.6182, df = 48.338, p-value = 0.0007079
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.190948 7.669314
## sample estimates:
## mean of x mean of y
## 12.145045  7.214914
```

[1] Data simulated from summary statistics in Rosner, *Fundamentals of Biostatistics*, 7^{th} ed., p. 310

3. **Global Warming.** Is there strong evidence of global warming? Let's consider a small-scale example, comparing how temperatures have changed in the US from 1968 to 2008. The daily high temperature reading on January 1 was collected in 1968 and 2008 for 51 randomly selected locations in the continental US. The readings are in `us_temps.Rdata`, in units of degrees Fahrenheit. We are interested in determining whether these data provide strong evidence of temperature warming in the continental US.

   According to the worldwide scientific community, emission of greenhouse gases like carbon dioxide are a major cause of global warming; emissions have increased over time due to use of fossil fuels for electricity, transportation, etc. Suppose that our analysis will be used to inform whether existing legislation to limit carbon dioxide emissions should be maintained.

   a) State the hypotheses and α. Consider whether an α different from 0.05 could be reasonable. Justify your choice of alternative hypothesis and α.

      The null hypothesis is that there is no difference in temperature between 1968 and 2008, $H_0 : \delta = 0$. The alternative hypothesis is that temperatures in 2008 are higher than in 1968, $H_A : \delta > 0$ (this argument implies calculating the differences as 2008 - 1968). It is defensible to use a one-sided hypothesis here; one could argue that the consequences of missing a difference in the other direction and concluding no difference are the same (i.e., both result in not maintaining the legislation), or that if temperatures are different, they are probably higher.[2]

      Let $\alpha = 0.10$. It is arguably more harmful to remove necessary legislation (i.e., incorrectly fail to reject H_0) than it is to keep unnecessary legislation (i.e., incorrectly reject H_0).[3] In other words, a Type II error is more dangerous than a Type I error. Raising α implies that the observed difference does not need to be as extreme in order for H_0 to be rejected.

   b) Conduct the hypothesis test and interpret your conclusions.

      The data are paired, since the differences can be calculated between the 1968 measurement and 2008 measurement for each location.

      The p-value is 0.038. Since $p < \alpha$, there is sufficient evidence at $\alpha = 0.10$ to reject the null hypothesis that there has been no change in temperature from 1968 to 2008. The data provide evidence of temperature warming in the continental United States since 1968 at $\alpha = 0.10$.

   c) Calculate and interpret a confidence interval that corresponds to the test conducted in part b).

      The one-sided 90% confidence interval is $(0.37, \infty)$ degrees Fahrenheit; we can be 90% confident that the mean increase in temperature is at least 0.37 degrees Fahrenheit. However, this is not especially informative, so the 80% two-sided interval could also be calculated. We can be 80% confident that the mean increase in temperature is within the interval (0.37, 2.28) degrees Fahrenheit.

[2] Note that it is also reasonable to argue for an unbiased perspective, in which we do not anticipate a difference in either direction and use the two-sided alternative $H_A : \delta \neq 0$.

[3] It is also defensible to argue the other perspective, that a Type I error is more dangerous than a Type II error and that α should be lowerered to 0.01.

```r
#load the data
load("us_temps.Rdata")

#conduct the test
t.test(us.temps$temps.2008, us.temps$temps.1968, paired = TRUE,
   conf.level = 0.90, alternative = "greater")
```

```
##
##  Paired t-test
##
## data:  us.temps$temps.2008 and us.temps$temps.1968
## t = 1.8068, df = 50, p-value = 0.03841
## alternative hypothesis: true difference in means is greater than 0
## 90 percent confidence interval:
##  0.3730991       Inf
## sample estimates:
## mean of the differences
##                1.326758
```

```r
#calculate a two-sided confidence interval
t.test(us.temps$temps.2008, us.temps$temps.1968, paired = TRUE,
   conf.level = 0.80)$conf.int
```

```
## [1] 0.3730991 2.2804162
## attr(,"conf.level")
## [1] 0.8
```

4. **Work Hours and Education.** The General Social Survey collects data on demographics, education, and work, among many other characteristics of U.S. residents. The dataset `gss_work.Rdata` contains the simulated responses for 954 participants. The variable `work.hours` contains the number of hours the respondent works in an average week, while `education.level` contains the highest level of education attained.

   Investigate whether the average amount of hours worked per week differs between individuals who have different levels of educational attainment, at significance level $\alpha = 0.05$.

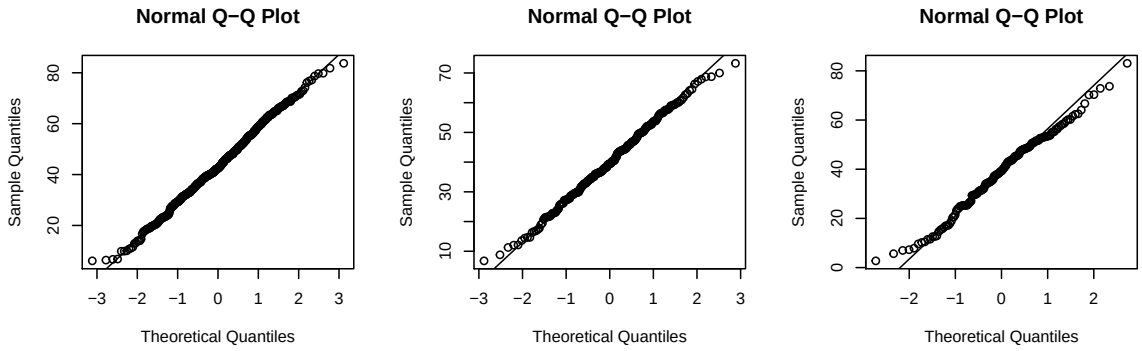
   a) Evaluate the assumptions for ANOVA.

      The data are from a large survey of United States residents; it is reasonable to assume the responses are independent within and across groups. The Q-Q plots show no evidence of non-normality; the dots closely follow the diagonal line. The variability across groups is roughly equal.

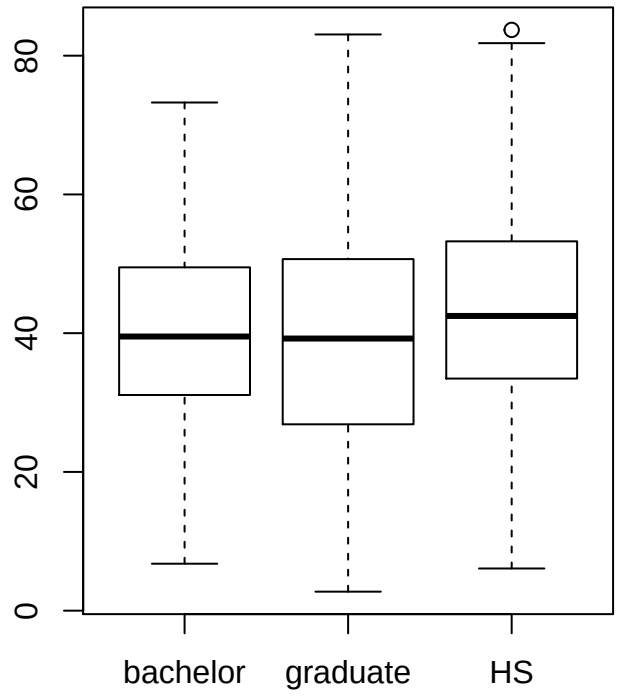
```r
#load the data
load("gss_work.Rdata")

#assess normality within each group
par(mfrow = c(1, 3))
qqnorm(gss.work$work.hours[gss.work$education.level == "high school"])
qqline(gss.work$work.hours[gss.work$education.level == "high school"])
qqnorm(gss.work$work.hours[gss.work$education.level == "bachelor"])
```

```r
qqline(gss.work$work.hours[gss.work$education.level == "bachelor"])
qqnorm(gss.work$work.hours[gss.work$education.level == "graduate"])
qqline(gss.work$work.hours[gss.work$education.level == "graduate"])
```




```r
#evaluate the variability across groups
boxplot(gss.work$work.hours ~ gss.work$education.level,
    names = c("bachelor", "graduate", "HS"))
```



```
tapply(gss.work$work.hours, gss.work$education.level, var)
```

```
##    bachelor    graduate high school
##    170.1626    256.3360    211.8925
```

b) State the null and alternative hypotheses.

The null hypothesis is that the mean amount of hours worked per week does not differ between the different educational attainment groups, $H_0 : \mu_1 = \mu_2 = \mu_3$, where the values 1-3 represent the three groups, respectively (high school, bachelor's, and graduate). The alternative hypothesis is that at least one group mean differs from the others.

c) Conduct the F-test and summarize the results.

The F-statistic is 8.87 and the p-value is 0.00015. Since $p < \alpha$, there is sufficient evidence to reject H_0 in favor of H_A. The evidence suggests that there is at least one group with a population mean hours worked per week that is different from the other groups.

```
#conduct the F-test
summary(aov(gss.work$work.hours ~ gss.work$education.level))
```

```
##                            Df Sum Sq Mean Sq F value   Pr(>F)
## gss.work$education.level    2   3689    1844   8.866 0.000153 ***
## Residuals                951 197838     208
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

d) Complete the analysis using pairwise comparisons.

i. What is the appropriate significance level $\alpha^\star$ for the individual comparisons, as per the Bonferroni correction?

The appropriate significance level $\alpha^\star = 0.05/3 = 0.0167$, since there are three possible pairwise comparisons between three groups.

ii. Conduct pairwise comparisons and summarize the results.

Evidence suggests that the mean hours worked per week for the graduate degree group is not different from the mean hours worked per week for the bachelor's degree group. However, there is evidence at the $\alpha = 0.05$ significance level that mean hours worked per week for individuals in the high school degree group differs from both the hours worked per week of the bachelor's degree group and graduate degree group. The observed data suggests that individuals who attained at most a high school degree work more than individuals who attained at most a bachelor's degree or a graduate degree. In the sample, individuals with at most a high school degree worked an average of 43.4 hours per week; in contrast, individuals with at most a bachelor's degree worked an average of 40.0 hours per week and those with at most a graduate degree worked an average of 38.7 hours per week.

```r
#compare results to alpha*
pairwise.t.test(gss.work$work.hours, gss.work$education.level,
            p.adj = "none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  gss.work$work.hours and gss.work$education.level
##
##             bachelor graduate
## graduate    0.37438  -
## high school 0.00207  0.00036
##
## P value adjustment method: none
```

```r
#alternatively... compare results to alpha
pairwise.t.test(gss.work$work.hours, gss.work$education.level,
            p.adj = "bonf")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  gss.work$work.hours and gss.work$education.level
##
##             bachelor graduate
## graduate    1.0000   -
## high school 0.0062   0.0011
##
## P value adjustment method: bonferroni
```

```r
#calculate group means
tapply(gss.work$work.hours, gss.work$education.level, mean)
```

```
##    bachelor    graduate high school
##    40.02348    38.71600    43.41174
```

5. **Corn Yield.** A large farm wants to try out a new type of fertilizer to evaluate whether it will improve the farm's corn production. The land is broken into plots that produce an average of 1,215 pounds of corn with a standard deviation of 94 pounds per plot. The owner is interested in detecting any average difference of at least 40 pounds per plot. Assume each plot of land gets treated with either the current fertilizer or the new fertilizer. Let $\alpha = 0.05$.

   a) How many plots of land would be needed for the experiment if the desired power level is 90%?

   At least 118 plots of land would be needed for each group for a power of 90%.

```r
#calculate sample size
power.t.test(n = NULL, delta = 40, sd = 94, sig.level = 0.05, power = .90,
         type = "two.sample", alternative = "two.sided")
```

```
##
##      Two-sample t test power calculation
##
##               n = 117.0232
##           delta = 40
##              sd = 94
##       sig.level = 0.05
##           power = 0.9
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

b) What would the power of the study be if 160 total plots of land were used for the experiment, with an equal number being treated with the current fertilizer versus new fertilizer?

If there were 80 plots of land for each group, then the power of the test would be 0.76.

```
#calculate power
power.t.test(n = 80, delta = 40, sd = 94,
         sig.level = 0.05, power = NULL,
         type = "two.sample",
         alternative = "two.sided")
```

```
##
##      Two-sample t test power calculation
##
##               n = 80
##           delta = 40
##              sd = 94
##       sig.level = 0.05
##           power = 0.7626783
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

c) Would the power of the test increase or decrease if the significance level were increased to $\alpha = 0.10$? Explain your answer.

The power of the test would increase. Increasing α increases the number of times that the null hypothesis is rejected, which both increases the number of rejections when the null is true and when the null is false. By definition, an increase in rejecting H_0 when the null is false is an increase in power.

To prove the point: changing the significance level to 0.10 increases power to 0.85 whereas before it was only 0.76.

```
power.t.test(n = 80, delta = 40, sd = 94, sig.level = 0.10, power = NULL,
         type = "two.sample", alternative = "two.sided")$power
```

```
## [1] 0.8496409
```