MODULE 17

2-SAMPLE T-TEST

Objectives:

1. Identify when a 2-Sample t-Test is appropriate.
2. Describe what a homogeneity of variance test is and why it is required within a 2-Sample t-Test.
3. Perform the 11 steps of a significance test in a 2-Sample t-Test situation.

Contents

While it is often useful to test whether a population mean differs from a specific value (i.e., with the 1-Sample t-Test of Module 16), there are many instances where interest is determining whether the means from two populations differ. For example, is there a difference in income between males and females, in test scores between students from high- and low-income families, in percent body fat between raccoons from southern and northern Wisconsin, or in amount of milk produced from cows provided with a hormone or a placebo. In all of these situations, interest is identifying if a difference in population means exists between the two populations (males and females, students from high- and low-income families, raccoons from southern and northern Wisconsin, cows given a hormone or a placebo). A **2-Sample t-Test** is used in these situations and is the subject of this module.

17.1 2-Sample t-Test Specifics

In a 2-Sample t-Test, $H_0 : \mu_1 = \mu_2$ states that the two population means are equal. This can be rewritten as $H_0 : \mu_1 - \mu_2 = 0$, because the difference between two population means should be zero if the two population means are equal. With this H_0, the "parameter" is $\mu_1 - \mu_2$ and the corresponding statistic is $\bar{x}_1 - \bar{x}_2$. Thus, a 2-Sample t-Test is focused on the difference in populatio means.

◇ **The parameter in a 2-Sample t-Test is the difference in population means ($\mu_1 - \mu_2$). The corresponding statistic is the difference in sample means ($\bar{x}_1 - \bar{x}_2$).**

When looking at the "general" test statistic formula (i.e., Equation (13.3.1)) of

$$\text{Test Statistic} = \frac{\text{Observed Statistic} - \text{Hypothesized Parameter}}{SE_{\text{Statistic}}}$$

it is apparent that the SE of $\bar{x}_1 - \bar{x}_2$ (i.e., the statistic) is needed. Unfortunately, the calculation of this standard error depends on whether the two population variances are equal or not. When the variances are approximately equal (discussed in Section 17.2), the standard error of $\bar{x}_1 - \bar{x}_2$ is

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

where n_1 and n_2 are the sample sizes for the two populations and s_p^2 is the "pooled sample variance" computed as a weighted average of the two sample variances (s_1^2 and s_2^2), or

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

The degrees-of-freedom for the 2-Sample t-Test with equal variances come from the denominator of the pooled variance calculation; i.e., $df = n_1 + n_2 - 2$. The specifics of the 2-Sample t-Test are in Table 17.1.

◇ **The s_p^2 calculation can be "checked" by determining if the value of s_p^2 is between s_1^2 and s_2^2 or if the value of $\sqrt{s_p^2}$ is between s_1 and s_2.**

Many times the 2-Sample t-Test will be used to test an alternative hypothesis of simply finding a difference between the two populations. However, if the null hypothesis is rejected in these instances (thus, identifying a significant difference between the two populations), then care should be taken to specifically describe how the two populations differ. If the statistic is negative, then the mean of the first population is lower than the mean of the second population and, if the statistic is positive, then the mean of the first population is larger than the mean of the second population. The values of the confidence region should be used to identify how much larger or smaller the mean from one population is compared to the mean of the other population.

◇ **Use the statistic and confidence region results to specifically determine which population has a larger or smaller mean when the null hypothesis of the 2-Sample t-Test has been rejected in favor of the "not equals" alternative hypothesis.**

Table 17.1. Characteristics of a 2-Sample t-Test with equal variances.

- **Hypothesis:** $H_0 : \mu_1 - \mu_2 = 0$
- **Statistic:** $\bar{x}_1 - \bar{x}_2$
- **Test Statistic:** $t = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$ where $s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$.
- **Confidence Region:** $\bar{x}_1 - \bar{x}_2 + t^*\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$
- **df:** $n_1 + n_2 - 2$
- **Assumptions:** $n_1+n_2 > 40$, $n_1+n_2-2 > 15$ and **each sample** (i.e., histogram) is not strongly skewed, OR **each sample** is normally distributed.
- **When to Use:** Quantitative response, two populations, individuals are independent between populations.

17.2 Testing for Equal Variances

As noted above, the methods of a 2-Sample t-Test differ depending on whether the two population variances are equal or not. This should present a problem to you because the population variances are parameters and are typically not known.[1] The question of whether these parameters are equal or not will be handled with a hypothesis test, as has been done with all other questions about parameters.

◇ A hypothesis test must be used to determine if two population variances are equal.

A Levene's Test is used to determine whether two population variances are equal. The specifics of the Levene's test are not examined in detail here, rather you only need to know that $H_0 : \sigma_1^2 = \sigma_2^2$ is tested against $H_A : \sigma_1^2 \neq \sigma_2^2$. We will rely on computer software to compute the p-value for this test (without further detail). If the Levene's Test $p - value < \alpha$, then H_0 is rejected and the population variances are considered unequal. If the $p - value > \alpha$, then H_0 is not rejected and the population variances are at least approximately equal.

◇ If the Levene's Test p-value is greater than α, then H_0 is not rejected and the population variances are at least approximately equal.

[1] Actually, the population variances don't have to be known exactly, it just needs to be known whether they are equal or not.

17.2.1 Example - Corn and Fertilizers

Consider the following situation,

> *An agricultural researcher thought that corn plants grown in pots exposed to a certain type of synthetic fertilizer would grow taller than plants exposed to an organic fertilizer. To collect data to test this idea, he grew 50 corn plants in individual pots – 25 were treated with organic fertilizer and 25 were treated with synthetic fertilizer. Each pot contained soil from a well-mixed common source and was planted in the same greenhouse. Each plant was similar in all regards (similar genetics, age, etc.). Use the results (heights of individual plants) in Table 17.2 to test the researcher's hypothesis (at the $\alpha = 0.05$ level).*

Table 17.2. Summary statistics of the corn plant height in two treatments.

```
               Synthetic  Organic
means:           51.46     47.49
SD:              5.975     6.721
Levene's Test:  p=0.1341
```

The 11-steps (Section 15.1) for a hypothesis test for this example are as follows:

1. α=0.05.
2. $H_0 : \mu_s - \mu_o = 0$ vs $H_A : \mu_s - \mu_o > 0$, where μ is the mean plant height, s represents the synthetic fertilizer, and o represents the organic fertilizer. [*Note that positive differences represent larger values for the synthetic fertilizer and, with this, the H_A represents the synthetic fertilizer producing taller plants.*]
3. A 2-Sample t-Test is required because (i) a quantitative variable (height) was measured, (ii) two populations were sampled (synthetic and organic fertilizers), and (iii) plants in the two populations were **IN**dependent as the plants were not paired, plants were not tested over time, etc.
4. The data appear to be part of an experiment (the researcher imposed the treatments on the plants) with no clear indication of random selection of plants or random allocation of plants to the two treatments.
5. (i) $n_s + n_o$=50> 40, (ii) the individuals in the two populations are independent as discussed above, and (iii) the two population variances appear to be equal because the Levene's Test p-value of 0.1341 is > 0.05.
6. $\bar{x}_s - \bar{x}_0$= 51.46 − 47.49= 3.97. Additionally,

   $$s_p^2 = \frac{(25-1)5.975^2 + (25-1)6.721^2}{25+25-2} = 40.44$$

   and

   $$SE_{\bar{x}_s - \bar{x}_o} = \sqrt{40.44\left(\frac{1}{25} + \frac{1}{25}\right)} = 1.799$$

7. t=$\frac{3.97-0}{1.799}$=$\frac{3.97}{1.799}$=2.207 with $25 + 25 - 2 = 48$ df.
8. p-value=0.0161.
9. The H_0 is rejected because the p-value $< \alpha = 0.05$.
10. The average height of the corn plants appears to be greater for plants grown with synthetic fertilizer than for plants grown with organic fertilizer.
11. I am 95% confident that plants grown with synthetic fertilizer are more than 0.95 cm taller, on average, than plants grown with the organic fertilizer. [Note $3.97 - 1.677 * 1.799$=$3.97 - 3.02$=0.95.]

R Appendix:

```
( distrib(2.207,distrib="t",df=48,lower.tail=FALSE) )
( distrib(0.95,distrib="t",type="q",df=48,lower.tail=FALSE) )
```

17.2.2 Example - Music and Anxiety

Consider the following situation,

> *An oral surgeon conducted an experiment to determine if background music decreased the anxiety level of patients during a tooth extraction. Over a one-month period, 32 patients had a tooth removed while listening to music and 36 had a tooth removed with no music to listen to. Each patient was given a questionnaire following the extraction. Answers to the questionnaire were converted to a numeric scale to measure the patient's level of anxiety (larger numbers mean greater anxiety). For those given background music, the mean anxiety level was 4.2 (with a standard deviation of 1.2), while the group without music had a mean of 5.9 (with a standard deviation of 1.9). The surgeon also reported a Levene's test p-value of 0.089. Test the surgeon's hypothesis using* $\alpha = 0.05$.

The 11-steps (Section 15.1) for a full hypothesis test for this example are as follows:

1. α=0.05.
2. $H_0 : \mu_w - \mu_o = 0$ vs $H_A : \mu_w - \mu_o < 0$, where μ is the mean anxiety level, w represents patients "with", and o represents "without" music. [*Note that negative numbers represent lower anxiety values in patients in the "with music" treatment. Thus, H_A suggests lower anxiety in paients with music.*]
3. A 2-Sample t-Test is required because (i) a quantitative variable (anxiety level) was measured, (ii) two populations were sampled (music or no music), and (iii) individuals in the two populations are independent (i.e., they were not paired, were not otherwise related, etc.).
4. The data appear to be an experiment as the music treatment was imparted by the surgion, but there is no obvious random selection or allocation in this study.
5. (i) $n_w + n_o$=68> 40, (ii) individuals in the two populations are independent as described above, and (iii)the two population variances appear to be equal because the Levene's Test p-value of 0.089 is α).
6. $\bar{x}_w - \bar{x}_o$= 4.2 − 5.9= −1.7. Additionally,

   $$s_p^2 = \frac{(32-1)1.2^2 + (36-1)1.9^2}{32+36-2} = 2.59$$

   and

   $$SE_{\bar{x}_w - \bar{x}_o} = \sqrt{2.59\left(\frac{1}{32} + \frac{1}{36}\right)} = 0.391$$

7. t=$\frac{-1.7-0}{0.391}$=−4.348 with $32 + 36 - 2 = 66$ df.
8. p-value=< 0.00005
9. H_0 is rejected because the $p - value < \alpha = 0.05$.
10. The mean anxiety level appeared to be lower when music was played for the patients.
11. I am 95% confident that the mean anxiety level is more than -1.05 points lower, on average, when music is played than when it is not. [Note $-1.7 + 1.668 * 0.391$=$-1.7 + 0.65$=-1.05.

R Appendix:

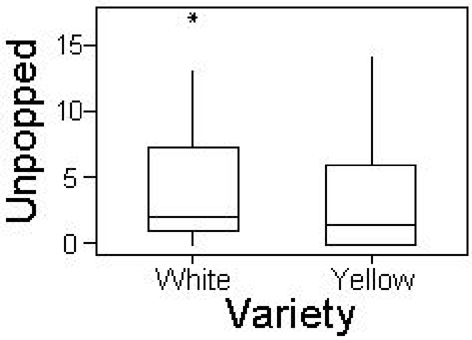
```
( distrib(-4.348,distrib="t",df=66) )
( distrib(0.95,distrib="t",type="q",df=66) )
```

Review Exercises

17.1 Erville Redenbacher wanted to see if the number of unpopped kernels differed between yellow and white varieties of his grandpa's famous popcorn. To test this, he would put 100 kernels of either white or yellow popcorn into a standard air popper, pop the corn until no "pops" were heard, and then count the number of unpopped kernels. He tested 30 randomly selected groups of 100 kernels for both white and yellow varieties (Erville is very thorough). Use the results below to test, at the 10% level, Erville's hypothesis. Answer

```
Variable  N  Mean  Median  StDev  SE Mean
White    30 4.267   2.000  4.456    0.814
Yellow   30 3.567   1.500  4.485    0.819

Levene's Test -- P-Value = 0.972
```

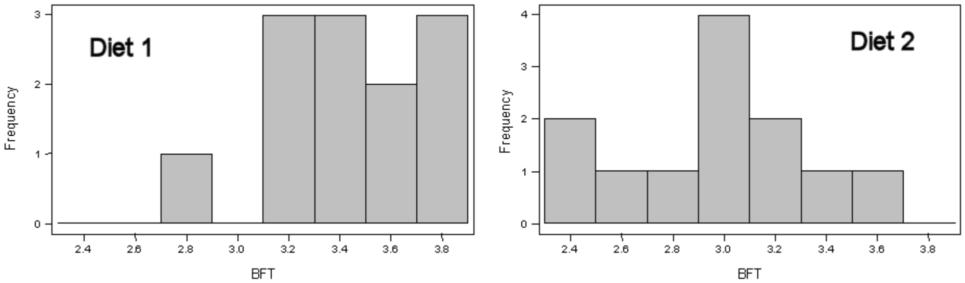


17.2 A study was performed in order to evaluate the effectiveness of two devices for improving the efficiency of gas home-heating systems. Energy consumption in houses was measured after one of the two devices was installed. The two devices were an electric vent damper (DampVent=Electric) and a thermally activated vent damper (DampVent=ThermAct). Energy consumption (in BTUs) was measured for a variety of houses fitted with the two devices. Compare, at the 10% level, the effectiveness of these two devices by determining if a difference exists in energy consumption between houses fitted with the devices. Note that Levene's test p-value is 0.996. Answer

```
Variable DampVent   N    Mean   Median   StDev SE Mean Minimum Maximum     Q1     Q3
BTU.In   Electric  40   9.908    9.590   3.020   0.477   4.000  18.260  7.885 11.555
         ThermAct  50  10.143   10.290   2.767   0.391   2.970  16.060  8.127 12.212
```

17.3 A pig diet manufacturer wants to determine if the backfat thickness differs between pigs raised on two different diets. Backfat thickness is an indicator of pork quality; smaller thicknesses mean better quality. A group of 24 pigs was randomly allocated to two groups which differed only in the diet received. Test the results from this experiment to see if a difference in backfat thickness is evident at the $\alpha = 0.05$ level. Note that Levene's test p-value is 0.532. Answer

```
Var Diet  N   Mean Median  StDev SE Mean  Min  Max
BFT    1 12  3.420  3.390  0.295  0.0850  2.87 3.87
       2 12  2.989  3.035  0.375  0.108   2.40 3.62
```



17.3 2-Sample t-Tests in R

17.3.1 Data Format

The data for a 2-Sample t-Test must be entered in stacked format, as described in Section 4.3.2. In stacked format the measurements are in one column and labels for which group the measurement was recorded from is in another column. Thus, each row corresponds to a measurement and the group for a single individual. As an example, BOD measurements from either the inlet or outlet to an aquaculture facility are shown below. These data are stacked because each row corresponds to one individual (a water sample) with one column of (BOD) measurements and another for which group the individual belongs.

```
   BOD    src
 6.782  inlet
 5.809  inlet
 6.849  inlet
 8.545 outlet
 8.063 outlet
 8.001 outlet
```

Δ **Stacked Data**: Data where the quantitative measurements of two groups are "stacked" on top of each other and a second variable is used to record to which group the measurement belongs.

◇ **Stacked data is the preferred format for 2-Sample data, because each vector corresponds to a variable and each row corresponds to only one individual.**

17.3.2 Levene's Test

Before conducting a 2-Sample t-Test, the assumption of equal population variances must be tested with Levene's test. The Levene's test is computed in R with `levenesTest()`, where the first argument is a model formula of the form `response~group` where `response` represents the quantitative measurements and `group` represents the group factor variable.[2] As per usual, the corresponding data.frame of stacked data is given to `data=`.

[2] This is the same model formula introduced in Section 5.8 for summarizing multiple groups of data.

17.3.3 2-Sample t-Test

A 2-Sample t-Test is computed in R with `t.test()`, where the first argument is the same formula as in `levenesTest()` (and, thus, the same `data=`. Additionally, the following arguments may be specified for a 2-Sample t-Test.

- `mu=`: The specific value in H_0. For 2-Sample t-Tests this is usuall 0, which is the default. Thus, `mu=` is rarely used for a 2-Sample t-Test.
- `alt=`: A character string that indicates the type of H_A (i.e., `alt="two.sided"` (default), `alt="greater"`, or `alt="less"`).
- `conf.level=`: The level of confidence (`0.95` is the default) to be used when constructing the confidence region for $\mu_1 - \mu_2$.
- `var.equal=`: A logical value that indicates whether the two population variances should be considered to be equal or not. If `var.equal=TRUE`, then the pooled sample variance is calculated and used in the standard error. The default value is to assume unequal variances; thus, this argument must be set to `TRUE` if the result from `levenesTest()` suggests that the population variances are equal.

◇ **The `var.equal=TRUE` argument must be used in `t.test()` to assume equal variances. This is NOT the default setting in R.**

R computes the difference among populations as the alphabetically "first" level minus the alphabetically "second" level. For example, if the two levels are *inlet* and *outlet*, then R will compute $\mu_{inlet} - \mu_{outlet}$. If this is not the order you want, then you need to change the order of the levels by using `levels=` in `factor()` (as described in Modules 6 and 9). For example, the order of the levels of *src* in the *aqua* data.frame is changed below.

```
> aqua$src <- factor(aqua$src,levels=c("outlet","inlet"))
> levels(aqua$src)
[1] "outlet" "inlet"
```

17.3.4 Example - BOD in Aquaculture Water

Consider the following situation (which was examined in parts above),

> *An aquaculture farm takes water from a stream and returns it to the stream after it has circulated through the fish tanks. The owner is concerned that the water may contain heightened levels of organic matter when it is released into the stream after it has circulated in the tanks. He has taken steps to reduce this possibility, i.e., circulated the water rather quickly through the tanks, but is still concerned about the increase in organic material in the effluent. To determine if this is true, he takes samples of the water at the intake and, at other times, downstream from the outlet and measures the biological oxygen demand (BOD) as a measure of the organics in the effluent (a higher BOD at the outlet would imply that organics are taken up from the tanks). The farmers data are recorded in BOD.csv.* Test for any evidence (i.e., at the 10% level) of support for the farmer's concern.

The 11-steps (Section 15.1) for a hypothesis test for this example are as follows:

1. α=0.10.
2. $H_0 : \mu_{outlet} - \mu_{inlet} = 0$ vs $H_A : \mu_{outlet} - \mu_{inlet} > 0$, where μ is the mean BOD, *outlet* represents the outlet source, and *inlet* represents the inlet source. [*Positive differences represent larger values at the outlet, which implies that BOD is higher in the water released from the facility. Thus, the H_A represents the farmer's concern. Further note that order of subtract could have been reversed and then the farmer's concern would require a "less than" H_A. This is simply a matter of choice. However, note that the order of the levels had to be changed to use my choice of hypotheses.*]
3. A 2-Sample t-Test is required because (i) a quantitative variable (BOD level) was measured, (ii) two populations were sampled (outlet or inlet), and (ii) the sets of samples were **IN**dependent (note that it said that the outlet samples came from different times then the inlet samples).
4. The data appear to be part of an observational study with no obvious randomization.
5. (i) n =20> 15 and the histograms (Figure 17.1) are inconclusive about the shape because of the small sample sizes in each group (it appears that the *inlet* data is not strongly skewed whereas the *outlet* data is skewed, which may invalidate the results of this hypothesis test; however, I continued to make a complete example), (ii) individuals in the two samples are independent as discussed above, and (iii) the variances appear to be equal because the Levene's test p-value ($p = 0.5913$) is larger than α.
6. $\bar{x}_{outlet}$-$\bar{x}_{inlet}$=8.69-6.65=2.03 (Table 17.3).
7. t=8.994 with 18 df (Table 17.3).
8. p-value=$<$ 0.00005 (Table 17.3).
9. H_0 is rejected because the p-value $< \alpha$.
10. The average BOD is greater at the outlet than at the inlet to the aquaculture facility. Thus, the aquaculture facility appears to add to the biological oxygen demand of the water and the farmer's concern is warranted.
11. I am 90% confident that the BOD measurement at the outlet is AT LEAST 1.73 GREATER than the BOD measurement at the inlet (Table 17.3).

R Appendix:

```
aqua <- read.csv("data/BOD.csv")
aqua$src <- factor(aqua$src,levels=c("outlet","inlet"))
hist(BOD~src,data=aqua,main="",xlab="BOD Measurement")
levenesTest(BOD~src,data=aqua)
( aqua.t <- t.test(BOD~src,data=aqua,var.equal=TRUE,alt="greater",conf.level=0.90) )
plot(aqua.t)
```

Table 17.3. Results from the 2-Sample t-Test for differences in BOD between the inlet and outlet of an aquaculture facility.

```
t = 8.994, df = 18, p-value = 2.224e-08
90 percent confidence interval:
 1.732704      Inf
sample estimates:
mean in group outlet  mean in group inlet
              8.6873               6.6538
```

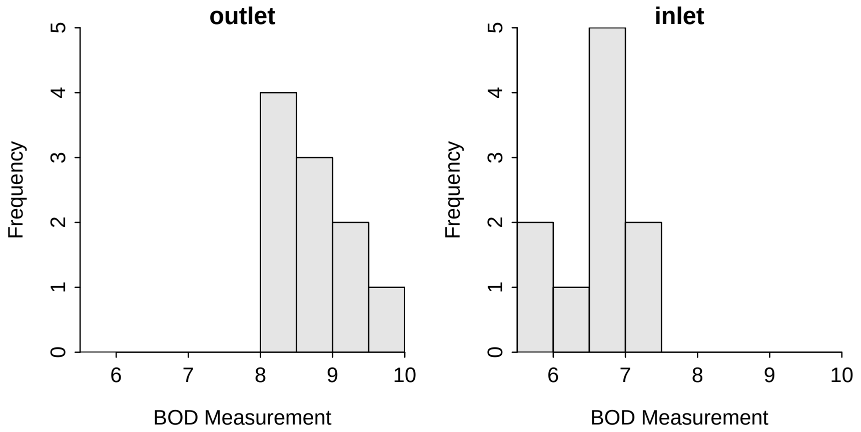


Figure 17.1. Histogram of the BOD measurements at the outlet and inlet of the aquaculture facility.

Review Exercises

17.4 A study[3] examined the effectiveness of foil-lined milk cartons to reduce "leakage" of dioxins from the carton to the milk (dioxins were found in milk cartons due to the bleaching process). The dioxin content (parts per thousand, ppt) in milk from 50 unlined and 50 lined cartons of milk are recorded in MilkCartons.csv. Determine, at the 1% level, if lining the cartons with foil significantly reduced the amount of dioxin in the milk. *Answer*

17.5 The University of North Carolina math department is noted for "giving out" low grades. To examine this, the GPA from a random sample of 22 math classes and 29 "other" university classes (from the last year) are in UNCGrades.csv. Determine, at the 10% level, if grades are significantly lower in math than in other classes. *Answer*

17.6 A health commissioner needs to determine if the number of hours worked per week by medical interns differs between two cities. To examine this, the commissioner found the mean number of hours worked by interns in the first city for a random sample of 13 weeks and the same for a random sample of 16 weeks from the second city. Her data are in MedInternHrs.csv. Determine if the hours worked by the interns differs, at the 10% level, between the two cities. *Answer*

17.7 Agronomists are interested in determining conditions that increase crop yield. In one experiment, 80 one-acre plots of corn were randomly divided into two groups of 40 plots. An insecticide was used on each plot in one group and sterilized male individuals of an insect pest were released on each of the other plots. The resulting yields are recorded in CropYield.csv. Is there a difference, at the 10% level, in yield between the two treatments. *Answer*

17.8 Templer's Death Anxiety Scale (DAS) is a measure of an individual's anxiety concerning death. Robbins (1990) recorded the DAS score for 25 organ donors and 69 non-organ donors in DeathAnxiety.csv. Determine, at the 1% level, if there is a difference in anxiety levels concerning death between organ and non-organ donors. *Answer*

[3] Data was recreated from Blaisdell 1998.