MODULE 18

2-SAMPLE T-TEST

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 17), there are many instances where interest is in whether means from two populations differ. For example, is there a difference in mean income between males and females, in mean test scores between students from high- and low-income families, in mean percent body fat between raccoons from southern and northern Wisconsin, or in mean 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 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.

18.1 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 population means.

When looking at the "general" test statistic formula (i.e., Equation (14.3.1)) 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 18.2), 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 18.1.

Table 18.1. 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:** $n_1+n_2 \geq 40$, $n_1+n_2-2 \geq 15$ 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.

◇ **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.**

A 2-Sample t-Test is often 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.

18.2 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 is answered with a hypothesis test, as has been done with all other questions about parameters.

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 use 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 considered equal.

18.2.1 Example - Corn and Fertilizers

Below are the 11-steps (Section 16.1) for completing a full hypothesis test for 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 18.2 to test the researcher's hypothesis at the 5% level.*

Table 18.2. 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
```

1. $\alpha = 0.05$.
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 synthetic fertilizer, and o represents organic fertilizer. [*Note that positive differences represent larger values for synthetic fertilizer; thus, H_A represents 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) individuals in the two populations are independent as discussed above, and (iii) the population variances appear to be equal because the Levene's Test p-value (0.1341) is $> \alpha$.
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

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

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

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 < α.
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:

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

18.2.2 Example - Music and Anxiety

Below are the 11-steps (Section 16.1) for completing a full hypothesis test for the following situation:

> *An oral surgeon conducted an experiment to determine if background music decreased the anxiety level of patients during tooth extraction. Over a one-month period, 32 patients had a tooth removed while listening to music and 36 had a tooth removed without listening to music. 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 more 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 at the 5% level.*

1. $\alpha = 0.05$.
2. $H_0 : \mu_w - \mu_{wo} = 0$ vs $H_A : \mu_w - \mu_{wo} < 0$, where μ is the mean anxiety level, w represents patients "with", and wo 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) $n_w + n_{wo} = 68 > 40$, (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 greater than α.
6. $\bar{x}_w - \bar{x}_{wo} = 4.2 - 5.9 =$ -1.7. Additionally,

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

and

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

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 $< \alpha$.
10. The mean anxiety level appears 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:

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

18.3 2-Sample t-Tests in R

18.3.1 Data Format

Data must be in stacked format (as described in Section 4.3.2) for a 2-Sample t-Test. Stacked data has measurements in one column and group labels for the measurement 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 column for which group the individual belongs.

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

18.3.2 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 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] The data.frame containing `response` and `group` is given in `data=`.

18.3.3 2-Sample t-Test

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

- `mu=`: The specific value in H_0. For a 2-Sample t-Test this is usually 0, which is the default.
- `alt=`: A string that indicates the type of H_A (i.e., `"two.sided"` (default), `"greater"`, or `"less"`).
- `conf.level=`: The level of confidence (default is `0.95`) used for the confidence region of $\mu_1 - \mu_2$.
- `var.equal=`: A logical value that indicates whether the two population variances should be considered equal or not. If `TRUE`, then the pooled sample variance is calculated and used in the standard error. The default `FALSE`, to assume UNequal variances.

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

[2] This is the same model formula introduced in Section 5.3 for summarizing multiple groups of data.

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 $\bar{x}_{outlet} - \bar{x}_{inlet}$. 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 7 and 10). 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"
```

18.3.4 Example - BOD in Aquaculture Water

Below are the 11-steps (Section 16.1) for completing a full hypothesis test for the following situation:

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

1. $\alpha = 0.10$.
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, H_A represents the owner's concern. Further note that the order of subtraction could have been reversed such that the owner's concern would require a "less than" H_A. This is simply a matter of choice. However, note that the order of the levels has to be changed in R 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 and 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 18.1) are inconclusive about the shape because of the small sample size 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 (=0.5913) is greater than α.
6. $\bar{x}_{outlet} - \bar{x}_{inlet} = 8.69 - 6.65 = 2.03$ (Table 18.3).
7. $t = 8.994$ with 18 df (Table 18.3).
8. p-value < 0.00005 (Table 18.3).
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 mean BOD measurement at the outlet is AT LEAST 1.73 GREATER than the mean BOD measurement at the inlet (Table 18.3).

Table 18.3. 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
```

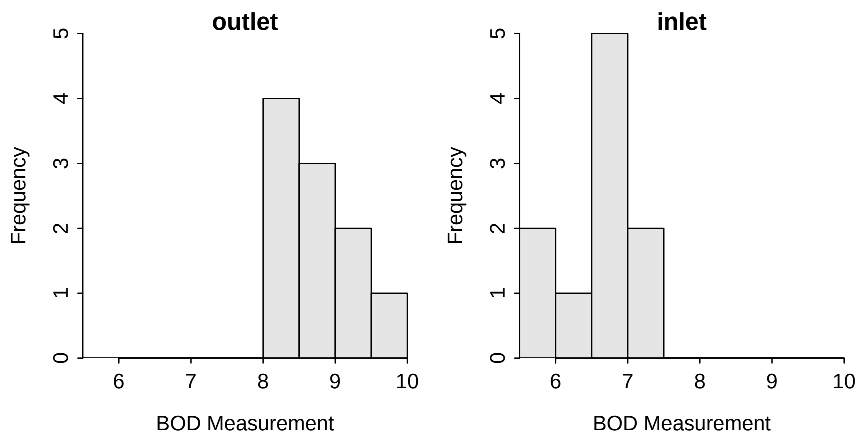

Figure 18.1. Histogram of the BOD measurements at the outlet and inlet of the aquaculture facility.

R Appendix:

```
aqua <- read.csv("data/BOD.csv")
aqua$src <- factor(aqua$src,levels=c("outlet","inlet"))
hist(BOD~src,data=aqua,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) )
```