Robust t Tests

James H. Steiger

Department of Psychology and Human Development
Vanderbilt University

Robust t Tests

Introduction

Statistical Assumptions for the *t*-Test

- In Psychology 310, we discussed the *statistical assumptions* of the classic multi-sample t statistics, of which the two-sample independent sample t is the simplest and best known special case.
  1. *Independence of observations.* Each observation is independent. As we emphasized in Psychology 310, the classic formula for the sampling variance of the sample mean, $\text{Var}(\overline{X}_\bullet) = \sigma^2/n$, is based on this assumption.
  2. *Normality.* The distribution of the populations is assumed to be normal.
  3. *Homogeneity of variances.* The populations are assumed to have equal variances.
- We need to consider, in turn,
  1. How violations of these assumptions affect performance of the *t*-test.
  2. What methods are available to produce reasonable inferential performance when assumptions are violated.
  3. How to detect violations of assumptions.

Effect of Violations

Independence

- If the n observations are independent, then $\overline{X}_\bullet$ has a sampling variance of σ^2/n. Otherwise, the sampling variance may be quite different.
- Since most classic tests assume the formula σ^2/n is correct, they can be seriously in error if this assumption is violated.
- Exactly what the affect of the error is depends on the precise nature of the dependency.
- *If* the pattern of dependency is known, it may be possible to correct for it, using linear combination theory as taught in Psychology 310.

Effect of Violations

Normality

- A key fact about the normal distribution is that the sample mean and sample variance of a set of observations taken randomly from a normal population are independent.
- This independence of the mean and variance are crucial in the derivation of Student's t distribution.
- When populations are not normal, this lack of independence can lead to poor performance of the t-test.

Effect of Violations

Normality

- Violations of normality can occur in several distinct ways.
- The general shape of the distribution can be skewed, in some cases for obvious reasons related to the nature of the measurement process.
- There can be *contamination* by *outliers*. These extreme and unusual observations lead to the distribution having tails that are much longer than seen with a normal distribution.
- Yet, if the contamination probability is small, it may be difficult to diagnose outlier problems when they occur. For example, are the outliers the result of:
  1. A mixture of two or more processes (or subgroups) that characterize the population of interest?
  2. A random measurement error?

Effect of Violations

Normality

- High skewness or kurtosis can lead to Type I error rates that are either much higher or much lower than the nominal rates.
- Contamination by outliers can lead to a significant loss of power when the null hypothesis is false.

Effect of Violations

Homogeneity of Variances

- Consider the two-sample independent sample t-statistic.
- In Psychology 310, we saw that the denominator of the statistic explicitly assumes equal variances.
- Recall that, with independent samples, the variance of $_1-_2$ is

$$\text{Var}(_1-_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \tag{1}$$

- The t statistic replaces this formula with one that assumes equal variances, i.e.,

$$\text{Var}(_1-_2) = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\sigma^2 \tag{2}$$

and then substitutes the estimate $\hat{\sigma}^2$ for σ^2, where

$$\hat{\sigma}^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} \tag{3}$$

Effect of Violations

Homogeneity of Variances

- Notice that, in the preceding formula, we are *essentially* substituting the (weighted) average of the two variances for each variance in the formula for the sampling variance of $_1 - _2$.
- If the assumption of equal variances is correct, the resulting formula will be a consistent estimate of the correct quantity.
- What will the effect be if the assumption of equal variances is incorrect?
- How can we approximate the impact of a violation of the equal variances assumption on the true Type I error rate of the t-test when the null hypothesis is true?

Effect of Violations

Homogeneity of Variances

- One simplified approach would be two assume that there is no sampling error in the sample variances, i.e., that $s_1^2 = \sigma_1^2$ and $s_2^2 = \sigma_2^2$, and measure the result of the violation of assumptions.
- For example, suppose $\sigma_1^2 = 40$, and $\sigma_2^2 = 10$, while $n_1 = 10$ and $n_2 = 20$. What will the approximate effect on the true α?

Effect of Violations

Homogeneity of Variances

- Using our simplified assumption that the sample variances would perfectly estimate the population variances, let's compute the ratio of the obtained denominator to the correct denominator.
- First, let's compute $\hat{\sigma}^2$.

$$\begin{aligned}\hat{\sigma}^2 &= \frac{(10-1)40+(20-1)(10)}{10+20-2} \\ &= \frac{360+190}{28} \\ &= 19.64286\end{aligned}$$

- The obtained denominator is then

$$\begin{aligned}\sqrt{\hat{\text{Var}}(_1-_2)} &= \sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\hat{\sigma}^2} \\ &= \sqrt{\left(\frac{1}{10}+\frac{1}{20}\right)19.64286} \\ &= \sqrt{2.946429} \\ &= 1.716516\end{aligned}$$

Effect of Violations

Homogeneity of Variances

- However, the correct denominator is

$$\sqrt{\frac{40}{10} + \frac{10}{20}} = \sqrt{4.5} = 2.12132 \tag{4}$$

- The obtained denominator is considerably smaller than it should be. So the t statistic will, in general, be larger in value than it should be, and will therefore reject more often than it should.
- The critical value of the t statistic with 28 degrees of freedom is

```
> qt(.975,28)

[1] 2.048407
```

- Since obtained values of the t statistic are, in effect, expanded by the ratio 2.12132/1.71516, the true α can be approximated as the area outside absolute t values of

```
> 1.71516/2.12132 * qt(.975,28)

[1] 1.656207
```

- This is

```
> 2*pt(1.71516/2.12132 * qt(.025,28),28 )

[1] 0.1088459
```

Effect of Violations

Homogeneity of Variances

- The above estimate of .109 was obtained with a simplifying assumption, and is an approximation.
- An alternative approach is Monte Carlo simulation.

```
> js.t.test <- function(data1,data2,alpha){
+ n1 <- length(data1)
+ n2 <- length(data2)
+ df <- n1 + n2 - 2
+ crit.t <- qt(1-alpha/2,df)
+ sigma.hat.squared <- ((n1-1)*var(data1) + (n2-1)*var(data2))/df
+ t.obtained <- (mean(data1)-mean(data2))/sqrt(((1/n1)+(1/n2))*sigma.hat.squ
+ return(c(t.obtained,abs(t.obtained) > crit.t))
+ }
> set.seed(12345)
> rowMeans(replicate(10000,js.t.test(rnorm(10,0,sqrt(40)),rnorm(20,0,sqrt(10

[1] 0.01248439 0.11550000
```

- This confirms that the true α is more than twice as large as the nominal α of .05.

Effect of Violations

Homogeneity of Variances

- Using the approaches just demonstrated, we can verify the following general principles, as described by RDASA3 (p.137):
  1. If the two sample sizes are equal, the difference between nominal and true α will be minimal, unless sample sizes are really small and the variance ratio really large.
  2. If the two samples sizes are unequal, and the variances are inversely related to sample sizes, then the true α will substantially exceed the nominal α.
  3. If the two sample sizes are unequal, and the variances are directly related to sample sizes, then the true α will be substantially lower than the nominal α.

Dealing with Non-Normality

- When data show a recognized non-normal distribution, one has recourse to several options:
  1. *Do nothing.* If violation of normality is not severe, the *t*-test may be reasonably robust.
  2. *Transform the data.* This seems especially justifiable if the data have a similar non-normal shape. With certain kinds of shapes, certain transformations will convert the distributions to be closer to normality. However, this approach is generally not recommended, for a variety of reasons.
  3. *Trim the data.* By trimming a percentage of the more extreme cases from the data, the skewness and kurtosis may be brought more into line with those of a normal distribution.
  4. *Use a non-parametric procedure.* Tests for equality of means that do not assume normality are available. However, they generally assume that the two samples have equal distributions, not that they simply have equal means (or medians).
- Although the "jury is still out" on these matters, a number of authors writing on robustness in social science statistical journals (e.g., Algina, Keselman, Lix, Wilcox) have promoted the use of trimmed means.
- In the preceding lecture module, we described a single sample test and confidence interval using a trimmed mean.
- One could examine the data and then choose a trimming proportion γ, but many authors recommend using a fixed value of $\gamma = 0.20$ to avoid the general problems connected with *post hoc* analysis.

Two-Sample Trimmed Mean Test

- In their Tech Note 7.1, MWL discuss an analog of the two-sample independent sample t-test, assuming equal population variances.
- This formula has an obvious error, in that it allows the sample sizes to be unequal while tacitly assuming that the number of trimmed observations taken from each tail is the same (i.e., k in the MWL notation) in each sample.
- Of course, if the sample sizes are unequal and γ remains fixed in the two samples, then the number of trimmed observations will not be the same.
- The corrected formula is given below.

Two-Sample Trimmed Mean Test

- Define the number of observations trimmed from each tail of the groups as g_1 and g_2. Define the "effective n" for group j as $h_j = n_j - 2g_j$, and the Winsorized sum of squared deviations for each group as $SSW_j = (n_j - 1)s^2_{w_j}$.
- The pooled Winsorized variance estimate is

$$\hat{\sigma}^2_w = \frac{SSW_1 + SSW_2}{h_1 + h_2 - 2} \tag{5}$$

- Then the t statistic is

$$t_{h_1+h_2-2} = \frac{\bar{Y}_{t1} - \bar{Y}_{t2}}{\sqrt{(\frac{1}{h_1} + \frac{1}{h_2})\hat{\sigma}^2_w}} \tag{6}$$

Dealing with Unequal Variances

The Welch Test

- With unequal variances, a standard approach, recommended in many textbooks, is to employ the Welch test, which can be generalized to the analysis of variance setting.
- In the case of the two-sample t statistic, the Welch test employs a modified test statistic,

$$t' = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{7}$$

and modified degrees of freedom

$$df' = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}} \tag{8}$$

Dealing with Unequal Variances

The Welch Test

- The Welch test is implemented in the R function `oneway.test()`.
- It is also implemented in the MASS library R function `t.test`.
- Suppose we had an experimental population that if, untreated, would have a normal distribution with a mean of 50 and a standard deviation of 10.
- However, if exposed to a treatment, the variance of scores would be increased by a factor of 9, while the mean would remain unchanged.
- From our principles of linear transformation, we know that the resulting experimental group would have a mean of 50 and a standard deviation of 30.
- Suppose further that the experimental treatment is very expensive to administer, and so the ratio of experimental to control subjects in the study is 1 to 5, and that you end up with 100 control subjects and just 20 experimental subjects.
- Let's simulate some data from this situation.

Dealing with Unequal Variances

The Welch Test

- We begin by simulating some data from the control group and experimental group, and combining them into a single matrix.

```
>   make.data <-function(){
+   y <- c(rnorm(100,50,10),rnorm(20,50,30))
+   group <- c(rep(1,100),rep(2,20))
+   data <- cbind(y,group)
+   return(data)
+   }
>   tdata<-make.data()
>   t.test(y~group,var.equal=TRUE,data=tdata)


Two Sample t-test

data:  y by group
t = -0.3953, df = 118, p-value = 0.6933
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -8.917425  5.949373
sample estimates:
mean in group 1 mean in group 2
       49.28534        50.76937
```

- Next, we perform the ordinary two-sample t-test.

Dealing with Unequal Variances

The Welch Test

- Note that we must specify equal variances for the standard test to be done. The default is to always use the Welch test.

```
> t.test(y~group,var.equal=TRUE,data=tdata)


Two Sample t-test

data:  y by group
t = -0.3953, df = 118, p-value = 0.6933
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -8.917425  5.949373
sample estimates:
mean in group 1 mean in group 2
       49.28534        50.76937
```

- Compare the above results to those obtained from the Welch test.

```
> t.test(y~group,data=tdata)


Welch Two Sample t-test

data:  y by group
t = -0.2331, df = 20.269, p-value = 0.818
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -14.75195  11.78390
sample estimates:
mean in group 1 mean in group 2
       49.28534        50.76937
```

Dealing with Unequal Variances

The Welch Test

- How do the two tests perform in general under these circumstances?
- Simulation can tell us. First, we have to learn how to get the *p*-value from our output.
- We can do that by using the `str()` function to examine the contents of the `t.test` object.

```
> output <- t.test(y~group,var.equal=TRUE,data=make.data())
> str(output)

List of 9
 $ statistic  : Named num -2.91
  ..- attr(*, "names")= chr "t"
 $ parameter  : Named num 118
  ..- attr(*, "names")= chr "df"
 $ p.value    : num 0.00437
 $ conf.int   : atomic [1:2] -18.06 -3.42
  ..- attr(*, "conf.level")= num 0.95
 $ estimate   : Named num [1:2] 48.7 59.4
  ..- attr(*, "names")= chr [1:2] "mean in group 1" "mean in group 2"
 $ null.value : Named num 0
  ..- attr(*, "names")= chr "difference in means"
 $ alternative: chr "two.sided"
 $ method     : chr " Two Sample t-test"
 $ data.name  : chr "y by group"
 - attr(*, "class")= chr "htest"
```

Dealing with Unequal Variances

The Welch Test

- Soon, we're on our way to a quick simulation.

```
> replicate(5,t.test(y~group,var.equal=TRUE,
+                    data=make.data())$p.value)

[1] 0.12059612 0.01241230 0.01288572 0.12910297 0.34593059

> replicate(5,t.test(y~group,var.equal=TRUE,
+                    data=make.data())$p.value<.05)

[1]  TRUE FALSE FALSE FALSE FALSE
```

Dealing with Unequal Variances

The Welch Test

- As we see below, clearly the Welch test controls Type I error in this situation much more effectively than the standard *t*-test.

```
> set.seed(12345)
> mean(replicate(2000,t.test(y~group,
+ var.equal=TRUE,data=make.data())$p.value<.05))

[1] 0.288

> set.seed(12345)
> mean(replicate(2000,t.test(y~group,
+ var.equal=FALSE,data=make.data())$p.value<.05))

[1] 0.0475
```

Dealing with Unequal Variances

The Welch Test

- However, it is important to realize that the Welch test is not an exact test. As we shall see later, it does not perform well in some extreme situations.

General Testing Strategy

The Welch Test

- How should one employ the Welch test?
- Some authors advocate a sequential strategy, in which one first tests for equal variances. If the equal variance test rejects, employ the Welch test, otherwise employ the standard t-test.
- This is, at its foundation, an "Accept-Support" strategy, in which one employs the standard test if the null hypothesis is not rejected.
- The fact that tests on variances have low power compromises this strategy.
- As a result, some authors advocate always doing a Welch test.

Non-normality *and* Heterogeneity

The Yuen-Welch Test

- The Yuen-Welch Test is a combination of the Welch test and the two-sample test on trimmed means.
- Define the squared estimated standard error as

$$d_j = \frac{SSW_j}{h_j(h_j - 1)} \tag{9}$$

- The Yuen-Welch test statistic is

$$t_{Yuen} = \frac{\bar{Y}_{t1} - \bar{Y}_{t2}}{\sqrt{d_1 + d_2}} \tag{10}$$

An adjusted degrees of freedom parameter is used for the test. This is estimated as

$$\nu_y = \frac{(d_1 + d_2)^2}{\frac{d_1^2}{h_1 - 1} + \frac{d_2^2}{h_2 - 1}} \tag{11}$$

Non-normality *and* Heterogeneity

The Yuen-Welch Test

- The Yuen-Welch test is implemented in the WRS library function yuen.
- Here, we simulate two samples with unequal n, unequal variances, and long tails.

```
> contaminated.sample <- function(mu,sigma,n,p,mu2,sigma2)
+ {
+ x<-1:n
+ for(i in 1:n) {if(rbinom(1,1,p)==1)
+   x[i] <- rnorm(1,mu2,sigma2) else
+     x[i] <- rnorm(1,mu,sigma)}
+ return(x)
+ }
> x <- contaminated.sample(0,1,20,.10,0,10)
> y <- contaminated.sample(1,3,10,.10,1,15)
> js.t.test(x,y,.05)

[1] -0.9760167  0.0000000
```

Non-normality *and* Heterogeneity

The Yuen-Welch Test

```
> ## Load augmented Wilcox Library Functions
> source("http://www.statpower.net/R311/Rallfun-v27.txt")
> source("http://www.statpower.net/R311/WRS.addon.txt")
> yuen(x,y,tr=0.20,alpha=.05)

$n1
[1] 20

$n2
[1] 10

$est.1
[1] 0.1862195

$est.2
[1] 1.052019

$ci
[1] -4.719953  2.988355

$p.value
[1] 0.5962929

$dif
[1] -0.8657992

$se
[1] 1.540801

$teststat
[1] 0.5619148

$crit
[1] 2.501396

$df
[1] 5.503793
```