Intermediate Econometrics

V. Gioia

Comparing two population means

We have two populations and we collect data about samples from these populations. Let's denote with $y_{i1}, i = 1, \dots, n_1$, a certain observed characteristic in the first sample and let y_{i2}, $i = 1, \dots, n_2$, the measurements on the same characteristic of the second sample.

Let's consider a quantitative variable. Goal: **compare the population means μ_1 and μ_2**

- Parameter: $\mu_1 - \mu_2$
- Parameter estimate: $\bar{y}_1 - \bar{y}_2$
- Estimator: $\bar{Y}_1 - \bar{Y}_2$

Suppose to assume $Y_{i1} \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $i = 1, \dots, n_1$ and $Y_{i2} \sim \mathcal{N}(\mu_2, \sigma_2^2)$, $i = 1, \dots, n_2$, with Y_{i1} and Y_{i2} independent. We will also assume $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

Two-sided confidence interval for the difference of two means

Let $\overline{Y}_1$ and $\overline{Y}_2$ be the sample mean for the two groups, respectively. Let us identify the pivotal-quantity

- You just know that $\sigma_1^2 = \sigma_2^2 = \sigma^2$ but it is a parameter. You need to estimate it. Let's define the **pooled variance estimator**

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

$$T = \frac{\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Now we know that $T \sim t_{n_1+n_2-2}$ and the $(1 - \alpha)$ confidence interval for $\mu_1 - \mu_2$ is given by

$$IC^{1-\alpha}_{\mu_1-\mu_2} = (\bar{y}_1 - \bar{y}_2 - t_{n_1+n_2-2;1-\alpha/2} \times s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \bar{y}_1 - \bar{y}_2 + t_{n_1+n_2-2;1-\alpha/2} \times s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}})$$

Example

Let us consider this example. A study makes use of a cognitive behavioral therapy to treat a sample of teenage girls who suffered from anorexia. The study, like most such studies, also had a control group that received no treatment. Then researchers analyzed how the mean weight change compared for the treatment and control groups. The girls in the study were randomly assigned to the cognitive behavioral therapy (Group1) or to the control group (Group2).

Let μ_1 and μ_2 denote the mean weight gains (in pounds) for these groups.

```r
Anor <- read.table("http://stat4ds.rwth-aachen.de/data/Anorexia.dat",
                   header=TRUE)
# Get difference post-pre treatment for the group cb and c
cogbehav <- Anor$after[Anor$therapy == "cb"] - Anor$before[Anor$therapy == "cb"]
control <- Anor$after[Anor$therapy == "c"] - Anor$before[Anor$therapy == "c"]


# Get the 95% CI via t.test function
res <- t.test(cogbehav, control, var.equal = TRUE, conf.level = 0.95)
res$conf.int
```

```
## [1] -0.680137  7.593930
## attr(,"conf.level")
## [1] 0.95
```

- The mean weight change for the cognitive behavioral therapy could be as much as 0.68 pounds lower or as much as 7.59 pounds higher than the mean weight change for the control group.
- The interval includes 0: it is plausible that the population means are identical (we can also see the results underlying the hypothesis test)
- The confidence interval is relatively wide: sample sizes are not large.

Obtain the confidence interval by hand

```r
n1 <- length(cogbehav)
n2 <- length(control)

s2 <- ((n1 - 1) * var(cogbehav) + (n2 - 1) * var(control))/(n1 + n2 - 2)

CI <- mean(cogbehav) - mean(control) +
  c(-1,1) * qt(0.975, df = n1 + n2 - 2) *  sqrt(s2 * (1/n1 + 1/n2))
CI
```

```
## [1] -0.680137  7.593930
```

Test for the mean difference

Let's consider again the example on Anorexia, illustrated above. We have two groups and we want ask if the mean weight change between the two groups can be considered as equal.

We could set up a test with the following aim: do the cognitive behavioral group therapy have mean weight change equal to the mean weight change of the control group? Or, do the cognitive behavioral group therapy have mean weight greater than the control group?

Under the same assumptions detailed above we aim to compare their means, μ_1 and μ_2 through the following hypothesis test (consider $\alpha = 0.05$)

two-sided two-sample test

$$\begin{cases} H_0 : \mu_1 - \mu_2 = 0 \\ H_1 : \mu_1 - \mu_2 \neq 0 \end{cases}$$

one-sided two-sample test

$$\begin{cases} H_0 : \mu_1 - \mu_2 = 0 \quad (\text{equivalently} \quad \mu_1 - \mu_2 \leq 0) \\ H_1 : \mu_1 - \mu_2 > 0 \end{cases}$$

Assuming that $\sigma_1^2 = \sigma_2^2 = \sigma^2$, and then the test statistic, under H_0 has the form

$$T = \frac{\overline{X} - \overline{Y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \underset{H_0}{\sim} t_{n_1+n_2-2}$$

The results obtained by using the t.test() function for the two-sided two-sample test are

```r
res.two <- t.test(cogbehav, control, var.equal = TRUE)
res.two
```

```
##
##  Two Sample t-test
##
## data:  cogbehav and control
## t = 1.676, df = 53, p-value = 0.09963
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.680137  7.593930
## sample estimates:
## mean of x mean of y
##  3.006897 -0.450000
```

and for the one-sided two-sample test are

```
res.one <- t.test(cogbehav, control, var.equal = TRUE, alternative = "greater")
res.one
```

```
##
##  Two Sample t-test
##
## data:  cogbehav and control
## t = 1.676, df = 53, p-value = 0.04981
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.003879504         Inf
## sample estimates:
## mean of x mean of y
##  3.006897 -0.450000
```

By hand, we obtain

```
testStat <- (mean(cogbehav) - mean(control))/sqrt(s2 * (1/n1 +1/n2))
## two-sided
2 * pt(testStat, df = n1+n2-2, lower = FALSE )
```

```
## [1] 0.09962901
```

```
## one-sided
pt(testStat, df = n1+n2-2, lower = FALSE )
```

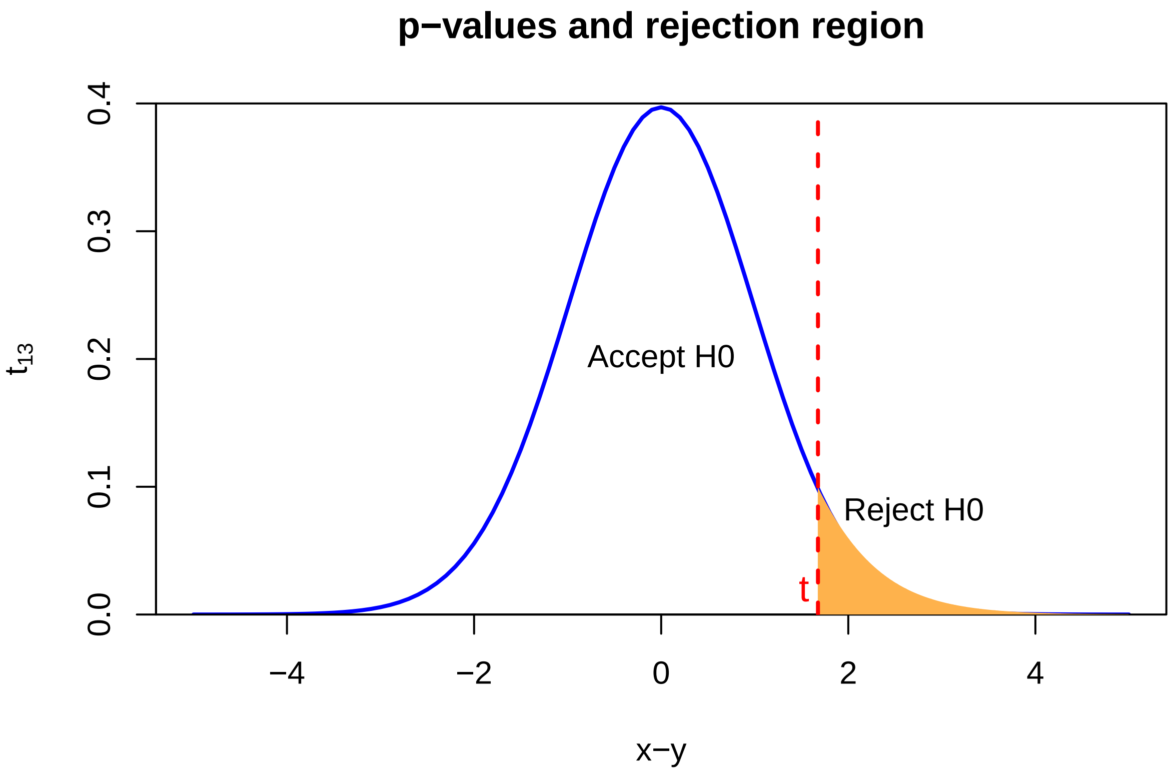
```
## [1] 0.04981451
```

What are the conclusions?

```r
library(RColorBrewer)
plotclr <- brewer.pal(6, "YlOrRd")

curve(dt(x, n1 + n2 - 2), xlim = c(-5, 5), ylim = c(0, 0.4),
      main = "p-values and rejection region", col = "blue",
      lwd = 2, xlab = "x-y",  ylab = expression(t[13]),  yaxs="i")
cord.x <- c(qt(0.95, n1 + n2 - 2),seq(qt(0.95, n1 + n2 - 2), 5, 0.01), 5)
cord.y <- c(0, dt(seq(qt(0.95, n1 + n2 - 2), 5, 0.01), 13), 0)
polygon(cord.x, cord.y, col = plotclr[3], border = NA )


abline(v = res.one$statistic, lty = 2, lwd = 2, col = "red")
text(0, 0.2, paste("Accept", expression(H0)))
text(2.7, 0.08, paste("Reject", expression(H0)))
text(as.double(res.one$statistic) - 0.15, 0.02, "t", col = "red", cex = 1.2)
```



On the plot above note that the value of observed test statistic vary as function of the sample (only for this particular case is too close to alpha that you cannot detect a graphical difference). Different data (having same sample size) will produce a value of the observed test statistic completely different, while considering the same α (and of course considering the same test statistic distribution) the area in orange will remain the same.