Lecture 10

Ciprian Crainiceanu

Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
Johns Hopkins University

September 8, 2020

Table of contents

Outline

1. Introduce independent group t confidence intervals
2. Define the pooled variance estimate
3. Derive the distribution for the independent group, common variance, statistic
4. Cover likelihood methods for the change in the group means per standard deviation
5. Discuss remedies for unequal variances

Independent group t confidence intervals

- Suppose that we want to compare the mean blood pressure between two groups in a randomized trial; those who received the treatment to those who received a placebo
- We cannot use the paired t CI because the groups are independent and may have different sample sizes
- We now present methods for comparing independent groups

Notation

- Let $X_1, \ldots, X_{n_x}$ be iid $N(\mu_x, \sigma^2)$
- Let $Y_1, \ldots, Y_{n_y}$ be iid $N(\mu_y, \sigma^2)$
- Let $\bar{X}$, $\bar{Y}$, S_x, S_y be the means and standard deviations
- Using the fact that linear combinations of normals are again normal, we know that $\bar{Y} - \bar{X}$ is also normal with mean $\mu_y - \mu_x$ and variance $\sigma^2(\frac{1}{n_x} + \frac{1}{n_y})$
- The pooled variance estimator

$$S_p^2 = \{(n_x - 1)S_x^2 + (n_y - 1)S_y^2\}/(n_x + n_y - 2)$$

is a good estimator of σ^2

Note

- The pooled estimator is a mixture of the group variances, with greater weight on whichever has a larger sample size
- If the sample sizes are the same then the pooled variance estimate is the average of the group variances
- The pooled estimator is unbiased

$$\begin{aligned} E[S_p^2] &= \frac{(n_x-1)E[S_x^2]+(n_y-1)E[S_y^2]}{n_x+n_y-2} \\ &= \frac{(n_x-1)\sigma^2+(n_y-1)\sigma^2}{n_x+n_y-2} \end{aligned}$$

- The pooled variance estimate is independent of $\bar{Y}-\bar{X}$ since S_x is independent of $\bar{X}$ and S_y is independent of $\bar{Y}$ and the groups are independent

Result

- The sum of two independent Chi-squared random variables is Chi-squared with degrees of freedom equal to the sum of the degrees of freedom of the summands
- Therefore

$$\begin{aligned}(n_x + n_y - 2)S_p^2/\sigma^2 &= (n_x - 1)S_x^2/\sigma^2 + (n_y - 1)S_y^2/\sigma^2 \\ &= \chi^2_{n_x-1} + \chi^2_{n_y-1} \\ &= \chi^2_{n_x+n_y-2}\end{aligned}$$

Putting this all together

- The statistic

$$\frac{\frac{\bar{Y}-\bar{X}-(\mu_y-\mu_x)}{\sigma\left(\frac{1}{n_x}+\frac{1}{n_y}\right)^{1/2}}}{\sqrt{\frac{(n_x+n_y-2)S_p^2}{(n_x+n_y-2)\sigma^2}}} = \frac{\bar{Y}-\bar{X}-(\mu_y-\mu_x)}{S_p\left(\frac{1}{n_x}+\frac{1}{n_y}\right)^{1/2}}$$

  is a standard normal divided by the square root of an independent Chi-squared divided by its degrees of freedom
- Therefore this statistic follows Gosset's t distribution with $n_x + n_y - 2$ degrees of freedom
- Notice the form is (estimator - true value) / SE

Confidence interval

- Therefore a $(1-\alpha)\times 100\%$ confidence interval for $\mu_y - \mu_x$ is

$$\bar{Y} - \bar{X} \pm t_{n_x+n_y-2,1-\alpha/2} S_p \left(\frac{1}{n_x} + \frac{1}{n_y}\right)^{1/2}$$

- Remember this interval is assuming a constant variance across the two groups
- If there is some doubt, assume a different variance per group, which we will discuss later

Likelihood method

- Exactly as before,

$$\frac{\bar{Y} - \bar{X}}{S_p \left(\frac{1}{n_x} + \frac{1}{n_y} \right)^{1/2}}$$

follows a non-central t distribution with non-centrality parameter $\frac{\mu_y - \mu_x}{\sigma \left(\frac{1}{n_x} + \frac{1}{n_y} \right)^{1/2}}$

- Therefore, we can use this statistic to create a likelihood for $(\mu_y - \mu_x)/\sigma$, a standardized measure of the change in group means

Example

A common use for two-sample t-tests is to create a table comparing two groups of interest on a nuisance variable that is not of interest. This is done to see how alike the two groups are for subsequent comparisons. In AJE vol 64 page 529-537, Zhang et al. would like to compare current smokers to never smokers with respect to sleep characteristics. They first compare the ages of the two groups. The mean (sd) age (in years) for the 10 current smokers was 59.6 (9.5) while it was 63.5 (11.5) for 10 never smokers. (Note that the sample sizes have been fudged a little bit for our purposes.) Test the hypothesis that the two age groups are the same.

Example continued

- Pooled sd

$$\sqrt{\frac{9.5^2(10-1)+11.5^2(10-1)}{20-2}} = 10.55$$

- .975 quantile of a t-distribution with 18 df
  `qt(.975,18)=2.10`
- .95 confidence interval

$$59.6 - 63.5 \pm 2.10 \times 10.55 \sqrt{\frac{1}{10}+\frac{1}{10}}$$

Example, Page 304 Rosner

The systolic blood pressure of a group of 8 oral contraceptive users (mean 132.6 mmHg, sd 15.34 mmHg) is compared to that of 21 controls (mean 127.4 mmHg, sd 18.23 mmHg). What can be said about the difference in mean SBP between the two groups?

Unequal variances

- Note that under unequal variances

$$\bar{Y} - \bar{X} \sim N\left(\mu_y - \mu_x, \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}\right)$$

- The statistic

$$\frac{\bar{Y} - \bar{X} - (\mu_y - \mu_x)}{\left(\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}\right)^{1/2}}$$

approximately follows Gosset's t distribution with degrees of freedom equal to

$$\frac{\left(S_x^2/n_x + S_y^2/n_y\right)^2}{\left(\frac{S_x^2}{n_x}\right)^2/(n_x - 1) + \left(\frac{S_y^2}{n_y}\right)^2/(n_y - 1)}$$

General formula

```
t.test(x, ...)

## Default S3 method:
t.test(x, y = NULL,
       alternative = c("two.sided","less","greater"),
       mu = 0, paired = FALSE, var.equal = FALSE,
       conf.level = 0.95, ...)
```

- One sample t-test (e.g. $H_0 : \mu = 3$)
- Two sample t-test (e.g. $H_0 : \mu_X = \mu_Y$)
- Alternative hypotheses (e.g. $H_A : \mu < 3$ or $H_A : \mu_X > \mu_Y$)

One sample t-test

```
x<-1:10
t.test(x,alternative="two.sided",mu=2)

One Sample t-test

data:  x
t = 3.6556, df = 9, p-value = 0.005271
alternative hypothesis: true mean is not equal to 2
95 percent confidence interval:
 3.334149 7.665851
sample estimates:
mean of x
      5.5
```

Two sample t-test

```
x<-1:10
y<-c(7:20)
t.test(x,y,alternative="less",var.equal = TRUE)

Two Sample t-test

data:  x and y
t = -5.1473, df = 22, p-value = 1.845e-05
alternative hypothesis: true difference in means is
                        less than 0
95 percent confidence interval:
      -Inf -5.331188
sample estimates:
mean of x mean of y
      5.5      13.5
```

Two sample t-test

```
sleep$extra
 [1]  0.7 -1.6 -0.2 -1.2 -0.1  3.4  3.7  ...
[16]  4.4  5.5  1.6  4.6  3.4

sleep$group
 [1] 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2
Levels: 1 2
```

Two sample t-test

```
t.test(extra ~ group, data = sleep)

        Welch Two Sample t-test

data:  extra by group
t = -1.8608, df = 17.776, p-value = 0.07939
alternative hypothesis: true difference in means is
                        not equal to 0
95 percent confidence interval:
 -3.3654832  0.2054832
sample estimates:
mean in group 1 mean in group 2
           0.75            2.33
```

Paired two sample t-test

```
t.test(extra ~ group, paired=TRUE, data = sleep)

Paired t-test

data:  extra by group
t = -4.0621, df = 9, p-value = 0.002833
alternative hypothesis: true difference in means is
                        not equal to 0
95 percent confidence interval:
 -2.4598858 -0.7001142
sample estimates:
mean of the differences
                  -1.58
```

Paired two sample t-test

```
extra=sleep$extra
group=sleep$group
before=extra[group == 1]
after=extra[group == 2]

t.test(before, after, paired=TRUE)
```

Observations

- The t-test can refer to many different types of tests
- It is crucial to know a-priori what kind of test will be applied (especially in clinical trials)
- Paired t-tests are more powerful than un-paired ones; use only when pairing makes sense
- Pairing is done intrinsically via the group variable
- Unpaired two-sample t-tests do not require the same sample size and can be conducted using the assumption of equal variance (pooled estimator) or without (un-pooled)