Two group intervals

Statistical Inference

Brian Caffo, Jeff Leek, Roger Peng

## Independent group 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 test because the groups are independent and may have different sample sizes
* We now present methods for comparing independent groups

## Notation

* Let be iid
* Let be iid
* Let , , , be the means and standard deviations
* Using the fact that linear combinations of normals are again normal, we know that is also normal with mean and variance
* The pooled variance estimator
* is a good estimator of

## Note

* The pooled estimator is a mixture of the group variances, placing greater weight on whichever has a larger sample size
* If the sample sizes are the same the pooled variance estimate is the average of the group variances
* The pooled estimator is unbiased
* The pooled variance estimate is independent of since is independent of and is independent of 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

## Putting this all together

* The statistic
* 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 distribution with degrees of freedom
* Notice the form is (estimator - true value) / SE

## Confidence interval

* Therefore a confidence interval for is
* 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

## Example

### Based on Rosner, Fundamentals of Biostatistics

* Comparing SBP for 8 oral contraceptive users versus 21 controls
* mmHg with mmHg
* mmHg with mmHg
* Pooled variance estimate

sp <- sqrt((7 \* 15.34^2 + 20 \* 18.23^2) / (8 + 21 - 2))  
132.86 - 127.44 + c(-1, 1) \* qt(.975, 27) \* sp \* (1 / 8 + 1 / 21)^.5

## [1] -9.521097 20.361097

data(sleep)  
x1 <- sleep$extra[sleep$group == 1]  
x2 <- sleep$extra[sleep$group == 2]  
n1 <- length(x1)  
n2 <- length(x2)  
sp <- sqrt( ((n1 - 1) \* sd(x1)^2 + (n2-1) \* sd(x2)^2) / (n1 + n2-2))  
md <- mean(x1) - mean(x2)  
semd <- sp \* sqrt(1 / n1 + 1/n2)  
md + c(-1, 1) \* qt(.975, n1 + n2 - 2) \* semd

## [1] -3.363874 0.203874

t.test(x1, x2, paired = FALSE, var.equal = TRUE)$conf

## [1] -3.363874 0.203874  
## attr(,"conf.level")  
## [1] 0.95

t.test(x1, x2, paired = TRUE)$conf

## [1] -2.4598858 -0.7001142  
## attr(,"conf.level")  
## [1] 0.95

## Unequal variances

* Under unequal variances
* The statistic
* approximately follows Gosset's distribution with degrees of freedom equal to
* ---

## Example

* Comparing SBP for 8 oral contraceptive users versus 21 controls
* mmHg with mmHg
* mmHg with mmHg
* ,
* Interval
* In R, t.test(..., var.equal = FALSE)

## Comparing other kinds of data

* For binomial data, there's lots of ways to compare two groups
* Relative risk, risk difference, odds ratio.
* Chi-squared tests, normal approximations, exact tests.
* For count data, there's also Chi-squared tests and exact tests.
* We'll leave the discussions for comparing groups of data for binary and count data until covering glms in the regression class.
* In addition, Mathematical Biostatistics Boot Camp 2 covers many special cases relevant to biostatistics.