1 Lab 2: $t$ tests, confidence intervals, & effect size

1.1 initialize R

Create the folder Rlab2 inside the PSY710 folder located in your home directory. Then launch R and enter the following commands:

```r
setwd("~/PSY710/Rlab2") # set working directory
```

1.2 $t$ tests

In this section we will use $t$ tests to evaluate differences between two means. You will use the `cohen.d` command in the `effsize` package, so you need to install and load the package with the following commands:

```r
install.packages("effsize") # if effsize is not on your computer
library(effsize) # load package into R memory
# help(package="effsize") # to see list of commands
```

Use the following R code to create and store two samples of data in the variables d1 and d2:

```r
mu1 <- 0; sd1 <- 1 # population 1
mu2 <- 0.9; sd2 <- 1 # population 2
n1 <- 20; n2 <- 20 # sample sizes
set.seed(30912) # set random number generator
d1 <- rnorm(n1,mu1,sd1)
d2 <- rnorm(n2,mu2,sd2)
```

Next, we use `t.test` to evaluate the null hypothesis that the two samples were selected from populations with the same mean:

```r
t.test(d2,d1,paired=F,alternative="two.sided")
```

Use the results of your $t$ test to answer the following questions:

- What are the null and alternative hypotheses being evaluated by this test?
  
  **Answer:** The null hypothesis is that the two population means are equal (i.e., the difference between means is zero) and the alternative hypothesis is that the two population means are not equal (i.e., the difference between means is not equal to zero).

- What does the $p$ value mean?
  
  **Answer:** The $p$ value ($p = 0.059$) means that the probability of obtaining a difference that is at least as extreme as the observed difference *when the null hypothesis is true* is 0.059.
• What is your decision regarding the null hypothesis?

**Answer:** If I want my Type I error rate to be 0.05, then I will reject the null hypothesis when $p \leq 0.05$. In this case, $p = 0.059$ so I do not reject the null hypothesis. I would report this result in the following way: “The difference between means was not significant ($t(34.77) = 1.945, p = 0.059$).”

• What is the 95% confidence interval? Explain what this interval means.

**Answer:** A *point estimate* of the difference between population means equals the difference between sample means: $\hat{\delta}\mu = 0.8747 - 0.2382 = 0.636$. A confidence interval is an *interval estimate* of the difference between population means: it says “we think the difference between population means lies in this range/interval”. In this case, the 95% confidence interval is [-0.028, 0.238]. The reason we call it the 95% interval is because the statistical formula used to estimate the interval (from our data) are designed in such a way that the interval – which will vary across replications of this experiment – will (in the long run) contain the true value of $\delta\mu$ 95% of the time.

The previous $t$ test did not assume that the variance in the two populations was the same. We can make that assumption by setting the `var.equal` parameter to `TRUE`:

```
set.seed(8719)
d3 <- rnorm(n2,1.5*mu2,4*sd2)
t.test(d3,d1,paired=F,alternative="two.sided",var.equal=F)
```

• How do the results of this $t$ differ from the one that did not make the equal variance assumption?

**Answer:** When we assume the variances are equal, then the degrees of freedom equals $n_1 + n_2 - 2$, which in this case is 38. This change in the degrees of freedom alters the $p$ value from 0.0599 to 0.592 and produces subtle changes in the confidence interval (because the calculation of the CI uses the degrees of freedom). In this case, the effects on the $p$ value and CI are small because the two samples have very similar variances. When the variances differ, the changes will be larger. For example, compare the following two $t$ tests...

```
set.seed(8719)
d3 <- rnorm(n2,1.5*mu2,4*sd2)
t.test(d3,d1,paired=F,alternative="two.sided",var.equal=F)
```

```
##
## Welch Two Sample t-test
##
## data: d3 and d1
## t = 2.466, df = 21.44, p-value = 0.0222
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## 0.305579 3.566099
## sample estimates:
## mean of x mean of y
## 2.174046 0.238207
```

```
set.seed(8719)
d3 <- rnorm(n2,1.5*mu2,4*sd2)
t.test(d3,d1,paired=F,alternative="two.sided",var.equal=T)
```

```
##
## Two Sample t-test
##
## data: d3 and d1
## t = 2.466, df = 38, p-value = 0.0183
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## 0.346905 3.524772
## sample estimates:
## mean of x mean of y
## 2.174046 0.238207
```
• In general, do you think it is better to assume equal variance or to not assume equal variance?

**Answer:** It is probably more reasonable to not assume that the population variances are equal.

Next, calculate Cohen’s \(d\) and the Common Language Effect Size (CLES) for our data:

```r
# install.packages("effsize") # install package onto computer
# library(effsize) # load into R's memory
# install.packages("lsr") # install package onto computer
# library(lsr) # contains another cohen's d command
cohen.d(d=d2,f=d1,pooled=T,paired=F) # cohen's d

##
## Cohen's d
##
## d estimate: 0.614993 (medium)
## 95 percent confidence interval:
##    lower    upper
## -0.0401344 1.2701208

cohensD(x=d2,y=d1,method="unequal") # in lsr; assumes unequal variance

## Error in cohensD(x = d2, y = d1, method = "unequal"): could not find function "cohensD"

VD.A(d=d2,f=d1) # version of CLES

##
## Vargha and Delaney A
##
## A estimate: 0.69 (medium)
```

• What do Cohen’s \(d\) and CLES represent?

**Answer:** Cohen’s \(d\) represents an estimate of the difference between population means in terms of a pooled estimate of the population standard deviation. The value can differ depending on how the population standard deviation is estimated from the two samples. The \(VD.A\) command calculates an estimate of the Common Language Effect Size. It calculates the percentage of times a randomly selected value from \(d_2\) will be greater than a randomly selected value from \(d_1\).

Next, re-do our original \(t\) test but increase sample size from 20 (per sample) to 100:

```r
# re-do t test with much larger sample sizes:
n1 <- 100
n2 <- 100
set.seed(30912) # initialize random number generator
d1 <- rnorm(n1,mu1,sd1) # sample 1
d2 <- rnorm(n2,mu2,sd2) # sample 2
t.test(d2,d1,paired=F,alternative="two.sided")

##
## Welch Two Sample t-test
##
## data:  d2 and d1
## t = 4.497, df = 195.3, p-value = 1.18e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.381965  0.978792
## sample estimates:
## mean of x mean of y
##  0.7092620  0.028834
```
**cohen.d(d2,d1, pooled=T, paired=F)**  # in effsize; assumes equal variance

```r
##
## Cohen's d
##
## d estimate: 0.635909 (medium)
## 95 percent confidence interval:
##  lower  upper
## 0.350062 0.921756
```

**cohensD(x=d2, y=d1, method="unequal")**  # in lsr; assumes unequal variance

```r
## Error in cohensD(x = d2, y = d1, method = "unequal"): could not find function "cohensD"
```

- What are the results of the *t* test? Explain why altering sample size causes these differences.

**Answer:** Unlike before, the difference between means is significant (*t*(195.3) = 4.4966, *p* < 0.001). The reason the difference is statistically significant is that increasing the sample size reduced the standard error of the mean *dramatically*. Having large samples reduces the expected variation between random samples. Consequently, our estimate of the difference between population means becomes much more precise (look at the 95% confidence interval) and we can declare the observed difference between sample means (∆µ = 0.709 − 0.0288 = 0.68) as being unusual given the assumption that the true difference between means is zero.

- What is the value of Cohen’s *d*? How does the effect of sample size on Cohen’s *d* compare to the effect on the *p* value? Explain.

**Answer:** The change in sample size produce a large change in the *p* value of our *t* test. However, the values of Cohen’s *d* (0.68 and 0.64) are similar to the ones obtained with small sample sizes. The reason for this lack of change is that Cohen’s *d* is an estimate of the difference between *population* means (in standardized units), and this difference does not depend on sample size.

Next, we return to sample sizes of 20, but do a one-tailed test:

```r
n1 <- 20
n2 <- 20
set.seed(30912)
d1 <- rnorm(n1,mu1,sd1)
d2 <- rnorm(n2,mu2,sd2)
t.test(d2,d1,paired=F,alternative="greater")
```
### Example

```r
cohensD(d2,d1,method="unequal")
```

```
## Error in cohensD(d, d1, method = "unequal"): could not find function "cohensD"
```

- Compare the \( t \), \( p \), and \( d \) values to the values obtained with the two-tailed test.

**Answer:** The values of \( t \) and the degrees of freedom are the same, but the \( p \) value is one-half of the previous value (\( p = 0.05992/2 = 0.02996 \approx 0.03 \)). Note that if we had set `alternative` to `less`, the \( p \) value would have been \( p = 1 - (0.0599/2) = 0.97 \). The values of Cohen’s \( d \) remain the same as before.

- What are the null and alternative hypotheses?

**Answer:** The null hypothesis is that \( \mu_2 \leq \mu_1 \) and the alternative hypothesis is that \( \mu_2 > \mu_1 \).

- What is the 95% confidence interval of the mean difference? How is it related to the 95% CI that we obtained with the two-tailed test?

**Answer:** The 95% confidence interval is \([0.083, \infty]\). This is a one-sided interval: in the long run, we expect this interval to contain the true difference 95% of the time, and for the lower bound of the interval to be too high (i.e., for the true difference to fall outside the interval) 5% of the time.

Finally, we do several paired-sample \( t \) tests. Note the change in the way we generate the \( d_2 \) data sample, and in the way we call the `t.test` and `cohen.d` commands:

```r
n1 <- 20
c2 <- 20
set.seed(912)
d1 <- rnorm(n1,mu1,sd1)
d2 <- d1+rnorm(n2,mu2,sd2) # notice difference here
t.test(d2,d1,paired=T,alternative="two.sided")
```

```
## Paired t-test
##
data: d2 and d1
## t = 4.022, df = 19, p-value = 0.000728
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## 0.506624 1.605974
## sample estimates:
## mean of the differences
## 1.0563
```

**Answer:** The paired-sample \( t \) test is equivalent to doing a one-sample \( t \) on the difference scores:
diff.scores <- d2-d1

t.test(diff.scores,alternative="two.sided")

# cohen.d(diff.scores) # produces error

The above commands introduced a correlation between the two samples to mimic the type of correlation that is found in within-subject design. Visualize the correlation by plotting the two samples in a scatter plot, adding the regression line to the scatter plot, and evaluating the correlation with `cor.test`.

```r
plot(x=d1,y=d2,type="p",xlab="d1",ylab="d2") # scatter plot
lm.01 <- lm(d2~d1) # linear regression
abline(lm.01,lty=2) # add regression line to plot
```

## 2 effect size & power

In this section we will review how to calculate power. To do these exercises, you will need to load the `effsize` and `pwr` libraries:

```r
install.packages("pwr") # if pwr is not on your computer
library(pwr) # load package into R memory
# help(package="pwr") # to see list of commands
```

Suppose we believe that the phenomenon that we are study will produce a difference between group means that will correspond to an effect size \( d = 1 \). How big must our samples be in order to have a power of 0.8 if we use a between-subjects design? If we use a within-subjects design?

```r
# between-subjects:
pwr.t.test(n=NULL,d=1,sig.level=.05,power=0.8,type="two.sample",alternative="two.sided")
# within-subjects:
pwr.t.test(n=NULL,d=1,sig.level=.05,power=0.8,type="paired",alternative="two.sided")
```

**Answer:** Note the change in the required sample size!

Use `pwr.t.test` to answer the following questions.

- Suppose we can test 20 subjects in a within-subjects design. What is the smallest effect size that we can investigate if we want to have a power of 0.8?

```r
pwr.t.test(n=20,d=NULL,sig.level=.05,power=0.8,type="paired",alternative="two.sided")
```
Repeat the previous question, but assume that we will test subjects in a between-subjects design.

```r
pwr.t.test(n = 20, d = NULL, sig.level = .05, power = 0.8, type = "two.sample", alternative = "two.sided")
```

```
## Two-sample t test power calculation
##
##    n = 20
##    d = 0.909159
##  sig.level = 0.05
##    power = 0.8
## alternative = two.sided
##
## NOTE: n is number in *each* group
```