Outline

Exercise

1. Assume two independent samples:

```
set.seed(123)
a=rnorm(1368, 0.065, 1 )
b=rnorm(1315, -0.06, 0.99)
```

Assume, population variances are unknown (and unequal). Test the null hypothesis, $H_0 : \mu_a - \mu_b = 0$ against the alternative, $H_1 : \mu_a \neq \mu_b$

R can directly solve this problem:

```
t.test(a,b)
```

Make sure, your results can match the ones directly calculated by R.

2. Assume, in the above example, population variances are unknown but assume to be **equal**. Test the null hypothesis, $H_0 : \mu_a - \mu_b = 0$ against the alternative, $H_1 : \mu_a \neq \mu_b$

R can directly solve this problem:

```
t.test(a,b, var.equal = T)
```

Make sure, your results can match the ones directely calculated by R

Exercise

3. 3 Use Chile data as on slide no. 21. There is a categorical variable education in which PS refers to post secondary education (or higher education). Assume, we are interested in investigating the association between education and voting YES.

   Create a category of highly educated people using variable PS (make only 2 categories, poportion with high education (PS) and a proportion without high education)

   Create a category of voters who voted YES (make only 2 categories, poportion who voted YES and a proportion who did not vote YES)

   - Assume the proportion of people who have higher education and voted yes is P_x
   - Assume the proportion of people who have no higher education and voted YES is P_y
   - Test the null hypothesis, $H_0 : P_y - P_x = 0$ against two sided alternative
   - Also calculate the confidence interval of the difference in two proportions (P_y - P_x).

Outline

Exercise

Solution no. 1:

- We want to test:
  - Null hypothesis: $H_0 : \mu_a - \mu_b = 0$
  - Two sided alternative: $H_1 : \mu_a - \mu_b \neq 0$
- Now, calculate the mean and the standard deviations of a and b

```
a_bar=mean(a)
b_bar=mean(b)
s_a= sd(a)
s_b=sd(b)
```

Exercise

- Now calculate t statistics:

```
n_a=1368;       n_b=1315
se=sqrt((s_a^2/n_a) + (s_b^2/n_b))
se

## [1] 0.0379701

t_stats= (a_bar-b_bar)/se
t_stats

## [1] 3.49125
```

- Now we calculate the corresponding p values:

```
2*(1 - pnorm(t_stats))

## [1] 0.000480767
```

In question, no 1, we are told that population variances are unknown (and unequal) which requires calculating degrees of freedom. However, I used pnorm (which uses a z score) because $n > 30$ in this example. If $n < 30$, then we need to calculate degrees of freedom using formula discussed in the lecture.

Exercise

Solution no. 2:

- In question, no 2, we are told that population variances are unknown but equal which requires calculating standard errors using formula: $se_d = \sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}}$

  And s_p^2, for in the formula is calculated as follows:

$$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}$$

```
sep = ((n_a - 1)*s_a^2 + (n_b - 1)*s_b^2)/(n_a + n_b -2)
se=sqrt((sep/n_a) + (sep/n_b));         se

## [1] 0.0379866

t_stats= (a_bar-b_bar)/se;              t_stats

## [1] 3.48973
```

- Now we calculate the corresponding p values:

```
2*(1 - pnorm(t_stats))

## [1] 0.000483507
```

Exercise

Solution no. 3:

- We first tell R that people with PS education should be assigned TRUE, otherwise FALSE.

```
library(car);                                    data("Chile")
Chile$PS_educ <- factor(Chile$education=="PS")
```

We also tell R that people who voted YES should be assigned TRUE, otherwise FALSE.

```
Chile$voteYES <- factor(Chile$vote=="Y")
```

Before defining the table it is important to understand the question. Our interest is in finding the association of education and voting YES.

```
tab <- xtabs(~PS_educ + voteYES, data=Chile)
tab <- tab[ , c("TRUE", "FALSE")];                tab

##         voteYES
## PS_educ TRUE FALSE
##   FALSE  733  1351
##   TRUE   130   308
```

- Proportion who have no higher education and voted YES: $\hat{p}_y = \frac{733}{1351+733} = 0.352$
- Proportion who have higher education and voted YES: $\hat{p}_x = \frac{130}{130+308} = 0.297$
- Difference of the two propotions: $\hat{p}_y - \hat{p}_x = 0.297 - 0.352 = -0.054$
- Formulae for calculation of CI, se_d, se_0, and z score can be seen in the main lecture

Exercise

- I use R to directly calculate

```
prop.test(tab)
```

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  tab
## X-squared = 4.61, df = 1, p-value = 0.0318
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.00609977 0.10374782
## sample estimates:
##   prop 1   prop 2
## 0.351727 0.296804
```

P value is less than 5%, we can clearly reject the null hypothesis. This means there is statistical difference in the two proportions. This implies that people who have no higher education have a higher tendency to vote for a YES as compared to people who have higher education.