# Hypothesis tests: Comparing two populations
# Lecture: 10

Hamid Raza
Assistant Professor in Economics
raza@business.aau.dk

Aalborg University

Statistics - Statistik

# Outline

# Tests of the Difference Between Two Normal Population Means: Dependent Samples case

We develop procedures for testing the differences between two population means, proportions, and variances

## Comparing two population means: Dependent Samples case

- An instructor is interested in knowing whether giving individual assignments increases or decreases students test scores in the course.
- To answer this question the instructor gives individual assignments to half of the class
- After exams, we test for evidence whether giving individual assignments improved test scores or not
- Let's say mean score of group with individual assignments is $\mu_x$ and the other group is $\mu_y$

# Comparing two population means: Dependent Samples case

## Comparing two population means: Dependent Samples case

- **Hypothesis setup:**
  - Null hypothesis: $H_0 : \mu_x - \mu_y = 0$
  - One sided alternative $H_1 : \mu_x - \mu_y > 0$ or $(\mu_x - \mu_y < 0)$
    or
  - Two sided alternative: $H_1 : \mu_x - \mu_y \neq 0$
- **Procedure:**
  - We simply calculate the difference of the two sample means:

    $$d = \bar{x} - \bar{y} \quad \text{(we also calculate standard deviation } s_d\text{)}$$

  - Test statistic is calculated as: $t = \frac{d}{se_d}$ where $se_d = \frac{s_d}{\sqrt{n}}$
  - **Confidence Interval:** CI can also be calculated using formula:

    $d \pm t_{v,\alpha/2} se_d$ OR $d \pm ME$ (see lecture no. 8)

# Comparing two population means: Dependent Samples (cont)

## Comparing two population means: Dependent Samples (cont)

- **Decision rule for one sided alternatives**
  - Reject $H_0$: if $t < -t_{n-1,\alpha}$ (for left (or lower) tailed test)
    or
  - Reject $H_0$: if $t > t_{n-1,\alpha}$ (for right (or upper tailed test))
- **Decision rule for two sided alternatives**
  - Reject $H_0$: if $t < -t_{n-1,\alpha/2}$
    or
  - Reject $H_0$: if $t > t_{n-1,\alpha/2}$

Note: Like before you can calculate the corresponding p values for a t dist.

# Example in R

## Example in R:

- Assume we have two random normally distributed variables

```
set.seed(123)
x=rnorm(25, 25, 5)
y=rnorm(25, 24, 6)
```

- Assume we want to test:
  - Null hypothesis: $H_0 : \mu_x - \mu_y = 0$
  - Two sided alternative: $H_1 : \mu_x - \mu_y \neq 0$
- First, we calculate the difference in x and y:

```
z=x - y
```

- Now, calculate the mean and the standard deviation of the difference (z)

```
d= mean(z)
s_d=sd(z)
```

- Note: for two statistically dependent samples, from the property of variance, we know that:

$$SD(x - y) \neq SD(x) - SD(y)$$

- Avoid the common mistake of calculating $s_x$ and $s_y$ separately and then subtracting one standard deviation from the other

# Example in R

## Example in R:

- Now calculate t statistics:

```
n=25
se=s_d/sqrt(25)
t_stats= d/se
```

- Now we calculate the corresponding p values (with $n-1$ degrees of freedom):

```
2*(1 - pt(t_stats, df = n-1))

## [1] 0.86684
```

- Conclusion: *Null hypothesis cannot be rejected*
- The above result can be directly achieved in R by typing the following function:

```
t.test(x,y, paired = T)
```

# Dependent sample: Paired t-test

## Dependent sample: Paired t-test

- You choose at random 10 Netto stores, where you measure the average speed time by the cash registers over some period of time.
- Now, new cash registers are installed in all 10 stores, and you repeat the experiment.

It is interesting to investigate whether or not the new cash registers have changed the expedition time. So we have 2 samples corresponding to old and new technology. In this case we have dependent samples, since we have 2 measurement in each store.

- We use the following strategy for analysis:
- For each store calculate the change in average expedition time when we change from old to new technology.
- The changes $d_1, d_2, ..., d_{10}$ are simply treated as **ONE** sample from a population with mean $\mu$.

# Dependent sample: Paired t-test

## Dependent sample: Paired t-test (Example in R)

- We can directly do a paired t test in R

```
Netto <- read.csv("~/Dropbox/Teaching/Statistics/Lecture 10/Netto.csv", header = T)
head(Netto, n = 3) # this shows the first 3 observations

##   before  after
## 1 3.7306 3.4402
## 2 2.6233 2.3147
## 3 3.7953 3.5863
```

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

##
##  Paired t-test
##
## data:  Netto$before and Netto$after
## t = 5.72, df = 9, p-value = 0.00029
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.11227 0.25916
## sample estimates:
## mean of the differences
##                 0.18572
```

# Outline

# Comparing two population means: Independent Samples case

## Mean of Independent Samples: Population Variances known

- **Hypothesis setup:**
  - Null hypothesis: $H_0 : \mu_x - \mu_y = 0$
  - One sided alternative $H_1 : \mu_x - \mu_y > 0$ or $(\mu_x - \mu_y < 0)$
    or
  - Two sided alternative: $H_1 : \mu_x - \mu_y \neq 0$
- **Procedure:**
  - z score is calculated as: $z = \frac{\bar{x} - \bar{y}}{se_d}$

    where $se_d = \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}$

# Comparing two population means: Independent Samples case

## Mean of Independent Samples: Population Variances unknown

When the population variance is unknown, we need to use the Student's t distribution. There are some theoretical problems when we use the Student's t distribution for differences between sample means.

- Here we will address two possible cases:
  - Population Variances for x and y are unknown (but assumed to be unequal)
  - Population Variances for x and y are unknown (but assumed to be equal)
- **Important note:** When we assume the variances of x and y are unknown but equal, this does not mean that the sample variances ($s_x^2$ and $s_y^2$) would be equal. This assumption simply mean that x and y has a common variance.
- Remember sample variances would always be different if we draw a radnom sample from a population

# Comparing means of Independent Samples: Population Variances unknown but **unequal**

- **Hypothesis setup:**
  - Null hypothesis: $H_0 : \mu_x - \mu_y = 0$
  - One sided alternative $H_1 : \mu_x - \mu_y > 0$ or $(\mu_x - \mu_y < 0)$
    or
  - Two sided alternative: $H_1 : \mu_x - \mu_y \neq 0$
- **Procedure:**
  - t statistics is calculated as: $t = \frac{\bar{x}-\bar{y}}{se_d}$

    where $se_d = \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$
  - In this case: the main difference lies in the calculation of the degrees of freedom, for which the formula is as follows:

$$v = \frac{[(\frac{s_x^2}{n_x}) + (\frac{s_y^2}{n_y})]^2}{(\frac{s_x^2}{n_x})^2/(n_x - 1) + (\frac{s_y^2}{n_y})^2/(n_y - 1)}$$

Note: The decision rules for hypothesis testing is the same

Note: If $n_x$ and $n_y$ are greater than 30 then we can directly use z score

## Mean of Independent Samples: Population Variances unknown but **equal**

- **Hypothesis setup:**
  - Null hypothesis: $H_0 : \mu_x - \mu_y = 0$
  - One sided alternative $H_1 : \mu_x - \mu_y > 0$ or $(\mu_x - \mu_y < 0)$
    or
  - Two sided alternative: $H_1 : \mu_x - \mu_y \neq 0$
- **Procedure:**
  - t statistics is calculated as: $t = \frac{\bar{x} - \bar{y}}{se_d}$

    where $se_d = \sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}}$
  - In this case: the main difference lies in the calculation of the $s_p^2$, for which the formula is as follows:

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

Note: $s_p^2$ is the weighted average of the two sample variances

The decision rules for hypothesis testing is the same

# Outline

# Tests of the difference between two population proportions

## Tests of the difference between two population proportions

- We consider the situation, where we have two qualitative samples and we investigate whether a given property is present or not:
- The proportion of population 1 has the property $P_x$, which is estimated by $\hat{p}_x$ based on a sample of size $n_x$
- The proportion of population 2 has the property $P_y$, which is estimated by $\hat{p}_y$ based on a sample of size $n_y$
- We are interested in the difference $p_y - p_x$, which is estimated by $d = \hat{p}_y - \hat{p}_x$

# Tests of the difference between two population proportions

## Tests of the difference between two population proportions

- **Hypothesis setup:**
  - Null hypothesis: $H_0 : P_x - P_y = 0$
  - One sided alternative $H_1 : P_x - P_y > 0$ or $(P_x - P_y < 0)$
    or
  - Two sided alternative: $H_1 : P_x - P_y \neq 0$
- **Procedure:**
  - z score is calculated as: $z = \frac{\hat{p}_x - \hat{p}_y}{se_0}$
  - The important question is how do we calculate $se_0$ in this case?

# Tests of the difference between two population proportions (cont)

- In the situation where we have independent population proportions we have:

$$se_d = \sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}}$$

- **Confidence Interval:** CI can also be calculated using formula:

  $(\hat{p}_y - \hat{p}_x) \pm z_{\alpha/2} se_0$ OR $(\hat{p}_y - \hat{p}_x) \pm ME$ (see lecture no. 8)

- When we assume that the population proportions are equal, then we estimate the common proportion as follows:

$$\hat{p}_0 = \frac{n_x\hat{p}_x + n_y\hat{p}_y}{n_x + n_y}$$

- Therefore we can write: $se_0 = \sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n_x} + \frac{\hat{p}_0(1-\hat{p}_0)}{n_y}}$
- So Z score can be calculated as: $z = \frac{\hat{p}_x - \hat{p}_y}{\sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n_x} + \frac{\hat{p}_0(1-\hat{p}_0)}{n_y}}}$
- **Caution:** The approximation is only good, when $n_x\hat{p}_0$, $n_x(1-\hat{p}_0)$, $n_y\hat{p}_0$, $n_y(1-\hat{p}_0)$ all are greater than 5

## Tests of the difference between two population proportions (cont)

- **Decision rule for one sided alternatives**
  - Reject $H_0$: if $z < -z_\alpha$ (for left (or lower) tailed test)
    or
  - Reject $H_0$: if $z > z_\alpha$ (for right (or upper tailed test))
- **Decision rule for two sided alternatives**
  - Reject $H_0$: if $z < -z_{\alpha/2}$
    or
  - Reject $H_0$: if $z > z_{\alpha/2}$

Note: Like before you can calculate the corresponding p values for a normal dist.

# Example in R

Assume we want to find out whether a person intends to vote NO or something else. We take two proportions male and female

```
library(car);                                    data("Chile")
```

The code below tells R, if a person is a "NO VOTER" (then TRUE), all other types of voters (then FALSE)

```
Chile$voteNo <- factor(Chile$vote=="N")
tab <- xtabs(~sex+voteNo, data=Chile)
tab <- tab[ , c("TRUE", "FALSE")] # This command swaps the coloumns and treats VoteNo (TRUE) as success.
# I do it for consistency as R runs prop.test treating first column as success (see prop.test code below).
# But the results do not matter regardless of which column is treated as success event.

tab
##    voteNo
## sex TRUE FALSE
##   F  363   946
##   M  526   697
```

- The proportion of men who vote NO: $\hat{p}_x = \frac{526}{526+697} = 0.43008$
- The proportion of women who vote NO: $\hat{p}_y = \frac{363}{363+946} = 0.2773$
- Difference of the two propotions: $\hat{p}_y - \hat{p}_x = 0.2773 - 0.43008 = -0.15277$
- Standard error of the difference:

$$se_d = \sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}} = \sqrt{\frac{0.430(1-0.430)}{1223} + \frac{0.277(1-0.277)}{1309}} = 0.0188$$

- Confidence interval at 95%: $\hat{p}_y - \hat{p}_x \pm 1.96 se_d = (-0.189; -0.1151)$

# Example in R (cont)

- We now test the null hypothesis: $H_0 : P_x - P_y = 0$
- Two sided alternative: $H_1 : P_x - P_y \neq 0$
- First we calculate: $\hat{p}_0 = \frac{n_x \hat{p}_x + n_y \hat{p}_y}{n_x + n_y}$ and then:

$$se_0 = \sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n_x} + \frac{\hat{p}_0(1-\hat{p}_0)}{n_y}} = 0.0190$$

- Now we calculate $z = \frac{\hat{p}_x - \hat{p}_y}{se_0} = -8.06$
- R can directly calculate this:

```
prop.test(tab, correct = FALSE)


##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  tab
## X-squared = 64.8, df = 1, p-value = 8.4e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.18963 -0.11593
## sample estimates:
##  prop 1  prop 2
## 0.27731 0.43009
```

# Outline

# Comparing variance of two populations

## Comparing variance of two populations

- We now develop a procedure for testing the assumption that population variances from independent samples are equal.
- To perform such tests, we introduce the F probability distribution, represented as follows:

$$F = \frac{s_x^2/\sigma_x^2}{s_y^2/\sigma_y^2}$$

- The F distribution is constructed as the ratio of two chi-square random variables, each divided by its degrees of freedom
- We need to emphasize that this test is quite sensitive to the assumption of normality.

# Comparing variance of two populations

## Comparing variance of two populations

- **Hypothesis setup:**
  - Null hypothesis: $H_0 : \sigma_x^2 = \sigma_y^2$
  - One sided alternative $H_1 : \sigma_x^2 > \sigma_y^2$ or $(\sigma_x^2 < \sigma_y^2)$ or
  - Two sided alternative: $H_1 : \sigma_x^2 \neq \sigma_y^2$
- **Procedure:**
  - F stats is calculated as: $F = \frac{s_x^2}{s_y^2}$

## Comparing variance of two populations

- **Decision rule for one sided alternatives**
  - Reject $H_0$: if $F < F_{n_x-1,n_y-1,1-\alpha}$ (for left (or lower) tailed test)
    or
  - Reject $H_0$: if $F > F_{n_x-1,n_y-1,\alpha}$ (for right (or upper tailed test))
- **Decision rule for two sided alternatives**
  - Reject $H_0$: if $F < F_{n_x-1,n_y-1,1-\alpha/2}$
    or
  - Reject $H_0$: if $F > F_{n_x-1,n_y-1,\alpha/2}$

## Example in R:

- Assume two variables $x$ and $y$, with variance of x ($s_x^2$) greater than variance of y ($s_y^2$)
- We test the null hypothesis, $H_0 = \sigma_x^2 = \sigma_y^2$ against two sided alternative, $H_1 = \sigma_x^2 \neq \sigma_y^2$

```
set.seed(123);        x <- rnorm(50, mean = 1, sd = 2);        y <- rnorm(30, mean = 1, sd = 1.5)
var_x=var(x);          var_y=var(y)
```

Now we calculate the F statistics. But when calculating f stats, make sure the larger vaiance is the numerator, and lower variance is the denominator

```
f_stats= var(x)/var(y)
```

Define the degrees of freedom for each sample

```
dof_x= length(x) -1;            dof_y=length(y) - 1
```

Check the p values using f stats and the degrees of freedom for each sample

```
2*(1-pf(f_stats, df1=dof_x, df2= dof_y))

## [1] 0.11223
```

We can directly do this in R

```
var.test(x, y) # the variable with greater variance (x in this case) should come first.
```

# 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, proportion with high education (PS) and a proportion without high education)

   Create a category of voters who voted YES (make only 2 categories, proportion 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 $H_1 : P_y - P_x \neq 0$
   - Also calculate the confidence interval of the difference in two proportions ($P_y$ - $P_x$).