

# ECON3150/4150 Spring 2015

## Lecture 2 - Review of statistics

Siv-Elisabeth Skjelbred

University of Oslo

March 9, 2015

# Overview

In this lecture we will:

- Learn about estimators and their properties
- Go through the statistical methods for estimation, hypothesis testing and confidence intervals

# Statistics

A random sample can be used to:

- Estimate the population mean
- Test hypotheses about the population mean
- Compute a confidence interval for the population mean

# Estimates

- An estimate is the numerical value computed using an estimator on a specific sample.
- An estimator is a mathematical rule used to calculate an estimate.

# Estimating the population mean

Potential estimators of the population mean:

- The sample average ( $\bar{Y}$ ).
- The value of the first observation ( $Y_1$ ).
- The sample median.
- The sample mode.
- ...

## Choosing estimator

The best estimator is the one that is as close as possible to the unknown true value. More specifically a 'good' estimator is:

- Unbiased:  $E(\hat{\mu}_y) - \mu_y = 0$  which means that  $E(\hat{\mu}_y) = \mu_y$ .
- Consistent:  $\hat{\mu}_y \xrightarrow{P} \mu_y$ .
- Efficient:  $var(\hat{\mu}_y) < var(\tilde{\mu}_y)$  where  $\tilde{\cdot}$  denotes another estimator.

### Random sampling

Non-random sampling may result in unbiased, inconsistent, estimators even if these estimator are unbiased under random sampling.

## The sample average as estimator

The sample average is a unbiased estimator of the population parameter:

$$E(\bar{Y}) = E\left(\frac{1}{n} \sum_{i=1}^n Y_i\right) = \mu_y$$

The expected value of the sample average is the true population parameter. Thus the sample average is unbiased.

### Rules of summation:

- $\sum aX = a \sum X$
- $E(X + Y) = E(X) + E(Y)$
- $\sum a = na$

# The sample average as estimator

Consistency:

- The sample mean is consistent if the probability that  $\bar{Y}$  is in the range  $(\mu_y - c)$  to  $(\mu_y + c)$  becomes arbitrarily close to 1 as  $n$  increases for any constant  $c > 0$ .
- Mathematically:  $\bar{Y} \xrightarrow{P} \mu_y$

Law of large numbers:

If the sample consist of independently and identically distributed observations with  $E(Y_i) = \mu_Y$  and large outliers are unlikely then the sample average is consistent.



# The sample average as estimator

Efficiency:

$$\begin{aligned} \text{var}(\bar{Y}) &= \text{var}\left(\frac{1}{n} \sum_{i=1}^n Y_i\right) \\ &= \frac{1}{n^2} \sum_{i=1}^n \text{var}(Y_i) + \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1, j \neq i}^n \text{cov}(Y_i, Y_j) \\ &= \frac{\sigma_y^2}{n} \end{aligned} \quad (1)$$

Rules of variance:

- $\text{Var}(aX + bY) = a^2 \text{var}(X) + 2abcov(X, Y) + b^2 \text{var}(Y)$
- $\text{cov}(Y_i, Y_j) = 0$  when i.i.d. draws.

# The sample average as estimator

The sample average vs  $Y_1$

- Both are unbiased:  $E(Y_1) = \mu_y = E(\bar{Y})$
- Both are consistent under the same assumptions
- $Var(Y_1) = \sigma_y^2 > Var(\bar{Y}) = \frac{\sigma_y^2}{n}$

The sample average is a better estimator than using the first observation.

# The sample average as estimator

## BLUE

The sample average is the best linear unbiased estimator of the population mean. It is the most efficient among all unbiased estimators that are linear functions of  $Y_1, \dots, Y_n$ .

# Hypotheses

- Given the sample average we can perform hypothesis tests about the true population mean
- The null hypothesis is the hypothesis being tested typically denoted  $H_0$
- We use the data to compare the null hypothesis to an alternative hypothesis which holds if the null hypothesis does not hold.
- The null hypothesis can be a conjecture about the sample average

# Hypotheses

For the same null hypothesis about the population mean ( $H_0 : E(Y) = \mu_{Y,0}$ ) we can formulate three alternative hypotheses:

- $H_1 : E(Y) > \mu_{Y,0}$  The true value is larger than the null hypothesis value (1-sided.)
- $H_1 : E(Y) < \mu_{Y,0}$  The true value is smaller than the null hypothesis value (1-sided.)
- $H_1 : E(Y) \neq \mu_{Y,0}$  The true value is different (smaller or larger) from the null hypothesis value (2-sided.)

## Important:





A null hypothesis is never accepted, it is either rejected or failed to be rejected.

# Hypothesis testing

- To evaluate the hypothesis the sample average is compared to the hypothesized value
- The sample average ( $\bar{Y}$ ) can differ from the hypothesized value ( $\mu_{y,0}$ ) either because:
  - The null hypothesis is false, i.e. the true mean does not equal  $\mu_{y,0}$
  - Random sampling, i.e. the true mean does equal  $\mu_{y,0}$ , but the sample average differs from the true value due to the nature of random sampling.

# Errors in statistical hypothesis tests

- Type I error: Rejecting the null hypothesis when it is true
- Type II error: Not rejecting the null hypothesis when it is false

HYPOTHESIS TESTING OUTCOMES		Reality	
		The Null Hypothesis Is True	The Alternative Hypothesis is True
R e s e a r c h	The Null Hypothesis Is True	Accurate $1 - \alpha$ 	Type II Error $\beta$ 
	The Alternative Hypothesis is True	Type I Error $\alpha$ 	Accurate $1 - \beta$ 



# Significance level

## Significance level

The prespecified rejection probability of a statistical hypothesis test when the null hypothesis is true.

- We determine the extent to which we can accept making a type I error by specifying the significance level.

## Significance probability

The probability of drawing a statistic at least as adverse to the null hypothesis as the one you actually computed in your sample, assuming that the null hypothesis is correct.

- Each significance level has an associated critical value which is the threshold value for rejecting the null hypothesis at that significance level.

# Power

## Power

The probability that a test correctly rejects the null hypothesis when the alternative is true.

- Power =  $1 - P(\text{type II error})$
- There is a trade-off between the significance level and power, the higher the significance level the lower the power. Because the stronger the evidence needed to reject the null hypothesis the lower the chance that the null hypothesis will be rejected.

# Hypothesis testing

- 1 Choose a desired significance level.
- 2 Perform a hypothesis test.
  - a) Compute t-statistics.
  - b) Compute p-value.

## Population variance

- The population variance ( $\sigma_Y$ ) is required for hypothesis testing.
- However,  $\sigma_Y$  is typically not known.
- The sample variance is an estimator for the population variance

$$s_Y^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2 = \text{"sample variance of Y"}$$

- The standard error of the sample average is given by:

$$SE(\bar{Y}) = \hat{\sigma}_{\bar{Y}} = \frac{s_Y}{\sqrt{n}}$$

### Sample variance

The sample variance is an unbiased and consistent estimate of the population variance as long as the observations are i.i.d. and large outliers are unlikely. ( $E(Y^4) < \infty$ )

# T-test for a population mean

The t-test is used when you want to test whether your data is consistent with a hypothesized population average, when the population standard deviation is unknown.

## T-statistic

$$t = \frac{\bar{Y} - \mu_{Y,0}}{SE(\bar{Y})} = \frac{\bar{Y} - \mu_{Y,0}}{s_y/\sqrt{n}}$$

- The t-statistic is the number of standard errors your sample average is from the hypothesized mean.
- The t-statistic is t-distributed when  $\bar{Y}$  is normally distributed
- The t-statistic has heavier tails than the normal distribution.

# T-test for a population mean

Using the t-statistic for hypothesis testing:

- 1) Compute the t-statistic ( $t^{act}$ )
- 2) Compute the degrees of freedom, which is  $n-1$
- 3) Look up the critical value of your desired significance level ( $t^c$ ) (Table 2, page 805)
- 4) Reject the null hypothesis if:
  - Two sided test:  $|t^{act}| > t_{\alpha, v}^c$
  - One sided test,  $H_1 : \mu_y > \mu_{y0}$   $t > t_{\alpha, v}^c$
  - One sided test,  $H_1 : \mu_y < \mu_{y0}$   $t < -t_{\alpha, v}^c$

Note: Two-sided  $t_{5\%, v}$  equals the one sided  $t_{2.5\%, v}$

## T-test for a population mean

### Example

200 college graduates are asked about their wage. Mean wage in the sample is \$ 22.64 and the sample standard deviation is \$ 18.14. Is this evidence for or against the hypothesis that college graduates earn on average \$ 20 an hour?

$$t^{act} = \frac{22.64 - 20}{\frac{18.14}{\sqrt{200}}} = 2.06$$

Degrees of freedom	5% two sided
120	1.98
$\infty$	1.96

$2.06 > 1.96$  the null hypothesis is rejected at a 5% significance level.

# T-test for a population mean

- It is easy and quick to perform an hypothesis test at the desired significance by calculating the t-statistic.
- However, it conveys less information than if you calculate the p-value.



# P-value for population mean

## P-value

The p-value is the probability of obtaining a test statistic, by random sampling variation, at least as adverse to the null hypothesis value as is the statistic actually observed, assuming that the null hypothesis is correct.

- The probability that we would observe a statistic at least as large as the sample average computed if the null hypothesis is true.
- The smaller the p-value the more unlikely it is to obtain the calculated statistic by random sampling if the null hypothesis is true.
- Assuming that the null is true you would obtain the a difference at least as large as the one observed in  $p\%$  of studies due to random sampling error.

## P-value for population mean

P-value when  $\bar{Y}$  is  $N(\mu_{Y,0}, \sigma_{\bar{Y}}^2)$

$$p - \text{value} = Pr_{H_0}(|Z| > |Z^{act}|) = 2\phi(-|Z^{act}|)$$

- $Pr_{H_0}$  is the denotation for the probability computed under the null hypothesis, i.e. assuming that  $E(Y_i) = \mu_{Y,0}$ .

- 

$$Z = \frac{\bar{Y} - \mu_{Y,0}}{\sigma_{\bar{y}}}$$

- The p-value is the area in the tails of the distribution of  $\bar{Y}$  under the null hypothesis.
- $\phi$  is the standard normal cumulative distribution function.

# P-value for population mean

## P-value when distribution is unknown

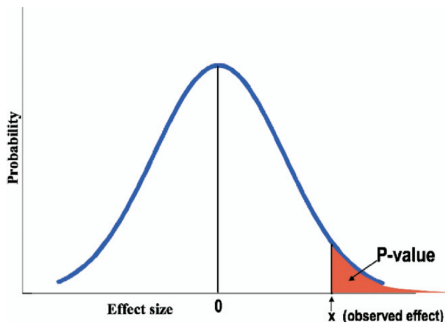
$$p - \text{value} = Pr_{H_0}(|t| > |t^{\text{act}}|) = 2\phi(-|t^{\text{act}}|)$$

$$\begin{aligned} p\text{-value} &= Pr_{H_0} \left( \left| \frac{\bar{Y} - \mu_{Y,0}}{\frac{\sigma_Y}{\sqrt{n}}} \right| > \left| \frac{\bar{Y}^{\text{act}} - \mu_{Y,0}}{\frac{\sigma_Y}{\sqrt{n}}} \right| \right) \\ &\cong Pr_{H_0} \left( \left| \frac{\bar{Y} - \mu_{Y,0}}{\frac{s_Y}{\sqrt{n}}} \right| > \left| \frac{\bar{Y}^{\text{act}} - \mu_{Y,0}}{\frac{s_Y}{\sqrt{n}}} \right| \right) \\ &= 2\phi \left( - \left| \frac{\bar{Y}^{\text{act}} - \mu_{Y,0}}{SE(\bar{Y})} \right| \right) \end{aligned}$$

$\cong$  probability under normal tails.

- When  $n$  is large  $t$  is approximately distributed  $N(0,1)$  (CLT) thus the distribution of the  $t$ -statistic is approximately the same as  $(\bar{Y} - \mu_{Y,0})/\sigma_{\bar{Y}}$ .

## P-value for population mean



**Figure:** Graphical depiction of the definition of a (one-sided) p-value. The curve represents the probability of every observed outcome under the null hypothesis. The p-value is the probability of the observed outcome ( $x$ ) plus all "more extreme" outcomes, represented by the shaded "tail area".

# Rejection rules

Reject the null hypothesis if:

- If  $|t^{act}| > t^c$
- If p-value  $<$  desired significance level

What significance level?

# Confidence interval

- A 95% confidence interval for  $\mu_y$  is an interval that contains the true value of  $\mu_y$  in 95% of repeated samples.
- The confidence interval is the analogue of the significance level, the 95% confidence interval is the set of values of  $\mu_y$  not rejected by a hypothesis test with a 5% significance level.
- 95% confidence interval:  $\mu_y = \{\bar{Y} + / - 1.96SE(\bar{Y})\}$

# Comparing means from two populations

Examples of questions one may ask:

- Are white applicants more likely to be called in for a job interview than African Americans?
- Do men earn more than women?
- Do people with a college degree earn more than those without?

The answer to all these questions involve comparing means of two different population distributions.

## Comparing means from two populations

Let  $m$  denote men and  $w$  denote women. The null hypothesis is that men and women in the population we investigate have the same mean earnings, i.e.  $d_0 = 0$

$$H_0 : \mu_m - \mu_w = d_0 \text{ v.s. } H_1 : \mu_m - \mu_w \neq d_0$$

- Estimate the means,  $\bar{Y}_m - \bar{Y}_w$  is an estimator for  $\mu_m - \mu_w$
- Calculate the standard error

$$SE(\bar{Y}_m - \bar{Y}_w) = \sqrt{\frac{s_m^2}{n_m} + \frac{s_w^2}{n_w}} \text{ (due to CLT, two independent RNV)}$$

- Calculate t-statistic, p-value or confidence interval as normal

$$t = \frac{\bar{Y}_m - \bar{Y}_w - d_0}{SE(\bar{Y}_m - \bar{Y}_w)}$$



## When $n$ is small

- The p-value calculations conducted is based on the assumption that the statistic is approximately normal (CLT and large  $n$ ).
- When  $n$  is small the standard normal distribution can be a poor approximation to the distribution of the t-statistic.
- The exact distribution of the t-statistic depends on the distribution of  $Y$  and it can be very complicated.
- If the population distribution is normally distributed the student t distribution can be used for hypothesis testing.
- However, it is rare that economic variables are normally distributed.