

Econ4570/6560  
Econometrics/Introduction to Econometrics  
Slide 2: Review of Statistics

**Huaming Peng**

Stock and Watson Chapter 2-3

# Lecture outline

- Simple random sampling
- Distribution of the sample average
- Large sample approximation to the distribution of the sample mean
  - Law of large numbers
  - central limit theorem
- Estimation of the population mean
  - unbiasedness
  - consistency
  - efficiency
- Hypothesis test concerning the population mean
- Confidence intervals for the population mean

# Simple random sampling

Simple random sampling means that  $n$  objects are drawn randomly from a population and each object is equally likely to be drawn

Let  $Y_1, Y_2, \dots, Y_n$  denote the 1st to the  $n$ th randomly drawn object.

Under simple random sampling:

- The marginal probability distribution of  $Y_i$  is the same for all  $i = 1, 2, \dots, n$  and equals the population distribution of  $Y$ .
  - because  $Y_1, Y_2, \dots, Y_n$  are drawn randomly from the same population.
- $Y_1$  is distributed independently from  $Y_2, \dots, Y_n$ 
  - knowing the value of  $Y_i$  does not provide information on  $Y_j$  for  $i \neq j$

When  $Y_1, \dots, Y_n$  are drawn from the same population and are independently distributed, they are said to be **i.i.d random variables**

## Simple random sampling: Example

- Let  $G$  be the gender of an individual ( $G = 1$  if female,  $G = 0$  if male)
- $G$  is a Bernoulli random variable with  $E(G) = \mu_G = \Pr(G = 1) = 0.5$
- Suppose we take the population register and randomly draw a sample of size  $n$ 
  - The probability distribution of  $G_i$  is a Bernoulli distribution with mean 0.5
  - $G_1$  is distributed independently from  $G_2, \dots, G_n$
- Suppose we draw a random sample of individuals entering the building of the physics department
  - This is not a sample obtained by simple random sampling and  $G_1, \dots, G_n$  are not i.i.d
  - Men are more likely to enter the building of the physics department!

# The sampling distribution of the sample average

The **sample average**  $\bar{Y}$  of a randomly drawn sample is a random variable with a probability distribution called the **sampling distribution**.

$$\bar{Y} = \frac{1}{n} (Y_1 + Y_2 + \dots + Y_n) = \frac{1}{n} \sum_{i=1}^n Y_i$$

Suppose  $Y_1, \dots, Y_n$  are i.i.d and the mean & variance of the population distribution of  $Y$  are respectively  $\mu_Y$  &  $\sigma_Y^2$

- The mean of  $\bar{Y}$  is

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

- The variance of  $\bar{Y}$  is

$$\begin{aligned} \text{Var}(\bar{Y}) &= \text{Var}\left(\frac{1}{n} \sum_{i=1}^n Y_i\right) \\ &= \frac{1}{n^2} \sum_{i=1}^n \text{Var}(Y_i) + 2 \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1, j \neq i}^n \text{Cov}(Y_i, Y_j) \\ &= \frac{1}{n^2} n \text{Var}(Y) + 0 \\ &= \frac{1}{n} \sigma_Y^2 \end{aligned}$$

## The sampling distribution of the sample average:example

- Let  $G$  be the gender of an individual ( $G = 1$  if female,  $G = 0$  if male)
- The mean of the population distribution of  $G$  is

$$E(G) = \mu_G = p = 0.5$$

- The variance of the population distribution of  $G$  is

$$\text{Var}(G) = \sigma_G^2 = p(1 - p) = 0.5(1 - 0.5) = 0.25$$

- The mean and variance of the average gender (proportion of women)  $\bar{G}$  in a random sample with  $n = 10$  are

$$E(\bar{G}) = \mu_G = 0.5$$

$$\text{Var}(\bar{G}) = \frac{1}{n} \sigma_G^2 = \frac{1}{10} 0.25 = 0.025$$

# The finite sample distribution of the sample average

The finite sample distribution is the sampling distribution that exactly describes the distribution of  $\bar{Y}$  for any sample size  $n$ .

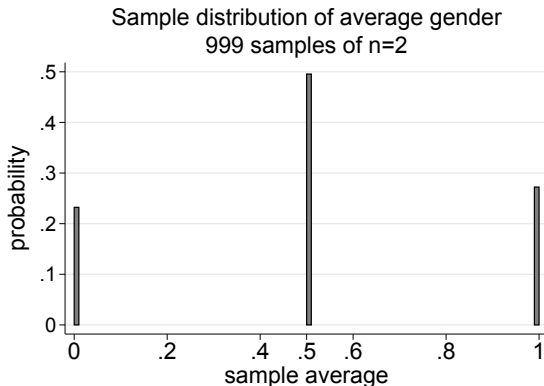
- In general the exact sampling distribution of  $\bar{Y}$  is complicated and depends on the population distribution of  $Y$ .
- A special case is when  $Y_1, Y_2, \dots, Y_n$  are i.i.d draws from the  $N(\mu_Y, \sigma_Y^2)$ , because in this case

$$\bar{Y} \sim N\left(\mu_Y, \frac{\sigma_Y^2}{n}\right)$$

# The finite sample distribution of average gender $\bar{G}$

Suppose we draw 999 samples of  $n = 2$ :

Sample 1			Sample 2			Sample 3			.....	Sample 999		
$G_1$	$G_2$	$\bar{G}$	$G_1$	$G_2$	$\bar{G}$	$G_1$	$G_2$	$\bar{G}$		$G_1$	$G_2$	$\bar{G}$
1	0	0.5	1	1	1	0	1	0.5		0	0	0





# The asymptotic distribution of $\bar{Y}$

- Given that the exact sampling distribution of  $\bar{Y}$  is complicated
- and given that we generally use large samples in econometrics
- we will often use an approximation of the sample distribution that relies on the sample being large

The **asymptotic distribution** is the approximate sampling distribution of  $\bar{Y}$  if the sample size  $n \rightarrow \infty$

We will use two concepts to approximate the large-sample distribution of the sample average

- The law of large numbers.
- The central limit theorem.

# Law of Large Numbers

The Law of Large Numbers states that if

- $Y_i, i = 1, \dots, n$  are independently and identically distributed with  $E(Y_i) = \mu_Y$
- and large outliers are unlikely;  $Var(Y_i) = \sigma_Y^2 < \infty$

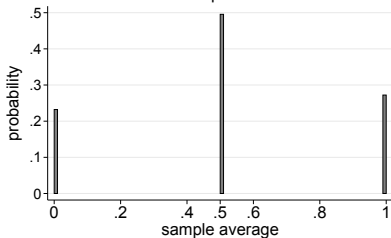
$\bar{Y}$  will be near  $\mu_Y$  with very high probability when  $n$  is very large ( $n \rightarrow \infty$ )

$$\bar{Y} \xrightarrow{p} \mu_Y$$

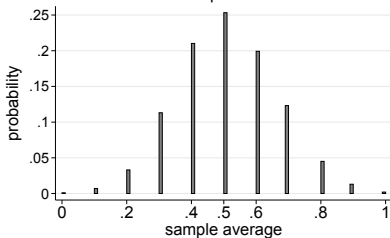
# Law of Large Numbers

Example: Gender  $G \sim \text{Bernoulli}(0.5, 0.25)$

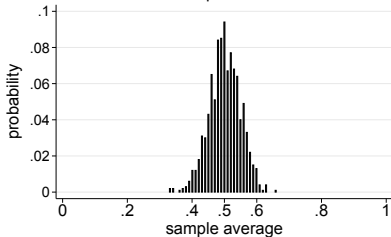
Sample distribution of average gender  
999 samples of  $n=2$



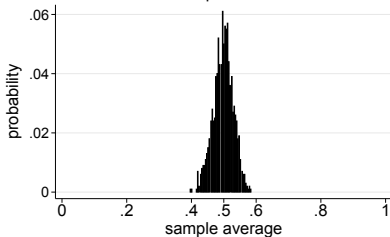
Sample distribution of average gender  
999 samples of  $n=10$



Sample distribution of average gender  
999 samples of  $n=100$



Sample distribution of average gender  
999 samples of  $n=250$



# The Central Limit theorem

The Central Limit Theorem states that if

- $Y_i, i = 1, \dots, n$  are i.i.d. with  $E(Y_i) = \mu_Y$
- and  $\text{Var}(Y_i) = \sigma_Y^2$  with  $0 < \sigma_Y^2 < \infty$

The distribution of the sample average is approximately normal if  $n \rightarrow \infty$

$$\bar{Y} \sim N\left(\mu_Y, \frac{\sigma_Y^2}{n}\right)$$

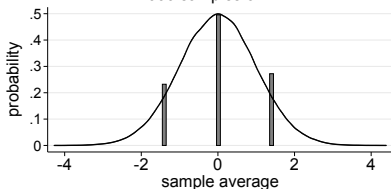
The distribution of the standardized sample average is approximately standard normal for  $n \rightarrow \infty$

$$\frac{\bar{Y} - \mu_Y}{\sigma_{\bar{Y}}^2} \sim N(0, 1)$$

# The Central Limit theorem

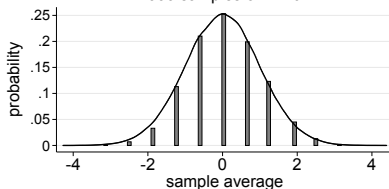
Example: Gender  $G \sim \text{Bernoulli}(0.5, 0.25)$

Sample distribution of average gender  
999 samples of  $n=2$



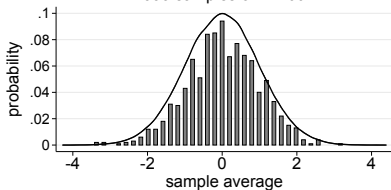
Finite sample distr. standardized sample average  
Standard normal probability density

Sample distribution of average gender  
999 samples of  $n=10$



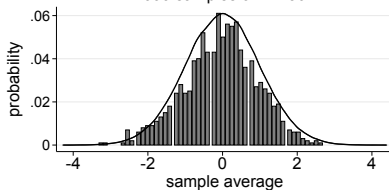
Finite sample distr. standardized sample average  
Standard normal probability density

Sample distribution of average gender  
999 samples of  $n=100$



Finite sample distr. standardized sample average  
Standard normal probability density

Sample distribution of average gender  
999 samples of  $n=250$



Finite sample distr. standardized sample average  
Standard normal probability density

# The Central Limit theorem

How good is the large-sample approximation?

- If  $Y_i \sim N(\mu_Y, \sigma_Y^2)$  the approximation is perfect
- If  $Y_i$  is not normally distributed the quality of the approximation depends on how close  $n$  is to infinity
- For  $n \geq 100$  the normal approximation to the distribution of  $\bar{Y}$  is typically very good for a wide variety of population distributions

# Estimation

# Estimators and estimates

**An Estimator** is a function of a sample of data *to be* drawn randomly from a population

- An estimator is a random variable because of randomness in drawing the sample

**An Estimate** is the numerical value of an estimator when it is actually computed using a specific sample.



## Estimation of the population mean

Suppose we want to know the mean value of  $Y$  ( $\mu_Y$ ) in a population, for example

- The mean wage of college graduates.
- The mean level of education in Norway.
- The mean probability of passing the econometrics exam.

Suppose we draw a random sample of size  $n$  with  $Y_1, \dots, Y_n$  i.i.d

Possible estimators of  $\mu_Y$  are:

- The sample average  $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$
- The first observation  $Y_1$
- The weighted average:  $\tilde{Y} = \frac{1}{n} \left( \frac{1}{2} Y_1 + \frac{3}{2} Y_2 + \dots + \frac{1}{2} Y_{n-1} + \frac{3}{2} Y_n \right)$

# Estimation of the population mean

To determine which of the estimators,  $\bar{Y}$ ,  $Y_1$  or  $\tilde{Y}$  is the best estimator of  $\mu_Y$  we consider 3 properties:

Let  $\hat{\mu}_Y$  be an estimator of the population mean  $\mu_Y$ .

**Unbiasedness:** The mean of the sampling distribution of  $\hat{\mu}_Y$  equals  $\mu_Y$

$$E(\hat{\mu}_Y) = \mu_Y$$

**Consistency:** The probability that  $\hat{\mu}_Y$  is within a very small interval of  $\mu_Y$  approaches 1 if  $n \rightarrow \infty$

$$\hat{\mu}_Y \xrightarrow{p} \mu_Y$$

**Efficiency:** If the variance of the sampling distribution of  $\hat{\mu}_Y$  is smaller than that of some other estimator  $\tilde{\mu}_Y$ ,  $\hat{\mu}_Y$  is more efficient

$$Var(\hat{\mu}_Y) < Var(\tilde{\mu}_Y)$$

# Example

Suppose we are interested in the mean wages  $\mu_w$  of individuals with a master degree

We draw the following sample ( $n = 10$ ) by simple random sampling

$i$	$W_i$
1	47281.92
2	70781.94
3	55174.46
4	49096.05
5	67424.82
6	39252.85
7	78815.33
8	46750.78
9	46587.89
10	25015.71

The 3 estimators give the following estimates:

$$\overline{W} = \frac{1}{10} \sum_{i=1}^{10} W_i = 52618.18$$

$$W_1 = 47281.92$$

$$\widetilde{W} = \frac{1}{10} \left( \frac{1}{2} W_1 + \frac{3}{2} W_2 + \dots + \frac{1}{2} W_9 + \frac{3}{2} W_{10} \right) = 49398.82.$$

# Unbiasedness

All 3 proposed estimators are unbiased:

- As shown on slide 5:  $E(\bar{Y}) = \mu_Y$

- Since  $Y_i$  are i.i.d.  $E(Y_1) = E(Y) = \mu_Y$

- $$\begin{aligned}
 E(\tilde{Y}) &= E\left(\frac{1}{n} \left(\frac{1}{2} Y_1 + \frac{3}{2} Y_2 + \dots + \frac{1}{2} Y_{n-1} + \frac{3}{2} Y_n\right)\right) \\
 &= \frac{1}{n} \left(\frac{1}{2} E(Y_1) + \frac{3}{2} E(Y_2) + \dots + \frac{1}{2} E(Y_{n-1}) + \frac{3}{2} E(Y_n)\right) \\
 &= \frac{1}{n} \left[\left(\frac{n}{2} \cdot \frac{1}{2}\right) E(Y_i) + \left(\frac{n}{2} \cdot \frac{3}{2}\right) E(Y_i)\right] \\
 &\qquad\qquad\qquad E(Y_i) \qquad\qquad\qquad = \mu_Y
 \end{aligned}$$

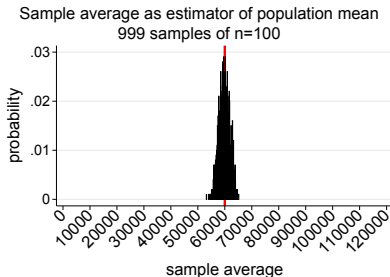
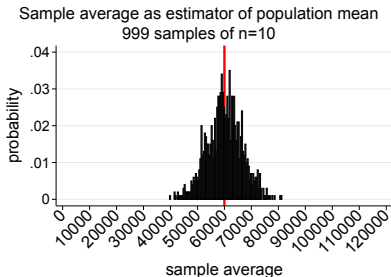
# Consistency

Example: mean wages of individuals with a master degree with  $\mu_W = 60\,000$

By the law of large numbers

$$\overline{W} \xrightarrow{p} \mu_W$$

which implies that the probability that  $\overline{W}$  is within a very small interval of  $\mu_W$  approaches 1 if  $n \rightarrow \infty$

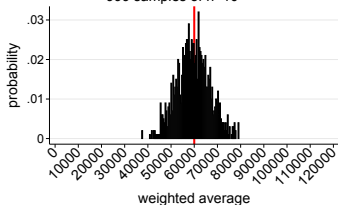


# Consistency

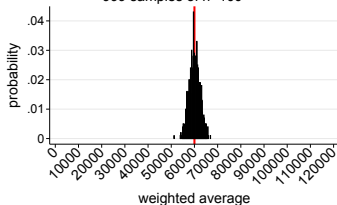
Example: mean wages of individuals with a master degree with  $\mu_W = 60\,000$

$\widetilde{W} = \frac{1}{n} \left( \frac{1}{2} W_1 + \frac{3}{2} W_2 + \dots + \frac{1}{2} W_{n-1} + \frac{3}{2} W_n \right)$  is also consistent

Weighted average as estimator of population mean  
999 samples of  $n=10$

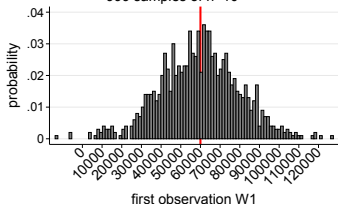


Weighted average as estimator of population mean  
999 samples of  $n=100$

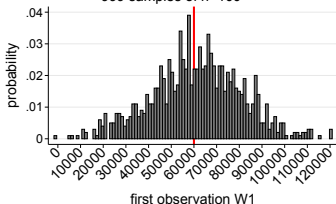


However  $W_1$  is not a consistent estimator of  $\mu_W$ :

First observation  $W_1$  as estimator of population mean  
999 samples of  $n=10$



First observation  $W_1$  as estimator of population mean  
999 samples of  $n=100$



# Efficiency

Efficiency entails a comparison of estimators on the basis of their variance

- The variance of  $\bar{Y}$  equals

$$\text{Var}(\bar{Y}) = \frac{1}{n} \sigma_Y^2$$

- The variance of  $Y_1$  equals

$$\text{Var}(Y_1) = \text{Var}(Y) = \sigma_Y^2$$

- The variance of  $\tilde{Y}$  equals

$$\text{Var}(\tilde{Y}) = 1.25 \frac{1}{n} \sigma_Y^2$$

For any  $n \geq 2$   $\bar{Y}$  is more efficient than  $Y_1$  and  $\tilde{Y}$

# BLUE: Best Linear Unbiased Estimator

- $\bar{Y}$  is not only more efficient than  $Y_1$  and  $\tilde{Y}$ , but it is more efficient than any unbiased estimator of  $\mu_Y$  that is a weighted average of  $Y_1, \dots, Y_n$

$\bar{Y}$  is the Best Linear Unbiased Estimator (BLUE) it is the most efficient estimator of  $\mu_Y$  among all unbiased estimators that are weighted averages of  $Y_1, \dots, Y_n$

- Let  $\hat{\mu}_Y$  be an unbiased estimator of  $\mu_Y$

$$\hat{\mu}_Y = \frac{1}{n} \sum_{i=1}^n a_i Y_i \quad \text{with } a_1, \dots, a_n \text{ nonrandom constants}$$

then  $\bar{Y}$  is more efficient than  $\hat{\mu}_Y$ , that is

$$\text{Var}(\bar{Y}) < \text{Var}(\hat{\mu}_Y)$$



# Hypothesis tests concerning the population mean

# Hypothesis tests concerning the population mean

Consider the following questions:

- Is the mean monthly wage of college graduates equal to NOK 60 000?
- Is the mean level of education in Norway equal to 12 years?
- Is the mean probability of passing the econometrics exam equal to 1?

These questions involve the population mean taking on a specific value  $\mu_{Y,0}$

Answering these questions implies using data to compare a null hypothesis

$$H_0 : E(Y) = \mu_{Y,0}$$

to an alternative hypothesis, which is often the following two sided hypothesis

$$H_1 : E(Y) \neq \mu_{Y,0}$$

# Hypothesis tests concerning the population mean

## p-value

Suppose we have a sample of  $n$  i.i.d observations and compute the sample average  $\bar{Y}$

The sample average can differ from  $\mu_{Y,0}$  for two reasons

- 1 The population mean  $\mu_Y$  is not equal to  $\mu_{Y,0}$  ( $H_0$  not true)
- 2 Due to random sampling  $\bar{Y} \neq \mu_Y = \mu_{Y,0}$  ( $H_0$  true)

To quantify the second reason we define the p-value

**The p-value** is the probability of drawing a sample with  $\bar{Y}$  at least as far from  $\mu_{Y,0}$  given that the null hypothesis is true.

# Hypothesis tests concerning the population mean

p-value

$$p - value = Pr_{H_0} \left[ |\bar{Y} - \mu_{Y,0}| > |\bar{Y}^{act} - \mu_{Y,0}| \right]$$

To compute the p-value we need to know the sampling distribution of  $\bar{Y}$

- Sampling distribution of  $\bar{Y}$  is complicated for small  $n$
- With large  $n$  the central limit theorem states that

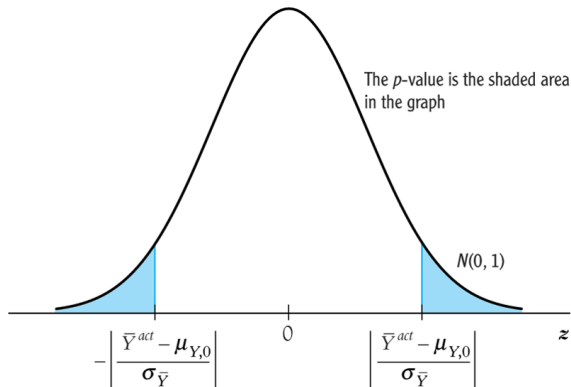
$$\bar{Y} \sim N \left( \mu_Y, \frac{\sigma_Y^2}{n} \right)$$

- This implies that if the null hypothesis is true:

$$\frac{\bar{Y} - \mu_{Y,0}}{\sqrt{\frac{\sigma_Y^2}{n}}} \sim N(0, 1)$$

# Computing the p-value when $\sigma_Y$ is known

$$p - value = Pr_{H_0} \left[ \left| \frac{\bar{Y} - \mu_{Y,0}}{\sqrt{\frac{\sigma_Y^2}{n}}} \right| > \left| \frac{\bar{Y}^{act} - \mu_{Y,0}}{\sqrt{\frac{\sigma_Y^2}{n}}} \right| \right] = 2\Phi \left( - \left| \frac{\bar{Y}^{act} - \mu_{Y,0}}{\sqrt{\frac{\sigma_Y^2}{n}}} \right| \right)$$



- For large  $n$ ,  $p$ -value = the probability that  $Z$  falls outside  $\left| \frac{\bar{Y}^{act} - \mu_{Y,0}}{\sqrt{\frac{\sigma_Y^2}{n}}} \right|$

# Estimating the standard deviation of $\bar{Y}$

- In practice  $\sigma_Y^2$  is usually unknown and must be estimated

The sample variance  $s_Y^2$  is the estimator of  $\sigma_Y^2 = E[(Y_i - \mu_Y)^2]$

$$s_Y^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$$

- division by  $n - 1$  because we “replace”  $\mu_Y$  by  $\bar{Y}$  which uses up 1 degree of freedom
- if  $Y_1, \dots, Y_n$  are i.i.d. and  $E(Y^4) < \infty$ ,  $s_Y^2 \xrightarrow{p} \sigma_Y^2$   
(Law of Large Numbers)

The sample standard deviation  $s_Y = \sqrt{s_Y^2}$  is the estimator of  $\sigma_Y$

# Computing the p-value using $SE(\bar{Y}) = \hat{\sigma}_{\bar{Y}}$

The standard error  $SE(\bar{Y})$  is an estimator of  $\sigma_{\bar{Y}}$

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

- Because  $s_Y^2$  is a consistent estimator of  $\sigma_Y^2$ , we can (for large  $n$ ) replace  $\sqrt{\frac{\sigma_Y^2}{n}}$  by  $SE(\bar{Y}) = \frac{s_Y}{\sqrt{n}}$
- This implies that when  $\sigma_Y^2$  is unknown and  $Y_1, \dots, Y_n$  are i.i.d. the p-value is computed as

$$p - value = 2\Phi \left( - \left| \frac{\bar{Y}^{act} - \mu_{Y,0}}{SE(\bar{Y})} \right| \right)$$

# The t-statistic and its large-sample distribution

- The standardized sample average  $(\bar{Y}^{act} - \mu_{Y,0}) / SE(\bar{Y})$  plays a central role in testing statistical hypothesis
- It has a special name, the **t-statistic**

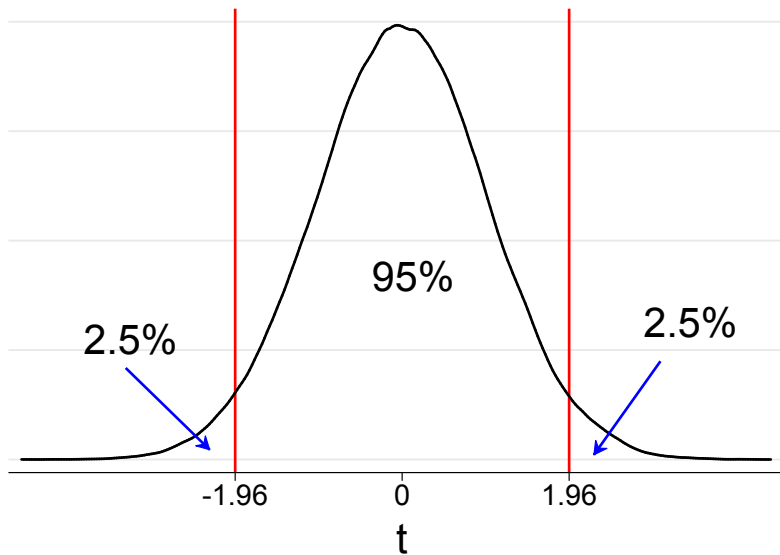
$$t = \left| \frac{\bar{Y} - \mu_{Y,0}}{SE(\bar{Y})} \right|$$

- $t$  is approximately  $N(0, 1)$  distributed for large  $n$
- The p-value can be computed as

$$p - value = 2\Phi(-|t^{act}|)$$



# The t-statistic and its large-sample distribution



## Type I and type II errors and the significance level

There are 2 types of mistakes when conducting a hypothesis test

**Type I error** refers to the mistake of rejecting  $H_0$  when it is true

**Type II error** refers to the mistake of not rejecting  $H_0$  when it is false

- In hypothesis testing we usually fix the probability of a type I error

The **significance level**  $\alpha$  is the probability of rejecting  $H_0$  when it is true

- Most often used significance level is 5% ( $\alpha = 0.05$ )

Since area in tails of  $N(0, 1)$  outside  $\pm 1.96$  is 5%:

- We reject  $H_0$  if p-value is smaller than 0.05.
- We reject  $H_0$  if  $|t^{act}| > 1.96$

## 4 steps in testing a hypothesis about the population mean

$$H_0 : E(Y) = \mu_{Y,0} \quad H_1 : E(Y) \neq \mu_{Y,0}$$

Step 1: Compute the sample average  $\bar{Y}$

Step 2: Compute the standard error of  $\bar{Y}$

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

Step 3: Compute the t-statistic

$$t^{act} = \frac{\bar{Y} - \mu_{Y,0}}{SE(\bar{Y})}$$

Step 4: Reject the null hypothesis at a 5% significance level if

- $|t^{act}| > 1.96$
- or if  $p\text{-value} < 0.05$

# Hypothesis tests concerning the population mean

Example: The mean wage of individuals with a master degree

Suppose we would like to test

$$H_0 : E(W) = 60000 \quad H_1 : E(W) \neq 60000$$

using a sample of 250 individuals with a master degree

$$\text{Step 1: } \bar{W} = \frac{1}{n} \sum_{i=1}^n W_i = 61977.12$$

$$\text{Step 2: } SE(\bar{W}) = \frac{s_W}{\sqrt{n}} = 1334.19$$

$$\text{Step 3: } t^{act} = \frac{\bar{W} - \mu_{W,0}}{SE(\bar{W})} = \frac{61977.12 - 60000}{1334.19} = 1.48$$

Step 4: Since we use a 5% significance level, we do not reject  $H_0$  because  $|t^{act}| = 1.48 < 1.96$

*Note: We do never accept the null hypothesis!*

# Hypothesis tests concerning the population mean

Example: The mean wage of individuals with a master degree

This is how to do the test in Stata:

\_\_\_\_\_ (R)  
 /\_/\_/\_/\_/\_/  
 /\_/\_/\_/\_/\_/  
 Statistics/Data Analysis

```
. ttest wage=60000
```

One-sample t test

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
wage	250	61977.12	1334.189	21095.37	59349.39	64604.85

```
mean = mean( wage)          t = 1.4819
Ho: mean = 60000             degrees of freedom = 249
```

Ha: mean < 60000  
 Pr(T < t) = 0.9302

Ha: mean != 60000  
 Pr(|T| > |t|) = 0.1396

Ha: mean > 60000  
 Pr(T > t) = 0.0698

# Hypothesis test with a one-sides alternative

- Sometimes the alternative hypothesis is that the mean exceeds  $\mu_{Y,0}$

$$H_0 : E(Y) = \mu_{Y,0} \quad H_1 : E(Y) > \mu_{Y,0}$$

- In this case the p-value is the area under  $N(0, 1)$  to the right of the t-statistic

$$p - value = Pr_{H_0} (t > t^{act}) = 1 - \Phi (t^{act})$$

- With a significance level of 5% ( $\alpha = 0.05$ ) we reject  $H_0$  if  $t^{act} > 1.64$
- If the alternative hypothesis is  $H_1 : E(Y) < \mu_{Y,0}$

$$p - value = Pr_{H_0} (t < t^{act}) = 1 - \Phi (-t^{act})$$

and we reject  $H_0$  if  $t^{act} < -1.64$  /  $p - value < 0.05$

Statistics/Data Analysis (R)

One-sample t test

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
wage	250	61977.12	1334.189	21095.37	59349.39	64604.85

```
mean = mean( wage)                                t = 1.4819
Ho: mean = 60000                                degrees of freedom = 249
```

```
Ha: mean < 60000
Pr(T < t) = 0.9302
```

$$\begin{aligned} H_a: \text{mean} & \neq 60000 \\ \Pr(|T| > |t|) & = 0.1396 \end{aligned}$$

```
Ha: mean > 60000
Pr(T > t) = 0.0698
```

## Confidence intervals for the population mean

- Suppose we would do a two-sided hypothesis test for many different values of  $\mu_{Y,0}$
- On the basis of this we can construct a set of values which are not rejected at a 5% significance level
- If we were able to test all possible values of  $\mu_{Y,0}$  we could construct a 95% confidence interval

A **95% confidence interval** is an interval that contains the true value of  $\mu_Y$  in 95% of all possible random samples.

- Instead of doing infinitely many hypothesis tests we can compute the 95% confidence interval as

$$\left\{ \bar{Y} - 1.96 \cdot SE(\bar{Y}) \quad , \quad \bar{Y} + 1.96 \cdot SE(\bar{Y}) \right\}$$

- Intuition: a value of  $\mu_{Y,0}$  smaller than  $\bar{Y} - 1.96 \cdot SE(\bar{Y})$  or bigger than  $\bar{Y} + 1.96 \cdot SE(\bar{Y})$  will be rejected at  $\alpha = 0.05$



# Confidence intervals for the population mean

Example: The mean wage of individuals with a master degree

When the sample size  $n$  is large:

$$95\% \text{ confidence interval for } \mu_Y = \left\{ \bar{Y} \pm 1.96 \cdot SE(\bar{Y}) \right\}$$

$$90\% \text{ confidence interval for } \mu_Y = \left\{ \bar{Y} \pm 1.64 \cdot SE(\bar{Y}) \right\}$$

$$99\% \text{ confidence interval for } \mu_Y = \left\{ \bar{Y} \pm 2.58 \cdot SE(\bar{Y}) \right\}$$

Using the sample of 250 individuals with a master degree:

95% conf. int. for  $\mu_W$  is

$$\{61977.12 \pm 1.96 \cdot 1334.19\} = \{59349.39, 64604.85\}$$

90% conf. int. for  $\mu_W$  is

$$\{61977.12 \pm 1.64 \cdot 1334.19\} = \{59774.38, 64179.86\}$$

99% conf. int. for  $\mu_W$  is

$$\{61977.12 \pm 2.58 \cdot 1334.19\} = \{58513.94, 65440.30\}$$