Applied Statistics and Econometrics
Lecture 3

Giuseppe Ragusa

Luiss University

gragusa@luiss.it
http://gragusa.org/

February 27, 2016

Where are we?

1. The probability framework for statistical inference
2. Estimation
3. **Hypothesis Testing**
4. Confidence intervals

Hypothesis Testing

The hypothesis testing problem (for the mean): make a provisional decision, based on the evidence at hand, whether a null hypothesis is true, or instead that some alternative hypothesis is true. That is, test

$$H0 : E(Y) = \mu_{Y,0} \text{ vs. } H1 : E(Y) > \mu_{Y,0} \text{ (1-sided, ¿)}$$

$$H0 : E(Y) = \mu_{Y,0} \text{ vs. } H1 : E(Y) < \mu_{Y,0} \text{ (1-sided, ¡)}$$

$$H0 : E(Y) = \mu_{Y,0} \text{ vs. } H1 : E(Y) \neq \mu_{Y,0} \text{ (2-sided, } \neq \text{)}$$

Some terminology for testing statistical hypotheses:

The significance level of a test

is a pre-specified probability of incorrectly rejecting the null, when the null is true.

Hypothesis

Testing (Two sided)

$$H_0 : \mu_Y = \mu_{Y,0} \quad \text{vs.} \quad H_1 : \mu_Y \neq \mu_{Y,0}$$

We reject the null hypothesis at the **5%** significance level if:

$$\frac{\bar{Y} - \mu_{Y,0}}{s_Y/\sqrt{n}} > 1.96 \quad \text{or} \quad \frac{\bar{Y} - \mu_{Y,0}}{s_Y/\sqrt{n}} < -1.96$$

or, more compactly, if

$$\left| \frac{\bar{Y} - \mu_{Y,0}}{s_Y/\sqrt{n}} \right| > 1.96$$

Hypothesis Testing

Testing (Two sided)

$$H_0 : \mu_Y = \mu_{Y,0} \quad \text{vs.} \quad H_1 : \mu_Y \neq \mu_{Y,0}$$

We reject the null hypothesis at the **10%** significance level if:

$$\frac{\bar{Y} - \mu_{Y,0}}{s_Y / \sqrt{n}} > 1.64 \qquad \text{or} \qquad \frac{\bar{Y} - \mu_{Y,0}}{s_Y / \sqrt{n}} < -1.64$$

or, more compactly, if

$$\left| \frac{\bar{Y} - \mu_{Y,0}}{s_Y / \sqrt{n}} \right| > 1.64$$

Hypothesis Testing

Testing (one sided)

$$H_0 : \mu_Y = \mu_{Y,0} \quad \text{vs.} \quad H_1 : \mu_Y > \mu_{Y,0}$$

We reject the null hypothesis at the **5%** significance level if:

$$\frac{\bar{Y} - \mu_{Y,0}}{s_Y/\sqrt{n}} \geq 1.64$$

$$H0 : \mu_Y = \mu_{Y,0} \text{ vs. } H1 : \mu_Y < \mu_{Y,0}$$

We reject the null hypothesis at the **5%** significance level if:

$$\frac{\bar{Y} - \mu_{Y,0}}{s_Y/\sqrt{n}} \leq 1.64$$

Terminology

The quantity:

$$t = \frac{\bar{Y} - \mu_{Y,0}}{s_Y / \sqrt{n}}$$

is referred to as the (Student's) *t*-statistics.

The same quantity can be equivalently expressed as:

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

where $SE(\bar{Y}) = s_Y / \sqrt{n}$ is called the standard error of $\bar{Y}$.

CORRIERE DELLA SERA

Una città, un Paese
Napoli metropoli ribelle cerca l'orgoglio perduto

Risparmio
Investire nei fondi
La classifica dei migliori

LA CLASSE DIRIGENTE ITALIANA
UN CLUB ESCLUSIVO DI UNA CERTA ETÀ

Le cifre Eurostat sui salari. Draghi: alcuni Paesi Ue tutelano il posto e non i dipendenti
Stipendi tra i più bassi d'Europa
Agli italiani la metà dei tedeschi. Fornero: sistema da scardinare

Sconfitta e Ranieri in bilico

In primo piano
I troppi passaggi tra busta paga e costo del lavoro

Banche-imprese: moratoria dei debiti per i piccoli

Adesso l'Inter non c'è più

La linea del governo: esenzione per la stragrande maggioranza degli istituti
Così l'Ici sulle scuole cattoliche
Non pagheranno se aperte a tutti e senza profitti

I veleni di Milan e Juve

Il gesto mancato di Buffon

Lega in affanno
La politica (smarrita) del Carroccio

Dirigenti pubblici e intreccio di regole
Maxi compensi, tetto in bilico

Il caso Conti in rosso, sprechi e ritardi. Emergenza per Lazio, Puglia e Campania
Sanità, rischio bancarotta in 7 Regioni

Il processo Lea Garofalo
Sciolta nell'acido: ma per la legge non è vittima della 'ndrangheta

L'ipotesi corruzione
«Spesa eccessiva» Inchiesta in India su dodici elicotteri dell'Agusta

SEAT
Ibiza a € 8.950 solo a febbraio.

Figure 1: Front page Corriere della Sera, March 1, 2012.

The actual OCSE data

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Belgium (2)	31,644	33,109	34,330	34,643	35,704	36,673	37,674	38,659	40,698	:	:
Bulgaria (2)	1,430	1,514	1,588	1,678	1,784	1,978	2,195	2,626	3,328	4,085	:
Czech Republic (3)	4,616	5,142	6,016	6,137	6,569	7,405	8,284	9,071	10,930	10,596	11,312
Denmark	40,962	41,661	43,577	44,692	46,122	47,529	48,307	53,165	55,001	56,044	:
Germany	34,400	35,200	36,400	37,200	38,100	38,700	39,364	40,200	*41,400*	*41,100*	*42,400*
Estonia (2)(3)	3,887	4,343	4,778	5,278	5,658	6,417	:	:	10,045	9,492	9,712
Ireland	:	:	:	:	:	40,462	:	39,858	45,893	45,207	:
Greece	14,723	15,431	16,278	16,739	:	:	:	:	25,915	29,160	:
Spain	17,432	17,874	18,462	19,220	19,931	20,333	21,402	21,891	25,208	26,316	:
France (2)	26,712	27,418	28,185	28,847	29,608	30,521	31,369	32,413	33,574	34,132	:
Italy (3)	19,991	20,583	21,076	21,494	:	22,657	23,406	:	:	:	:
Cyprus (3)	16,086	16,736	17,431	18,165	19,290	20,549	21,310	:	:	24,775	25,251
Latvia (2)	3,247	3,426	3,523	3,515	3,806	4,246	5,211	6,690	8,676	8,728	8,596
Lithuania (3)(4)	3,591	3,726	4,046	4,195	4,367	4,770	5,543	6,745	7,398	7,406	7,234
Luxembourg (2)	35,875	37,745	38,442	39,587	40,575	42,135	43,621	45,284	47,034	48,174	49,316
Hungary	4,173	4,898	5,846	6,447	7,119	7,798	7,866	8,952	10,237	9,603	10,100
Malta (2)	13,461	13,791	14,068	14,096	14,116	14,706	15,278	15,679	16,158	:	:
Netherlands	31,901	33,900	35,200	36,600	37,900	38,700	40,800	42,000	43,146	44,412	:
Austria (2)	:	:	:	:	34,995	36,032	36,673	37,716	32,787	33,384	:
Poland (3)(4)	6,226	7,510	7,173	6,434	6,230	6,270	8,178	:	10,787	8,399	:
Portugal	12,620	13,338	13,322	13,350	13,700	14,042	14,893	15,345	16,691	17,129	17,352
Romania (2)(3)	1,748	1,993	2,075	2,142	2,414	3,155	3,713	4,825	5,457	*5,450*	*5,891*
Slovenia (3)	10,316	10,851	11,461	11,932	12,466	12,985	13,687	14,625	15,997	16,282	17,168
Slovakia	3,583	3,837	4,582	4,945	5,706	6,374	7,040	8,400	9,707	10,387	10,777
Finland (2)	27,398	28,555	29,916	30,978	31,988	33,290	34,080	36,114	37,946	39,197	:
Sweden	31,621	30,467	31,164	32,177	33,344	34,027	35,084	36,871	37,597	34,746	40,008
United Kingdom	37,676	39,233	40,553	38,793	41,286	42,866	44,496	46,051	:	38,047	:
Iceland	37,641	34,100	:	:	:	:	:	:	:	:	:
Norway (2)	36,202	38,604	43,750	40,883	42,152	45,560	47,221	:	:	51,343	
Switzerland (3)	43,682	:	48,499	:	45,760	:	46,058	:	47,088	:	:
Croatia (3)	:	:	:	8,491	9,036	9,634	:	:	11,979	11,969	:

Figure 2: Earnings in the business economy (average gross annual earnings of full-time employees), 2000-2010

Italian labor force survey

The Italian Labour Force Survey (Lfs)

provides data on labour market variables (employment status, type of work, work experience, job search, etc.), disaggregated by gender, age and territory (up to regional detail on a quarterly base).

```
##    RETRIC ETAM DETIND
## 1    1530   50      2
## 6    1600   61      2
## 7    1500   46      2
## 10   2800   43      2
## 11   1300   33      2
## 12    940   38      1
## 16   1700   57      2
## 21   2180   32      2
## 25   1470   52      2
## 26    700   50      2
## 45   1800   46      2
## 46   1100   42      2
## 48   1550   50      2
## 49   1250   44      2
```

Italian LFS

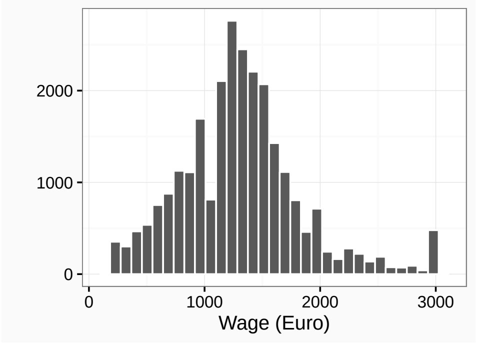


Figure 3: Italian (net) monthly wages.
Source: ISTAT, 2015Q3

Table 1: Wage of italian employees

Statistic	N	Mean	St. Dev.
RETRIC	26,127	1,306.757	522.510

Example: Italian wages, ctd.

The sample mean of Italian (net) wages is

$$\overline{Wage} = 1307,$$

with a standard deviation of

$$s_W = 523.$$

Example: Italian wages, ctd.

The sample mean of Italian (net) wages is

$$\overline{Wage} = 1307,$$

with a standard deviation of

$$s_W = 523.$$

Test on the population mean

$$H_0 : E[Wage] = 1300, \quad vs \quad H_1 : E[wage] \neq 1300$$

Example: Italian wages, ctd.

Steps:

1. Calcualte the t-statistics

$$t = \frac{\overline{Wage} - 1300}{s_W/\sqrt{n}} = 2.1$$

Example: Italian wages, ctd.

Steps:

1. Calcualte the t-statistics

$$t = \frac{\overline{Wage} - 1300}{s_W/\sqrt{n}} = 2.1$$

2. Compare the absolute value with the critical value

$$|t| > 1.96$$

Example: Italian wages, ctd.

Steps:

1. Calcualte the t-statistics

$$t = \frac{\overline{Wage} - 1300}{s_W/\sqrt{n}} = 2.1$$

2. Compare the absolute value with the critical value

$$|t| > 1.96$$

3. Draw a conclusion
    - We reject the null hypothesis that Italian average monthly net wages are 1,300 euro (at the 5% significance level)

The p-value

p-value

probability of drawing a statistic (e.g. $\bar{Y}$) at least as adverse to the null as the value actually computed with your data, assuming that the null hypothesis is true. Calculating the p-value based on $\bar{Y}$:

$$p - value = \Pr[|\bar{Y} - \mu_{Y,0}| > |\bar{Y}^{act} - \mu_{Y,0}|]$$

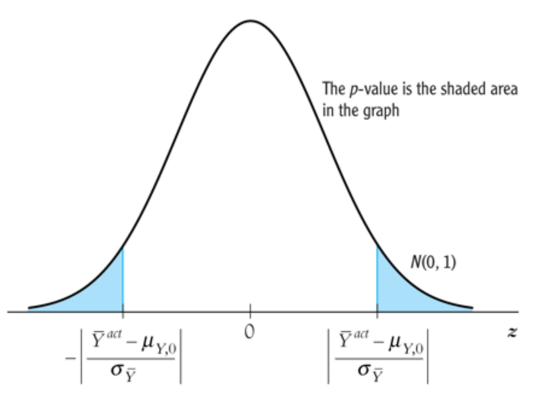
where $\bar{Y}^{act}$ is the value of $\bar{Y}$ actually observed (nonrandom)

Calculating the p-value, ctd.

- To compute the p-value, you need the to know the sampling distribution of $\bar{Y}$, which is complicated if n is small.
- If n is large, you can use the normal approximation (CLT):

$$
\begin{aligned}
p - value &= \Pr_{H_0}[|\bar{Y} - \mu_{Y,0}| > |\bar{Y}^{act} - \mu_{Y,0}|] \\
&= \Pr_{H_0}\left[\frac{|\bar{Y} - \mu_{Y,0}|}{\sigma_Y/\sqrt{n}} > \frac{|\bar{Y}^{act} - \mu_{Y,0}|}{\sigma_Y/\sqrt{n}}\right] \\
&\approx \text{probability under left+right of } N(0,1) \text{ density}
\end{aligned}
$$

Calculating the p-value with σ_Y known:



- For large n, p-value = the probability that a N(0,1) random variable falls outside $\sqrt{n}\frac{|(\bar{Y}-\mu_{Y,0})|}{\sigma_Y}$
- In practice, is unknown - it must be **estimated**

Computing the p-value with σ_Y estimated:

$$p - value = \Pr_{H_0}[|\bar{Y} - \mu_{Y,0}| > |\bar{Y}^{act} - \mu_{Y,0}|]$$
$$= \Pr_{H_0}[\frac{|\bar{Y} - \mu_{Y,0}|}{s_Y/\sqrt{n}} > \frac{|\bar{Y}^{act} - \mu_{Y,0}|}{s_Y/\sqrt{n}}]$$
$$\approx \text{probability under left+right N(0,1) density}$$

so,

$$p - value = \Pr_{H_0}[|t| > |t^{act}|] \qquad (\sigma_Y \text{ estimated})$$
$$= \text{probability under normal N(0,1) tails outside } |t^{act}|$$

where t is the **t-statistic** seen as a random variable.

What is the link between the p-value and the significance level?

Computer programs often communicate the p-value since the p-value contains more information.

For example, if the prespecified significance level is 5%,

- you reject the null hypothesis if $|t| > 1.96$
- equivalently, you reject H_0 if $p - value < 0.05$.

In general:

If $p - value < (\alpha \times 100)\%$ we reject the null hypothesis at $(\alpha \times 100)\%$.

t-test and t-table

At this point, you might be wondering,. . .

What happened to the t-table and the degrees of freedom?

Digression: the Student t distribution

If Y_i, $i = 1, \ldots, n$ is i.i.d. $N(\mu_Y, \sigma_Y^2)$, then the t-statistic has the Student t-distribution with n-1 degrees of freedom.

- The critical values of the Student t-distribution is tabulated in the back of all statistics books.

Comments on the Student t-distribution

1. The theory of the t-distribution was one of the early triumphs of mathematical statistics. It is astounding, really: if Y is i.i.d. normal, then you can know the exact, finite-sample distribution of the t-statistic - it is the Student t. So, you can construct confidence intervals (using the Student t critical value) that have exactly the right coverage rate, no matter what the sample size. This result was really useful in times when "computer" was a job title, data collection was expensive, and the number of observations was perhaps a dozen. It is also a conceptually beautiful result, and the math is beautiful too - which is probably why stats profs love to teach the t-distribution. But. . . .

Comments on Student t distribution, ctd.

1. If the sample size is moderate (several dozen) or large (hundreds or more), the difference between the t-distribution and N(0,1) critical values are negligible. Here are some 5% critical values for 2-sided tests:

degrees of freedom (n - 1)	5% t-distribution critical value
10	2.23
20	2.09
30	2.04
60	2.00
∞	1.96

Comments on Student t distribution, ctd.

1. So, the Student-t distribution is only relevant when the sample size is very small; but in that case, for it to be correct, you must be sure that the population distribution of Y is normal. In economic data, the normality assumption is rarely credible. Here are the distributions of some economic data.
2. Do you think earnings are normally distributed?
3. Suppose you have a sample of $n = 10$ observations from one of these distributions — would you feel comfortable using the Student t distribution?

Comments on Student t distribution, ctd.

1. You might not know this. Consider the t-statistic testing the hypothesis that two means (groups s, l) are equal:

$$t = \frac{\bar{Y}_s - \bar{Y}_l}{\sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}}} = \frac{\bar{Y}_s - \bar{Y}_l}{SE(\bar{Y}_s - \bar{Y}_l)}$$

- Even if the population distribution of Y in the two groups is normal, this statistic doesn't have a Student t distribution!
- There is a statistic testing this hypothesis that has a normal distribution, the "pooled variance" t-statistic - see SW (Section 3.6) - however the pooled variance t-statistic is only valid if the variances of the normal distributions are the same in the two groups.
- Would you expect this to be true, say, for men's v. women's wages?

The Student-t distribution - summary

- The assumption that Y is distributed $N(\mu_Y, \sigma_Y^2)$ is rarely plausible in practice (income? number of children?)
- For n ¿ 30, the t-distribution and $N(0,1)$ are very close (as n grows large, the $t(n-1)$ distribution converges to $N(0,1)$)
- The t-distribution is an artifact from days when sample sizes were small and "computers" were people
- For historical reasons, statistical software typically uses the t-distribution to compute p-values - but this is irrelevant when the sample size is moderate or large.
- For these reasons, in this class we will focus on the large-n approximation given by the CLT

Where are we?

1. The probability framework for statistical inference
2. Estimation
3. Testing
4. **Confidence intervals**

Confidence Intervals

Definition

- A 95% confidence interval for μ_Y is an interval that contains the true value of $\bar{Y}$ in 95% of repeated samples.

Confidence Intervals

Definition

- A 95% confidence interval for μ_Y is an interval that contains the true value of $\bar{Y}$ in 95% of repeated samples.
- In general, a $(\alpha \times 100)\%$ confidence interval for μ_Y is an interval that contains the true value of $\bar{Y}$ in $(\alpha \times 100)\%$ of repeated samples.

Confidence intervals, ctd.

Digression:

- What is random here? The values of $Y_1, \ldots, Y_n$ and thus any functions of them - including the confidence interval.

Confidence intervals, ctd.

Digression:

- What is random here? The values of $Y_1, \ldots, Y_n$ and thus any functions of them - including the confidence interval.
- The confidence interval it will differ from one sample to the next.

Confidence intervals, ctd.

Digression:

- What is random here? The values of $Y_1, \ldots, Y_n$ and thus any functions of them - including the confidence interval.
- The confidence interval it will differ from one sample to the next.
- The population parameter, μ_Y, is not random, we just don't know it.

Confidence intervals, ctd.

A **95% confidence interval** has the followig form:

$$\left\{\mu_Y : \left|\frac{Y-\mu_Y}{s_Y/\sqrt{n}}\right| > 1.96\right\} = \left\{\mu_Y \in \left(\bar{Y} - 1.96 \times \frac{s_Y}{\sqrt{n}}, \bar{Y} + 1.96 \times \frac{s_Y}{\sqrt{n}},\right)\right\}$$

This confidence interval relies on the large-n results that $\bar{Y}$ is **approximately normally** distributed and $\sigma_Y \xrightarrow{p} \sigma_Y$.

Confidence intervals, ctd.

A **90% confidence interval** has the followig form:

$$\left\{\mu_Y : \left|\frac{Y - \mu_Y}{s_Y/\sqrt{n}}\right| > 1.64\right\} = \left\{\mu_Y \in \left(\bar{Y} - 1.64 \times \frac{s_Y}{\sqrt{n}}, \bar{Y} + 1.64 \times \frac{s_Y}{\sqrt{n}},\right)\right\}$$

Example: Italian wages, ctd.

Sample information

$$\overline{Wage} = 1307, \quad s_W = 523$$

95% confidence interval

The 95% confidence interval for the (population) monthly wage is

$$\left[\overline{Wage} - 1.96 \times \frac{s_W}{\sqrt{n}}, \overline{Wage} + 1.96 \times \frac{s_W}{\sqrt{n}}\right]$$

Example: Italian wages, ctd.

Sample information

$$\overline{Wage} = 1307, \quad s_W = 523$$

95% confidence interval

The 95% confidence interval for the (population) monthly wage is

$$\left[\overline{Wage} - 1.96 \times \frac{s_W}{\sqrt{n}}, \overline{Wage} + 1.96 \times \frac{s_W}{\sqrt{n}}\right]$$
$$= \left[1307 - 1.96 \times \frac{523}{\sqrt{26127}}, 1307 + 1.96 \times \frac{523}{\sqrt{26127}}\right]$$

Example: Italian wages, ctd.

Sample information

$$\overline{Wage} = 1307, \quad s_W = 523$$

95% confidence interval

The 95% confidence interval for the (population) monthly wage is

$$\left[\overline{Wage} - 1.96 \times \frac{s_W}{\sqrt{n}}, \overline{Wage} + 1.96 \times \frac{s_W}{\sqrt{n}}\right]$$
$$= \left[1307 - 1.96 \times \frac{523}{\sqrt{26127}}, 1307 + 1.96 \times \frac{523}{\sqrt{26127}}\right]$$
$$\approx [1300.4, 1313.1]$$

Example: Italian wages, ctd.

Sample information

$$\overline{Wage} = 1307, \quad s_W = 523$$

90% confidence interval

The 90% confidence interval for the (population) monthly wage is

do it as an exercise

Summary:

From the two assumptions of:

1. simple random sampling of a population, that is, $\{Y_i, i = 1, \ldots, n\}$ are i.i.d.
2. $0 < E(Y^2) < \infty$

we developed, for large samples (large n):

- Theory of estimation (sampling distribution of)
- Theory of hypothesis testing (large-n distribution of t-statistic and computation of the p-value)
- Theory of confidence intervals (constructed by inverting test statistic)

Are assumptions 1. & 2. plausible in practice? Yes