# Statistica inferenziale- III

STATISTICA NUMERICA

A.Y. 2022-2023

# Outline

## 1)Intervalli di confidenza

STATISTICA NUMERICA, CAP. 6.2.5

## 2) Test di ipotesi

STATISTICA NUMERICA CAP. 6.3

# 1)Intervalli di confidenza

STATISTICA NUMERICA, CAP. 6.2.5

# Intervalli di confidenza per la stima della media campionaria: popolazione normale

Siano $X_1$, $X_{2,\dots}$ $X_n$ SRS(n) da una distribuzione normale con (media=mu, sd=sigma).
Consideriamo come statistica campionaria la media $\overline{X}$.
Si commette rispetto alla media esatta un errore non noto.

Fissato $\alpha$ in [0,1] e $z_{\alpha/2}$ il quantile di indice $\alpha/2$ della distribuzione normale standard,

l'intervallo:

$$\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

È detto **intervallo di confidenza** $100(1-\alpha)\%\, di\, \mu.$
$\alpha$ è il **livello di confidenza**.

# Intervalli di confidenza per la stima della media campionaria: popolazione normale

Possiamo anche dire che la probabilita' che il valore esatto. $\mu$ Sia nell' intervallo di confidenza e' del $100(1-\alpha)\%$.

L'intervallo è casuale perché dipende dalla variabile aleatoria

Possiamo fare le seguenti considerazioni.

- Per un livello di confidenza fissato $1-\alpha$, se $n$ aumenta, l'intervallo di confidenza diminuisce.
- Per $n$ fissato, se $1-\alpha$ aumenta, l'intervallo di confidenza aumenta.

Esempio 6.11 di Statistica Numerica.

# Intervalli di confidenza per la stima della media campionaria: popolazione normale

**Se la deviazione standard della popolazione non e' nota?**

Si utilizza la deviazione standard campionaria:

$$\bar{X} \pm z_{\alpha/2}\frac{S}{\sqrt{n}}$$

$\bar{X}$

# Intervalli di confidenza per la stima della media campionaria: popolazione normale

Cosa posso dire di questa approssimazione?

- SE n e' grande S tende alla deviazione standard esatta
- Se n e' piccolo, sostituisco il valore del quantile della distribuzione normale con quello della distribuzione t di student (df=1):

$$\bar{X} \pm t_{\alpha/2}(df = 1)\frac{S}{\sqrt{n}}.$$

$\bar{X}$

# Intervalli di confidenza per la stima della media campionaria: popolazione non normale

Siano $X_1$, $X_{2,\dots}$ $X_n$ SRS(n) da una distribuzione **qualunque** con (media=mu, sd=sigma).
Consideriamo come statistica campionaria la media $\overline{X}$.
Per il Teorema del Limite Centrale, se n è sufficientemente grande la variabile aleatoria media campionaria si comporta come nel caso di campioni estratti da distribuzione normale.
Quindi l'intervallo di confidenza si calcola sempre come:

I

$$\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

# Intervalli di confidenza per la stima della media campionaria: popolazione non normale

Purtroppo è molto difficile nel caso di distribuzioni **non** normali avere la deviazione standard esatta quindi la sostituisco con quella campionaria S.

I

$$\bar{X} \pm z_{\alpha/2}\frac{S}{\sqrt{n}}$$

# Funzioni python

Funzione Python per calcolare il quantile di ordine p della distribuzione normale standard:

```
from scipy import stats
Q=stats.norm.ppf(p)
```

Nel caso di distribuzione normale con media=m e deviazione standard=s:

```
Q=stats.norm.ppf(p, loc=m,scale=s)
Q=stats.norm.ppf(0.95, loc=0,scale=1.5)
```

Nel caso di distribuzione t di student con gradi di libertà  n:

```
Q=stats.t.ppf(p, df)
Q=stats.t.ppf(0.99, 4)
```

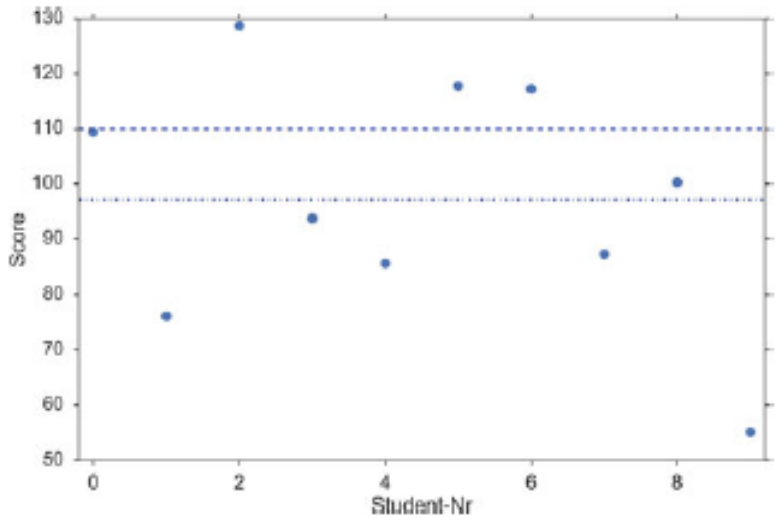
# 2) Test di ipotesi

STATISTICA NUMERICA CAP. 6.3

# Test di ipotesi

Supponiamo che tu sia il direttore di una scuola. Se gli studenti ottengono un punteggio di 110 nell'esame finale, visto che la media nazionale è di 100, ottieni un incentivo.
Se il voto è significativamente minore non lo ottieni (devi assumere piu insegnanti), ma anche se è significativamente maggiore non lo ottieni (hai speso troppo e quindi devi licenziare delle insegnanti).

Come fare a decidere?



# Test di ipotesi: procedura generale

Step:

1. Considero un SRS(n) *(il campione dei voti considerato)*
2. Formulo una ipotesi nulla $H_0$ *(La media dei voti è 110)*
3. Formulo una ipotesi alternativa $H_1$ *(la media dei voti è diversa da 110)*
4. Calcolo una statistica *(la media campionaria)*
5. Confronto il valore della statistica calcolata con quello della ipotesi nulla e calcolo un valore detto **p-value**
6. Interpreto il p-value e decido se l'ipotesi nulla è da rigettare oppure no

# Test di ipotesi: interpretazione del p-value

Il p-value è un valore *p* nell'intervallo [0,1].

Se $p<0.05$ possiamo interpretare così:

SE l'ipotesi nulla è vera, la probabilità di trovare un valore della media campionaria piu estremo (maggiore o minore) di quello osservato è del 5%.

N.B. Non significa che l'ipotesi nulla sia falsa e nemmeno che l'ipotesi alternativa sia vera!

# Test di ipotesi: interpretazione del p-value

In pratica però di solito il p-value si utilizza così per decidete il test di ipotesi:

**Se p<0.05 si rigetta l'ipotesi nulla a favore di quella alternativa, per una *differenza statisticamente significativa del valore osservato rispetto all'ipotesi nulla*.**

Ricordiamo però che il p-value rappresenta solo una verosimiglianza che l'ipotesi nulla sia vera , niente di piu!

# Test di ipotesi: tipi di errore

**Errori del primo tipo (Type I errors):**
Sono gli errori che si commettono quando l'ipotesi nulla è vera anche se la statistica misurata Differisce significativamente.
Sono detti *producer risk errors* perché si rigettano dati che in realtà rispondono ai requisiti richiesti.

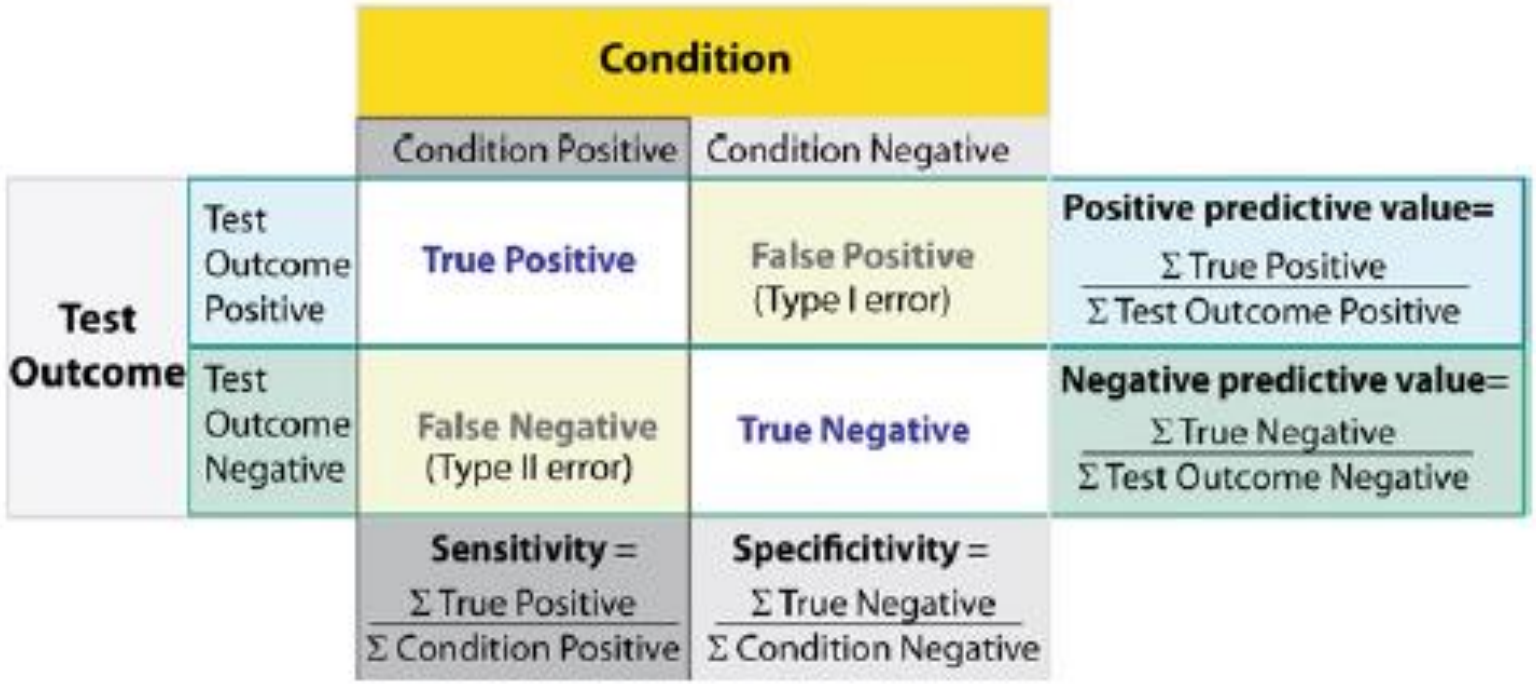
**Errori del secondo tipo (Type II errors):**
Sono gli errori che si commettono quando si accetta l'ipotesi nulla perché la statistica non differisce da essa significativamente anche *se l'ipotesi nulla è falsa*.
Sono detti *consumer risk errors* perchè si accettano misure anche se non conformi ai requisiti richiesti.

# Test di ipotesi: sensitivtà e specificità

Sopponiamo di fare un test di ipotesi utilizzanpo una grandezza che identifica l'esistenza o meno di un tumore. Quindi:
Ipotesi nulla: il paziente ha il tumore
Ipotesi alternativa: il paziente NON ha il tumore.

**Sensitività**: anche detta *potenza*. Proporzione di **positivi** correttamente identificati da un test.
**Specificità**: Proporzione di **negativi** correttamente identificati da un test.
**Positive Predicted Value (PPV):** proporzione di pazienti con test **positivo** che sono stati correttamente diagnosticati
**Negative Predicted Value (NPV):** proporzione di pazienti con test **negativo** che sono stati correttamente diagnosticati

# Test di ipotesi: sensitivtà e specificità



# Test di ipotesi: sensitivtà e specificità

Mentre la sensitività e la specificità caratterizzano un test, non indicano quale sensitività e specificità caratterizza un test, non indicao quale porzione di pazienti con test “anormale” sono veramente “non normali”.  Questa informazione è data dai valori PPV e NPV.

Tuttavia, questi valori da soli non sono sufficienti per una valutazione affidabile.

Per rispondere a questa domanda:
*Se un paziente ha un test positivo, qual è la probabilità che sia veramente ammalato?*

Dobbiamo considerare altri due fattori: la **prevalenza** e **l’incidenza**.

# Test di ipotesi: prevalenza e incidenza

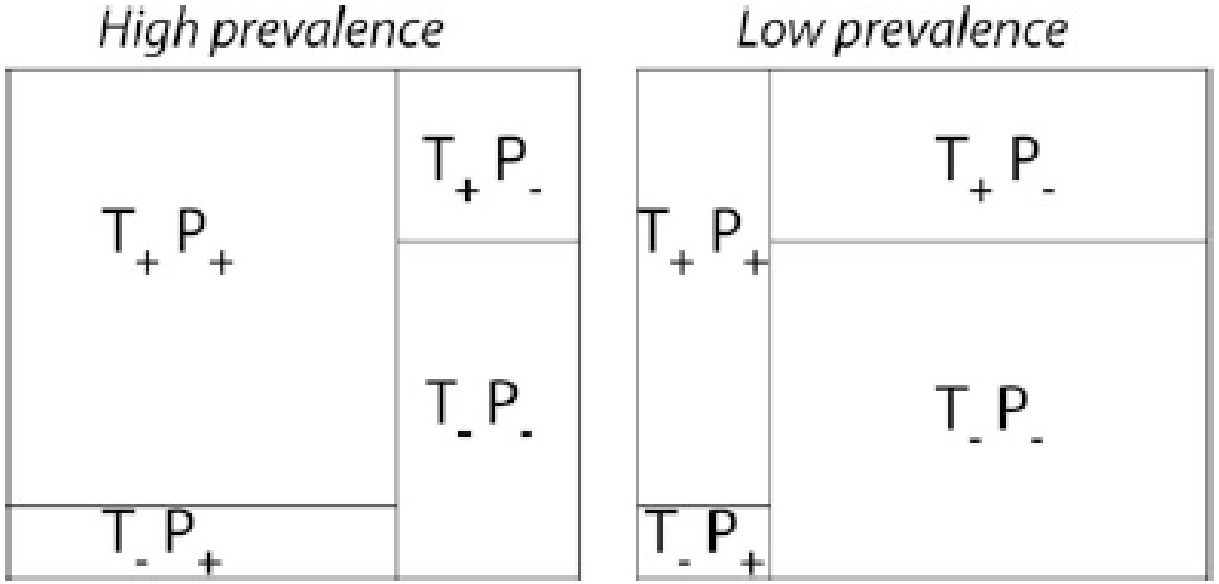



**Fig. 7.8** Effect of prevalence on PPV and NPV. "T" stands for "test," and "P" for "patient." (For comparison with below: T+P+ = TP, T−P− = TN, T+P− = FP, and T−P+ = FN.)

*Prevalenza*: quante persone su 100000(numero fissato) è ammalata

*incidenza*: il numero di nuovi casi diagnosticati su 100000 persone.

# Test di ipotesi: risultati di un esempio



| | | Condition | | |
|---|---|---|---|---|
| | | Condition Positive | Condition Negative | |
| **Test Outcome** | Test Outcome Positive | **True Positive** (TP) = 25 | False Positive (FP) = 175 | **Positive predictive value=** = TP / (TP+FP) = 25 / (25+175) = 12.5% |
| | Test Outcome Negative | False Negative (FN) = 10 | **True Negative** (TN) = 2000 | **Negative predictive value=** = TN / (FN+TN) = 2000 / (10+2000) = 99.5% |
| | | **Sensitivity =** = TP / (TP+FN) = 25 / (25+10) = 71% | **Specificitivity =** = TN / (FP+TN) = 2000 / (175+2000) = 92% | |

# Test di ipotesi: risultati di un esempio

- False positive rate ($\alpha$) = type I error = $1 - specificity = \frac{FP}{FP+TN} = \frac{175}{175+2000} = 8\,\%$
- False negative rate ($\beta$) = type II error = $1 - sensitivity = \frac{FN}{TP+FN} = \frac{10}{25+10} = 29\,\%$
- Power = sensitivity = $1 - \beta$
- $positive\ likelihood\ ratio = \frac{sensitivity}{1-specificity} = \frac{71\,\%}{1-92\,\%} = 8.9$
- $negative\ likelihood\ ratio = \frac{1-sensitivity}{specificity} = \frac{1-71\,\%}{92\,\%} = 0.32$

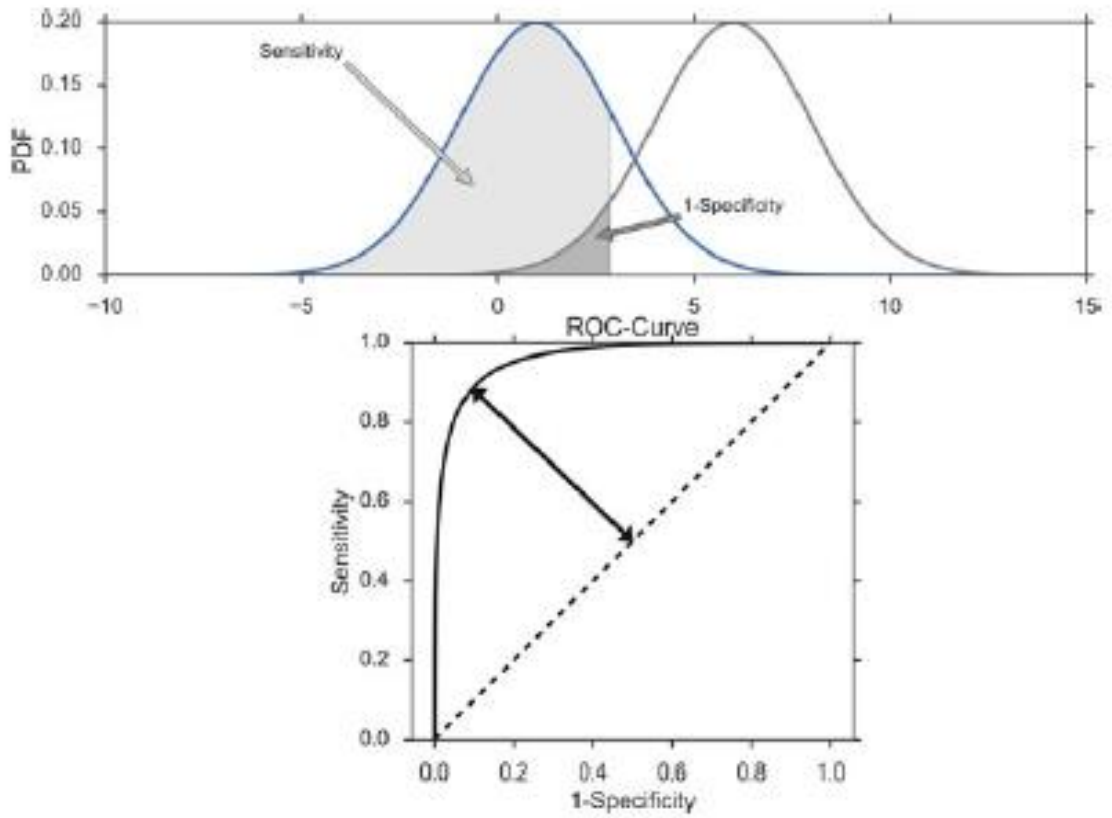



**Fig. 7.10** *Top*: Probability density functions for two distributions. *Bottom*: Corresponding *ROC-curve*, with the largest distance between the diagonal and the curve indicated by the *arrow*

# Curva di ROC (Receiving Operating Characteristic)

Mostra la relazione fra True Positive Rate (sull'asse y) e False Positive rate (sull'asse x)

# Funzioni Python

- Test di ipotesi sulla media di un campione di dati distribuiti in modo normale: t-test

t, pVal = stats.ttest_1samp scipy.stats.ttest_1samp(*a*, *popmean*, *axis=0*, *nan_policy='propagate'*, *alternative='two-sided'*, ***, *keepdims=False*)

scipy.stats.ttest_1samp — SciPy v1.10.1 Manual

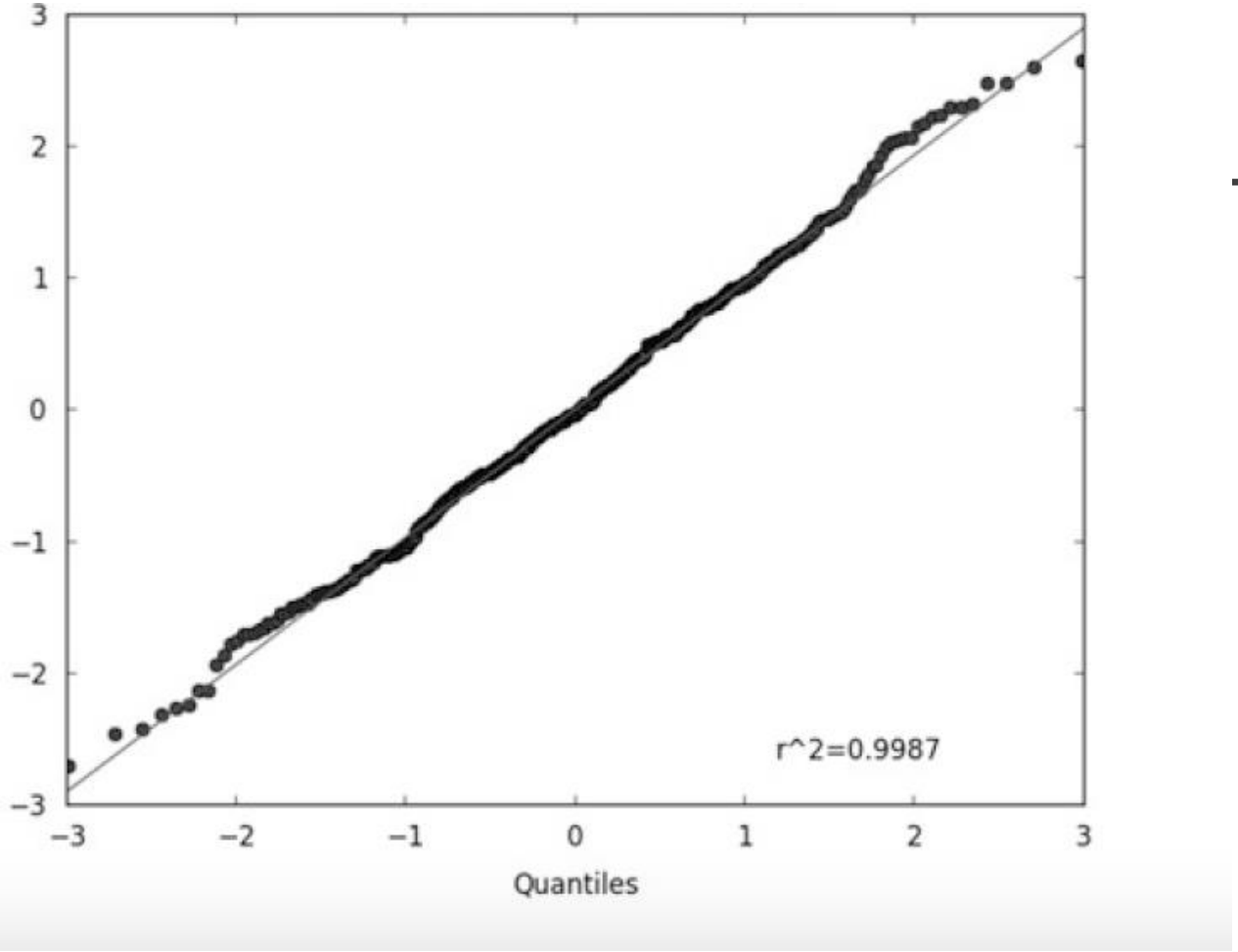
# Funzioni Python

- Test di ipotesi per decidere se due campioni x e y provengono dalla stessa distribuzione:

**Res=scipy.stats.wilcoxon(*x, y=None, zero_method='wilcox', correction=False, alternative='two-sided', method='auto', *, axis=0, nan_policy='propagate', keepdims=False*)**

Res.statistic
Res.pvalue

scipy.stats.wilcoxon — SciPy v1.10.1 Manual




# Test di  di normalità

QQ-plot: Il quantile del data set considerato e' plottato rispetto al quantile di una

distribuzione (normale in questo caso) di riferimento.

Se le due distribuzioni sono simili, I punti devono stare molto vicini alla retta.

# Test di ipotesi di normalità

Ci sono diversi test di ipotesi di normalita' basati sul confronto della distribuzione stimata dei dati rispetto alla distribuzione normale.

Uno dei piu famosi e' il test di Shapiro-Wilk, che si basa sulla matrice di covarianza delle statitsiche ordinate delle osservazioni e puo' essere utilizzato anche con un numero ridotto (<+50) di osservazioni.

H_0: residui normali
H_a: residui non normali.