

# Formulario *non ufficiale* del corso di Statistica

Questo documento contiene una raccolta sintetica delle formule più utilizzate negli esercizi di Statistica. *Non si tratta di materiale ufficiale del corso*: è stato preparato indipendentemente da uno studente e non verificato dai docenti e, pertanto, potrebbe contenere errori e/o omissioni (specialmente nelle sezioni dedicate all'inferenza statistica).

## 1 Campionamento

**Schemi di campionamento** Estrazione di  $n$  palline da un'urna che ne contiene  $M$ .

	Ordinati	Non ordinati
Con reinserimento	$M^n$	$\binom{M+n-1}{n}$
Senza reinserimento	$\frac{M!}{(M-n)!}$	$\binom{M}{n}$

**Campionamento duale** Assegnazione di  $n$  oggetti a  $M$  celle *distinte*.

	Distinti	Non distinti
Senza esclusione	$M^n$	$\binom{M+n-1}{n}$
Con esclusione	$\frac{M!}{(M-n)!}$	$\binom{M}{n}$

**Corrispondenze** Assegnazione di  $n$  oggetti distinti a  $n$  celle distinte.  $C(r, n)$  permutazioni con  $r$  oggetti capitano al posto giusto.

$$\frac{C(0, n)}{n!} = \sum_{k=0}^n (-1)^k \frac{1}{k!} = \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \frac{1}{5!} + \dots$$

$n$	1	2	3	4	5	6	7	8
$C(0, n)$	0	1	2	9	44	265	1854	14833
$n!$	1	2	6	24	120	720	5040	40320

$$C(r, n) = \binom{n}{r} C(0, n-r)$$

**Principio di inclusione ed esclusione** Probabilità dell'unione di  $n$  eventi. Si sottraggono le probabilità delle intersezioni di ordine pari e si sommano quelle delle intersezioni di ordine dispari.

$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i) - \sum_{i < k} P(A_i \cap A_k) + \dots + (-1)^{n-1} P(A_1 \cap A_2 \cap \dots \cap A_n)$$

## 2 Probabilità condizionata

**Probabilità condizionata** Probabilità che si verifichi  $A$  sapendo che si è verificato  $B$ .

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

**Teorema delle probabilità totali** Dati  $C_1, C_2, \dots, C_n$  partizione di  $\Omega$ ,

$$P(E) = \sum_{i=1}^n P(E|C_i)P(C_i)$$

**Teorema di Bayes** Sapendo che si è verificato  $E$ , probabilità che il suo verificarsi sia dovuto alla causa  $C_i$ .

$$P(C_i|E) = \frac{P(E|C_i)P(C_i)}{P(E)}$$

## 3 Variabili casuali discrete

**Valore atteso**  $E(X) = \mu$

**Linearità**  $E(aX + bY) = aE(X) + bE(Y)$

**Varianza**  $Var(X) = E[(X - \mu)^2] = EX^2 - \mu^2 = \sigma^2$

**Invarianza alla traslazione**  $Var(a + X) = Var(X)$

**Cambiamento di scala**  $Var(bX) = b^2 Var(X)$

**Problema diretto** Fare attenzione, l'estremo superiore è *incluso*, quello inferiore è *escluso*.

$$P(a < X \leq b) = F(b) - F(a)$$

### 3.1 Uniforme discreta $U(k)$

$$P(X = x) = \frac{1}{k}, \quad x = 1, \dots, k$$
$$E(X) = \frac{k(k+1)}{2} \quad Var(X) = \frac{(k^2-1)}{12}$$

### 3.2 Binomiale $Bin(n, p)$

$n$  prove ripetute *indipendenti*. Ogni prova ha esito dicotomico (successo/insuccesso) con probabilità di successo  $p$  *omogenea*.

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \dots, n$$
$$E(X) = np \quad Var(X) = np(1-p)$$

### 3.3 Ipergeometrica $IG(n, M, K)$

Contatore di successi (palline rosse) in  $n$  estrazioni *senza reinserimento* da una popolazione di  $M$  palline di cui  $K$  rosse.

$$P(X = x) = \frac{\binom{K}{x} \binom{M-K}{n-x}}{\binom{M}{n}}, \quad x = 0, 1, \dots, n, \quad n < K, \quad n < M - K$$

$$E(X) = np, \quad p = \frac{K}{M} \quad Var(X) = np(1-p) \left(1 - \frac{n-1}{M-1}\right)$$

Approssimazione:  $n \ll M \Rightarrow IG(n, M, K) \cong Bin(n, \frac{K}{M})$

### 3.4 Poisson $Poiss(\lambda t)$

Numero di arrivi in un intervallo di tempo lungo  $t$ .

$$P(X = x) = \frac{(\lambda t)^x}{x!} e^{-\lambda t}, \quad x = 0, 1, 2, \dots$$
$$E(X) = \lambda t \quad Var(X) = \lambda t$$

**Formula di ricorrenza**  $P(X = x) = \frac{\lambda t}{x} P(X = x-1)$   
Approssimazione:  $n \rightarrow \infty, p \rightarrow 0 \Rightarrow Bin(n, p) \cong Poiss(np)$

### 3.5 Geometrica $G(p)$

Prove ripetute indipendenti con probabilità di successo  $p$  omogenea. Conta il numero di tentativi da effettuare per ottenere il primo successo.

$$P(X = x) = (1-p)^{x-1} p, \quad x = 1, 2, 3, \dots$$

$$E(X) = \frac{1}{p} \quad Var(X) = \frac{1-p}{p^2}$$

## 4 Variabili casuali continue

**Valore atteso** È il momento di ordine 1.

$$E(X) = \int_{-\infty}^{+\infty} x f(x) dx = \mu$$

**Varianza** È il momento centrato di ordine 2.

$$Var(X) = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx = \sigma^2$$

**Kurtosi**

$$K = \frac{\bar{\mu}_4}{\sigma^4}$$

**Problema diretto** Non importa se l'estremo è incluso o escluso.

**Problema inverso** Quantile di ordine  $p$ : lascia alla sua sinistra un'area pari a  $p$ .

$$\tilde{x}_p = F^{-1}(p)$$

### 4.1 Uniforme o Rettangolare $U(a, b)$

$$f(x) = \frac{1}{b-a}, \quad a \leq x \leq b$$

$$E(X) = \frac{a+b}{2} \quad Var(X) = \frac{(b-a)^2}{12}$$

## 4.2 Esponenziale $Exp(\lambda)$

Istante del primo arrivo, tempo tra un arrivo e l'altro.

$$f(x) = \lambda e^{-\lambda x}, \quad x > 0, \quad \lambda > 0 \quad F(x) = 1 - e^{-\lambda x}$$

$$E(X) = \frac{1}{\lambda} \quad Var(X) = \frac{1}{\lambda^2}$$

## 4.3 Gamma $Gamma(r, \lambda)$

Istante del  $r$ -esimo arrivo, tempo tra arrivi distanti  $r$ .

$$f(x) = \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x}, \quad x > 0, \quad \lambda > 0, \quad r = 1, 2, 3, \dots$$

$$E(X) = \frac{r}{\lambda} \quad Var(X) = \frac{r}{\lambda^2}$$

### 4.3.1 Funzione Gamma

Funzione Gamma con  $r$  gradi di libertà

$$\Gamma(r) = \int_0^{+\infty} t^{r-1} e^{-t} dt, \quad r > 0$$

**Formula di ricorrenza**  $\Gamma(r) = (r-1)\Gamma(r-1)$

Per  $n$  intero,  $\Gamma(n) = (n-1)!$

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$$

## 4.4 Normale standard $N(0, 1)$

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

$$E(Z) = 0 \quad Var(Z) = 1 \quad K = E(Z^4) = 3$$

**Intervalli caratteristici**

$$P(-1 < Z < 1) \cong 0.68$$

$$P(-2 < Z < 2) \cong 0.945$$

$$P(-3 < Z < 3) \cong 0.997$$

## 4.5 Normale (generica) $N(\mu, \sigma^2)$

$$E(X) = \mu \quad Var(X) = \sigma^2$$

**Standardizzazione** Tutti i calcoli si fanno sulla normale standard.

$$Z = \frac{X - \mu}{\sigma} \quad X = \mu + \sigma Z$$

## 4.6 Chi quadro $\chi_n^2$

Date  $Z_1, Z_2, \dots, Z_n \text{ iid } N(0, 1)$ , si ha:

$$\sum_{i=1}^n Z_i^2 \sim \chi_n^2$$

$$E(X) = n \quad Var(X) = 2n$$

**Caso particolare della Gamma** La  $\chi_n^2$  è una Gamma con  $\lambda = 1/2$  e  $r = n/2$ .

$$\chi_n^2 \equiv Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$$

## 4.7 T di Student $t_n$

Date  $Z \sim N(0, 1)$  e  $X \sim \chi_n^2$ , si ha:

$$\frac{Z}{\sqrt{\frac{X}{n}}} \sim t_n$$

# 5 Stimatori e loro proprietà

La distribuzione  $f(x; \theta)$  della popolazione è nota a meno di uno o più parametri  $\theta$ .  $T_n$  è una v.c. utilizzata per stimare il parametro  $\theta$ . Spesso le proprietà di  $T_n$  dipendono dalla dimensione  $n$  del campione.

## 5.1 Correttezza

$$E[T_n] = \theta$$

### 5.1.1 Distorsione o bias

Esprime l'errore sistematico della stima.

$$b[T_n] = E[T_n] - \theta$$

### 5.1.2 Correttezza asintotica

$$E[T_n] \rightarrow \theta, \quad n \rightarrow +\infty$$

## 5.2 Errore quadratico medio

$$MSE[T_n] = E[(T_n - \theta)^2] = Var[T_n] + b[T_n]^2$$

## 5.3 Consistenza

L'incertezza sulla stima scompare all'aumentare di  $n$ .

$$T_n \rightarrow \theta, \quad n \rightarrow +\infty$$

### 5.3.1 Condizioni sufficienti di consistenza

1.  $E[T_n] \rightarrow \theta$ , per  $n \rightarrow +\infty$
2.  $Var[T_n] \rightarrow 0$ , per  $n \rightarrow +\infty$

### 5.3.2 Consistenza in media quadratica

$$MSE[T_n] \rightarrow 0, \quad n \rightarrow +\infty$$

## 5.4 Efficienza

$$MSE[T_1] < MSE[T_2] \implies T_1 \text{ è più efficiente di } T_2$$

### 5.4.1 Soglia di Cramér-Rao

Stabilisce la varianza minima per uno stimatore corretto.  $\tau(\theta)$  indica il parametro da stimare: se è semplicemente  $\theta$ , il numeratore vale sempre 1.

$$Var[T_n] = \frac{[\tau'(\theta)^2]}{nE\left\{\left[\frac{\partial}{\partial\theta} \log f(x;\theta)\right]^2\right\}}$$

## 6 Stima di massima verosimiglianza

Osservato un certo campione, il valore più plausibile del parametro  $\theta$  è dato dalla stima di massima verosimiglianza.

### 6.1 Funzione di verosimiglianza

È la distribuzione congiunta espressa in funzione di  $\theta$  (il campione  $x_1, x_2, \dots, x_n$  si considera fissato).

$$L(\theta) = f(x_1, x_2, \dots, x_n; \theta) = \prod_{i=1}^n f(x_i; \theta)$$

### 6.2 Logaritmo

Il logaritmo semplifica la ricerca di minimi e massimi, senza cambiare l'andamento della funzione (è un'operazione monotona).

$$\log L(\theta)$$

### 6.3 Equazione di verosimiglianza

Si cerca il valore  $\hat{\theta}$  che massimizza la funzione.

$$\frac{\partial \log L(\theta)}{\partial \theta} = 0$$

### 6.4 Verifica del vincolo

Si controlla che il punto trovato sia un massimo (e non un minimo).

$$\frac{\partial^2 \log L(\theta)}{\partial \theta^2} \Big|_{\theta=\hat{\theta}} < 0$$

## 7 Intervalli di confidenza

Esprimono l'incertezza sulla stima. Il parametro  $\theta$  appartiene all'intervallo con una confidenza pari a  $1 - \alpha$ .

Gli intervalli unilaterali si costruiscono considerando solo un estremo (ed il valore critico opportuno:  $z_\alpha$  invece di  $z_{\frac{\alpha}{2}}$ ,  $\chi^2_{1-\alpha, n-1}$  invece di  $\chi^2_{1-\frac{\alpha}{2}, n-1}$ ).

Il valore critico di ordine  $\alpha$  è quel valore  $x_\alpha$  tale che  $\Pr(X > x_\alpha) = \alpha$ .

## 7.1 Media $\mu$ di una popolazione Normale

Siano  $X_1, X_2, \dots, X_n$  iid  $\mathcal{N}(\mu, \sigma^2)$ , da cui si calcolano le statistiche

**Media campionaria** 
$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

**Varianza campionaria** 
$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n-1} \left( \sum_{i=1}^n X_i^2 - n\bar{X}^2 \right)$$

### 7.1.1 Varianza $\sigma^2$ nota

**Distribuzione della statistica** 
$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

Per  $n$  sufficientemente grande ( $n > 30$ ), questa formula vale approssimativamente anche per popolazioni non Normali.

**Intervallo di confidenza** 
$$\mu \in \left[ \bar{x} \pm z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right]$$

**Numerosità campionaria** Il minimo  $n$  che soddisfi  $\Pr(|\bar{X} - \mu| < \varepsilon) = 1 - \alpha$  è dato da

$$n \geq \left( \frac{z_{\frac{\alpha}{2}} \cdot \sigma}{\varepsilon} \right)^2$$

### 7.1.2 Varianza ignota

La varianza campionaria  $S^2$  fornisce una stima non distorta per  $\sigma^2$ .

**Distribuzione della statistica** 
$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

**Intervallo di confidenza** 
$$\mu \in \left[ \bar{x} \pm t_{\frac{\alpha}{2}, n-1} \cdot \frac{s}{\sqrt{n}} \right]$$

**Numerosità campionaria** Il minimo  $n$  che soddisfi  $\Pr(|\bar{X} - \mu| < \varepsilon) = 1 - \alpha$  è dato da

$$n \geq \left( \frac{t_{\frac{\alpha}{2}, n-1} \cdot s}{\varepsilon} \right)^2$$

## 7.2 Varianza $\sigma^2$ di una popolazione Normale

A differenza degli altri, si tratta di un intervallo di confidenza asimmetrico.

**Distribuzione della statistica** 
$$(n-1) \frac{S^2}{\sigma^2} \sim \chi_{n-1}^2$$

Per  $n$  sufficientemente grande ( $n > 100$ ), questa formula vale approssimativamente anche per popolazioni non Normali.

**Intervallo di confidenza** 
$$\sigma^2 \in \left[ \frac{(n-1)s^2}{\chi_{\frac{\alpha}{2}, n-1}^2}, \frac{(n-1)s^2}{\chi_{1-\frac{\alpha}{2}, n-1}^2} \right]$$

## 7.3 Proporzione $\pi$ di una popolazione Bernoulliana

Siano  $X_1, X_2, \dots, X_n$  iid  $\text{Bin}(1, \pi)$  e  $X$  il numero di successi nel campione, da cui si calcola la statistica

**Percentuale campionaria** 
$$\hat{p} = \frac{X}{n}$$

**Distribuzione della statistica** 
$$\hat{p} \sim \frac{1}{n} \text{Bin}(n, \pi) \approx \mathcal{N}\left(\pi, \frac{\pi(1-\pi)}{n}\right)$$

L'approssimazione con la Normale vale solo per  $n$  sufficientemente grande ( $n > 30$ ).

**Intervallo di confidenza** 
$$\pi \in \left[ \hat{p} \pm z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right]$$

**Numerosità campionaria** Il minimo  $n$  che soddisfi  $\Pr(|\hat{p} - \pi| < \varepsilon) = 1 - \alpha$  è dato da

$$n \geq \left( \frac{z_{\frac{\alpha}{2}}}{\varepsilon} \right)^2 \cdot \hat{p}(1-\hat{p}), \quad \hat{p}(1-\hat{p}) \leq \frac{1}{4}$$

## 7.4 Differenza tra le medie $\mu_1$ e $\mu_2$ di due popolazioni Normali

Siano  $X_1, X_2, \dots, X_n$  iid  $\mathcal{N}(\mu_x, \sigma_x^2)$  e  $Y_1, Y_2, \dots, Y_m$  iid  $\mathcal{N}(\mu_y, \sigma_y^2)$ , indipendenti tra loro, su cui si calcolano le medie campionarie  $\bar{X}$  e  $\bar{Y}$ .

### 7.4.1 Varianze $\sigma_x^2$ e $\sigma_y^2$ note

**Distribuzione della statistica** 
$$\frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}} \sim \mathcal{N}(0, 1)$$

**Intervallo di confidenza** 
$$(\mu_x - \mu_y) \in \left[ (\bar{X} - \bar{Y}) \pm z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}} \right]$$

### 7.4.2 Varianze ignote ma uguali ( $\sigma_x^2 = \sigma_y^2$ )

Si può effettuare la stima solo nel caso *omoschedastico*, cioè  $\sigma_x^2 = \sigma_y^2$ . Dopo aver calcolato le varianze campionarie  $S_x^2$  e  $S_y^2$ , si stima la varianza di  $\bar{X} - \bar{Y}$  con

$$S_p = \frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}$$

**Distribuzione della statistica** 
$$\frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{S_p \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2}$$

**Intervallo di confidenza** 
$$(\mu_x - \mu_y) \in \left[ (\bar{X} - \bar{Y}) \pm t_{\frac{\alpha}{2}, n+m-2} \cdot S_p \sqrt{\frac{1}{n} + \frac{1}{m}} \right]$$

## 8 Test di ipotesi

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta \neq \theta_0 \quad (\text{oppure } H_1 : \theta \gtrless \theta_0)$$

Si vuole verificare se il parametro  $\theta$  è *significativamente* diverso dal valore  $\theta_0$ . I possibili risultati sono due: accettazione o rifiuto di  $H_0$ .

**Errore di I tipo**  $\alpha = \Pr(\text{Rifiuto } H_0 \mid H_0 \text{ vera})$ .

Il valore di  $\alpha$  è ben noto, perché si conosce la distribuzione della popolazione sotto  $H_0$ .

**Errore di II tipo**  $\beta = \Pr(\text{Accetto } H_0 \mid H_0 \text{ falsa})$ .

*Non si conosce il valore di  $\beta$  perché dipende dall'ignoto valore vero  $\theta_1$  del parametro.*

**p-value** esprime la credibilità dell'ipotesi nulla in base ai dati osservati.

*Ottenere un p-value  $\hat{\alpha}$  molto piccolo equivale a dimostrare con molta forza la falsità di  $H_0$ .*

Se la significatività nominale  $\alpha$  è fissata, si rifiuta  $H_0$  per ogni p-value osservato  $\hat{\alpha} < \alpha$ .

### 8.1 Media $\mu$ di una popolazione Normale

Test di  $H_0 : \mu = \mu_0$  vs.  $H_1 : \mu > \mu_0$  ad un livello di significatività  $\alpha$ .

#### 8.1.1 Varianza $\sigma^2$ nota

**Statistica sotto  $H_0$**  
$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

**Regione d'accettazione** 
$$\bar{x} \leq \mu_0 + z_\alpha \cdot \frac{\sigma}{\sqrt{n}}$$

**Potenza del test**

$$\pi(\mu_1) = \Pr\left(\bar{X} > \mu_0 + z_\alpha \cdot \frac{\sigma}{\sqrt{n}} \mid \mu = \mu_1\right) = \Phi\left(\sqrt{n}\delta + z_\alpha\right), \quad \delta = \frac{\mu_1 - \mu_0}{\sigma}$$

**P-value per  $Z = z_0$**  
$$\hat{\alpha} = \Pr(\mathcal{N}(0, 1) > z_0)$$

#### 8.1.2 Varianza ignota

**Statistica sotto  $H_0$**  
$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$$

**Regione d'accettazione** 
$$\bar{x} \leq \mu_0 + t_{\alpha, n-1} \cdot \frac{s}{\sqrt{n}}$$

**P-value per  $T = t_0$**  
$$\hat{\alpha} = \Pr(t_{n-1} > t_0)$$

## 8.2 Varianza $\sigma^2$ di una popolazione Normale

Test di  $H_0 : \sigma^2 = \sigma_0^2$  vs.  $H_1 : \sigma^2 > \sigma_0^2$  ad un livello di significatività  $\alpha$ . A differenza degli altri, si tratta di un test asimmetrico.

Statistica sotto  $H_0$  
$$X = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi_{n-1}^2$$

Regione d'accettazione 
$$s^2 \leq \frac{\sigma_0^2}{n-1} \chi_{\alpha, n-1}^2,$$

P-value per  $X = \chi_0^2$  
$$\hat{\alpha} = \Pr(\chi_{n-1}^2 > \chi_0^2)$$

## 8.3 Proporzione $\pi$ di una popolazione Bernoulliana

Test approssimato (valido per  $n$  grande) di  $H_0 : \pi = \pi_0$  vs.  $H_1 : \pi > \pi_0$  ad un livello di significatività  $\alpha$ .

Statistica sotto  $H_0$  
$$Z = \frac{\hat{p} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \approx \mathcal{N}(0, 1)$$

Regione d'accettazione 
$$\hat{p} \leq \pi_0 + z_\alpha \cdot \sqrt{\frac{\pi_0(1-\pi_0)}{n}}$$

## 8.4 Differenza tra le medie $\mu_1$ e $\mu_2$ di due popolazioni Normali

Test di  $H_0 : \mu_x - \mu_y = 0$  vs.  $H_1 : \mu_x - \mu_y > 0$  ad un livello di significatività  $\alpha$ .

### 8.4.1 Varianze $\sigma_x^2$ e $\sigma_y^2$ note

Statistica sotto  $H_0$  
$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}} \sim \mathcal{N}(0, 1)$$

Regione d'accettazione 
$$(\bar{x} - \bar{y}) \leq z_\alpha \cdot \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}$$

### 8.4.2 Varianze ignote ma uguali $\sigma_x^2 = \sigma_y^2$

Statistica sotto  $H_0$  
$$T = \frac{\bar{X} - \bar{Y}}{S_p \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2}$$

Regione d'accettazione 
$$(\bar{x} - \bar{y}) \leq t_{\alpha, n+m-2} \cdot S_p \sqrt{\frac{1}{n} + \frac{1}{m}}$$

## 8.5 Test di omoschedasticità $\sigma_x^2 = \sigma_y^2$ per due popolazioni Normali

Test di  $H_0 : \sigma_x^2 = \sigma_y^2$  vs.  $H_1 : \sigma_x^2 \neq \sigma_y^2$  ad un livello di significatività  $\alpha$ .

Statistica sotto  $H_0$  
$$F = \frac{S_x^2}{S_y^2} \sim F_{n-1, m-1}$$

Regione d'accettazione 
$$F \in [F_{1-\frac{\alpha}{2}, n-1, m-1}, F_{\frac{\alpha}{2}, n-1, m-1}]$$

## 9 Regressione

Modello lineare  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon = \mathbf{x}'\boldsymbol{\beta} + \varepsilon$   
(Modello lineare semplice  $k = 1$ )  $y = \beta_0 + \beta_1 x_1 + \varepsilon$

- $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$  è un errore casuale indipendente da  $\mathbf{x}$ .
- $\hat{y} = \mathbf{x}'\boldsymbol{\beta}$  sono i valori interpolati.

### 9.1 Forma matriciale

Date  $n$  osservazioni, si scrive  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_0 \\ \vdots \\ \varepsilon_k \end{pmatrix}$$

## Matrici utili

$$X'X = \begin{pmatrix} n & \sum x_1 & \sum x_2 & \dots \\ \sum x_1 & \sum x_1^2 & \sum x_1 x_2 & \dots \\ \sum x_2 & \sum x_1 x_2 & \sum x_2^2 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

$$(X'X)^{-1} = (v_{ij}), \quad \text{Cov}(\hat{\beta}) = \sigma_\varepsilon^2 (X'X)^{-1}$$

$$X'y = \begin{pmatrix} \sum y \\ \sum x_1 y \\ \sum x_2 y \\ \vdots \end{pmatrix}$$

## 9.2 Stima ai minimi quadrati (Least Squares)

La migliore stima di  $\beta$  è il valore che minimizza

$$Q(\beta) = (y - X\beta)'(y - X\beta) = \sum (y - \beta_0 - \beta_1 x_1 - \dots - \beta_k x_k)^2$$

Stima LS di  $\beta$

$$\hat{\beta} = (X'X)^{-1} X'y$$

$$(\text{Stima nel caso } k=1) \quad \hat{\beta}_0 = a = \bar{y} - b\bar{x}, \quad \hat{\beta}_1 = b = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{S_{xy}}{SS_x}$$

## 9.3 Scomposizione della varianza e adattamento

- Devianza  $D_x = SS_x = \sum x^2 - n\bar{x}^2$
- Codevarianza  $D_{xy} = S_{xy} = \sum xy - n\bar{x}\bar{y}$

Devianza totale

$$D_{\text{tot}} = SS_y = \sum y^2 - n\bar{y}^2 \sim \sigma_\varepsilon^2 \chi_{n-1}^2$$

Devianza residua

$$D_{\text{res}} = SS_e = \sum (y - \hat{y})^2 = \sum e_t^2 \sim \sigma_\varepsilon^2 \chi_{n-k-1}^2$$

Devianza spiegata

$$D_{\text{sp}} = D_{\text{tot}} - D_{\text{res}}, \sim \sigma_\varepsilon^2 \chi_k^2$$

Varianza residua, stima di  $\sigma_\varepsilon^2$

$$S^2 = \frac{SS_e}{n - k - 1}$$

Coefficiente di determinazione multipla

$$R^2 = 1 - \frac{SS_e}{SS_y}$$

Coefficiente di correlazione multipla

$$R = r(y, \hat{y}) = +\sqrt{R^2}$$

## 9.4 Test e intervalli di confidenza

### 9.4.1 Test $T$ sui coefficienti

La variabile  $x_i$  serve? Verifica di  $H_0 : \beta_i = 0$

Statistica sotto  $H_0$

$$T = \frac{\beta_i}{s\sqrt{v_{i+1,i+1}}} \sim t_{n-k-1}$$

### 9.4.2 Intervallo di confidenza sui coefficienti

Intervallo di confidenza

$$\beta_j \in \left[ \hat{\beta}_j \pm t_{\frac{\alpha}{2}, n-k-1} \cdot s\sqrt{v_{i+1,i+1}} \right]$$

### 9.4.3 Intervallo di confidenza per $E[y|x] = x'\beta$

Intervallo di confidenza

$$x'\beta \in \left[ x'\hat{\beta} \pm t_{\frac{\alpha}{2}, n-k-1} \cdot s\sqrt{x'(X'X)^{-1}x} \right]$$

### 9.4.4 Intervallo di confidenza sulle previsioni

Intervallo di confidenza

$$\hat{y} \in \left[ x'\hat{\beta} \pm t_{\frac{\alpha}{2}, n-k-1} \cdot s\sqrt{1 + x'(X'X)^{-1}x} \right]$$