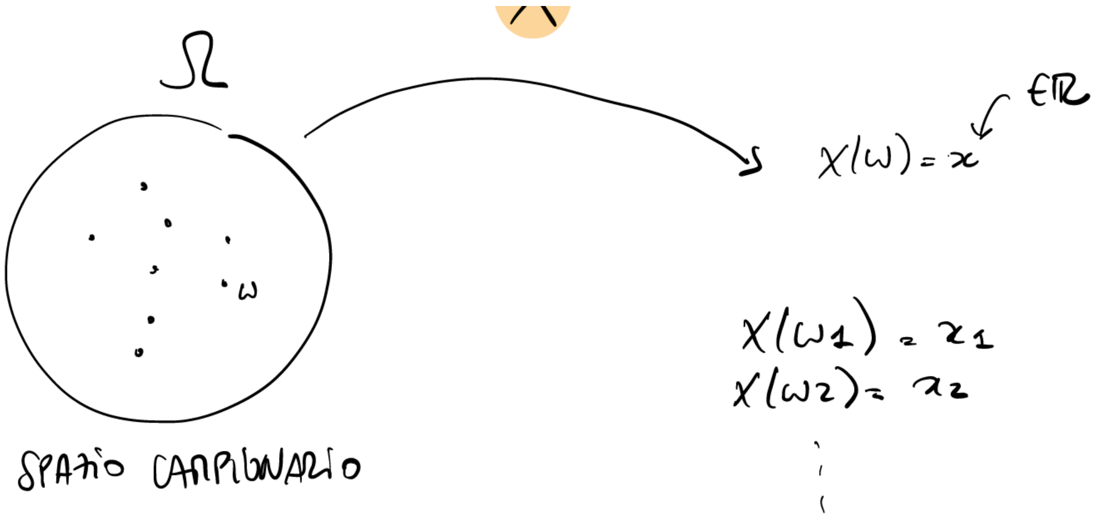
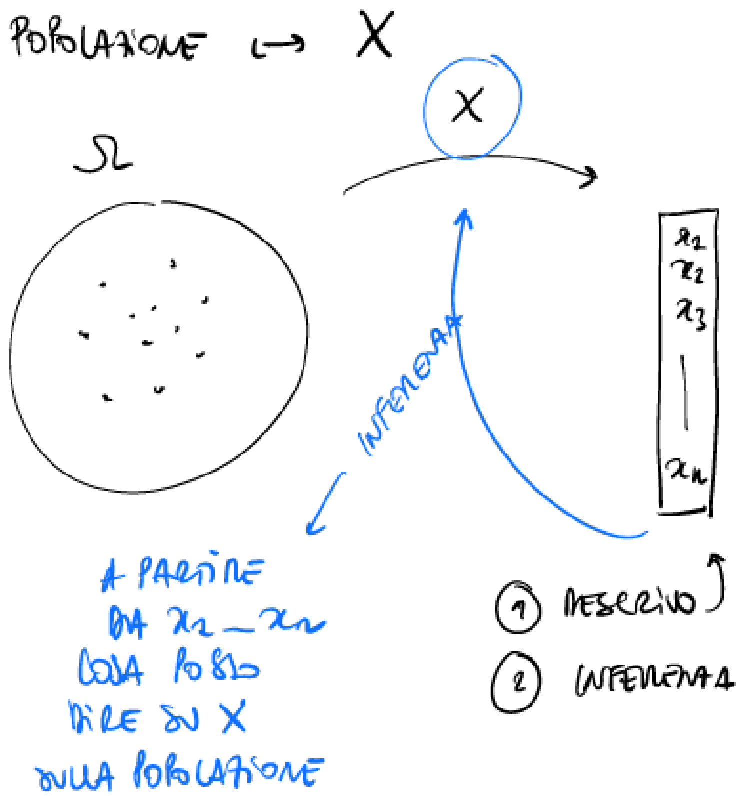
Lezione 1

1. **Statistica descrittiva :** descrivere il campione casuale aka in questo caso la variabile Luigi e' il campione casuale
2. **Statistica inferenziale :** generalizzare all'intera popolazione alla quale appartengono i soggetti

Campione casuale e cosa sono le variabili :



In uno spazio campionario (omega) con all'interno degli omega piccolo
La variabile aleatoria e' il valore che viene associato al punto omega dell'esperimento
Campionare gli elementi di omega tram ite estrazione con reimbussolamento (rimetto il campione nello spazio campionario)



Le variabili del dataset x1,x2,xn sono le n misurazioni di una variabile aleatoria. le descrivo e le inferenze(passare da n osservazioni a una qualche informazione sulla popolazione)

Statistica descrittiva: tipologia delle variabili

1. **Variabili quantitative**:misurano i numeri i quali sono interessanti come quantita'
    - **Discrete:** variabili di conteggio
    - **Continue :** misurano grandezze fisiche
2. **Variabili qualitative / fattori:** misurano le caratteristiche del soggetto, codificate con etichette / livelli per appartenenza di una certa classe (colore degli occhi)

Character Data : variabili identificativi e non si analizzano

Lezione 2

Statistica descrittiva univariata : lavoriamo su una variabile alla volta,divideremo le analisi secondo la tipologia di analisi
Variabili quantitative :
Calcolare degli indici → Summaries (riassuntivi)
collezione di pochi indici che siano pieni di significato(sostituiamo 130 numeri in 5)

1. Categoria degli **INDICI DI POSIZIONE:** collezione di indici che si occupano di indicarci dove troviamo i dati sull'asse reale
- **MEDIA CAMPIONARIA** (x sbarra sopra), e' pero influenzata da eventuali **valori estremi**(lontano dagli altri)

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

(MEDIA ARITMETICA DELLE OSSERVAZIONI)

$$\bar{x} = 3.72$$

- **MEDIANA CAMPIONARIA :** ci da l'indice di posizione del valore centrale delle osservazioni
  su 7 numeri [1,2,3,4,5,6,7] il valore centrale e' 4
  su 6 numeri [1,2,3,4,5,6] il valore centrale e' ¾

- **QUANTILI CAMPIONARI: Q25=** pedice indica un numero tra 0 - 100, la percentuale di osservazioni che lasci a sinistra

- **QUARTILI :** Q25 - Q50 - Q75 = Quartili, perche' dividono in quarti le osservazioni

2. Categoria degli **INDICI DI DISPERSIONE:**
- **Range:**
- **Varianza campionaria:** definita con s quadro, la varianza e' in un'altra scala perche' eleviamo al quadrato
  → per tornare all'unita di misura delle osservazioni s = √s²

$$s^2 = \frac{1}{\underset{\downarrow\ 1/n}{n-1}} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

MEDIA CAMPIONARIA

SCARTI QUADRATICI DALLA MEDIA CAMPIONARIA

$\bar{x}$ .... MEDIA ARITMETICA

- **Coefficiente di variazione:**

$$CV = \frac{s}{\bar{x}} = \begin{cases} <1 & \text{VARIAB.} < \text{EXP} \\ 1 & \text{VARIABILITÀ PARI A QUELLA DELL'ESPONENZIALE} \\ >1 & \text{VARIAB.} > \text{EXP} \end{cases}$$

3. categoria degli **Indici di forma:**

$$\frac{1}{n}\sum_{i=1}^{n} z_i^3$$

$$z_i = \frac{x_i - \bar{x}}{s}$$

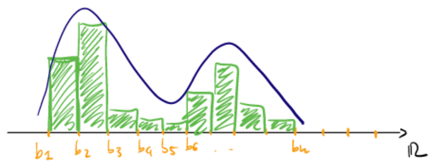
Z-SCORE

MEDIA CAMPIONARIA PARI A 0 e DEV. ST. CAMP. PARI A 1

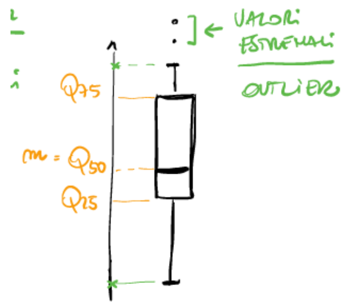
- **Skewness** : non simmetria / asimmetria

LEZIONE 3

Descrittiva univariata per variabili quantitative:



- **Istogramma:** grafico che rappresenta la distribuzione dei dati
  1) Si divide la retta in intervalli (Pins) in genere tutti uguali
  2) di ciascun intervallo una colonna alta quanto il n di osservazioni nell'intervallo



- **Boxplot:** grafico orizzontale o verticale con:
  1) una scatola : 50% dei valori centrali
  2) dei baffi : 25% dei valori piu piccoli e piu grandi,esclusi gli outliner
  3) Outliner: valori che si allontanano troppo dagli altri dati Troppo = maggiori di 1.5(75-q25) , maggiori della distanza interquartile

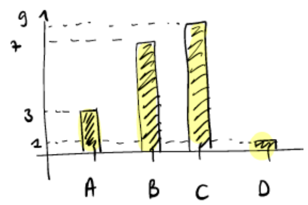
Descrittiva univariata per dati qualitativi/categoria:

A	B	C	D
3	7	9	1

si tabula la variabile : conto quante osservazioni in ciascuna categoria, ovvero quante volte compare una certa label
Tabella di conteggio

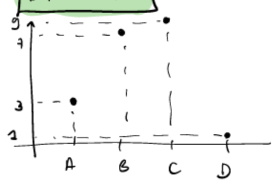
Grafici:

BAR CHARTS



- **Bar charts:** altezza variabile di ciascuna variabile

DOT CHART



- **Dot chart:**

- **Pie Chart:**DAI CHE LO SAI E NON LO DEVO METTERE IL DISEGNINO, che tanto sono inutili

Descrittiva BIVARIATA:

Variabile quantitativa / Variabile qualitativa
Siamo interessati alle distribuzioni delle distribuzioni nella v.quantitativa nelle diverse classi

VAR 1	VAR 2
x_1	A
x_2	B
	A
⋮	⋮
x_n	A

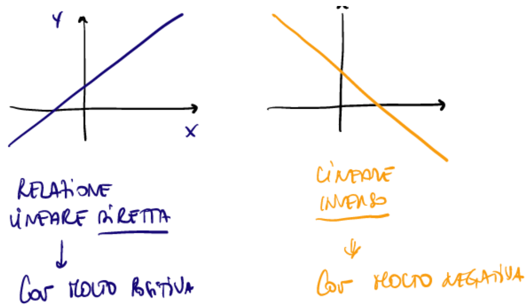
LEZIONE 4

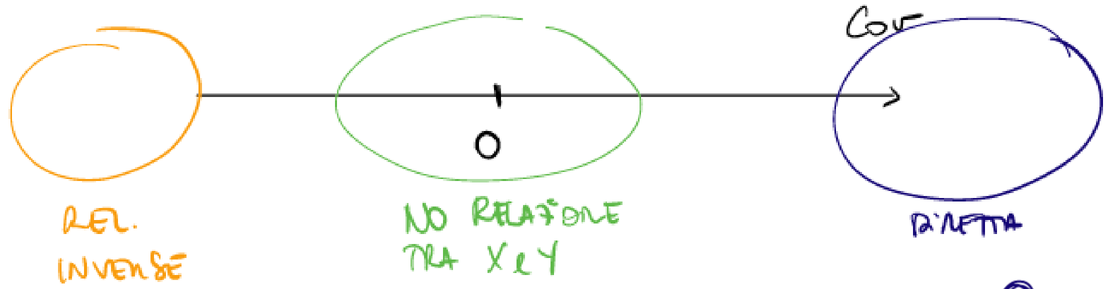
V.quantitativa / V. quantitativa : che relazione hanno X e Y,
hanno un qualche legame ?

- **Scatterplot**: grafico a punti delle informazioni
  1) **nuvola informe :**i punti si spargono in modo causale (non c'e' relazione tra X e Y)
  2) Osservazioni si dispongono lungo il grafico di una funzione

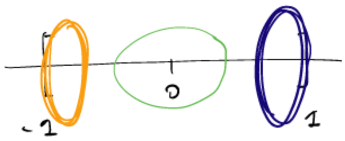
Anche degli indici che riassumano e quantifichino la relazione:

Media varianza - MOMENTI
Covarianza(x,y), media x* y - la media di x - la media di Y
Covarianza positiva , Covarianza Negativa
Versione Normalizzata , correlazione(X,Y) [-1 ,

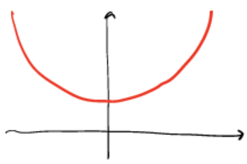




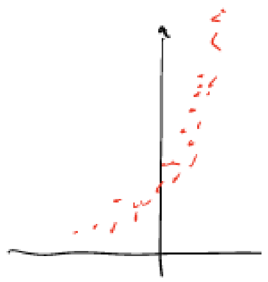
1]



LA COV. PUÒ VENIRE UGUALE A 0 ANCHE SE LE VAR. SONO LEGATE es: $Y=X^2$

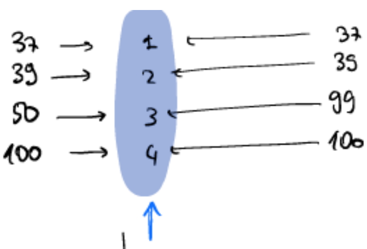


INDICE DI CORRELAZIONE CAMPIONARIA DI PEARSON:
stima la correlazione di prima appartenente all'intervallo [-1 , 1] , piu si avvicina allo zero e meno sono correlati ,piu si avvicina a uno e piu sono correlato , a -1 sono negativamente correlati



CORRELAZIONE CAMPIONARIA DI SPEARMAN: meglio di quella di Person, la calcolo se la relazione non e' lineare ma sembra monotona

Calcolo i ranghi: posizioni di ciascuno osservazione nel vettore ordinata in maniera crescente



CORRELAZIONE CAMPIONARIA DI PEARSON, MA DEI RANGHI

V.QUALI / V.QUALI :

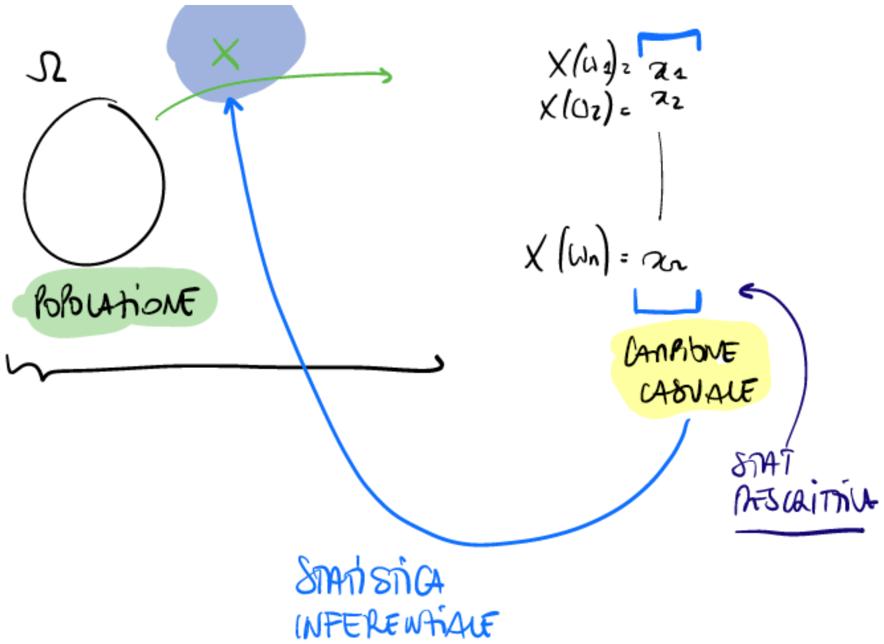
TABELLE DI FREQUENZA

	A	B	C	D
PIPPO	32	1	0	0
PLUTO	5	5	3	21

LEZIONE 5:

Statistica inferenziale:

Spazio campionario : la popolazione
campiono gli individui : numerandoli → estrazioni con reimbussolamento
Campione casuale(dati): il nome della persona → statistica descrittiva
Parleremo di campione di misure, di popolazione delle misure
Confronto individui e misure , ci interessano le misure
A partire dal campione casuale, vorrei dare informazioni sulla popolazione, la statistica descrittiva non cambia, ti dice sempre quei numeri delle persone della popolazione, la statistica inferenziale invece va oltre, a partire dal campione casuale vorrei dare informazione sulla popolazione, ovvero dalla variabile aleatoria X dalla quale sto campionando (generalizzazione)



Statistica inferenziale parametrica :
statistica inferenziale → vogliamo informazioni su X , la popolazione
parametrica → x e' una v.aleatoria

1) Famiglia : v.a → x e' Bernoulli (evento con successo 1,0 insuccesso), x e' Binomiale, geometrica, Poisson, normale, esponenziale, uniforme // e' come se le conoscessi
2) Parametri : x e' Bernoulli(p) , binomiale(n,p), normale(μ,sigmaquadro), esponenziale(lambda)

La statistica inferenziale parametrica si occupa di dare informazioni sui parametri della popolazione X
x binomiale (n,p) = media = n*p e poi Varianza = n*p(1-p)
x normale (μ,sigma quadro) = media= μe poi varianza = sigma quadro
media e varianza → funzioni dei parametri delle famiglia

Parametri sono 3 : Media - Varianza - Proporzione

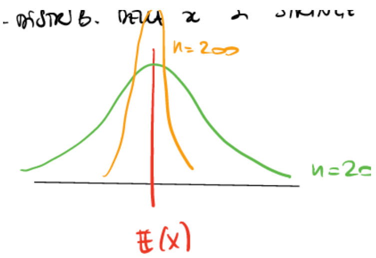
Proporzione : ci riferiamo al parametro **P** della Xn Bernoulli(p)
ci servono per codificare le variabili di tipo qualitativo
→ x = 1 →p 1=capelli rossi
→ x = 0 → (1 - p) 0=capelli non rossi

Variabilità campionaria: do informazioni sui parametri che controllano la variabilità campionaria, delle tecniche che tengano conto che il campione e' causale e che avrebbero potuto darci altri 130 numeri diversi

IDEA: Valutare quanto diversi potrebbero essere altri 130 numeri a partire dalla variabilità nei 130 che mi hanno dato

COME: 1) **calcolare una stima puntuale**(calcolato a partire dal campione casuale e che aspettiamo che sia vicino al parametro che non conosciamo) per il parametro che ci interessa ,ma lo abbiamo gia' fatto :
Media popolazione → Media campionaria
Varianza popolazione → Varianza campionaria
Proporzione → proporzione campionaria // frequenze relative vengono chiamati stimatori, sono funzioni del campione casuale(devono essere calcolabili, senza incognite), che forniscono stime puntuali,ovvero valori *__vicini__* al valore del parametro incognito
vicini PERCHE? (non lo dice nessuno) :

1) la media campionaria cambia quando cambio il campione casuale, e' una v.a. ha la sua distribuzione
2) la distribuzione della x cappello e' centrata
   la media campionaria e' uno stimatore corretto per la media della popolazione
3) quando n (taglia del campione) aumenta la distribuzione della media campionaria si stringe
4) La forma a campana c'e' sempre qualunque sia la distribuzione della popolazione x

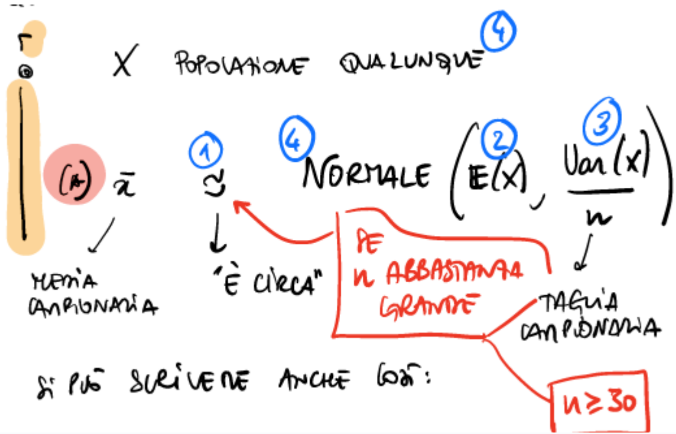


2) **controllare la variabilita':**siccome non viene calcolata dalla stima puntuale,
- intervalli di confidenza
- test di ipotesi

LEZIONE 6

Le 4 osservazioni di prima sono riassunte in 2 teoremi:

1) **TEOREMA DEL LIMITE CENTRALE:** X la popolazione qualunque



pNorm calcola la parte a sx della X

p norm

$$P(X \le x) = \int_{-\infty}^{x} PDF$$

$X \sim Pois(\lambda = 5)$

$E(X) = \lambda = 5$

$Var(X) = \lambda = 5$

1) $n = 20$

2) $n = 200$

$$P(\bar{x} > 5.5) \underset{(1)}{=}$$

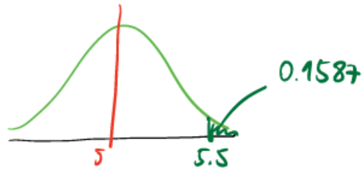
USO IL TLC

$$\bar{x} \approx N\left(5, \frac{5}{20}\right)$$

R) $\int_{5.5}^{+\infty} PDF = 1 - P(\bar{x} \le 5.5) =$

$= 1 - pnorm\left(5.5, 5, \sqrt{\frac{5}{20}}\right)$

$= 0.1587$



$$\bar{x} \xrightarrow[n \to +\infty]{} E(X)$$

2) **LEGGE DEI GRANDI NUMERI:** x medio converge alla media della popolazione, sono risultati asintotici, le approssimazioni vanno bene quando n e' grande

1) squadro cambia al cariare del campione casuale ,e' una v.a. anche s quadro → ha una distribuzione ,come si comporta?

2) s quadro e' corretto cioe' $E(s^2) = Var(X)$

3) c'e' una leggera asimmetria

4) se n tende a infinito (cresce) la distribuzione di s quadro si stringe cioe' var (s quadro) decresce

LGN: s quadro tende alla Var(x)

Per s quadro non vale il TLC cioe non posso dire che e' circa Gaussiana

$\boxed{\hat{p} = \bar{x}}$

Per la p cappello valgono tutti i risultati che abbiamo visto per la media campionaria

TLC per p cappello: la popolazione e' un Bernoulli

LGN per p cappello :

$$\hat{p} \xrightarrow[n \to +\infty]{} p$$

METODI DI CONTROLLO DELLA VARIABILITA' → INTERVALLI DI CONFIDENZA:
due metodi che fanno la stessa cosa
[a,b] → intervallo di numeri
restituiamo una confidenza indicata come (1-∝)=0.95/0.90/0.99 →alta probabilità di essere vero

Come leggerli: la probabilita' che l'intervallo [a,b] contenga il valore vero del parametro (il valore vero del parametro) e' pari a (1- ∝) aka la confidenza

La confidenza non e' un risultato,mauna scelta , la fisso prima di iniziare
(1- ∝) = 0.95/0.90/0.99

In pratica di tico che il numero sta all'interno dell'intervallo con probabilita' (1- ∝)

[a,b]?
IC per proporzioni:

- **Metodo della quantità pivotale**
  Passo 1) Ti danno la quantità, e' una funzione del campione casuale,che contiene il parametro incognito MA della quale conosci la distribuzione

$$\frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \sim N(0,1) \quad \hookrightarrow \boxed{n \geq 30}$$

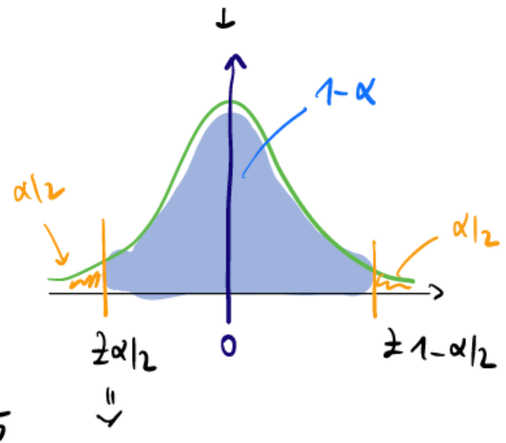
tipo qua e' una normale (0,1)

Passo 2) posso calcolare due valori q1 e q2 tali che :
la probabilità(q1<q.Pivotale<q2) = (1-∝)

$$\mathbb{P}\left(q_1 < \text{Q.PIVOTALE} < q_2 \right) = 1-\alpha$$

Praticamente sono i QUANTILI, potendo esplorare tutte le aree
a pedice l'area che lasciano a sx
(1-∝) = 0.95
q2= z0.95

Scelgo q1= z ∝ /2
q2 = z 1-∝ /2



$1-\alpha = 0.95$
$\alpha = 0.05$
$\alpha/2 = 0.025$
$1-\alpha/2 = 0.975$

$z_{0.025} = -1.96 = \text{qnorm}(0.025)$
$z_{0.975} = +1.96 = \text{qnorm}(0.975)$

Cosa vuol dire? la scelta simmetrica, l'area ∝ la dividiamo e in mezzo 1-∝

Perche' e' la scelta ottimale perche si dimostra che restituisce gli IC più stretti

$$\mathbb{P}\left(-1.96 < \frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})}} < 1.96 \right) = 0.95$$

LEZIONE 7:

Metodo della quantità pivotale(continuo esercizio lezione 6), ti danno una variabile aleatoria la quale contiene un oggetto che tu sai come viene distribuito

$$P\left(\underset{z_{\alpha/2}}{-1.96} < \frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} < \overset{z_{1-\alpha/2}}{1.96}\right) = \underbrace{0.95}_{1-\alpha}$$

PASSO 3:

TRASFORMIAMO QUELLO CHE È SCRITTO NELLA FORMA CON LA QUALE ABBIAMO VISTO L'IC

"LA P CHE IL VALORE VERO DEL PARAMETRO SIA COMPRESO TRA [a,b] È PARI A 1-α"

$$P(a < p < b) = 1-\alpha$$

$$P\left(z_{\alpha/2}\cdot\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < \hat{p}-p < z_{1-\alpha/2}\cdot\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 1-\alpha$$

$$P\left(z_{\alpha/2}\cdot\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} - \hat{p} < -p < z_{1-\alpha/2}\cdot\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} - \hat{p}\right) = 1-\alpha$$

$$P\left(\underbrace{\hat{p} - z_{1-\alpha/2}\cdot\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}_{a} < p < \underbrace{\hat{p} - z_{\alpha/2}\cdot\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}_{b}\right) = 1-\alpha$$

$$IC: \left[\hat{p} - z_{1-\alpha/2}\cdot\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} - z_{\alpha/2}\cdot\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

E= errore standard, simmetrico alla p cappello.

se n aumenta Errore standard diminuisce, l'IC diventa più piccolo

se la confidenza aka (1-∝) aumenta z1- ∝/2, allora z1-∝ /2 si alza
IC diventa più largo

$$E = z_{1-\alpha/2}\cdot\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = -z_{\alpha/2}\cdot\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

ERRORE STANDARD

$$[\hat{p}-E,\ \hat{p}+E]$$

IC per la media:
Passo 0: fisso 1-∝

$$\frac{\bar{x} - \mathbb{E}(X)}{\frac{s}{\sqrt{n}}}$$

Passo 1 : mi danno una quantità pivotale →
x medio= il valor medio
EE(x)=media della popolazione
s=Deviazione Standard campionaria
se hai X una taglia maggiore di 30, n>30

$$\frac{\bar{x} - \mathbb{E}(X)}{\frac{s}{\sqrt{n}}} \approx t(n-1)$$

t di STUDENT CON n-1 GRADI DI LIB.

puoi considerarlo t di student con n-1

$$\frac{\bar{x} - \mathbb{E}(X)}{\frac{s}{\sqrt{n}}} \sim t(n-1)$$

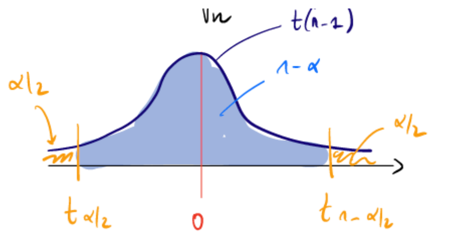
se X normale , puoi affermare che
Chi e' una t(n-1) :

1) Una variabile aleatoria continua
2) PDF

sulle code rimane più massa, ha le code alte, quando degrees of freedom (grado di liberta) aumentano, assomiglia sempre piu alla n(0,1)
E simile alle n(0,1)

$$\mathbb{P}\left(t_{\alpha/2} < \underbrace{\frac{\bar{x} - \mathbb{E}(X)}{\frac{s}{\sqrt{n}}}}_{t(n-1)} < t_{1-\alpha/2} \right) = 1-\alpha$$

PASSO 2 :



$(1-\alpha) = 0.95$ $\alpha = 0.01$ $\alpha/2 = 0.005$

$t_{\alpha/2} = t_{0.005}$:(R) $= -2.68$

$n=50$ $t(49)$

$t_{1-\alpha/2} = 2.68$

Passo 3: Riscrivo , cioe'

$$IC \left[\bar{x} - t_{1-\alpha/2} \cdot \frac{s}{\sqrt{n}} \, , \, \bar{x} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} \right]$$

$$\left[\bar{x} - E, \ \bar{x} + E\right]$$

Errore Standard:

1) Se n cresce → IC piu piccolo
2) se (1-∝) cresce → t1-∝/2 → IC piu grande

Prima di fare gli esercizi : Check → 1) se n>= 30 oppure 2) X normale

LEZIONE 8 :

Intervalli di confidenza per la varianza:

s quadro = varianza campionaria

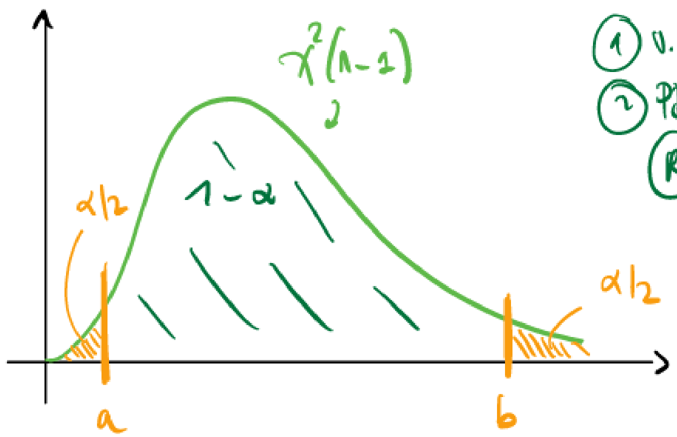
var(x)

x popolazione

Passo 1: (1-∝)

Passo 2: ricevo la variabile Pivotale → $\frac{(n-1)s^2}{Var(X)}$ chi quadro con n-1 gradi si liberta' se X e' normale . $\chi^2(n-1)$

Passo 3 : trovare a e b tali che $P\left(a < \text{V. PIVOTALE} < b\right) = 1-\alpha$



① V.A. CONTINUA
② PDF
ℝ

la Chi quadro

scelta l'area ∝ a meta' nelle due code :

$a = \chi_{\alpha/2} = \chi_{0.025} = 31.5549$

$b = \chi_{1-\alpha/2} = \chi_{0.975} = 70.2224$

$(1-\alpha) = 0.95$
$\alpha = 0.05$
$\alpha/2 = 0.025$
$n = 50$
$\chi^2(49)$

QUINDI :

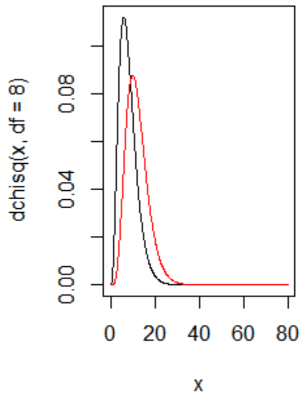
$$P\left(\chi_{\alpha/2} < \frac{(n-1)s^2}{Var(X)} < \chi_{1-\alpha/2}\right) = 1-\alpha$$

Passo 4 : Riscrivo →

$$\text{IC PER } VAR(X) \rightarrow \left[\frac{(n-1)s^2}{\chi_{1-\alpha/2}} , \frac{(n-1)s^2}{\chi_{\alpha/2}} \right] \quad \text{CONFIDENZA } (1-\alpha)$$

se aumento (1 - ∝) : chi ∝ /2 → diventa piu piccolo e chi ∝/2 → diventa piu grande IC si allarga

se n aumenta : (n-1) aumenta ,ma IC diventano piu stretti
VarCI comando per calcolare la Varianza



Esercizio 12
 Istogramma delle degenze delle neonate femmine non e' normale , X non può essere ipotizzata normale

TEST DI IPOTESI : secondo metodo di controllo della variabilita' oltre IC
Ipotesi : sono affermazioni che riguardano i parametri della popolazione

Abbiamo sempre 2 ipotesi : Ipotesi nulla → H0 & Ipotesi alternativa → H1

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

idk mo cambia gli indici ma vabb
Testiamo H0 contro H1, ci chiediamo se e' opportuno abbandonare (rifiutare) H0 a favore di H1?
Credo in H0, raccolgo un campione casuale e mi domando se cio che e' contenuto nei numeri mi costringe ad abbandonare H0 a favore di H1
H0=5 - H1 >5 O H1<5 Test unilaterali a una coda // cerco valori solo da una parte
H0=5 H1 != 5 Test bilaterali a due code
Test di ipotesi sulle proporzioni:
passo 1 : H0 : p=0.5 H1: p>0.5 scrivi direttamente H0 e H1

passo 2 : Fisso la significativa alfa piccolo : 0.05 - 0.01 - 0.1

Errore di 1 specie: rifiuti H0 quando H0 e' vera, alfa e' la soglia Max per l'errore di prima specie

Errore di 2 specie: Non rifiuti H0 quando e' falsa

Matrice confusione	H0 vera	H0 falsa
Rifiuto H0	❌	✅
Non rifiuto H0	✅	❌

Passo 3 : calcolo la statistica del test, sono quantita' aleatorie che dipendono dal campione e dal parametro p che non e' incognito per chi fa il test perche CREDO in H0(la mia ipotesi)

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \underset{(TLC)}{\approx} N(0,1) \quad \boxed{n \geq 30} \qquad \frac{\frac{83}{125} - 0.5}{\sqrt{\frac{0.5 \cdot 0.5}{125}}} = 3.6672$$

0.5 lo abbiamo deciso noi ,83 e' il numero di successi ed n=125

Passo 4 : e' un valore ragionevole con ciò che mi aspettavo ?
H1,sarebbe stato meglio sotto alternativa H1?

CALCOLO

P (STATISTICA del TEST > 3.6672) = PVALUE

↑

N(0,1)

Calcolo la probabilita' aka statistica del test :
> e' il nostro H1
1-p (stat test< 3.6672) = pnorm(3.6672) = 0.0001
PValue < ∝ ? mi fissa la soglia dell'errore di prima specie
se e' piu' piccolo rifiuto H0 a favore di H1

LEZIONE 9 → inizio lezione con riassunto lezione 8

Non possono essere vere entrambe
binom.test.
∝ = 0.01 succ=83 p.value= 0.0002< ∝ =0.01?
Si rifiuto H0 a favore di H1 ovvero p>0.5 posso affermare con significativa ∝ =0.01 che la popolazione e' d'accordo con me

OSS.supponiamo
H0= p=0.5   H1= p>0.5   n=125   succ=20   p cappello = 2/125 = 0.16
pvalue = 1< ∝ =0.01
No non posso rifiutare H0 in favore di H1 – Pvalue = 1

ESERCIZIO 9.10 : The variable sat.m in th data set stud.recs (Using R) contains math SAT scores for a group of students sampled from a larger population.Test the null hypothesis that the population mean score is 500 against a two-sided alternative.
Would you accept or reject at a 0.05 signification level? Pvalue

H0 = μ=500   H1= μ!= 500   ∝ = 0.05

t.test(stud.recs$sat.m, μ= 500, alternative = "two.sided")
# otteniamo che `t = -2.5731, df = 159(n casi meno 1 ), p-value = 0.01099`
`IC = 475.1437 sx 496.7313 dx`

P Value = 0.011 < 0.05 ?   SI) rifiuto H0 a favore di H1 μ != 500

SE ∝ = 0.05 - P Value = 0.011 < 0.01 ?   NO) Non posso rifiutare H0 a favore di H1

ESERCIZIO 9.11: In the babies (Using R) data set, the variable dht records the father's height for the sampled cases. Do a significance test of the null hypothesis that the population mean height is 68 inches against an alternative that it is taller. Remove the values of 99 from the data, as these indicate missing data

H0 = μ=68 H1 = μ>68 ∝ = 0.01

t.test(altezze, μ= soglia, alternative = "greater")
# otteniamo che `t = 20.796, df = 743, p-value < 2.2e-16`
`IC = 177.8755       Inf`

P value < 2.2e-16 < 0.01? SI)rifiuto H0 a favore di H1, affermo che la popolazione e i padri ha altezza media maggiore di 68 inches

Test su differenza di medie: Proporzioni medie e varianze
2 casi
1) Campioni appaiati: Le diete, soggetti con pesi iniziali, e soggetti con pesi dopo la dieta
vorrei dire che la media dei pesi dopo, e' inferiore alla media dei pesi prima
Ipotesi μprima>μdopo
H0: μp=μd H1: μp>μd
H0: μp-μd =0 H1: μp-μd>0
sulla stessa riga ho lo stesso soggetto

APPAIATI ?

PRIMA	DOPO
•	x

2) Campioni indipendenti: sono interessato al peso e alla differenza di peso tra maschio e femmina, mi interessa sapere se il medio dei maschi e' maggiore del peso medio delle femmine
H0: μm - μf = 0 H1: μm > μf
H0 : μm - μf =0 H1: μm - μf >0 → differenza di medie
sulla stessa riga stesso soggetto, avro 2 var una quantitativa Sex

INDIPENDENTI ?

PESO	SEX

Passo 1 : il mio problema ha campioni appaiati o indipendenti?

LEZIONE 10 :

1) **campioni appaiati:**

PESO PRIMA	PESO DOPO	DIFFER.
x	o	x-o

DIFFERENZA = PESO PRIMA - PESO DOPO

$$E(X-Y) = E(X) - E(Y)$$

media della differenza e' uguale alla differenza delle medie
H0= μp-μd = 0 H1: μp-μd>0
H0: μdiff=0 H1: μdiff>0 → quando i campioni sono appaiati ,ha senso costruire la variabile aleatoria e testare la differenza di medie equivale a testare la media della differenza
→TEST SULLA MEDIA

2) **Campioni indipendenti:** invece non ha senso calcolare le differenze del peso di una donna e di un uomo, magari ho preso 2 campioni femmine,

Passo 1: Scrivo bene H0 e H1
H0: μm - μf = 0 H1: μm-μf>0
(μm>μf)

Passo 2: fisso ∝ = 0.01 / 0.05 / 0.1

Passo 3: calcolo la statistica del test
se nm&nf>30 oppure xm & xf Normali
Calcolo il valore campionario e sotto H0

$$\frac{\bar{x}_M - \bar{x}_F - (\mu_M - \mu_F)}{\sqrt{\frac{s_M^2}{n_M} + \frac{s_F^2}{n_F}}}$$

Passo 4: probabilità che la statistica del testo assuma valori pari al valore campionari o più estremi nel verso dell'alternativa
P(stat del test > valore campionario) = Pvalue

Passo 5: Pvalue < ∝?
SE SI rifiuto H0 in favore di H1
SE NO non posso rifiutare H0

Esercizio 9.31: For the babies (UsingR) data set, the variable age contains the recorded mom's age and dage contains the dad's age for several different cases in the sample. Do a significance test of the null hypothesis of equal ages against a one-sided alternative that the dads are older in the sampled population.

H0 : μd - μm=0 H1: μd - μm > 0 (μd-μm) ∝=0.05

```
t = 28.092, df = 1226, p-value < 2.2e-16          IC = 2.943902       Inf
mean difference = 3.127139 // MEDIA CAMPIONARIA DELLE DIFFERENZE, i padri
sono più anziani delle madri di 3 anni
```

Pvalue = 2.2e-16< ∝=0.01 SI) posso rifiutare H0 a favore di H1 ovvero che i padri siano piu vecchi delle madri

Controllo ipotesi
n>30 : df=1226 n = df+1 = 1227✅
oppure
Xnormale = differenze <- babies$dage - babies$age

Esercizio 9.32 : The data set normtemp(UsingR) contains body measurements for 130 healthy ,randomly selected individuals from some parent population. The variable temperature contains normal body temperature data and the variable gender contains gender information,with male coded as 1 and female as 2. Is the sample difference across the two groups statistically significant?

H0 : μf=μm H1: μf != μm
μf - μm = 0 μf - μm != 0 →modo formale per scrivere il test

∝ = 0.05
Campioni indipendenti

```
normtemp$temp <- (normtemp$temperature - 32)/1.8        # trasformazione gradi
tempf <- normtemp$temp[normtemp$gender == 2]            # temperature femmine
tempm <- normtemp$temp[normtemp$gender == 1]            # temperature maschi

hist(tempf)            # istogrammi rispettivi
hist(tempm)

boxplot(tempf, tempm)          # boxplot affiancati per vedere differenze

t.test(tempf, tempm, mu = 0, alternative = "two.sided", paired = FALSE) # t di student per
# femmine meno maschi , x - y
# mu = 0 ipotesi nulla
# alternativa con una two sided
# gli dobbiamo dire se sono appaiati
t = 2.2854, df = 127.51, p-value = 0.02394
95 percent confidence interval: 0.02156277 0.29980476
mean of x mean of y : 36.88547  36.72479
```

appaiati P value : 0.02394< 0.05 ? SI)rifiuto H0 a favore di H1
OSS: µf>30? µm>30? SI   oppure Xm normale ? Xf Normale ?

Esempio 9.36 : The Galton (HisData) data set contains data used by Francis Galton in 1885. Each data point contains a child's height and an average of his or her parents heights. Assuming the data is a random sample for a population of interest,perform a t-test to see if there is a difference in the population mean height. assume the paired t-test is appropriate

c= child   p=parent

H0 : µc - µp = 0   H1: µc - µp > 0   ∝ = 0.01

Campioni appaiati (testo)

```
boxplot(Galton$parent, Galton$child)

t.test(Galton$child, Galton$parent, mu = 0, paired = TRUE,
       alternative = "two.sided")
# child - parent ,x-y
# mu = 0
# paired= True (detto dal testo)
# alternative = 2 sided
t = -2.8789, df = 927, p-value = 0.004082
95 percent confidence interval: -0.36949983 -0.06993982
mean difference: -0.2197198
```

Pvalue= 0.004082< ∝ = 0.01 ? SI) posso rifiutare H0 a favore di H1
Controllo ipotesi : n >30 ? SI) OK , ma cosi non fosse dovrei controllare la Normale guardando le altezze nell’istogramma, come qui sotto

```
differenze <- Galton$child - Galton$parent
hist(differenze)                    # se a forma di campana allora e' corretto
```