Statistica – i miei appunti

Popolazione: insieme di tutti e soli gli elementi che hanno una caratteristica comune.

I campioni devono essere casuali e rappresentativi.

- ✓ Probabilità matematica: probabilità calcolabile a priori, cioè prima che si manifesti l'evento. Esempio: so a priori che ci possono essere sei eventi in un dado.
- ✓ Probabilità campionaria o frequenza relativa: ha bisogno dell'esperimento per essere valutata. Solo all'infinito io sarò sicuro di aver trovato la probabilità reale.
- ✓ Probabilità assiomatica: presuppone che noi abbiamo un insieme di eventi: tutto l'insieme vale 1 e a ogni sottoinsieme verrà attribuito un valore la cui somma totale con gli altri sottoinsieme farà 1.

Il teorema di Bayes: P(A|B) P(B) = P(B|A) P(A)

Tabella di verità/contingenza sui risultati dei test

Test/verità	Falso	Vero	Totale
Risultato positivo	Falso positivo FP	Vero positivo VP	TP
Risultato negativo	Falso negativo FN	Falso positivo VN	TN
Totale	TF	TV	T

Sensibilità (Se): proporzione dei soggetti realmente malati e positivi al test (veri positivi) rispetto all'intera popolazione dei malati. $Se = \frac{VP}{VP + FN} = \frac{VP}{TM}$

Specificità (Sp): proporzione dei soggetti realmente sani e negativi al test (veri negativi) rispetto all'intera popolazione dei sani. $Sp = \frac{VN}{VN + FP} = \frac{VN}{TS}$

Precisione: bassa dispersione dei valori attorno alla media.

Accuratezza: bassa distanza tra il valore medio dei dati e il valore vero.

Errore sistematico: δ = μ (media) – θ (valore vero). La misura è tanto più accurata quanto minore è l'errore sistematico.

Errore casuale: ε = x (una misura) – μ. La misura è tanto più precisa quanto più è piccolo l'errore casuale.

L'attendibilità o errore totale è la somma dell'errore sistematico e dell'errore casuale.

Le stime di accuratezza sono basate sulla media campionaria, le stime di precisione sono basate sulla deviazione standard.

Media: μ (popolazione) e ẍ (campione)

Deviazione standard: σ (popolazione) e S_x (campione)

La varianza è il quadrato della deviazione standard.

$S_x = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}}$ La deviazione standard va da 0 a $+\infty$. 0 corrisponde alla precisione massima.

La mediana M di un insieme di n dati ordinati in ordine di grandezza crescente è il valore centrale dei dati, se il numero di dati è dispari, o la media aritmetica dei due valori centrali, se il numero dei dati è pari.

La moda di un insieme di n dati è il valore o la classe a cui corrisponde la massima frequenza assoluta.

Il primo quartile Q_1 è un valore tale che il 25 % dei dati ordinati è minore o uguale a Q_1. Il primo quartile Q_1 è detto anche 25-esimo percentile e indicato con $P_{0.25}$.

Il terzo quartile Q_3 è un valore tale che il 75 % dei dati ordinati è minore o uguale a Q_3 ed è detto anche 75-esimo percentile e indicato con $P_{0.75}$.

Il secondo quartile Q_2 (50-esimo percentile) coincide con la mediana.

Boxplot: grafico in cui sono rappresentati tutti i dati che ci servono. Il boxplot ha nel suo centro un segno (barra) che è la mediana. I baffi sono due quartili. I valori molto distanti dalla mediana sono rappresentati con pallini.

Una matrice è costituita da un insieme ordinato di vettori, normalmente vettori colonna.

La funzione di densità di probabilità: è la curva che, per un numero di misure che tende a infinito, viene tracciata per mezzo di un istogramma, cioè un grafico sulla cui ascissa troviamo classi di misure simili e sulla cui ordinata troviamo la frequenza specifica di ciascuna classe.

Nel caso di misure con errori casuali, si può dimostrare che la funzione di densità di probabilità coincide con la distribuzione gaussiana e che la curva a campana è centrata attorno al valore vero, mentre la larghezza è legata alla precisione della misura.

La statistica è il rapporto adimensionale tra segnale e rumore. $Statistica = \frac{segnale}{rumore}$ La deviazione standard porta il rumore; la differenza tra le medie è l'informazione o segnale.

L'ipotesi nulla è di solito la predizione che non c'è nessuna relazione nella popolazione o nessuna differenza tra gruppi. L'ipotesi nulla dice che si suppone la differenza tra le medie pari a zero. Prima di passare all'ipotesi alternativa, bisogna scartare l'ipotesi nulla.

Errore standard della media: è la deviazione standard delle medie dei campioni casuali nella distribuzione gaussiana. $S_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$. La deviazione standard è definita come valore medio dell'errore stimato su ognuna delle misure; essa rappresenta l'errore che possiamo associare a

ciascuna delle misure. L'errore standard dalla media, minore della deviazione standard, ci porta a stabilire che la notazione finale di una misura è: $x_0 = \bar{x} \pm S_{\bar{x}}$

Standardizzazione: porre media nulla e deviazione standard uguale a 1.

La distribuzione di probabilità o funzione cumulativa è l'integrale della funzione di densità di probabilità e il suo valore massimo è 1. $F_x(x) = P(X \leq x)$ L'ordinata esprime la probabilità di trovare un valore inferiore a quello che abbiamo fissato e quindi cercato in ascissa.

La distribuzione normale o gaussiana è continua e dipende da media e deviazione standard. Ha la caratteristica forma a "campana".

La distribuzione binomiale è una distribuzione discreta che viene usata se l'esperimento aleatorio può assumere solo due possibili risultati. Esempio: viene usata in relazione al quadrato di Punnet.

La distribuzione t-Student è una distribuzione continua e simmetrica; ha media nulla e deviazione standard maggiore di 1; è caratterizzata da un unico parametro detto gradi di libertà. A differenza della distribuzione normale, alla quale assomiglia molto, ha code più pesanti e quindi è caratterizzata da più dati estremi o anomali.

α: livello di significatività: è la probabilità soglia che sono disponibile ad accettare con l'errore che sto commettendo. Se p.value < α, allora posso rigettare l'ipotesi nulla. Internazionalmente α può essere uguale a: 0,05; 0,01; 0,001. α è l'errore di falso positivo; è la probabilità di errore di tipo I.

p.value è la probabilità di errore che abbiamo se consideriamo l'ipotesi nulla. p.value = p.coda destra + p.coda sinistra. Il p.value è la probabilità di fare falso positivo. Siamo contenti che il p.value sia alto quando cerchiamo di ottenere la conferma che due elementi sono uguali.

β è definita come probabilità di errore di tipo II; è l'errore di falso negativo.

1- β è definito potenza del test; è la probabilità di scelta corretta: rifiutare l'ipotesi nulla quando l'ipotesi alternativa è vera.

		Stato effettivo dell'ipotesi nulla	
Decisione presa		Ipotesi nulla vera	Ipotesi nulla non vera
	Accetta ipotesi nulla	Scelta corretta p = 1- α	Commette errore di tipo II. p = β
	Rifiuta ipotesi nulla	Commette errore di tipo I. p = α	Scelta corretta p = 1- β

Maggiore è il valore sperimentale della statistica, più favorevole è la conclusione sulla significatività in quanto il segnale prevale sul rumore.

α e β sono dipendenti l'uno dall'altro: se abbassi α, allora alzi β.

I test parametrici sono quei test di significatività che si basano sui parametri fondamentali della statistica: media e deviazione standard.

t-test

- ✓ Unico campione paragonato alla popolazione

- ✓ Campioni indipendenti comparati con il controllo
- ✓ Campioni dipendenti o accoppiati

Il t-test è obbligatorio per piccoli gruppi di campioni < 30, ma è preferibile usarlo anche per numeri maggiori.

Se le varianze dei campioni sono diverse, il t-test deve apporre una correzione chiamata Welch.

Nell'analisi della potenza del test, n indica qual è il numero di elementi che serve per avere una determinata potenza. Quando abbiamo troppi pochi elementi rischiamo di rigettare l'ipotesi giusta.

Test del chi-quadrato

È un test non parametrico; serve a verificare:

- ✓ Se un campione di dati si adatta a una distribuzione teorica attesa (bontà di adattamento)
- ✓ Se due o più fattori, applicati allo stesso insieme di dati, siano indipendenti (test di indipendenza)

Nel test del chi-quadrato si devono mette le frequenze in una tabella, nel caso del test di indipendenza.

Modelli e regressione

Un modello deve essere in grado di spiegare molti aspetti della realtà e usare meno concetti possibili. Un polinomio, in un grafico a dispersione x-y, non può essere il modello giusto perchè il suo grado dipende sostanzialmente dal numero di punti sperimentali che hanno variabilità casuale anche dovuta al rumore. Una retta, che chiamiamo di regressione, è un compromesso tra complessità e riduzione della ridondanza.

Regressione lineare semplice: una variabile dipendente e una variabile indipendente.

Regressione lineare multipla: una variabile dipendente e diverse variabili indipendenti.

Regressione per la popolazione: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ dove β_0 è l'intercetta per la popolazione, β_1 è l'inclinazione per la popolazione e ε_i è l'errore casuale relativo all'osservazione i-esima.

Regressione per il campione: $\hat{Y}_i = b_0 + b_1 X_i$ dove $\hat{Y}_i$ è il valore stimato di Y per l'osservazione i-esima, b_0 è l'intercetta campionaria e b_1 è il coefficiente di regressione.

Metodo dei minimi quadrati: con iterazioni successive, il metodo cambia b_0 e b_1 sino a minimizzare la somma dei quadrati dei residui. Il metodo consiste nell'elevare al quadrato i residui e sommarli, ottenendo un valore numerico che viene chiamato valore della funzione obiettivo. Minimizzando tale valore, troviamo la retta giusta.

Residuo: è la distanza di un y misurato sperimentalmente e l'y stimato grazie alla retta di regressione.

In una qualsiasi nuvola di dati, vi è una retta di regressione e questa passa per $(\bar{x}, \bar{y})$, cioè il punto caratterizzato dalla media delle due variabili.

La forza di correlazione è maggiore se i punti si trovano il più vicino possibile alla retta di regressione; se la retta di regressione ha una buona forza di correlazione, la retta è un buon modello.

Devianza totale: somma dei quadrati degli scarti. $\sum_{i=1}^{N}(y_i - \bar{y})^2$

Devianza residua: somma dei quadrati dei residui. Se la devianza residua è 0 vuol dire che non ci sono residui e quindi i dati si trovano esattamente sulla retta di regressione. $\sum_{i=1}^{N}(y_i - y_{stim})^2$

Devianza spiegata: $\sum_{i=1}^{N}(y_{stim} - \bar{y})^2$ Se la devianza spiegata è uguale a 0 vuol dire che la retta è parallela all'asse x, il coefficiente di correlazione è 0 e non c'è dipendenza tra x e y.

Devianza totale = devianza residua + devianza spiegata

Il coefficiente di correlazione indica la forza di correlazione. r va da –1 (rappresenta una retta in discesa) a 1 (rappresenta una retta in salita). Se r = 0, non c'è correlazione tra x e y. Più $r \to |1|$ più la retta di regressione si avvicina alla realtà. La correlazione di una variabile con se stessa è uguale a 1. Per dire che c'è forza di correlazione, r deve arrivare a 0,7-0,75. $r = \pm\sqrt{\frac{Devianza \cdot spiegata}{Devianza \cdot totale}}$

Per verificare l'utilità del modello di regressione, si usa un'analisi basata sul calcolo del parametro F. se F è alto, la devianza spiegata prevale e quindi il modello lineare risulta accettabile. $F = \frac{(devianza \cdot spiegata)/k}{(devianza \cdot residua)/(n-k-1)}$ dove k è il numero di variabili indipendenti e n è il numero di campioni.

Intervallo di confidenza: intervallo di valori centrato sul valore medio entro il quale noi abbiamo una certa probabilità (di solito si fissa 95 %) che vada a trovarsi il valore di tutte le altre possibili medie. Per calcolarlo, $\bar{x} - z_{\alpha/2} \cdot \sigma_{\bar{x}} \le \mu \le \bar{x} + z_{\alpha/2} \cdot \sigma_{\bar{x}}$ dove $z_{\alpha/2}$ è il fattore critico che va calcolato con qnorm in R (dentro qnorm bisogna mettere 1- α, cioè se lavoro a due code 1-0,025 mentre se lavoro con una 1-0,05) e $\sigma_{\bar{x}}$ è l'errore standard.

ANOVA: analisi della varianza

ANOVA generalizza il t-test per più di due gruppi contemporaneamente. Analizza:

- ✓ La varianza nel gruppo (within)
- ✓ La varianza tra i gruppi (between)

ANOVA a una via: quando c'è una sola variabile dipendente e una indipendente.

- ✓ Ipotesi nulla: le medie dei gruppi sono uguali; questo significa che tutti i gruppi sono stati estratti dalla stessa popolazione.
- ✓ Ipotesi alternativa: almeno una media è diversa; questo significa che almeno un gruppo è estratto da una differente popolazione.

La devianza tra le medie è la somma delle differenze tra la media di un gruppo e la media totale (si fa la media delle medie di tutti i gruppi) al quadrato. Per ottenere la varianza si divide per il numero di gradi di libertà.

Per fare la verifica di ANOVA, usiamo il F-test.

Quando rifiutiamo l'ipotesi nulla, ANOVA non ci dice qual è il gruppo che ha media diversa dalle altre perciò dobbiamo capire quale è: bisogna analizzare tutte le coppie di medie con il test Turkey-Kramer.

Se i gruppi hanno numerosità uguale tra loro (hanno cioè lo stesso numero di elementi), il disegno sperimentale si dice bilanciato. Se non è così si può comunque usare ANOVA ma ci sono alcuni svantaggi.