Stats tutorial

Michele Scandola

last modified: Wed Dec 01 09:28:47 2021

# Types of statistical variables 30 November 2021

Quelle che vediamo sono:

* scala continua (sotto: intervallare e numerica) parametrica
* scala continua non parametrica
* frequenze
* conteggi
* ordinali

Però classicamente queste variabili sono rappresentate in modo diverso. Per questo motivo qua sotto ti riporto qualcosa di più “classico”, poi riprendiamo quello che abbiamo visto.

Wikipedia: <https://it.wikipedia.org/wiki/Variabile_(statistica)>

|  |  |
| --- | --- |
| Tipologia dei dati | Contenuto informativo |
| Scala nominale (es: the type of brain lesion) | Relazione d’equivalenza |
| Scala ordinale (es: likert scale) | Relazione d’equivalenza + Relazione d’ordine |
| Scala intervallare (es: neurological level of spinal cord injury, number of lesioned voxels) | Relazione d’equivalenza + Relazione d’ordine + Relazione d’uguaglianza |
| Scala di rapporto (es: reaction times) | Relazione d’equivalenza + Relazione d’ordine + Relazione d’uguaglianza + Relazione di rapporto |

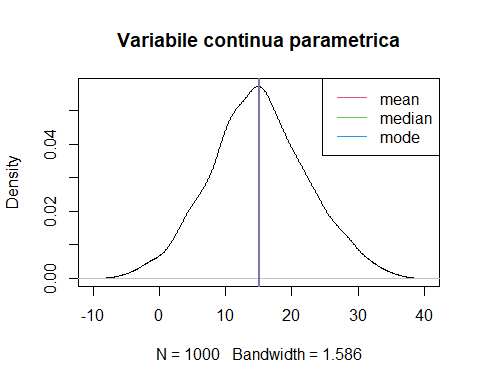
relazione d’equivalenza: you can say if two values are equal or not relazione d’ordine: you can say one value is greater or less than another relazione d’uguaglianza: if the difference between two values is the same between two different values es: 17 - 7 and 34 - 24, the two differences have the same meaning relazione di rapporto: you can compute the decimals

Torno ancora sulla tipologia di variabili, sulla differenza fra scala intervallare e di rapporto. La intervallare è una scala continua ma sulla quale non esistono i decimali. Il numero di anni compiuti ad esempio, il valore in euro degli oggetti (sotto il centesimo non vai… è un po’ tirata ma vabbé), se lo trasformiamo in numerico l’altezza della lesione spinale, oppure il numero di voxel di una lesione. E’ come se su questa misura non potessi aumentare la precisione a livello infinito… Ad un certo punto ti fermi, ma non per la precisione dello strumento, proprio perché non esiste altro. Su una scala invece a rapporto puoi aumentare questa precisione all’infinito. Pensa ai tempi di reazione, puoi misurarli in secondi, millisecondi, microsecondi, ecc… Potenzialmente potresti migliorare la precisione della misura in modo infinito. Ma questa differenza fra scala intervallare e ordinale non fa molta differenza a livello di analisi.

## continuous parametric variables

è una variabile che può essere su scala intervallare o di rapporto. è caratterizzata dal rispettare gli assiomi di Gauss-Markov, in pratica ha media, moda e mediana equivalenti e la funzione di frequenza di probabilità simmetrica.

varcont <- rnorm( n = 1000, mean = 15, sd = 7 )  
plot( density( varcont ), main = "Variabile continua parametrica" )  
abline(v=mean(varcont), col = 2)  
abline(v=median(varcont), col = 3)  
abline(v=my\_mode(varcont), col = 4 )   
legend("topright",  
 lty = c(1,1,1),  
 col = 2:4,  
 legend = c("mean", "median", "mode"))



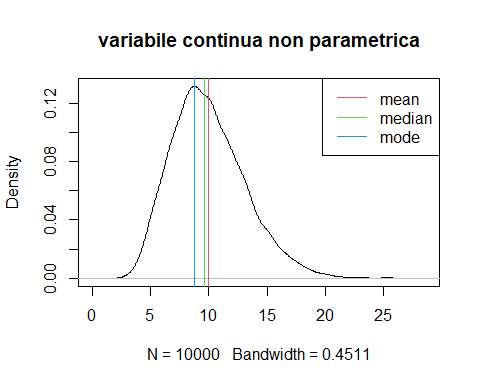
test che si possono utilizzare per analizzare questi dati

* se 2 gruppi indipendenti t.test(gruppo1, gruppo2, paired = FALSE) t.test( y ~ group, data = dat, paired = FALSE)
* se 2 condizioni dipendenti t.test(cond1, cond2, paired = TRUE) t.test( y ~ cond, data = dat, paired = TRUE)
* se regressioni o tutti fattori indipendenti: lm
* se fattori within-subjects, ANCOVA, fattori nested: aov, da afex aov\_4 aov\_car, mixed, da lme4 lmer

## continuous non-parametric variables

è una variabile che può essere su scala intervallare o di rapporto. è caratterizzata dal NON rispettare gli assiomi di Gauss-Markov, in pratica ha media, moda e mediana NON equivalenti e la funzione di frequenza di probabilità NON è simmetrica.

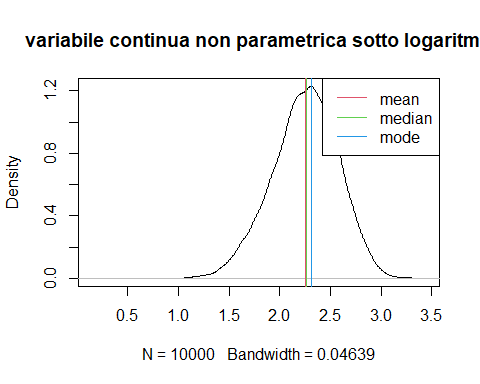
varcont2 <- rgamma( n = 10000, rate = 1, shape = 10 )  
plot( density( varcont2 ) , main = "variabile continua non parametrica" )  
abline(v=mean(varcont2), col = 2)  
abline(v=median(varcont2), col = 3)  
abline(v=my\_mode(varcont2), col = 4 )   
legend("topright",  
 lty = c(1,1,1),  
 col = 2:4,  
 legend = c("mean", "median", "mode"))

 per analizzarla:

1. provare a trasformare la variabile e vedere se diventa normale, come col logaritmo

* se 2 gruppi indipendenti wilcox.test(gruppo1, gruppo2, paired = FALSE) wilcox.test( y ~ group, data = dat, paired = FALSE)
* se 2 condizioni dipendenti wilcox.test(cond1, cond2, paired = TRUE) wilcox.test( y ~ cond, data = dat, paired = TRUE)
* se anova fattore unico indipendente: kruskal.test
* se anova fattore unico dipendente: friedman.test aov, da afex aov\_4 aov\_car, mixed, da lme4 lmer
* se modelo con effetti within e between, modelli lineari generalizzati glm o glmer da pacchetto lme4, mixed da pacchetto afex tempi di reazione: glmer con family = “Gamma”

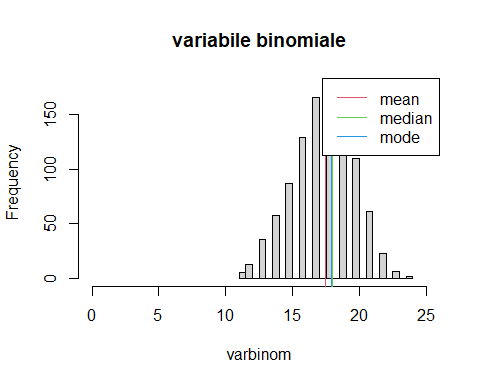
varlogcont2 <- log( varcont2 )  
plot( density( varlogcont2 ) , main = "variabile continua non parametrica sotto logaritmo" )  
abline(v=mean(varlogcont2), col = 2)  
abline(v=median(varlogcont2), col = 3)  
abline(v=my\_mode(varlogcont2), col = 4 )   
legend("topright",  
 lty = c(1,1,1),  
 col = 2:4,  
 legend = c("mean", "median", "mode"))



## frequency or binomial variables

è una di frequenza caratterizzata da un massimo totale questo a livello teorico comporta la possibilità di poter calcolare una probabilità di successo/fallimento

varbinom <- rbinom(n = 1000, size = 25, prob = 0.7)  
hist( varbinom , main = "variabile binomiale", breaks = 25,  
 xlim = c(0,25))  
abline(v=mean(varbinom), col = 2)  
abline(v=median(varbinom), col = 3)  
abline(v=my\_mode(varbinom), col = 4 )   
legend("topright",  
 lty = c(1,1,1),  
 col = 2:4,  
 legend = c("mean", "median", "mode"))



test che si possono utilizzare per analizzare questi dati

* se 2 gruppi indipendenti chisq.test(M)
* se 2 condizioni dipendenti mcnemar.test(M)
* se unico fattore between chisq.test(M)
* se covariate, fattori multipli, ecc… da lme4: glmer, da afex: mixed, da base: glm, sempre con family = “binomial”

### Esempi:

M <- cbind(  
 c( 10, 5),  
 c( 7, 19)  
)  
  
chisq.test(M)

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: M  
## X-squared = 4.6613, df = 1, p-value = 0.03085

mcnemar.test(M)

##   
## McNemar's Chi-squared test with continuity correction  
##   
## data: M  
## McNemar's chi-squared = 0.083333, df = 1, p-value = 0.7728

#### da database ottenere la matrici per chisquare

dat <- data.frame(  
 yb <- c(  
 rbinom(n = 50, size = 1, prob = 0.7),  
 rbinom(n = 50, size = 1, prob = 0.3)  
 ),  
 group = rep(c("A", "B"), each = 50)  
 )  
  
a <- aggregate( yb ~ group, data = dat, FUN = sum)  
a$no <- aggregate( yb ~ group, data = dat, FUN = function(x){length(x)-sum(x)})$yb  
   
chisq.test(a[,2:3])

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: a[, 2:3]  
## X-squared = 19.747, df = 1, p-value = 8.84e-06

mcnemar.test(as.matrix(a[,2:3]))

##   
## McNemar's Chi-squared test with continuity correction  
##   
## data: as.matrix(a[, 2:3])  
## McNemar's chi-squared = 1.3333, df = 1, p-value = 0.2482

#### unico fattore between subjects

M <- rbind(  
 c( 10, 5, 7, 8),  
 c( 7, 19, 2, 100)  
)  
  
chisq.test(M)

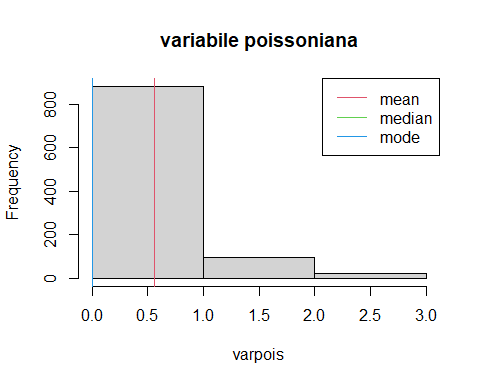
## Warning in chisq.test(M): Chi-squared approximation may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: M  
## X-squared = 47.229, df = 3, p-value = 3.107e-10

## count or poissonian variables

è un conteggio, non abbiamo un numero fisso massimo es: numero di nascite / morti numero di macchine rosse che passano in un’ora

varpois <- rpois(n = 1000, lambda = 0.5)  
hist( varpois , main = "variabile poissoniana", breaks = max(varpois),  
 xlim = c(0,max(varpois)))  
abline(v=mean(varpois), col = 2)  
abline(v=median(varpois), col = 3)  
abline(v=my\_mode(varpois), col = 4 )   
legend("topright",  
 lty = c(1,1,1),  
 col = 2:4,  
 legend = c("mean", "median", "mode"))



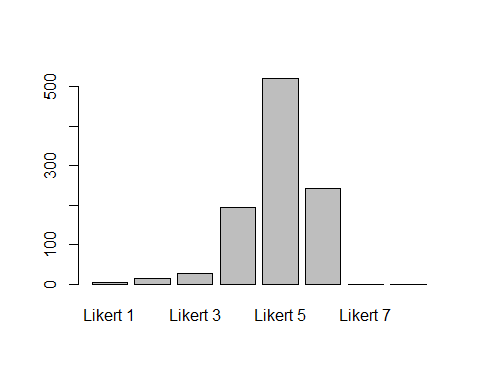
test che si possono utilizzare per analizzare questi dati

* se 2 gruppi indipendenti chisq.test(M)
* se 2 condizioni dipendenti mcnemar.test(M)
* se unico fattore between chisq.test(M)
* se covariate, fattori multipli, ecc… da lme4: glmer, da afex: mixed, da base: glm, sempre con family = **“poisson”**

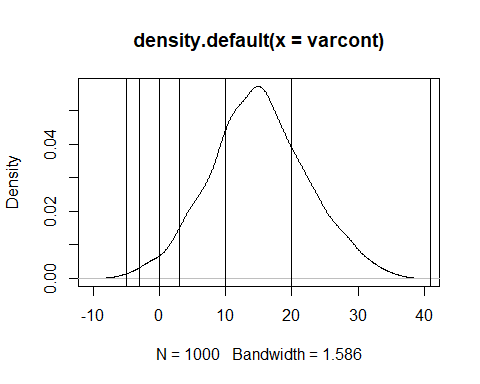
## ordinal variables

è una scala likert, ma non il punteggio totale di un questionario validato es: singola domanda di embodiment

varorder <- cut( varcont, breaks = c(-5, -3, 0, 3, 10, 20, 41))  
levels(varorder) <- paste("Likert", 1:8)  
barplot(table(varorder))



plot( density( varcont ) )  
abline(v=c(-5, -3, 0, 3, 10, 20, 41))



per analizzarla:

* se 2 gruppi indipendenti wilcox.test(gruppo1, gruppo2, paired = FALSE) wilcox.test( y ~ group, data = dat, paired = FALSE)
* se 2 condizioni dipendenti wilcox.test(cond1, cond2, paired = TRUE) wilcox.test( y ~ cond, data = dat, paired = TRUE)
* se anova fattore unico indipendente: kruskal.test
* se anova fattore unico dipendente: friedman.test
* se modelo con effetti within e between, modelli lineari generalizzati clmm o clmm2 da pacchetto ordinal