# #Probabilità

## #Binomiale

#P → somma delle probabilità di tutti i possibili esiti (0, 1, 2, n)
pbinom(n_successi, n_tentativi, probabilità)

#1-pbinom → successi maggiori di n
pbinom(n_successi, n_tentativi, probabilità, lower.tail = FALSE)

#D → probabilità di n_successi esatti
dbinom(n_successi, n_tentativi, probabilità)

## #Geometrica → numero di fallimenti PRIMA del successo

#P → somma delle probabilità di tutti i possibili insuccessi
pgeom(n_insuccessi, probabilità)

#1-pgeom → insuccessi maggiori di n
pgeom(n_insuccessi, probabilità, lower.tail = FALSE)

#D → probabilità di n_insuccessi esatti
dgeom(n_insuccessi, probabilità)

## #Ipergeometrica → estrazione di palline senza reimbussolamento

#P → somma dei possibili esiti delle estrazioni
phyper(n_successi, n_favoverevoli, n_sfavorevoli, n_tentativi)

#1-phyper → successi superiori a n_successi
phyper(n_successi, n_favoverevoli, n_sfavorevoli, n_tentativi, lower.tail = FALSE)

#D → probabilità di n_successi esatti
dhyper(n_successi, n_favoverevoli, n_sfavorevoli, n_tentativi)

## #Poisson → prove ripetute con probabilità molto piccola (dato dal testo)

#P → somma dei possibili esiti delle per i successi
ppois(n_successi, labda = n_eventi_medio)

#1-ppois → successi superiori a n_successi
ppois(n_successi, labda = n_eventi_medio, lower.tail = FALSE)

#D → probabilità di n_successi esatti
dpois(n_successi, labda = n_eventi_medio)

## #Normale

#P → valore sotto la curva tra valore minimo della curva e il valore indicato
pnorm(valore, mean = media, sd = deviazione_standard)

#1-pnorm → valore sotto la curva tra il valore indicato e il valore massimo
pnorm(valore, mean = media, sd = deviazione_standard, lower.tail = FALSE)

#dnorm → valore della probabilità al punto indicato
dnorm(valore, mean = media, sd = deviazione_standard)
#Q→ valore del quantile (valore per cui n% è minore di tale valore) (inversa della distribuzione (??))
qnorm(valore percentuale richiesto, mean = media, sd = deviazione_standard)

## #Uniforme

#P → valore sotto la curva tra valore minimo e il valore indicato
punif(valore, min = a, max = b)

#1-punif → valore sotto la curva tra il valore indicato e il valore massimo
punif(valore, min = a, max = b, lower.tail = FALSE)

#D → valore della probabilità al punto indicato

dnorm(valore, min = a, max = b)

**#Esponenziale** (lamba è sull'asse y, media = 1/lamba)

#P → valore sotto la curva tra valore minimo e il valore indicato

pexp(valore, rate = 1/media)

#1-pexp → valore sotto la curva tra il valore indicato e il valore massimo

pexp(valore, rate = 1/media, lower.tail = FALSE)

#D → valore della probabilità al punto indicato

dexp(valore, rate = 1/media)

#Q → valore del quantile (valore per cui n% è minore di tale valore) data una probabilità restituisce il numero che rappresenta nella funzione quella probabilità, cioè l'area sottesa dalla curva.

qexp(valore percentuale richiesto, rate = 1/media)

proprietà assenza di memoria: P(X>M | X>N) = P(X>M-N)
####################################

# #Statistica

data("InsectSprays") #selezione dataset

str(InsectSprays) #anteprima dataset

datiWithoutB = subset(dati, dati$group != "2") #rimozione dato gruppo (seleziona solo i dati con gruppo diverso da 2)

newdata <- InsectSprays$count[InsectSprays$spray != "A"] #considero solo un gruppo (da count stai selezionando solo quelli che non sono associati allo spray A.)

summary(InsectSprays$count) #sommario dataset

quantile(datiwithoutB$var, 0.6, na.rm = TRUE) #solo il 60%

dei dati ha var inferiore, se 60% var sup, 0.4 al posto di 0.6

(boxplot(InsectSprays$count) #grafico

abline(h = 20, col = "Blue") #visualizzazione valori outlier

hist(InsectSprays$count) #istogramma

mean(InsectSprays$count[InsectSprays$spray == "A"]) #media

median(InsectSprays$count[InsectSprays$spray == "A"]) #mediana t.test(A$var1, A$var2, alternative = "two.sided/greater/less", mu = media, paired = TRUE conf.level = 0.05) #test due campioni confidenza 0.05 (e due variabili non indipendenti il valore medio di una distribuzione si avvicina ad un valore di riferimento?)

t.test(A$var1, alternative = "two.sided/greater/less", mu = media, conf.level = 0.05) #test un campione confidenza 0.05

binom.test(c(N successi, N fallimenti), alternative = "two.sided/greater/less", ipotesi nulla = value)

binom.test(vettore lunghezza 2(numero successi e fallimenti), conf.level=0.95)
c(5,95) -> vettore lungo 2, 5 successi e 95 fallimenti
binom.test(c(5,95), conf.level = 0.95)

ATTENZIONE: PARTE POCO CHIARA, VAI A VEDERE GLI ESEMPI

###########
#distribuzione-> media, mediana
media>mediana -> asimmetrica con valori più piccoli
mediana>media -> asimmetrica con valori più grandi
media = mediana -> all'incirca è simmetrica
varianza->sd(dataset)^2

#PER CONFRONTARE ELEMENTI IN MANIERA QUALITATIVA

Il trattamento insetticida A sembra eliminare gli insetti in maniera simile al trattamento C
boxplot(count~spray, data = InsectSprays)

#MEDIA SOLO SU UN ELEMENTO

newdataA <- InsectSprays$count[InsectSprays$spray == "A"]
mean(newdataA)

datiWithoutB = subset(dati, dati$group != "2") #eliminare elemento

table(InsectSprays$spray, useNA="always") #per vedere tutti gli elementi con somme

prima boxplot, poi abline(h=valore rischiesto, col = "red") #outlier

#rinomino
A<-InsectSprays$count[InsectSprays$spray=='A']
sum(table(A)) o length(A) #somma le categorie

#contare elementi nulli
na<-InsectSprays$count[InsectSprays$spray=='na'] #per elementi vuoti

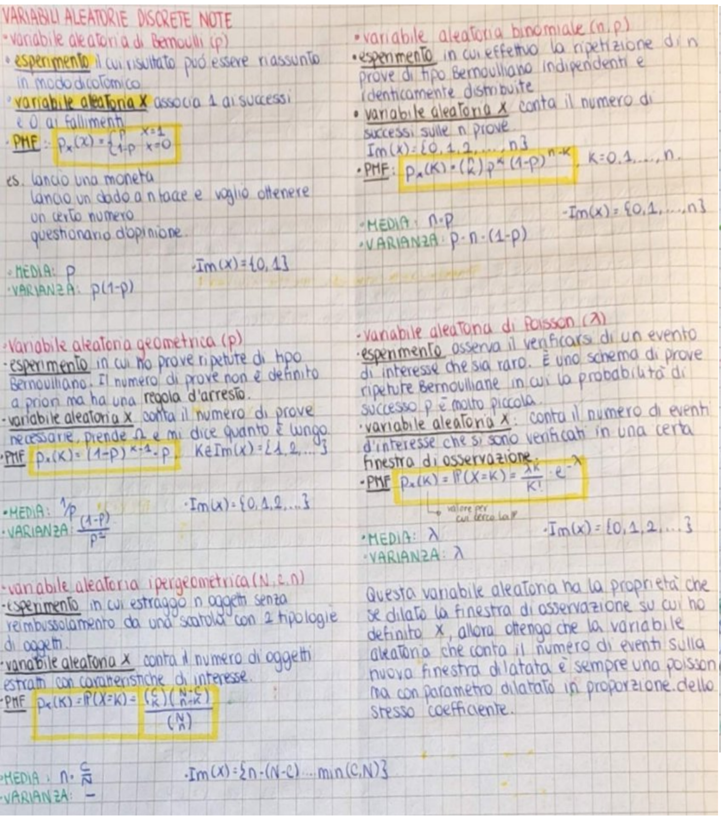
length(na)
oppure
table(InsectSprays$spray.useNa=="always")
#MEDIA CAMPIONARIA, ricordarsi il na.rm=true

#ipotesi alternativa
- pvalue > significatività
a favore dell'abbandono ipotesi alternativa e a favore ipotesi nulla [o nessuna delle precedenti o stessa media o differenza delle medie = 0 ->risposte esercitazioni]
-p value < significatività a favore ipotesi alternativa e all'abbandono ipotesi nulla [o gruppo 1 fa meno del gruppo 2?] [media x != media y]

se osservazioni sono <30 devi ipotizzare la normalità, se sono >30 non aggiungo altro
#####################################

# #Probabilità

P(A ∪ B) = P(A) + P(B) − P(A ∩ B) (intersezione può essere uguale a 0 (esce 1 e esce 2 non hanno intersezioni))
P(A ∩ B) = P(A) * P(B)  #eventi **indipendenti**
P(A ∩ B) = P(A) * P(B|A) = P(B) * P(A|B) #eventi **dipendenti**
P($A^c$∩ $B^c$) = 1 - P(A∪B)
P(A) = P(A|B) * P(B) + P(A|$B^C$) * P($B^C$)
P(A|B) = P(B ∩ A)/P(A) #Bayes
$P\,(A \cup B)^c = P(A^C \cap B^C)$ #DeMorgan
$P\,(A \cap B)^c = P(A^C) \cup P(B^C)$#DeMorgan


VARIABILI ALEATORIE DISCRETE NOTE
- variabile aleatoria di Bernoulli (p)
- esperimento il cui risultato può essere riassunto in modo dicotomico
- variabile aleatoria X associa 1 ai successi e 0 ai fallimenti
- PMF: $p_x(x) = \begin{cases} p & x=1 \\ 1-p & x=0 \end{cases}$

es. lancio una moneta
lancio un dado a n facce e voglio ottenere un certo numero
questionario d'opinione.

- MEDIA: p
- VARIANZA: p(1-p)
- Im(x) = {0, 1}

- variabile aleatoria geometrica (p)
- esperimento in cui ho prove ripetute di tipo Bernoulliano. Il numero di prove non è definito a priori ma ha una regola d'arresto.
- variabile aleatoria X conta il numero di prove necessarie, prende Ω e mi dice quanto è lungo.
- PMF $p_x(K) = (1-p)^{K-1} \cdot p$, K∈Im(x) = {1, 2, ...}

- MEDIA: 1/p
- VARIANZA: $\frac{(1-p)}{p^2}$
- Im(x) = {0, 1, 2, ...}

- variabile aleatoria ipergeometrica (N, C, n)
- esperimento in cui estraggo n oggetti senza reimbussolamento da una scatola con 2 tipologie di oggetti.
- variabile aleatoria X conta il numero di oggetti estratti con caratteristiche di interesse.
- PMF $p_x(K) = \mathbb{P}(X=K) = \frac{\binom{C}{K}\binom{N-C}{n-K}}{\binom{N}{n}}$

- MEDIA: $n \cdot \frac{C}{N}$
- VARIANZA: –
- Im(x) = {n-(N-C) ... min(C,N)}

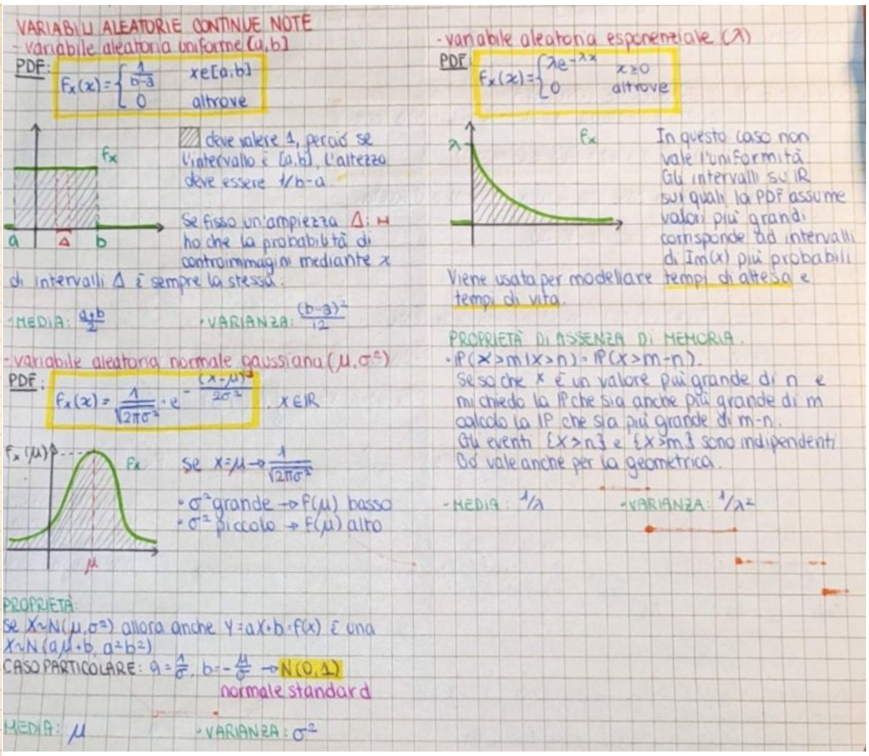
- variabile aleatoria binomiale (n, p)
- esperimento in cui effettuo la ripetizione di n prove di tipo Bernoulliano indipendenti e identicamente distribuite
- variabile aleatoria X conta il numero di successi sulle n prove.
Im(x) = {0, 1, 2, ..., n}
- PMF: $p_x(K) = \binom{n}{k} p^K (1-p)^{n-K}$, K = 0, 1, ..., n.

- MEDIA: n·p
- VARIANZA: p·n·(1-p)
- Im(x) = {0, 1, ..., n}

- variabile aleatoria di Poisson (λ)
- esperimento osserva il verificarsi di un evento di interesse che sia raro. È uno schema di prove ripetute Bernoulliane in cui la probabilità di successo p è molto piccola.
- variabile aleatoria X: conta il numero di eventi d'interesse che si sono verificati in una certa finestra di osservazione.
- PMF $p_x(K) = \mathbb{P}(X=K) = \frac{\lambda^K}{K!} \cdot e^{-\lambda}$
valore per cui cerco la P

- MEDIA: λ
- VARIANZA: λ
- Im(x) = {0, 1, 2, ...}

Questa variabile aleatoria ha la proprietà che se dilato la finestra di osservazione su cui ho definito X, allora ottengo che la variabile aleatoria che conta il numero di eventi sulla nuova finestra dilatata è sempre una poisson ma con parametro dilatato in proporzione dello stesso coefficiente.


VARIABILI ALEATORIE CONTINUE NOTE
- variabile aleatoria uniforme [a, b]
PDF: $f_x(x) = \begin{cases} \frac{1}{b-a} & x \in [a,b] \\ 0 & \text{altrove} \end{cases}$



deve valere 1, perciò se l'intervallo è [a, b], l'altezza deve essere 1/b-a

Se fisso un'ampiezza Δ: H ho che la probabilità di controimmagini mediante x di intervalli Δ è sempre la stessa.

- MEDIA: $\frac{a+b}{2}$
- VARIANZA: $\frac{(b-a)^2}{12}$

- variabile aleatoria normale gaussiana (μ, σ²)
PDF: $f_x(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, x ∈ ℝ



se $x = \mu \rightarrow \frac{1}{\sqrt{2\pi\sigma^2}}$
- σ² grande → f(μ) basso
- σ² piccolo → f(μ) alto

PROPRIETÀ:
se X~N(μ, σ²) allora anche Y = aX+b = f(x) è una X~N(aμ+b, a²b²)
CASO PARTICOLARE: $a = \frac{1}{\sigma}$, $b = -\frac{\mu}{\sigma}$ → N(0, 1)
normale standard

MEDIA: μ
- VARIANZA: σ²

- variabile aleatoria esponenziale (λ)
PDF: $f_x(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{altrove} \end{cases}$



In questo caso non vale l'uniformità. Gli intervalli su ℝ sui quali la PDF assume valori più grandi corrisponde ad intervalli di Im(x) più probabili

Viene usata per modellare tempi di attesa e tempi di vita.

PROPRIETÀ DI ASSENZA DI MEMORIA.
- ℙ(X>m | X>n) = ℙ(X>m-n).
Se so che x è un valore più grande di n e mi chiedo la ℙ che sia anche più grande di m calcolo la ℙ che sia più grande di m-n.
Gli eventi {X>n} e {X>m} sono indipendenti
Ciò vale anche per la geometrica.

- MEDIA: 1/λ
- VARIANZA: 1/λ²

E(X^2) = E(X)^2 + Var(X)
E(X^2-n) = E(X)^2 + Var(X) – n
Var(aX+n)= Var(X)^2
StDev(aX+n) = sqrt[(Var(X))^2]

MEDIA

$\mathbb{E}(k) = k$

$\mathbb{E}(kX) = k \cdot \mathbb{E}(X)$

$\mathbb{E}(kX + k) = k \cdot \mathbb{E}(X) + k$

-> k -> Costante
X -> Variabile Aleatoria

VARIANZA

$Var(k) = 0$

$Var(aX) = a^2 \cdot X$

$Var(aX + b) = a^2 \cdot X$

-> k, a, b = Costanti
X = variabile Aleatoria

DEVIAZIONE STANDARD

$SD(a) = 0$

$SD(aX) = a \cdot \sqrt{X}$

$SD(aX + b) = a \cdot \sqrt{X}$

-> a, b = Costanti
X = Variabile Aleatoria

# #t.test e binom.test: esempi

**9.10** The variable `sat.m` in the data set `stud.recs` (UsingR) 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 significance level?

$H_0: \mu = 500$
$H_1: \mu \neq 500$

```
t.test(stud.recs$sat.m,alternative = "two.sided",mu = 500)
```

One Sample t-test
data: stud.recs$sat.m
t = -2.5731, df = 159, p-value = 0.01099
alternative hypothesis: true mean is not equal to 500
95 percent confidence interval:
475.1437 496.7313
sample estimates:
mean of x
485.937

0.05>0.01099? sì! Allora rifiuto ipotesi nulla in favore dell'alternativa
Two.sided->mi chiede se è diverso, non < o >
475.1437 496.7313 sono tutti valori plausibili, infatti 500 non c'è->alternativa

on base percentage is 0.550 against a two sided alternative.

$H_0: \mu = 98.6$
$H_1: \mu < 98.6$

**9.14** The data set `normtemp` (UsingR) contains measurements of 130 healthy, randomly selected individuals. The variable `temperature` contains normal body temperature. Does the data appear to come from a normal distribution? If so, perform a *t*-test to see if the commonly assumed value of 98.6°F is correct. (Studies have suggested that 98.2°F is actually more accurate.)

```
t.test(normtemp$temperature, alternative = "less", mu = 98.6)
```

One Sample t-test
data: normtemp$temperature
t = -5.4548, df = 129, p-value = 1.205e-07
alternative hypothesis: true mean is less than 98.6
95 percent confidence interval:
-Inf 98.35577
sample estimates:
mean of x
98.24923

0.05>1.205e-07? sì! 0.01>? sì! rifiuto ipotesi nulla in favore dell'alternativa
Less-> mi dice che la nulla è 98.6, ma si pensa che possa essere 98.2 (98.2<98.6)->less
-Inf 98.35577 sono tutti valori plausibili, infatti 98.6 non c'è->alternativa

$H_0: p = 0.75$ $H_1: p > 0.75$

**9.3** A new drug therapy is tested. Of 50 patients in the study, 40 had no recurrence in their illness after 18 months. With no drug therapy, the expected percentage of no recurrence would have been 75%. Does the data support the hypothesis that this percentage has increased? What is the *p*-value?

```
binom.test(c(40,10),alternative = "greater", p = 0.75)
```

Exact binomial test
data: c(40, 10)
number of successes = 40, number of trials = 50, p-value = 0.2622
alternative hypothesis: true probability of success is greater than 0.75
95 percent confidence interval:
0.6844039 1.0000000
sample estimates:
probability of success
0.8

0.05>0.2622? no! 0.01>? no! Rifiuto l'ipotesi alternativa in favore della nulla
Greater->"percentage has increased?"

**9.7** Historically, a car from a given company has a 10% chance of having a significant mechanical problem during its warranty period. A new model of the car is being sold. Of the first 25,000 sold, 2,700 have had an issue. Perform a test of significance to see whether the proportion of all of these new cars that will have a problem is more than 10%. What is the $p$-value?

$H_0: \; p = 0.1$

$H_1: \; p > 0.1$

```
binom.test(c(2700,(25000-2700)),alternative = "greater", p = 0.1)
```

Exact binomial test
data: c(2700, (25000 - 2700))
number of successes = 2700, number of trials = 25000, p-value = 1.588e-05
alternative hypothesis: true probability of success is greater than 0.1
95 percent confidence interval:
0.1047849 1.0000000
sample estimates:
probability of success
0.108

0.05>1.588e-05? Sì! 0.01>? Sì! Rifiuto l'ipotesi nulla a favore dell'alternativa
Greater->"more than"

PAIRED

**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.

$Z = X - Y$

DAD

$H_0: \; \mu_Z = 0$

$H_1: \; \mu_Z > 0$

```
t.test(babies$dage, babies$age, alternative = "greater", mu = 0, paired = TRUE)
```

Paired t-test
data: babies$dage and babies$age
t = 17.392, df = 1235, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
3.047148 Inf
sample estimates:
mean of the differences
3.365696

Paired=TRUE!->si riferiscono entrambe ad un bambino. Misuro il bimbo e dal bimbo ricavo età mamma e papà
Motivo per cui uso due variabili (babies$dage, babies$age)
Greater->età papà>età mamma
0.05> 2.2e-16? Sì! 0.01> 2.2e-16? Sì->abbandonare ipotesi nulla a favore dell'alternativa

**9.36** The Galton (HistData) 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. What problems are there with

$Z = X - Y$

$H_0: \; \mu_Z = 0$

$H_1: \; \mu_Z \neq 0$

```
t.test(Galton$child, Galton$parent, alternative = "two.sided", mu = 0, paired =
TRUE)
```

Paired t-test
data: Galton$child and Galton$parent
t = -2.8789, df = 927, p-value = 0.004082
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.36949983 -0.06993982
sample estimates:
mean of the differences
-0.2197198

Two.sided->mi interessa la differenza, non se è > o <
Paired=true ->si riferiscono entrambe ad un unico bambino
0.05>0.004082? sì! 0.01>? sì! Abbandonare ipotesi nulla a favore dell'alternativa