Estadística

$\sum vf$

Conceptes generals d'estadística

Població és el conjunt format per tots els elements que són objecte d'estudi.

Mostra és un subconjunt representatiu de la població, que serveix per inferir característiques de tota la població.

Individu és cadascun dels elements de la pobació o de la mostra.

Variable estadística(X) és cadascuna de les característiques de la població que volem estudiar.

Variable qualitativa es refereix a una característica que no pot ser mesurada amb números.

Varibla quantitativa és la que pren valors numèrics.
- **Variable discreta** sols pren valors aïllats.
- **Variable contínua** pot prendre qualsevol valor entre dos valors donats de la variable.

Dada estadística(X_i) cadascun dels valors que pren la variable estadística.

Organització i ordenació de les dades. Taula de freqüències. Gràfics estadístics

Distribució de freqüències. És una ordenació en forma de taula de les dades estadístiques, assignant-li a cada dada la seua freqüència. S'empra en variables estadístiques discretes.

Freqüència absoluta(f_i) nombre de vegades que es dóna una dada estadística. $f_1+f_2+f_3+\cdots+f_n=N$ o també $\sum_{i=1}^{n} f_i=N$, on N representa el número total de dades.

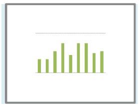
Freqüència relativa(n_i) és el valor del quocient $n_i=\frac{f_i}{N}$. Evidentment.$\sum_{i=1}^{n} n_i=1$.

Freqüència acumulada(F_i) és la suma de les freqüències absolutes de totes les dades iguals o inferiors a x_i.

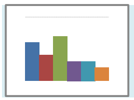
Freqüència relativa acumulada(N_i) és el valor del quocient $N_i=\frac{F_i}{N}$.

Distribució de freqüències agrupades. És una ordenació en forma de taula de les dades agrupades. S'utilitza quan la variable és contínua o quan pren un gran número de valors. Els valors s'agrupen en intervals que tenen la mateixa amplitud anomenats ***classes***. A cada classe se li assigna la seua freqüència. Cada ***classe*** és un interval tancat per l'esquerre i obert per la dreta$[L_{i-1}, L_i[$. És convenient que hi haja entre 5 i 15 intervals o classes. **Marca de classe** és el punt mitjà de cada interval i és el valor que representa a tot l'interval per calcular els paràmetres estadístics. $X_i=\frac{L_{i-1}+L_i}{2}$

Diagrama de barres. S'utilitza en variables estadístiques discretes. Les dades es representen a l'eix X i en cada dada es representa una barra d'altura igual a la seua freqüència.

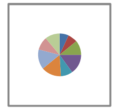


Histograma de freqüències. S'utilitza en variables estadístiques contínues. En cada interval classe es representa un rectangle que té d'altura la freqüència de cada classe.



Polígon de freqüències. S'utilitza també en variables estadístiques contínues. Es construeix unint els punts mitjans de la base superior de cada rectangle de l'histograma.

Diagrama de sectors. Com el seu nom indica, cada dada està representada en un sector circular, l'angle del qual és proporcional a la freqüència de la dada.



Paràmetres estadístics

Univariant

Mesures de centralització

Moda(Mo). Valor de la variable de major freqüència.

Mediana(Me). Valor que ocupa el lloc central de la distribució.

Mitjana aritmètica($\bar{x}$) $\bar{x}=\dfrac{\sum x_i\cdot f_i}{n}$

Mesures de dispersió

Recorregut o Rang (R). És la diferència entre el major valor de les dades i el menor valor de les dades.

Desviació mitjana(DM)	$DM=\dfrac{\sum f_i\cdot\lvert x_i-\bar{x}\rvert}{N}$
Variància(σ_x^2)	$\sigma_x^2=\dfrac{\sum (x_i-\bar{x})^2\cdot f_i}{n}=\dfrac{\sum x_i^2\cdot f_i}{n}-\bar{x}^2$
Desviació Típica(σ_x)	$\sigma_x=+\sqrt{\sigma_x^2}$
Coeficient de Variació(CV)	$CV=\dfrac{\sigma_x}{\bar{x}}$

Mesures de posició. Centils. Quartils

El percentil k, P_k, és el valor que deixa per sota d'ell el k % de la població.

Quartil 1, Q_1, és el valor que deixa per sota d'ell el 25 % de la població. $Q_1=P_{25}$

Quartil 2, Q_2, és el valor que deixa per sota d'ell el 50 % de la població. $Q_2=P_{50}$=Mediana.

Quartil 3, Q_3, és el valor que deixa per sota d'ell el 75 % de la població. $Q_3=P_{75}$

Recorregut interquartílic=Q_3-Q_1.

Bivariant

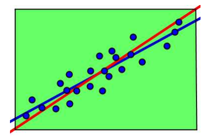
Mesures. Núvol de punts

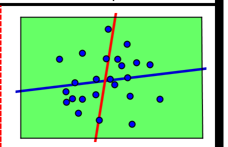
Covariança(σ_{xy})	$\sigma_{xy}=\dfrac{\sum\sum (x_i-\bar{x})\cdot(y_j-\bar{y})\cdot f_{ij}}{n}=\dfrac{\sum\sum x_i\cdot y_j\cdot f_{ij}}{n}-\bar{x}\cdot\bar{y}$
Coeficient de correlació(r)	$r=\dfrac{\sigma_{xy}}{\sigma_x\cdot\sigma_y}$ $-1\le r\le 1$
Recta de regressió de y sobre x	$y-\bar{y}=\dfrac{\sigma_{xy}}{\sigma_x^2}\cdot(x-\bar{x})$
Recta de regressió de x sobre y	$x-\bar{x}=\dfrac{\sigma_{xy}}{\sigma_y^2}\cdot(y-\bar{y})$

Si $\lvert r\rvert\cong 1$ l'angle que formen les rectes és quasi **0º**

Si $\lvert r\rvert\cong 0$ l'angle que formen les rectes és quasi **90º**

Les dues rectes de regressió es tallen en el punt $(\bar{x},\bar{y})$





Distribucions

Binomial de paràmetres n i p. $B(n,p)$

Experiència dicotòmica. En ella només considerem dos resultats possibles: o bé ocorre A, o bé: ocorre el seu contrari $\bar{A}$.
Si $P(A)=p$, $P(\bar{A})=1-p=q$. L'esdeveniment A s'anomena èxit.

La Binomial $B(n,p)$
- Es repeteix n vegades la mateixa experiència dicotòmica.
- La probabilitat d'èxit, p, és la mateixa cada vegada, $P(A)=p$
- Ens preguntem pel nombre d'èxits, X.
- Els valors que pot prendre la variable X són *0,1,2,…,n* (variable aleatòria discreta).
- La variable X segueix una distribució binomial de paràmetres n i p. Es denota $X\sim B(n,p)$.

Probabilitats en una distribució binomial
- La probabilitat per a cada valor de X, des de 0 fins a n, ve donada per: $P(X=x)=f(x)=\binom{n}{x}\cdot p^x\cdot(1-p)^{n-x}$
- Per calcular aquest valor, existeix una taula (Taula de La Binomial), encara que quasi sempre, si es pot, s'aproxima a una normal.
- El valors de la mitjana aritmètica i la variància són: $E[X]=\mu=n\cdot p$ i $Var[X]=\sigma_X^2=n\cdot p\cdot(1-p)=n\cdot p\cdot q$

Ajustament d'un conjunt de dades a una distribució binomial. Donat un conjunt de dades obtingudes experimentalment, volem saber si és raonable acceptar la hipòtesi que les dades provinguen d'una distribució binomial. En aquest cas, calculem la mitjana aritmètica de les nostres dades, en igualar-la a la mitjana teòrica, $\mu=np$, com que n és conegut, obtenim p. Això ens permet obtenir les probabilitats teòriques de cada valor de la variable i comparar-les amb les obtingudes experimentalment, per valorar les diferències. Depenent d'aquestes, conclourem si les nostres dades s'ajusten o no a una distribució binomial.

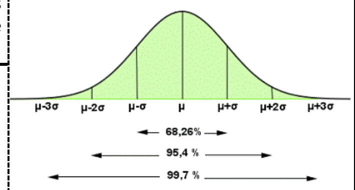
Aproximació a la normal. En general, quan el paràmetre n és molt gran, la variable binomial $B(n,p)$ pot ser aproximada per la distribució normal, de mitjana μ i desviació típica σ, essent μ i σ els paràmetres corresponents de la variable binomial inicial.

Si n és molt gran i $\begin{cases} n\cdot p>5 \\ n\cdot p\cdot q>5\end{cases}\longrightarrow X\approx N(\mu,\sigma)$

Normal de paràmetres μ i σ. $N(\mu,\sigma)$

Identificació. Moltes variables es distribueixen *normalment*. Exemple: talla, pes, de persones; notes en un examen; nombre de visitants a un museu; la durada en hores de bombetes; … La variable aleatòria és contínua, és a dir, entre dos valors donats, pot prendre qualsevol valor intermedi.

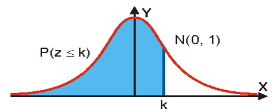
Probabilitats en una distribució normal
- Per calcular $P(a\le x\le b)$, obtindrem l'àrea davall la corba a l'interval $[a,b]$.
(Les probabilitats puntuals són nul·les a les variables aleatòries contínues.)
- La funció de densitat és $f(x)=\frac{1}{\sigma\sqrt{2\pi}}\cdot e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, x\in R$. $P(x\le k)=\int_{-\infty}^{k} f(t)dt$ i $\int_{-\infty}^{\infty} f(t)\,dt=1$
- D'aquesta funció no existeix primitiva. Aleshores per obtenir probabilitats fem ús d'una taula.



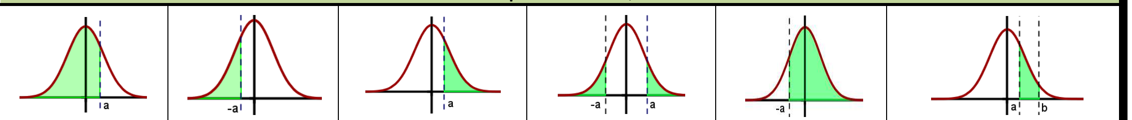
Normal tipificada. Siga $X\sim N(\mu,\sigma)$ una variable normal. Les probabilitats en dues normals qualssevol es reparteixen de forma anàloga. Per obtenir $P(x\le k)$, cal *tipificar* la variable.

Per tipificar la variable X, considerem: $Z=\frac{X-\mu}{\sigma}\sim N(0,1)$.

Per poder obtenir la probabilitat $P(x\le k)=P\left(z\le\frac{x-\mu}{\sigma}\right)$, utilitzem la ***taula de la normal tipificada*** (disponible en: llibres de text, Internet, …)



Càlcul de probabilitats. $a,b\ge 0$



$P(Z\le a)$	$P(Z<-a)=1-P(Z\le a)$	$P(Z>a)=1-P(Z\le a)$	$P(\lvert Z\rvert>a)=2\,P(Z>a)$	$P(Z>-a)=P(Z\le a)$	$P(a\le Z\le b)=P(Z\le b)-P(Z\le a)$