1 - Estadística descriptiva

Estadística descriptiva

- Describe una muestra sin extrapolar datos al resto de la población.
- **Población:** conjunto de individuos que constituyen el objetivo del estudio
- **Variable:** rasgo medible de los elementos de la población.
    - Variable cualitativas: Valores no numéricos
        - Nominales: Los valores no tienen orden (sexo, color de ojos)
        - Ordinales: Orde subyacente entre categorías (nivel de estudios)
    - Variables cuantitativas: Toman valores numéricos
        - Discretas: Valores enteros y separados
        - Continuas: Valores reales (incluye racionales)
- **Muestra:** subconjunto de la población para el que se conocen los valores de las variables a analizar. Debe ser representativa.
- **Individuo:** cada uno de los elementos de la muestra

Tablas de frecuencias

- Utilizadas para representar la información de una muestra tamaño n
- **Clase (c_i):** Cada uno de los valores que puede tomar una variable.
- **Frecuencia absoluta (n_i):** Número de individuos en la clase c_i. $0<=n_i<=n$
- **Frecuencia relativa (f_i):** = $\frac{n_i}{n}$ $0<=f_i<=1$

c_i	n_i	f_i	N_i	F_i
Baixa	2	0.2	2	0.2
Media	5	0.5	7	0.7
Alta	3	0.3	10	1.0

- **Frecuencia absoluta acumulada (N_i):** Número de individuos en c_i o en valores anteriores. $N_k = n$
- **Frecuencia relativa acumulada (F_i)**: $\frac{N_i}{n}$. $F_k = 1$
- Para unha variable cualitativa nominal non se inclúen valores absolutos.

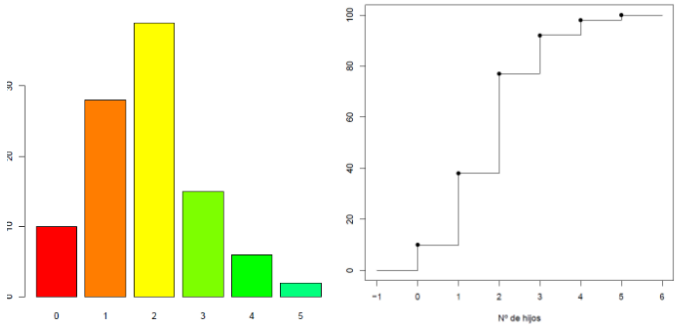
Tablas de frecuencias (continua)

- Denominados intervalos de clase: $[e_i, e_{i+1})$[1] sendo o punto medio a marca de clase (c_i)
- Se suelen tomar como número de intervalos $\sqrt{n}$ aproximado.
- Suelen ser todos de igual longitud y contiguos. La longitud se calcula aproximando hacia arriba para evitar excluir datos.

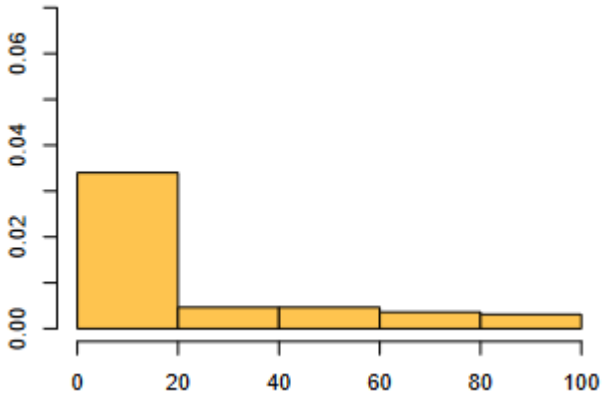
e_i	n_i	f_i	N_i	F_i
[2.5, 3.1)	1	0.1	1	0.1
[3.1, 3.7)	4	0.4	5	0.5
[3.7, 4.3)	5	0.5	10	1.0

Representacións gráficas

- Variables cualitativas o cuantitativas discretas: **diagrama de barras** o sectores, y diagrama de frecuencias acumuladdas



- Variables cuantitativas continuas: **histograma** o diagrama de caja



[1] el último intervalo, como excepción, puede ser cerrado por la derecha: $[e_{k-1}, e_k]$

Medidas de posición

- Medidas de posición: indican a posición que ocupa a mostra. Poden ser de tendencia central (indican o centro da mostra, media, moda) ou non central.
- Media: $\frac{\sum_{i=1}^{n} x_i}{n}$. En caso de intervalos se puede realizar elemento por elemento o intervalo por intervalo.
- Media truncada: elimínanse un porcentaxe dos datos máis extremos
    - Media recortada: os datos extremos son reemprazados polo punto de corte
- Cuantiles: Medidas non centrais. Exemplo: O cuantil $q_{0.45}$ será o dato que deixa á súa esquerda, como mínimo, o 45% dos datos.
    - En un conxunto de 10 datos, $q_{0.45}$ e $q_{0.4}$ serán ambos o 5º dato.
    - Os cuartiles (Q_1, Q_2, Q_3) son cuantiles de orden (0.25, 0.5, 0.75). Existen tamén deciles e percentiles.
    - A mediana é o segundo cuartil. Se hai un dato impar de datos, é a media dos dous centrais.

Medidas de dispersión

- Recorrido: diferencia entre o máximo e mínimo
    - Recorrido intercuartílico: diferenza entre Q_3 e Q_1
- Varianza (**s^2**): $\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2$. A desviación típica é a raíz da varianza, s
    - Siempre positiva.
    - Si se modifican los datos haciendo **$y_i=ax_i+b$**, se cumplirá que **$s_y^2=a^2*s_x^2$**
- Cuasivarianza(**s_{n-1}**): Igual, pero tomando n-1. (ainda asi con todos os datos)
- Coeficiente de variación: CV: s / media. Util porque a magnitude da desviación típica é maior en datos de media maior.

Medidas de forma

- Coeficiente de asimetría: 0 se é simétrica respecto á media. Valores positivos se hai máis datos por encima da media, negativos se hai máis por debaixo.

Figura 5. Interpretación del coeficiente de asimetría.

  - $$\gamma_F = \frac{1}{s^3}\frac{(x_1-\overline{x})^3+\ldots+(x_n-\overline{x})^3}{n} = \frac{1}{s^3}\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^3$$

- Coeficiente de curtosis: mide o grao de apuntamento da distribución
  - Nunha curva normal é de 3. Maior de 3 é leptocúrtica, menor é platicúrtica.

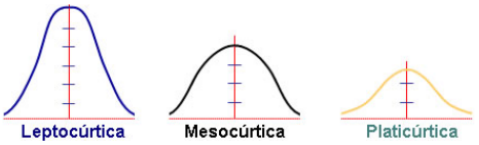
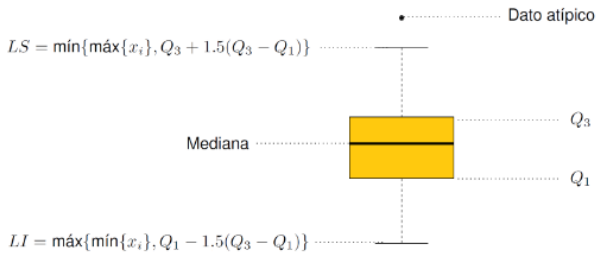


Figura 6. Interpretación del coeficiente de asimetría.

  - $$\gamma_C = \frac{1}{s^4}\frac{(x_1-\overline{x})^4+\ldots+(x_n-\overline{x})^4}{n} = \frac{1}{s^4}\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^4$$

Diagrama de caixa

- Representa variables cuantitativas continuas.
- A caixa está delimitada polos cuartiles, sendo a raia do medio o segundo cuartil (a mediana). A altura da caixa é o rango intercuartílico, e contén un 50% dos datos.
- Os bigotes (LS e LI) calcúlanse coa fórmula do gráfico.
- Se quedan datos fóra dos bigotes, considérense atípicos:
  - Defínense outras cotas inferiores e superiores que sexan Q_1-3RI e Q_3+3RI
  - Se están entre bigotes e a cota especificada, son moderados (*)
  - Se están fóra das cotas, son atípicos extremos (º)
- 

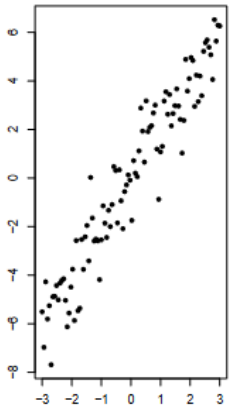
Estadística descriptiva bivariante

$X\backslash Y$	rubio	pelirrojo	castaño	oscuro	negro
claros	688	116	584	188	4
azules	326	38	241	110	3
castaños	343	84	909	412	26
oscuros	98	48	403	681	85

$X\backslash Y$	y_1	$\cdots$	y_j	$\cdots$	y_l	
x_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1l}	$n_{1\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{il}	$n_{i\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_k	n_{k1}	$\cdots$	n_{kj}	$\cdots$	n_{kl}	$n_{k\bullet}$
	$n_{\bullet 1}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet l}$	n

- Permite analizar simultáneamente 2 o máis variables
- Se son categóricas ou discretas emprégase a tabla de continxencia
- Con variables continuas realizase unha recta de regresión
- Se temos variables de distinto tipo, por exemplo unha categórica e unha continua, creamos distintos grupos de estudo segundo a variable categórica

Diagrama de dispersión



- Represéntase o conxunto de individuos como puntos nun plano onde os eixos x e y son as dúas variables a medir
  - Permite describir a relación entre variables
- **Covarianza:** $$S_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{xy}.$$
  - O signo describe a relación entre as variables: se é directa $S_{xy}>0$, se é inversa $S_{xy}<0$, se non hai relacion é proxima a 0
- Para evitar que a covarianza esté influida pola media dos datos, calcúlase o **coeficiente de correlación lineal:** $$r_{xy} = \frac{S_{xy}}{s_x s_y},$$
  - Dise que a relación é significativa se $|r_{xy}| \geq 0.7$, e que existe algunha se é maior de 0.3

Recta de regresión

- Consiste en calcular a recta que mellor representa a mostra. A recta será do tipo Y = a + bX + ε, sendo ε o erro cometido.
- **Método de mínimos cuadrados:** Consiste en minimizar a suma dos cadrados dos residuos.
  - Os residuos son a diferenza entre o valor y de cada punto é o da recta.
  - $\sum_{i=1}^{n}(y_i - a - bx_i)^2$. A partir de esto: $\widehat{b} = \frac{S_{xy}}{s_x^2}, \quad \widehat{a} = \overline{y} - \widehat{b}\overline{x},$

    - Os valores deben ser **homcedasticos:** A variabilidade dos residuos debe ser constante.

- **Coeficiente de regresión:** Parámetro **b** de la recta de regresión.
    - Si b>0, aumentar los valores de X también aumenta Y.
- **Coeficiente de determinación (R^2):** Mide la proporción de variabilidad de Y que explica X. Entre 0 y 1.
    - Si R es próximo a 1, la recta es un buen ajuste de la muestra.
    - $$R^2 = r_{xy}^2 = \frac{S_{xy}^2}{s_x^2 s_y^2}.$$

2 - Probabilidade

Experimento aleatorio

- Los posibles resultados son conocidos, pero es imprevisible cuál será.
- El experimento se puede repetir en las mismas condiciones con distintos resultados
- **Determinista:** no aleatorio (mismo resultado en mismas condiciones)

Sistemas exhaustivos e completos

- **Sistema:** familia de sucesos en un experimento
- **Sistema exhaustivo de sucesos:** Si su unión cubre el espacio muestral
- **Sistema completo de sucesos:** Si es exhaustivo y las intersecciones entre elementos son 0 para todos los pares de elementos.

Probabilidad

- **Definición de Laplace:** casos favorables/casos posibles
- **Definición frecuentista:** repetimos n veces un experimento, y el suceso ocurre n_A veces.
    - La frecuencia relativa del suceso será n_A/n. para A elevado, será una aproximación de la probabilidad

Definición axiomática

- Sea el par (Ω, A) un espacio probabilizable. Se dice que P es una probabilidad sobre (Ω, A) si cumple:
    - $P(\Omega)=1$
    - La probabilidad de la unión de sucesos disjuntos es la suma de las probabilidades de cada uno
    - La probabilidad de todo suceso está entre 0 y 1.

- A partir de los axiomas, se pueden obtener:
  - P(∅) = 0
  - $P(A^c)$ = 1-P(A)
  - Si A C B, P(A)<=P(B), y P(B\A) = P(B)-P(A)
  - P(AUB) = P(A) + P(B) - P(AnB)

Probabilidad condicionada

- Probabilidad de A si ha ocurrido B: $\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$
- Sucesos **independientes:** Si P(AnB) = P(A)*P(B), por lo que P(A|B)=P(A).

Regla del producto

- P(AnB) = P(A) * P(B|A)
- P(A ∩ B ∩ C ∩ D) = P(A)*P(B|A)*P(C|A ∩ B)*P(D|A ∩ B ∩ C).
- $\mathbb{P}(\cap_{i=1}^{n} A_i) = \mathbb{P}(A_1)\mathbb{P}(A_2|A_1)\mathbb{P}(A_3|A_1 \cap A_2)\ldots\mathbb{P}(A_n|\cap_{i=1}^{n-1} A_i)$

Teorema de Probabilidades totales

- Si $\{A_1, \ldots, A_n\}$ es un conjunto completo de sucesos de probabilidad no nula, y B es un suceso cualquiera:
  - $\mathbb{P}(B) = \sum_{i=1}^{n} \mathbb{P}(B|A_i)\mathbb{P}(A_i)$ (suma de intersecciones de B con cada A_i)

Teorema de Bayes

- Si $\{A_1, \ldots, A_n\}$ es un conjunto completo de sucesos de probabilidad no nula, y B€A es otro suceso:
- $$\mathbb{P}(A_j|B) = \frac{\mathbb{P}(A_j \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A_j)\mathbb{P}(A_j)}{\sum_{i=1}^{n} \mathbb{P}(B|A_i)\mathbb{P}(A_i)}.$$

3 - Variables aleatorias discretas

Definición

- Dado un experimento aleatorio, unha **variable aleatoria X** é unha aplicación que asocia a cada elemento dun elemento muestral un número.
- Exemplo: no experimento de lanzar un dado, a variable será 'valor do dado'.
- $X : (\Omega, \mathcal{A}, \mathbb{P}) \longrightarrow (\mathbb{R}, \mathbb{B}, \mathbb{P}_X)$ Sendo **B** o conxunto de todos os intervalos de R.

Propiedades

- Se X é unha variable sobre $(\Omega, \mathcal{A}, \mathbb{P})$ e c é unha constante, c*X tamén é unha variable.
- Se X e Y son variables sobre $(\Omega, \mathcal{A}, \mathbb{P})$, X+Y e X*Y tamén.

Variable aleatoria discreta

- Unha **variable aleatoria discreta** é aquela que toma valores nun conxunto finito (ou infinito numerable).
- O conxunto de posibles valores que toma denomínase **soporte** (**Sop(X)={x_1, …, x_k})**
- O conxunto de probabilidades de cada suceso denomínase **masa de probabilidade** {p_1, …, p_k}, onde $p_i = P(X=x_i)$
    - A suma de todos os p_i debe ser 1.
    - Represéntase mediante un diagrama de barras.
    - Exemplo: nº de caras ao tirar dúas moedas: {0.25, 0.5, 0.25}
    - Menor valor ao lanzar 2 dados: {11/36, 9/36, 7/36, 5/36, 3/36, 1/36}
- As variables discretas pódense caracterizar polo seu soporte e as respectivas posibilidades.

Función de distribución

- A **función de distribución** de unha variable, continua ou discreta, asocia a cada número coa probabilidade de que X acade un valor menor ou igual ca ese número. **$F(x)=P_x(X<=x)$**
- **Propiedades:**
  - F(x)€ [0,1]
  - F(x) no es decreciente y es continua por la derecha
  - F(+inf)=1 y F(-inf)=0

Medidas características (discreta)

- **Esperanza matemática** (media): E(X)= μ = $\sum_{i=1}^{K} x_i p_i$. Sendo K o número de valores x_i posibles e p_i as súas respectivas probabilidades.
  - E(aX + b) = a*E(X) + b
  - E(X+Y) = E(X)+E(Y)
  - Exemplo (moedas): 1/4 *0 + 1*2*1 + ¼*2 = 1
- **Mediana** (M_e): Valor x que divide a idstribución en dúas metades iguais. $F(M_e)=0.5$
- **Varianza:** $\sigma^2 = \mathbb{V}\mathrm{ar}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2]$, sendo E(X) a esperanza matemática
  - **Desviación típica:** Raíz da varianza
  - Var(ax + b) = a^2Var(x)
  - Exemplo (moedas): $(0-1)^2$*¼ + (1-1)*½ + $(2-1)^2$*¼ = 0.5
- **CV:** Desviación típica / media
- **Moda:** Valor para el cual la masa de probabilidad es máximo

Variables tipificadas

- Si tenemos una variable X con media μ y desviación típica σ, podemos transformarla en una variable Y de media 0 y varianza 1.
- Para **estandarizar** una variable, se resta la media y divide por la desviación típica: $Y = \frac{X - \mu}{\sigma}$.

Independencia de variables

- Dos variables aleatorias X e Y son independientes si $\mathbf{F_{X,Y}(x,y) = F_X(x) * F_Y(y)}$ para todos x,y reales.
- En el caso discreto, se dice que X e Y son independientes si $\mathbf{P(X=x_i,Y=y_j)=P(X=x_i)*P(Y=y_i).}$
- En el caso continuo, son independientes si $f_{X,Y}(x,y) = f_X(x) * f_Y(y)$
- De ser independientes, se cumple que **E(XY)=E(X)*E(Y)**, y **Var(X+Y)=Var(X)+Var(Y)**

Experimento de Bernoulli

- Aquel que solo presenta dos posibles resultados.
- Llamaremos éxito a uno de los resultados y fracaso al otro. La probabilidad de éxito será p.
- A variable X será o resultado, que vale 1 con probabilidad p e 0 con 1-p
- E(X)=p, Var(x) = p*(1-p)
- Ejercicio: de los alumnos, un 20% se han estudiado el tema
    - Si asisten a clase 40 alumnos, el número medio que han estudiado es 40*0.2 = 8 alumnos
    - Si asisten 20, 20*0.2 = 4 alumnos

Distribución binomial

- Repetimos un experimento de bernoulli n veces e consideramos a variable X=nº de éxitos. $\mathbf{X \in Bi(n,p)}$ [2]
- A súa función masa de probabilidade será $\mathbf{P(X=x) = (n \mid x)\ p^x\ (1-p)^{n-x}}$
- E(X) = np
- Var(X) = np(1-p)
- **Propiedades:**
    - Para n=1, Ber(p) = Bi(1,p)
    - Si $X \in Bi(n_1, p)$, $Y \in Bi(n_2, p)$ independientes, entonces $X+Y \in Bi(n_1+n_2, p)$.

[2] Significa 'X sigue una distribución binomial de parámetros n y p'.

- Si X ∈ Bi(n, p) entonces X = $\sum_{i=1}^{n} X_i$, donde X_i ∈ Ber(p) independientes

Distribución geométrica

- Repetimos un experimento de Bernoulli e consideramos a variable X=nº de fracasos ata o primeiro éxito. **X ∈ G(p).**
- A súa función masa de probabilidade será **P(X=x) = (1-p)^x * p**
- E(X) = (1-p) / p
- Var(X) = (1-p) / p^2

Distribución binomial negativa

- Repetimos un experimento de Bernoulli e consideramos a variable X=nº de fracasos ata o éxito **n**. **X ∈ BN(n,p).**
- A súa función masa de probabilidade será **P(X=x) = (n+x-1 | x) * (1-p)^x * p^n**
- E(X) = n(1-p) / p
- Var(X) = n(1-p) / p^2
- Nota: G(X) equivale a BN(1,p)

Distribución de Poisson

- Un proceso de Poisson consiste en observar la aparición de sucesos en un soporte continuo, como el tiempo, cuando el número de sucesos en intervalo de tiempo, denotado por λ, se mantiene constante.
- Definimos X como el nº de sucesos en un intervalo fijado. **X ∈ Pois(λ)**
- La masa de probabilidad será P(X=x) = $\frac{e^{-\lambda}\lambda^x}{x!}$
- E(X) = Var(X) = λ
- A medida que aumenta el valor de λ, la variable se hace más simétrica.

Distribución Hipergeométrica

- Existe unha poboación de N elementos, dos cales k son de clase D e (N-k) son de clase D'. Tomamos unha mostra aleatoria de n elementos, sin reemplazamiento.
- Sea X la variable que indica el nº de elementos de clase D en la muestra. **X ∈ H(N, n, k).**
- O soporte será x ∈ {máx(0, n + k − N), mín(k, N)}
- $$\mathbb{P}(X = x) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}$$
- E(X) = $\frac{nk}{N}$ = np
    - Idéntico a una distribución binomial, pero la varianza es distinta. Debido a esto, si N es muy grande respecto a n, se puede aproximar por una distribución binomial.
- Var(X) = npq * $\frac{N-n}{N-1}$

Distribución uniforme discreta

- Si X toma valores {x_1, …, x_k} y todos tienen la misma probabilidad, se dice que X tiene una distribución **uniforme discreta, X ∈ U{x_1, …, x_k}**
- P(X=x) = 1/K para todos los x.
- E(x) = 1/K * $\sum_{i=1}^{K} x_i$,
- Var(x) = 1/K * $\sum_{i=1}^{K} (x_i - \mu)^2$

Resumen distribuciones discretas

Nombre	Definición	MOP	E(X)	Var(X)
Binomial Bi(n,p)	nº de éxitos tras repetir bernoulli n veces	$\binom{n}{x} p^x q^{n-x}$	np	np(1-p)
Geométrica G(p)	nº de fracasos hasta el primer éxito	$(1-p)^x p$	q/p	q/p^2
Binomial negativa BN(n,p)	nº de fracasos hasta el n-ésimo éxito	$\binom{n+x-1}{x} q^x p^n$	nq/p	nq/p^2
Poisson Pois(λ)	nº de sucesos en un intervalo λ	$\frac{e^{-\lambda}\lambda^x}{x!}$	λ	λ
Hiper - geométrica H(N,n,k)	con una población de N elementos, elementos de clase K en una muestra de tamaño n	$\frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}$	nk/N = np	$npq\frac{N-n}{N-1}$
Uniforme U{x_1,...x_k}	valores con igual probabilidad	1/K	1/K * $\sum_{i=1}^{K} x_i$	1/K* $\sum_{i=1}^{K} (x_i - \mu)^2$

Resumen distribuciones continuas

Nombre	Definición	f(x)	F(x)	E(X)	Var(X)
Uniforme U(a,b)	probabilidad constante dentro de un intervalo	$\begin{cases} \frac{1}{b-a} & x \in (a,b), \\ 0 & \text{en otro caso.} \end{cases}$	$\begin{cases} 0 & x < a, \\ \frac{x-a}{b-a} & a \le x < b \\ 1 & x \ge b. \end{cases}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Normal N(0, 1)	-	$\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$	área bajo la curva	-	-
Exponencial Exp(λ)	Tiempo promedio entre sucesos de un proceso de Poisson de pár.λ	$\begin{cases} \lambda e^{-\lambda x} & x > 0, \\ 0 & \text{en otro caso.} \end{cases}$	$1 - e^{-\lambda x}$	1/λ	$1/\lambda^2$
Gamma Γ(n,λ)	tiempo hasta la aparición del n-ésimo suceso	$\begin{cases} \frac{\lambda^n}{\Gamma(n)} e^{-\lambda x} x^{n-1} & x > 0, \\ 0 & \text{en otro caso.} \end{cases}$	-	n/λ	n/λ^2

4 - Variables aleatorias continuas

Variable aleatoria continua

- Una variable aleatoria continua es aquella que toma valores en un **intervalo** de la recta real.
- La función de distribución **F(x)** se define de misma forma que con una variable discreta: la prob. de que la variable sea menor o igual que x.
- La función de densidad **f(x)** funciona como generalización de la masa de probabilidad[3] para el caso continuo. Se define como **f(x) = F'(x)**
    - La función de distribución en un punto se calcula como el área bajo la función de densidad, desde ese punto hasta el inicio del soporte de la variable.
    - **Propiedades:**
        - f(x)>=0
        - El área total bajo la función de densidad es 1
        - La probabilidad de un intervalo (a,b), (a,b] o [a,b] será el área bajo la curva f(x) entre a y b, es decir, $\int_a^b f(x)dx = F(b) - F(a)$.

Medidas características (continua)

- **Esperanza matemática** (media): E(X)= μ = $\int_{-\infty}^{\infty} x f(x)dx$
    - Exemplo (moedas): 1/4 *0 + 1*2*1 + ¼*2 = 1
- **Mediana** (M_e): Análogo a caso discreto. $F(M_e)=0.5$
- **Moda:** Valor para el cual la densidad alcanza un máximo relativo. $M_o=x_0$ si se cumple que $f'(x_0)=0$, y $f''(x_0)<0$.
- **Varianza:** $\sigma^2 = \mathbb{V}\text{ar}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2]$, sendo E(X) a esperanza matemática.
    - $\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx$, $\sigma = +\sqrt{\int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx}$.

[3] Al ser una variable continua, se considera que la probabilidad de que la variable alcance un valor concreto es siempre 0. Por esto, se debe calcular siempre la probabilidad de que la variable se sitúe en un intervalo.

Distribución uniforme continua

- Si X toma valores en un intervalo (a,b), se dice que X tiene una distribución **uniforme continua, X ∈ U(a,b),** si su función de densidad es:
- Su función de distribución asociada $F(x) = \begin{cases} 0 & x < a, \\ \frac{x-a}{b-a} & a \leq x < b, \\ 1 & x \geq b. \end{cases}$ será:
- E(X) = $\frac{a+b}{2}$
- Var(X) = $\frac{(b-a)^2}{12}$
- Su función de densidad será un segmento plano sobre el soporte de la variable, cuya altura será inversa a la longitud del soporte.

Distribución Normal

- Una variable aleatoria tiene distribución **normal**, X ∈ N (μ, σ^2), si su función de densidad es: $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- Se suele considerar la **distribución normal estándar**, de media 0 y varianza 1, X ∈ N(0,1), denotada por Φ(z).
  - $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$.
  - Todas las variables normales se pueden transformar a una estándar, restando la media y dividiendo por la desviación típica. $X \in N(\mu, \sigma^2) \Rightarrow Z = \frac{X-\mu}{\sigma} \in N(0,1)$
    - Ejemplo: si queremos calcular P(X<28), siguiendo X una distribución normal, es lo mismo que calcular P(Z<$\frac{28-\mu}{\sigma}$), siendo Z∈(0,1).
- Las distribuciones normales son aditivas: si tomamos X ∈ N(μ_1, σ^2_1) e Y∈N(μ_2,σ^2_2), entonces X+Y ∈ N ($\mu_1+\mu_2$, $\sigma^2_1+\sigma^2_2$)

Distribución exponencial

- Tenemos un proceso de Poisson de parámetro λ, y consideramos la variable X como el tiempo entre sucesos consecutivos. X sigue una distribución **exponencial** de parámetro λ: **X ∈ Exp(λ)**.
- X será continua y positiva.
- Su función de densidad será: $f(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0, \\ 0 & \text{en otro caso.} \end{cases}$
- $F(x) = 1 - e^{-\lambda x}$
- $E(X) = 1/\lambda$ (λ es el número esperado de sucesos por unidad de tiempo)
- $Var(X) = 1/\lambda^2$

Distribución Gamma[4]

- Generalización de la distribución exponencial: mide el tiempo hasta la aparición del n-ésimo suceso. **X ∈ Γ(n, λ)**
- Su función de densidad será: $f(x) = \begin{cases} \frac{\lambda^n}{\Gamma(n)} e^{-\lambda x} x^{n-1} & x > 0, \\ 0 & \text{en otro caso.} \end{cases}$
  - Si $n \in \mathbb{Z}^+$ entonces $\Gamma(n+1) = n!$
- $E(X) = n/\lambda$
- $Var(X) = n/\lambda^2$
- Dos variables Gamma con el mismo λ son aditivas.

[4] quoted saying que non entra no examen

Teorema central del límite

- El promedio de n variables independientes y que siguen la misma distribución (sea cual sea) con varianza finita sigue una **distribución normal**.
- Sea $\{X_i\}_{i\in N}$ una sucesión de variables aleatorias independientes e idénticamente distribuidas, con $E(X_i)=\mu$ y $Var(X_i)=\sigma^2$ para todo i. Entonces, para n suficientemente grande:

$$\sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2) \iff \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \sim N(0,1),$$

- Por ejemplo, si $X\in Bi(n,p)$, se puede escribir como $X = \sum_{i=1}^{n} X_i$, siendo cada $X_i\in Ber(p)$.
    - Entonces, $E(X_i)=p$ y $Var(X_i)=pq$ para todo i.
    - Aplicamos el teorema central del límite, y obtenemos que X se puede aproximar por: $N(np, npq).$
- Otro ejemplo, una distribución Gamma se puede expresar como $X = \sum_{i=1}^{n} X_i$, siendo cada $X_i\in Exp(\lambda)$
    - $E(X_i)=1/\lambda$ y $Var(X_i)=1/\lambda^2$ para todo i.
    - Aplicamos el teorema central del límite y obtenemos que X se puede aproximar por: $X \sim N(n/\lambda, n/\lambda^2)$

Corrección de Yates

- Al aproximar distribuciones por la normal, debemos tener en cuenta que la binomial y la Poisson son distribuciones discretas y la Normal es continua.
    - En una binomial, podemos calcular P(X=20), pero en una normal esta probabilidad será nula.
- Para solucionarlo, se aplica la **corrección de Yates**: $\mathbb{P}(X = 20) = \mathbb{P}(19.5 < X < 20.5)$

Aproximación de distribuciones

Distribución	Caso	Aproximación
Bi(n,p)	n>=30, p<0.1	Pois(np)
	n>=30, p∈[0.1, 0.9]	N(np, npq)
	n>=30, p>0.9	Se transforma en Bi(n,1-p)
Pois(λ)	λ>=10	N(λ,λ

5 - Inferencia estadística

Conceptos

- Una **muestra** es un subconjunto representativo de la población.
- Sea **X** la variable que estudiamos. Denotaremos por $\{X_1, \ldots, X_n\}$ una **muestra aleatoria simple (m.a.s)**, donde cada variable X_i tomará un valor x_i.
    - Al conjunto de valores que toma la muestra $\{x_1, \ldots, x_n\}$ se denomina realización muestral.
- Identificamos la población con la distribución de la variable que analizamos. Por ejemplo, 'una población Normal' hace referencia a la distribución de nuestra variable, no al conjunto de individuos.
- **Parámetro:** Característica de la población. Por ejemplo, en una población Binomial, el parámetro que caracteriza a nuestra variable es p[5].
    - El parámetro puede ser unidimensional (caso previo) o bidimensional (por ejemplo, en una $N(\mu, \sigma^2)$)
    - Denotamos por θ al parámetro de interés en la población. Se dirá que la variable tiene distribución $F(\theta)$.
- **Estadístico:** Cualquier función de la muestra (media, varianza, máximo…) Se denotan por $T(X_1, \ldots, X_n)$
    - **Estimadores:** Estadísticos independientes de los parámetros de la población. Por ejemplo, la media de la muestra es un estimador de la media poblacional, y la varianza muestral lo es de la poblacional
- **Método de muestreo:** Procedimiento por el cual se selecciona la muestra.

Distribuciones en el muestreo

- Sea $X_1, \ldots, X_n$ una muestra aleatoria simple de una v.a. X con distribución $F(\theta)$.
- Se considera que la muestra sigue la misma distribución que la variable aleatoria.
- Consideremos que F es una distribución normal. Entonces, debemos calcular su media y su varianza.

[5] Se asume que n es conocido.

- Definimos $T(X_1, \ldots, X_n)$ como un estadístico. Dado que es una función de la muestra aleatoria, tendrá también un comportamiento aleatorio y por tanto una distribución.
    - Se debe tener en cuenta que el valor del será distinto para distintas muestras de la misma población.
    - Si realizamos m muestras de $X_1,..X_n$, denotamos por $T^1, \ldots, T^m$ los distintos valores del estadístico en cada una de las muestras.

$$\begin{array}{cccc|cc} x_1^1 & x_2^1 & \ldots & x_n^1 & \rightarrow & T^1 \\ x_1^2 & x_2^2 & \ldots & x_n^2 & \rightarrow & T^2 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ x_1^m & x_2^m & \ldots & x_n^m & \rightarrow & T^m \end{array}$$

    - Estos valores de $T^1, \ldots, T^m$ seguirán una distribución.

Proporción muestral

- La **proporción muestral** es la relación de casos de éxitos en una muestra respecto al tamaño de la muestra. Se denota con $\hat{p}$. (p^ nestes apuntes)
- Si me interesa conocer la proporción de elementos p de una población X ~ Ber(p), seleccionamos una MAS $X_1,...,X_n$ de variables Ber(p).
- **Distribución:** $\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n} \sim \frac{Bi(n,p)}{n} \rightarrow \boxed{\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)}$

Media muestral (varianza conocida)

- Supongamos que disponemos de una m.a.s $X_1, \ldots, X_n$ tal que X~ N (μ, σ2). La **media muestral** $X_{barra} = \frac{1}{n}\sum_{i=1}^{n} X_i$ se puede escribir como la suma total de n términos $Y_i = \frac{X_i}{n}$.
    - **Nota:** no confundir con la **media real** de la población, denotada por **μ**
- Debido a esto, la media muestral sigue una distribución normal: $\boxed{\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)}$ que se puede tipificar como $\boxed{\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1).}$
    - Es decir, los posibles valores de X_{barra} se distribuyen según una Normal, centrada en la media real **μ** y cuya varianza disminuye a medida que aumenta el tamaño n de la muestra.

Distribución χ^2 (ji-cuadrado)

- Supongamos que disponemos de una m.a.s $X_1, \ldots, X_n$ de variables Xi~N(μ,σ2)

  Entonces, $\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi^2_n.$

- Cuando la muestra es lo suficientemente grande, una distribución χ^2_n se puede aproximar por una N(n,2n)

- La distribución χ^2_n **no** es simétrica.

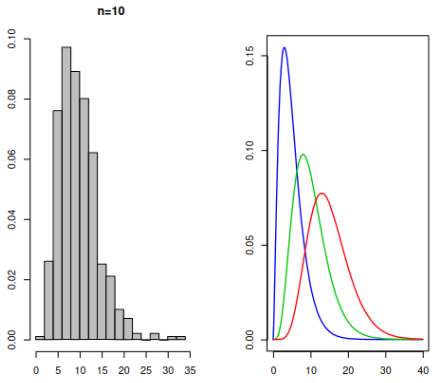


Figura 3: Distribución en el muestreo de la suma de los cuadrados de $n = 10$ variables Normales estándar. Gráficas de la densidad χ^2_n: línea azul: $n = 5$; línea verde: $n = 10$; línea roja: $n = 15$.

Distribución de la varianza/cuasivarianza muestral

- El **Teorema de Fisher** establece que si $X_1, \ldots, X_n$ es una m.a.s. de variables normales con varianza σ^2, entonces X y s^2 son independientes y además: (imagen) (s^2 es la varianza muestral)

$$\frac{ns^2}{\sigma^2} \sim \chi^2_{n-1}.$$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

- Si en su lugar queremos utilizar la **cuasivarianza** S^2,

Distribución t de Student

- Consideremos una variable X ~ N (0, 1) y otra v.a. Y ~ χ^2_n independientes, el cociente: $\frac{X}{\sqrt{\frac{Y}{n}}} \sim t_n$ sigue una **distribución t** con **n grados de libertad**.

- Cuando n es suficientemente grande, se aproxima a una N(0,1).

- Permite describir la distribución de la **media muestral** si no se conoce la varianza poblacional.

$$\frac{\overline{X}-\mu}{s/\sqrt{n-1}} \sim t_{n-1}, \quad \text{o bien} \quad \frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}.$$

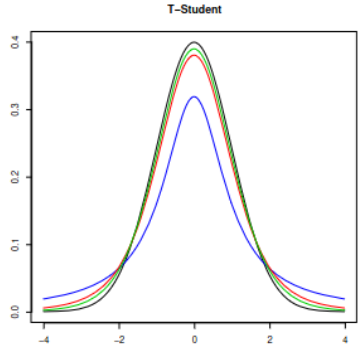


Figura: Distribución t de Student con distintos grados de libertad. Azul: $n = 1$ (Cauchy); roja: $n = 5$; verde: $n = 10$; negra: $N(0, 1)$.

Distribución de la diferencia de medias[6]

- Supongamos ahora que tenemos dos poblaciones y las correspondientes muestras X1, . . . , Xn m.a.s. de X ~ N (μX , σ^2_X) e Y1, . . . , Ym una m.a.s. de Y ~ N (μY , σ^2_Y).

- Si σ^2_X y σ^2_y son conocidas: $\frac{(\overline{X}-\overline{Y})-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}} \sim N(0,1)$

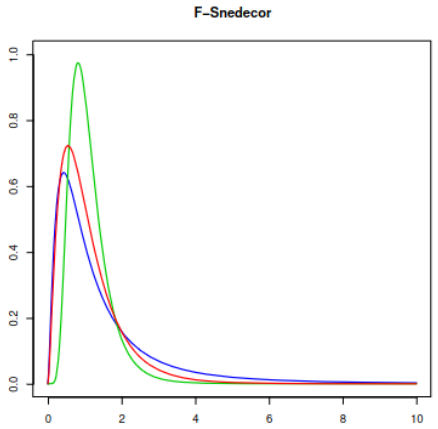
- Si σ^2_X y σ^2_y son desconocidas, pero iguales: $S_T^2=\frac{(n-1)S_X^2+(m-1)S_Y^2}{n+m-2}$, $\frac{(\overline{X}-\overline{Y})-(\mu_X-\mu_Y)}{S_T\sqrt{\frac{1}{n}+\frac{1}{m}}} \sim t_{n+m-2}$

- Si σ^2_X y σ^2_y son desconocidas, y distintas

$\frac{(\overline{X}-\overline{Y})-(\mu_X-\mu_Y)}{\sqrt{\frac{S_X^2}{n}+\frac{S_Y^2}{m}}} \sim t_{n+m-2-\delta}$, sendo δ o enteiro máis proximo a $\frac{(m-1)\frac{S_X^2}{n}+(n-1)\frac{S_Y^2}{m}}{(m-1)\left(\frac{S_X^2}{n}\right)^2+(n-1)\left(\frac{S_Y^2}{m}\right)^2}$

Distribución F de Snedecor

- Sexa X unha va con distribución χ^2_n e Y con distribución χ^2_m, ambas independientes. O cociente $\frac{X}{n}/\frac{Y}{m} \sim F_{n,m}$ sigue unha distribución F con n e m graos de liberdade.



- Azul: n = 5, m = 5
- Roja: n = 5, m = 20
- Verde: n = 20, m = 20.

[6] estas formulas danas no examen let's fucking GOOOO

Distribución do cociente de varianza

- Polo teorema de Fisher, coñecemos que $\frac{(n-1)S_X^2}{\sigma_X^2} \sim \chi_{n-1}^2$ y $\frac{(m-1)S_Y^2}{\sigma_Y^2} \sim \chi_{m-1}^2$
- Entón, o cociente de varianzas ten distribución F con n-1 e m-1 graos de liberdade: $\boxed{\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim F_{n-1,m-1}.}$

Intervalo de confianza

- Dada unha m.a.s $X_1, \ldots, X_n$ de $X \sim F_\theta$, un **intervalo de confianza de nivel (1-α)** con α € (0,1) é un intervalo aleatorio tal que[7]
- $$\mathbb{P}\left(\hat{\theta}_1(X_1,\ldots,X_n) < \theta < \hat{\theta}_2(X_1,\ldots,X_n)\right) = 1-\alpha, \quad \forall\theta \in \Theta.$$
    - Os extremos dos intervalos son aleatorios porque dependen da mostra.
- É dicir, hai unha probabilidade 1-α de que a m.a.s pertenza a ese intervalo.
    - Desta forma, aínda que non coñezamos o valor dun estadístico, podemos calcular un intervalo ao que existe unha determinada probabilidade ao que pertenzan, estimando o seu valor.

Intervalo de confianza para a proporción p (p descoñecido)[8]

- Para unha proporción p que descoñecemos, grazas ao teorema central do límite coñecemos que p^~N(p, p(1-p)/n).
- Tipificando a variable obtenemos que $\frac{(p^\wedge - p)}{\sqrt{p(1-p)/n}}$ ~ N(0,1). Denominamos esta nova m.a.s como p_2.
    - p_2 coñécese como o **estadístico pivote:** o elemento do que partimos para obter o intervalo de confianza.

---

[7] non confundir nivel de significación (α) con nivel de confianza (1-α). 'Por ejemplo, si el nivel de confianza de un intervalo de confianza es del 95%, su nivel de significación es del 5%. Esto significa que si repetimos 100 veces el estudio estadístico, 95 veces obtendremos un resultado que coincide con el de la población real, mientras que 5 veces obtendremos un resultado erróneo.'

[8] Lembrar que p denota a proporción da poboación (descoñecida, queremos estimar o seu valor), e p^ denota a proporción muestral (coñecida)

- Non podemos calcular directamente o intervalo $P(L_1<=p<=L_2)=1-\alpha$ porque non coñecemos a distribución de p, pero si coñecemos a de p_2.
- $P(-z_{1-\alpha/2} <= p_2 <= z_{1-\alpha/2}) = 1-\alpha$. Ahora, despexamos p de dentro de p_2. Finalmente, substituímos todas as instancias de p por p^ para que o intervalo non dependa de p (que é descoñecido)
    - **Nota:** z_p denota o valor tal que $P(N(0,1)<=z_p) = p$.
    - Por exemplo, se se pedise un intervalo de confianza 90%, teríamos $1-\alpha=0.9 \rightarrow \alpha=0.1 \rightarrow$Calculamos $z_{0.95}$, é decir, consultamos a tabla para ver que valor de z_p se corresponde coa p=0.95 (ver en que casilla se atopa este valor). Neste caso, o máis aproximado sería entre 1.75 e 1.76, poderiamos tomar $z_p=1.755$.
- O intervalo final obtido é simétrico respecto de p^:[9]

$$(\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}})$$

- Debido a que é simétrico, podemos expresar o intervalo como:

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

$$L = 2z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} \Leftrightarrow n = \frac{4z^2_{1-\alpha/2}\hat{p}(1-\hat{p})}{L^2}$$

Intervalo de confianza para a media con varianza coñecida

- Sexa $X_1, \ldots, X_n$ unha m.a.s de $X \sim N(\mu,\sigma^2)$ de varianza coñecida. Entón, $\sqrt{n}\frac{Xmedia-\mu}{\sigma} \sim N(0,1)$. $\sqrt{n}\frac{Xmedia-\mu}{\sigma}$ será o noso estadístico pivote.[10]
- O intervalo do que partimos é $P(-z_{1-a/2}, \sqrt{n}\frac{Xmedia-\mu}{\sigma}, z_{1-a/2}) = 1-a$.

$$I_c = Xmedia \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$L = 2z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \Leftrightarrow n = \frac{4z^2_{1-\alpha/2}\sigma^2}{L^2}$$

Intervalo de confianza para a media con varianza descoñecida

- Se non coñecemos a varianza, odemos utilizar $\sqrt{n}\frac{Xmedia-\mu}{S} \sim t_{n-1}$. Este será o estadístico pivote.
- O intervalo do que partimos é $P(-t_{n-1,\,1-a/2}, \sqrt{n-1}\frac{Xmedia-\mu}{s}, t_{n-1,\,1-a/2}) = 1-a$.

$$L = 2t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}} \Leftrightarrow n = \frac{4t^2_{n-1,1-\alpha/2}S^2}{L^2}$$

[9] en realidade, os p^ de dentro das raices serían p, pero como non o coñecemos asumimos que p^ é close enough

[10] X_{media} é a x cunha barra arriba non sei escribila en gogoeld cdocs

$$\mathbf{I_c} = Xmedia \pm t_{n-1,\,1-a/2}\frac{S}{\sqrt{n}}$$

Exemplo de exercicio (intervalo de confianza)

- Dada unha variable que estamos analizando, tomamos unha mostra de n=16 elementos. Nesta mostra, a varianza é de 25 e a media de 503. Calcular o intervalo tal que hai un 90% de probabilidade de que inclúa á media da poboación.
    - 1-a = 0.9 → a = 0.1.
    - Calculase $z_{1-0.1/2}$= $z_{0.95}$. mirase na tabla o valor de z que da p=0.95, aproximase a 1.645.
    - I_c = (503-1.645*5/$\sqrt{16}$, 503+1.645*5/$\sqrt{16}$) = (501.355, 504.645)
- Se non coñecemos a varianza, senon que só coñecemos que S=38.47, empregase a distribucion t de student (ten unha tabla distinta)
    - Trátase dunha $t_{15,\,0.95}$. O valor da tabla correspondente é 1.753. (neste caso é ao reves, hai que consultar o valor da tabla na casilla indicada)
    - $I_c = 503 \pm 1.753\frac{38.47}{\sqrt{16}}$

Intervalo de confianza para a varianza

- Coñecemos que $nS^2/\sigma^2 \sim \chi^2_{n-1}$. Este será o noso estadístico pivote.

$$\left(\frac{ns^2}{\chi^2_{n-1,\,1-a/2}}, \frac{ns^2}{\chi^2_{n-1,\,a/2}}\right)$$

Lonxitude dun intervalo de confianza

- Para calqueira intervalo excepto o da varianza[11], o valor da lonxitude do I_c será 2*(o extremo superior do intervalo).
- Despexando n na fórmula resultante, podemos calcular n para que un I_c teña unha lonxitude L. (ver fórmulas)

[11] Debido a que a distribución que emprega non é simétrica e as outras si.

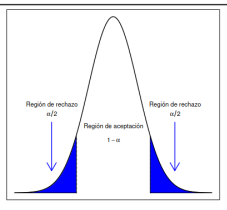
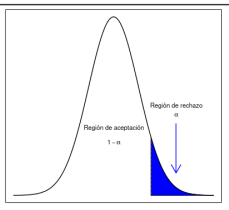
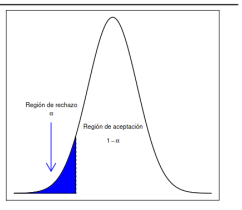
6 - Contrastes de hipótese

Conceptos

- **Contraste de hipótesis:** procedimiento estadístico mediante el cual se investiga la veracidad o falsedad de una hipótesis sobre una o varias poblaciones.
- Tipos de hipótesis:
    - **Paramétrica:** Afirmación sobre alguno de los valores de los parámetros poblacionales. Pueden ser bilaterales ($\mu=\mu_0$) o unilaterales ($\mu\leq\mu0$ o $\mu\geq\mu0$) a la derecha e izquierda, respectivamente.
        - **Simple:** Especifica un único valor para cada parámetro de la población
        - **Compuesta:** Especifica un conjunto de posibles valores para los par.
- **Hipótesis nula (H_0):** Hipótesis que se desea contrastar. Se compara con la **hipótesis alternativa (H_a)**. Por ejemplo, si H_0: $\mu=\mu_0$, entonces la alternativa será H_a: $\mu\neq\mu_0$.

Estadístico de contraste

- Llamamos **estadístico de contraste** al que tiene distribución conocida cuando H_0 es cierta.
    - Por ejemplo, si nuestra hipótesis H_0 es p=0.5, sabemos que si H_0 es cierta, $\frac{\hat{p}-0,5}{\sqrt{\frac{0,5(1-0,5)}{n}}}\sim N(0,1)$ entonces:
- **Región crítica:** Conjunto de valores del estadístico que nos llevan a rechazar H_0
- **Región de aceptación:** Conjunto de valores del estadístico que nos llevan a aceptar H_0.
- El **nivel de significación (α)** es la área de la región de rechazo. A mayor α, más exigente el estudio.

Bilateral	Unil. izquierda	Unil. derecha
		
$H_0:\ \theta=\theta_0$	$H_0:\ \theta\leq\theta_0$	$H_0:\ \theta\geq\theta_0$

Errores

- **Tipo I:** Rechazar H_0 cuando es válida
    - **Nivel de significación (α):** Probabilidad de cometer el error de tipo I[12].
- **Tipo II:** Aceptar H_0 cuando es falsa
    - **β:** Probabilidad de cometer el error de tipo II.
    - **Potencia del contraste (1-β):** Probabilidad de rechazar H_0 cuando es falsa.

Procedimiento de contraste (con ejemplo)

- **Enunciado:** Comprobar si la variabilidad muestral de 0.005 con una muestra de 20 elementos es mayor que la variabilidad del mes pasado, que fue de 0.004, con un nivel de significación del 10%.

1. Se establecen la hipótesis nula H_0 y la alternativa H_a.
    - H_0: $s^2>=0.004$, H_1. $s^2<0.004$
2. Fijamos un nivel de significación. Normalmente, α=0.01, α=0.05, α=0.10
    - α=0.10
3. Especificamos el tamaño muestral n
    - n=20
4. Consideramos un estadístico de contraste y establecemos su distribución considerando cierta la hipótesis nula.
    - $\frac{ns^2}{\sigma^2} \sim \chi^2_{n-1},$
5. Calculamos el valor del estadístico, asumiendo que H_0 es cierta.
    - $\frac{ns^2}{\sigma} = \frac{20 * 0.005}{0.004} = 25$
6. Construimos las regiones de aceptación/rechazo
    - Consultamos la tabla de la distribución Chi-cuadrado, para k=19 y p=α=0.10. El valor es 11.65.
    - Al ser un contraste unilateral a la izquierda ($s^2>=0.004$), la región de aceptación es la que queda por la derecha del valor.
        i. RR: (-inf, 11.65). RA:(11.65, +inf)
7. Observamos a qué región pertenece, y concluímos el contraste.

[12] como suele ser mui pequeno, generalmente tendense a aceptar muitos. a non ser que sea como unha barbaridad. daseme mal este tema nn sei

- 25 ∈ RA. Sí, es mayor.

8. Otro método es calcular el **p-valor**: el valor de p que devuelve el valor obtenido del estadístico, y comparar este p-valor con α.

Anexo: Estadísticos de contraste (Resumen)

Distribución de X	Estadístico a aproximar	Distribución de estadístico
Ber(p)	p	$\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$
X~Ber(p_X), Y~Ber(p_Y)	(p_X - p_Y)	$\frac{(\hat{p}_X-\hat{p}_Y)-(p_X-p_Y)}{\sqrt{\frac{p_X(1-p_X)}{n}+\frac{p_Y(1-p_Y)}{m}}} \sim N(0,1)$
N (μ, σ^2)	μ	$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$
		$\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}, \quad \frac{\overline{X}-\mu}{s/\sqrt{n-1}} \sim t_{n-1}$
	σ^2	$\frac{ns^2}{\sigma^2} \sim \chi^2_{n-1}, \quad \frac{(n-1)S^2}{\sigma^2}\chi^2_{n-1}$

Boletín 2. Fundamentos de probabilidad
Estadística - Grado en Ingeniería Informática

1. Sean A, B sucesos independientes en un mismo espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$. Si $\mathbb{P}(A^c) = 0.3$ y $\mathbb{P}(A \cup B) = 0.75$, calcula:

   a) $\mathbb{P}(B)$

   b) $\mathbb{P}(A \backslash B)$

   c) Sea $C \in (\Omega, \mathcal{A}, \mathbb{P})$. ¿Qué condiciones deberían darse sobre C para que A y C formen un sistema completo de sucesos?

2. Sean A, B y C sucesos en un mismo espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$ donde $\mathbb{P}(A) = 0.3$, $\mathbb{P}(B) = 0.6$ y $\mathbb{P}(C) = 0.25$. Si A y C son independientes, A y B son incompatibles y $\mathbb{P}(B \cup C) = 0.8$ calcula $\mathbb{P}(A \cup B|C)$.

3. Sean A y B dos sucesos asociados a un mismo experimento aleatorio con $\mathbb{P}(A) = 0.7$, $\mathbb{P}(B) = 0.6$ y $\mathbb{P}(A^c \cup B^c) = 0.58$.

   a) ¿Son A y B independientes?

   b) Si $E \subset A$, calcula $\mathbb{P}(E^c|A^c)$.

4. Sean A y B dos sucesos tales que $\mathbb{P}(A) = 1/4$, $\mathbb{P}(B|A) = 1/2$ y $\mathbb{P}(A|B) = 1/4$. Decir si son ciertas o falsas las siguientes relaciones:

   a) A y B son independientes.

   b) $\mathbb{P}(A \backslash B) = 1/2$.

5. Sean A y B dos sucesos independientes. Probar que:

   a) A y B^c son independientes.

   b) A^c y B^c son independientes.

6. Sean A y B dos sucesos de un espacio de probabilidad, de manera que: $\mathbb{P}(A) = 0.4$, $\mathbb{P}(B) = 0.3$ y $\mathbb{P}(A \cap B) = 0.1$. Calcula:

   a) $\mathbb{P}(A \cup B)$

   b) $\mathbb{P}(\bar{A} \cup \bar{B})$

   c) $\mathbb{P}(A|B)$

   d) $\mathbb{P}(\bar{A} \cap \bar{B})$

7. Sean A y B dos sucesos de un experimento aleatorio. ¿Es posible que $\mathbb{P}$ sea una probabilidad si $\mathbb{P}(A) = 2/5$, $\mathbb{P}(B) = 1/5$ y $\mathbb{P}(\bar{A} \cap \bar{B}) = 3/10$?

8. Tres alumnos del Grado en Ingeniería Informática quedan los viernes por la tarde para jugar a los dardos. Si la probabilidad de acertar en el centro de la diana para cada uno de ellos es $1/6$, $1/4$ y $1/3$ respectivamente, calcula la probabilidad de que tan sólo uno de ellos acierte en la diana.

9. Un lote de ordenadores se compone de 20 ordenadores TECNO y 10 ordenadores PROTÓN. La mitad de los TECNO y la mitad de los PROTÓN tienen una tarjeta gráfica Nvidia. Calcula la probabilidad de que, eligiendo un ordenador al azar sea:

   a) Un TECNO con tarjeta Nvidia.
   b) Un PROTÓN sin tarjeta Nvidia.
   c) Sabiendo que es un PROTÓN, calcula la probabilidad de que tenga tarjeta Nvidia.
   d) ¿Son independientes los sucesos *PROTÓN* y *tener tarjeta Nvidia*?

10. Una caja contiene 3 bolas blancas y 7 bolas negras. Otra caja tiene 9 bolas blancas y una negra. Se elige una caja al azar y se extrae una bola. Calcula:

    a) La probabilidad de que la bola sea blanca, habiendo elegido la primera caja.
    b) La probabilidad de que la bola sea blanca, habiendo elegido la segunda caja.
    c) La probabilidad de que hayamos elegido la primera caja y la bola extraida sea blanca.
    d) La probabilidad de que hayamos elegido la primera caja y la bola extraida sea negra.
    e) La probabilidad de que la bola sea blanca.
    f) Sabiendo que la bola elegida es blanca, ¿cuál es la probabilidad de que venga de la segunda caja?

11. En una empresa, el $40\,\%$ de los ordenadores son hp, el $25\,\%$ tienen Windows 10 y el $15\,\%$ son hp con Windows 10. Se elige un ordenador al azar:

    a) Si es un hp, ¿cuál es la probabilidad de que tenga Windows 10?
    b) Si tiene Windows 10, ¿cuál es la probabilidad de que sea un hp?
    c) ¿Cuál es la probabilidad de que ni sea hp ni tenga Windows 10?

12. La evaluación de una materia consiste en dos pruebas. La probabilidad de pasar la primera es de 0.6, mientras que la probabilidad de pasar la segunda es de 0.8 y de pasar ambas es 0.5. Calcula:

    a) La probabilidad de que un alumno pase al menos una prueba.
    b) La probabilidad de que no pase ninguna prueba.
    c) ¿Son las pruebas sucesos independientes?
    d) La probabilidad de que pase la segunda, sabiendo que ha superado la primera.

13. Unos conocidos productores de café de Colombia utilizan compañías aéreas locales para enviar el café producido desde las montañas al aeropuerto internacional más cercano. Por razones de coste, el $65\,\%$ de las veces contratan a la compañía AirWings, mientras que los viajes restantes los realizan con LifeFlight. Ambas compañías poseen aviones Tupolev (la mitad de las aeronaves de AirWings y el $75\,\%$ de las de LifeFlight son de este fabricante). Calcula:

    a) La probabilidad de que uno de los envíos no se realice en Tupolev.

    b) Si el envío desde las montañas ha sido realizado en un Tupolev, calcula la probabilidad de que la compañía que lo ha transportado sea LifeFlight.

    c) La probabilidad de que el envío sea con AirWings o en Tupolev.

14. Una empresa de reparto de mercancías cubre diariamente el servicio entre Ferrol y Ribadeo por la N642. Supongamos que la probabilidad de que sufra un accidente en un día con niebla es 0.02 y en un día sin niebla es 0.004. Cierto día de un mes, en el que hubo 12 días con niebla y 18 sin niebla, se produjo un accidente. Calcula:

    a) La probabilidad de que el accidente ocurriese un día con niebla.

    b) La probabilidad de que el accidente ocurriese un día sin niebla.

15. En un análisis sobre la efectividad de programas antivirus para el troyano Bifrost, el $57\,\%$ de los usuarios utiliza el antivirus tipo A (con una efectividad del $99\,\%$), el $38\,\%$ tienen el antivirus tipo B (efectividad del $95\,\%$) y el resto de los usuarios, tienen otros antivirus con una efectividad del $93\,\%$. Si elegimos un usuario al azar y comprobamos que el antivirus fue efectivo, calcula la probabilidad de que ese usuario no tenga el antivirus tipo A.

16. En una prueba de evaluación continua, los profesores dan a elegir dos opciones, A y B. El $80\,\%$ del alumnado elige la opción A. Cuando corrigen la prueba, se dan cuenta de que la distribución de aprobados y suspensos en las opciones no es la misma. Así, el $90\,\%$ de los que eligen B aprueban, mientras que para la prueba A, el porcentaje de aprobados se reduce al $60\,\%$.

    a) Calcula el porcentaje de estudiantes aprobados.

    b) Si un estudiante ha suspendido, calcula la probabilidad de que haya elegido la opción A.

    c) ¿Cuál es la probabilidad de que un estudiante haya elegido A y haya aprobado?

17. Una empresa de venta de material informático por internet ha comprobado que uno de cada cien clientes, con fondos en su cuenta bancaria, pone el número de cuenta erróneo. Sin embargo, todo cliente sin fondos en la cuenta pone un número de cuenta incorrecto. Sabiendo que el $90\,\%$ de los clientes tiene fondos en la cuenta, calcula:

    a) Si hoy se ha realizado una venta, ¿cuál es la probabilidad de que el número de cuenta aportado sea incorrecto?

    b) Si en la venta realizada se comprueba que el número de cuenta es incorrecto, ¿cuál es la probabilidad de que el cliente non tenga fondos?

18. En un servidor de correo electrónico, la probabilidad de que llegue SPAM es de 0.1. Si el correo es SPAM, la probabilidad de que el software anti–SPAM del servidor lo borre es de 0.95. Además, el software anti–SPAM del servidor también elimina, por error, el $3\,\%$ de los correos que no son SPAM. Calcula:

    a) Porcentaje de correo eliminado por el software anti–SPAM.

    b) Porcentaje de SPAM en los correos eliminados.

19. El $6.7\,\%$ de las declaraciones de la renta presentan errores numéricos. Este porcentaje de errores es del $90\,\%$ entre las declaraciones fraudulentas y del $5\,\%$ en las no fraudulentas.

    a) Determina la probabilidad de que una declaración elegida al azar sea fraudulenta.

    b) Determina la probabilidad de que una declaración con error numérico sea fraudulenta.

Boletín 3 y 4. Variables aleatorias
Estadística - Grao en Ingeniería Informática

1. Sea la variable aleatoria $X =$menor valor observado al lanzar dos dados. Calcula:

   a) La función de masa de probabilidad.

   b) La función de distribución.

   c) La media y la varianza de X.

   d) La probabilidad de que X sea par.

   e) $\mathbb{P}(X \leq 4)$ y $\mathbb{P}(3 < X \leq 5)$.

2. El número de virus X detectados por un programa antivirus sigue una distribución dada por:

x_i	0	1	2	3	4
p_i	0.9	0.05	?	0.015	0.005

   a) Calcula la número medio de virus detectados.

   b) Calcula Var(X).

   c) Obtén la función de distribución.

   d) $\mathbb{P}(X \leq 2)$ y $\mathbb{P}(X > 3)$.

3. En la primera parte del examen de Estadística del curso 2019-2020 se hizo una prueba tipo test de 20 preguntas. Un alumno que ha preparado la materia concienzudamente, tiene un 80 % de posibilidades de responder bien a cada pregunta.

   a) ¿Cuál es la probabilidad de que un alumno conteste correctamente a 10 preguntas?

   b) ¿Cuál es la probabilidad de que la primera pregunta correcta sea la cuarta?

   c) ¿Cuál es la probabilidad de que la undécima pregunta sea la quinta correcta que contesta un alumno?

4. Un fabricante de faros para coche informa que en un envío de 4000 faros a un distribuidor, 500 tenían un defecto. Si se compran al distribuidor 20 faros elegidos al azar, ¿cuál es la probabilidad de que haya exactamente 2 con defecto?

5. Para revisar la calidad de una cadena de producción de tarjetas gráficas, se seleccionan diariamente 10 unidades y se observa y si son defectuosas o no. De un estudio más exhaustivo, se ha concluido que la probabilidad de tener una tarjeta defectuosa es 0.05. El responsable de producción decide que el proceso se para si en una muestra de 10 tarjetas, hay 2 o más defectuosas.

   a) ¿Cuál es la v.a. adecuada? Plantea su distribución.

b) ¿Cuál es la probabilidad de obtener una pieza defectuosa?

c) ¿Cuál es la probabilidad de que se pare la cadena?

d) ¿Cuál es la probabilidad de que tengamos que revisar 5 tarjetas antes de obtener la primera defectuosa?

e) ¿Cuál es la probabilidad de que tengamos que revisar 7 tarjetas antes de obtener 2 defectuosas?

6. Sea X una variable aleatoria Binomial de media 20 y varianza 16. Calcula la probabilidad de que X tome el valor 25.

7. Consideremos la variable X=*número de personas que se suben a un autobús* y sea la variable Y_i la que codifica si una persona sube al autobús o si no lo hace, tomando valores 1 y 0, respectivamente. Sea p_i la probabilidad de que la persona i-ésima se suba al autobús. Si tenemos n personas cuya decisión de subirse o no al autobús no depende de los demás:

   a) ¿Cómo se puede escribir la variable X en función de las Y_i? ¿Qué distribución tiene?

   b) Calcula el valor esperado y la varianza de la variable X.

8. Un proveedor de computación en la nube recibe una media de 6 programas por minuto para ejecutar. Si la estación ha estado parada durante 45 segundos, ¿cuál es la probabilidad de que hayan quedado sin atender más de 3 programas? ¿Cuál es la probabilidad de que el primer programa después de solucionar la avería llegue antes de 15 segundos?

9. El trayecto desde el centro de Santiago hasta el polígono del Tambre está cubierto por una línea de autobuses. Cada mañana, los 140 trabajadores de una fábrica del polígono deciden, de forma independiente de sus compañeros, si cogerán el autobús o si no lo harán. La probabilidad de que un trabajador coja el autobús para desplazarse es 0.15. Sea X la variable que cuantifica el número de trabajadores que suben al autobús.

   a) Determina la distribución de la variable X. ¿Cuál es el número esperado de trabajadores que suben al autobús?

   b) ¿Cuál es la probabilidad de que suban exactamente 21 trabajadores?

   c) ¿Cuál es la probabilidad de que suban menos de 28?

10. Sean X e Y dos variables aleatorias con $\mathbb{E}(X) = \mathbb{E}(Y) = \mathbb{E}(XY) = 1$ y $\text{Var}(X) = \text{Var}(Y) = 1$, y definamos $W = bX + cY$. ¿Qué valores deben tomar b y c para que $\mathbb{E}(W) = \text{Var}(W) = 1$?

11. Dada la función:

$$f(x) = \begin{cases} 0 & \text{si} \quad x \leq -1, \\ kx^2 & \text{si} \quad -1 < x \leq 1, \\ 0 & \text{si} \quad 1 < x. \end{cases}$$

   a) Calcular k para que f sea función de densidad.

b) Calcular $\mathbb{P}(-0.5 < X \leq 0.5)$ y $\mathbb{P}(-2 < X \leq 0)$.

c) Calcular la función de distribución asociada a f.

d) Comprobar que, el segundo resultado del apartado $b)$ puede obtenerse como $F(0) - F(-2)$.

12. El diámetro de una pieza de la CPU varía según la densidad:

$$f(x) = k(x-1)(x-3), \quad \text{si} \quad 1 \leq x \leq 3.$$

La pieza sirve para un determinado modelo si su diámetro está entre 1.9 y 2.1. ¿Cuál es la probabilidad de que una pieza elegida al azar sea útil?

13. Sea X una v.a. con distribución Normal de media μ y varianza σ^2. Calcula las siguientes probabilidades:

a) $\mathbb{P}(X \leq 25)$ con $(\mu, \sigma^2) = (20, 9)$.

b) $\mathbb{P}(X > 5)$ con $(\mu, \sigma^2) = (4, 4)$.

c) $\mathbb{P}(X < 70)$ con $(\mu, \sigma^2) = (70, b)$.

d) $\mathbb{P}(13 < X < 17)$ con $(\mu, \sigma^2) = (15, 4)$.

e) $\mathbb{P}(|X| < 3)$ con $(\mu, \sigma^2) = (4, 1)$.

f) $\mathbb{P}(|X| > 7)$ con $(\mu, \sigma^2) = (5, 1)$.

14. Un alumno viene a la ETSE andando todos los días, saliendo de casa en algún momento entre las 8:40 y las 8:50. Si llegar a la ETSE le lleva 15 minutos y la clase comienza a las 9:00, ¿cuál es la probabilidad de que llegue tarde? ¿Cuál es la hora media de llegada?

15. El 5 % de los alumnos de Estadística supera los 180cm. de altura, mientras que el 6 % no llega a 160. Suponiendo que la altura sigue una distribución Normal, ¿qué proporción de alumnos mide más de 190cm? ¿Qué proporción de alumnos mide entre 170 y 175cm?

16. Un determinado tipo de bombillas tienen una vida media de 8 meses.

a) ¿Cuál es la probabilidad de que una bombilla dure entre 3 y 12 meses?

b) ¿Cuál es la probabilidad de que una bombilla que ya ha durado 10 meses, siga funcionando más de 25 meses?

17. En el ascensor de un centro comercial caben aproximadamente 20 personas, y tiene un límite de seguridad de 1200 Kg. Si el peso de una persona es una variable aleatoria con distribución $N(65, 225)$, calcula la probabilidad de que se supere el límite de seguridad.

18. El tiempo de duración del ventilador de una CPU sigue una distribución Normal con media y desviación típica de 10 y 3.5 años, respectivamente. El fabricante reemplaza todos los ventiladores que han fallado durante el período de garantía. Si no tiene intención de reemplazar más del 4 % de los ventiladores, ¿cuál es el máximo período de garantía que puede fijar?

19. En un colegio de primaria se pasa un test de inteligencia y los resultados siguen una distribución normal con media 100 y desviación típica 10.

    a) ¿Cuál es la probabilidad de que un niño elegido al azar haya sacado entre 90 y 115 puntos?

    b) ¿Qué nota se necesita sacar para estar entre el $10\,\%$ de los mejores?

    c) Calcula c para que $\mathbb{P}(105 \leq X \leq c) = 0.10$.

20. La media de llegadas de coches a una estación de servicio es de 3 por minuto.

    a) Calcula la probabilidad de que pase más de 1 minuto hasta que llegue el primer coche.

    b) Si ya han atendido a un coche, ¿cuál es la probabilidad de que el siguiente tarde más de un minuto?

    c) La probabilidad de esperar más de media hora por el vigésimo coche.

    d) La probabilidad de que al menos 5 coches lleguen en 2 minutos.

    e) La probabilidad de que en una hora pasen más de 200 coches.

21. La aplicación de solicitud de cita previa para renovación del DNI soporta una media de 4 peticiones cada 10 minutos. Calcula:

    a) La probabilidad de que en dos minutos se reciba alguna petición.

    b) La probabilidad de que haya que esperar más de dos minutos entre dos peticiones consecutivas.

    c) Si debido al procedimiento de autentificación empleado, la aplicación soporta un máximo de 40 peticiones cada media hora, ¿cuál es la probabilidad de que el servicio falle?

22. La USC pone a disposición de su profesorado un nuevo programa que detecta en qué páginas de un documento, escrito en castellano o gallego, hay faltas de ortografía. Tras la revisión de los trabajos entregados por el alumnado en los últimos anos, se establece que la probabilidad de que una página contenga alguna falta de ortografía es del $30\,\%$.

    a) En un documento de 5 páginas, ¿cuál es la probabilidad de que 3 presenten faltas de ortografía?

    b) En un documento de 50 páginas, ¿cuál es la probabilidad de que, como mucho 21 páginas contengan faltas de ortografía?

23. El tiempo de espera para ser atendido telefónicamente en el servicio de averías de una compañía eléctrica se modela según una exponencial de media 2 minutos.

    a) ¿Cuál es la probabilidad de que una persona tarde menos de 2 minutos en ser atendida?

    c) Calcula la probabilidad de que el tiempo total de espera de 50 clientes sea menor a una hora.

24. El número de fallos de acceso a un servidor de base de datos de Xescampus sigue una distribución de Poisson de media 5 fallos cada 10 minutos. Calcula:

    a) La probabilidad de que en 2 minutos haya al menos 1 fallo.

    b) La probabilidad de que el tiempo entre dos fallos consecutivos sea inferior a 30 segundos.

    c) La probabilidad de que haya que esperar más de dos horas para tener 50 fallos.

25. El Servicio de Atención al Cliente de una empresa de telefonía móvil recibe, en promedio, 6 llamadas cada cuarto de hora. Calcula:

    a) La probabilidad de que en una hora se reciban más de 20 chamadas.

    b) La probabilidad de que haya que esperar menos de 1 minuto entre dos chamadas consecutivas.

    c) El tiempo medio de espera hasta recibir 40 llamadas.

26. El tiempo de espera para ser atendido telefónicamente en el servicio de averías de una compañía eléctrica se modela según una exponencial de media 2 minutos.

    a) ¿Cuál es la probabilidad de que una persona tarde menos de 2 minutos en ser atendida?

    c) Calcula la probabilidad de que el tiempo total de espera de 50 clientes sea menor a una hora.

27. En un colegio de primaria se pasa un test de inteligencia y los resultados siguen una distribución normal con media 100 y desviación típica 10.

    a) ¿Cuál es la probabilidad de que un niño elegido al azar haya sacado entre 90 y 115 puntos?

    b) ¿Qué nota se necesita sacar para estar entre el $10\,\%$ de los mejores?

    c) Calcula c para que $\mathbb{P}(105 \leq X \leq c) = 0.10$.

28. La media de llegadas de coches a una estación de servicio es de 3 por minuto.

    a) Calcula la probabilidad de que pase más de 1 minuto hasta que llegue el primer coche.

    b) Si ya han atendido a un coche, ¿cuál es la probabilidad de que el siguiente tarde más de un minuto?

    c) La probabilidad de esperar más de media hora por el vigésimo coche.

    d) La probabilidad de que al menos 5 coches lleguen en 2 minutos.

    e) La probabilidad de que en una hora pasen más de 200 coches.

Boletín 5. Introducción a la inferencia estadística
Estadística - Grado en Ingeniería Informática

1. Las máquinas de una planta de envasado de arroz no están bien calibradas y llenan bolsas según una distribución Normal con desviación típica de 2kg por bolsa.

   a) ¿Cuál es la probabilidad de que en una muestra de 9 bolsas, la media de los pesos supere a la media teórica en más de 1kg? ¿Y en una muestra de 100 bolsas?

   b) ¿Cuál es la probabilidad de que la media de los pesos diste de la media poblacional en menos de medio kilo, en una muestra de 25 bolsas?

2. Un cafetal en Colombia envasa café recomendado para las nuevas cafeteras de Nexpresso, en bolsas estancas cuyo peso medio es de 250gr con una desviación típica de 20gr. Los trabajadores del cafetal echan en las bolsas paladas de café, que en media pesan 1kg, con una desviación típica de 40gr, antes de proceder a cerrarlas a vacío. Después, las bolsas se transportan en cajas de madera, conteniendo 10 bolsas cada una y cuyo peso medio es de 2kg y medio, con una desviación típica de 500gr. Si suponemos que todas las variables siguen una distribución Normal:

   a) Calcula el peso medio de las cajas llenas. ¿Cuál es su varianza?

   b) Para transportar las cajas desde las montañas hasta el puerto más cercano, se utilizan avionetas de la compañía LifeFlight. Por razones de seguridad, no permiten transportar más de 150kg en cada vuelo. ¿Cuál es la probabilidad de que un lote de 11 cajas no sea aceptado?

   c) La empresa A-Wings posee aviones de transporte de mayor tamaño y sólo mide el peso de la mercancía un 5 % de las veces. Si esta supera los 2000kg, se debe reducir la cantidad. ¿Cuál es el número máximo de cajas que enviaríamos a Nexpresso?

3. Un distribuidor de software desea conocer la proporción de clientes dispuestos a adquirir un nuevo programa, del que han descargado la demo. Para ello, se consultó al azar a 100 de ellos, resultando que 30 estarían dispuestos a comprar el programa.

   a) ¿Cuál es la probabilidad de que la proporción muestral de interesados difiera de la real en menos de un 15 %?

   b) Otro estudio de mercado indica que la proporción de usuarios que adquirirían el software es de un 40 %. ¿De qué tamaño es la muestra que debemos elegir para que la proporción muestral de usuarios interesados en el producto diste de la real en menos de un 10 % con una probabilidad del 95 %?

4. Un pozo petrolífero produce diariamente un cierto número de barriles de petróleo, variando de día a día. Se ha observado la producción en 60 días a lo largo de un año, elegidos al azar. Por información de otros pozos, se supone que la desviación típica es de 16 barriles por día.

a) ¿Cuál es la probabilidad de que el número medio de barriles en esos 60 días supere en 10 barriles a la producción media real?

b) ¿Cuál es la probabilidad de que la media de barriles se diferencia de la producción media real en más de 4?

5. Un grupo musical cuelga de su web el single de su último disco, y se cuenta el número de descargas en 17 días, obtiendo que la varibilidad (medida en desviación típica muestral) es de 5 canciones diarias. ¿Cuál es la probabilidad de que el número medio de canciones descargadas en esos 17 días se diferencie de la media real de descargas en menos de 2 canciones?

6. Sabemos que el número de horas que los alumnos de una facultad pasan al día conectados a internet sigue una distribución Normal de desviación típica 3 horas. En una muestra de 10 alumnos, ¿cuál es la probabilidad de que la varianza muestral sea mayor que 3/4 de la varianza real (poblacional)? ¿Y si la muestra fuese de 50 alumnos?

Boletín 6. Intervalos de confianza y contrastes de hipótesis
Estadística - Grado en Ingeniería Informática

1. El peso en gramos de unos envases de detergente varía con una distribución Normal, de desviación típica 5 gramos. Si en una muestra de 16 envases se obtiene una media igual a 503 gramos.

   a) Intervalos de confianza de nivel $90\,\%$, $95\,\%$ y $99\,\%$ para el peso medio.

   b) Si quisiésemos obtener un intervalo de confianza de longitud 1 al $95\,\%$, ¿qué tamaño de muestra deberíamos seleccionar?

   c) Si tampoco conocemos la varianza poblacional, pero de la muestra obtenemos que $S^2 = 38.47$ (gr^2), calcula los intervalos de confianza para el peso medio al $90\,\%$, $95\,\%$ y $99\,\%$.

2. El diámetro de una pieza del motor de turismo, fabricada por una empresa española, debe ser de 3 cm. Sin embargo, dado que las máquinas no son del todo precisas, el diámetro varía según una distribución Normal. De una muestra de 12 piezas, se obtiene una varianza muestral $s^2 = 0.00005$. Obtener un intervalo de confianza al $99\,\%$ para la varianza del diámetro.

3. En una encuesta realizada a 50 internautas que se han descargado la demo de un nuevo videojuego, se ha obtenido que el $76\,\%$ estarían dispuestos a comprarlo.

   a) ¿Entre qué valores tendríamos la proporción real de posibles compradores, con un $95\,\%$ de probabilidad?

   b) ¿Qué tamaño muestral deberíamos considerar para que la longitud del intervalo fuese menor que 2 centésimas?

4. En un reciente estudio sobre la inserción laboral de los titulados universitarios, se ha publicado que el $70\,\%$ encuentran trabajo relacionado con su titulación. De una muestra de 500 titulados, 400 afirman que el trabajo que tienen está relacionado.

   a) Si fijamos un nivel de significación del $5\,\%$, ¿tenemos evidencias suficientes para considerar que la hipótesis de que el $70\,\%$ trabaja en algo relacionado es cierta?

   b) Obtén un intervalo confianza al $99\,\%$ para la proporción teórica de titulados que tienen trabajo relacionado con sus estudios.

5. En la materia de Estadística se estima que el tiempo medio para la realización de una prueba es de 30 minutos, con una desviación de 5 minutos. En la prueba realizada a 36 alumnos, el tiempo medio que necesitaron fue de 35 minutos. Suponemos que el tiempo de realización de la prueba sigue una distribución Normal.

   a) Con una significación del $5\,\%$, ¿era correcta la planificación del tiempo medio?

b) Si no conocemos la variabilidad en la duración de la prueba y sobre la muestra de 36 alumnos se ha obtenido una desviación típica muestral de 7 minutos, ¿podemos considerar correcta la planificación?

6. La oficina de reclamaciones de un productor de grandes electrodomésticos ha recibido bastantes quejas debido a la variabilidad en el diámetro de los tambores de las lavadoras que produce. El responsable de calidad quiere comprobar si la variabilidad en el diámetro (que sigue una distribución Normal) durante este mes es mayor que la que se registró en el mes anterior, que fue de 0.004 cm^2. De una muestra de 20 lavadoras, se obtiene una varianza de los diámetros de 0.005 cm^2. Si fijamos un nivel de significación del $10\,\%$, ¿debe preocuparse el responsable de calidad?

7. De los 100 alumnos de Estadística de primer curso, un grupo de 60 practican algún deporte, mientras que los otros 40 afirman no realizar actividades deportivas. Si medimos la altura de los alumnos en ambos grupos, obtenemos una altura media de 1.80 m en el primer grupo y una altura media de 1.76 m en el segundo. Supondremos que la altura sigue una distribución Normal. Entonces:

   a) Si la variabilidad en las alturas del primer grupo es $\sigma_x^2 = 0.02$ y para el segundo grupo $\sigma_y^2 = 0.015$, ¿podemos afirmar que los alumnos que practican algún deporte son más altos? (significación del $5\,\%$)

   b) Si no conocemos la variabilidad pero admitimos que en ambos grupos es la misma, y las desviaciones típicas que observamos son 0.08m y 0.10m respectivamente, ¿podemos afirmar que los alumnos que practican deporte son más altos, con una significación del $5\,\%$?

8. Un investigador quiere estudiar si la proporción de delincuentes juveniles que llevan gafas es diferente de la proporción de chicos no delincuentes con gafas. La información que recoge se muestra en la siguiente tabla. Si fijamos una significación del $5\,\%$, ¿podemos afirmar que la proporción de no delincuentes con gafas es mayor que la de delincuentes?

	Con gafas	Sin gafas
Delincuente juvenil	1	8
No delincuente juvenil	5	2

9. En un proceso de depuración de las bases de datos de la Agencia Tributaria, se detectaron fallos en el campo correspondiente al DNI. De 500 registros revisados, 35 contenían un DNI erróneo. (En este ejercicio NO puedes utilizar comandos de R para responder)

   a) Obtener un intervalo de confianza con un $95\,\%$ de cobertura para la proporción de DNIs erróneos en las bases de la Agencia Tributaria. Interpreta el resultado obtenido.

b) El ministro responsable defiende que, tras una auditoría independiente, el porcentaje DNIs almacenados incorrectamente es inferior al $4\,\%$. ¿Podemos confiar en el ministro? (considera una significación del $5\,\%$).

10. En una encuesta realizada a 50 alumnos/as de primer curso del GrEI, se deduce que han dedicado al estudio una media de 10.5 horas diarias durante el mes de junio, con una cuasi-desviación de 3.25 horas.

    a) Suponiendo que el tiempo dedicado al estudio se distribuye según una normal, calcula un intervalo de confianza al $99\,\%$ para el tiempo medio diario de horas dedicadas al estudio.

    Por otro lado, en una encuesta realizada a 100 alumnos/as de otras titulaciones, se obtiene que el tiempo medio dedicado es de 9 horas diarias, con una cuasidesviación de 3.5 horas. Suponiendo que las desviaciones sobre el tiempo medio son las mismas en ambos grupos

    b) ¿Se puede afirmar, con una significación del $5\,\%$, que el tiempo medio diario dedicado al estudio es mayor en el alumnado del GrEI que en otras titulaciones?

11. En un análisis sobre contaminantes en aguas subterráneas, se han realizado mediciones en 100 pozos en la comarca del Barbanza, obteniéndose una media de concentración de benceno de 1.61 con una cuasivarianza de 2.3 (en escala logarítmica, para poder suponer una distribución normal).

    a) Obtén un intervalo de confianza para la concentración media de benceno, con un nivel del $95\,\%$.

    b) En la comarca de Santiago, también se han registrado 200 mediciones en otros tantos pozos, con una media de benceno de 1.5 y una cuasivarianza de 4. Ante estos resultados, los gestores del Barbanza indican que la media de benceno es la misma en ambas comarcas. Con un nivel de significación del $10\,\%$, ¿están en lo cierto? ¿Tendríamos la misma conclusión si cambiamos el nivel de significación al $5\,\%$ o al $1\,\%$?

12. En un estudio medioambiental en la comarca de Valdeorras, se han registrado mediciones de benceno en 45 pozos, calificando como contaminados 12 pozos.

    a) Obtener un intervalo de confianza para la proporción de pozos contaminados, con un nivel del $99\,\%$.

    b) Cuando el porcentaje de pozos contaminados supera el $30\,\%$ se produce una intervención por parte del gobierno regional. Con un nivel de significación del $5\,\%$, ¿dirías que es necesaria dicha intervención?