**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

**Extracción aleatoria**

* Permite pasar de población a muestra. Consiste en el cálculo de probabilidades.

**Inferencia estadística**

* Permite pasar de muestra a población
* Consiste en obtener conclusiones sobre la población a partir de la muestra.

**Tablas de frecuencias**

* Utilizadas para representar la información de una muestra tamaño n
* **Clase (ci):** Cada uno de los valores que puede tomar una variable.
* **Frecuencia absoluta (ni):** Número de individuos en la clase ci. 0<=ni<=n
* **Frecuencia relativa (fi):** = 0<=fi<=1
* **Frecuencia absoluta acumulada (Ni):** Número de individuos en ci o en valores anteriores. Nk = n
* **Frecuencia relativa acumulada (Fi)**: . Fk = 1
* Para unha variable cualitativa nominal non se inclúen valores absolutos.
* Para unha variable cuantitativa continua defínense **intervalos**.
  + Denominados intervalos de clase: [ei, ei+1)[[1]](#footnote-0) sendo o punto medio a marca de clase (ci)
  + Se suelen tomar como número de intervalos aproximado.
  + Suelen ser todos de igual longitud y contiguos. La longitud se calcula aproximando hacia arriba para evitar excluir datos.

**Ejemplos tabla de frecuencia**

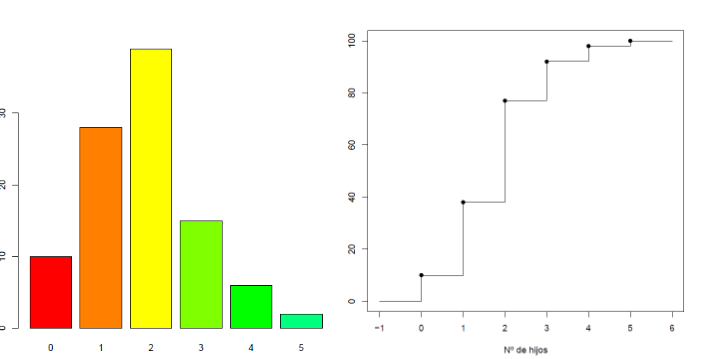
Variable cualitativa ordinal Variable cuantitativa continua

| ci | ni | fi | Ni | Fi |
| --- | --- | --- | --- | --- |
| Baixa | 2 | 0.2 | 2 | 0.2 |
| Media | 5 | 0.5 | 7 | 0.7 |
| Alta | 3 | 0.3 | 10 | 1.0 |

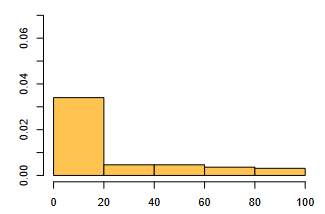
| ei | ni | fi | Ni | Fi |
| --- | --- | --- | --- | --- |
| [2.5, 3.1) | 1 | 0.1 | 1 | 0.1 |
| [3.1, 3.7) | 4 | 0.4 | 5 | 0.5 |
| [3.7, 4.3) | 5 | 0.5 | 10 | 1.0 |

**Representacións gráficas**

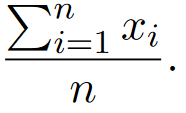
* Variables cualitativas o cuantitativas discretas: **diagrama de barras** o sectores, y diagrama de frecuencias acumuladdas



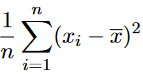
* Variables cuantitativas continuas: **histograma** o diagrama de caja



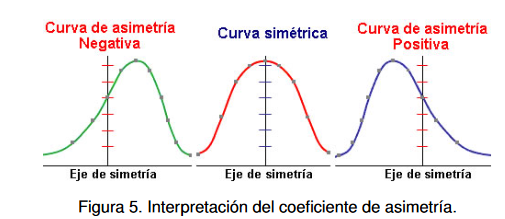
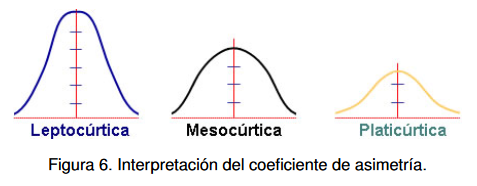
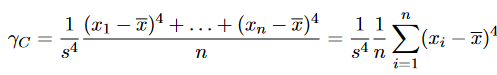
**Medidas de posición**

* 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: 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 q0.45 será o dato que deixa á súa esquerda, como mínimo, o 45% dos datos.
    - En un conxunto de 10 datos, q0.45 e q0.4 serán ambos o 5º dato.
    - Os cuartiles (Q1, Q2, Q3) 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.

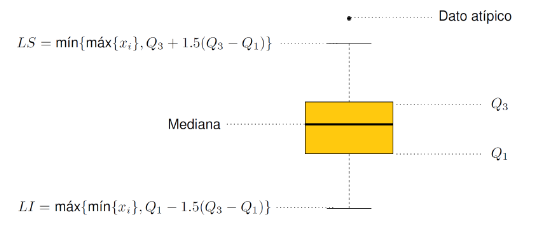
**Medidas de dispersión**

* Recorrido: diferencia entre o máximo e mínimo
  + Recorrido intercuartílico: diferenza entre Q3 e Q1
* Varianza (**s2**):  . A desviación típica é a raíz da varianza, s
  + Siempre positiva.
  + Si se modifican los datos haciendo yi=axi+b, se cumplirá que sy2=a2\*sx2
* Cuasivarianza(**sn-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.
  + 
* Coeficiente de curtosis: mide o grao de apuntamento da distribución 
  + Nunha curva normal é de 3. Maior de 3 é leptocúrtica, menor é platicúrtica.
  + 

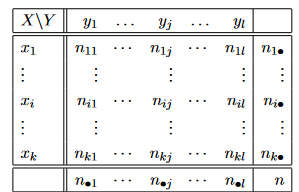
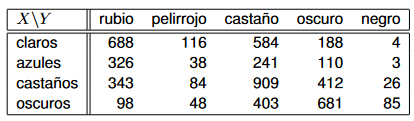
**Diagrama de caixa**

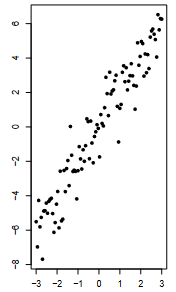
* Representa variables cuantitativas continuas.
* A caixa está delimitada polos cuartiles, sendo a raia do medio o segundo cuartil (a mediana). A altura da caixa é o rango intercuartílico, e contén un 50% dos datos.
* Os bigotes (LS e LI) calcúlanse coa fórmula do gráfico.
* Se quedan datos fóra dos bigotes, considérense atípicos:
  + Defínense outras cotas inferiores e superiores que sexan Q1-3RI e Q3+3RI
  + Se están entre bigotes e a cota especificada, son moderados (\*)
  + Se están fóra das cotas, son atípicos extremos (º)
* 

**Estadística descriptiva bivariante**

* 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

**Tablas de continxencia**

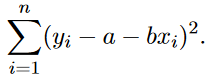
* 
* A suma dos datos dunha columna ou fila concreta denomínanse distribucións marxinais (n**.**1)



**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:** 
  + O signo describe a relación entre as variables: se é directa Sxy>0, se é inversa Sxy<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:** 
  + Dise que a relación é significativa se |rxy| ≥ 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.
  + . A partir de esto: 
  + 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 (R2):** 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.
  + 

**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)

**Conceptos básicos**

* **Espacio muestral (Ω):** conjunto de los posibles sucesos elementales (resultados)
* **Suceso:** Subconjunto del espacio muestral. Ω é un suceso seguro, Ø é imposible.
* **Complementario (Ac)**: cando non sucede A
* **Unión (AUB)**: cando sucede A ou sucede B
  + (AUB)c = Ac n Bc
* **Intersección (AnB)**: cando sucede A e B
  + Sucesos incompatibles se AnB=Ø
  + (AnB)c = Ac U Bc
* **Diferencia (A\B):** sucede A e non B

**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 nA veces.
  + La frecuencia relativa del suceso será nA/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(Ω)=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(Ac) = 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: 
* 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).
* 

**Teorema de Probabilidades totales**

* Si {A1, …, An} es un conjunto completo de sucesos de probabilidad no nula, y B es un suceso cualquiera:
  +  (suma de intersecciones de B con cada Ai)

**Teorema de Bayes**

* Si {A1, …, An} es un conjunto completo de sucesos de probabilidad no nula, y B€A es otro suceso:
* 

**Variables aleatorias**

**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’.
*  Sendo B o conxunto de todos os intervalos de R.

**Propiedades**

* Se X é unha variable sobre  e c é unha constante, c\*X tamén é unha variable.
* Se X e Y son variables sobre , 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)={x1, …, xk})**
* O conxunto de probabilidades de cada suceso denomínase **masa de probabilidade** {p1, …, pk}, onde pi = P(X=xi)
  + A suma de todos os pi 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)=Px(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
* Exemplos:
  + Lanzamento de 2 moedas: F(x)=0 se x<0, 0.25 se 0<=x<1, 0.75 se 1<=x<2, 1 se x>=2
  + Menor valor ao tirar 2 dados: F(x) =0 se x<1, 11/36 se 1<=x<2, 20/36 se 2<=x<3, 27/36 se 3<=x<4, 32/36 se 4<=x<5, 35/36 se 5<=x<6, 1 se x>=6.

**Medidas características (discreta)**

* **Esperanza matemática** (media): E(X)= μ = Sendo K o número de valores xi posibles e pi 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** (Me): Valor x que divide a idstribución en dúas metades iguais. F(Me)=0.5
* **Varianza:** , sendo E(X) a esperanza matemática
  + **Desviación típica:** Raíz da varianza
  + Var(ax + b) = a2Var(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: 

**Independencia de variables**

* Dos variables aleatorias X e Y son independientes si **FX,Y(x,y) = FX(x) \* FY(y)** para todos x,y reales.
* En el caso discreto, se dice que X e Y son independientes si **P(X=xi,Y=yj)=P(X=xi)\*P(Y=yi).**
* En el caso continuo, son independientes si fX,Y(x,y) = fX(x) \* fY(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. **X ∊** **Bi(n,p) [[2]](#footnote-1)**
* A súa función masa de probabilidade será **P(X=x) = (n | x) px (1-p)n-x**
* E(X) = np
* Var(X) = np(1-p)
* **Propiedades:**
  + Para n=1, Ber(p) = Bi(1,p)
  + Si X∈Bi(n1, p), Y∈Bi(n2, p) independientes, entonces X+Y∈Bi(n1+n2, p).
  + Si X ∈ Bi(n, p) entonces X = Xi, donde Xi ∈ 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) / p2

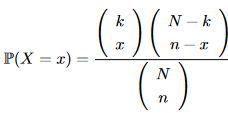
**Distribución binomial negativa**

* 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 \* pn**
* E(X) = n(1-p) / p
* Var(X) = n(1-p) / p2
* 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) =
* 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)}
* ****
* E(X) = = 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 \*

**Distribución uniforme discreta**

* Si X toma valores {x1, …, xk} y todos tienen la misma probabilidad, se dice que X tiene una distribución **uniforme discreta, X ∈ U{x1, …, xk}**
* P(X=x) = 1/K para todos los x.
* E(x) = 1/K \* xi,
* Var(x) = 1/K \* (xi - μ)2

**Resumen distribuciones discreta**s

| Nombre | Definición | MOP | E(X) | Var(X) |
| --- | --- | --- | --- | --- |
| Binomial Bi(n,p) | nº de éxitos tras repetir bernoulli n veces |  | np | np(1-p) |
| Geométrica G(p) | nº de fracasos hasta el primer éxito |  | q/p | q/p2 |
| Binomial negativa BN(n,p) | nº de fracasos hasta el n-ésimo éxito |  | nq/p | nq/p2 |
| Poisson Pois(λ) | nº de sucesos en un intervalo λ |  | λ | λ |
| Hiper - geométrica H(N,n,k) | con una población de N elementos, elementos de clase K en una muestra de tamaño n |  | nk/N = np |  |
| Uniforme U{x1,...xk} | valores con igual probabilidad | 1/K | 1/K \* xi | 1/K\*(xi - μ)2 |

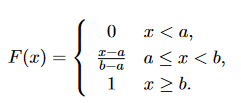
**Variable aleatoria continua**

* 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]](#footnote-2) 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, .

**Medidas características (continua)**

* **Esperanza matemática** (media): E(X)= μ = 
  + Exemplo (moedas): 1/4 \*0 + 1\*2\*1 + ¼\*2 = 1
* **Mediana** (Me): Análogo a caso discreto. F(Me)=0.5
* **Moda:** Valor para el cual la densidad alcanza un máximo relativo. Mo=x0 si se cumple que f’(x0)=0, y f’’(x0)<0.
* **Varianza:** , sendo E(X) a esperanza matemática.
  + , 

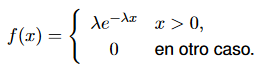
**Distribución uniforme continua**

* 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 será:
* E(X) =
* Var(X) =
* Su función de densidad será un segmento plano sobre el soporte de la variable, cuya altura será inversa a la longitud del soporte.

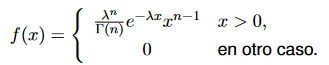
**Distribución Normal**

* Una variable aleatoria tiene distribución **normal**, X ∈ N (μ, σ2),si su función de densidad es:
* Se suele considerar la **distribución normal estándar**, de media 0 y varianza 1, X ∈ N(0,1), denotada por Φ(z).
  + 
  + Todas las variables normales se pueden transformar a una estándar, restando la media y dividiendo por la desviación típica. 
    - Ejemplo: si queremos calcular P(X<28), siguiendo X una distribución normal, es lo mismo que calcular P(Z<), siendo Z∈(0,1).
* Las distribuciones normales son aditivas: si tomamos X ∈ N(μ1, σ21) e Y∈N(μ2,σ22), entonces X+Y ∈ N (μ1+μ2, σ21+σ22)

**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) = 1 - e-λx
* E(X) = 1/λ (λ es el número esperado de sucesos por unidad de tiempo)
* Var(X) = 1/λ2

**Distribución Gamma[[4]](#footnote-3)**

* Generalización de la distribución exponencial: mide el tiempo hasta la aparición del n-ésimo suceso. **X ∈ Γ(n, λ)**
* Su función de densidad será:
* E(X) = n/λ
* Var(X) = n/λ2
* Dos variables Gamma con el mismo λ son aditivas.

**Resumen distribuciones continuas**

| Nombre | Definición | f(x) | F(x) | E(X) | Var(X) |
| --- | --- | --- | --- | --- | --- |
| Uniforme U(a,b) | probabilidad constante dentro de un intervalo |  |  |  |  |
| Normal N(0, 1) | - |  | área bajo la curva | - | - |
| ExponencialExp(λ) | Tiempo promedio entre sucesos de un proceso de Poisson de pár.λ |  |  | 1/λ | 1/λ2 |
| Gamma Γ(n,λ) | tiempo hasta la aparición del n-ésimo suceso |  | - | n/λ | n/λ2 |

**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 {Xi}i∊N una sucesión de variables aleatorias independientes e idénticamente distribuidas, con E(Xi)=μ y Var(Xi)=σ2 para todo i. Entonces, para n suficientemente grande:
* Por ejemplo, si X∊Bi(n,p), se puede escribir como , siendo cada Xi∊Ber(p).
  + Entonces, E(Xi)=p y Var(Xi)=pq para todo i.
  + Aplicamos el teorema central del límite, y obtenemos que X se puede aproximar por:
* Otro ejemplo, una distribución Gamma se puede expresar como , siendo cada Xi∊Exp(λ)
  + E(Xi)=1/λ y Var(Xi)=1/λ2 para todo i.
  + Aplicamos el teorema central del límite y obtenemos que X se puede aproximar por: 

**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**:

**Aproximación de distribuciones**

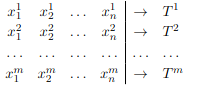
| **Distribución** | **Caso** | **Aproximación** |
| --- | --- | --- |
| Bi(n,p) | n>=30, p<0.1 | Pois(np) |
| n>=30, p∊[0.1, 0.9] | N(np, npq) |
| n>=30, p>0.9 | Se transforma en Bi(n,1-p) |
| Pois(λ) | λ>=10 | N(λ,λ) |

**Inferencia estadística**

**Conceptos**

* Una **muestra** es un subconjunto representativo de la población.
* Sea **X** la variable que estudiamos. Denotaremos por {X1, …, Xn} una **muestra aleatoria simple (m.a.s)**, donde cada variable Xi tomará un valor xi.
  + Al conjunto de valores que toma la muestra {x1, …, xn} 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]](#footnote-4).
  + El parámetro puede ser unidimensional (caso previo) o bidimensional (por ejemplo, en una N(μ, σ2))
  + Denotamos por θ al parámetro de interés en la población. Se dirá que la variable tiene distribución F(θ).
* **Estadístico:** Cualquier función de la muestra (media, varianza, máximo…) Se denotan por T(X1, …, Xn)
  + **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**

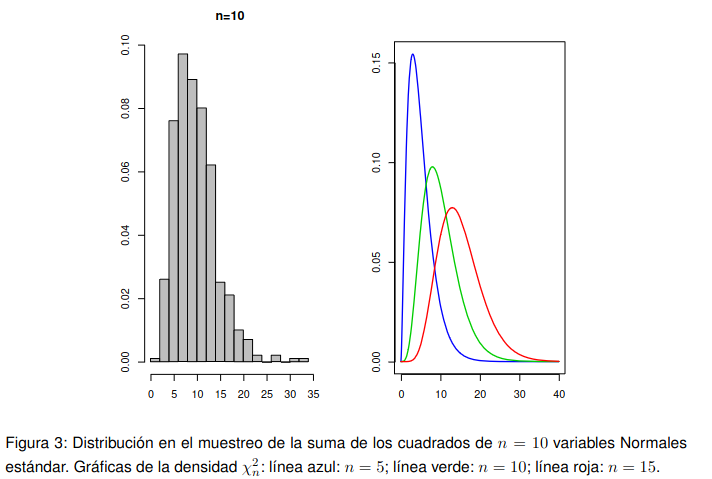
* Sea X1, …, Xn una muestra aleatoria simple de una v.a. X con distribución F(θ).
* Se considera que la muestra sigue la misma distribución que la variable aleatoria.
* Consideremos que F es una distribución normal. Entonces, debemos calcular su media y su varianza.
* Definimos T(X1, …, Xn) como un estadístico. 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 X1,..Xn, denotamos por T1, …, Tm los distintos valores del estadístico en cada una de las muestras.
  + Estos valores de T1, …, Tm seguirán una distribución.

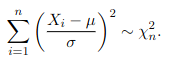
**Proporción muestral**

* La **proporción muestral** es la relación de casos de éxitos en una muestra respecto al tamaño de la muestra. Se denota con . (p^ nestes apuntes)
* Si me interesa conocer la proporción de elementos p de una población X ∼ Ber(p), seleccionamos una MAS X1,...,Xn de variables Ber(p).
* **Distribución:**  → 

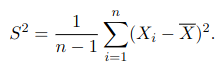
**Media muestral (varianza conocida)**

* Supongamos que disponemos de una m.a.s X1, …, Xn tal que X∼ N (μ, σ2). La **media muestral** Xbarra=  se puede escribir como la suma total de n términos .
  + **Nota:** no confundir con la **media real** de la población, denotada por **μ**
* Debido a esto, la media muestral sigue una distribución normal:  que se puede tipificar como 
  + Es decir, los posibles valores de Xbarra 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 χ2 (ji-cuadrado)**

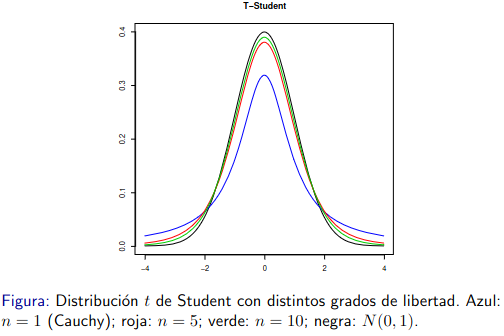
* Supongamos que disponemos de una m.a.s X1, …, Xn de variables Xi∼N(μ,σ2) Entonces, 
* Cuando la muestra es lo suficientemente grande, una distribución χ2n se puede aproximar por una N(n,2n)
* La distribución χ2n **no** es simétrica.

**Distribución de la varianza/cuasivarianza muestral**

* El **Teorema de Fisher** establece que si X1, ..., Xn es una m.a.s. de variables normales con varianza σ2, entonces X y s2 son independientes y además: (imagen) (s2 es la varianza muestral)
* Si en su lugar queremos utilizar la **cuasivarianza** S2,

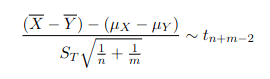
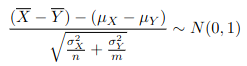
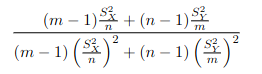
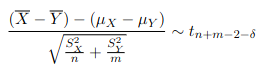
**Distribución t de Student**

* Consideremos una variable X ∼ N (0, 1) y otra v.a. Y ∼ χ2n independientes, el cociente: 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.



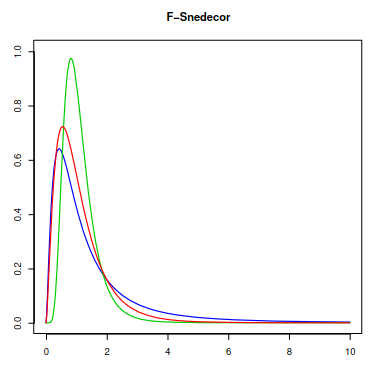


**Distribución de la diferencia de medias**[[6]](#footnote-5)

* Supongamos ahora que tenemos dos poblaciones y las correspondientes muestras X1, . . . , Xn m.a.s. de X ∼ N (μX , σ2X) e Y1, . . . , Ym una m.a.s. de Y ∼ N (μY , σ2Y ).
* Si σ2X y σ2y son conocidas: 
* Si σ2X y σ2y son desconocidas, pero iguales: ,
* Si σ2X y σ2y son desconocidas, y distintas

, sendo δ o enteiro máis proximo a

**Distribución F de Snedecor**

* Sexa X unha va con distribución χ2n e Y con distribución χ2m, ambas independientes. O cociente 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.

**Distribución do cociente de varianza**

* Polo teorema de Fisher, coñecemos que 
* Entón, o cociente de varianzas ten distribución F con n-1 e m-1 graos de liberdade:

**Intervalo de confianza**

* Dada unha m.a.s X1, …, Xn de X ∼ Fθ, un **intervalo de confianza de nivel (1-α)** con α € (0,1) é un intervalo aleatorio tal que[[7]](#footnote-6)
* 
  + 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]](#footnote-7)

* 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 ∼ N(0,1). Denominamos esta nova m.a.s como p2.
  + p2 coñécese como o **estadístico pivote:** o elemento do que partimos para obter o intervalo de confianza.
* Non podemos calcular directamente o intervalo P(L1<=p<=L2)=1-α porque non coñecemos a distribución de p, pero si coñecemos a de p2.
* P(-z1-α/2 <= p2 <= z1-α/2) = 1-α. Ahora, despexamos p de dentro de p2. Finalmente, substituímos todas as instancias de p por p^ para que o intervalo non dependa de p (que é descoñecido)
  + **Nota:** zp denota o valor tal que P(N(0,1)<=zp)=p.
  + Por exemplo, se se pedise un intervalo de confianza 90%, teríamos 1-α=0.9→α=0.1 →Calculamos z0.95, é decir, consultamos a tabla para ver que valor de zp 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 zp=1.755.
* O intervalo final obtido é simétrico respecto de p^:[[9]](#footnote-8)

| (p^ - z1-α/2 , p^ + z1-α/2 ) |
| --- |

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

| **Ic =** |
| --- |

**Intervalo de confianza para a media** con varianza coñecida

* Sexa X1, …, Xn unha m.a.s de X∼N(μ,σ2) de varianza coñecida. Entón, ∼N(0,1). será o noso estadístico pivote.[[10]](#footnote-9)
* O intervalo do que partimos é P(-z1-a/2, , z1-a/2) = 1-a.

| **Ic =** |
| --- |

**Intervalo de confianza para a media** con varianza descoñecida

* Se non coñecemos a varianza, odemos utilizar ∼tn-1. Este será o estadístico pivote.
* O intervalo do que partimos é P(-tn-1, 1-a/2, , tn-1, 1-a/2) = 1-a.

| **Ic =** |
| --- |

**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 z1-0.1/2= z0.95. mirase na tabla o valor de z que da p=0.95, aproximase a 1.645.
  + Ic = (503-1.645\*5/, 503+1.645\*5/) = (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 t15, 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)
  + Ic =

**Intervalo de confianza para a varianza**

* Coñecemos que nS2/σ2 ∼ χ2n-1.. Este será o noso estadístico pivote.

| ) |
| --- |

**Lonxitude dun intervalo de confianza**

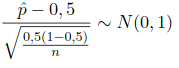
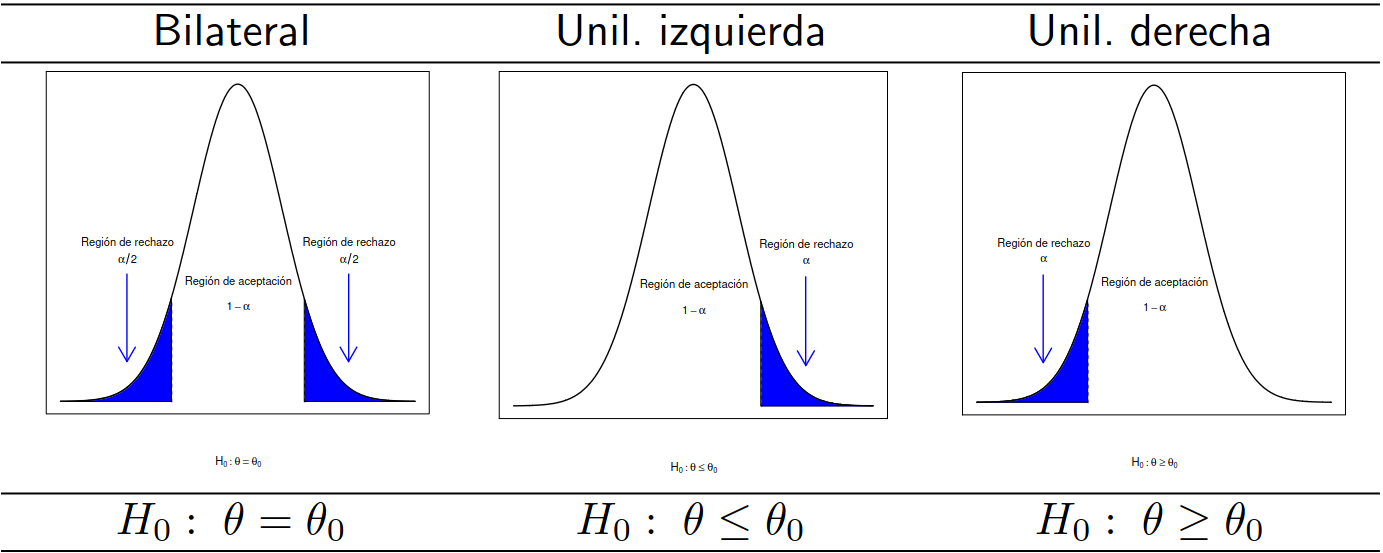
* Para calqueira intervalo excepto o da varianza[[11]](#footnote-10), o valor da lonxitude do Ic será 2\*(o extremo superior do intervalo).
* Despexando n na fórmula resultante, podemos calcular n para que un Ic teña unha lonxitude L. (ver fórmulas)

**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 (μ=μ0) o unilaterales (μ≤μ0 o μ≥μ0) 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 (H0):** Hipótesis que se desea contrastar. Se compara con la **hipótesis alternativa (Ha)**. Por ejemplo, si H0: μ=μ0, entonces la alternativa será Ha: μ≠μ0.

**Estadístico de contraste**

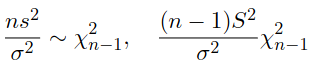
* Llamamos **estadístico de contraste** al que tiene distribución conocida cuando H0 es cierta.
  + Por ejemplo, si nuestra hipótesis H0 es p=0.5, sabemos que si H0 es cierta, entonces:
* **Región crítica:** Conjunto de valores del estadístico que nos llevan a rechazar H0
* **Región de aceptación:** Conjunto de valores del estadístico que nos llevan a aceptar H0.
* El **nivel de significación (α)** es la área de la región de rechazo. A mayor α, más exigente el estudio.

**Errores**

* **Tipo I:** Rechazar H0 cuando es válida
  + **Nivel de significación (α):** Probabilidad de cometer el error de tipo I[[12]](#footnote-11).
* **Tipo II:** Aceptar H0 cuando es falsa
  + **β:** Probabilidad de cometer el error de tipo II.
  + **Potencia del contraste (1-β):** Probabilidad de rechazar H0 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 H0 y la alternativa Ha.
   * H0: s2>=0.004, H1. s2<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.
   * 
5. 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 (s2>=0.004), la región de aceptación es la que queda por la derecha del valor.
     1. RR: (-inf, 11.65). RA:(11.65, +inf)
6. Calculamos el valor del estadístico, asumiendo que H0 es cierta.
7. Observamos a qué región pertenece, y concluímos el contraste.
   * 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 |  |
| X∼Ber(pX), Y∼Ber(pY) | (pX - pY) |  |
| N (μ, σ2) | μ |  |
|  |
| σ2 |  |

1. el último intervalo, como excepción, puede ser cerrado por la derecha: [ek-1, ek] [↑](#footnote-ref-0)
2. Significa ‘X sigue una distribución binomial de parámetros n y p’. [↑](#footnote-ref-1)
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. [↑](#footnote-ref-2)
4. quoted saying que non entra no examen [↑](#footnote-ref-3)
5. Se asume que n es conocido. [↑](#footnote-ref-4)
6. estas formulas danas no examen let’s fucking GOOOO [↑](#footnote-ref-5)
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.’ [↑](#footnote-ref-6)
8. Lembrar que p denota a proporción da poboación (descoñecida, queremos estimar o seu valor), e p^ denota a proporción muestral (coñecida) [↑](#footnote-ref-7)
9. en realidade, os p^ de dentro das raices serían p, pero como non o coñecemos asumimos que p^ é close enough [↑](#footnote-ref-8)
10. Xmedia é a x cunha barra arriba non sei escribila en gogoeld cdocs [↑](#footnote-ref-9)
11. Debido a que a distribución que emprega non é simétrica e as outras si. [↑](#footnote-ref-10)
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 [↑](#footnote-ref-11)