Matemática Estadística

Oihana Aristondo Echeberria
José David Núñez González
Lucía Porlán Ferrando

eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea

Matemática Estadística

Matemática Estadística

Oihana Aristondo Echeberria
José David Núñez González
Lucía Porlán Ferrando

eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

CIP. Biblioteca Universitaria

Aristondo Echeberria, Oihana

Matemática estadística [Recurso electrónico] / Oihana Aristondo Echeberria, José David Nuñez González, Lucía Porlán Ferrando. – Datos. – [Leioa] : Universidad del País Vasco / Euskal Herriko Unibertsitatea, Argitalpen Zerbitzua = Servicio Editorial, [2025]. – 1 recurso en línea: PDF (114 p.). – (Unibertsitateko Eskuliburuak = Manuales Universitarios)
Modo de acceso: World Wide Web
ISBN: 978-84-9082-947-9

1. Estadística matemática. I. Núñez González, José David, coaut. II. Porlán Ferrando, Lucía, coaut.

(0.034)519.2

Escuela de Ingeniería de Gipuzkoa, Sección Eibar
Departamento de Matemática Aplicada

ISBN: 978-84-9082-947-9

Índice general

1. Estadística descriptiva

1.1. Población y Muestra

Al grupo formado por los objetos principales de cada estudio estadístico se le llama **población** o **universo**. Para que una población esté bien definida es necesario saber si algún elemento en particular pertenece a la población o no. Además, las poblaciones pueden ser de dos tipos: finitas o infinitas.

$$\Omega = \text{población}$$

Una **muestra** es un subconjunto representativo de la población. El número de elementos de la muestra se denomina tamaño de muestra.

$$\Omega' = \text{muestra, donde } \Omega' \subset \Omega$$

Para que los resultados de una población sean confiables, elegimos como muestra la subpoblación más representativa.

1.2. Tipos de Características

Lo interesante en un investigación es analizar las características de la muestra de forma conjunta. Los valores de las características pueden ser números o atributos, por lo que, se tendrá la siguiente clasificación:

- **Características cualitativas:** cuando los valores son atributos.
- **Características cuantitativas:** cuando los valores son números. Esta categoría se subdivide en:
  - **Características discretas:** cuando los valores comprenden cantidades finitas de números.
  - **Características continuas:** cuando se toma cualquier valor dentro de un intervalo.

Ejemplos:

1. Categoría profesional de un taller → Característica cualitativa.
2. Tipos de combustible → Característica cualitativa.
3. Si un taller donde elaboran platos tomamos 30 platos y queremos saber cuántos defectos tiene cada uno. Los valores de la característica se van a tomar en el conjunto {0, 1, 2, 3, ...}. → Caraterística cuantitativa discreta.

4. Si se toma una muestra en un taller de tornillos para medir el diámetro de los tornillos. (El valor de la característica se puede tomar dentro de un intervalo) → Característica cuantitativa continua.

1.3. Características Cuantitativas

1.3.1. Distribución de frecuencias

Datos no agrupados o discretos

Supongamos que tenemos una muestra de tamaño n, tenemos n dimensiones y antes de trabajar tenemos que ordenar estos valores de menor a mayor, y solo tomaremos valores diferentes $\{x_1, x_2, ..., x_m\}$, donde $x_1 < x_2 < ... < x_m$.

Colocaremos estos valores y sus frecuencias en la tabla de distribución de frecuencias:

Valores	Frecuencia absoluta	Frecuencia relativa	Frecuencia absoluta acumulada	Frecuencia relativa acumulada
x_1	n_1	f_1	N_1	F_1
x_2	n_2	f_2	N_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_j	n_j	f_j	N_j	F_j
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_m	n_m	f_m	$N_m = n$	$F_m = 1$
Total	n	1		

Cuadro 1.1: Tabla de distribución de frecuencias de datos no agrupados

- **1ª columna:** Valores de las frecuencias.
- **2ª columna:** Frecuencia absoluta de cada valor x_j, y lo indicaremos con el símbolo n_j. Se completará $\sum_{j=1}^{m} n_j = n$. Es decir, el valor x_j se repite n_j veces.
- **3ª columna:** Frecuencias relativas. Utilizaremos el símbolo $f_j = \dfrac{n_j}{n}$, donde $\sum_{j=1}^{m} f_j = 1$. Estos datos se utilizan para comparar diferentes muestras.
- **4ª columna:** Frecuencias absolutas acumuladas, N_j. Su definición es la siguiente: $N_j = \sum_{k=1}^{j} n_k$. Entonces $N_m = n$.
- **5ª columna:** Frecuencias relativas acumuladas que se indicarán con F_j, donde $F_j = \dfrac{N_j}{n}$, y además, $F_j = \sum_{k=1}^{j} f_k$, entonces $F_m = 1$.

Nota: Las frecuencias relativas pueden expresarse como porcentajes.

Ejemplo: Si la frecuencia relativa f_i es 0,15, entonces al multiplicar f_i por 100 obtenemos 15. Esto significa que el valor x_i representa el 15 % de la muestra.

Por otro lado, si la frecuencia acumulada F_i es 0,46, al multiplicar F_i por 100 obtenemos 46. Esto indica que el valor x_i es menor o igual al 46 % de todos los valores de la muestra.

En resumen:

- La frecuencia relativa muestra el porcentaje de la muestra que corresponde al valor x_i.
- La frecuencia acumulada muestra el porcentaje de la muestra que es menor o igual al valor de x_i.

Datos agrupados o continuos

Cuando el número de valores diferentes de la muestra es mayor o igual a 30, los datos se suelen agrupar en intervalos de clases para facilitar el trabajo. Entonces, en lugar de x_j, normalmente se localiza el punto medio del intervalo, k_j. Este valor se denomina representante de clase.

Hay que tener en cuenta algunos puntos para disponer de una buena agrupación:

- Se debe seleccionar la anchura del intervalo de clases que proporciona el error mínimo.
- El recorrido de la muestra se redondea a un múltiplo del ancho de los intervalos de clase. Siendo el recorrido de la muestra la diferencia entre el valor más alto y el valor más bajo.

Ejemplos:

1. Si el recorrido es $x_m - x_1 = 0,33$ y el ancho del intervalo es $0,05$ entonces sumaremos $0,02$ al recorrido: en lugar de x_1 pondremos $x_1 - 0,01$ y en lugar de x_m pondremos $x_m = x_m + 0,01$. De esta forma el recorrido redondeado será de $0,35$ y los 7 espacios tendrán $0,05$ de ancho. Para que una medida no sea igual a un extremo, es muy conveniente que los extremos de los intervalos tengan un decimal más que el tamaño de la muestra.
2. En el ejemplo anterior, el recorrido $x_1 = 12,20$ y $x_m = 12,53$ es $x_m - x_1 = 12,54 - 12,19 = 0,35$ tendrá los siguientes intervalos:

(12,19 - 12,24); (12,24 - 12,29); (12,29 - 12,34); (12,34 - 12,39);
(12,39 - 12,44); (12,44 - 12,49); (12,49 - 12,54)

Sus representantes de clase:

$$k_1 = 12,215;\ k_2 = 12,265;\ k_3 = 12,315;\ k_4 = 12,365;\ k_5 = 12,415;$$
$$k_6 = 12,465;\ k_7 = 12,515$$

Después de hacer esto, se realiza la tabla de distribución de frecuencias como el caso anterior, pero dispondrá de una columna más.

Intervalos	Representante de clase	Frecuencia absoluta	Frecuencia relativa	Frecuencia absoluta acumulada	Frecuencia relativa acumulada
$(x_1, x_2]$	k_1	n_1	f_1	N_1	F_1
$(x_2, x_3]$	k_2	n_2	f_2	N_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$(x_j, x_{j+1}]$	k_j	n_j	f_j	N_j	F_j
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$(x_m, x_{m+1}]$	k_m	n_m	f_m	$N_m = n$	$F_m = 1$
Total		n	1		

Cuadro 1.2: Tabla de distribución de frecuencias de datos agrupados

- **1ª columna:** Intervalos de frecuencias x_i, x_{i+1}.
- **2ª columna:** Representantes de clase. Donde $k_i = \dfrac{x_i + x_{i+1}}{2}$ es el número de valores de muestra en el intervalo n_i, $(x_i,\ x_{i+1}]$. En general, para dejar claro dónde se ubica cada valor los intervalos son semiabiertos.
- El resto de columnas representan las frecuencias relativas, absolutas y acumuladas de la misma forma que en el Cuadro 1.1.

1.3.2. Descripciones gráficas

Aunque podemos ver en las tablas cómo se comporta la característica que estamos analizando, entenderemos los datos más fácilmente a través de descripciones gráficas.

Diagrama de barras y polígonos de frecuencia absoluta o diagramas de líneas

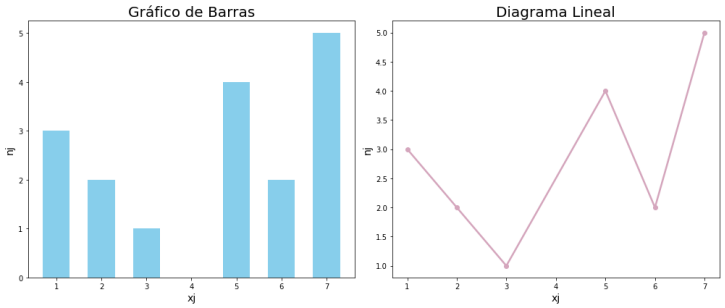
Cuando hay datos no agrupados en una muestra podemos construir dos gráficas cartesianas usando las dos primeras columnas de la tabla. Los diferentes valores de la característica se ubican en el eje OX y la frecuencia de cada valor en el eje OY, formando los puntos (x_j, n_j).

Para hacer **gráficos de barras** conectamos los puntos $(x_j, 0)$ y (x_j, n_j) con una barra.

Para hacer **polígonos de frecuencias** o **diagramas lineales** $(x_1, n_1); ...; (x_m, n_m)$ estos puntos están conectados por segmentos de recta.

Ejemplo: A partir de los datos de la tabla realizar un gráfico de barras y un diagrama lineal.

x_j	n_j
1	3
2	2
3	1
5	4
6	2
7	5



Nota: Se pueden realizar las mismas gráficas para representar tanto frecuencias relativas como acumuladas.

Histogramas y polígonos de frecuencia

Esta representación gráfica se utiliza cuando hay datos agrupados. La interpretación de la construcción de los **histogramas** la clasificaremos en función de la anchura de los intervalos:

1. Cuando todos los intervalos tienen la misma anchura:

   En los ejes cartesianos dibujaremos rectángulos de la misma base y de diferentes alturas, donde la base será la anchura del tramo y la frecuencia será la altura.

   Colocando los intervalos en el eje OX y la frecuencia del intervalo en el eje OY.

2. Cuando los intervalos tienen diferente anchura:

   En este caso se hará de forma similar al caso anterior, pero la altura de cada rectángulo será la medida de su frecuencia entre su anchura, es decir, $\frac{n_i}{z_i} = h_i$ donde z_i es el ancho del intervalo.

Nota: Los histogramas también se pueden utilizar para representar frecuencias relativas y acumuladas.

Para hacer **polígonos de frecuencia**, tomamos puntos en cada intervalo (k_i, n_i) y los conectamos con segmentos rectos.

1.4. Características Cualitativas

1.4.1. Distribución de frecuencias

Cuando la muestra sea cualitativa funcionará como datos discretos. En este caso pondremos los atributos o cualidades en la primera columna.

Atributos	Frecuencia absoluta	Frecuencia relativa	Frecuencia absoluta acumulada	Frecuencia relativa acumulada
1. Atributo	n_1	f_1	N_1	F_1
2. Atributo	n_2	f_2	N_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
j. Atributo	n_j	f_j	N_j	F_j
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
m. Atributo	n_m	f_m	$N_m = n$	$F_m = 1$
Total	n	1		

Cuadro 1.3: Tabla de distribución de frecuencias para características cualitativas.

- **1ª columna:** Atributos.
- El resto de columnas representan las frecuencias relativas, absolutas y acumuladas de la misma forma que en el Cuadro 1.1.

1.4.2. Descripciones gráficas

Para representar características cualitativas veremos dos tipos de gráficos:

- Diagrama de sectores
- Pictograma

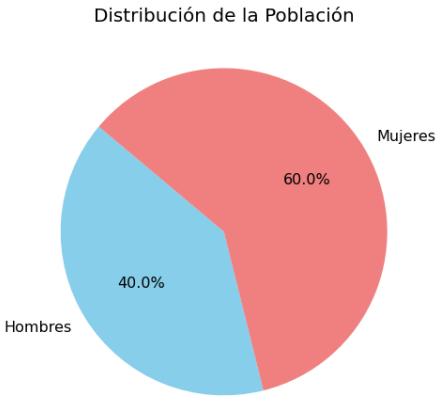
Diagrama de sectores

En esta expresión, la frecuencia de cada categoría está representada por una fracción del área de la circunferencia.

Para ello, calcularemos los ángulos usando la siguiente regla de tres:

$$\text{Total} = \sum_{i=1}^{n} x_i, \;\; \text{Ángulo}_i = \frac{x_i}{\text{Total}} \; 360^\circ \tag{1.1}$$

Ejemplo: Realizar el diagrama de sectores de una población donde 60.000 son hombres y 90.000 son mujeres.

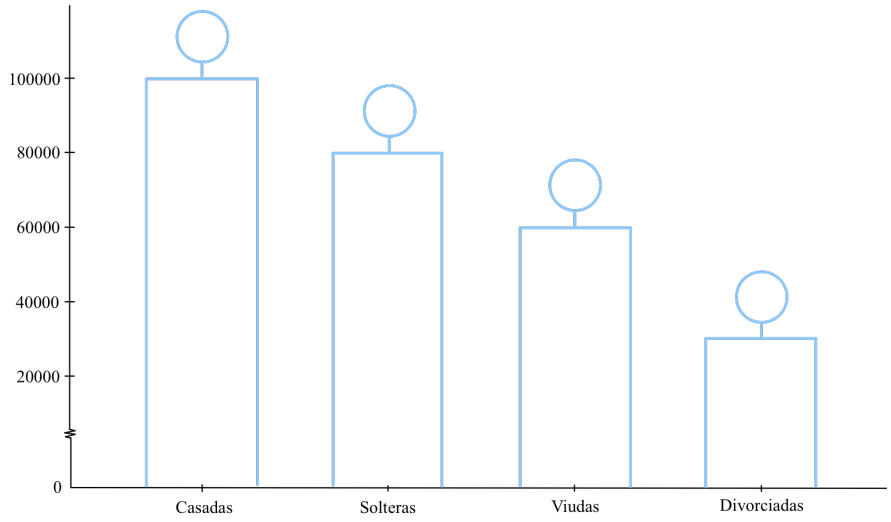


Pictograma

Los **pictogramas** se realizan mediante dibujos, el tamaño de los dibujos indica la frecuencia de la categoría.

Ejemplo: Se está analizando cuántas casadas, solteras, divorciadas y viudas hay en un pueblo y se han obtenido los siguientes resultados:

- Casadas: 100.000
- Solteras: 80.000
- Viudas: 60.000
- Divorciadas 30.000



1.5. Principales estatígrafos de muestras cuantitativas simples

Para describir las características de un conjunto de datos se utilizan tres tipos de estatígrafos:

- Medidas de centralización.
- Medidas de dispersión.
- Medidas de asimetría.

1.5.1. Medidas de centralización

Estos estatígrafos se utilizan para localizar el valor central de la muestra en términos de moda, mediana y media.

Moda

La característica más frecuente es la **moda**. Se toma de la tabla de distribución de frecuencias y se indica con el símbolo M_o. Si hay dos valores diferentes con la misma mayor frecuencia, la distribución será bimodal. En este caso se pueden tomar ambos valores, o establecer un criterio de selección (media, inferior, superior, al azar,...)

Pero cuando hay datos agrupados se calcula mediante la siguiente fórmula:

$$M_o = L_{i-1} + \frac{n_{i+1}}{n_{i+1} + n_{i-1}}\ z_i \tag{1.2}$$

donde L_{i-1} es el extremo inferior, n_{i+1} es la frecuencia absoluta del intervalo posterior, n_{i-1} es la frecuencia absoluta del intervalo anterior y z_i es la amplitud de la clase.

Mediana

Se indica con Me y representa el valor de la variable de posición central en un conjunto de datos ordenados. Los valores inferiores a este valor constituyen el 50 % de la muestra total. Es decir, cuando $F_{Me}\ 100 = 50$. Si la muestra tiene un número par de elementos, la mediana es la media entre los dos elementos centrales.

Cuando hay datos agrupados, la explicación anterior es válida para saber en qué intervalo se ubica la mediana, pero calcularemos el valor con la siguiente fórmula:

$$Me = L_{i-1} + \frac{n/2 - N_{i-1}}{n_i}\ z_i \tag{1.3}$$

donde L_{i-1} es el extremo más pequeño del intervalo en el que se encuentra la mediana, N_{i-1} es la frecuencia absoluta acumulada del intervalo anterior en el que se encuentra la mediana, n_i es la frecuencia absoluta del intervalo en el que se encuentra la mediana y z_i es la amplitud de la clase.

Media aritmética

La **media aritmética** es el estatígrafo dominante de una muestra, se denota con el símbolo $\overline{x}$ y se define mediante la siguiente fórmula:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{1.4}$$

Pero cuando hay valores repetidos en la muestra, su fórmula varía de la siguiente manera:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i\ n_i}{n} \tag{1.5}$$

Diferenciaremos en la notación la media aritmética muestral y poblacional con $\overline{x}$ y μ respectivamente.

Notas:

1. Cuando la muestra tiene pocos valores diferentes y no intercalados, utilizaremos las siguientes tablas:

   Ejemplo:

x_i	n_i	$x_i \cdot n_i$
2,3	10	23
4,5	6	27
1,5	4	10
Total	20	60

   En estos casos: $\overline{x} = \dfrac{60}{20} = 3$

2. Si los valores de la muestra están desagrupados y son muy grandes, se realizará un cambio de variable. A cada valor muestral se le restará un valor fijo "a" para que los resultados sean pequeños, es decir, $x'_i = x_i - a$.

   Entonces, $\sum_{i=1}^{m} x'_i = \sum_{i=1}^{m} x_i - a$ y dividiendo por n: $\overline{x}' = \overline{x} - a \rightarrow \overline{x} = \overline{x}' + a$.

   Ejemplo: a $= 126$ eligiendo el punto central $\overline{x} = \dfrac{-17}{30} = -0{,}566$ entonces $\overline{x} = -0{,}566 + 126 = 125{,}434$.

x_i	n_i	f_i	N_i	$x'_i = x_i - a$	$x'_i \cdot n_i$
121	5			-5	-25
122	4			-4	-16
123	3			-3	-9
124	2			-2	-4
126	1			0	0
127	6			1	6
128	4			2	8
130	3			4	12
131	1			5	5
132	1			6	6
Total	30				-17

3. Cuando los datos de la muestra sean intervalos agrupados de anchura b, los valores utilizados para calcular la media serán los representantes de la clase.

Media geométrica

La **media geométrica** rara vez se utiliza, uno de sus usos sería para el cálculo de los intereses bancarios. Se escribe con la letra G y su fórmula es la siguiente:

$$G = \sqrt[n]{x_1^{n_1}\ x_2^{n_2}\ ...\ x_m^{n_m}} \tag{1.6}$$

Media armónica

La **media armónica** al igual que el anterior, es muy rara y poco utilizada. Se escribe con la letra H y su expresión es la siguiente:

$$H = \frac{n}{\sum \frac{1}{x_i}\ n_i} \tag{1.7}$$

Media cuadrática

La **media cuadrática**, también conocida como desviación estándar o RMS (Root Mean Square), se utiliza frecuentemente en estadística y en análisis de datos para medir la variabilidad de un conjunto de valores. Se escribe con la letra Q y su fórmula es la siguiente:

$$Q = \sqrt{\frac{1}{n}\ \sum_{i=1}^{n} x_i^2} \tag{1.8}$$

1.5.2. Medidas de dispersión

Recorrido

Está entre el valor más alto y el más bajo.

$$\text{Recorrido} = x_m - x_i \tag{1.9}$$

Percentiles

Es un concepto similar al de la mediana, pero en lugar de dividir los datos por la mitad, los dividen en porcentajes. En concreto, si los datos se ordenan de menor a mayor, el percentil de orden k es el valor que deja $k\,\%$ de los datos a su izquierda. En este caso, tendremos $k = 1, 2, 3, ..., 99$ percentiles.

- En el caso discreto, simplemente se aplica la definición.
- En el caso continuo, se deberá aplicar la siguiente fórmula:

$$P_k = l_i + \frac{\frac{k\ n}{100} - N_{i-1}}{n_i}\ z_i \tag{1.10}$$

De manera similar a cómo se definen los percentiles, también se definen deciles y cuartiles.

- Deciles: Específicamente, el decil de orden k deja el 10 % de los datos de distribución a su izquierda, con $k = 1, 2, ..., 9$. Esto es: $D_1 = P_{10}$, $D_2 = P_{20}$, ..., $D_9 = P_{90}$, $D_{10} = P_{100}$.
- Cuartiles: Por definición, el cuartil de orden k deja el 25 % de la distribución con datos a su izquierda, con $k = 1, 2, 3$. Esto es: $Q_1 = P_{25}$, $Q_2 = P_{50}$, $Q_3 = P_{75}$, $Q_4 = P_{100}$.

Evidentemente, el segundo cuartil es la mediana: $Q_2 = Me$. En el caso de los cuartiles, se define el rango intercuartil, que es la distancia entre los cuartiles de tercer y primer orden: $R_1 = Q_3 - Q_1$.

Varianza y Desviación estándar

El estatígrafo que mide el nivel de dispersión respecto de la media aritmética de los valores muestrales se llama **varianza** y se denorta con el símbolo s_x^2 si es muestral o con el símbolo σ_x^2 si es poblacional. La expresión para la varianza poblacional es:

$$\sigma_x^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n} \tag{1.11}$$

Si desarrollamos la fórmula:

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum(x_i^2 + \overline{x}^2 - 2\ x_i\ \overline{x}) = \sum x_i^2 + n\ \overline{x}^2 - 2\ \overline{x}\ \sum x_i =$$

$$= \sum x_i^2 + n\ \overline{x}^2 - 2\ \overline{x}\ n\ \overline{x} = \sum x_i^2 - n\ \overline{x}^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

Entoces $s_x^2 = \frac{1}{n}\ [\sum x_i^2 - \frac{1}{n}\ (\sum x_i)^2] = \frac{1}{n}\ \sum x_i^2 - (\overline{x})^2$

La **desviación estándar** es simplemente la raíz cuadrada de la varianza:

$$\sigma_x = \sqrt{\frac{\sum(x_i - \overline{x})^2}{n}} = \sqrt{\frac{1}{n}\ \sum\left(x_i^2 - \overline{x}\right)^2} \tag{1.12}$$

Cuando la muestra contiene datos repetidos, las frecuencias absolutas se incluyen en las fórmulas:

$$\sigma_x^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2\ n_i}{n} = \frac{1}{n}\ \left[\sum x_i^2\ n_i - \frac{1}{n}\ \left(\sum x_i\ n_i\right)^2\right] \tag{1.13}$$

Un valor pequeño indica que hay poca dispersión, pero un valor alto indica que hay una gran dispersión.

En todas las definiciones y fórmulas comentadas anteriormente hacen referencia a datos poblacionales. La **corrección de Bessel** implica dividir entre *(n - 1)* en lugar de n en las fórmulas de la varianza muestral y de la desviación estándar muestral, puesto que al tratarse de una muestra al menos faltará un elemento. De esta manera, se corrigen posibles sesgos.

Veamos cuáles son las expresiones correspondientes a la varianza y desviación estándar muestral:

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} \tag{1.14}$$

$$s_x = \sqrt{\frac{\sum(x_i - \overline{x})^2}{n-1}} = \sqrt{\frac{1}{n-1}\sum x_i^2 - (\overline{x})^2} \tag{1.15}$$

1.5.3. Medidas de asimetría

Se utilizan para determinar si los datos están concentrados en una parte de la trayectoria de la variable. Es decir, si son simétricos con respecto a la media aritmética o no.

Asimetría

La **asimetría** es la medida que indica la simetría de la distribución de una variable respecto a la media aritmética, sin necesidad de realizar una representación gráfica.

Realizar la comparación directa de la media y la moda, $\overline{x} - M_o$, no es muy fiable. Por lo tanto, se usa el coeficiente de Pearson donde:

$$A_s = \frac{\overline{x} - M_o}{s_x}$$

Depenediendo el valor de A_s tenemos lo siguiente:

- Cuando $A_s = 0$, la muestra es simétrica.
- Cuando $A_s < 0$, la muestra tiene asimetría por la izquierda o asimetría negativa.
- Cuando $A_s > 0$, la muestra presenta asimetría por la derecha o positiva.

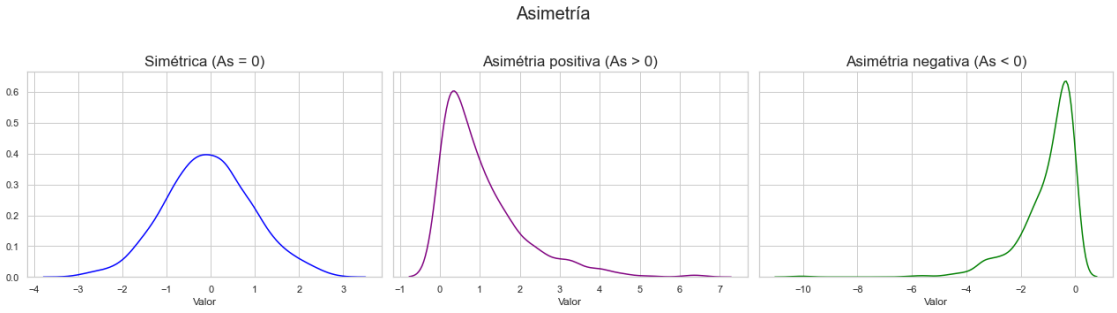


Figura 1.1: Medidas de asimetría. Asimetría.

Curtosis

La **curtosis** es la medida del achatamiento o apuntamiento de una distribución en relación a la distribución normal. Su expresión es la siguiente:

$$K = \frac{\sum \frac{(x_i - \overline{x})^4 \; n_i}{n}}{s_x^4} - 3 \tag{1.16}$$

Según el valor de K tendremos la siguiente clasificación:

- Cuando $K = 0$, tiene curtosis nula y se dice que la muestra es mesocúrtica.
- Cuando $K < 0$, tenemos curtosis negativa y la muestra es platicúrtica.
- Cuando $K > 0$, tenemos cutosis positiva y la muestra es leptocúrtica.

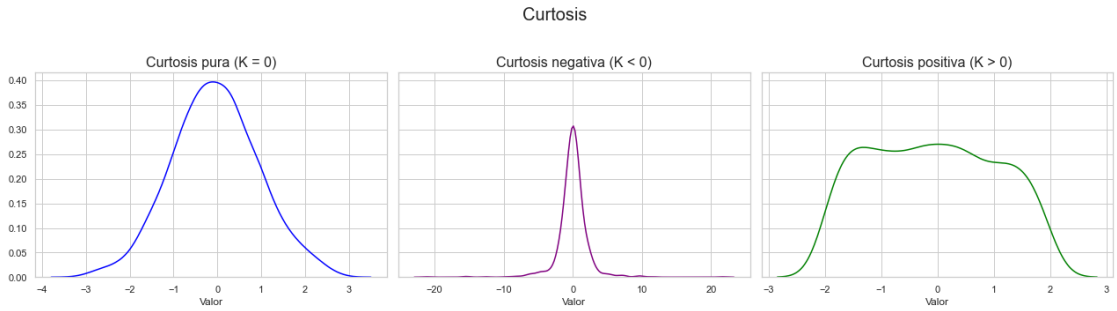


Figura 1.2: Medidas de asimetría. Curtosis.

1.5.4. Diagramas de caja

Los **diagramas de caja** muestran la distribución de datos para una variable numérica. Éstos ayudan a ver el centro y la extensión de los datos. También se pueden utilizar como herramienta visual para comprobrar la normalidad o identificar puntos que podrían ser valores atípicos.

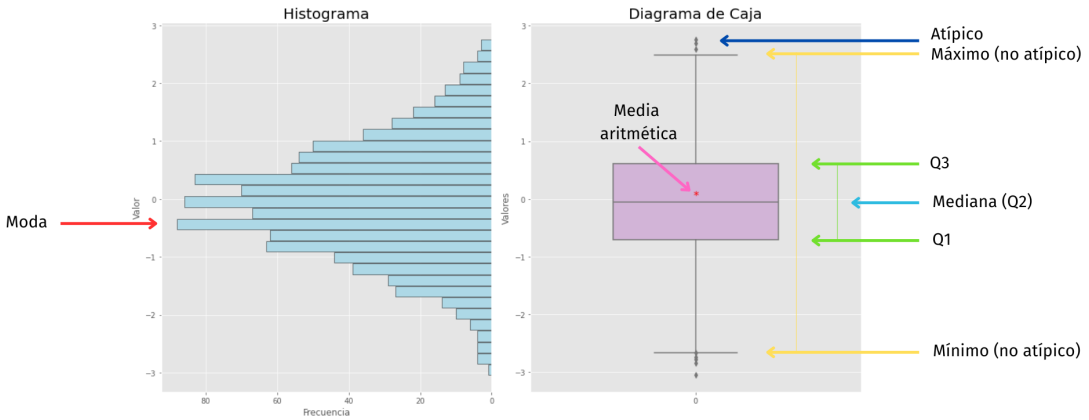


Figura 1.3: Diagrama de caja.

A continuación se van a comentar las partes básicas de un diagrama de caja:

- La línea central de la caja indica la mediana de los datos. Si los datos son simétricos, la mediana estará en el centro de la caja.
- Los extremos de arriba y abajo de la caja indican los cuantiles, o percentiles, 25 y 75. La longitud de la caja es la diferencia entre estos dos percentiles y se conoce como rango intercuartílico (IQR).
- Las líneas que se extienden desde la caja se llaman bigotes. Estos bigotes representan la varianza esperada de los datos.
- Los datos que queden por debajo o por encima de los extremos de los bigotes, se representan con puntos, siendo éstos los valores atípicos.
- El asterisco indica el valor de la media aritmética del conjunto de datos.

2. Regresión y Correlación

La **regresión** y la **correlación** son técnicas utilizadas para relacionar dos variables aleatorias definidas en la misma población. Escribiremos estas variables aleatorias bidimensionales como (X, Y).

Ejemplos:

1. Si tomáramos los datos de la altura y el peso de las personas de Eibar, de alguna manera estarían relacionados.
2. La altura y la edad de los niños también estarían relacionadas.
3. En finanzas, por ejemplo, los precios diarios de Telefónica y Euskaltel también estarán de alguna manera relacionados.

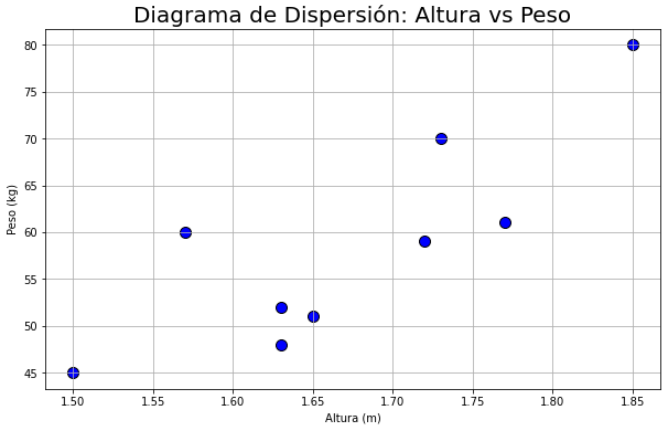
2.1. Diagramas de dispersión y regresión

Para representar gráficamente una variable aleatoria bidimensional (X, Y), se utilizan diagramas de dispersión. En estos, una de las variables se ubicará en el eje x y la otra en el eje y, y cada X se le ajustará una Y.

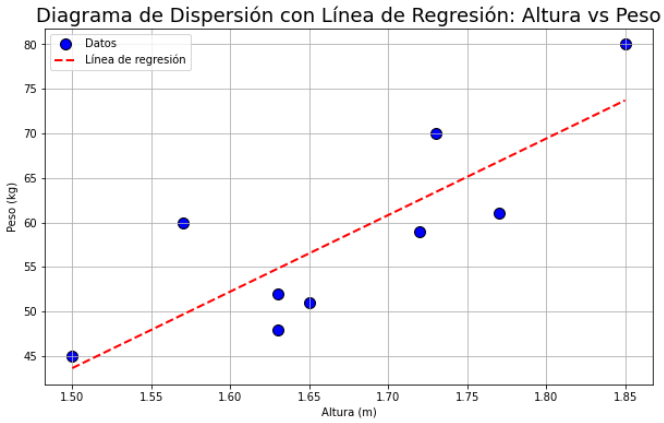
Ejemplo: Supongamos que preguntamos la altura y el peso a 10 mujeres de Eibar y nos dan los siguientes datos:

Persona	1	2	3	4	5	6	7	8	9
X = altura	1,65	1,85	1,72	1,57	1,73	1,77	1,63	1,63	1,50
Y = peso	51	80	59	60	70	61	48	52	45

Entonces, un diagrama de dispersión de estos datos se vería así:



Y la regresión correcta, que mejor resumiría esta serie de puntos, sería:



2.2. Regresión: ¿Qué es?

En este caso, ambas variables se relacionan mediante una ecuación concreta. Así, dando una de las dos, más o menos, podríamos encontrar una aproximación para la otra.

Antes de comenzar con la regresión veremos algunas definiciones:

$$S(x) = \sum x; \quad S(xy) = \sum xy; \quad S(y^2) = \sum y^2; \quad cov(x,y) = \frac{1}{n} \cdot \sum x \cdot y - \overline{x} \cdot \overline{y}$$

2.2.1. Tipos de regresión

La regresión puede adoptar muchas formas:

- $y = a + b \cdot x$
- $y = a + b \cdot x + c \cdot x^2$
- $y = a \cdot e^{b \cdot x}$
- $y = a \cdot x^b$

De esta manera, a partir de las x e y verdaderas, se pueden encontrar los valores estimados de y ($\hat{y}$).

2.2.2. Cálculo de parámetros

Podríamos decir que existe un bien método para calcular todos los parámetros, el llamado método de mínimos cuadrados. Lo que hace este método es encontrar parámetros que hagan que $\sum(y_i - \hat{y}_i)^2$ sea lo más pequeño posible.

2.3. Correlación: ¿Qué es?

De esta forma, se utiliza un único coeficiente para relacionar dos variables, que es el coeficiente de correlación r. El valor de r estará en el rango [-1, 1] y cuanto más cerca esté su valor absoluto del número 1, mayor será la relación entre las dos variables:

$$r = \frac{conv(x,y)}{s_x \cdot s_y} = \frac{\frac{1}{n} \cdot \sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{s_x \cdot s_y} = \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x - \overline{x}) \cdot (y - \overline{y}) \cdot f(x,y) \cdot dx \cdot dy}{s_x \cdot s_y} \tag{2.1}$$

Nota: En este caso, sabemos si existe o no relación entre las dos variables, pero no sabemos encontrar una dada la otra.

La **covarianza** se puede calcular de la siguiente manera:

$$cov(x,y) = \frac{1}{n}\sum_{i=1}^{n} x_i \cdot y_i - \overline{x} \cdot \overline{y} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x \cdot y \cdot f(x,y) \cdot dx \cdot dy - \overline{x} \cdot \overline{y} \tag{2.2}$$

2.4. Encontrar parámetros de regresión

Como se mencionó anteriormente, buscaremos valores que hagan que la aproximación y el valor verdadero sean lo más similares posible.

2.4.1. Regresión lineal

Siendo la expresión $y = a + b \cdot x$ hay que encontrar los parámetros a y b que hacen que $\sum(y - a - b \cdot x)^2$ sean el mínimo.

Elaboración de derivadas:

$$\begin{cases} \sum 2 \cdot (y - a - b \cdot x) \cdot (-1) = 0 \\ \sum 2 \cdot (y - a - b \cdot x) \cdot (-x) = 0 \end{cases} \rightarrow \begin{cases} \sum (y - a - b \cdot x) = 0 \\ \sum (y \cdot x - a \cdot x - b \cdot x^2) = 0 \end{cases}$$

Entonces los coeficientes serían:

$$b = \frac{cov(x,y)}{s_x^2}; \quad a = \overline{y} - b \cdot \overline{x} \tag{2.3}$$

2.4.2. Relación entre estos coeficientes y la correlación

$$r_z = \frac{s_x}{s_y} \cdot b \tag{2.4}$$

Estandarización

¿Qué pasa si tenemos $x' = \dfrac{x + x_o}{u}$ y $y' = \dfrac{y + y_o}{v}$?
Entonces,

- $r(x, y) = r(x', y')$
- $b(x, y) = \dfrac{u}{v} \cdot b(x', y')$

Error de estimación

El error de estimación se calcula como la varianza entre el valor real de y y el error de estimación.

$$s_{ey} = \sqrt{\frac{\sum y_i^2 - a \cdot \sum y_i - b \cdot \sum x_i \cdot y_i}{n - 2}} \tag{2.5}$$

2.4.3. Regresión exponencial

Siendo la expresión $y = a \cdot e^{b \cdot x}$ para realizar el cálculo de los parámetros aplicaremos logaritmos a ambos lados:

$$ln\, y = ln(a \cdot e^{b \cdot x}) = ln\, a + ln(e^{b \cdot x}) = ln\, a + b \cdot x$$

Así, tenemos una ecuación similar a la que teníamos en el caso anterior, y por tanto,

$$b = \frac{cov(x, ln\, y)}{s_x^2}; \;\; a = e^{\overline{s} - b \cdot \overline{x}} \;\text{ donde }\; s = ln\, y$$

El coeficiente de correlación tiene la siguiente expresión:

$$r_e = \frac{s_x}{s_{ln\, y}} \cdot b$$

2.4.4. Regresión potencial

Siendo la expresión $y = a \cdot x^b$ para realizar el cálculo de parámetros se realiza de forma similar a la anterior:

$$ln\, y = ln(a \cdot x^b) = ln\, a + ln(x^b) = ln\, a + b \cdot (ln\, x)$$

Así, tenemos una ecuación similar a la que teníamos en el caso anterior, y por tanto,

$$b = \frac{cov(ln\, x, ln\, y)}{s_{ln\, x}^2}; \;\; a = e^{\overline{s} - b \cdot \overline{t}} \;\text{ donde }\; s = ln\, y \;\text{ y }\; t = ln\, x$$

El coeficiente de correlación tiene la siguiente expresión:

$$r_p = \frac{s_{ln\, x}}{s_{ln\, y}} \cdot b$$

2.4.5. Regresión parabólica

Teniendo en cuenta la expresión $y = a + b \cdot x + c \cdot x^2$, al haber tres parámetros, no se puede dar ningún valor específico. En cada caso se debe resolver el siguiente sistema de ecuaciones:

$$\begin{cases} a \cdot n + b \cdot \sum_i x_i + c \cdot \sum_i x_i^2 = \sum_i y_i \\ a \cdot \sum_i x_i + b \cdot \sum_i x_i^2 + c \cdot \sum_i x_i^3 = \sum_i x_i \cdot y_i \\ a \cdot \sum_i x_i^2 + b \cdot \sum_i x_i^3 + c \cdot \sum_i x_i^4 = \sum_i x_i^2 \cdot y_i \end{cases}$$

En este caso, la estimación del error se realizará de la siguiente manera:

$$s_{ey} = \sqrt{\frac{\sum y^2 - a \cdot \sum y - b \cdot \sum x \cdot y - c \cdot \sum x^2 \cdot y}{n}}$$

3. Combinatoria

La combinatoria estudia los métodos para contar las distintas configuraciones de los elementos de un conjunto que cumplan ciertos criterios especificados.

En todo problema combinatorio hay conceptos claves que se debe distinguir:

- **Población** Conjunto de elementos que se está estudiando. Se denomina con $\boldsymbol{n}$ al número de elementos de este conjunto. De esta forma, diremos que el conjunto A contiene los elementos $\{a_1, ..., a_n\}$ y lo denotaremos de la siguiente forma: $A = \{a_1, a_2, ..., a_n\}$
- **Muestra** Subconjunto de la población. Se denomina con $\boldsymbol{m}$ al número de elementos que componen la muestra. Dos subconjuntos serán distintos si contienen al menos un elemento diferente. Por ejemplo $\{a_1, a_2\} \neq \{a_1, a_3\}$. Además, la muestra viene caracterizada por:
  - **Orden** Si importa o no que los elementos de la muestra aparezcan ordenados. Entonces:
    - Si el orden es importante entonces $\{a_1, a_2\} \neq \{a_2, a_1\}$ y por tanto se consideran distintos.
    - Si el orden NO es importante entonces $\{a_1, a_2\} = \{a_2, a_1\}$ y por tanto se consideran iguales.
  - **Repetición** La posibilidad de repetición o no de los elementos.

En definitiva, dos subconjuntos son distintos si contienen al menos un elemento diferente o, si el orden de los elementos es diferente cuando el orden es importante.

3.1. Variaciones

Las variaciones son aquellas formas de agrupar los elementos de un conjunto teniendo en cuenta el orden.

3.1.1. Sin repetición

Las **variaciones sin repetición** de n elementos tomados de m en m se definen como las distintas agrupaciones formadas con m elementos diferentes, eligiéndolos de entre los n elementos que disponemos. Se considera una variación distinta a otra si:

- $m \leq n$, aunque relegamos el caso $m = n$ para las permutaciones.
- dos subconjuntos son distintos si contienen al menos un elemento diferente o, si el orden de los elementos es diferente.

El número de variaciones que se pueden construir se puede calcular mediante la fórmula:

$$\mathrm{V}_{\mathrm{n}}^{\mathrm{m}} = \frac{n!}{(n-m)!} \tag{3.1}$$

Ejemplo: De un conjunto $A = \{\mathrm{a}_1, \mathrm{a}_2, \mathrm{a}_3\}$ los elementos se eligen de 2 en 2. ¿Cuantas variaciones se pueden realizar?

Tenemos un total de 3 elementos en el conjunto, por lo que el valor de n es 3. Y se escogen los elementos en grupos de 2, por lo tanto, el valor de m es de 2. Teniendo en cuenta los valores de estos dos parámetros las combinaciones que se pueden realizar son:

$$\mathrm{V}_3^2 = \frac{3!}{(3-2)!} = \frac{3!}{1!} = 6$$

Si realizamos estas combinaciones a mano, podemos comprobar que el cálculo está bien:

$$\begin{array}{ll} \{\mathrm{a}_1, \mathrm{a}_2\} & \{\mathrm{a}_2, \mathrm{a}_1\} \\ \{\mathrm{a}_1, \mathrm{a}_3\} & \{\mathrm{a}_3, \mathrm{a}_1\} \\ \{\mathrm{a}_2, \mathrm{a}_3\} & \{\mathrm{a}_3, \mathrm{a}_2\} \end{array}$$

3.1.2. Con repetición

Las **variaciones con repetición** de n elementos tomados de m en m se definen como las distintas agrupaciones formadas con m elementos que pueden repetirse, eligiéndolos de entre los n elementos que disponemos, considerando una variación distinta a otra tanto si difieren en algún elemento como si están situados en distinto orden. El número de variaciones que se pueden construir se pueden calcular mediante la fórmula:

$$\mathrm{VR}_{\mathrm{n}}^{\mathrm{m}} = n^m \tag{3.2}$$

Ejemplo: De un conjunto $A = \{1,\ 2,\ 3,\ 4,\ 5\}$ se quieren crear números de 5 cifras pudiéndose repetir.

Tenemos un total de 5 elementos en el conjunto, por lo que el valor de n es 5. Y se escogen los elementos en grupos de 5, por lo tanto, el valor de m es de 5. Teniendo en cuenta los valores de estos dos parámetros y que las cifras de los números buscados se pueden repetir las combinaciones que se pueden realizar son:

$$\mathrm{VR}_5^5 = 5^5 = 3125$$

3.2. Combinaciones

Las combinaciones son aquellas formas de agrupar los elementos de un conjunto teniendo en cuenta que:

- NO influye el orden en que se colocan.

3.2.1. Sin repetición

Las **combinaciones sin repetición** de n elementos tomados de m en m se definen como las agrupaciones formadas con m elementos distintos, eligiéndolos de entre los n elementos que disponemos, considerando una variación distinta a otra sólo si difieren en algún elemento. (No influye el orden de colocación de sus elementos) El número de combinaciones que se pueden construir se puede calcular mediante la fórmula:

$$\mathrm{C}_{\mathrm{n}}^{\mathrm{m}} = \binom{n}{m} = \frac{n!}{(n-m)!\, m!} \tag{3.3}$$

Ejemplo: De un conjunto $A = \{\mathrm{a}_1, \mathrm{a}_2, \mathrm{a}_3\}$ los elementos se eligen de 2 en 2. ¿Cuantas combinaciones se pueden realizar?

Tenemos un total de 3 elementos en el conjunto, por lo que el valor de n es 3. Se escogen los elementos en grupos de 2, por lo tanto, el valor de m es de 2. Teniendo en cuenta los valores de estos dos parámetros las combinaciones que se pueden realizar son:

$$\mathrm{C}_3^2 = \frac{3!}{(3-2)!\, 2!} = \frac{3!}{2!} = 3$$

Si realizamos estas combinaciones a mano, podemos comprobar que el cálculo está bien:

$$\{\mathrm{a}_1, \mathrm{a}_2\}$$
$$\{\mathrm{a}_1, \mathrm{a}_3\}$$
$$\{\mathrm{a}_2, \mathrm{a}_3\}$$

3.2.2. Con repetición

Las **combinaciones con repetición** de n elementos tomados de m en m se definen como las distintas agrupaciones formadas con m elementos que pueden repetirse, eligiéndolos de entre los n elementos que disponemos, considerando una variación distinta a otra sólo si difieren en algún elemento. (No influye el orden de colocación de sus elementos). El número de combinaciones que se pueden construir

se puede calcular mediante la fórmula:

$$\mathrm{CR}_{\mathrm{n}}^{\mathrm{m}} = \binom{n+m-1}{m} = \frac{(n+m-1)!}{(n-1)!\,m!} \tag{3.4}$$

Ejemplo: En una pastelería disponen 5 tipos de pasteles, y nos dejan elegir 4 de los cuales se pueden repetir, ¿cuántas combinaciones de pasteles se podrían escoger?

Tenemos un total de 5 elementos en el conjunto, por lo que el valor de n es 5. Y se escogen los elementos en grupos de 4, por lo tanto, el valor de m es de 4. Teniendo en cuenta los valores de estos dos parámetros las combinaciones que se pueden realizar son:

$$\mathrm{CR}_5^4 = \binom{5+4-1}{4} = \frac{(5+4-1)!}{(5-1)!\,4!} = 70$$

3.3. Permutaciones

Las permutaciones, o también llamadas ordenaciones, son aquellas formas de agrupar los elementos de un conjunto teniendo en cuenta que:

- Influye el orden en que se colocan.
- Tomamos todos los elementos de que se disponen.

3.3.1. Sin repetición

Las **permutaciones sin repetición** de n elementos se definen como las distintas formas de ordenar todos esos elementos distintos, por lo que la única diferencia entre ellas es el orden de colocación de sus elementos. El número de estas permutaciones será:

$$\mathrm{P}_{\mathrm{n}} = n! \tag{3.5}$$

Ejemplo: ¿De cuántas formas se pueden sentar 3 personas en 3 butacas?

$$\mathrm{P}_3 = 3! = 6 \tag{3.6}$$

3.3.2. Con repetición

Llamamos a las **permutaciones con repetición** de n elementos tomados de a en a, de b en b, de c en c, etc, cuando en los n elementos exixten elementos repetidos verificándose que $a + b + c + \ldots = n$. Si tomamos como ejemplo el conjunto

$A = \{a, a, a, ..., a, b, b, b, \ ..., b, c, c, c, ..., c\}$ el número de estas permutaciones será:

$$\mathrm{PR}_{\mathrm{n}}^{\alpha,\beta,\gamma} = \frac{n!}{\alpha!\ \beta!\ \gamma!} \tag{3.7}$$

Siendo:

- α: las veces que aparece el elemento a.
- β: las veces que aparece el elemento b.
- γ: las veces que aparece el elemento c.

Ejemplo: ¿Cuántos números se pueden construir utilizando todas las cifras del siguiente conjunto $A = \{4, 4, 4, 3, 3, 5\}$?,

Tenemos un total de 6 elementos en el conjunto, por lo que el valor de n es 6. Y para los valores de los parámetros α, β y γ hay que ver cuantas veces se repiten los elementos:

- $\alpha = 3$, se repite 3 veces el número 4.
- $\beta = 2$, se repite 2 veces el número 3.
- $\gamma = 1$, se repite 1 vez el número 5.

$$\mathrm{PR}_{6}^{3,2,1} = \frac{6!}{3!\ 2!\ 1!} = 60$$

3.4. Esquema

El siguiente esquema representa los pasos que hay que realizar para poder visualizar de que tipo de combinatoria se dispone.

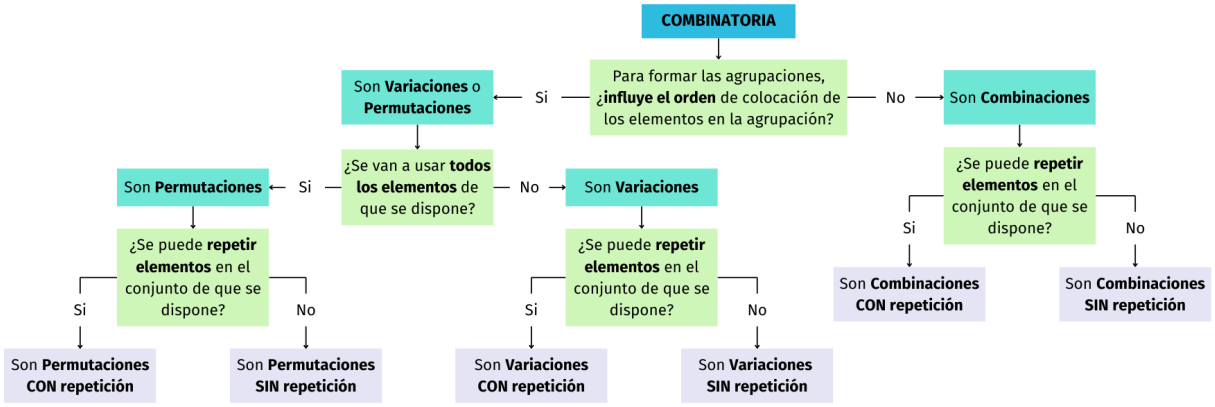


Figura 3.1: Esquema combinatoria.

3.5. Los números combinatorios

Los números combinatorios o coeficientes binomiales son los números de la forma $\binom{n}{m}$ siendo m y n enteros no negativos con $m \leq n$. Se calculan, como ya se ha dicho mediante la fórmula:

$$\binom{n}{m} = \frac{n\ (n-1)\ (n-2)...(n-m+1)}{m!} = \frac{n!}{m!\ (n-m)!} \tag{3.8}$$

Indican:

- El número de combinaciones sin repetición de n elementos tomados de m en m.
- El número de subconjuntos de m elementos que tiene un conjunto de n elementos.

Propiedades básicas:

1. $\binom{n}{m} = \binom{n}{n-m}$
2. $\binom{n}{m} = \binom{n-1}{m-1} + \binom{n-1}{m}$
3. $\binom{n}{0} = \binom{n}{n} = 1$
4. $\binom{n}{1} = n$

4. Probabilidad

El término **probabilidad** se refiere al estudio de azar y la incertidumbre en cualquier situación en la cual varios posibles sucesos pueden ocurrir; la disciplina de la probabilidad proporciona métodos de cuantificar las oportunidades y probabilidades asociadas con varios sucesos.

4.1. Definiciones

- **Experimento aleatorio:**
  - Se pueden repetir experimentos análogos de manera infinita.
  - Los resultados que se pueden obtener en cada experimento conforman el universo o espacio muestral.
  - Antes de realizar un experimento no se puede saber el resultado que se va a obtener (condición de aleatoriedad).
  - Cuando el número de experimentos crece, la frecuencia relativa de cada resultado (o cada punto del espacio muestral) tiende a estabilizarse.
- **Suceso aleatorio:** subconjunto del espacio muestral.

Ejemplos:

1. Probabilidad de que salgan números pares en un dado.

$$\text{Experimento aleatorio} \rightarrow E = \{1, 2, 3, 4, 5, 6\}$$
$$\text{Subconjunto aleatorio} \rightarrow S = \{2, 4, 6\}$$

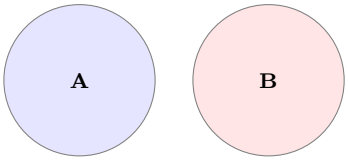
2. Experimento aleatorio: lanzamiento de dos dados. Suceso aleatorio: "la suma de los dados sea 7".

1, 1	1, 2	1, 3	1, 4	1, 5	1, 6
2, 1	2, 2	2, 3	2, 4	2, 5	2, 6
3, 1	3, 2	3, 3	3, 4	3, 5	3, 6
4, 1	4, 2	4, 3	4, 4	4, 5	4, 6
5, 1	5, 2	5, 3	5, 4	4, 5	5, 6
6, 1	6, 2	6, 3	6, 4	6, 5	6, 6

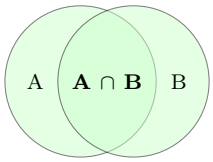
$$\text{N} = 36 = 6^2$$

$$\{(1, 6), (2, 5), (3, 4), (5, 2), (6, 1)\}$$

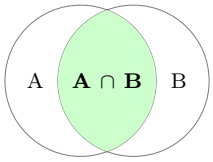
- **Operaciones:**
  - Los sucesos se representan con letras mayúsculas (A, B, C,...)



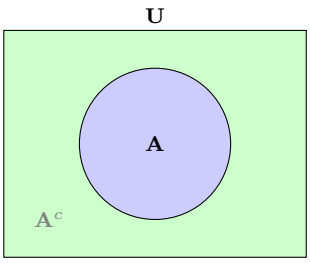
  - Los elementos pertenecen a los sucesos (a $\in$ A, b $\in$ B, ...)
  - Unión de sucesos: $A \cup B = \{x \in S \mid x \in A \vee x \in B\}$



  - Intersección de sucesos: $A \cap B = \{x \in S \mid x \in A \wedge x \in B\}$



  - Complemento: $\overline{A} = A^c = \{x \in S \mid x \notin A\}$



- **Observaciones:**
  - Cuando un suceso es imposible el conjunto que representa es un conjunto "vacío". $P(\emptyset) = 0$
  - Cuando un suceso ocurre siempre se denomina "suceso seguro". P = 100 %
  - Cuando un suceso no es ni imposible ni seguro se denomina "suceso probable".
- **Nota:** La unión y la intersección son conmutativas. Es decir, $A \cap B = B \cap A$, y análogamente, $A \cup B = B \cup A$

Ejemplos:

1. Lanzamiento de dos dados y el producto de ambos tiene que ser ≥ 40.
   Suceso imposible. Es imposible que salga ≥ 40, ya que el valor máximo del producto entre ambos dados es de 36.

2. Lanzamiento de dos dados y que la suma de ambos sea ≤ 12.
   Suceso seguro. Ya que la suma entre los valores de los dados estará comprendido entre: $1 + 1 \leq x \leq 6 + 6 = 2 \leq x \leq 12$.

3. Lanzamiento de dados y que la suma sea 8:

$$\mathrm{A} = \{(2,6),(4,4),(3,5),(5,3),(6,2)\}$$

   Lanzamiento de dados y que el producto sea 12:

$$\mathrm{B} = \{(2,6),(3,4),(4,3),(6,2)\}$$

   - $A \cup B$: $\{(2,6),(4,4),(3,5),(5,3),(6,2),(3,4),(4,3)\}$
   - $A \cap B$: $\{(2,6),(6,2)\}$
   - En el suceso A tenemos 5 posibilidades de las 36 que hay en total. Por lo tanto, en $\overline{A}$ tendremos 36 - 5 = 31 posibilidades.

4.2. Probabilidad

Regla de Laplace: en el caso de que todos los resultados de un experimento aleatorio sean equiprobables, Laplace define la probabilidad de un suceso A como el cociente entre el número de resultados favorables a que ocurra el suceso A en el experimento y el número de resultados posibles del experimento. Así, se puede expresar con la siguiente fórmula:

$$P(A) = \frac{\text{Casos favorables a A}}{\text{Total casos posibles}} \tag{4.1}$$

Propiedades:

- $0 \leq P(A) \leq 1$
- $\mathrm{P(E)} = 1$
- $\mathrm{P}(\emptyset) = 0$
- Sucesos compatibles: Cuando pueden ocurrir los dos sucesos, es decir, que tienen algún suceso elemental común.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

- Sucesos incompatibles: cuando no pueden ocurrir dos o más sucesos a la vez, es decir, no tienen ningún suceso elemental en común.

$$P(A \cup B) = P(A) + P(B)$$

- $\bigcup_{i=1}^{\infty} A_i = \sum_{i=1}^{\infty} P(A)$
- $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) + P(A \cap C) - P(A \cap B \cap C)$

Ejemplos:

1. Probabilidad de que un alumno apruebe matemáticas es $2/3$, la probabilidad de que apruebe inglés es $4/9$ y la probabilidad de que apruebe las dos es $1/2$. ¿Cuál es la probabilidad de aprobar al menos una?

$$P(M \cup I) = P(M) + P(I) - P(M \cap I) = 2/3 + 4/9 - 1/2 = 31/36$$

2. Al lanzar 2 dados la probabilidad de obtener 7 u 11.

   - "*obtener* 7" $\rightarrow A = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}$
   - "*obtener* 11" $\rightarrow B = \{(5,6),(6,5)\}$

$$P(A) = \frac{6}{36} = \frac{1}{6}\,; \quad P(B) = \frac{2}{36} = \frac{1}{18}$$

$A \cup B = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1),(5,6),(6,5)\}$
$A \cap B = \emptyset$
$P(A \cup B) = P(A) + P(B) = 1/6 + 1/18 = 2/9$
$P(A \cap B) = 0$

4.3. Probabilidad condicionada

La **probabilidad condicionada** es la posibilidad de que ocurra un evento, al que denominamos A, como consecuencia de que haya tenido lugar otro evento, al que denominamos B. Se puede expresar mediante la siguiente fórmula:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \tag{4.2}$$

La fórmula se lee que la probabilidad de que suceda A, dado que ha acontecido B, es igual a la probabilidad que ocurra A y B, al mismo tiempo, entre la probabilidad de B.

Ejemplos:

1. En un dado se tienen los siguientes sucesos:

$$\text{“obtener un número cuadrado”} \rightarrow A = \{1, 4\}$$
$$\text{“obtener un número mayor que 3”} \rightarrow B = \{4, 5, 6\}$$
$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}; \quad P(B \mid A) = \frac{P(B \cap A)}{P(A)}$$

   - $A \cap B : \{4\}$
   - $P(A) = \frac{2}{6} = \frac{1}{3}; \quad P(B) = \frac{3}{6} = \frac{1}{2}; \quad P(A \cap B) = \frac{1}{6}$

$$P(A \mid B) = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{2}{6} = \frac{1}{3}; \quad P(B \mid A) = \frac{\frac{1}{6}}{\frac{1}{3}} = \frac{3}{6} = \frac{1}{2}$$

2. Probabilidad de que un avión salga en hora es 0,83, probabilidad de que llegue en hora es 0,82 y probabilidad de que un avión salga y llegue en hora es 0,78.

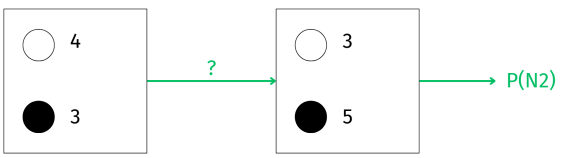
$$P(S) = 0{,}83; \quad P(Ll) = 0{,}82; \quad P(S \cap Ll) = 0{,}78$$
$$P(Ll \mid S) = \frac{P(S \cap Ll)}{P(S)}; \quad P(S \mid Ll) = \frac{P(S \cap Ll)}{P(Ll)}$$
$$P(Ll \mid S) = \frac{0{,}78)}{0{,}83} = 0{,}94; \quad P(S \mid Ll) = \frac{0{,}78}{0{,}82} = 0{,}95$$

4.4. Reglas de multiplicación

Sean dos sucesos A y B:

$$\begin{cases} P(A \cap B) = P(A) \cdot P(B \mid A) \\ P(B \cap A) = P(B) \cdot P(A \mid B) \end{cases} \rightarrow P(A) \cdot P(B \mid A) = P(B) \cdot P(A \mid B)$$

Ejemplo: Tenemos dos cajas, en la primera tenemos 4 bolas blancas y 3 bolas negras. Extraemos una bola de la primera caja y la introducimos en la segunda caja, donde inicialmente había 3 bolas blancas y 5 bolas negras. ¿Cuál es la probabilidad de que al extraer una bola de la segunda caja sea negra?



$$P(N_1 \cap N_2) = P(N_1) \cdot P(N_2 \mid N_1) = \frac{3}{7} \cdot \frac{6}{9} = \frac{18}{63}$$
$$P(B_1 \cap N_2) = P(B_1) \cdot P(N_2 \mid B_1) = \frac{4}{7} \cdot \frac{5}{9} = \frac{20}{63}$$
$$P(N_2) = P(N_1 \cap N_2) + P(B_1 \cap N_2) = \frac{18}{63} + \frac{20}{63} = \frac{38}{63}$$

Dos sucesos libres A y B $\iff$ $P(A \cap B) = P(A) \cdot P(B)$ *o* $P(A \mid B) = P(A)$ *y* $P(B \mid A) = P(B)$

Ejemplo: Lanzar una moneda y que salga cara y lanzar un dado y que salga número par.

No tienen ninguna relación estos sucesos, por lo que son sucesos libres.

4.5. Probabilidad total

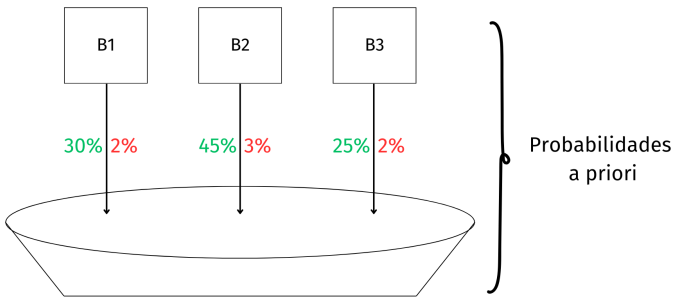
Existen situaciones en las cuales varios eventos intervienen en la realización de algún otro suceso del mismo espacio muestral. Sean B_1, B_2,..., B_n sucesos:

Suponemos que S es una partición del espacio, es decir, los sucesos son libres y $S = \bigcup_{i=1}^{n} B_i$ *o* $S = B_1 \cup ... \cup B_i \cup ... \cup B_n$

La siguiente fórmula permite calcular la probabilidad del suceso A perteneciente a S (A $\in$ S), conocidos los valores de probabilidad de los suceso B_1, B_2, ..., B_n.

$$P(A) = \sum_{i=1}^{n} P(B_i) \cdot P(A \mid B_i) \tag{4.3}$$

Ejemplo: En un taller tenemos 3 máquinas B_1, B_2 y B_3 que producen respectivamente el 30 %, 45 % y 25 % de la piezas. Dichas máquinas producen respectivamente 2 %, 3 % y 2 % de piezas defectuosas. Si tomamos una pieza al azar, ¿cuál es la probabilidad de que sea defectuosa?



Se van a identificar los siguientes sucesos:

- A: "la pieza tiene fallo"

- B_1: "pieza de la 1ª máquina"
- B_2: "pieza de la 2ª máquina"
- B_3: "pieza de la 3ª máquina"

$$P(A) = P(B_1) \cdot P(A \mid B_1) + P(B_2) \cdot P(A \mid B_2) + P(B_3) \cdot P(A \mid B_3) = \\ 0{,}3 \cdot 0{,}02 + 0{,}45 \cdot 0{,}03 + 0{,}25 \cdot 0{,}02 = 0{,}0245 \rightarrow 2{,}45\,\%$$

4.6. Teorema de Bayes

Sean $B_1, B_2, ..., B_i$ suceso mutuamente excluyentes y cuya unión es el espacio muestral E, esto es, $B_1 \cup B_2 \cup ... \cup B_n = E$. Si A es otro suceso, entonces:

$$P(B_i \mid A) = \frac{P(B_i) \cdot (P(A \mid B_i))}{\sum_{k=1}^{n} P(B_k) \cdot P(A \mid B_k)} = \frac{P(B_i \cap A)}{P(A)} \tag{4.4}$$

Ejemplo: En el ejecicio anterior (el de las 3 máquinas). ¿Cuál es la probabilidad de que la pieza defectuosa salga de la máquina B_2?

$P(A) = 0{,}0245$

$P(B_2 \mid A) = \frac{P(B_2) \cdot P(A|B_i)}{\sum_{k=1}^{n} P(B_k) \cdot P(A|B_k)} = \frac{0{,}45 \cdot 0{,}03}{0{,}0245} = 0{,}538 \rightarrow 53{,}8\,\%$

5. Distribución de Variables Aleatorias Discretas

En el material estudiado anteriormente aprendimos a calcular la probabilidad de sucesos de un espacio muestral S. En esta unidad estudiaremos las reglas para establecer correspondencias de elementos de S con los números reales, para luego asignarles un valor de probabilidad.

5.1. Variable aleatoria

La función que asigna un número real a cada punto del espacio muestral se denomina variable aleatoria.

La variable aleatoria se escribe en mayúsculas y sus valores en minúsculas. Entonces X es una variable aleatoria y x es un valor de la variable.

$$\text{X} = \text{x; siendo x el número real}$$

5.2. Variable aleatoria discreta

- Una variable aleatoria es discreta si sus posibles valores son finitos o infinitos numerables.
- ¿Qué NO sería una variable discreta? Calcular la probabilidad de tiempo de llegada a Donostia, puesto que el tiempo al ser un valor continuo no sería un valor finito o infinito numerable.

Función de Probabilidad

Cuando las probabilidades de los puntos de la muestra son conocidas, la función que asigna su probabilidad a cada valor de la variable aleatoria discreta se denomina **función de probabilidad**.

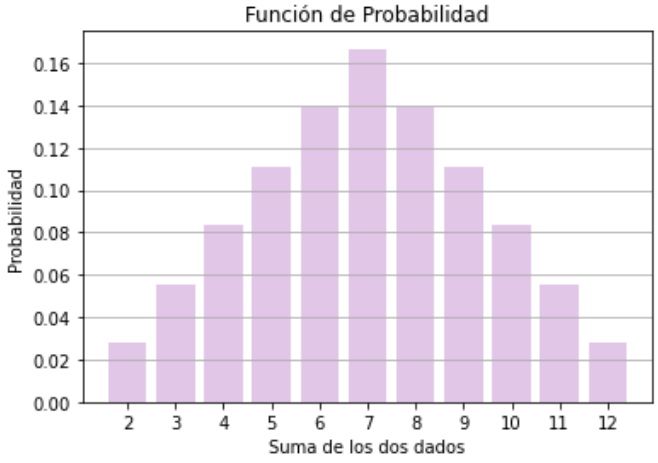
$$P(X = x) = P_i \tag{5.1}$$

Ejemplo: Función de probabilidad del lanzamiento de 2 dados:

- $x = 2 \rightarrow A_2 = \{(1,1)\} \rightarrow P(A_2) = \frac{1}{36}$
- $x = 3 \rightarrow A_3 = \{(1,2),(2,1)\} \rightarrow P(A_3) = \frac{2}{36} = \frac{1}{18}$
- $x = 4 \rightarrow A_4 = \{(1,3),(2,2),(3,1)\} \rightarrow P(A_4) = \frac{3}{36} = \frac{1}{12}$

- $x = 5 \rightarrow A_5 = \{(1,4),(2,3),(3,2),(4,1)\} \rightarrow P(A_5) = \frac{4}{36} = \frac{1}{9}$
- $x = 6 \rightarrow A_6 = \{(1,5),(2,4),(3,3),(4,2),(5,1)\} \rightarrow P(A_6) = \frac{5}{36}$
- $x = 7 \rightarrow A_7 = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\} \rightarrow P(A_7) = \frac{6}{36} = \frac{1}{6}$
- $x = 8 \rightarrow A_8 = \{(2,6),(3,5),(4,4),(5,3),(6,2)\} \rightarrow P(A_8) = \frac{5}{36}$
- $x = 9 \rightarrow A_9 = \{(3,6),(4,5),(5,4),(6,3)\} \rightarrow P(A_9) = \frac{4}{36} = \frac{1}{9}$
- $x = 10 \rightarrow A_{10} = \{(4,6),(5,5),(6,4)\} \rightarrow P(A_10) = \frac{3}{36} = \frac{1}{12}$
- $x = 11 \rightarrow A_{11} = \{(5,6),(6,5)\} \rightarrow P(A_{11}) = \frac{2}{36} = \frac{1}{18}$
- $x = 12 \rightarrow A_{12} = \{(6,6)\} \rightarrow P(A_{12}) = \frac{1}{36}$

x_i	2	3	4	5	6	7	8	9	10	11	12
$P(X = x_i)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$



Función de Distribución

La función de distribución de una variable aleatoria es la función de probabilidad acumulada.

$$F(x) = P(X \leq x) = \sum_{i=1}^{j} P(x = i) \tag{5.2}$$

Ejemplo: Calcular F(4) del ejemplo anterior.

$$F(4) = P(x \leq 4) = P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3) + P(x = 4) =$$

$$= 0 + 0 + \frac{1}{36} + \frac{2}{36} + \frac{3}{36} = \frac{6}{36} = \frac{1}{6}$$

Propiedades:

- $F(x_i) = P(x_1) + P(x_2) + P(x_3) + ... + P(x_i) = F(X_{i-1} + P(x_i)$
- $P(x_i \leq x \leq x_j) = P(x \leq x_j) - P(x < x_i) = F(x_j) - F(x_{i-1})$
- $P(x_i < x < x_j) = P(x < x_j) - P(x \leq x_i) = F(x_{j-1}) - F(x_i)$
- $P(x_i < x \leq x_j) = P(x \leq x_j) - P(x \leq x_i) = F(x_j) - F(x_i)$
- $P(x_i \leq x < x_j) = P(x < x_j) - P(x < x_i) = F(x_{j-1}) - F(x_{i-1})$

5.3. Media y varianza de distribuciones variables discretas

Media

Como sabemos, la media aritmética de una distribución discreta es $\overline{X} = \sum_{i=1}^{k} x_i \cdot \frac{n_i}{n}$, donde los valores de la variable x_i y n_i son su frecuencia absoluta.

Pero cuando la muestra es muy grande, por ejemplo se cumple $n \to \infty$, $\frac{n_i}{n} \to P(x_i)$ entonces $\lim_{x\to\infty} \overline{x} = \lim_{n\to\infty} \sum x_i \cdot \frac{n_i}{n} = \sum x \cdot P(x_i)$, y esta expresión se denomina media o esperanza matemática de la variable aleatoria y se expresa con el signo $E(X)$ o μ_x. En resumen es:

$$E(x) = \sum_{i=1}^{k} x_i \cdot P(x_i) \tag{5.3}$$

Ejemplo: En el ejemplo de los dados anterior:

$E(x) = 2 \cdot P(2) + 3 \cdot P(3) + 4 \cdot P(4) + 5 \cdot P(5) + 6 \cdot P(6) + 7 \cdot P(7) + 8 \cdot P(8) + 9 \cdot P(9) + 10 \cdot P(10) + 11 \cdot P(11) + 12 \cdot P(12) = 2 \cdot \frac{1}{36} + 3 \cdot \frac{2}{36} + 4 \cdot \frac{3}{36} + 5 \cdot \frac{4}{36} + 6 \cdot \frac{5}{36} + 7 \cdot \frac{6}{36} + 8 \cdot \frac{5}{36} + 9 \cdot \frac{4}{36} + 10 \cdot \frac{3}{36} + 11 \cdot \frac{2}{36} + 12 \cdot \frac{1}{36} = \frac{6}{36} = \frac{1}{6}$

Varianza y Desviación típica

Como hemos hecho anteriormente, conocemos que la varianza de la distribución discreta es $s_x^2 = \frac{\sum (x_i - \overline{x})^2 \cdot n_i}{n}$, entonces cuando $n \to \infty$ se cumplirá $\overline{X} \to \mu_x$ y $\frac{n_i}{n} \to P(x_1)$, por lo que el límite de la expresión anterior se llama varianza de la

variable azar y se escribe σ_x^2 o $Var(x)$.

$Var(x) = \lim_{x\to\infty} s_x^2 = \sum_{i=1}^{k}(x_i - \overline{x})^2 \cdot P(x_i)$ y el cuadrado de este binomio desarrollando $Var(x) = \sum_{i=1}^{k} x_i^2 \cdot P(x_i) - \overline{x}^2$.

5.4. Distribución de variables aleatorias discretas

5.4.1. Distribución hipergeométrica

Características de la distribución hipergeométrica:

- Tenemos una población finita de N elementos.
- Todos los elementos están clasificados en dos clases excluyentes: A y $\overline{A}$.
- $P(A) = p; \quad P(\overline{A}) = 1 - p$
- Configuramos muestras de tamaño n a partir de la población finita. **SIN REEMPLAZO**

Nota: Tomar una muestra sin reemplazo significa que los elementos son tomados uno a uno, sin devolución. Podemos concluir entonces que los ensayos ya no pueden ser considerados independientes porque la probabilidad de "éxito" al tomar cada nuevo elemento es afectada por el resultado de los ensayos anteriores debido a que la cantidad de elementos de la población cambia en cada ensayo.

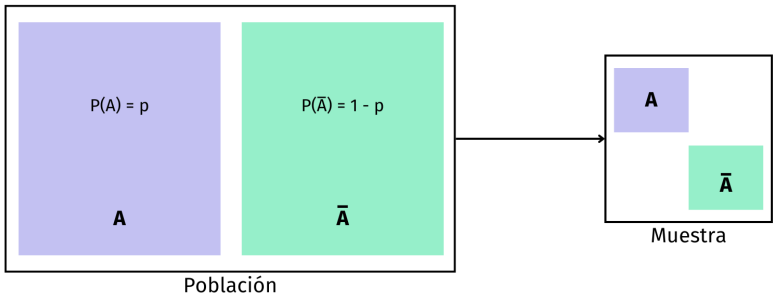


Figura 5.1: Muestra sacada de una población.

X (variable aleatoria): número de elementos de la clase A que hay en la muestra.

$$X = \{x \mid max[0, n - N(1-p)] \leq x \leq min[n, Np]\} \tag{5.4}$$

$$P(X=x)=\frac{\binom{Np}{x}\binom{N(1-p)}{n-x}}{\binom{N}{n}};\quad E(x)=np;\quad Var(x)=\frac{N-n}{N-1}np(1-p)\quad (5.5)$$

Ejemplo: Una caja contiene 100 camisas, donde hay 60 buenas y 40 malas. Configuro una muestra de 70 camisas.

$$A:\text{"}camisa\ buena\text{"};\quad \overline{A}:\text{"}camisa\ mala\text{"};\quad N=100;\quad A=60;\quad \overline{A}=40;\quad n=70$$

a. ¿Cuántas camisas buenas puede contener la muestra?

$$X=\{x \mid max[0,70-100\cdot(1-0{,}6)]\leq x\leq min[70,100\cdot 0{,}6]\}\rightarrow 30\leq x\leq 60$$

b. Probabilidad de que en la muestra haya exactamente 40 camisas buenas.

$$P(x=40)=\frac{\binom{100\cdot 0{,}6}{40}\binom{100(1-0{,}6)}{70-40}}{\binom{100}{70}};$$

$$P(x=40\ buenas)=P(x=30\ malas)$$

5.4.2. Distribución binomial

Esta distribución es muy importante y de uso frecuente. El interés de la variable aleatoria está relacionada con la cantidad de "éxitos" que se obtienen en los diferentes ensayos.

Características:

- Se hacen n ensayos exactamente iguales.
- En cada ensayo hay solo dos posibles soluciones: éxito y fracaso.
- Si la probabilidad del éxito es *p*, la probabilidad de fracaso es *1 - p*. Esto es, $P(exito)=p$ y $P(fracaso)=1-p$.
- Las probabilidades son constantes siempre, no cambian de un ensayo a otro.
- Todos los ensayos son libres. (si en el primero ha salido fracaso, no condiciona a que en el segundo salga éxito o fracaso)

Se define el suceso aleatorio X: "número de éxitos en n ensayos";
Dominio: $\{0, ..., n\}$

$$X = B(n, p) \tag{5.6}$$

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}; \quad \sum P(X = x) = 1 \tag{5.7}$$

$$E(x) = np; \quad Var(x) = np(1-p) \tag{5.8}$$

Tipos de distribución binomial:

- P(X = x_i) → Binomial puntual
- P(X ≤ x_i) → Binomial acumulada → $P(X \leq x_i) = 1 - P(X \geq x_i)$, en este caso se define como $X = B(n, p, x)$

Ejemplo: "X: Obtener tres caras en cuatro lanzamientos"

$$P(X = 3) = \binom{4}{3} 0,5^3 (1 - 0,5)^{4-3}; \tag{5.9}$$

5.4.3. Distribución de Poisson

La distribución de Poisson es un modelo que puede usarse para calcular la probabilidad correspondiente al número de "éxitos" que se obtendrían en una región o en intervalo de tiempo especificados, si se conoce el número promedio de "éxitos" que ocurren.

Probabilidad de x sucesos en un tiempo determinado, cuando la probabilidad del suceso es m.

$$\text{m} \to \infty \qquad \text{p} \to 0$$

$$P(X = x) = \frac{m^x e^{-m}}{x!}; \quad m = np \tag{5.10}$$

$$E(x) = Var(x) = m = np \tag{5.11}$$

5.4.4. Distribución multinomial / polinomial

Es una generalización de la distribución binomial para el caso de que en cada pruba se consideren n sucesos excluyentes $A_1, A_2, ..., A_n$ con probabilidades respectivas $p_1, p_2, ..., p_n$ siendo la suma de todas igual a la unidad.

$$P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n) = \frac{n!}{x_1!\ x_2!\ ...\ x_n!}\ p_1^{x_1} p_2^{x_2} ... p_n^{x_n} = \frac{n!}{\prod_{i=1}^{n} x_i!} \prod_{i=1}^{n} p_i^{x_i} \tag{5.12}$$

5.4.5. Distribución binomial negativa

Este modelo de probabilidad tienen características similares al modelo binomial:

- Ensayos independientes, cada ensayo tiene únicamente dos resultados posibles.
- P(éxito) = p; P(fracaso) = 1 - p
- X: "número de ensayos hasta obtener el r-abo éxito"

$$P(X = x) = \binom{x-1}{r-1}(1-p)^{x-r}p^r;\ \ E(x) = \frac{r(1-p)}{p};\ \ Var(x) = \frac{r(1-p)}{p^2} \tag{5.13}$$

5.4.6. Distribución geométrica

Es un caso particular de la distribución binomial negativa, cuando r = 1. Es decir, interesa conocer la probabilidad respecto a la cantidad de ensayos que se realizan hasta obtener el primer "éxito".

- Ensayos independientes.
- p(éxito) = p; p(fracaso) = 1 - p
- X: "número de ensayos hasta obtener el primer éxito"

$$P(X = x) = (1-p)^{x-1} \cdot p;\ \ E(x) = \frac{1}{p};\ \ Var(x) = \frac{1-p}{p^2} \tag{5.14}$$

5.5. Esquema resumen

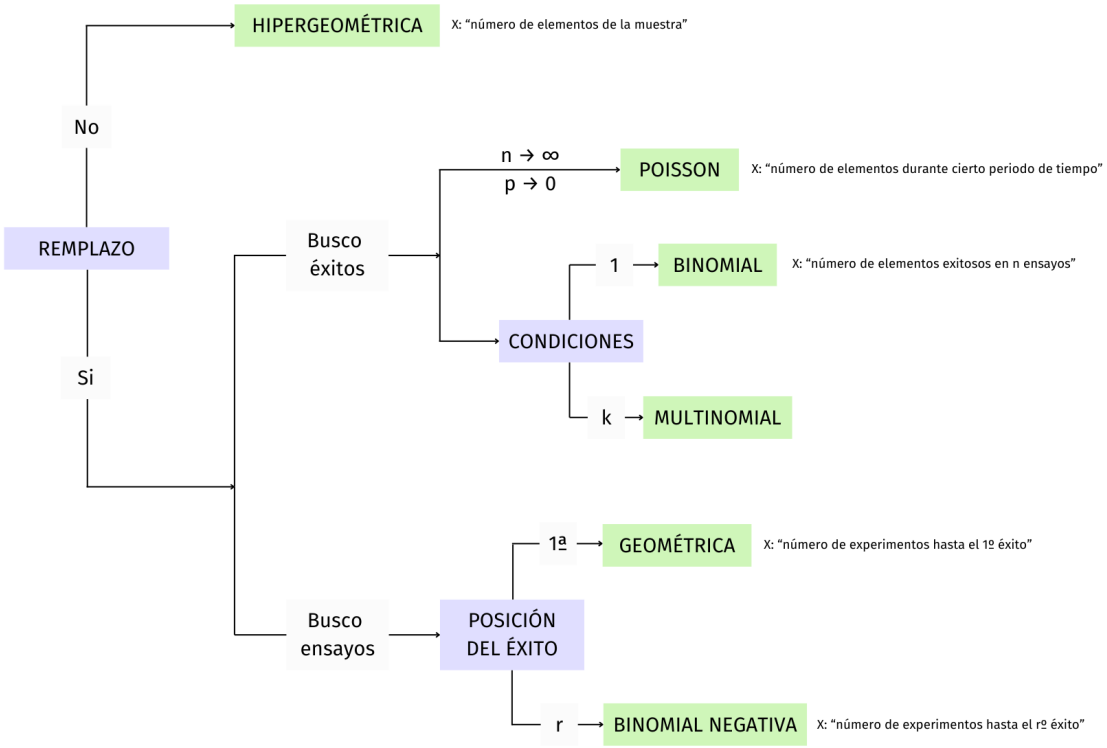


Figura 5.2: Esquema. Distribución de Variable Aleatoria Discreta.

6. Distribución de Variables Aleatorias Continuas

Recuerda lo que era una variable continua: una variable que puede tomar cualquier valor en la recta real. Por lo tanto, la distribución continua será la distribución utilizada para calcular las probabilidades de una variable continua.

6.1. Función de Distribución

Sea X una variable aleatoria discreta cuyos valores suponemos ordenados de menor a mayor. Llamaremos función de distribución de la variable X, y escribiremos *F(x)* a la función:

- $F(x) = P(X \leq x)$
- $P(x_1 \leq x \leq x_2) = F(x_2) - F(x_1)$
- $P(x > x_2) = 1 - P(x \leq x_2)$

6.2. Función de Densidad

Recuerda que nos interesan las probabilidades, pero aún no hemos definido una función de probabilidad. En variables continuas no podemos utilizar la función de probabilidad puesto que el caso favorable es único y los casos posibles son infinitos. $P(X = a) = \frac{1}{\infty} = 0$.

La **función de densidad** describe la probabilidad relativa según la cual dicha variable aleatoria tomará determinado valor. Se define como $f(x)$, como derivada de la función de distribución:

$$\text{f(x)} = \frac{dF(x)}{dx} \rightarrow F(x) = \int_{-\infty}^{x} f(x)dx \tag{6.1}$$

A continuación, se van a mostrar alguna de las propiedades de la función de densidad:

- $P(x_1 \leq x \leq x_2) = \int_{x_1}^{x_2} f(x)dx$
- $\int_{-\infty}^{\infty} f(x)dx = 1$
- $E(x) = \mu = \int_{-\infty}^{\infty} xf(x)dx$
- $Var(x) = \sigma_x^2 = \int_{-\infty}^{\infty} x^2 f(x)dx - \mu^2$

6.3. Distribuciones

6.3.1. Distribución normal

Una variable aleatoria continua X, sigue una distribución normal de media μ y desviación típica σ, y se designa por $\mathrm{N}(\mu, \sigma)$. Este tipo de distribuciones tienen la característica de que la media, la moda y la mediana tienen el mismo valor y se denomina μ.

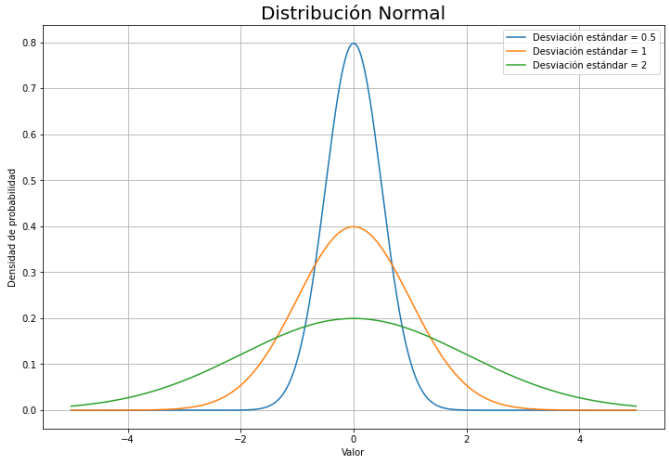


Figura 6.1: Distribución normal.

Funciones

La función de densidad de esta distribución y, en consecuencia, la función de distribución son las siguientes:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \rightarrow F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx \tag{6.2}$$

Como esta distribución es simétrica:

$$P(-z < Z < z) = F(z) - F(-z) = F(z) - [1 - F(z)] = 2F(z) - 1$$

Tipificación de la variable

Sin embargo, la más popular de todas las distribuciones normales es $N(0, 1)$. Esta es una distribución normal con una media de 0 y una desviación típica de 1, que denominaremos normal estándar. Además, las tablas que se van a utilizar serán las de distribución estándar.

Cuando no se dispone de una distribución normal estándar, es decir, la media no es 0 y la desviación no es 1, hay que tipificar la variable para disponer de esta distribución.

$$N(\mu \neq 0,\ \sigma \neq 1) \rightarrow N(\mu = 0,\ \sigma = 1)$$
$$P(X < x) \rightarrow P(Z < z)$$

Para ello tenemos que transformar la variable X que sigue una distribución N(μ, σ) en otra variable Z que siga una distribución N(0, 1).

$$F_{\mu,\,\sigma}(x) = P(X \leq x_i) = P\left(\frac{X-\mu}{\sigma} \leq \frac{x_i-\mu}{\sigma}\right) = P\left(Z \leq \frac{x_i-\mu}{\sigma}\right) = F_{0,\,1}\left(\frac{x-\mu}{\sigma}\right)$$

$$P(x_1 \leq X \leq x_2) = P\left(\frac{x_1-\mu}{\sigma} \leq Z \leq \frac{x_2-\mu}{\sigma}\right)$$

Teorema de Moivre

Sea X una variable binomial donde n es grande. Entonces la variable binomial X puede aproximarse mediante una variable normal. Esto es:

$$B \sim N(\mu,\ \sigma);\ X = N(np,\ \sqrt{np(1-p)}) \tag{6.3}$$

A la hora de tipificar la variable nos quedará:

$$z = \frac{x-np}{\sqrt{np(1-p)}} \tag{6.4}$$

Si queremos calcular la probabilidad de un punto, $P(X = x_i)$, como Z es una variable normal, la probabilidad de un punto será cero. Para superar este problema calcularemos $P(X = x_i) = P(x_i - 0{,}5 \leq x \leq x_i + 0{,}5)$.

$$B \sim N(\mu,\ \sigma) \rightarrow P(X = x_i) = P(x_i - 0{,}5 \leq Y \leq x_i + 0{,}5) \tag{6.5}$$

Así tenemos:

$$P(X = x_i) = P(x_i{-}0{,}5 \leq Y \leq x_i{+}0{,}5) = P\left(\frac{x_i - 0{,}5 - np}{\sqrt{np(1-p)}} < Z < \frac{x_i + 0{,}5 - np}{\sqrt{np(1-p)}}\right)$$

- $P(x_1 \leq X \leq x_2) = P(x_1 - 0{,}5 < Y < x_2 + 0{,}5)$
- $P(x_1 < X < x_2) = P(x_1 + 0{,}5 \leq Y \leq x_2 - 0{,}5)$

Ejemplos: Sea X una variable aleatoria binomial donde $X = B(16, 0{,}5)$. Ahora usando las tablas binomiales y el teorema de Moivre calcularemos las siguientes probabilidades:

1. $P(X = 8);\ X = B(16;\ 0{,}5)$

$$P(X = 8) = P(8 - 0{,}5 < Y < 8 + 0{,}5) = P(7{,}5 < Y < 8{,}5)$$

Tipificamos la variable:

$$P\left(\frac{7{,}5 - 160{,}5}{\sqrt{160{,}5(1 - 0{,}5)}} \leq Z \leq \frac{8{,}5 - 160{,}5}{\sqrt{160{,}5(1 - 0{,}5)}}\right) = P\left(-0{,}25 \leq Z \leq 0{,}25\right) =$$

$$= 1 - 2 \cdot P(z \geq 0{,}25) = 1 - 2 \cdot 0{,}4013 = 0{,}1974$$

2. $P(X < 6);\ X = B(16,\ 0{,}5)$

$$P(X < 6{+}0{,}5) = P(Y \leq 6{,}5) \sim P\left(Z < \frac{6{,}5 - 16 \cdot 0{,}5}{\sqrt{16 \cdot 0{,}5 \cdot (1 - 0{,}5)}}\right) = P(Z < -0{,}75) =$$

$$= P(Z > 0{,}75) = 0{,}2266$$

3. $P(6 < X < 10);\ X = B(16,\ 0{,}5)$

$$P(6 < X < 10) = P(6 - 0{,}5 \leq Y \leq 10 + 0{,}5) = P(5{,}5 \leq Y \leq 10{,}5) \sim$$

$$\sim P\left(\frac{5{,}5 - 16 \cdot 0{,}5}{\sqrt{16 \cdot 0{,}5 \cdot (1 - 0{,}5)}} < Z < \frac{10{,}5 - 16 \cdot 0{,}5}{\sqrt{16 \cdot 0{,}5 \cdot (1 - 0{,}5)}}\right) = P(-1{,}25 < Z < 1{,}25) =$$

$$= 1 - 2 \cdot P(Z > 1{,}25) = 1 - 2 \cdot 0{,}1056 = 0{,}7888$$

6.3.2. Distribución chi-cuadrado de Pearson

La función de distribución de chi-cuadrado, χ^2 , se genera como la suma de los cuadrados de n variables normales. Al ser la suma de cuadrados, siempre será positiva. Si $z_1,\ z_2,\ ...,\ z_n$ son variables aleatorias independientes, todas ellas con distribución normal estándar, entonces la variable aleatoria $z_1^2 + z_2^2 + ... + z_n^2$ sigue una distribución denominada Chi-cuadrado de Pearson con n grados de libertad, que se denota por χ_n^2.

$$\chi_n^2 = z_1^2 + z_2^2 + ... + z_n^2 \quad (z_i \sim N(0,\ 1)) \tag{6.6}$$

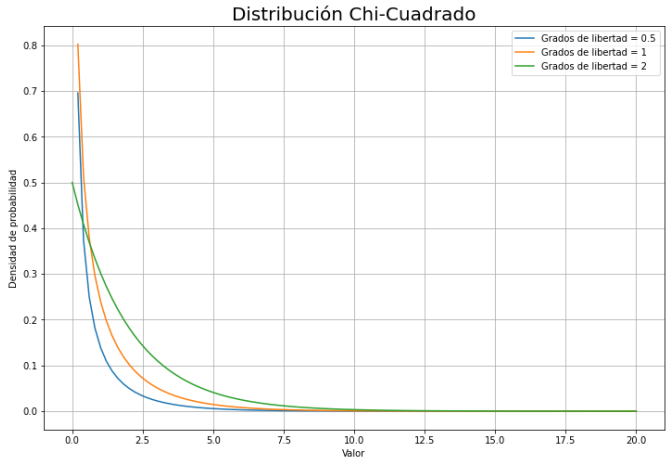


Figura 6.2: Distribución chi-cuadrado.

En este caso depende de un único parámetro, denominado grado de libertad n: χ^2_n. Esto también se calculará mediante tablas.

6.3.3. Distribución T de Student

Si z sigue una distribución normal estándar y x^2_n es independiente de z, entonces la variable aleatoria sigue una distribución denominada **t de Student** con n grados de libertad, que se denomina t_n.

$$t_n = \frac{z \cdot \sqrt{n}}{\sqrt{x^2_n}} \tag{6.7}$$

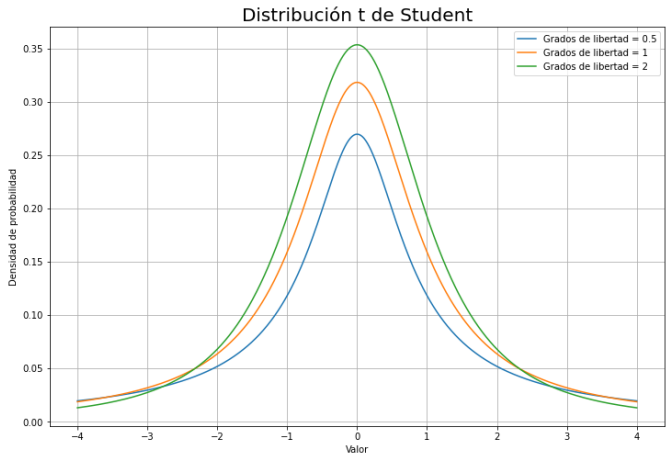


Figura 6.3: Distribución t student.

Tiene una forma similar a la distribución normal, pero tiene más punta. Esta distribución también es simétrica y está definida en el intervalo $(-\infty, \infty)$. Útil para muestras de tamaño pequeña.

6.3.4. Distribución F de Snedecor

Sean χ^2_{v1} y χ^2_{v2} dos distribuciones chi-cuadrado con v_1 y v_2 grados de libertad. La distribución Fisher-Snedecor se genera de la siguiente forma:

$$X^2_{v_1},\ X^2_{v_2};\quad F_{v_1;\,v_2} = \frac{X^2_{v_1}/v_1}{X^2_{v_2}/v_2} \tag{6.8}$$

La función de distribución se define en el intervalo $(0, \infty)$.

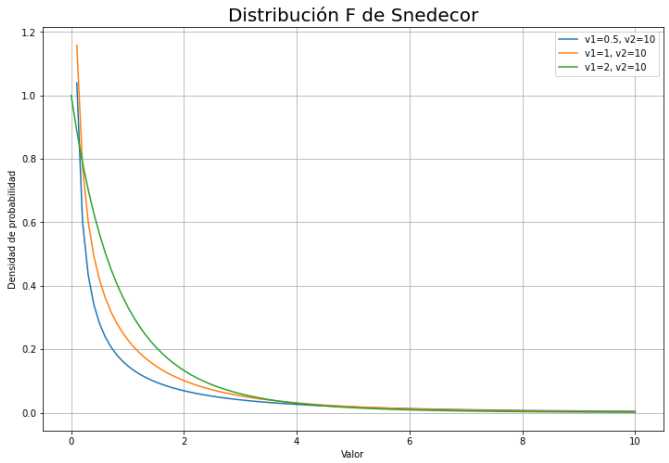


Figura 6.4: Distribución F de Snedecor.

Nota: Las tablas solo contienen datos $\alpha < 0{,}5$. Cuando $\alpha > 0{,}5$ se debe considerar la siguiente propiedad:

$$\frac{1}{F_{1-\alpha;\,v_2;\,v_1}} = F_{\alpha;\,v_1;\,v_2} \tag{6.9}$$

7. Muestreo y Estimación

7.1. Estimación de parámetros

El **parámetro de una población** es un número que expresa información sobre una característica importante de esa población.

Ejemplo: La media y la varianza son parámetros de una población; en un *t-student*, en cambio, el parámetro son los grados de libertad. Al tener una muestra de esa población vamos a estimar dichos parámetros.

7.1.1. Estimación puntual

El **estimador puntual** es la fórmula que nos dará una estimación del parámetro de la muestra.

Estimador de media y varianza

La media muestral ($\overline{x}$) y la varianza muestral (s^2) son estimadores de la media (μ) y la varianza (σ^2) poblacional respectivamente. Esto es:

$$\hat{y} = \overline{x} = \frac{1}{n} \cdot \sum_i x_i; \quad \hat{\sigma}^2 = s^2 = \frac{1}{n} \cdot \sum_i x_i^2 \tag{7.1}$$

La estimación puntual presenta un gran inconveniente: aún utilizando el mejor estimador de una característica poblacional o parámetro, no sólo no acertaremos en la estimación (la posibilidad de acertar es remota), sino que desconoceremos el grado de precisión y fiabilidad de la misma. Por ello, realizaremos la estimación basada en intervalos de confianza.

7.1.2. Intervalos de confianza

En los **intervalos de confianza** es donde se encuentra la estimación, para ello hay que ajustar el nivel de confianza o significancia (1 - α).

Nota: Si por ejemplo definimos un nivel de confianza al 95 %, significa que el intervalo de confianza contendrá el valor de verdad del parámetro poblacional (media o varianza) de nuestro interés con una probabilidad del 95 %. Dicho de forma, si construimos todas muestras posibles de tamaño n en la población dada, el 95 % de los intervalos de confianza calculados contendrán el valor de verdad del parámetro poblacional de nuestro interés y en el 5 % restante no se encontrará dicho parámetro.

INTERVALO DE CONFIANZA PARA LA MEDIA DE 1 POBLACIÓN NORMAL

1. Varianza poblacional conocida y muestra grande ($n \geq 30$)

$$T_\mu = \left[\overline{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}},\ \overline{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right] \tag{7.2}$$

2. Varianza poblacional desconocida y muestra grande ($n \geq 30$)

$$T_\mu = \left[\overline{x} - z_{\alpha/2} \cdot \frac{s}{\sqrt{n}},\ \overline{x} + z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}\right] \tag{7.3}$$

3. Varianza poblacional desconocida y muestra pequeña ($n < 30$)

$$T_\mu = \left[\overline{x} - t_{\alpha/2;\, n-1} \cdot \frac{s}{\sqrt{n}},\ \overline{x} + t_{\alpha/2;\, n-1} \cdot \frac{s}{\sqrt{n}}\right] \tag{7.4}$$

INTERVALO DE CONFIANZA PARA DIFERENCIA DE MEDIAS DE 2 POBLACIONES NORMALES

1. Varianzas poblacionales conocidas y muestras grandes ($n_x \geq 30, n_y \geq 30$)

$$T_{\mu_1-\mu_2} = \left[(\overline{x_1} - \overline{x_2}) - z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},\ (\overline{x_1} - \overline{x_2}) + z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right] \tag{7.5}$$

2. Varianza poblacionales desconocidas y muestras grandes ($n_x > 30, n_y > 30$)

$$T_{\mu_1-\mu_2} = \left[(\overline{x_1} - \overline{x_2}) - z_{\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},\ (\overline{x_1} - \overline{x_2}) + z_{\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right] \tag{7.6}$$

3. Varianzas poblaciones desconocidas pero iguales y muestras pequeñas ($n_x < 30, n_y < 30$)

$$T_{\mu_1-\mu_2} = \left[(\overline{x_1} - \overline{x_2}) \mp t_{\alpha/2;\, n_1+n_2-2} \cdot \sqrt{\frac{(n_1-1)\cdot s_1^2 + (n_2-1)\cdot s_2^2}{n_1+n_2-2}} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right] \tag{7.7}$$

4. Varianzas poblacionales desconocidas y muestras pequeñas ($n_x < 30, n_y < 30$)

$$T_{\mu_1-\mu_2} = \left[(\overline{x_1} - \overline{x_2}) \mp t_{\alpha/2;\, m} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right] \tag{7.8}$$

$$c = \frac{\frac{s_1^2}{n_1}}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}; \quad \frac{1}{m} = \frac{c^2}{n_1 - 1} + \frac{(1-c)^2}{n_2 - 1} \tag{7.9}$$

VARIANZA O DESVIACIÓN

1. VARIANZA O DESVIACIÓN DE UNA POBLACIÓN NORMAL

$$\text{Varianza: } T_{\sigma^2} = \left[\frac{(n-1)\cdot s^2}{\chi^2_{\alpha/2;\, n-1}}, \frac{(n-1)\cdot s^2}{\chi^2_{1-\alpha/2;\, n-1}} \right] \tag{7.10}$$

$$\text{Desviación típica: } T_{\sigma} = \left[\frac{\sqrt{n-1}\cdot s}{\sqrt{\chi^2_{\alpha/2;\, n-1}}}, \frac{\sqrt{n-1}\cdot s}{\sqrt{\chi^2_{1-\alpha/2;\, n-1}}} \right] \tag{7.11}$$

2. RAZÓN DE VARIANZAS O DESVIACIONES DE DOS POBLACIONES NORMALES

$$\text{Varianza: } I = \left[\frac{s_1^2}{s_2^2} \cdot \frac{1}{F_{\alpha/2;\, n_1-1;\, n_2-1}}, \frac{s_1^2}{s_2^2} \cdot F_{\alpha/2;\, n_1-1;\, n_2-1} \right] \tag{7.12}$$

$$\text{Desviación típica: } I = \left[\frac{s_1}{s_2} \cdot \frac{1}{\sqrt{F_{\alpha/2;\, n_1-1;\, n_2-1}}}, \frac{s_1}{s_2} \cdot \sqrt{F_{\alpha/2;\, n_1-1;\, n_2-1}} \right] \tag{7.13}$$

Notas:

1. Cuando se disponen de 2 muestras hay que ver de dónde provienen los datos
   - Muestras no pareadas: los datos provienen de dos poblaciones distintas.
   - Muestras pareadas: los datos provienen de una misma población. En ese caso, se calculará la diferencia $(X_i - Y_i)$ y se calculará el intervalo de confianza como si se tratase de una única población. La interpretación de los resultados se hará como si fueran dos poblaciones.
2. Cuando se calcula el intervalo de confianza de dos poblaciones normales pueden ocurrir los siguientes casos:
   - todo el intervalo positivo: $\hat{\mu}_A > \hat{\mu}_B$. En función del contexto comparativo podremos concluir qué población es mejor.

- un trozo del intervalo negativo y otro trozo positivo. Ocurre que a veces $\hat{\mu}_A > \hat{\mu}_B$ y a veces $\hat{\mu}_A < \hat{\mu}_B$. En este caso no se puede concluir nada.
- todo el intervalo negativo : $\hat{\mu}_A < \hat{\mu}_B$. En función del contexto comparativo podremos concluir qué población es mejor.

3. En la razón de varianzas o desviaciones:

   - todo el intervalo es mayor que 1: $\hat{\sigma}_A^{(2)} > \hat{\sigma}_B^{(2)}$ En función del contexto comparativo podremos concluir qué población es mejor.
   - todo el intervalo contiene a 1: no podemos concluir nada.
   - todo el intervalo es menor que 1: $\hat{\sigma}_A^{(2)} < \hat{\sigma}_B^{(2)}$ En función del contexto comparativo podremos concluir qué población es mejor.

7.1.3. Contrastes de normalidad

Para comprobar si una población sigue una distribución normal a través de una muestra, se realizan las siguientes pruebas:

$$\text{muestra} \begin{cases} n \geq 50 \rightarrow \text{Kolmogórov-Smirnov} \\ n < 50 \rightarrow \text{Shapiro-Wilk} \end{cases} \rightarrow \begin{cases} H_0 : \text{sigue una distribución normal} \\ H_1 : \text{no sigue una distribución normal} \end{cases} \rightarrow$$

	2 poblaciones	3 ó más poblaciones
H_0	t-Student	ANOVA
H_1	Wilcoxon	Kruskal Wallis

7.2. Muestreo

Se conoce como muestreo a la técnica para la selección de una muestra a partir de una población estadística.

7.2.1. Tipos de muestreo

- **Muestreo probabilístico:** Cada unidad de la población tiene una probabilidad de ser seleccionada para la muestra, la cual puede determinarse con precisión. Todos los elementos tienen la misma probabilidad de ser seleccionados.

  - Muestreo aleatorio simple: Este método permite calcular la probabilidad de extracción para cualquier muestra posible.

    - Sin reemplazo de los elementos.
    - Con reemplazo de los elementos.
  - Muestreo sistemático: Se utiliza cuando la población es grande o debe ser muestreada a lo largo del tiempo.
  - Muestreo estratificado: Consiste en dividir previamente la población en grupos homogéneos con respecto a alguna característica relevante para el estudio.
    - Asignación proporcional: El tamaño de la muestra es proporcional al tamaño de cada estrato.
    - Asignación óptima: La muestra prioriza estratos con mayor variabilidad, requiriendo conocimiento previo de la población.

- **Muestreo no probabilístico:** En este tipo de muestreo no se puede calcular la probabilidad de extracción de una muestra específica, ya que no todos los sujetos tienen la misma probabilidad de ser seleccionados.
  - Muestreo por cuotas: El investigador elige a los sujetos de la muestra dentro de cada estrato de manera libre.
  - Muestreo de bola de nieve: Partiendo de un grupo inicial de individuos que cumplen ciertos criterios, estos sirven como localizadores para otros con características similares.
  - Muestreo subjetivo por decisión razonada: Las unidades de la muestra se seleccionan en función de características específicas de manera racional y no al azar.

7.2.2. Variables

A parte de la media, la varianza y la desviación se pueden calcular metavariables.

- **Poblaciones infinitas:**

$$\text{De las medias: } \begin{cases} \mu_{\overline{x}} = \mu \\ \sigma^2_{\overline{x}} = \dfrac{\sigma^2}{n} \rightarrow \sigma_{\overline{x}} = \dfrac{\sigma}{\sqrt{n}} \end{cases} \rightarrow \overline{x} = \frac{x_1 + x_2 + x_3 + ... + x_n}{n}$$

$$\text{De las sumas: } \begin{cases} \mu_T = n \cdot \mu \\ \sigma^2_T = n \cdot \sigma^2 \rightarrow \sigma_T = \sqrt{n} \cdot \sigma \end{cases} \rightarrow x_1 + x_2 + x_3 + ... + x_n$$

- **Poblaciones finitas:**

$$\text{De las medias: } \begin{cases} \mu_{\overline{x}} = \mu \\ \sigma^2_{\overline{x}} = \dfrac{\sigma^2}{n} \cdot \dfrac{N-n}{N-1} \end{cases}$$

$$\text{De las sumas: } \begin{cases} \mu_T = n \cdot \mu \\ \sigma_T^2 = n \cdot \sigma^2 \cdot \dfrac{N-n}{N-1} \end{cases}$$

Ejemplo: $\mu = 5{,}2$; $\sigma = 0{,}5$

a. Media y desviación de las sumas y medias (n = 4).

$$\text{De las medias: } \begin{cases} \mu_{\overline{x}} = \mu = 5{,}2 \\ \sigma_{\overline{x}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{0{,}5}{\sqrt{4}} = 0{,}25 \end{cases}$$

$$\text{De las sumas: } \begin{cases} \mu_T = n \cdot \mu = 4 \cdot 5{,}2 = 20{,}8 \\ \sigma_T = \sqrt{n} \cdot \sigma = \sqrt{4} \cdot 0{,}5 = 1 \end{cases}$$

b. Probabilidad de que la suma de calificaciones obtenidas por 4 alumnos sea superior o igual a 22.

$$P(x \geq 22) = P\left(z \geq \frac{22 - 20{,}8}{1}\right) = P(z \geq 1{,}2) = 1 - P(z \leq 1{,}2) = 1 - 0{,}8849 = 0{,}1151$$

c. Probabilidad de que la media de las calificaciones sea menor a 4.5.

$$P(x < 4{,}5) = P\left(z < \frac{4{,}5 - 5{,}2}{0{,}25}\right) = P(z < -2{,}8) = 1 - P(z < 2{,}8) = 1 - 0{,}9974 = 0{,}0026$$

8. Contrastes de Hipótesis

Muchas veces nos interesa saber si un parámetro tomará un valor en concreto o si superará un umbral.

Ejemplo: Si un ingeniero dice que la longitud media de las piezas encargadas en su fábrica son de 5 cm, y otro no. Tendremos que aceptar o rechazar esta hipótesis.

8.1. Tipos de errores

Consideramos dos tipos de errores:

- Tipo 1: admitir la hipótesis cuando no es cierta.
- Tipo 2: rechazar la hipótesis cuando es cierta.

8.2. Tipos de hipótesis

- **Hipótesis unilateral:**

$$\begin{cases} H_0 : \theta \leq \theta_0 \\ H_1 : \theta > \theta_0 \end{cases} \qquad \begin{cases} H_0 : \theta \geq \theta_0 \\ H_1 : \theta < \theta_0 \end{cases}$$

- **Hipótesis bilateral:**

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{cases}$$

8.3. Intervalos de aceptación de hipótesis

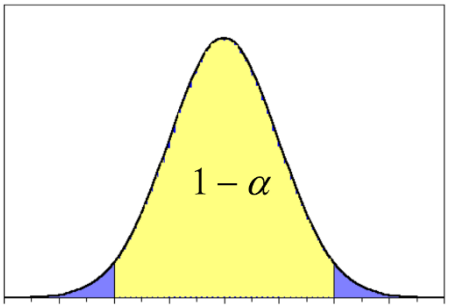
Son los intervalos estimados donde podría estar el estimador. Están directamente relacionados con intervalos de aceptación de la hipótesis. Para evitar duplicaciones en las fórmulas lo veremos a la vez. En ambos casos hay que fijar un nivel de confianza (n.c. = 1 - α).

1 POBLACIÓN NORMAL

1. **Varianza poblacional conocida y muestra grande** ($n \geq 30$)

   - Hipótesis bilateral

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases} \; ; \quad z_0 = \frac{(\overline{x} - \mu_0)}{\frac{\sigma}{\sqrt{n}}}; \quad z_0 \in [-Z_{\alpha/2};\ Z_{\alpha/2}]$$



Nota: Como se puede ver, el rango de aceptación es mucho mayor que el rango de rechazo. Por tanto, lo aceptaremos con preferencia, pero en lugar de aceptarlo diremos que no lo rechazamos.

- Hipótesis unilateral

$$\begin{cases} H_0 : \mu \leq \mu_0 \\ H_1 : \mu > \mu_0 \end{cases} ; \quad z_0 = \frac{(\overline{x} - \mu_0)}{\frac{\sigma}{\sqrt{n}}}; \quad z_0 \in [-\infty;\ Z_\alpha]$$

$$\begin{cases} H_0 : \mu \geq \mu_0 \\ H_1 : \mu < \mu_0 \end{cases} ; \quad z_o = \frac{(\overline{x} - \mu_0)}{\frac{\sigma}{\sqrt{n}}}; \quad z_o \in [-Z_\alpha;\ +\infty]$$

2. Varianza poblacional desconocida y muestra grande $(n \geq 30)$

- Hipótesis bilateral

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases} ; \quad z_0 = \frac{(\overline{x} - \mu_0)}{\frac{s}{\sqrt{n}}}; \quad z_0 \in [-Z_{\alpha/2};\ Z_{\alpha/2}]$$

- Hipótesis unilateral

$$\begin{cases} H_0 : \mu \leq \mu_0 \\ H_1 : \mu > \mu_0 \end{cases} ; \quad z_0 = \frac{(\overline{x} - \mu_0)}{\frac{s}{\sqrt{n}}}; \quad z_0 \in [-\infty;\ Z_\alpha]$$

$$\begin{cases} H_0 : \mu \geq \mu_0 \\ H_1 : \mu < \mu_0 \end{cases} ; \quad z_0 = \frac{(\overline{x} - \mu_0)}{\frac{s}{\sqrt{n}}}; \quad z_0 \in [-Z_\alpha;\ +\infty]$$

3. Varianza poblacional desconocida y muestra pequeña $(n < 30)$

- Hipótesis bilateral

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases} ; \quad t_0 = \frac{(\overline{x} - \mu_0)}{\frac{s}{\sqrt{n}}}; \quad t_0 \in [-t_{\alpha/2;\, n-1};\ Z_{\alpha/2;\, n-1}]$$

- Hipótesis unilateral

$$\left\{ \begin{array}{l} H_0 : \mu \leq \mu_0 \\ H_1 : \mu > \mu_0 \end{array} \right. ; \quad t_0 = \frac{(\overline{x} - \mu_0)}{\frac{s}{\sqrt{n}}}; \quad t_0 \in [-\infty;\ t_{\alpha;\, n-1}]$$

$$\left\{ \begin{array}{l} H_0 : \mu \geq \mu_0 \\ H_1 : \mu < \mu_0 \end{array} \right. ; \quad t_0 = \frac{(\overline{x} - \mu_0)}{\frac{s}{\sqrt{n}}}; \quad t_0 \in [-t_{\alpha;\, n-1};\ +\infty]$$

2 POBLACIONES NORMALES

1. **Varianzas poblacionales conocidas y muestras grandes** ($n_x \geq 30; n_y \geq 30$)

   - Hipótesis bilateral

$$\left\{ \begin{array}{l} H_0 : \mu_x = \mu_y \\ H_1 : \mu_x \neq \mu_y \end{array} \right. ; \quad z_o = \frac{(\overline{x} - \overline{y})}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}; \quad z_o \in [-Z_{\alpha/2};\ Z_{\alpha/2}]$$

   - Hipótesis unilateral

$$\left\{ \begin{array}{l} H_0 : \mu_x \leq \mu_y \\ H_1 : \mu_x > \mu_y \end{array} \right. ; \quad z_o = \frac{(\overline{x} - \overline{y})}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}; \quad z_o \in [-\infty;\ Z_{\alpha}]$$

$$\left\{ \begin{array}{l} H_0 : \mu_x \geq \mu_y \\ H_1 : \mu_x < \mu_y \end{array} \right. ; \quad z_o = \frac{(\overline{x} - \overline{y})}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}; \quad z_o \in [-Z_{\alpha};\ +\infty]$$

2. **Varianzas poblacionales desconocidas y muestras grandes ($n_x >$30, $n_y >$30)**

   - Hipótesis bilateral

$$\left\{ \begin{array}{l} H_0 : \mu_x = \mu_y \\ H_1 : \mu_x \neq \mu_y \end{array} \right. ; \quad z_o = \frac{(\overline{x} - \overline{y})}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}}; \quad z_o \in [-Z_{\alpha/2};\ Z_{\alpha/2}]$$

- Hipótesis unilateral

$$\begin{cases} H_0 : \mu_x \leq \mu_y \\ H_1 : \mu_x > \mu_y \end{cases} ; \quad z_o = \frac{(\overline{x} - \overline{y})}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}}; \quad z_o \in [-\infty;\ Z_\alpha]$$

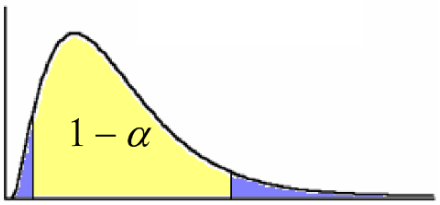
$$\begin{cases} H_0 : \mu_x \geq \mu_y \\ H_1 : \mu_x < \mu_y \end{cases} ; \quad z_o = \frac{(\overline{x} - \overline{y})}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}}; \quad z_o \in [-Z_\alpha;\ +\infty]$$

VARIANZA O DESVIACIÓN

1. De una población

- Hipótesis bilateral

$$\begin{cases} H_0 : \sigma = \sigma_0 \\ H_1 : \sigma \neq \sigma_0 \end{cases} ; \quad \chi_0^2 = \frac{(n-1) \cdot s^2}{\sigma^2}; \quad \chi_0^2 \in [\chi^2_{1-\alpha/2;\, n-1};\ \chi^2_{1-\alpha/2;\, n-1}]$$



- Hipótesis unilateral

$$\begin{cases} H_0 : \sigma \leq \sigma_0 \\ H_1 : \sigma > \sigma_0 \end{cases} ; \quad \chi_0^2 = \frac{(n-1) \cdot s^2}{\sigma^2}; \quad \chi_0^2 \in [0,\ \chi^2_{\alpha;\, n-1}]$$

$$\begin{cases} H_0 : \sigma \geq \sigma_0 \\ H_1 : \sigma < \sigma \end{cases} ; \quad \chi_0^2 = \frac{(n-1) \cdot s^2}{\sigma^2}; \quad \chi_0^2 \in [\chi^2_{1-\alpha;\, n-1};\ +\infty]$$

Nota: Con la distribución chi-cuadrado nos encontramos con una región de incertidumbre.

Si χ_0 cae dentro del la región de incertidumbre $[\chi^2_{1-\alpha;\, n-1}, \chi^2_{\alpha;\, n-1}]$ no podemos sacar ninguna conclusión de las varianzas. Una forma de solucionar este problema es aumentar la muestra.

9. Análisis de la Varianza (ANOVA)

Esta prueba se utiliza para determinar si las medias muestrales provienen de poblaciones con medias iguales, cuando hay más de dos poblaciones en estudio.

Supongamos que queremos comparar k poblaciones diferentes. Entonces nuestro contraste de hipótesis será el siguiente:

$$\begin{cases} H_o : \mu_1 = \mu_2 = \mu_3 = ... = \mu_n \\ H_i : \text{caso contrario} \end{cases}$$

Para este contraste, necesitamos hacer las siguientes suposiciones:

- Todas las poblaciones siguen una distribución normal. En el caso de que no siguieran una distribución normal, habría que aplicar el análisis de Kruscal-Wallis.
- Todas las poblaciones son independientes.
- Todas las poblaciones tienen la misma varianza.

Si se cumplen las tres suposiciones. El contraste de hipótesis se realizará mediante tablas ANOVA.

9.1. Construcción de la tabla ANOVA

	A_1	A_2	...	A_n
	x_{11}	x_{12}	...	x_{ik}
	x_{21}	x_{22}	...	x_{2k}
	...	...	...	...
		$x_{n_2 2}$	...	$x_{n_k k}$
	$x_{n_1 1}$		...	
B	B_1	B_2	...	B_k
N	n_1	n_2	...	n_k

$$B = \sum_{i=1}^{k} B_i; \quad \text{Tamaño muestral total: } N = \sum_{i=1}^{k} n_i \tag{9.1}$$

Dado que hemos asumido la misma varianza σ^2 para todas las poblaciones, este valor se puede estimar de dos maneras: usando la varianza interclase o usando la varianza intraclase.

- **Varianza interclase:**

$$S_A^2 = \frac{1}{k-1} \cdot \left[\sum_{j=1}^{k} \frac{B_j^2}{n_j} - \frac{B^2}{N} \right] \tag{9.2}$$

donde k es el número de poblaciones y *(k - 1)* es el grado de libertad.

- **Varianza intraclase:**

$$S_B^2 = \frac{1}{N-k} \cdot \sum_{j=1}^{n} \sum_{i=1}^{n} x_{ij}^2 - \sum_{j=1}^{n} \frac{B_j^2}{n_j} \tag{9.3}$$

donde *(N - k)* es el grado de libertad.

Por lo tanto, dividiendo las varianzas interclase e intraclase, obtenemos la distribución de Fisher Snedecor, con $(k-1)$ y $(N-k)$ grados de libertad.

$$F = \frac{S_A^2}{S_B^2} \qquad \begin{cases} \mathrm{H}_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_n \quad \rightarrow \quad F \leq F_{\alpha;\, k-1;\, N-k} \\ \mathrm{H}_1 : \text{caso contrario} \quad \rightarrow \quad F > F_{\alpha;\, k-1;\, N-k} \end{cases} \tag{9.4}$$

9.2. ANOVA con dos factores

Dispondremos de dos tipos de contrastes por columnas y filas.

$$\begin{cases} H_o : \mu_{Z_1} = \mu_{Z_2} = \mu_{Z_3} = ... = \mu_{Z_k} \\ H_i : \text{caso contrario} \end{cases} \quad \text{y} \quad \begin{cases} H_o : \mu_{E_1} = \mu_{E_2} = \mu_{E_3} = ... = \mu_{E_n} \\ H_i : \text{caso contrario} \end{cases}$$

Dado que hemos asumido la misma varianza σ^2 para todas las poblaciones, este valor se puede estimar de dos maneras: usando la varianza interclase o usando la varianza intraclases.

- **Varianza interclase:** La **varianza interclase** se calculará como en el caso anterior, pero en este caso por columna o fila.

$$KB_{AZ} = \sum_{j=1}^{k} \frac{B_j^2}{n} - \frac{B^2}{N}; \;\; k-1 \text{ grados de libertad}; \;\; S_Z^2 = \frac{KB_{AZ}}{k-1} \tag{9.5}$$

$$KB_{AE} = \sum_{j=1}^{k} \frac{D_j^2}{n} - \frac{D^2}{N}; \;\; n-1 \text{ grados de libertad}; \;\; S_E^2 = \frac{KB_{AE}}{n-1} \tag{9.6}$$

- **Varianza intraclase:**

$$KB_{B,H} = \sum_{j=1}^{k} \sum_{i=1}^{n} x_{ij}^2 - \sum_{j=1}^{k} \frac{B_j^2}{n} - \sum_{i=1}^{n} \frac{B_i^2}{n} - \frac{B^2}{N}; \;\; (n-1)(k-1) \text{ grados de libertad} \tag{9.7}$$

$$S_B^2 = \frac{KB_{B,H}}{(n-1)(k-1)}$$

Dado que esta última varianza intraclase es muy grande, podemos calcular la varianza total,

$$KB_T = \sum_{j=1}^{k}\sum_{i=1}^{n} x_{ij}^2 - \frac{B^2}{N} \text{ donde } KB_T = KB_{B,H} + KB_{AZ} + KB_{AE} \quad (9.8)$$

9.3. Análisis de variación

Variación	SK	Grado de libertad	S^2	F
Interclase (KB_A)	$\sum_{j=1}^{k} \frac{B_j^2}{n_j} - \frac{B^2}{N}$	k - 1	$\frac{KB_A}{k-1} = S_A^2$	$\frac{S_A^2}{S_B^2}$
Intraclase (KB_B)	$\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}^2 - \sum_{j=1}^{2} \frac{B_j^2}{n_j}$	N - k	$\frac{KB_B}{N-k} = S_B^2$	
Total	$\sum_{j=1}^{k}\sum_{i=i}^{n_j} x_{ij}^2 - \frac{B^2}{N}$	N - 1		

Variación	SK	Grado de libertad	S^2	F
Entre filas	KB_{AE}	n - 1	$\frac{KB_{AE}}{n-1} = S_E^2$	$\frac{S_E^2}{S_B^2}$
Entre columnas	KB_{AZ}	k - 1	$\frac{KB_{AZ}}{k-1} = S_Z^2$	$\frac{S_Z^2}{S_B^2}$
Variación residual	$\sum_{j=1}^{k}\sum_{i=i}^{n} x_{ij}^2 - \frac{B_{Z_j}^2}{n} - \sum_{i=1}^{n} \frac{B_{E_i}^2}{k} + \frac{B^2}{n \cdot k}$	(n - 1)(k - 1)	$\frac{KB_{AH}}{(n-1)(k-1)} = S_B^2$	

10. Ejercicios

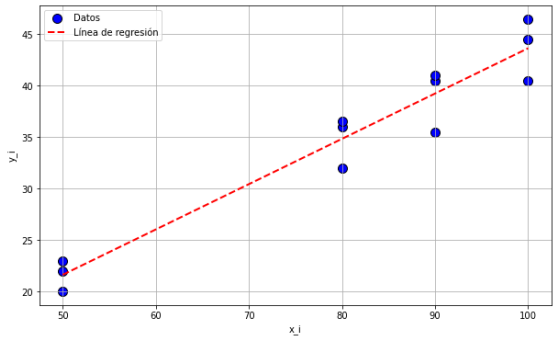
10.1. Regresión y Correlación

1. **Se realiza un análisis de sangre entre 12 personas. El volumen de sangre y el recuento de glóbulos rojos de cada individuo se miden en porcentajes. Los resultados se almacenan en la siguiente tabla:**

Persona	x_i	y_i
1	100	40,5
2	100	44,5
3	100	46,5
4	90	35,5
5	90	40,5
6	90	41,0
7	80	32,0
8	80	32,0
9	80	36,5
10	50	20,0
11	50	22,0
12	50	23,0

Calcule la recta y el error de regresión, y el coeficiente de correlación.

Antes de comenzar, veamos gráficamente cómo puede verse la regresión.



Se parece a la regresión lineal. Por lo tanto, se va a crear una tabla auxiliar:

Persona	x_i	y_i	$x_i \cdot y_i$	x_i^2	y_i^2
1	100	40,5	4050	10000	1640,25
2	100	44,5	4450	10000	1980,25
3	100	46,5	4650	10000	2162,25
4	90	35,5	3195	8100	1260,25
5	90	40,5	3645	8100	1640,25
6	90	41,0	3690	8100	1681,00
7	80	32,0	2560	6400	1024,00
8	80	36,0	2880	6400	1296,00
9	80	36,5	2920	6400	1332,25
10	50	20,0	1000	2500	400,00
11	50	22,0	1100	2500	484,00
12	50	23,0	1150	2500	529,00
Total	960	418,0	35290	81000	15429,50

Entonces,

$$b = \frac{cov(x,\ y)}{s_x^2} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i \cdot y_i - \overline{x} \cdot \overline{y}}{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} = \frac{35290/12 - 960/12 \cdot 418/12}{81000/12 - (960/12)^2} = 0{,}44047$$

$$a = \overline{y} - b \cdot \overline{x} = 418/12 - 0{,}44047 \cdot 960/12 = -0{,}40476$$

Por lo tanto,

$$\hat{y} = -0{,}40476 + 0{,}44047 \cdot x$$

Entonces la cantidad estimada de glóbulos rojos para una persona con una presión arterial de 100 será:

$$\hat{y} = -0{,}40476 + 0{,}44047 \cdot 100 = 43{,}642$$

En consecuencia, ¿cuál es el error cuadrático?

$$s_{ey} = \sqrt{\frac{\sum y_i^2 - a \cdot \sum y_i - b \cdot \sum x_i \cdot y_i}{n-2}} =$$

$$= \sqrt{\frac{15429{,}5 - (-0{,}40476) \cdot 418 - 0{,}44047 \cdot 35290}{12-2}} = 2{,}359$$

Calculemos el coeficiente de correlación:

$$r = \frac{s_x}{s_y} \cdot b = \frac{\sqrt{81000/12 - (960/12)^2}}{\sqrt{15429,5/12 - (418/12)^2}} \cdot 0,44047 = 0,968$$

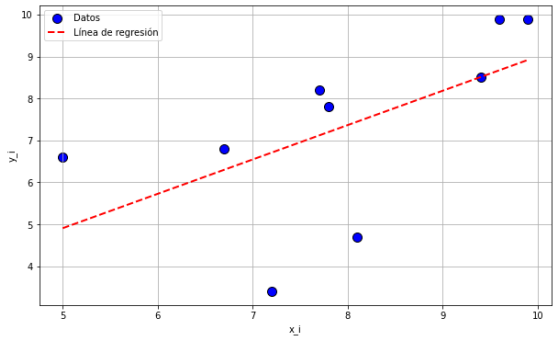
El valor del coeficiente de correlación obtenido es muy cercano a la unidad, por lo tanto hay una relación muy grande.

2. Las calificaciones de 9 estudiantes se muestran en la siguiente tabla. X son las calificaciones del primer cuatrimestre e Y del segundo:

Persona	x_i	y_i
1	7,7	8,2
2	5,0	6,6
3	7,8	7,8
4	7,2	3,4
5	8,1	4,7
6	9,4	8,5
7	9,6	9,9
8	9,9	9,9
9	6,7	6,8
Total	71,4	6,8

Para ver el efecto de la nota obtenida en el primer cuatrimestre sobre el segundo cuatrimestre, calcule la recta y el error de regresión.

Antes de comenzar, veamos gráficamente cómo se vería la regresión:



Se parece a la regresión lineal. Por lo tanto, se va a crear una tabla auxiliar:

Persona	x_i	y_i	$x_i \cdot y_i$	x_i^2	y_i^2
1	7,7	8,2	63,14	59,29	67,24
2	5,0	6,6	33,00	25,00	43,56
3	7,8	7,8	60,84	60,84	60,84
4	7,2	3,4	24,48	51,84	11,56
5	8,1	4,7	38,07	65,61	22,09
6	9,4	8,5	79,90	88,36	72,25
7	9,6	9,9	95,04	92,16	98,01
8	9,9	9,9	98,01	98,01	98,01
9	6,7	6,8	45,56	44,89	46,24
Total	71,4	65,8	538,04	586,00	519,80

Entonces,

$$b = \frac{cov(x,\ y)}{s_x^2} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i \cdot y_i - \overline{x} \cdot \overline{y}}{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}} = \frac{538,04/9 - 71,4/9 \cdot 65,8/9}{586/9 - (71,4/9)^2} = 0,819$$

$$a = \overline{y} - b \cdot \overline{x} = 65,8/9 - 0,819 \cdot 71,4/9 = 0,814$$

Por lo tanto,

$$\hat{y} = 0,814 + 0,819 \cdot x$$

Entonces la persona que obtuvo 8,5 en el primer cuatrimestre, la calificación estimada que obtendrá en el segundo será:

$$\hat{y} = 0,814 + 0,819 \cdot 8,5 = 7,776$$

En consecuencia, ¿cuál es el error cuadrático?

$$s_{ey} = \sqrt{\frac{\sum y_i^2 - a \cdot \sum y_i - b \cdot \sum x_i \cdot y_i}{n-2}} =$$

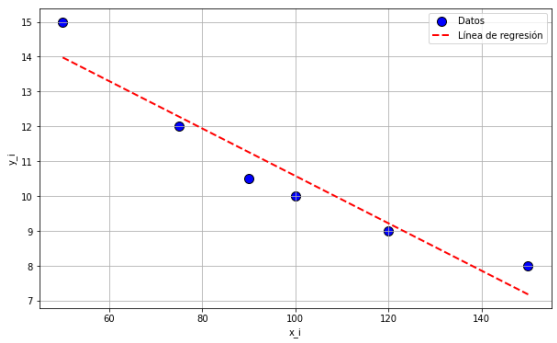
$$= \sqrt{\frac{519,8 - 0,814 \cdot 65,8 - 0,819 \cdot 71,4}{9-2}} = 7,632$$

3. La potencia (X) de distintos tipos de coches, mediada en caballos, y la capacidad de aceleración (Y), medida en segundos para llegar de 0 a 100, se resume en la siguiente tabla:

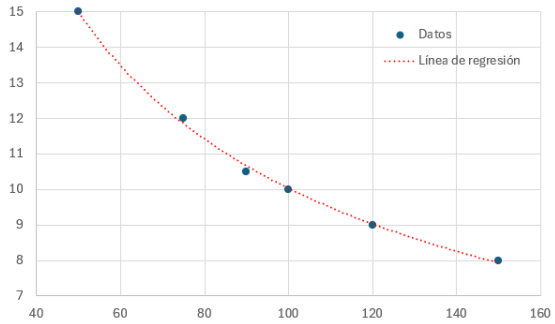
Coche	x_i	y_i
1	50	15,0
2	75	12,0
3	90	10,5
4	100	10,0
5	120	9,0
6	150	8,0

Calcule la curva y el error de regresión.

Antes de comenzar, veamos gráficamente cómo puede verse la regresión.



No se parece mucho a la regresión lineal. Veamos si está más cerca de la regresión potencial:



Es evidente que esta recta se acerca mucho más a nuestros datos. Entonces, ¿cómo calculamos los coeficientes?

$$y = \alpha \cdot x^{\beta} \rightarrow ln\, y = ln(\alpha \cdot x^{\beta}) = ln\, \alpha + ln(x^{\beta}) = ln\, \alpha + \beta \cdot (ln\, x)$$

Tenemos una regresión potencial. En la siguiente tabla se redondea a dos decimales pero los cálculos posteriores se realizan con todos los decimales.

Coche	x_i	y_i	$ln\, x_i$	$ln\, y_i$	$x_i \cdot y_i$	$x_i \cdot ln\, y_i$	$ln\, x_i \cdot ln\, y_i$	x_i^2	$(ln\, x_i)^2$	y_i^2	$(ln\, y_i)^2$
1	50	15,0	3,91	2,71	750	135,40	10,59	2500	15,30	225,00	7,33
2	75	12,0	4,32	2,48	900	186,37	10,73	5625	18,64	144,00	6,17
3	90	10,5	4,49	2,35	945	211,62	10,58	8100	20,25	110,25	5,53
4	100	10,0	4,61	2,30	1000	230,26	10,60	10000	21,21	100,00	5,30
5	120	9,0	4,79	2,19	1080	263,67	10,52	14400	22,92	81,00	4,83
6	150	8,0	5,01	2,08	1200	311,92	10,42	22500	25,11	64,00	4,32
Total	585	64,5	27,13	14,12	5875	1339,24	63,45	63125	123,43	724,25	33,49

Entonces,

$$b = \frac{cov(ln\, x, ln\, y)}{s^2_{ln(x)}} = \frac{\frac{1}{n}\sum_{i=1}^{n} ln\, x_i \cdot ln\, y_i - \overline{ln\, x} \cdot \overline{ln\, y}}{\frac{\sum_{i=1}^{n}(ln\, x_i - \overline{ln\, x})^2}{n}} =$$

$$= \frac{63,45/6 - 27,13/6 \cdot 14,12/6}{123,43/6 - (27,13/6)^2} = -0,578646$$

$ln\, a = \overline{ln\, y} - b \cdot \overline{ln\, x} = (14,12/6) - (-0,578646) \cdot (27,13/6) = 4,969777663 \rightarrow a = 143,994868$

Por lo tanto,

$$\hat{y} = 143,994868 \cdot x^{-0,578646}$$

Entonces el tiempo estimado para pasar de 0 a 100 para el coche de 76 caballos será:

$$\hat{y} = 143,994868 \cdot 76^{-0,578646} = 11,7495^{***} \text{ segundos}$$

En consecuencia, ¿cuál es el error cuadrático?

$$s_{ey} = \sqrt{\frac{\sum (ln\ y_i)^2 - ln\ a \cdot \sum ln\ y_i - b \cdot \sum ln\ x_i \cdot ln\ y_i}{n-2}} =$$

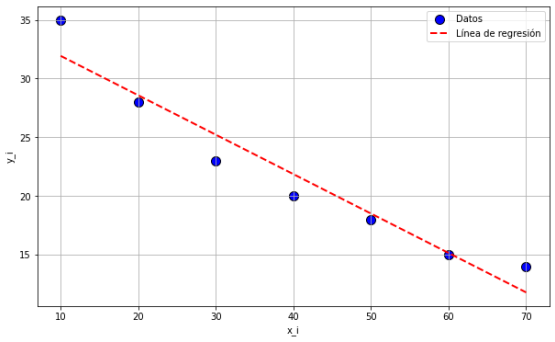
$$= \sqrt{\frac{33,49 - 4,97 \cdot 14,12 - (-0,578646) \cdot 63,45}{6-2}} = 0{,}048339251$$

4. **Han empezado a fabricar una nueva pieza en una fábrica. Para analizar la eficiencia de los trabajadores a lo largo del tiempo se han añadido los siguientes datos: el número de minutos para realizar la pieza (X) y el número de días que llevan fabricando la pieza en la fábrica (Y).**

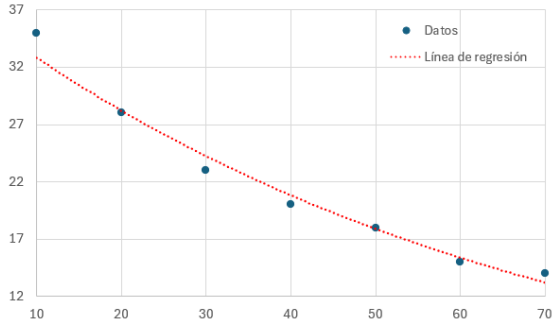
Persona	x_i	y_i
1	10	35
2	20	28
3	30	23
4	40	20
5	50	18
6	60	15
7	70	14

Calcule la curva y el error de regresión.

Antes de comenzar, veamos gráficamente cómo puede verse la regresión.



No se parece mucho a una regresión lineal. Veamos si se acerca más a la regresión exponencial:



Está claro que esta línea se acerca mucho más a nuestros datos. Entonces, ¿cómo calculamos los coeficientes?

$$y = \alpha \cdot e^{\beta \cdot x} \rightarrow ln\ y = ln\ (\alpha \cdot e^{\beta \cdot x}) = ln\ \alpha + ln\ (e^{\beta \cdot x}) = ln\ \alpha + \beta \cdot x$$

Tenemos una regresión exponencial. Por lo tanto, creemos la siguiente tabla auxiliar:

Persona	x_i	y_i	$ln\ x_i$	$ln\ y_i$	$x_i \cdot y_i$	$x_i \cdot ln\ y_i$	$ln\ x_i \cdot ln\ y_i$	x_i^2	$(ln\ x_i)^2$	y_i^2	$(ln\ y_i)^2$
1	10	35	2,30	3,56	350	35,55	8,19	100	5,30	1225	12,64
2	20	28	2,99	3,33	560	66,64	9,98	400	8,97	784	11,10
3	30	23	3,40	3,14	690	94,06	10,66	900	11,57	529	9,83
4	40	20	3,69	2,99	800	119,83	11,05	1600	13,61	400	8,97
5	50	18	3,91	2,89	900	144,52	11,31	2500	15,30	324	8,35
6	60	15	4,09	2,71	900	162,48	11,09	3600	16,76	225	7,33
7	70	14	4,25	6,34	980	184,73	11,21	4900	18,05	196	6,96
Total	280	153	24,64	21,25	5180	807,82	73,49	14000	89,56	3683	65,20

Entonces,

$$b = \frac{cov(x,\ ln\ y)}{s_x^2} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i \cdot ln\ y_i - \overline{x} \cdot \overline{ln\ y}}{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}} =$$

$$= \frac{807,82/7 - 280/7 \cdot 21,25/7}{14000/7 - (280/7)^2} = -0,0151$$

$$ln\ a = \overline{ln\ y} - b \cdot \overline{x} = (21,25/7) - (-0,0151) \cdot (280/7) = 3,639 \rightarrow a = e^{ln\ a} = e^{3,639} = 38,05376393$$

Por lo tanto,

$$\hat{y} = 38,05 \cdot e^{-0,0151 \cdot x}$$

En consecuencia, ¿cuál es el error cuadrático?

$$s_{ey} = \sqrt{\frac{\sum (ln\ y_i)^2 - ln\ a \cdot \sum ln\ y_i - b \cdot \sum x_i \cdot ln\ y_i}{n-2}} =$$

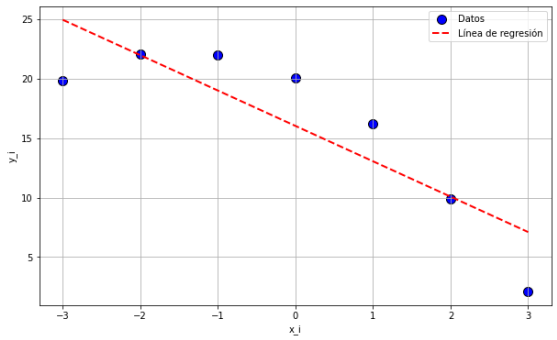
$$= \sqrt{\frac{65,20 - 3,639 \cdot 21,25 - (-0,0151) \cdot 807,82}{7-2}} = 0,117755679$$

5. **La altura de un proyectil disparado desde un acorazado se midió en los primeros 7 segundos después del lanzamiento. Para simplificar los cálculos se ha tomado el inicio de los tiempos como se deseaba. Los datos obtenidos son los siguientes:**

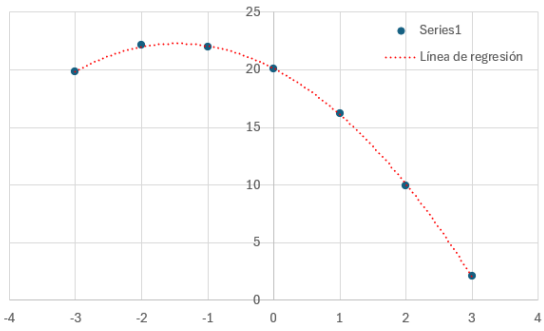
x_i: Tiempo (s)	y_i: Altura (m)
-3	19,8
-2	22,1
-1	22,0
0	20,1
1	16,2
2	9,9
3	2,1

Calcule la curva de regresión o parábola en este caso y el error de regresión.

Antes de comenzar, veamos gráficamente cómo puede verse la regresión.



No se parece mucho a una regresión lineal. Veamos si se acerca más a la regresión parabólica:



Está claro que esta curva se acerca mucho más a nuestros datos. Entonces, ¿cómo calculamos los coeficientes? Resolviendo el siguiente sistema de ecuaciones:

$$\begin{cases} a \cdot n + b \cdot \sum_i x_i + c \cdot \sum_i x_i^2 = \sum_i y_i \\ a \cdot \sum_i x_i + b \cdot \sum_i x_i^2 + c \cdot \sum_i x_i^3 = \sum_i x_i \cdot y_i \\ a \cdot \sum_i x_i^2 + b \cdot \sum_i x_i^3 + c \cdot \sum_i x_i^4 = \sum_i x_i^2 \cdot y_i \end{cases}$$

Para ello, creemos una tabla auxiliar:

Nº	x_i	y_i	$x_i \cdot y_i$	x_i^2	x_i^3	x_i^4	$x_i^2 \cdot y_i$	y_i^2
1	-3	19,8	-59,4	9	-27	81	178,2	329,04
2	-2	22,1	-44,2	4	-8	16	88,4	488,41
3	-1	22,0	-22	1	-1	1	22	484
4	0	20,1	0	0	0	0	0	404,01
5	1	16,2	16,2	1	1	1	16,2	262,44
6	2	9,9	19,8	4	2	16	39,6	98,01
7	3	2,1	6,3	9	27	81	18,9	4,41
Total	0	112,2	-83,3	28	0	196	363,3	2133,32

Entonces,

$$\begin{cases} a \cdot 7 + b \cdot 0 + c \cdot 28 = 112,2 \\ a \cdot 0 + b \cdot 28 + c \cdot 0 = -83,3 \\ a \cdot 28 + b \cdot 0 + c \cdot 196 = 363,3 \end{cases} \rightarrow \begin{cases} a = 20,1 \\ b = -2,975 \\ c = -1,0178 \end{cases}$$

Por lo tanto,

$$\hat{y} = 20,1 - 2,975 \cdot x - 1,0178 \cdot x^2$$

Finalmente, calcularemos el error cuadrático:

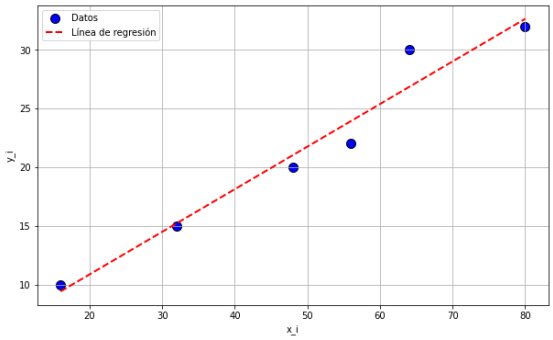
$$s_{ey} = \sqrt{\frac{\sum y^2 - a \cdot \sum y - b \cdot \sum x \cdot y - c \cdot \sum x^2 \cdot y}{n-2}} =$$

$$= \sqrt{\frac{2133,32 - 20,1 \cdot 112,2 - (-2,975) \cdot (-83,3) - (-1,0178) \cdot 363,3}{7-2}} = 0,09923$$

6. De una determinada empresa se conocen los siguientes datos, referidos al volumen de ventas (en miles de euros) y el gasto en publicidad (en euros) de los últimos 6 años:

x_i: Gastos publicidad (euros)	y_i: Volumen de ventas (miles euros)
16	10
32	15
48	20
56	22
64	30
80	32

a. ¿Existe relación lineal entre las ventas de la empresa y sus gastos en publicidad?



Observándolo podemos decir que existe relación lineal entre ambas variables. Ahora calculamos el coeficiente de determinación lineal para obtener una medida descriptiva del grado de asociación lineal que existen entre las variables. Para realizar este cálculo hay que crear la siguiente tabla:

	x_i	y_i	$x_i \cdot y_i$	x_i^2	y_i^2
	16	10	160	256	100
	32	15	480	1024	225
	48	20	960	2304	400
	56	22	1232	3136	484
	64	30	1920	4096	900
	80	32	2560	6400	1024
Total	296	129	7312	17216	3133

$$b = \frac{cov(x,\ y)}{s_x^2} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i \cdot y_i - \overline{x} \cdot \overline{y}}{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}} = \frac{7312/6 - 296/6 \cdot 129/6}{17216/6 - (296/6)^2} = 0{,}363$$

$$r = \frac{s_x}{s_y} \cdot b = \frac{\sqrt{17216/6 - (296/6)^2}}{\sqrt{3133/6 - (129/6)^2}} \cdot 0{,}363 = 0,9787 \rightarrow r^2 = 0{,}957$$

b. Obtener las rectas de regresión mínimo cuadrático.

Si expresamos como $\hat{y} = a + b \cdot x$ los coeficientes se calculan como:

$$b = \frac{cov(x,y)}{s_x^2} = 0,363; \quad a = \overline{y} - b \cdot \overline{x}$$

$$a = \frac{129}{6} - 0{,}363 \cdot \frac{296}{6} = 3{,}592 \rightarrow \hat{y} = 3{,}592 + 0{,}363 \cdot x$$

c. ¿Qué volumen de ventas de la empresa se podría esperar en un año que se gaste de publicidad 60€?¿Y para un gasto en publicidad de 200€?

Para realizar la predicción del volumen de ventas utilizamos la recta de regresión que tienen las ventas en función de los gastos en publicidad. Para un gasto en publicidad de 60€ obtendremos un volumen de ventas de:

$$\hat{y} = 3{,}592 + 0{,}363 \cdot x = 3{,}592 + 0{,}363 \cdot 60 = 25,372$$

Si el gasto es de 200€ no podemos utilizar la recta de regresión puesto que el valor 200 esta fuera del recorrido del gasto en publicidad.

10.2. Combinatoria

1. Se distribuyen tres regalos distintos entre cinco chicos. De cuántas formas pueden hacerlo si:

Tenemos un total de 5 personas y 3 regalos a repartir, por lo que el valor de n es de 5 y el valor de m es 3.

a. Cada persona sólo puede recibir un regalo.

Para identificar de que tipo de combinatoria se trata hay que tener en cuenta los siguientes aspectos:

- Importa el orden.
- No mencionan que se usen todos los elementos.
- Sólo puede recibir un regalo cada persona, por lo tanto no se repiten los elementos.

$$V_n^m = \frac{n!}{(n-m)!} \rightarrow V_5^3 = \frac{5!}{(5-3)!} = 60$$

b. A cada persona le puede tocar más de un regalo.

- Importa el orden.
- No mencionan que se usen todos los elementos.
- A cada persona le puede tocar más de un regalo, por lo tanto los elementos se pueden repetir.

$$VR_n^m = n^m \rightarrow VR_5^3 = 5^3 = 125$$

c. Cada persona sólo puede recibir un regalo pero los tres son idénticos.

- El orden no importa, ya que los elementos son idénticos.
- No mencionan que se usen todos los elementos.
- Sólo puede recibir un regalo cada persona, por lo tanto no se repiten los elementos.

$$C_n^m = \frac{n!}{(n-m)!\,m!} \rightarrow C_5^3 = \frac{5!}{(5-3)!\,3!} = 10$$

2. Con las cifras impares, ¿cuántos números de tres cifras se pueden formar pudiéndose repetir cifras?

Las cifras impares son 1, 3, 5, 7 y 9, y queremos agruparlas de tres en tres. Como influye el orden y las cifras se pueden repetir se trata de variaciones con repetición. Por lo tanto el resultado es el siguiente:

$$VR_5^3 = 5^3 = 125$$

Es decir, que hay 125 posibilidades.

3. Belén necesita seleccionar 4 personas, entre los 20 candidatos que tiene, para formar su equipo de trabajo. ¿De cuántas maneras puede hacer la selección?

El orden en que se haga la selección no influye:

$$C_{20}^4 = \frac{20!}{(20-4)!\,4!} = 4845$$

Tiene 4845 formas distintas de hacer la selección.

4. Ocho ciclistas van por el carril bici en fila. ¿De cuántas formas pueden ir ordenados?

El orden en la fila influye y se utilizan todos los elementos. Por lo tanto, se trata de una permutación:

$$P_8 = 8! = 40320$$

Se pueden colocar de 40320 formas distintas.

5. **En el palo de señales de un barco se pueden izar tres banderas rojas, dos azules y cuatro verdes. ¿Cuántas señales distintas pueden indicarse con la colocacióon de las nueve banderas?**

   Aquí utilizamos todos los elementos, importa el orden y hay elementos repetidos un cirto número de veces. Es decir, se trata de una permutacióon con repetición:

$$PR_9^{3,\,2,\,4} = \frac{9!}{3!\;2!\;4!} = 1260 \text{ posibilidades}$$

6. **Un número telefónico consta de siete cifras enteras. Supongamos que:**

   - **La primera cifra debe ser un númeor entre 2 y 9, ambos inclusive.**
   - **La segunda y la tercera cifra deben ser números entre 1 y 9, ambos inclusive.**
   - **Cada una de las restantes cifras es un número entre 0 y 9, ambos inclusive.**

   Por lo tanto, tenemos que para cada condición que se ha puesto se tratan de variaciones con repetición, debido a que influye el orden, no se utilizan todos los elementos y se pueden repetir. La fórmula de las variaciones con repetición es la siguiente: $V_n^m = n^m$.

   - Para la primera cifra tenemos 8 casos posibles, ya que:

$$V_8^1 = 8^1 = 8$$

   - Para la segunda y la tercera juntas tenemos:

$$V_9^2 = 9^2 = 81$$

   - En las cifras siguientes:

$$V_{10}^4 = 10^4 = 10000$$

   En consecuencia, el número de teléfonos es: $8 \cdot 81 \cdot 10000 = 6480000$.

7. **En una fábrica, hay 5 tipos diferentes de piezas que se pueden utilizar para ensamblar un producto: piezas A, B, C, D y E. Un ingeniero de producción necesita seleccionar un conjunto de 8 piezas para construir un prototipo, pero no es necesario que todas las piezas sean diferentes. Además, hay una disponibilidad ilimitada**

de cada tipo de pieza.

El orden en que se haga la selección no influye y se pueden repetir las piezas, por lo tanto se trata de combinaciones con repetición.

$$\mathrm{CR}_5^8 = \binom{5+8-1}{8} = \frac{(5+8-1)!}{(5-1)!\ 8!} = 495$$

Hay 495 maneras de seleccionar un conjunto de 8 piezas de los 5 tipos disponibles cuando se permite la repetición.

10.3. Probabilidad

1. **En una academia de artes escénicas se imparten clases de danza y teatro. De danza, hay modalidad de danza clásica y cabaret. En la academia, un 17 % de individuos practica danza clásica, un 45 % cabaret y un 5 % ambas modalidades. Si elegimos un individuo que asiste a dicha academia:**

   Sean los sucesos: A = "practicar danza clásica", B = "practicar cabaret" y T = "practicar teatro".

   a. **Calcula la probabilidad de que practique algún tipo de danza (o las dos).**

   $$p(A \cup B) = p(A) + p(B) - p(A \cap B) = 0,17 + 0,45 - 0,05 = 0,57$$

   b. **Calcula la probabilidad de que practique solamente teatro.**

   $$p(T) = p(\overline{A \cup B}) = 1 - 0,57 = 0,43$$

2. **En un departamento de calidad se analiza el funcionamiento del software del motor de vehículos eléctricos e híbridos. Se revisaron 85 coches eléctricos y 145 coches híbridos. En total, 43 coches tenían errores en el software de sus motores. Además, de los motores con software defectuoso, 12 correspondían a coches eléctricos.**

   Organizamos los datos en una tabla de doble entrada.

	Eléctricos (E)	Híbridos (H)	TOTALES
Defectuoso (D)	12	31	43
No defectuoso ($\overline{D}$)	73	114	187
TOTALES	85	145	230

a. **Calcule la probabilidad de que un coche revisado seleccionado al azar, sea híbrido y presente el software de su motor correcto.**

Del total de 230 coches revisados, hay 114 híbridos con el software del motor correcto:

$$p(H \cap \overline{D}) = \frac{114}{230} = 0,4957$$

b. **Calcule la probabilidad de que un coche híbrido seleccionado al azar tenga defectuoso el software del motor.**

Entre los 145 coches híbridos hay 31 con el software del motor defectuoso:

$$p(D|H) = \frac{31}{145} = 0,2138$$

3. Un estudiante hace dos pruebas en un mismo día. La probabilidad de que pase la primera prueba es de 0.6; la de que pase la segunda es de 0.8, y la de que pase ambas es de 0.5. Halla las siguientes probabilidades:

Sean los sucesos:

- A = "pasar la primera prueba" $\rightarrow$ P(A) = 0.6
- B = "pasar la segunda prueba" $\rightarrow$ P(B) = 0.8
- A $\cap$ B = "pasar las pruebas A y B" $\rightarrow$ P(A $\cap$ B) = 0.5

a. **Que pase al menos una prueba.**

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0,6 + 0,8 - 0,5 = 0,9$$

b. **Que no pase ninguna prueba.**

$$P(\overline{A} \cap \overline{B}) = P(\overline{A \cup B}) = 1 - P(A \cup B) = 1 - 0,9 = 0,1$$

c. **¿Son las pruebas sucesos independientes?**

Para que A y B sean sucesos independientes se tiene que cumplir lo siguiente: $P(A \cap B) = P(A) \cdot P(B)$.

$$P(A \cap B) = 0,5 \neq 0,6 \cdot 0,8 = 0,48 = P(A) \cdot P(B)$$

Por lo tanto, los sucesos A y B no son sucesos independientes.

d. **Que pase la segunda prueba en caso de no haber superado la primera.**

$$P(B|\overline{A}) = \frac{P(B \cap \overline{A})}{P(\overline{A})} = \frac{0{,}3}{0{,}4} = 0{,}75$$

Siendo $B = (A \cap B) \cup (\overline{A} \cap B) \rightarrow P(B) = P(A \cap B) + P(\overline{A} \cap B) \rightarrow 0{,}8 = 0{,}5 + P(\overline{A} \cap B) \rightarrow P(\overline{A} \cap B) = 0{,}3$

4. **Un dado con las caras numeradas del 1 al 6 está trucado de modo que la probabilidad de obtener un número es directamente proporcional a dicho número.**

$$P(1) = x \quad P(2) = 2x \quad P(3) = 3x \quad P(4) = 4x \quad P(5) = 5x \quad P(6) = 6x$$

$$P(\Omega) = 1 = x + 2x + 3x + 4x + 5x + 6x = 21x \rightarrow x = \frac{1}{21}$$

a. **Halla la probabilidad de que salga 3 si se sabe que salió impar.**

$$P(3|\text{impar}) = \frac{P(3)}{P(1,3,5)} = \frac{\frac{3}{21}}{\frac{1}{21} + \frac{3}{21} + \frac{5}{21}} = \frac{3}{9} = \frac{1}{3}$$

b. **Calcula la probabilidad de que salga par si se sabe que salió mayor que 3.**

$$P(\text{par}|\text{mayor que } 3) = \frac{P(4,6)}{P(4,5,6)} = \frac{\frac{4}{21} + \frac{6}{21}}{\frac{4}{21} + \frac{5}{21} + \frac{6}{21}} = \frac{10}{15} = \frac{2}{3}$$

5. **En una planta de producción de microchips, se producen tres tipos de chips: Tipo A, Tipo B y Tipo C. Las probabilidades de que un chip producido sea de cada tipo son:**

- **Tipo A:** $P(A) = 0,5$
- **Tipo B:** $P(B) = 0,3$
- **Tipo C:** $P(C) = 0,2$

Además, existe una probabilidad de que cada tipo de chip presente presente defectos:

- **Si el chip es del Tipo A, la probabilidad de que esté defectuoso es** $P(D|A) = 0,02$.
- **Si el chip es del Tipo B, la probabilidad de que esté defectuoso es** $P(D|B) = 0,05$.
- **Si el chip es del Tipo C, la probabilidad de que esté defectuoso es** $P(D|C) = 0,1$.

Responde las siguientes preguntas:

a. ¿Cuál es la probabilidad de que un chip seleccionado al azar esté defectuoso?

Utilizamos la Ley de la Probabilidad Total para calcular $P(D)$, la probabilidad de que un chip esté defectuoso:

$$P(D) = P(D|A) \cdot P(A) + P(D|B) \cdot P(B) + P(D|C) \cdot P(C) =$$

$$= 0,02 \cdot 0,5 + 0,05 \cdot 0,3 + 0,1 \cdot 0,2 = 0,045$$

Entonces, las probabilidad de que un chip esté defectuoso es 0,045 o 4,5 %.

b. Si se selecciona un chip defectuoso, ¿cuál es la probabilidad de que sea de cada tipo de chip?

Usamos la probabilidad condicional para calcular la probabilidad de que un chip defectuoso sea de cada tipo, es decir, $P(A|D), P(B|D)$, y $P(C|D)$.

$$P(A|D) = \frac{P(D|A) \cdot P(A)}{P(D)} = \frac{0,02 \cdot 0,5}{0,045} = 0,222$$

$$P(B|D) = \frac{P(D|B) \cdot P(B)}{P(D)} = \frac{0,05 \cdot 0,3}{0,045} = 0,333$$

$$P(C|D) = \frac{P(D|C) \cdot P(C)}{P(D)} = \frac{0,1 \cdot 0,2}{0,045} = 0,444$$

Por lo tanto, las probabilidades de que un chip defectuoso sea de cada tipo son aproximadamente:

- Tipo A: 22,2 %.
- Tipo B: 33,3 %.
- Tipo C: 44,4 %.

10.4. Distribución de Variables Aleatorias Discretas

1. **Una fábrica produce piezas electrónicas en lotes de 100 unidades. La probabilidad de que una pieza sea defectuosa es de 0,1 (10 %). Se selecciona un lote y se revisan 5 piezas al azar, donde la variable aleatoria X representa el número de piezas defectuosos en la muestra seleccionada**.

   La función de probabilidad de X está dada por la siguiente distribución binomial:

   $$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}; \ \sum P(X = x) = 1$$

   donde n = 5 y p = 0,1.

   Dado que X sigue una distribución binomial $X \sim \text{Binomial}(n = 5, p = 0,1)$, podemos calcular sus parámetros utilizanod las fórmulas para la distribución binomial.

   a. **Calcula la media $E(X)$ del número de piezas defectuosas en la muestra.**

   $$E(X) = n \cdot p = 5 \cdot 0,1 = 0,5$$

   b. **Calcula la varianza $Var(X)$ del número de piezas defectuosas en la muestra.**

   $$Var(X) = n \cdot p \cdot (1-p) = 5 \cdot 0,1 \cdot (1-0,1) = 0,45$$

   c. **Calcula la desviación estándar σ del número de piezas defectuosasa en la muestra.**

   $$\sigma = \sqrt{Var(X)} = \sqrt{0,45} = 0,67$$

2. **De los huevos que se producen diariamente en una granja, deben desecharse el 20 % por no ser aptos para su consumo. Se seleccionan de manera aleatoria e independiente 5 huevos:**

   Se trata de una distribución binomial. Los sucesos son: A = "el huevo no es apto para el consumo" y B = "el huevo es apto para el consumo". La probabilidad de que un huevo sea no apto para el consumo es p(A) = p = 0,20 y la probabilidad de su contrario (es apto para el consumo) es p(B) = q = 0,80. Como se eligen al azar 5 huevos, estamos ante una distribución binomial B(5; 0,20).

a. **Calcula la probabilidad de que tengamos que desechar alguno de los huevos seleccionados (al menos 1).**

El suceso "desechar alguno de los huevos" es el suceso contrario a "todos son aptos para el consumo".

$$\text{En la binomial B(5; 0,80): } p[x=5] = \binom{5}{5} \cdot 0,80^5 \cdot 0,20^0 = 0,32768$$

$$\text{luego: } p = 1 - 0,32768 = 0,67232$$

b.
- **¿Qué es más probable, que haya exactamente 2 huevos no aptos, o que haya exactamente 3 huevos no aptos? Obtén estas probabilidades.**

$$p[x=2] = \binom{5}{2} \cdot 0,20^2 \cdot 0,80^3 = \frac{5!}{2! \cdot 3!} \cdot 0,20^2 \cdot 0,80^3 = 0,2048$$

$$p[x=3] = \binom{5}{3} \cdot 0,20^3 \cdot 0,80^2 = \frac{5!}{3! \cdot 2!} \cdot 0,20^3 \cdot 0,80^2 = 0,0512$$

Luego es más probable que haya exactamente 2 huevos no aptos para el consumo.

- **¿Cómo razonarías la respuesta a la pregunta anterior son hacer uso de la calculadora?**

Al sacar al zar un huevo es menor la probabilidad de que no sea apto para el consumo que la probabilidad de que sí lo sea. Es más probable entonces sacar dis huevos no aptos para el consumo que sacar tres.

3. Si la probabilidad de que ocurra un suceso A es 1/5.

a. **¿Cuál es el mínimo de veces que hay que repetir el experimento para que la probabilidad de que ocurra al menos una vez el suceso A sea mayor que 1/2?**

$$\text{X = "número de éxitos en n pruebas", X} \sim \text{B(n, 0.2), p = 0.2, q = 0.8}$$

$$\text{P(X} \geq 1) > \frac{1}{2} \rightarrow \text{P(X} < 1) \leq \frac{1}{2} \rightarrow \text{P(X} = 0) \leq \frac{1}{2}$$

$$P(X=0) = \binom{n}{0} \cdot 0{,}2^0 \cdot 0{,}8^n = 0{,}8^n \leq \frac{1}{2} \rightarrow (0{,}8^4 = 0{,}4096) \rightarrow n = 4$$

Para que el suceso A tenga al menos una probabilidad mayor a 1/2, hay que repetir el proceso un mínimo de 4 veces.

b. **¿Cuál es la probabilidad de que ocurra al menos dos veces A al realizar 5 veces el experimento?**

Se trata de una distribución binomial B(5; 0.2)

$$P(X \geq 2) = 1 - [P(X = 0) + P(X = 1)] = 1 - \left[\binom{5}{0} \cdot 0{,}2^0 \cdot 0{,}8^5 + \right.$$

$$\left. + \binom{5}{1} \cdot 0{,}2^1 \cdot 0{,}8^4 \right] = 1 - (0{,}3277 + 0{,}4096) = 0{,}2627$$

4. **En una encuesta sobre el uso de transportes, 50 personas eligen su modo de transporte favorito entre 4 opciones: automóvil, autobús, bicicleta y tren. La probabilidad de que una persona elija cada modo es de 0.4, 0.3, 0.2 y 0.1, respectivamente. ¿Cuál es la probabilidad de que, entre las 50 personas encuestadas, 20 prefieran el automóvil, 15 el autobús, 10 la bicicleta y 5 el tren?**

Este es un caso de distribución multinomial.

$$P(X_1 = 20, X_2 = 15, X_3 = 10, X_4 = 5) =$$

$$= \frac{50!}{20! \cdot 15! \cdot 10! \cdot 5!} \cdot 0{,}4^{20} \cdot 0{,}3^{15} \cdot 0{,}2^{10} \cdot 0{,}1^5 = 0{,}00355$$

5. **Un ingeniero prueba una máquina que tiene una probabilidad de 0.2 de fallar en cada intento. Define la variable aleatoria X como el núméor de intentos hasta observar la primera falla.**

Se trata de una distribución geométrica con parámetro $p = 0{,}2$.

a. **¿Cuál es la probabilidad de que la primera falla ocurra en el tercer intento?**

$$P(X = 3) = (1 - 0{,}2)^{3-1} \cdot 0{,}2 = 0{,}8^2 \cdot 0{,}2 = 0{,}128$$

Entonces, la probabilidad de que la primera falla ocurra en el tercer intento es 12.8 %.

b. **¿Cuál es la probabilidad de que la primera falla ocurra en o antes del quinto intento?**

$$P(X \leq 5) = P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5) =$$

$$= (1-0{,}2)^{1-1} \cdot 0{,}2 + (1-0{,}2)^{2-1} \cdot 0{,}2 + (1-0{,}2)^{3-1} \cdot 0{,}2 + (1-0{,}2)^{4-1} \cdot 0{,}2 +$$

$$+ (1 - 0{,}2)^{5-1} \cdot 0{,}2 = 0{,}67232$$

6. Una tienda de artículos eléctricos tiene 20 planchas, de las cuales 5 son amarillas. Si se extraen aleatoriamente y sin sustitución 10 planchas ¿Cuál es la probabilidad de que dos de ellas sean amarillas?

Como las planchas se extraen aleatoriamente y luego no tienen sustitución (reemplazo) se trata de una distribución hipergeométrica. La probabilidad de que una plancha sea amarilla es de: $p = \frac{5}{20} = 0{,}25$.

$$P(X = 2) = \frac{\binom{20 \cdot 0{,}25}{2}\binom{20 \cdot (1 - 0{,}25)}{10 - 2}}{\binom{20}{10}} = 0{,}3483$$

7. Sí la probabilidad de que un cierto dispositivo de medición muestre una desviación excesiva es de 0.05, ¿cuál es la probabilidad de que:?

a. El sexto de estos dispositivos de medición sometidos a prueba sea el tercero en mostrar una desviación excesiva.

Sea:

- x = 6 dispositivos de medición.
- r = 3 dispositivos que muestran desviación excesiva.
- p = p(dispositivo muestre una desviación excesiva) = 0.05
- q = p(dispositivo no muestre una desviación excesiva) = 0.95

$$P(X = 6) = \binom{6 - 1}{3 - 1}(1 - 0{,}05)^{6-3}0{,}05^3 = 0{,}001072$$

b. El séptimo de estos dispositivos de medición sometidos a prueba, sea el cuarto que no muestre una desviación excesiva.

Sea:

- x = 7 dispositivos de medición.
- r = 4 dispositivos que muestran desviación excesiva.
- p = p(dispositivo no muestre una desviación excesiva) = 0.95
- q = p(dispositivo muestre una desviación excesiva) = 0.05

$$P(X = 7) = \binom{7 - 1}{4 - 1}(1 - 0{,}95)^{7-4}0{,}95^4 = 0{,}002036$$

8. **La probabilidad de que en una mascletà (tipo de fuegos artificiales que estallan con gran estruendo) en fallas una persona se desmaye es de 0.001. Considerando que acuden unas 5000 personas a ver la mascletà el días de San José, ¿cuál es la probabilidad de que se desmayen 25 personas?**

   Se trata de una distribución de Poisson debido a que la probabilidad se aproxima al 0 y que el tamaño de la muestra se puede aproximar a ∞. Es por ello que $m = 5000 \cdot 0{,}001 = 50$.

$$P(X = 25) = \frac{50^{25} \cdot e^{-50}}{25!} = 3{,}70 \cdot 10^{-5}$$

10.5. Distribución de Variables Aleatorias Continuas

1. **Sea X una variable aleatoria continua con la siguiente función de densidad:**

$$f(x) \begin{cases} 3 \cdot x^2 & \text{si } 0 \leq \text{x} \leq 1 \\ 0 & \text{en otro caso} \end{cases}$$

   a. **Verifica que $f(x)$ es una unción de densidad de probabilidad.**

   Para que $f(x)$ sea una función de densidad debe cumplir:

   - $f(x) \geq 0$ para todo x.
   - La integral de $f(x)$ sobre todos los valores posibles de X debe ser 1.

$$\int_{-\infty}^{\infty} f(x)dx = \int_0^1 3 \cdot x^2 dx = 3 \cdot \int_0^1 x^2 dx = 3 \cdot \left[\frac{x^3}{3}\right]_0^1 = 3 \cdot \left(\frac{1^3}{3} - \frac{0^3}{3}\right) = 1$$

   Como la integral es 1, $f(x)$ es una función de densidad válida.

   b. **Calcula la función de distribución $F(x)$ para X.**

   La función de distribución $F(x)$ se define como:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\, dt$$

   Como $f(x) = 0$ para $x < 0$, tenemos:

$$F(x)\begin{cases} 0 & \text{si } x < 0, \\ \int_0^x 3\cdot t^2\ dt & \text{si } 0 \leq \text{x} \leq 1, \\ 1 & \text{si } x > 1. \end{cases}$$

Calculamos $\int_0^x 3\cdot t^2\ dt$ para $0 \leq x \leq 1$:

$$\int_0^x 3\cdot t^2\ dt = 3\cdot\int_0^x t^2\ dt = 3\cdot\left[\frac{t^3}{3}\right]_0^x = x^3$$

Por lo tanto,

$$F(x)\begin{cases} 0 & \text{si } x < 0, \\ x^3 & \text{si } 0 \leq \text{x} \leq 1, \\ 1 & \text{si } x > 1. \end{cases}$$

c. Encuentra la probabilidad de que X tome valores entre 0,2 y 0,8.

Queremos calcualr $P(0,2 \leq X \leq 0,8)$, que es:

$$P(0,2 \leq X \leq 0,8) = F(0,8) - F(0,2)$$

Usamos $F(x) = x^3$ en el intervalo $0 \leq x \leq 1$:

$$F(0,8) = 0,8^3 = 0{,}512, \quad F(0{,}2) = 0{,}2^3 = 0{,}008$$

Así,

$$P(0{,}2 \leq X \leq 0{,}8) = F(0{,}8) - F(0{,}2) = 0{,}512 - 0{,}008 = 0{,}504$$

2. El peso de los recién nacidos de una localidad, sigue una distribución normal de media 3300 gramos y desviación típica 465 gramos. Un recién nacido tiene bajo peso si su peso es inferior a 2500 gramos.

a. ¿Cuál es la probabilidad de que un recién nacido en esta localidad tenga bajo peso?

Los datos de que disponemos pertenecen a una distribución normal de media $\mu = 3300$ g y desviación típica $\sigma = 465$ gr., es decir una normal N(3300, 465).

Para hacer uso de la tabla correspondiente a la normal N(0, 1) debemos tipificar las variables $x \in N(3300,\ 465)$ convirtiéndolas en variables $z \in N(0,\ 1)$:

$$P[x \leq 2500] = P\left[z \leq \frac{2500 - 3300}{465}\right] = P[z \leq -1,72] = P[z \geq -1,72] =$$

$$= 1 - P[z \leq 1,72] = 1 - 0,9573 = 0,0427$$

b. ¿Cuál es la probabilidad de que un recién nacido en esta localidad tenga un peso entre 3500 y 4000 gramos?

Las variables pertenecen a la distribución normal N(3300, 465) por lo que debemos tipificarlas:

$$P[3500 \leq x \leq 400] = P\left[\frac{3500 - 3300}{465} \leq z \leq \frac{4000 - 3300}{465}\right] =$$

$$= P[0,43 \leq z \leq 1,51] = P[z \leq 1,51] - P[z \leq 0,43] = 0,9345 - 0,6664 = 0,2681$$

3. La cantidad de hierro en suero de una mujer adulta sigue una distribución normal de media 120 $\mu g/dl$ y desviación típica 30 $\mu g/dl$. Se considera que una mujer tiene un tipo de anemia por falta de hierro si su cantidad de hierro no llega a 75 $\mu g/dl$.

a. ¿Cuál es la probabilidad de que una mujer adulta tenga anemia por falta de hierro?

Los datos de que disponemos pertenecen a una distribución normal de media $\mu = 120\ \mu g/dl$ y desviación típica $\sigma = 30\ \mu g/dl$, es decir una normal N(120, 30). Para hacer uso de la tabla correspondiente a la normal N(0, 1) debemos tipificar las variables $x \in N(120,\ 30)$ convirtiéndolas en variables $z \in N(0,\ 1)$.

$$P[x \leq 75] = P\left[z \leq \frac{75 - 120}{30}\right] = P[z \leq -1,5] = P[z \geq 1,5] =$$

$$= 1 - P[z \leq 1,5] = 1 - 0,9332 = 0,0668$$

b. El 45 % de las mujeres adultas tienen una cantidad de hierro en suero superior a k. Averigüe el valor de k.

Las variables pertenecen a la distribución normal N(120, 30) por lo que debemos tipificarlas:

$$P[x > k] = P\left[\frac{x - 120}{30} > \frac{k - 120}{30}\right] = P\left[z > \frac{k - 120}{30}\right] = 0,45 \rightarrow$$

$$\rightarrow 1 - P\left[z < \frac{k-120}{30}\right] = 0,45 \rightarrow P\left[z < \frac{k-120}{30}\right] = 0,55 \rightarrow$$
$$\rightarrow \frac{k-120}{30} = 0,125 \rightarrow k = 123,75\ \mu g/dl$$

4. La estatura de los estudiantes en una universidad sigue una distribución normal con media $\mu = 170$ cm y desviación estándar $\sigma = 10$ cm.

a. Calcula la probabilidad de que un estudiante tenga una estatura menor a 160 cm.

$$P[X < 160] = P\left[Z < \frac{160-170}{10}\right] = P[Z < -1] = P[Z > 1] =$$
$$= 1 - P[Z < 1] = 1 - 0{,}8413 = 0{,}1587$$

La probabilidad de que un estudiante tenga una estatura menor a 160 cm es de un 15.87 %.

b. ¿Cuál es la probabilidad de que un estudiante tenga una estatura entre 165 cm y 185 cm?

$$P[165 < X < 185] = P\left[\frac{165-170}{10} < Z < \frac{185-170}{10}\right] = P[-0{,}5 < Z < 1{,}5] =$$
$$= P[Z < 1{,}5] - P[Z < -0{,}5] = P[Z < 1{,}5] - P[Z > 0{,}5] = P[Z < 1{,}5] -$$
$$-(1 - P[Z < 0{,}5]) = 0{,}9332 - (1 - 0{,}6915) = 0{,}6247$$

La probabilidad de que un estudiante tenga una estatura entre 165 cm y 185 cm es de 62,47 %.

c. Encuentra la estatura que corresponde al percentil 90.

Para el percentil 90, buscamos el valor X tal que $P[X < x_{90}] = 0{,}90$.

Primero, encontramos el valor Z correspondiente al percentil 90 en la distribución normal estándar:

$$Z_{0,90} \approx 1{,}28$$

Luego, transformamos este valor Z a la escala de distribución $N(170, 10)$:

$$Z = \frac{X - \mu}{\sigma} \rightarrow x_{90} = \mu + Z_{0,90} \cdot \sigma = 170 + 1{,}28 \cdot 10 = 182{,}8$$

La estatura que corresponde al percentil 90 es aproximadamente 182.8 cm.

5. **En la distribución Chi-Cuadrado con 20 grados de libertad, ¿cuánto es el percentil 10?**

$$\chi^2_{0{,}10;\,20} = P(\chi^5_{20} \geq 0{,}10) = P(\chi^5_{20} \leq (1 - 0{,}10)) = P(\chi^5_{20} \leq 0{,}90) = 12{,}443$$

6. **En la distribución F de Snedecor con 10 grados de libertad en el numerador y 10 en el denominador, ¿que percentil le corresponde al valor 2.3226?**

   Vemos en la tabla VII de la *Distribución F de Snedecor* para $n_1 = 10$ y $n_2 = 10$, que el valor 3.3226 corresponde a $\alpha = 0.10$. Esto quiere decir, que la variable deja por debajo el 90 % de las observaciones, por lo que le corresponde el percentil 90 (P_{90}).

10.6. Muestreo y Estimación

1. **Una empresa desea estudiar el tiempo promedio en (horas) que sus empleados dedican a tareas de formación cada mes. La empresa cuenta con un total de 200 empleados. Se toma una muestra aleatoia simple de 30 empleados y se obtienen los siguientes datos de tiempo de formación en horas:**

   4.5, 6.3, 3.2, 5.1, 7.5, 4.8, 6.0, 5.5, 3.9, 4.2, 5.6, 4.3, 6.1, 5.8, 4.0, ,7.0, 4.7, 3.6, 5.2, 4.4, 6.5, 4.1, 5.3, 3.7, 6.4, 5.7, 4.9, 3.8, 5.4, 6.2

   a. **Calcula la media muestral del tiempo de formación.**

$$\overline{x} = \frac{\sum x}{n} = 4{,}81$$

   b. **Calcula la varianza muestral y la desviación muestral.**

$$s^2 = \frac{\sum (x - \overline{x})^2}{n - 1} \approx 1{,}394 \rightarrow s \approx 1{,}18$$

   c. **Utilizando los datos obtenidos, estima la media poblacional de tiempo de formación.**

$$\mu_{\overline{x}} = \overline{x} = 4{,}81$$

   d. **¿Cuál sería la varianza de las medias muestrales si se tomaran muestras de 30 empleados de manera repetida?**

$$\sigma^2_{\overline{x}} = \frac{s^2}{n} \cdot \frac{N - n}{N - 1} = \frac{1{,}394}{30} \cdot \frac{200 - 30}{200 - 1} \approx 0{,}0429$$

2. **En una plantación, se cuenta con 1000 árboles frutales y se quiere estimar la cantidad total de frutos producidos. Se selecciona una muestra aleatoria simple de 50 árboles y se contabilizan los frutos en cada uno de ellos, obteniendo un promedio de 120 frutos por árbol y una desviación estándar de 15 frutos. Calcula la suma estimada de frutos producidos en toda la plantación.**

$$\mu = \overline{x} \cdot n = 120 \cdot 1000 = 120000$$

3. **Una fábrica de dulces produce barras de chocolate que deben pesar 100 gramos. Se toma una muestra aleatoria de 50 barras, y se encuentra un peso promedio de 101 gramos. Se sabe que la varianza poblacional es de 4 gramos^2. Encuentra el intervalo de confianza del 95 % para el peso promedio de las barras de chocolate.**

   - Tamaño de la muestra (n) = 50
   - Media muestral ($\overline{x}$) = 101 gramos
   - Varianza poblacional (σ^2) = 4 gramos^2 $\rightarrow$ σ = 2 gramos
   - Nivel de confianza = 95 %

   Con los datos dados, se trata de un intervalo de confianza para la media de una población normal, teniendo la varianza poblacional conocida y una muestra grande. Para un 95 % de confianza, α = 1 - 0,95 = 0,05.

$$T_\mu = \left[\overline{x} - z_{0,05/2} \cdot \frac{\sigma}{\sqrt{n}},\ \overline{x} + z_{0,05/2} \cdot \frac{\sigma}{\sqrt{n}}\right] =$$

$$= \left[101 - 1{,}96 \cdot \frac{2}{\sqrt{50}},\ 101 + 1{,}96 \cdot \frac{2}{\sqrt{50}}\right] = [100{,}446, 101{,}554]$$

   Con un 95 % de confianza, podemos decir que el peso promedio de las barras de chocolate está entre 100,45 gramos y 101,55 gramos.

4. **Un estudio de hábitos de lectura mide el número de libros leídos al mes por los estudiantes de una universidad. Se toma una muestra de 100 estudiantes, obteniéndose una media de 5 libros leídos y una desviación estándar muestral de 1,5 libros. Calcula el intervalo de confianza del 90 % para el número promedio de libros leídos por los estudiantes al mes.**

   - Tamaño de la muestra (n) = 100
   - Media muestral ($\overline{x}$) = 5 libros
   - Desviación muestral (s) = 1.5 libros

- Nivel de confianza = 90 %

Con los datos dados, se trata de un intervalo de confianza para la media de una población normal, teniendo la varianza poblacional desconocida y una muestra grande. Para un 90 % de confianza, α = 1 - 0,90 = 0,10.

$$T_\mu = \left[\overline{x} - z_{0,1/2} \cdot \frac{s}{\sqrt{n}},\ \overline{x} + z_{0,1/2} \cdot \frac{s}{\sqrt{n}}\right] =$$

$$= \left[5 - 1{,}645 \cdot \frac{1{,}5}{\sqrt{100}},\ 5 + 1{,}645 \cdot \frac{1{,}5}{\sqrt{100}}\right] = [4{,}753, 5{,}247]$$

Con un 90 % de confianza, podemos decir que el número promedio de los libros leídos por los estudiantes de esta universidad está entre 4.753 y 5.247 libros por mes.

5. Un investigador quiere estimar el tiempo promedio que un tipo específico de batería dura en una prueba de resistencia. Se prueban 9 baterías, encontrando un tiempo promedio de 18 horas con una desviación de 2 horas. Encuentra el intervalo de confianza del 95 % para la duración promedio de las baterías.

- Tamaño de la muestra (n) = 9
- Media muestral ($\overline{x}$) = 18 horas
- Desviación muestral (s) = 2 horas
- Nivel de confianza = 95 %

Con los datos dados, se trata de un intervalo de confianza para la media de una población normal, teniendo la varianza poblacional desconocida y una muestra pequeña. Para un 95 % de confianza, α = 1 - 0,95 = 0,05.

$$T_\mu = \left[\overline{x} - t_{0{,}05/2;\ 9-1} \cdot \frac{s}{\sqrt{n}},\ \overline{x} + t_{0{,}05/2;\ 9-1} \cdot \frac{s}{\sqrt{n}}\right] =$$

$$= \left[18 - 2{,}306 \cdot \frac{2}{\sqrt{18}},\ 18 + 2{,}306 \cdot \frac{2}{\sqrt{18}}\right] = [16{,}913, 19{,}087]$$

Con un 95 % de confianza, el tiempo promedio de duración de las baterías en la población se encuentran entre 16.913 y 19.087 horas.

6. Se estudian los salarios de trabajadores en dos ciudades. En una muestra de 200 trabajadores en la Ciudad A, el salario promedio es de 1200€ con una desviación estándar conocida de 100€. En una muestra de 150 trabajadores de la Ciudad B, el salario promedio es

de 1150€ con una desviación estándar conocida de 120€. Calcula el intervalo de confianza del 95 % para la diferencia en los salarios promedio entre ambas ciudades.

- Datos de la Ciudad A:
  - Tamaño de la muestra (n_A) = 200
  - Media muestral ($\overline{x}_A$) = 1200
  - Desviación poblacional (σ_A) = 100
- Datos de la Ciudad B:
  - Tamaño de la muestra (n_B) = 150
  - Media muestral ($\overline{x}_B$) = 1150
  - Desviación poblacional (σ_B) = 120

Con los datos dados, se trata de un intervalo de confianza para la diferencia de medias de dos poblaciones normales, teniendo las varianzas poblacionales conocidas y muestras grandes. Para un 95 % de confianza, α = 1 - 0,95 = 0,05.

$$T_{\mu_A-\mu_B} = \left[(\overline{x_A} - \overline{x_B}) \mp z_{0,05/2} \cdot \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_2^B}{n_B}}\right] =$$

$$= \left[(1200 - 1150) \mp 1{,}96 \cdot \sqrt{\frac{100^2}{200} + \frac{120^2}{150}}\right] = [26{,}33, 73{,}67]$$

Con un 95 % de confianza, la diferencia entre los salarios promedios entre ambas ciudades está entre 26,33€ y 73,67€.

7. En una investigación sobre el tiempo que los estudiantes dedican al ejercicio, se seleccionan muestras de estudiantes de dos universidades. En la universidad X, 100 estudiantes tienen un promedio de 6 horas de ejercicio a la semana, con una desviación estándar muestral de 1,8 horas. En la universidad Y, 120 estudiantes reportan un promedio de 5 horas, con una desviación estándar muestral de 2 horas. Calcula el intervalo de confianza del 90 % para la diferencia de tiempo dedicado al ejercicio entre ambas universidades.

- Datos de la Universidad X:
  - Tamaño de la muestra (n_X) = 100
  - Media muestral ($\overline{x}_X$) = 6 horas
  - Desviación muestral (s_X) = 1,8 horas
- Datos de la Universidad Y:
  - Tamaño de la muestra (n_Y) = 120

- Media muestral ($\overline{x}_Y$) = 5 horas
- Desviación muestral (s_Y) = 2 horas

Con los datos dados, se trata de un intervalo de confianza para la diferencia de medias de dos poblaciones normales, teniendo las varianzas poblacionales desconocidas y muestras grandes. Para un 90 % de confianza, α = 1 - 0,90 = 0,10.

$$T_{\mu_X-\mu_Y} = \left[(\overline{x_X} - \overline{x_Y}) \mp z_{0,10/2} \cdot \sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}\right] =$$

$$= \left[(6-5) \mp 1{,}645 \cdot \sqrt{\frac{1{,}8^2}{100} + \frac{2^2}{120}}\right] = [0{,}578, 1{,}422]$$

Con un 90 % de confianza, podemos afirmar que la diferencia en el tiempo promedio dedicado al ejercicio entre los estudiantes de la Universidad X y la Universidad Y se encuentra entre 0.578 y 1.422 horas. Esto sugiere que, en promedio, los estudiantes de la Universidad X dedican entre 0.578 y 1.422 horas a la semana al ejercicio que los estudiantes de la Universidad Y.

8. **Se comparan los tiempos de respuesta de dos servicios de atención al cliente. Se seleccionan muestras de 12 respuestas de cada servicio. En el servicio A, el tiempo promedio es de 2,5 minutos con una desviación estándar de 0,8 minutos. En el servicio B, el tiempo promedio es de 3 minutos con una desviación estándar de 0,7 minutos. Suponiendo que las varianzas son iguales en ambas poblaciones. Encuentra el intervalo de confianza del 95 % para la diferencia en el tiempo de respuesta promedio entre los dos servicios.**

   - Datos del servicio A:
     - Tamaño de la muestra (n_A) = 12
     - Media muestral ($\overline{x}_A$) = 2,5 minutos
     - Desviación muestral (s_A) = 0,8 minutos
   - Datos del servicio B:
     - Tamaño de la muestra (n_B) = 12
     - Media muestral ($\overline{x}_B$) = 3 minutos
     - Desviación muestral (s_B) = 0,7 minutos

   Con los datos dados, se trata de un intervalo de confianza para la diferencia de medias de dos poblaciones normales, teniendo las varianzas poblacionales

desconocidas pero iguales y muestras grandes. Para un 95 % de confianza, α = 1 - 0,95 = 0,05.

$$T_{\mu_A-\mu_B} = \left[(\overline{x_A} - \overline{x_B}) \mp t_{0,05/2;\ 12+12-2} \cdot \sqrt{\frac{(n_A-1)\cdot s_A^2 + (n_B-1)\cdot s_B^2}{n_A+n_B-2}} \cdot \sqrt{\frac{1}{n_A}+\frac{1}{n_B}}\right] =$$

$$= \left[(2{,}5-3) \mp 2{,}074 \cdot \sqrt{\frac{(12-1)\cdot 0{,}8^2 + (12-1)\cdot 0{,}7^2}{12+12-2}} \cdot \sqrt{\frac{1}{12}+\frac{1}{12}}\right] = [-1{,}137, 0{,}137]$$

Con un 95 % de confianza, la diferencia en los tiempos de respuesta promedio entre el Servicio A y el Servicio B está entre -1.137 y 0.137 minutos. Esto indica que, en promedio, el Servicio A podría ser hasta 1.137 minutos más rápido que el servicio B, o podría ser solo 0.137 minutos más lento que el servicio B. Como el intervalo incluye el valor 0, esto sugiere que no hay una diferencia estadísticamente significativa en los tiempos de respuesta entre ambos servicios.

9. Una empresa evalúa la efectividad de dos métodos de capacitación diferentes. Se toma una muestra de 10 empleados que usan el Método A, con un puntaje promedio de 85 en un examen posterior, y una desviación estándar de 5. Se toma otra muestra de 8 empleados que usan el Método B, con un punjuaje promedio de 80 y una desviación estándar de 6. Calcula el intervalo de confianza del 95 % para la diferencia en los puntajes promedio entre los dos métodos de capacitación.

- Datos del Método A:
  - Tamaño de la muestra (n_A) = 10
  - Media muestral ($\overline{x}_A$) = 85
  - Desviación muestral (s_A) = 5 minutos
- Datos del Método B:
  - Tamaño de la muestra (n_B) = 8
  - Media muestral ($\overline{x}_B$) = 80
  - Desviación muestral (s_B) = 6

Con los datos dados, se trata de un intervalo de confianza para la diferencia de medias de dos poblaciones normales, teniendo las varianzas poblacionales desconocidas y muestras pequeñas. Para un 95 % de confianza, α = 1 - 0,95 = 0,05.

$$T_{\mu_A-\mu_B} = \left[(\overline{x_A} - \overline{x_B}) \mp t_{0,05/2;\ m} \cdot \sqrt{\frac{s_A^2}{n_A}+\frac{s_B^2}{n_B}}\right]$$

$$c = \frac{\frac{s_A^2}{n_A}}{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}} = \frac{\frac{5^2}{10}}{\frac{5^2}{10} + \frac{6^2}{8}} = \frac{5}{14} \approx 0{,}3571$$

$$\frac{1}{m} = \frac{c^2}{n_A - 1} + \frac{(1-c)^2}{n_B - 1} \rightarrow \frac{1}{m} = \frac{0{,}3571^2}{10 - 1} + \frac{(1 - 0{,}3571)^2}{8 - 1} \rightarrow m = 13{,}65 \approx 13$$

$$T_{\mu_A - \mu_B} = \left[(85 - 80) \mp 2{,}160 \cdot \sqrt{\frac{5^2}{10} + \frac{6^2}{8}} \right] = [-0{,}715, 10{,}715]$$

Con un 95 % de confianza, podemos afirmar que la diferencia en los puntajes promedio entre los dos métodos de capacitación está entre -0.715 y 10.715 puntos. Esto indica que, aunque el puntaje promedio para el Método A parece ser más alto, el intervalo de confianza incluye valores negativos, lo que sugiere que no hay diferencia estadísticamente significativa entre los dos métodos.

10. Una máquina produce piezas con un diámetro nominal de 5 cm. Para verificar la precisión, se toma una muestra de 15 piezas y se mide el diámetro. La varianza muestral resultante es de 0.04 cm^2. Calcula el intervalo de confianza del 95 % para la varianza del diámetro de las piezas producidas por la máquina.

- Tamaño de la muestra (n) = 15
- Varianza muestral (s^2) = 0.04 cm^2
- Nivel de confianza = 95 %

Con los datos dados, se trata de un intervalo de confianza para la varianza de una población normal. Para un 95 % de confianza, α = 1 - 0,95 = 0,05.

$$T_{\sigma^2} = \left[\frac{(n-1) \cdot s^2}{\chi^2_{0{,}05/2;\, 15-1}}, \frac{(n-1) \cdot s^2}{\chi^2_{1-0{,}05/2;\, 15-1}} \right] =$$

$$= \left[\frac{(15-1) \cdot 0{,}04}{26{,}119}, \frac{(15-1) \cdot 0{,}04}{5{,}629} \right] = [0{,}0214, 0{,}0995]$$

Con un 95 % de confianza, podemos decir que la varianza del diámetro de las piezas producidas por la máquina se encuentra entre 0.0214 cm^2 y 0.0995 cm^2. Esto indica que, si tomáramos muchas muestras de 15 piezas cada una, el 95 % de los intervalos calculados de la misma manera contendrían la verdadera varianza del diámetro en la producción de esta máquina.

11. Se comparan las variabilidades en el tiempo de entrega de dos proveedores. En una muestra de 12 entregas del Proveedor A, la

varianza es de 16 minutos^2, mientras que en una muestra de 15 entregas del Proveedor B, la varianza es de 25 minutos^2. Calcula el intervalo de confianza del 90 % para la razón de varianzas del tiempo de entrega entre ambos proveedores.

- Datos del Proveedor A:
  - Tamaño de la muestra (n_A) = 12
  - Varianza muestral (s_A^2) = 16 minutos
- Datos del Proveedor B:
  - Tamaño de la muestra (n_B) = 15
  - Varianza muestral (s_B^2) = 25 minutos

Con los datos dados, se trata de un intervalo de confianza para la razón de varianzas de dos poblaciones normales. Para un 90 % de confianza, α = 1 - 0,90 = 0,10.

$$I = \left[\frac{s_A^2}{s_B^2} \cdot \frac{1}{F_{0,1/2;\ 12-1;\ 15-1}},\ \frac{s_A^2}{s_B^2} \cdot F_{0,1/2;\ 12-1;\ 15-1}\right] =$$

$$= \left[\frac{16}{25} \cdot \frac{1}{2{,}484},\ \frac{16}{25} \cdot 2{,}484\right] = [0{,}258, 1{,}589]$$

Con un 90 % de confianza, podemos decir que la razón de las varianzas del tiempo de entrega entre los dos proveedores se encuentra entre 0.2577 y 1.5898. Esto sugiere que, aunque las varianzas del tiempo de entrega de los dos proveedores pueden diferir, la variabilidad en los tiempos de entrega de un proveedor no es significativamente mayor que la del otro dentro de este intervalo de confianza.

10.7. Contrastes de Hipótesis

1. **El consumo de carne de pollo parece haberse disparado desde que hace unos meses cundió la alarma sobre otros tipos de carne. En cierta carnicería, las ventas diarias de carne de pollo sigue hasta entonces una normal de media 19 kilos y desviación típica 3 kilos. En una muestra de 35 días posteriores a la citada alarma se obtuvo una media de 21 kilos de carne de pollo vendidos al día. Suponiendo que las ventas siguen una distribución normal con la misma desviación típica:**

   a. **Plantea un test para contrastar que la venta de pollo no ha aumentado, como parecen indicar los datos. ¿A qué conclusión se llega a un nivel de significancia del 5 %?**

   Se trata de un contraste unilaterial para la media poblacional con varianza conocida. Se establecen las hipótesis:

   $$\text{Hipótesis nula } H_o\text{: } \mu \leq 19 \text{ kg}$$
   $$\text{Hipótesis alternativa } H_i\text{: } \mu > 19 \text{ kg}$$

   Se acepta H_o si el estadístico de contraste, z_o, se encuentra dentro de la región de aceptación $(-\infty;\ z_\alpha)$.

   Teniendo en cuenta que: $\overline{x} = 21$ kg, n = 35 días, $\sigma = 3$ kg y $\alpha = 0.05$. Tenemos que $z_\alpha = z_{0,05} = 1.645$.

   Siendo $z_o = \dfrac{\overline{x} - \mu}{\sigma/\sqrt{n}} = \dfrac{21 - 19}{3/\sqrt{35}} = 3{,}94$. Se rechaza la hipótesis nula, aceptando en consecuencia la hipótesis alternativa, afirmando que el consumo medio de carne de pollo ha aumentado con la alarma, con una fiabilidad del 95 %.

   b. **Calcula un intervalo de confianza del 95 % para la venta diaria media de carne de pollo.**

   El intervalo de confianza para la media poblacional de una distribución de varianza conocida:

   $$T_\mu = \left[\overline{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}},\ \overline{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right]$$

   donde $z_{\alpha/2} = z_{0,025} = 1{,}96$ con lo cual, $T_\mu = \left[21 \mp 1{,}96 \cdot \frac{3}{\sqrt{35}}\right] =$ [20,01; 21,99]. Se observa que $19 \notin$ [20,01; 21,99], se encuentra fuera del intervalo de confianza.

2. Un fabricante asegura que el diámetro de sus balones tiene una media de 22 cm con una desviación estándar de 2 cm. Se selecciona una muestra aleatoria de 100 balones y se observa una media de 21,6 cm. ¿Podemos concluir que el diámetro medio de los balones es diferente del valor especificado? Usa un nivel de significancia del 5 %.

Queremos probar si el diámetro promedio es diferente al valor especificado, por lo que tenemos una prueba bilateral.

$$\text{Hipótesis nula } H_o\text{: } \mu = 22$$
$$\text{Hipótesis alternativa } H_1\text{: } \mu \neq 22$$

Datos del ejercicio:

- Media poblacional hipotética (μ_0) = 22 cm
- Desviación estándar poblacional (μ) = 2 cm
- Tamaño de la muestra (n) = 100
- Media muestral ($\overline{x}$) = 21.6 cm
- El nivel de significancia dado es $\alpha = 0{,}05$.

Dado que conocemos la desviación estándar poblacional ($\sigma = 2$) y el tamaño de la muestra es grande ($n = 100$), usamos la prueba Z. El estadístico de prueba Z se calcula como:

$$Z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{21{,}6 - 22}{2/\sqrt{100}} = -2$$

Como se trata de una prueba bilateral, dividimos entre 2, lo que da $\alpha/2 = 0{,}025$ para cada cola de la distribución normal. Buscando en la tabla de distribución normal estándar, encontramos que los valores críticos para $\alpha/2 = 0.025$ es 1,96. Por lo tanto $z_0 \notin [-1{,}96, 1{,}96]$.

Como Z = -2 <-1,96 el valor cae en la región de rechazo. Rechazamos la hipótesis nula H_0 al nivel de significancia del 5 %. Esto indica que hay suficiente evidencia estadística para concluir que el diámetro promedio de los balones es diferente de 22 cm.

3. En una planta industrial, el tiempo promedio de fabricación de un producto se estima en 15 minutos. Una muestra de 81 observaciones muestra una media de 14,5 minutos y una desviación estándar muestral de 2,8 minutos. ¿Es el tiempo de fabricación significativamente menor? Usa un nivel de significancia del 1 %.

Dado que queremos saber si el tiempo de fabricación es menor que el valor especificado (15 minutos), tenemos una prueba unilateral a la izquierda.

$$\text{Hipótesis nula } H_o\text{: } \mu \geq 15$$
$$\text{Hipótesis alternativa } H_1\text{: } \mu < 15$$

Datos del ejercicio:

- Media poblacional hipotética (μ_0) = 15 minutos
- Desviación estándar muestral (s) = 2,8 minutos
- Tamaño de la muestra (n) = 81
- Media muestral ($\overline{x}$) = 14,5 minutos
- El nivel de significancia dado es $\alpha = 0.01$

Dado que no conocemos la desviación estándar poblacional y el tamaño de la muestra es grande ($n = 81$), podemos usar una prueba Z aproximada, utilizando la desviación estándar muestral. El estadístico de prueba Z se calcula como:

$$Z = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{14{,}5 - 15}{2{,}8/\sqrt{81}} = -1{,}607$$

Como se trata de una prueba unilateral a la izquierda, buscamos el valor crítico para un nivel de significancia de $\alpha = 0{,}01$ en una cola de la distribución normal. En la tabla de la distribución normal estándar, el valor para $\alpha = 0{,}01$ es aproximadamente 2,33. Por lo tanto, $z_0 \in [-2{,}33, +\infty]$.

El valor de Z no cae en la región de rechazo, por ello no rechazamos la hipótesis nula H_0 al nivel de significancia del 1 %. Esto indica que no hay suficiente evidencia estadística para concluir que el tiempo de fabricación promedio es menor que 15 minutos.

4. **En un estudio sobre una especie de aves, se afirma que el paso promedio es de 300 gramos. Se toma una muestra de 16 aves y se obtiene una media de 290 gramos y una desviación estándar de 12 gramos. ¿Es razonable concluir que el peso promedio es menor a 300 gramos? Usa un nivel de significancia del 5 %.**

Hipótesis nula H_o: $\mu \geq 300$
Hipótesis alternativa H_1: $\mu < 300$

Datos del ejercicio:

- Media poblacional hipotética (μ_0) = 300 gramos
- Desviación estándar muestral (s) = 12 gramos
- Tamaño de la muestra (n) = 16
- Media muestral ($\overline{x}$) = 290 gramos
- El nivel de significancia dado es $\alpha = 0.05$

Dado que no conocemos la desviación estándar poblacional y el tamaño de la muestra es pequeño ($n = 16$), usamos la prueba t de Student. El estadístico de la prueba t se calcula como:

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{290 - 300}{12/\sqrt{16}} = -3{,}33$$

En la tabla t de Studenr el valor para $\alpha = 0{,}05$ con 15 grados de libertad es -1,753. Por lo tanto, $t_0 \notin [-1,753, +\infty]$.

Rechazamos la hipótesis nula H_0 al nivel de significancia del 5 %. Esto indica que hay suficiente evidencia estadística para concluir que el peso promedio de las aves es menor a 300 gramos.

5. **Un estudio compara el tiempo promedio para completar una tarea entre empleados de dos compañías, A y B. En una muestra de 100 empleados de A, se obtuvo una media de 30 minutos y una desviación estándar de 5 minutos. En una muestra de 80 empleados de B, se obtuvo una media de 32 minutos y una desviación estándar de 4 minutos. ¿Es diferente el tiempo promedio entre las dos compañías? Usa un nivel de significancia del 5 %.**

Queremos probar si el tiempo promedio para completar la tarea es diferente entre las dos compañías, por lo que se trata de una prueba bilateral.

Hipótesis nula H_o: $\mu_A = \mu_B$
Hipótesis alternativa H_1: $\mu_A \neq \mu_B$

Datos del ejercicio:

- El nivel de significancia es $\alpha = 0{,}05$.
- Datos de la Compañía A:
  - Tamaño de la muestra $(n_A) = 100$
  - Media muestral $(\overline{x}_A) = 30$ minutos
  - Desviación estándar muestral $(s_A) = 5$ minutos
- Datos de la Compañía B:
  - Tamaño de la muestra $(n_B) = 80$
  - Media muestral $(\overline{x}_B) = 32$ minutos
  - Desviación estándar muestral $(s_B) = 4$ minutos

Como tenemos muestras grandes queremos comparar las medias de dos poblaciones, podemos usar una prueba Z para dos muestras independientes. El estadístico de prueba Z se calcula como:

$$Z = \frac{\overline{x}_A - \overline{x}_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}} = \frac{30 - 32}{\sqrt{\frac{5^2}{100} + \frac{4^2}{80}}} \approx -2{,}99$$

Dado que es una prueba bilateral con un nivel de significancia de $\alpha = 0{,}05$, dividimos α entre 2 para asignarlo a ambas colas de la distribución, de manera que $\alpha/2 = 0{,}025$. Buscando en la tabla de la distribución normal estándar, encontramos que el valor para $\alpha/2 = 0{,}025$ es 1.96. Por lo tanto, $z_0 \notin [-1{,}96, 1{,}96]$.

Rechazamos la hipótesis nula H_0 al nivel de significancia del 5 %. Esto indica que hay suficiente evidencia estadística para concluir que el tiempo promedio para completar la tarea es significativamente diferente entre las dos compañías.

10.8. ANOVA

1. **Los miembros de un equipo ciclista se dividen al azar en tres grupos que entrenan con métodos diferentes. El primer grupo realiza largos recorridos a ritmo pausado, el segundo grupo realiza series cortas de alta intensidad y el tercero trabaja en el gimnasio con pesas y se ejercita en el pedaleo de alta frecuencia. Después de un mes de entrenamiento se realiza un test de rendimiento consistente en un recorrido cronometrado de 9 km. Los tiempos empleados fueron los siguientes:**

	Método 1	Método 2	Método 3
	15	14	13
	16	13	12
	14	15	11
	15	16	14
	17	14	11
B	77	72	61
N	5	5	5

A un nivel de confianza del 95 % ¿Puede considerarse que los tres métodos producen resultados equivalentes? O por el cntrario ¿Hay algún método superior a los demás?

$$k = 3; \quad B = \sum_{i=1}^{k} B_i = 77 + 72 + 61 = 210; \quad N = \sum_{i=1}^{n} n_i = 5 + 5 + 5 = 15$$

$$S_A^2 = \frac{1}{k-1} \cdot \left[\sum_{j=1}^{k} \left(\frac{B_i^2}{n_i} \right) - \frac{B^2}{N} \right] = \frac{1}{3-1} \cdot \left[\left(\frac{77^2}{5} + \frac{72^2}{5} + \frac{61^2}{5} \right) - \frac{210^2}{15} \right] = 13{,}4$$

$$S_B^2 = \frac{1}{N-k} \cdot \sum_{i=1}^{n} \sum_{j=1}^{n} x_{ij}^2 - \sum_{i=1}^{n} \frac{B_i^2}{n_i} = \frac{1}{15-3} \cdot \Big[(15^2+16^2+14^2+15^2+17^2+14^2+13^2+$$

$$+15^2+16^2+14^2+13^2+12^2+11^2+14^2+11^2) - \left(\frac{77^2}{5} + \frac{72^2}{5} + \frac{61^2}{5} \right) \Big] = 1{,}43$$

$$F = \frac{S_A^2}{S_B^2} = \frac{13{,}4}{1{,}43} = 9{,}37$$

$$F_{\alpha;\, k-1;\, N-k} = F_{0{,}05;\, 3-1;\, 15-3} = F_{0{,}05;\, 2;\, 12} = 3{,}8853$$

Como $F > F_{\alpha;\, k-1;\, N-k}$ se rechaza la hipótesis nula y se concluye que los tres métodos de entrenamiento producen diferencias significativas.

2. Una lista de palabras sin sentido se presenta en la pantalla del ordenador con cuatro procedimientos diferentes, asignados al azar a un grupo de sujetos. Posteriormente se les realiza una prueba de recuerdo de dichas palabras, obteniéndose los siguientes resultados:

	Procdmt. 1	Procdmt. 2	Procdmt. 3	Procdmt. 4
	5	9	8	1
	7	11	6	3
	6	8	9	4
	3	7	5	5
	9	7	7	1
	7		4	4
	4		4	
	2			
B	43	42	43	18
N	8	5	7	6

¿Qué conclusiones pueden sacarse acerca de las cuatro formas de presentación, con un nivel de significancia del 5 %?

$$k = 4; \quad B = \sum_{i=1}^{k} B_i = 43{+}42{+}43{+}18 = 146; \quad N = \sum_{i=1}^{n} n_i = 8{+}5{+}7{+}6 = 26$$

$$S_A^2 = \frac{1}{k-1} \cdot \left[\sum_{j=1}^{k} \left(\frac{B_i^2}{n_i} \right) - \frac{B^2}{N} \right] =$$

$$= \frac{1}{4-1} \cdot \left[\left(\frac{43^2}{8} + \frac{42^2}{5} + \frac{43^2}{7} + \frac{18^2}{6} \right) - \frac{146^2}{26} \right] = 27{,}41$$

$$S_B^2 = \frac{1}{N-k} \cdot \sum_{i=1}^{n} \sum_{j=1}^{n} x_{ij}^2 - \sum_{i=1}^{n} \frac{B_i^2}{n_i} = \frac{1}{26-4} \cdot \Big[(5^2{+}7^2{+}6^2{+}3^2{+}9^2{+}7^2{+}4^2{+}2^2{+}$$

$$+9^2{+}11^2{+}8^2{+}7^2{+}7^2{+}8^2{+}6^2{+}9^2{+}5^2{+}7^2{+}4^2{+}4^2{+}1^2{+}3^2{+}4^2{+}5^2{+}1^2{+}4^2) -$$

$$- \left(\frac{43^2}{8} + \frac{42^2}{5} + \frac{43^2}{7} + \frac{18^2}{6} \right) \Big] = 3{,}91$$

$$F = \frac{S_A^2}{S_B^2} = \frac{27{,}41}{3{,}91} = 7{,}01$$

$$F_{\alpha;\, k-1;\, N-k} = F_{0{,}05;\, 4-1;\, 26-4} = F_{0{,}05;\, 3;\, 22} = 3{,}0491$$

Como $F > F_{\alpha;\, k-1;\, N-k}$ se rechaza la hipótesis nula y se concluye que los cuatro procedimientos de presentación producen diferencias significativas.

Unibertsitateko eskuliburuak
Manuales universitarios

ISBN: 978-84-9082-947-9

UPV/EHUko Argitalpen Zerbitzua
argitaletxea@ehu.eus • 94 601 2227
Biblioteka eraikuntza, 1. solairua
Sarriena auzoa z/g. Bizkaiko campusa
www.ehu.eus/argitalpenak

eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea

Servicio Editorial de la UPV/EHU
editorial@ehu.eus • 94 601 2227
Edificio Biblioteca, 1ª planta
Bº Sarriena s/n. Campus de Bizkaia
www.ehu.eus/argitalpenak