Estadística

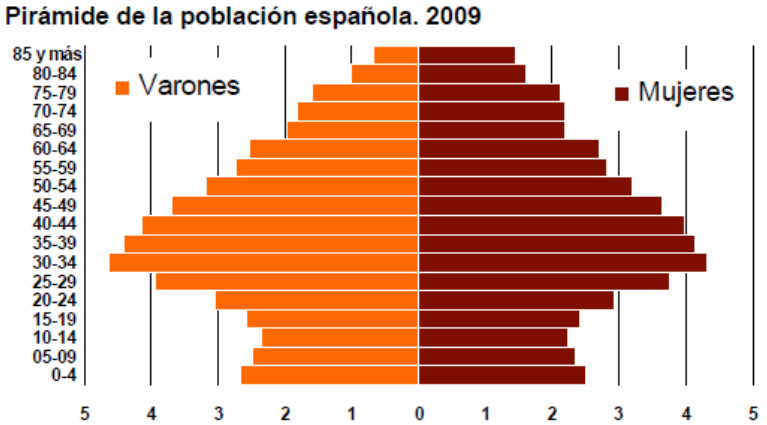
Eugenio Hernández

Universidad Autónoma de Madrid

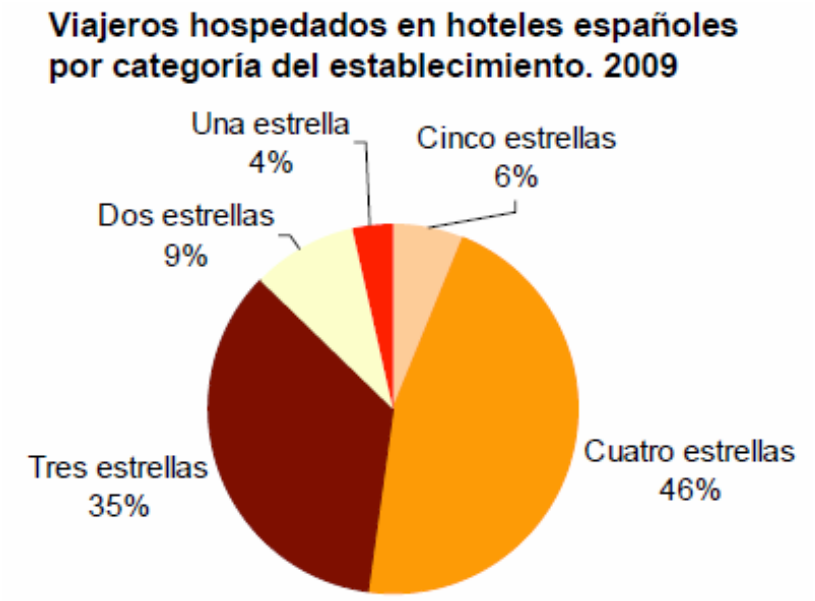
COMPLEMENTOS DE MATEMÁTICAS PARA LA EDUCACIÓN SECUNDARIA

ESTADÍSTICA: LA CIENCIA QUE ESTUDIA LOS DATOS.

Los datos se representan frecuentemente con diagramas de barras o de sectores.



Fuente: Explotación estadística del Padrón. INE



La campaña de Napoleón en Rusia en 1812

Hay gráficos muy variados específicos para cada conjuntos de datos.

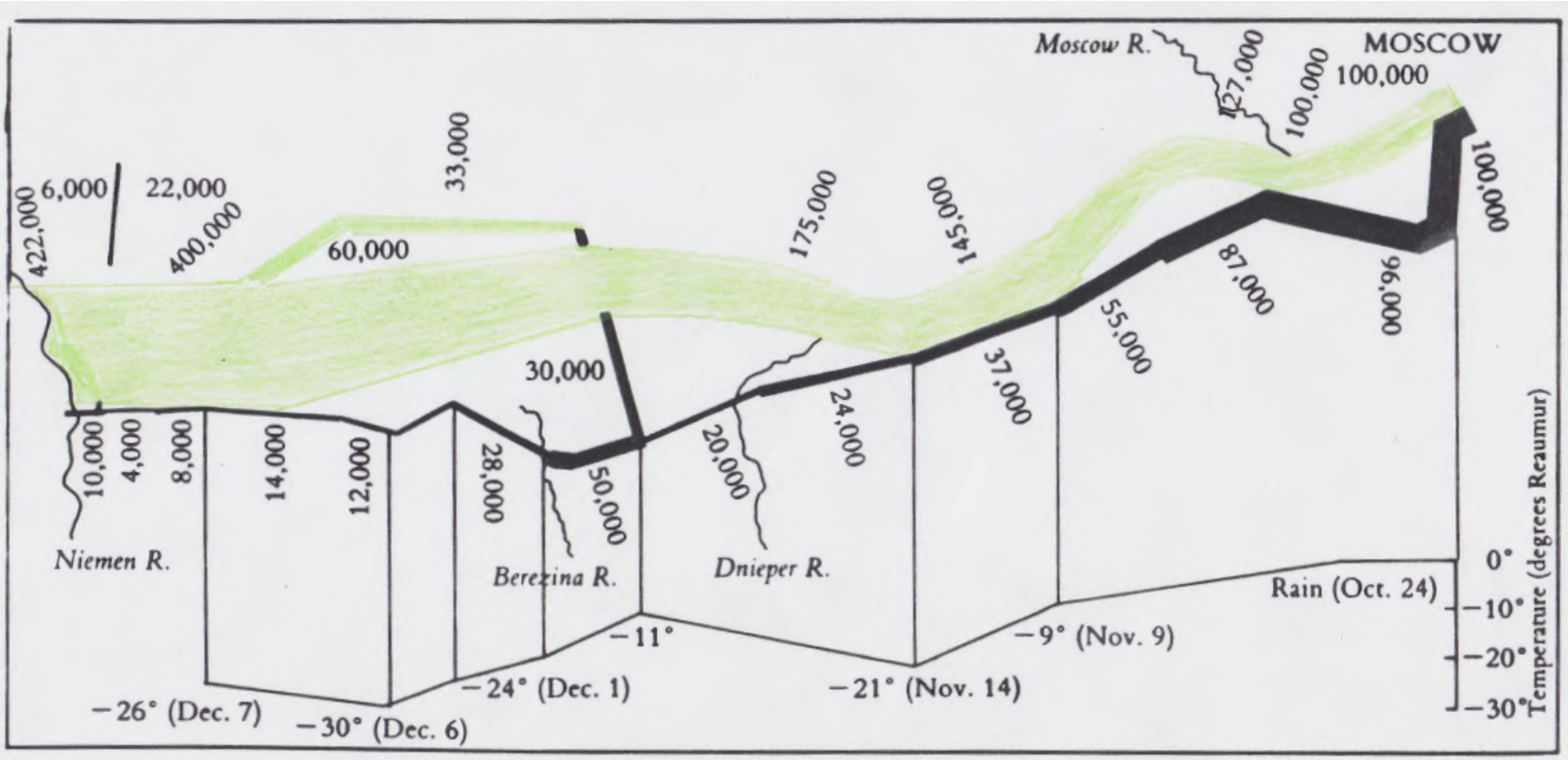
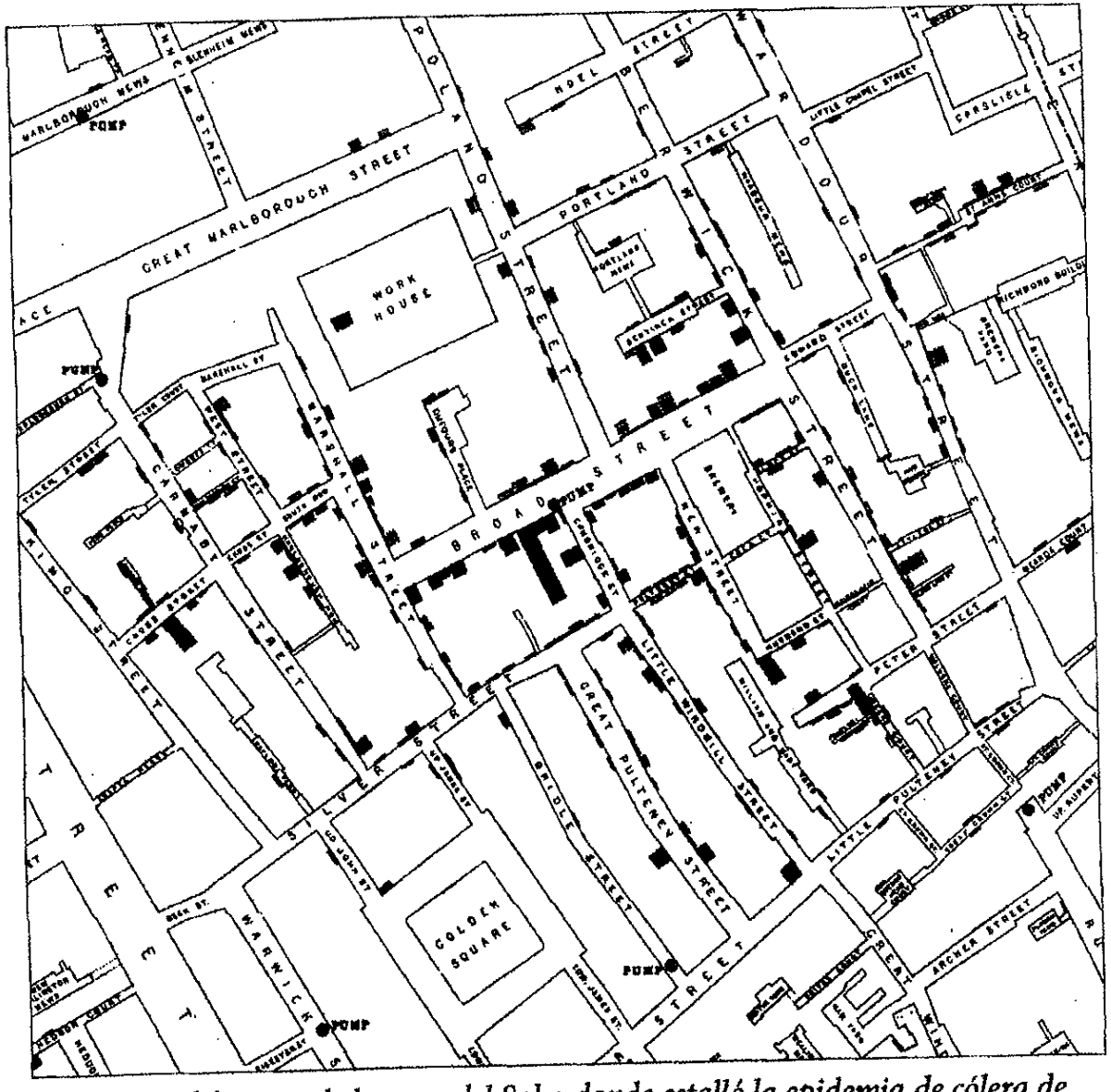


Figure 6.1 Redrawing of Charles Minard's 1861 graph of Napoleon's Russian campaign. (In the Reaumur temperature scale, water boils at 80°R and freezes at 0°R.)

JOHN SNOW Y LA FUENTE DEL CÓLERA - 1854



Fragmento del mapa de la zona del Soho donde estalló la epidemia de cólera de 1854. La fuente de la calle Broad se indica por medio de la leyenda «Pump», en el centro del mapa. Las rayas horizontales indican las víctimas de cada vivienda.

LA MEDIA - DATOS DISCRETOS

Para un conjunto de observaciones $x_1, x_2, \ldots, x_n$, con posibles repeticiones, la **media** es $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$.

LA MEDIA - DATOS CONTINUOS

Para una variable estadística **continua** con los datos agrupados en intervalos A_i se elige una marca de clase, x_i, generalmente el centro, y si el número de observaciones en A_i es n_i la **media** es $\overline{x} = \frac{1}{n} \sum_{i=1}^{k} n_i\, x_i = \sum_{i=1}^{k} \frac{n_i}{n}\, x_i = \sum_{i=1}^{k} f_i\, x_i$, donde $n = \sum_{i=1}^{k} n_i$ y $f_i = \frac{n_i}{n}$.

LA MEDIANA - DATOS DISCRETOS

Para una variable discreta, se ordenan los datos. Si hay un número impar de datos la mediana es el valor central; si hay un número par de datos la mediana es la media aritmética de los dos valores centrales.

LA MEDIANA - DATOS CONTINUOS

Cuando la variable estadística es continua y la distribución está dada con intervalos. la mediana es el intervalo que contiene a un punto por debajo del cual está el 50 % de los datos.

LA MODA - DATOS DISCRETOS

Para una variable discreta, es el valor o los valores más repetidos.

LA MODA - DATOS CONTINUOS

Cuando la variable estadística es continua con datos agrupados en intervalos de clase, la moda es el intervalo, o intervalos, en el que se acumulan mayor número de observaciones (el de mayor frecuencia absoluta).

LOS CUARTILES.

Primer cuartil: Valor por debajo del cual está el 25 % de las observaciones.

Segundo cuartil: Valor por debajo del cual está el 50 % de las observaciones.

Tercer cuartil: Valor por debajo del cual está el 75 % de las observaciones.

LA VARIANZA Y LA DESVIACIÓN TÍPICA: DATOS DISCRETOS.

Para una serie de datos discretos, $x_1, x_2, \ldots, x_n$, posiblemente algunos repetidos, la varianza y la desviación típica son:

$$V_n = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \qquad \text{y} \qquad \sigma_n = \sqrt{V_n}\,.$$

Nota: Algunos textos usan $V_{n-1} = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})^2$ y $\sigma_{n-1} = \sqrt{V_{n-1}}$ para calcular la varianza y la media.

Ejercicio 1. Prueba que $V_n = \dfrac{1}{n}(\displaystyle\sum_{i=1}^{n} x_i^2) - \overline{x}^2$.

En EXCEL, el cálculo de σ_n se puede hacer con la función DESVEST.P(Celda 1:Celda n). El cálculo de σ_{n-1} se puede hacer con la función DESVEST.M(Celda 1:Celda n).

LA VARIANZA Y LA DESVIACIÓN TÍPICA: DATOS CONTINUOS.

Para una variable estadística continua, distribuida en intervalos de clase con marcas $x_1, x_2, \ldots, x_k$ y frecuencias absolutas $n_1, n_2, \ldots, n_k$ la varianza y la desviación típica son:

$$V_n = \frac{1}{n}\sum_{i=1}^{k} n_i\,(x_i - \overline{x})^2 = \sum_{i=1}^{k} f_i\,(x_i - \overline{x})^2 \qquad \text{y} \qquad \sigma_n = \sqrt{V_n}\,.$$

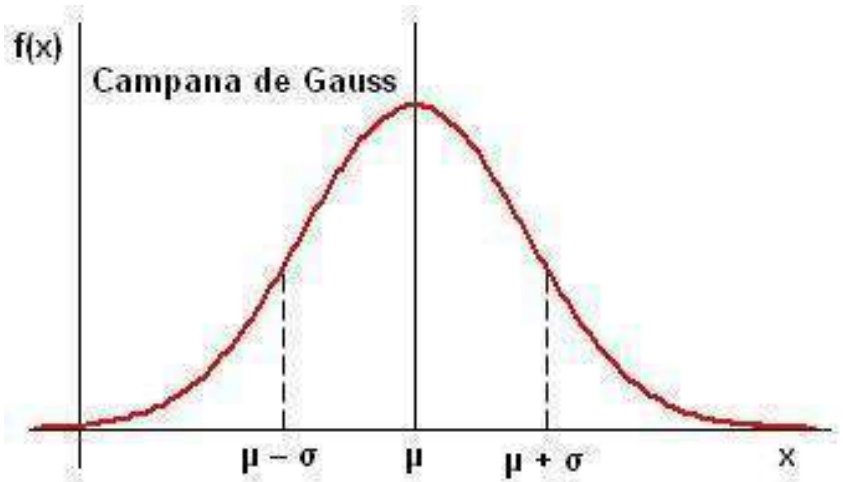
donde $n = \sum_{i=1}^{k} n_i$ y $f_i = \frac{n_i}{n}$.

Ejercicio 2. Comprueba que $V_n = (\sum_{i=1}^{n} f_i\, x_i^2) - \overline{x}^2$.

6.3.2. La distribución normal.

Una gran variedad de datos estadísticos provenientes de experimentos variados tienen histogramas que semejan una curva cuya fórmula es

$$N(\overline{x},\sigma): \quad f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\overline{x})^2}{2\,\sigma^2}}, \quad x\in(-\infty,\infty).$$



Ejercicio 3. Prueba que el área bajo una curva normal es 1.

Ejercicio 4. Prueba que para una distribución normal $N(\overline{x}, \sigma)(x)$ se tiene

$$\int_{-\infty}^{\infty} x\, N(\overline{x}, \sigma)(x)\, dx = \overline{x}\,, \quad \text{y} \quad \int_{-\infty}^{\infty} (x-\overline{x})^2 N(\overline{x}, \sigma)(x)\, dx = \sigma^2\,.$$

Ejercicio 5. Comprueba que los puntos de inflexión de la curva normal $N(\overline{x}, \sigma)(x)$ se tienen para los valores de x igual a $\overline{x} - \sigma$ y $\overline{x} + \sigma$.

Si una variable estadística X sigue una distribución normal $N(\overline{x}, \sigma)$, la probabilidad de que X tome valores entre x_0 y x_1 es

$$P(x_0 \leq X \leq x_1) = \int_{x_0}^{x_1} N(\overline{x}, \sigma)(x)\, dx.$$

El cálculo del área bajo una curva normal en un intervalo puede hacerse con integración numérica. Con el cambio de variable $\frac{x-\overline{x}}{\sigma} = z$ se pasa de $N(\overline{x}, \sigma)(x)$ a $N(0,1)(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$, y por tanto,

$$P(x_0 \leq X \leq x_1) = P(\frac{x_0 - \overline{x}}{\sigma} \leq Z \leq \frac{x_1 - \overline{x}}{\sigma})$$

donde Z es una variable estadística que sigue una distribución normal $N(0,1)$.

Hay tablas que dan los valores de $\int_{-\infty}^{k} N(0,1)(z)\,dz$ cuando $k \geq 0$ y $k < 4$ a incrementos de 0,01.
También puede usarse la función *Integral(Función, Extremo inferior, Extremo superior)* de Geogebra.

Ejercicio 6. Los resultados de un test sobre el CI siguen una distribución normal de media 100 y desviación típica 15. ¿Qué porcentaje de resultados está entre 80 y 130?

Ejercicio 7. Un niño de 12 años tuvo una puntuación de 152 en el test de inteligencia del ejercicio anterior. ¿Qué porcentaje de la población tendrá un CI superior o igual al suyo?

La regla de 1-2-3 desviaciones típicas.

Como $P(-1 \leq Z \leq 1) = 0,6826$, el 68,26 % de las observaciones de una distribución normal están a distancia de la media inferior a 1 desviación típica.

Como $P(-2 \leq Z \leq 2) = 0,9544$, el 95,44 % de las observaciones de una distribución normal están a distancia de la media inferior a 2 desviaciones típicas.

Como $P(-3 \leq Z \leq 3) = 0,9973$, el 99,73 % de las observaciones de una distribución normal están a distancia de la media inferior a 3 desviaciones típicas.

6.3.3. Distribuciones bidimensionales. Recta de regresión.

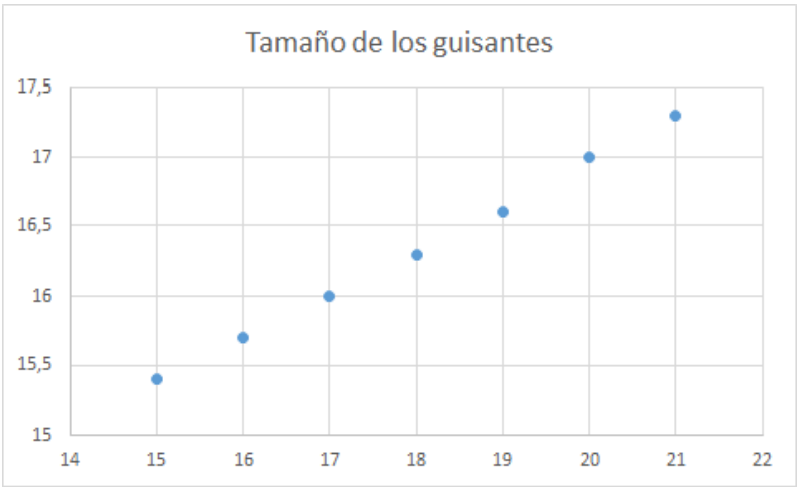
La relación entre dos variables es a veces exacta. En otras ocasiones, la relación no es tan precisa.

Ejemplo 1. Francis Galton (1822-1911) estaba convencido que la altura de los individuos era una propiedad hereditaria. Como no disponía de datos suficientes, por indicación de Charles Darwin, decidió experimentar con el diámetro de guisantes. En 1855 seleccionó guisantes y los dividió en 7 grupos de acuerdo con su diámetro. Después convenció a varios de sus amigos para que cada uno plantara 70 semillas, 10 de cada uno de los grupos. Una vez maduros, recopiló toda la cosecha, y midió el diámetro de los guisantes. Los resultados, para la media del diámetro de los guisantes, fueron los siguientes:

Padres	15	16	17	18	19	20	21
Hijos	15,4	15,7	16,0	16,3	16,6	17,0	17,3

Diámetro de la semilla $\times 0,01$ pulgadas (1 pulgada = 2,54 cm)

Se observa que hay una correlación positiva: padres pequeños tienen hijos pequeños y padres grandes tienen hijos grandes. Pero los hijos de padres pequeños no son tan pequeños, ni los hijos de padres grandes son tan grandes.



Galton llamó a este fenómeno reversión, pero más tarde, él y otros estadísticos usaron el término regresión porque la altura de los hijos *regresa* hacia la media.

Ejemplo 2. El cuadro contiene las calificaciones de 12 alumnos de Bachillerato en exámenes de Matemáticas, Física y Filosofía. Hacer el diagrama de dispersión de Matemáticas-Física y de Matemáticas-Filosofía, e indicar qué tipo de correlación hay entre ellas.

Matemáticas	2	3	4	4	5	6	6	7	7	8	10	10
Física	1	3	2	4	4	4	6	4	6	7	9	10
Filosofía	2	5	7	8	5	3	4	6	7	5	5	9

Ejercicio 8. Para una distribución bidimensional de datos $(x_1, y_1), \ldots (x_n, y_n)$, demuestra que $(\overline{x}, \overline{y})$ es el punto que minimiza el cuadrado de las distancias:

$$d^2(x, y) = D(x, y) = \sum_{i=1}^{n} (x_i - x)^2 + (y_i - y)^2 .$$

Coeficiente de correlación

Covarianza

La covarianza de un conjunto de datos $(x_1, y_1), \ldots (x_n, y_n)$ es

$$cov_{x,y} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}).$$

Ejercicio 9. Prueba que $cov_{x,y} = \frac{1}{n} \sum_{i=1}^{n} x_i y_i - \overline{x}\,\overline{y}$.

Coeficiente de correlación

El coeficiente de correlación de un conjunto de datos $(x_1, y_1), \ldots (x_n, y_n)$ es:

$$r = \frac{cov_{x,y}}{\sigma_x \, \sigma_y}.$$

Excel: COEF.DE.CORREL(Rango 1: Rango2)
COVARIANCE.P(Rango 1: Rango2)

La recta de regresión

La tendencia que muestran algunas nubes de puntos puede capturarse con una linea recta. Para encontrar esta recta, llamada recta de regresión, hay que hallar $y = a + bx$ de manera que se haga mínimo el error cuadrático medio (ECM), es decir, la media de los cuadrados de las distancias verticales entre los puntos y la recta:

$$ECM(a, b) = \frac{1}{n}\sum_{i=1}^{n}(y_1 - a - bx_i)^2 .$$

Ejercicio 10. Prueba que ECM se minimiza para $b = \frac{cov_{x,y}}{V_x}$ y $a = \overline{y} - \frac{cov_{x,y}}{V_x}\overline{x}$. Por tanto la recta de regresión es

$$y - \overline{y} = \frac{cov_{x,y}}{V_x}(x - \overline{x}) .$$

Ejercicio 11. El valor de ECM para los valores de a y b hallados en el ejercicio 10 se llama varianza residual. Prueba que la varianza residual es $V_y(1 - r^2)$.

Como la varianza residual es un número no negativo, se deduce del ejercicio 11 que el coeficiente de correlación es un número entre 1 y -1.

Recta de regresión con Excel.

- Selecciona ambas columnas de datos.
- En el menú *Insertar* elegir *Dispersión* para dibujar la nube de puntos.
- Pincha con el botón secundario del ratón sobre cualquier punto de la nube de puntos.
- Seleccionar *Agregar linea de tendencia.*
- En el menú que aparece a la derecha elegir *Lineal*.
- Para poner la ecuación en el gráfico, pincha con el botón secundario del ratón sobre la recta de regresión.
- Seleccionar *Formato de lineas de tendencia.*
- Marcar la casilla *Representar la ecuación en el gráfico.*
- También se puede añadir el valor de r^2 marcando la casilla correspondiente.

En un experimento aleatorio cuyo espacio muestral es E, una variable aleatoria X es una función $X : E \longrightarrow \mathbb{R}$ que a cada elemento del espacio muestral le asocia el valor numérico que nos interesa.

Ejemplo 3. Se lanza una moneda 3 veces. Describe la variable aleatoria X que indica el número de caras obtenidas.

Ejemplo 4. Se mide la altura de los miembros adultos de una comunidad. La variable aleatoria X es la que nos da la altura, que puede expresarse en cm.

Función de probabilidad.

Para una variable aleatoria su función de probabilidad es la función que indica la probabilidad de los diferentes valores que puede tomar la variable aleatoria.

Ejercicio 12. Describe la función de probabilidad de la variable aleatoria del ejemplo 3.

La función de probabilidad de la variable aleatoria del ejemplo 4 es una función Normal, de las estudiadas en la sección 6.3.2.

Una variable aleatoria X se dice discreta si solo toma un número finito de valores $x_1, x_2, \ldots, x_n$. Su función de probabilidad queda determinada por

$$P(X = x_1), \quad P(X = x_2), \ldots\ldots, P(X = x_n)\,.$$

Esperanza.

La esperanza de una variable aleatoria discreta X es $\mu = E[X] = \sum_{i=1}^{n} x_i\, P(X = x_i)$.

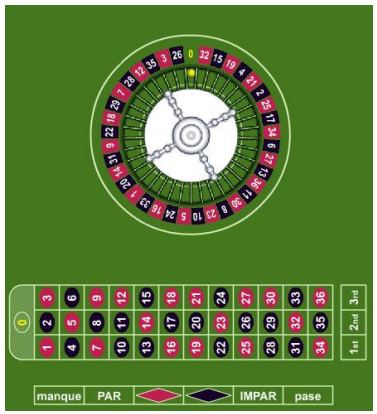
Varianza.

La varianza de una variable aleatoria discreta X es $V[X] = \sum_{i=1}^{n} (x_i - \mu)^2\, P(X = x_i)$.

Desviación típica.

La desviación típica de una variable aleatoria discreta X es $\sigma[X] = \sqrt{V[X]}$.

La ruleta



Una ruleta tiene 37 ranuras, de las cuales 18 son **rojas**, 18 **negras** y 1 **verde** (numerada 0). Cuando se gira la ruleta la bola tiene igual probabilidad de caer en cada una de las ranuras. Se pueden hacer muchas apuestas.
Una sencilla es elegir **Rojo** o **Negro**. Una apuesta de 1 euro al **Rojo** produce una ganancia de 1 euro si la bola cae en una ranura **roja**. Si cae en una ranura **negra** el croupier sonrie y se lleva tu euro. Si cae en la **verde**, sonrie aún más y se lleva cada euro de todos los participantes.

Ejercicio 13. Sea X la variable aleatoria *ganacias o pérdidas de cada apuesta de 1 euro al rojo*. Calcula la esperanza y la desviación típica de esta variable aleatoria.

Variables aleatorias independientes.

Sea X una v. a. discreta que puede tomar los valores $x_1, \ldots, x_n$. Sea Y otra v. a. discreta que puede tomar los valores $y_1, \ldots, y_m$. La función de probabilidad conjunta de X e Y es $P(X = x_i, Y = y_j), i = 1, \ldots, n, j = 1, \ldots, m$.

Variables aleatorias independientes

Dos variables aleatorias discreta X e Y se dicen independientes si para todo $i = 1, \ldots, n$, $j = 1, \ldots, m$ se tiene

$$P(X = x_i, Y = y_j) = P(X = x_i) \times P(Y = y_j).$$

Ejercicio 14. Se lanzan dos dados. Se consideran las variables aleatorias:
X = el mayor de los dos números.
Y = la suma de las puntuaciones.
Prueba que estas dos variables no son independientes.

Ejercicio 15. Sean X e Y dos variables aleatorias independientes.
a) Prueba que $E[X+Y] = E[X] + E[Y]$.
b) Prueba que $V[X+Y] = V[X] + V[Y]$.

A veces interesa conocer el total de ganancias o pérdidas cuando se hacen muchas apuestas. Sobre todo le interesa al casino y a las casas de apuestas.

Cuando la misma apuesta o experimento se repite n veces consecutivas, es decir $X_i = X, i = 1, \ldots, n$, los resultados son independientes y se tiene

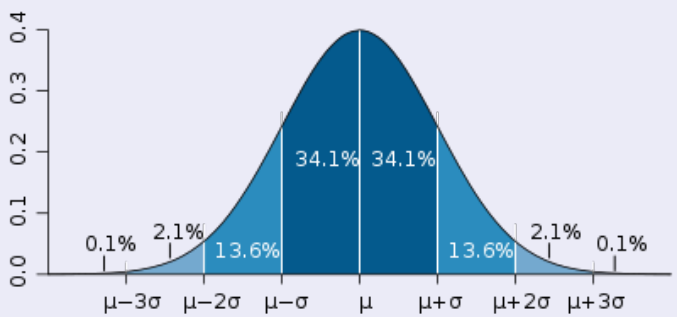
$$E[X_1 + \cdots + X_n] = nE[X], \quad \sigma[X_1 + \cdots + X_n] = \sqrt{nV[X]} = \sqrt{n}\,\sigma[X].$$

El Teorema Central del Límite

Supongamos ahora que estamos interesados en la v. a. que mide las ganancias o pérdidas después de n apuestas. Poniendo $X_i = X$ para $i = 1, 2, \ldots, n$, que se comportan independientemente, esta v. a. es $S_n = X_1 + X_2 + \cdots + X_n$.

Teorema Central del Límite

La suma S_n de n v. a. independientes e idénticamente distribuidas, con media μ y desviación típica σ, se aproxima, para n grande, a una Normal de media $n\mu$ y desviación típica $\sqrt{n}\sigma$. Es decir, la distribución de probabilidad de S_n se parece mucho a la Campana de Gauss:



Ejercicio 16. Supongamos que un jugador hace 100 apuestas de 1 euro al **Rojo** de la ruleta en una noche en el casino.

a) Prueba que la esperanza del jugador es perder en media 2,7 euros cada 100 apuestas. Calcula también la desviación típica de las 100 apuestas.

b) Con la regla de las 3 desviaciones típicas, prueba que el 99,73 % de las veces las *ganancias o pérdidas* del jugador estarán entre $-32,69$ euros y $27,29$ euros aproximadamente.

El casino siempre gana

La ruleta es una forma agradable, relajada y muy cómoda de perder dinero

Ejercicio 17. Pero en el casino hay muchos jugadores. Supongamos que en total se han hecho 100.000 apuestas de 1 euro al rojo en una noche.

a) Prueba que la esperanza del casino es ganar en media 2702,02 euros. Calcula la desviación típica de las 100.000 apuestas.

b) Con la regla de las 3 desviaciones típicas, prueba que el 99,73 % de las veces las ganancias del casino estarán entre $1754,37$ euros y $3651,03$ euros.

El fichero casino.xls, que se puede descargar desde el Moodle de la asignatura, contiene una simulación de este juego con Excel. Ha sido elaborado por Pablo Fernández Gallardo (Matemáticas, UAM).