Instituto Valenciano de Investigaciones Agrarias

Seminario

MÉTODOS ESTADÍSTICOS PARA LA INVESTIGACIÓN AGRONÓMICA

ORGANIZACIÓN DEL SEMINARIO

- Una sesión semanal de 2 horas (¿martes de 9:00 a 11:00?)
- Documentación básica para las primeras sesiones: libro MEPI y capítulos 11 y 12 del libro MEI (disponibles ya como documentos Word)
- En general se preparará para cada sesión una presentación en Power Point, que se hará llegar por adelantado a los asistentes
- **Importante:** los asistentes pueden aportar datos de sus investigaciones, con el fin de discutir entre todos sus posibles análisis estadísticos

Programa tentativo del Seminario (revisable)

- RECORDATORIO DE CONCEPTOS BÁSICOS
- ANÁLISIS DE LA VARIANZA
- MODELOS DE REGRESIÓN
- DISEÑO DE EXPERIMENTOS
- INTRODUCCIÓN A TEMAS AVANZADOS: Regresión Logística, Análisis de Supervivencia, Análisis Multivariante, … (en función del interés de los asistentes)

Recordatorio de Conceptos Estadísticos Básicos

- Poblaciones. Muestras. Variables aleatorias
- Representatividad y precisión. Tamaño muestral
- Estadística Descriptiva e Inferencia Estadística
- Estadística Descriptiva
  - Variables cualitativas: tabulación, gráficos de barras y de tarta
  - Parámetros de posición, dispersión y forma
  - Representaciones gráficas: Box-Whisker, Histogramas
  - Estadística Descriptiva en variables bidimensionales
- Ejercicio para los asistentes
- Variables aleatorias teórica
  - Caracterización. Esperanza matemática
  - Distribuciones discretas: Binomial y Poisson
  - La distribución Normal. Papel probabilístico
  - El papel de los modelos matemáticos en el Método Científico
- Distribuciones en el muestreo
- Ejercicio para los asistentes

Poblaciones

- El conjunto de entes o "individuos" para los que se desea obtener conclusiones en una investigación constituye la **población** objeto del estudio.
- En algunos casos las poblaciones existen realmente estando formadas por un número finito (quizás muy elevado) de individuos:
  - **Ejemplo1**: al controlar la calidad de una partida de piezas la población está constituida por todas las piezas de la partida
  - **Ejemplo 2**: en un estudio sobre la incidencia de una nueva virosis en una zona, la población podrían formarla todos los naranjos existentes en la zona
- En la mayoría de los casos, sin embargo, la población realmente relevante a estudiar no existe en la realidad, sino que tiene un carácter abstracto, siendo fruto del necesario proceso de conceptualización que debe preceder al estudio científico de cualquier problema real.
  - **Ejemplo 3**: todas las parcelas que podrían plantarse con una determinada variedad de chufa y cultivarse en ciertas condiciones
  - **Ejemplo 4**: todas las cajas de 20 tomates tratadas y almacenadas en ciertas condiciones
- ***Una adecuada y precisa definición de la población objeto del estudio es el primer paso en cualquier investigación.***

Subpoblaciones

- Las poblaciones pueden dividirse en **subpoblaciones**:
- bien a partir de ciertas características intrínsecas de los individuos que las forman:
  - en el Ejemplo 2 diferentes patrones definiría distintas subpoblaciones
  - en el Ejemplo 3 diferentes tipos de suelo definirían diferentes subpoblaciones
- bien a partir de los tratamientos que se hayan aplicado a los individuos
  - En el Ejemplo 3 diferentes tipos o dosis de abonado definirían diferentes subpoblaciones, según la que hubiera recibido la parcela
  - En el Ejemplo 4 las cajas podrían clasificarse en subpoblaciones en función del tratamiento fungicida que hubieran recibido o de la temperatura de almacenamiento

Muestras

- En general, es imposible estudiar todos los individuos de una población, por lo que las conclusiones deberán obtenerse analizando sólo un subconjunto de individuos de la población: la **muestra**
- Para que las conclusiones obtenidas de una muestra sean válidas para una población, la muestra debe ser **representativa** de ésta
- **Muestra aleatoria simple:** es el modelo matemático teórico de muestra representativa: se seleccionan al azar e independientemente los individuos que la van a componer, de forma que todos los individuos de la población tengan a priori la misma probabilidad de ser incluidos.
- Esta forma de proceder, sin embargo, muy rara vez es aplicable de forma estricta en la práctica, por lo que siempre hay que extremar las precauciones para que la forma real de obtener la muestra sea lo más parecida posible a la ideal.
- En general, ***un buen conocimiento previo sobre la población y el tema en estudio es indispensable para decidir si una muestra puede considerarse o no "representativa" de la misma***.

¿Cómo se eligen los individuos de la muestra?

- ¿En el caso (poco frecuente en investigación agraria) de que exista una población real finita, la selección de muestras puede llevarse a cabo de forma aleatoria (sorteo, tablas de números al azar, …)
- En el caso (más general) de las poblaciones abstractas, sus individuos pueden considerarse generados por un determinado proceso: plantar parcelas de chufas, preparar y almacenar una caja de tomates en ciertas condiciones
- En estos casos la forma de seleccionar una muestra de N individuos es realizar N veces el proceso que genera los individuos de la población.
- Cada prueba de una experiencia genera un individuo de la población hipotética que se obtendría si se realizase un número muy elevado de pruebas en las condiciones especificadas

Tamaño de la muestra

- Es muy frecuente la pregunta: **¿Qué tamaño mínimo debe tener mi muestra para que sea "representativa"?** (o, alternativamente y de una forma aun más genérica, para tener "buenos" resultados)?
- La **pregunta está mal planteada** y confunde dos conceptos diferentes: ***representatividad*** y ***precisión*** de una muestra
- Una vez que nuestros conocimientos de un problema nos permiten considerar que una muestra obtenida de una determinada forma es "representativa" de cierta población, siempre es posible analizarla estadísticamente, sea cual sea su tamaño
- El **tamaño muestral**, sin embargo, condiciona la precisión de las estimaciones que se obtengan a partir de la muestra, así como la **potencia** de los análisis estadísticos para detectar la posible existencia de los efectos de interés.

Tamaño de la muestra (continuación)

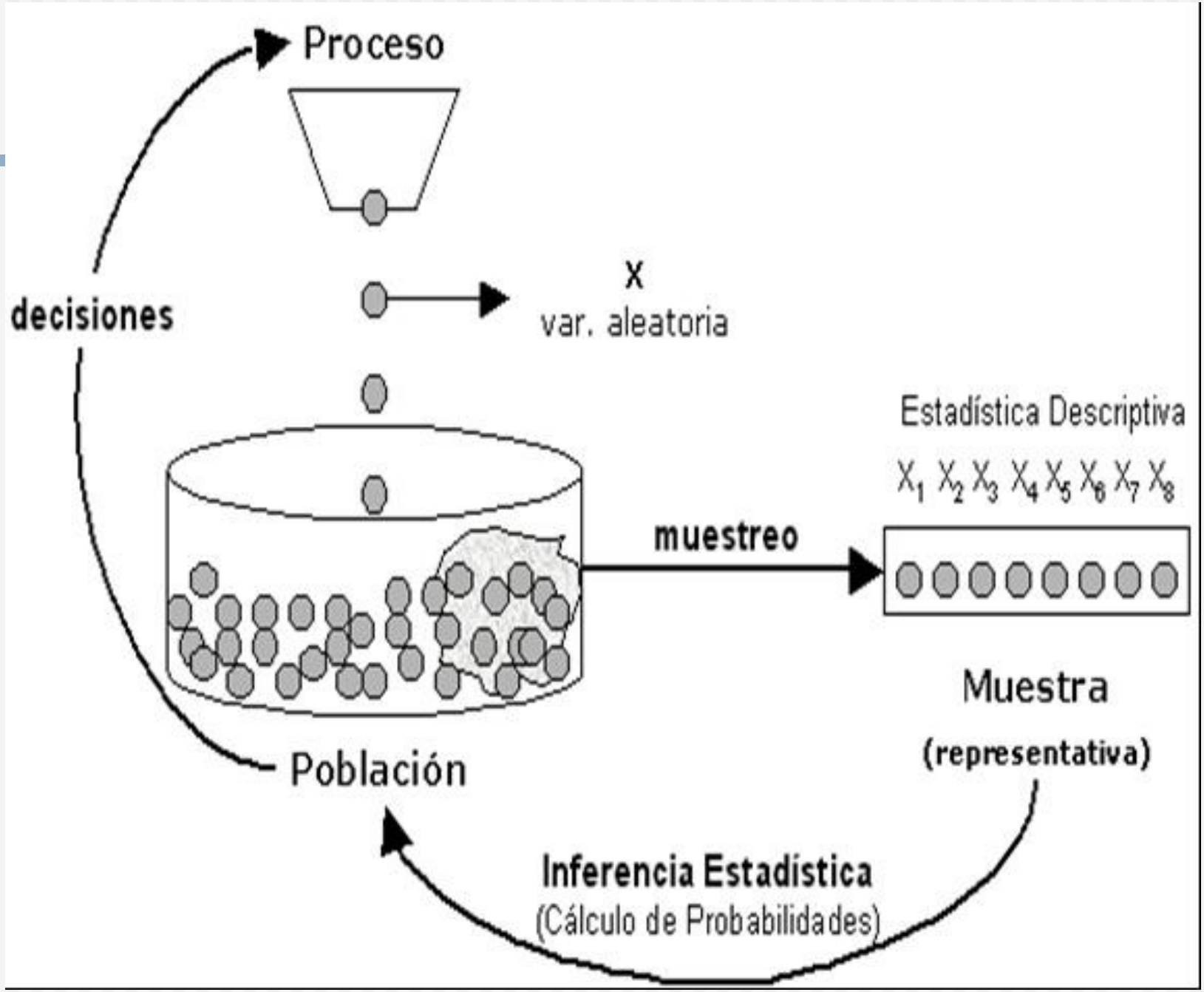
- Si la muestra es pequeña, los efectos se estimarán con poca precisión y sólo se detectarán como significativos (concepto que se estudia más adelante) aquéllos que realmente sean muy grandes.
- La discusión del tamaño mínimo que debe tener una muestra exige definir previamente la precisión deseada, concretando las magnitudes potenciales de los efectos para las que se desearía tener una "razonable seguridad" (que también habrá que concretar) de detectarlos como significativos.
- La determinación de dicho tamaño exige también, en general, tener una idea de la magnitud de la variabilidad existente en las poblaciones muestreadas.
- ***Debe quedar claro que una muestra "representativa" siempre puede ser analizada estadísticamente sea cual sea su tamaño (que, eso sí, condicionará la precisión de las estimaciones y la potencia de los análisis, pero no su legitimidad), y que una muestra no "representativa" no puede ser analizada estadísticamente, sea cual sea su tamaño***

Estadística Descriptiva

- Su objetivo es el tratamiento de los datos con el fin de poner de manifiesto sus características más relevantes.
- Con esta finalidad utiliza herramientas sencillas como:
  - tabulaciones adecuadas
  - cálculo de parámetros que sinteticen las distintas características de las pautas de variabilidad observada (posición, dispersión, asimetría, grado de relación, etc.)
  - representaciones gráficas (histogramas, diagramas Box-Whisker, gráficos en papel probabilístico, diagramas de dispersión, ...)
- En estos análisis meramente descriptivos no se pretende obtener conclusiones de carácter general sobre la población de la que procede la muestra, sino simplemente sintetizar los aspectos esenciales de las pautas de variabilidad existentes en los datos y que quedan frecuentemente enmascarados bajo un alud de números ("*los árboles no dejan ver el bosque*")
- **Un análisis descriptivo cuidadoso de los datos debe ser siempre el primer paso en cualquier estudio estadístico.**

Inferencia Estadística

- En general, sin embargo, el objetivo final de un estudio estadístico es la población muestreada, y la muestra es sólo un medio para llegar a conocerla.
- El análisis de los datos con el fin de obtener conclusiones que sean aplicables a la población, o poblaciones, de la que la muestra procede constituye el objeto de la ***Inferencia Estadística***.
- Estas conclusiones, al basarse en el análisis de sólo una parte de la población (la muestra), llevan siempre asociadas un cierto margen de incertidumbre
- Para realizar sus inferencias, y cuantificar ese margen de incertidumbre, la Ciencia Estadística se basa en los modelos matemáticos desarrollados en la Teoría del Cálculo de Probabilidades, que constituye por tanto una herramienta esencial para la Estadística


Proceso
decisiones
X
var. aleatoria
Estadística Descriptiva
X_1 X_2 X_3 X_4 X_5 X_6 X_7 X_8
muestreo
Muestra
(representativa)
Población
Inferencia Estadística
(Cálculo de Probabilidades)

Variables cuantitativas discretas y continuas

- **Variables discretas:** Cuando el conjunto de los valores que podría tomar la variable aleatoria es discreto (es decir, finito o infinito numerable)
- Ejemplos: todas aquéllas del tipo "número de …" y también cualquier variable que se origine al codificar las diferentes alternativas de una característica cualitativa

- **Variables continuas:** Cuando el conjunto de los valores que podría tomar la variable aleatoria es un infinito continuo
- Ejemplos: todas las características que se miden sobre una escala de naturaleza básicamente continua (por ejemplo, el rendimiento en una parcela, la longevidad de un parasitoide, …), al margen de que se expresen con más o menos decimales.

- Nota: Algunas variables discretas con un rango de variación muy amplio pueden tratarse matemáticamente como si fueran variables continuas.

Variables K-dimensionales

- Cuando sobre cada individuo de la población se estudian K características aleatorias diferentes, se tiene una variable aleatoria K-dimensional.
- Es importante darse cuenta de la diferencia entre una variable aleatoria K-dimensional, en la que las K variables se miden sobre los individuos de una única población, y un conjunto de K variables aleatorias unidimensionales, definidas sobre K poblaciones distintas.
- *Ejemplos:¿El contenido de leucocitos en la sangre de individuos alcohólicos y no alcohólicos es una variable aleatoria bidimensional? ¿Y las estaturas del marido y de la mujer en los matrimonios jóvenes de un país?*

Estadística Descriptiva

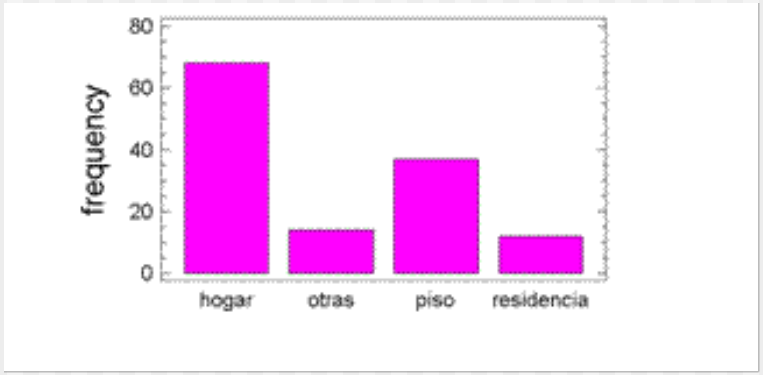
- El primer paso en el análisis de un conjunto de datos debe ser siempre un tratamiento descriptivo sencillo de los mismos, buscando poner de manifiesto las características y regularidades existentes y sintetizarlas en un número reducido de parámetros o mediante representaciones gráficas adecuadas
- Variables cualitativas: tabulación, diagramas de barras o de tarta
- Variables cuantitativas: cálculo de parámetros muestrales que sintetizan características relevantes de los datos
  - Posición: **media**, mediana, cuartiles
  - Dispersión: recorrido, **desviación típica**, intervalo entre cuartiles
  - Forma: coeficientes de asimetría y de curtosis
- Variables cuantitativas: representaciones gráficas:
  - Histogramas
  - Diagramas Box-Whisker
- Variables bidimensionales cuantitativas: diagramas de dispersión, coeficiente de correlación

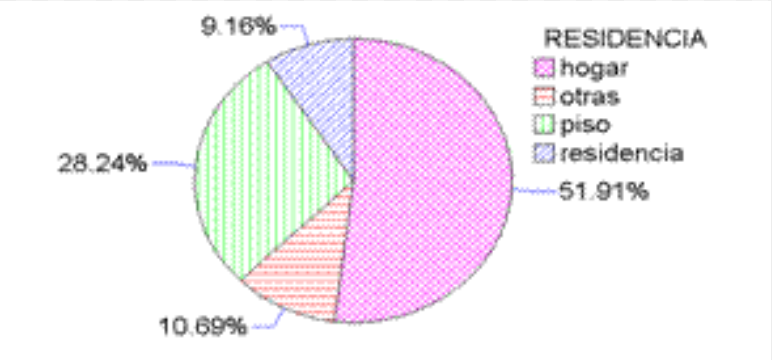
Descripción de variables cualitativas

- Las variables cualitativas tienen un número K, generalmente reducido, de alternativas posibles. La forma más inmediata de resumir la información de una muestra es una tabla que recoja las frecuencias (absolutas o relativas) observadas de cada categoría.
- Dicha tabla puede "visualizarse" mediante un **diagrama de barras** en el que a cada una de dichas alternativas se le hace corresponder una barra cuya altura se hace proporcional a la frecuencia observada de la misma.
- Con el mismo fin es muy corriente utilizar **diagramas de tarta**, repartiendo la superficie total de un círculo en sectores cuyas áreas sean proporcionales a las frecuencias observadas en la muestra para cada una de las alternativas

Ejemplo: lugar de residencia durante el curso de 131 alumnos de 1º de la ETSIA (1988)

			Frecuencia
Clase	*Valor*	*Frecuencia*	*Relativa*
1	hogar	68	0.5191
2	otras	14	0.1069
3	piso	37	0.2824
4	residencia	12	0.0916





Variables cuantitativas: Parámetros de posición

- **Media muestral**

media muestral $$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

- **Mediana**: es el valor central, una vez ordenados los datos. (Si resulta sensiblemente mayor que la media ➔ datos con asimetría positiva). En datos muy asimétricos la mediana puede ser un indicador de posición más adecuado que la media (justificarlo)

- **Cuartiles**:
  - Primer cuartil C1: la cuarta parte de los datos son inferiores a C1
  - Tercer cuartil C3: la cuarta parte de los datos son superiores a C3
  - Entre C1 y C3 están el 50% “central” de los datos

Variables cuantitativas: Parámetros de dispersión

- **Recorrido**: diferencia entre el máximo y el mínimo valor
- **Desviación típica**: raíz cuadrada del promedio de los cuadrados de las desviaciones respecto a la media

$$\text{Desviacion Tipica: } s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N-1}}$$

- En datos "normales" se cumple aproximadamente:
  - Dos tercios de los datos difieren de la media menos de s
  - El 95% de los datos difiere de la media menos de 2s
  - La práctica totalidad de los datos (en teoría más del 99.7%) difieren de la media menos de 3s

- **Intervalo entre cuartiles**: C3 – C1

Variables cuantitativas: Parámetros de forma: asimetría

- **Coeficiente de asimetría**:

Coeficiente de Asimetria: $$CA=\frac{\sum (x_i-\bar{x})^3 \big/ (N-1)}{s^3}$$

- Se utilizan sobre todo para ver si el modelo normal es adecuado
- En contextos inferenciales se utiliza el coeficiente de asimetría "estandarizado" (dividido por una estimación de su desviación típica en el muestreo) que en muestras que proceden de poblaciones normales está comprendido (en el 95% de los casos) entre -2 y +2.

Variables cuantitativas: Parámetros de forma: curtosis

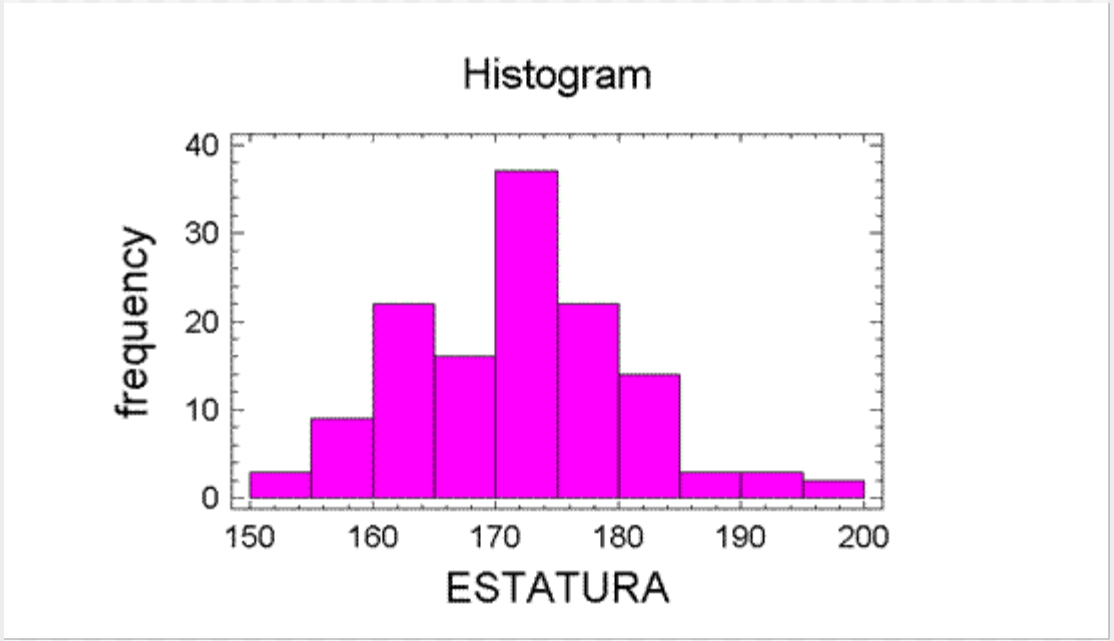
- **Coeficiente de curtosis**:

$$CC = \frac{\sum (x_i - \bar{x})^4 / (N-1)}{s^4} - 3$$

- En contextos inferenciales se utiliza el coeficiente de curtosis "estandarizado", que en muestras que proceden de poblaciones normales está comprendido (en el 95% de los casos) entre -2 y +2.
- Coeficiente de curtosis elevados positivos indican en general la presencia de datos anómalos en la muestra.
- Coeficiente de curtosis elevados negativos pueden deberse a datos artificialmente censurado (se han eliminado valores considerados extremos) o también a mezclas de poblaciones.

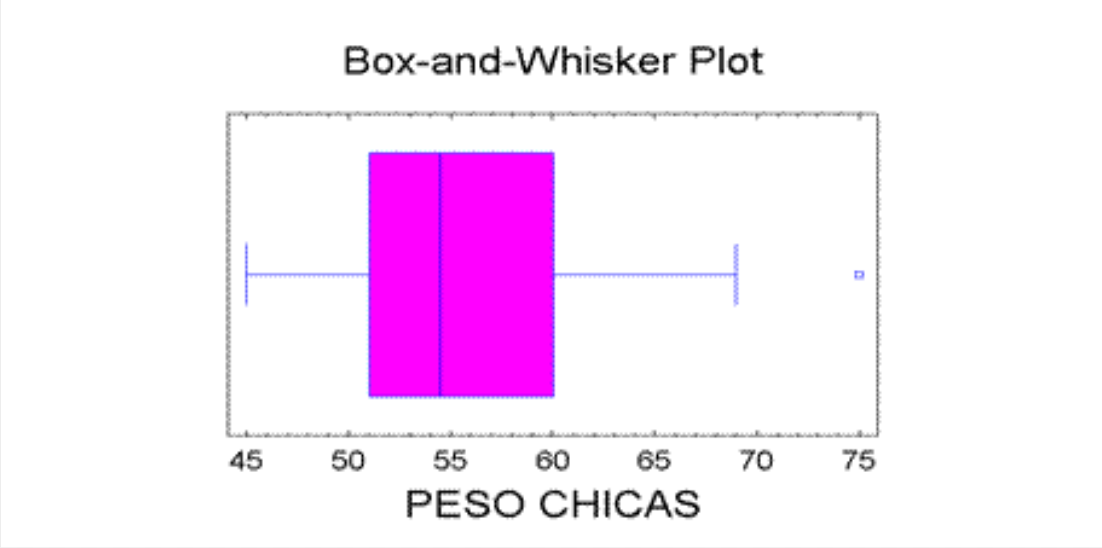
Histogramas

- Un histograma es una representación gráfica de un conjunto de valores observados de una variable cuantitativa continua (o discreta pero con un número elevado de valores diferentes). En el eje horizontal de las abscisas se representan los valores agrupados en tramos de igual tamaño. Sobre cada tramo se levanta una barra de altura proporcional a la frecuencia de valores observados en el mismo.
- Mínimo de 40 ó 50 datos. Número de tramos un entero cercano a la raíz cuadrada del número de datos (no más de 15 ó 20 tramos)



Diagramas Box-Whisker

- Presenta, frente a un histograma, la ventaja de no exigir un número elevado de datos para su construcción, además de resultar más sencillo su manejo cuando el objetivo es comparar distintos conjuntos de datos
- La "caja" comprende el 50% de los valores centrales de los datos, extendiéndose entre el primer cuartil y tercer cuartil; la línea central corresponde a la mediana. Los "bigotes" se extienden desde el menor al mayor de los valores observados y considerados "normales". Aquellos valores extremos que difieren del cuartil más próximo en más de 1.5 veces el intervalo intercuartílico, se grafican como puntos aislados



Variables bidimensionales

- Cuando sobre cada individuo de la muestra se “miden” dos características aleatorias se tiene una variable bidimensional
- En estas situaciones tiene especial interés, además de analizar por separado cada variable, estudiar la posible relación entre ambas
- Tres situaciones posibles:
  - Las dos variables son de naturaleza cualitativa (por ejemplo el sexo del alumno y su tendencia política)
  - Una variable es de tipo cualitativo y la otra de tipo cuantitativo (por ejemplo el sexo y la estatura del alumno)
  - Las dos variables son de naturaleza cuantitativa (por ejemplo el peso y la estatura de los alumnos)
- ¿Qué herramientas estadísticas se utilizan en cada caso para describir la posible relación entre las dos variables?

Relación entre dos variables cualitativas

- Construir una tabla con las frecuencias observadas para cada combinación de alternativas de la primera con la segunda variable
- La tabla se interpreta mejor si cada frecuencia se expresa en términos relativos, bien respecto al total de la fila o respecto al total de la columna (lo que sea más relevante para el objetivo perseguido)
- Ejemplo: posible relación del sexo con la tendencia política

Frecuencias relativas calculadas sobre el total de las columnas

	centro	derecha	izquier	pasan	Total por Fila
chicas	10	9	12	9	40
	41.67%	26.47%	40.00%	23.68%	31.75%
chicos	14	25	18	29	86
	58.33%	73.53%	60.00%	76.32%	68.25%
Total	24	34	30	38	126
	19.05%	26.98%	23.81%	30.16%	100.00%

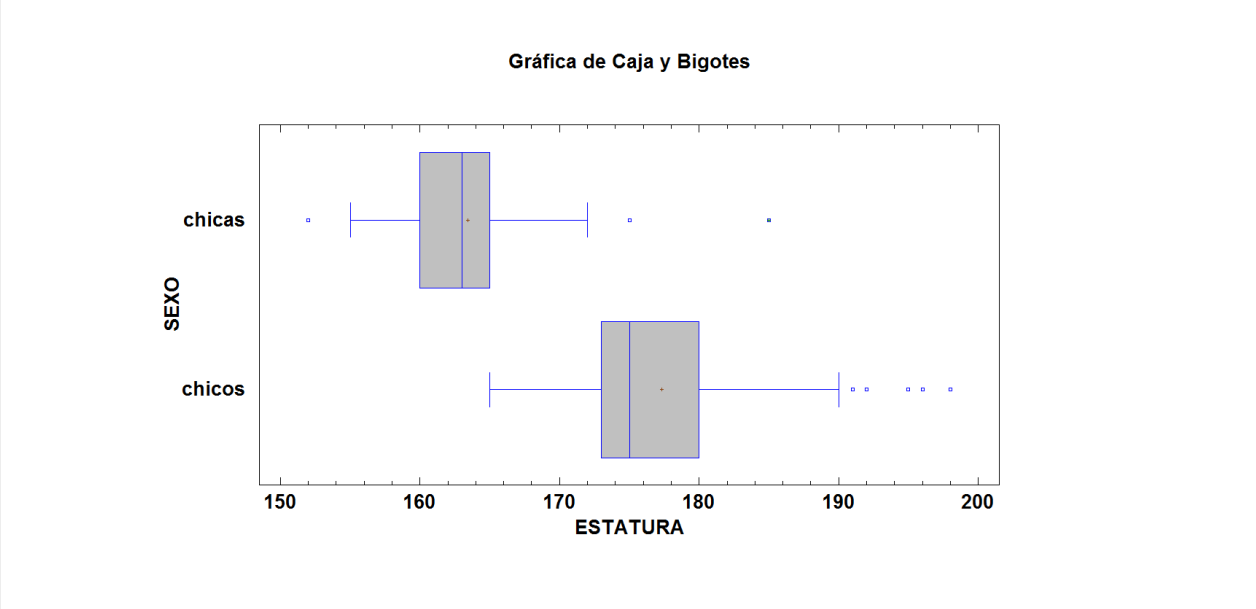
Frecuencias relativas calculadas sobre el total de las filas

	centro	derecha	izquier	pasan	Total por Fila
chicas	10	9	12	9	40
	25.00%	22.50%	30.00%	22.50%	31.75%
chicos	14	25	18	29	86
	16.28%	29.07%	20.93%	33.72%	68.25%
Total	24	34	30	38	126
	19.05%	26.98%	23.81%	30.16%	100.00%

- ¿Hay más tendencia hacia la izquierda en los chicos que en las chicas, porque el 60% de los alumnos de izquierda son chicos?

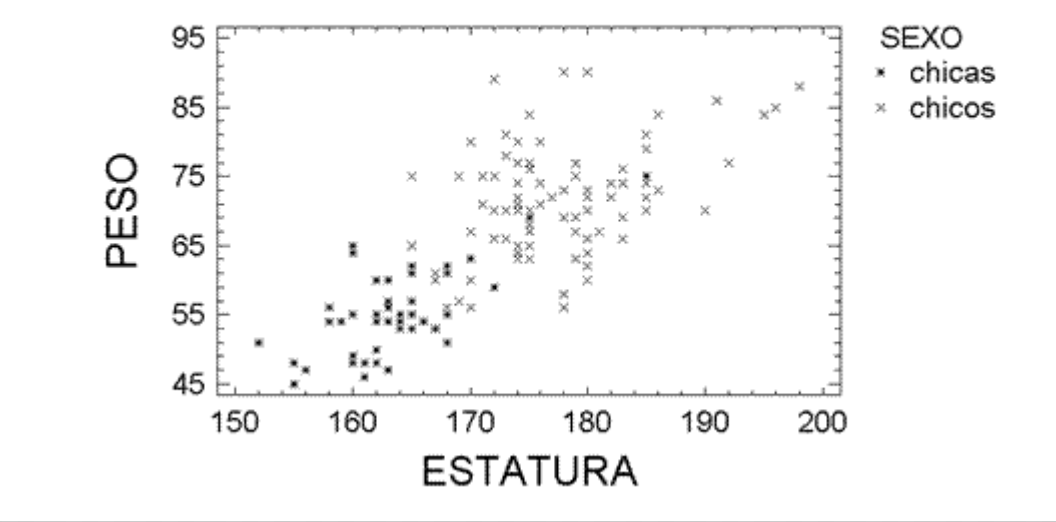
Relación entre una variable cualitativa y una cuantitativa

- La variable cualitativa particiona la población total en subpoblaciones correspondientes a cada alternativa.
- Se estudia y compara la pauta de variabilidad de la variable cuantitativa en las diferentes subpoblaciones, calculando los diferentes parámetros de posición, dispersión y forma, o gráficamente (por ejemplo mediante un diagrama Box-Whisker múltiple)
- Ejemplo: relación de la estatura de los alumnos con el sexo



Relación entre dos variables cuantitativas: diagramas de dispersión

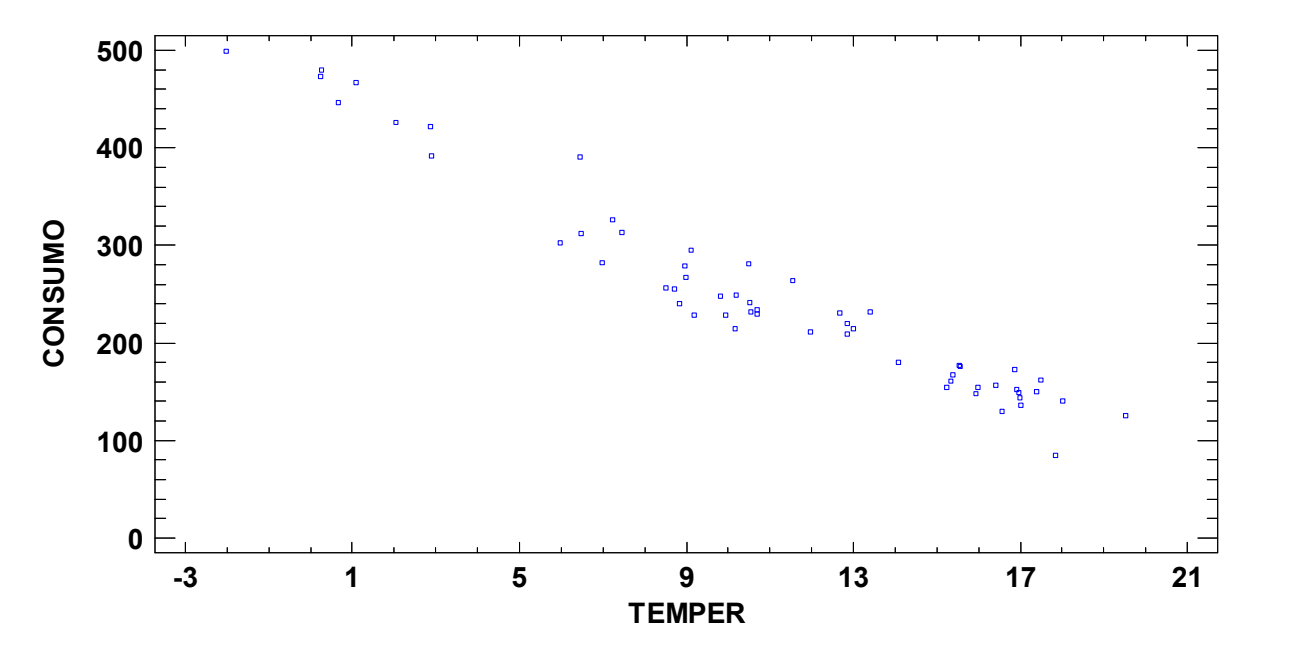
- Una forma de visualizar la relación entre dos variables cuantitativas es construir un **diagrama de dispersión** (scatter plot)
- La forma de la nube de puntos indica si existe una relación y, en caso afirmativo, informa sobre su sentido y "fuerza"
- Cuando una variable X1 se sospecha que depende otra variable X2, conviene poner esta última en el eje de abscisas
- Ejemplo 1:relación entre la estatura y el peso de los alumnos



- La forma de la nube indica una relación positiva medianamente fuerte entre Peso Y Estatura

Otro ejemplo de diagrama de dispersión

- **Ejemplo 2**:Relación entre la temperatura diaria y el consumo diario de energía en Ford en los meses de invierno



- La forma de la nube indica una relación negativa muy fuerte entre consumo diario de energía y temperatura diaria

Coeficiente de correlación

- Una forma de cuantificar la "fuerza" de la relación **(lineal)** existente entre dos variables es mediante el coeficiente de correlación, cuya fórmula es:

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

- r está siempre comprendido entre -1 y +1
- Los valores extremos, -1 y +1, sólo se alcanzan si existe una relación lineal exacta entre la Y y la X
- Cuando no existe relación alguna entre dos variables, el coeficiente de correlación es igual a cero. (En la práctica en una muestra de dos variables independientes r(X,Y) será "cercano" a cero, pues debido al azar del muestreo r fluctuará algo alrededor de su verdadero valor poblacional)
- En los dos ejemplos vistos:
  - r(Peso, Estatura) = + 0.74
  - r(Consumo, Temper) = -0.97

EJEMPLO

- En el archivo Excel granizo.xls se recogen datos sobre más de 300 nubes que fueron controladas a lo largo de varios años desde la estación meteorológica de Silla dentro de las campañas de lucha antigranizo mediante generadores de IAg realizadas por el Servicio de Defensa contra Plagas. De cada nube se recogen los siguientes datos:
- MES y AÑO de la observación
- RIESGO: 1 ó 0 según se hubiera o no dado aviso de riesgo a los encargados de encender los generadores
- INTENSIDAD: densidad de la nube (escala de 1 a 8)
- ALTURA: altura en metros de la zona de mayor densidad de la nube
- GRANIZO: indica si la nube finalmente provocó granizo
- DUREZA: dureza del granizo codificada como: blando, duro o mixto. En blanco para las nubes que no granizaron
- TAMAÑO: codificado como: lenteja, garbanzo, cereza o nuez. En blanco para las nubes que no granizaron
- DAÑOS: indica si el granizo produjo o no daños en la agricultura. En blanco para las nubes que no granizaron
- ALTURACOD: altura codificada en 4 tramos

Ejemplo (continuación)

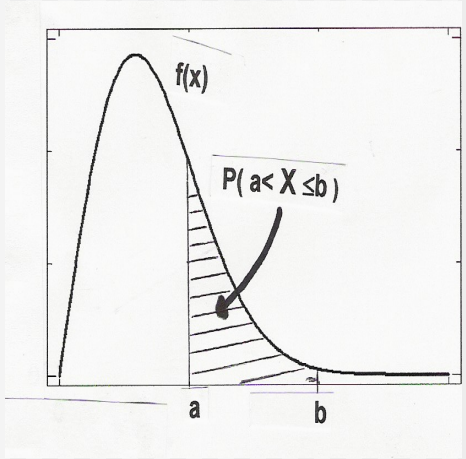
- Los asistentes utilizarán estos datos para practicar las diversas herramientas de Estadística Descriptiva vistas en el curso. Se plantean a continuación algunas cuestiones a analizar.
- 1. ¿Qué meses son los más conflictivos respecto al granizo?
- 2. ¿La variable ALTURA se adapta bien al modelo de la distribución Normal? ¿Por qué?
- 3. ¿Es aconsejable transformar la variable para normalizarla? ¿Con qué transformación? ¿Por qué?
- 4. Ésta es una pregunta para pensar (para los aficionados a las matemáticas): cuando en un estudio estadístico de comparación de dos poblaciones se ha aplicado una transformación logarítmica a los datos y se obtiene que hay diferencias significativas entre las medias de la variable transformada en las dos poblaciones, ¿se concluye también que hay diferencias significativas entre las medias o entre las medianas en las dos poblaciones de la variable primitiva?
- 5. ¿Existe alguna relación entre el tamaño y la dureza del granizo? ¿Y entre estas dos variables y la probabilidad de que el granizo produzca daños?
- 6. ¿Existe relación entre la altura de la nube y la probabilidad de que al granizar produzca daño? ¿Qué gráfico o tabla consideras más adecuado para exponer la conclusión obtenida?
- 7. El efecto de la altura de la nube sobre la probabilidad de daños del granizo ¿se debe a que la altura de la nube influye sobre el tamaño o sobre la dureza del granizo?
- 7. ¿Qué tal funcionaron las predicciones de riesgo?

Caracterización de variables aleatorias teóricas

- Como se ha indicado una variable aleatoria es de tipo **discreto** si su conjunto de "valores" posibles en la población es finito, o infinito numerable (Ejemplo: un parásito tratado con un plaguicida puede morir o no morir; el número de ceratitis capturado en una trampa puede ser 0, 1, 2, …)
- A la proporción de individuos en la población (expresada en tanto por uno) que tienen un determinado valor de la variable se le llama **probabilidad** de ese valor (Ejemplo: la probabilidad de que un parásito tratado muera es es 0.20, si el 20% de los parásitos mueren; la probabilidad de que el número de ceratitis (en el mosquero o trampa) sea 0 es 0.30 si el 30% de las trampas no tienen ninguna mosca)
- Una variable aleatoria discreta viene caracterizada por las probabilidades de los diferentes "valores" posibles (probabilidades cuya suma será siempre igual a 1)

Caracterización de variables aleatorias continuas

- Como se ha indicado, una variable aleatoria es de tipo **continuo** si su conjunto de "valores" posibles en la población es un infinito continuo.
- Una variable aleatoria continua viene caracterizada por su **función de densidad f(x)**



- La función de densidad de una variable aleatoria tiene la propiedad de que el área comprendida bajo dicha función entre dos valores "a" y "b", coincide con la probabilidad de que la variable aleatoria tome valores en dicho intervalo

Esperanza Matemática

- La media aritmética, o promedio, de un conjunto de valores es un concepto claramente intuitivo. Su idealización matemática conduce al concepto de **Esperanza Matemática**, o **Valor Medio**, de una función h(**X**) de una determinada variable aleatoria **X**
- **Variables discretas:**
  - En el caso de que X sea una variable aleatoria discreta, la esperanza matemática, o valor medio, de una función h(X) de dicha variable aleatoria se expresará como E(h(X)) y se define como:

$$E(h(X)) = \sum_i h(x_i)P(X = x_i)$$

  - sumatorio que se extiende para todos los valores x_i cuyas probabilidades no sean nulas.
- **Variables continuas:**
  - Si la variable X es continua, en la expresión del valor medio hay que sustituir la probabilidad $P(X=x_i)$ por el valor f(x) de la función de densidad en x, y el sumatorio por una integral definida

$$E\left(h(X)\right) = \int_{-\infty}^{+\infty} h(x)f(x)dx$$

Parámetros de una variable aleatoria

- **Parámetros de posición**
  - **Media:** se define como $m = E(\mathbf{X})$
  - **Mediana:** es un valor *med* que cumple $P(\mathbf{X}>med) = P(\mathbf{X}<med)$
  - **Cuartiles**: **C1** valor tal que $P(\mathbf{X}<C1) = 0.25$; **C3** valor tal que $P(\mathbf{X}>C3) = 0.25$

- **Parámetros de dispersión**
  - **Varianza:** se define como $\sigma^2 = E(\mathbf{X}-m)^2$
  - **Desviación típica:** se define como la raíz cuadrada σ de la varianza

- **Parámetros de forma:**
  - **Coeficiente de asimetría:** se define como $E(\mathbf{X}-m)^3/\sigma^3$
  - **Coeficiente de curtosis:** se define como $E(\mathbf{X}-m)^4/\sigma^4 - 3$

Modelos de distribuciones discretas: Distribución Binomial

La distribución Binomial

- Se define como el número **X** de veces que se presenta un suceso de probabilidad **p** en **N** individuos seleccionados al azar de la población (por ejemplo: número de frutas podridas en cajas de 20 unidades)
- Los valores posibles de **x** son: 0, 1, 2, …, N
- El valor medio de **X** es Np y la desviación típica $\sigma = (Np(1-p))^{1/2}$
- La probabilidad de que **X** tome un valor concreto x (x=0,…,N) es:

$$\mathrm{Prob}(X = x) = \binom{N}{x} p^x (1-p)^{N-x}$$

- <u>Nota 1</u>: la distribución binomial asume que el valor de **p** es el mismo para los N individuos. Si **p** varía de unos individuos a otros la σ puede ser muy inferior a la de una binomial con la misma **p** promedia (Ejemplo)
- <u>Nota 2</u>: Si se observan varios valores (p.e. varias cajas) y p varía de unas a otras, la σ observada será mayor que la de una Binomial

Transformación de datos binomiales

- En muchos experimentos los resultados en cada unidad son variables binomiales o derivadas de éstas (como porcentajes)
- Aunque existen técnicas estadísticas específicas para estos casos, a veces se prefiere utilizar técnicas estadísticas estándar, como el Análisis de la Varianza, que asumen la normalidad de los datos.
- Aunque una variable binomial X tiende en ciertas condiciones a distribuirse normalmente, se plantea el problema de que si un factor afecta a la media de X (que depende de p) también afectará a su varianza (que también depende de p) incumpliéndose la hipótesis de *homocedasticidad.*
- Para evitar este problema se aconseja utilizar como variable a analizar a la transformada $Y = arc\ sen\ (X)^{1/2}$ donde X viene expresada en tanto por uno.
- Se demuestra, en efecto, que si X es Binom(N,p) la varianza de Y casi no depende de p (aunque sí que depende de N)

Modelos de distribuciones discretas: Distribución de Poisson

- La distribución de Poisson se describe como el límite al que tiende la distribución de una variable binomial cuando n➔∞ y p➔0 de forma que el producto np tiende a una constante λ. (Ejemplo: número de ceratitis capturadas en una trampa una semana)
- Valores posibles: 0, 1, 2, … (no hay un N límite predeterminado)
- Media m = λ Desviación típica $\sigma = \lambda^{1/2}$
- Probabilidad de un valor concreto:

$$P\left(\text{Poisson}(\lambda) = x\right) = e^{-\lambda}\frac{\lambda^{x}}{x!}$$

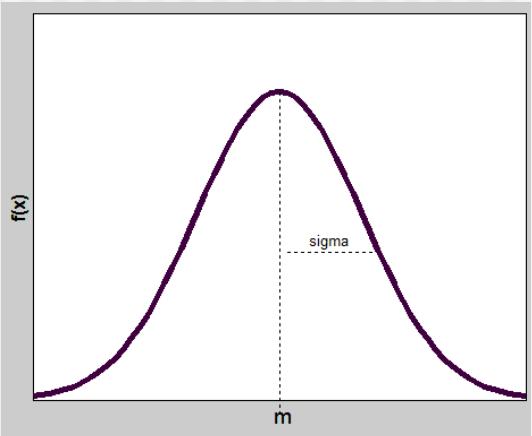
- <u>Nota</u>: Si se observan varios valores (p.e. varias trampas) y λ varía de unos casos a otros, la σ observada será mayor que la de una Poisson

Transformación de datos de Poisson

- En muchos experimentos los resultados en cada unidad son conteos que pueden tratarse con variables de Poisson
- Aunque existen técnicas estadísticas específicas para estos casos, a veces se prefiere utilizar técnicas estadísticas estándar, como el Análisis de la Varianza, que asumen la normalidad de los datos.
- Aunque una variable de Poisson X tiende a distribuirse de forma normal si λ es grande, se plantea el problema de que al ser iguales la media y la varianza de X si un factor afecta a la media también afectará a su varianza incumpliéndose la hipótesis de *homocedasticidad.*
- Para evitar este problema se aconseja utilizar como variable a analizar a la transformada $Y = (X)^{1/2}$ cuya varianza es aproximadamente igual a 0.25 sea cual sea λ

La distribución Normal

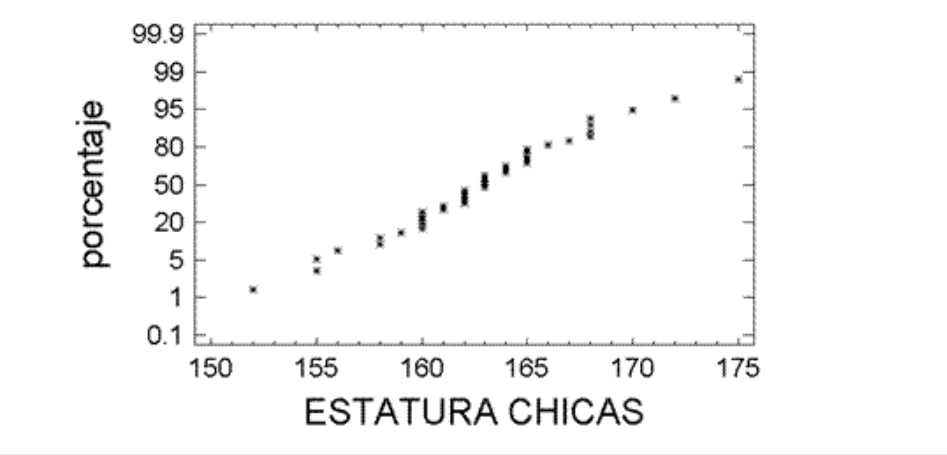
- La distribución Normal constituye el modelo matemático más importante para el análisis de variables aleatorias continuas (Otras distribuciones importantes en Análisis de Supervivencia se presentarán en la 3ª Parte del curso).
- La distribución Normal viene completamente caracterizada por su **media m** y su **desviación típica σ.** El hecho de que una variable aleatoria **X** sigue una distribución Normal con unas determinadas **m** y σ, se expresa como **X** ~ N(m,σ). A la distribución **N(0,1)** se le denomina distribución normal tipificada, viniendo tabuladas sus probabilidades.
- La función de densidad de una variable normal tiene forma de una campana simétrica en torno al valor m y cuya anchura depende σ



- Cuando se selecciona al azar una muestra de una población normal, el histograma de los datos tiene aproximadamente una forma de campana, parecida a la de dicha función de densidad.
- En estas muestras, los **coeficientes de asimetría y de curtosis estandarizados** tendrán valores cercanos a cero (generalmente comprendidos entre -2 y +2)

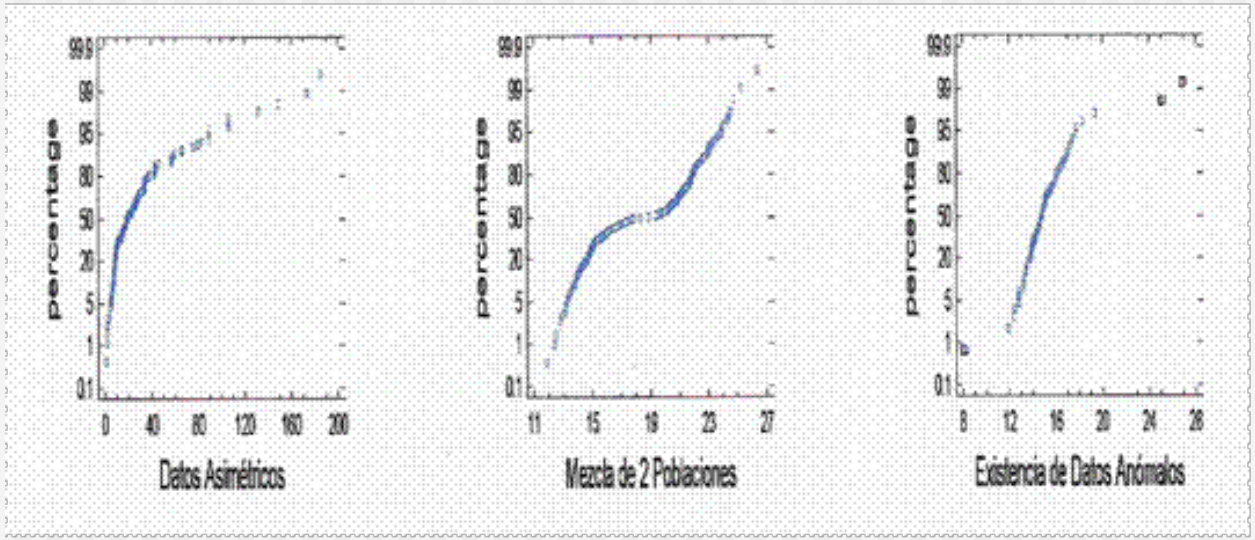
Representación de datos en papel probabilístico normal

- La mejor forma de comprobar si la distribución Normal es un modelo adecuado para unos datos, es representarlos en papel probabilístico normal
- Una representación de un conjunto de datos en papel probabilístico normal hace corresponder a cada observación un punto en el plano.
  - La abscisa del punto es el valor observado
  - La ordenada del punto corresponde al porcentaje de valores en la muestra que son menores o iguales que el considerado.
- La escala vertical del papel probabilístico está modificada, de forma que cuando los datos proceden de una distribución normal los puntos correspondientes se sitúan aproximadamente a lo largo de una recta.



Formas posibles con datos no normales

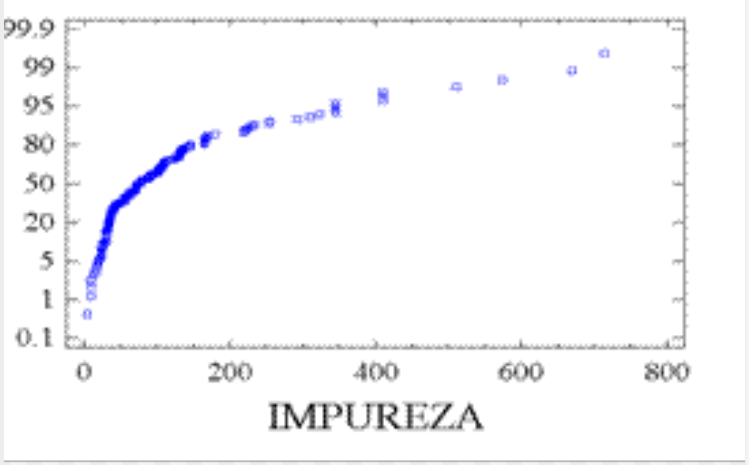
- Si la representación de los datos difiere claramente de una línea recta ello es una prueba de que la población muestreada no se adapta bien al modelo de la distribución Normal.

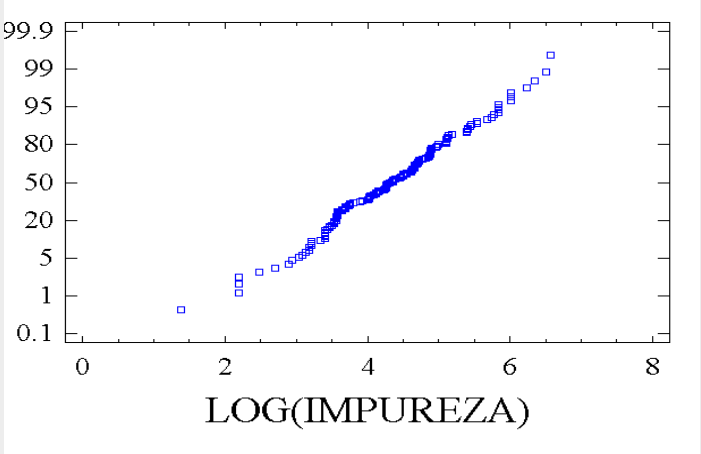


- Ejemplo 1: Datos con marcada asimetría positiva
- Ejemplo 2: datos procedentes de la mezcla de dos poblaciones
- Ejemplo 3: presencia de tres datos anómalos

Transformaciones

- En ciertos casos en los que los datos presentan una clara **asimetría** positiva que no hace adecuado el modelo Normal, frecuentemente el **logaritmo de los datos se ajusta mejor a una distribución normal**
- En la siguiente figura se recoge la representación en papel probabilístico normal del contenido en impurezas en un proceso químico y del logaritmo de dicho contenido





Los modelos matemáticos en el método científico

- Todo modelo matemático postulado para el estudio de un fenómeno real (sea la distribución normal, el péndulo matemático, o los rectángulos) implica necesariamente una simplificación y abstracción de ésta.
- ¡Carece por tanto de sentido la cuestión de si un determinado modelo es "verdadero" o "correcto" para aplicarlo a una realidad concreta!

Todos los modelos son **falsos**

pero …

Algunos modelos son **útiles**

Para que un modelo sea **útil** debe recoger de forma “razonablemente aproximada” los aspectos más relevantes de la realidad que se pretende analizar

Para usar correctamente un modelo estadístico es importante conocer hasta qué punto las conclusiones que se obtienen de los mismos son **“sensibles”** a graves incumplimientos de las hipótesis en las que se basa el modelo

Estudio de la normalidad de los datos

- Existen diversos tests estadísticos formales para estudiar la normalidad de unos datos (test Gi-dos, test de Kolmogorov, test de Saphiro, etc.).
- En nuestra opinión estos tests son de dudosa utilidad práctica, constituyendo **"la respuesta correcta a una pregunta equivocada".**
- En efecto, la pregunta a la que pretenden responder: ¿proceden mis datos **realmente** de una distribución normal teórica? carece de sentido.
- Sin embargo, sí que es importante plantearse si la pauta de variabilidad existente en los datos es "lo suficientemente parecida" a la postulada por el modelo, como para que la aplicación de éste conduzca a resultados "válidos" o "útiles" en la práctica.
- Con esta finalidad, en vez de usar estos tests estadísticos, es más útil recurrir a técnicas descriptivas como:
  - la construcción de un histograma (si hay bastantes datos),
  - el cálculo de los coeficientes de asimetría y curtosis
  - y, sobre todo, la representación de los datos en papel probabilístico normal
- Esta última es especialmente útil para detectar problemas como la presencias de datos anómalos (“outliers”), la mezcla de poblaciones y la asimetría marcada de los datos.

Distribuciones en el muestreo

- Sea una población a cuyos individuos va asociada una variable aleatoria X. Para obtener conclusiones sobre la distribución de esta variable se obtiene una muestra aleatoria constituida por N individuos de la población.
- De una misma población es posible "a priori" obtener una gran cantidad de muestras diferentes de tamaño N. Existe por tanto una población de posibles muestras, o sea una **nueva población cuyos individuos son dichas muestras**.
- A cada individuo de esta nueva población se le pueden hacer corresponder diferentes características numéricas (por ejemplo: la media muestral o la desviación típica muestral) que serán por tanto nuevas variables aleatorias.
- Cualquier función de los valores muestrales (“estadístico”) es una variable aleatoria, cuya distribución dependerá en general de la distribución de la población y del tamaño de la muestra.

Propiedades de la media y la varianza muestral

- Una variable aleatoria **X** se distribuye en una población con media **m** y varianza σ^2.
- De dicha población se toma una muestra (aleatoria simple) de tamaño N. La media muestral (x_{bar}) y la varianza muestral (s^2) son valores observados de dos variables aleatorias (definidas sobre la población de todas las posibles muestras). **Es importante entender bien lo que esto quiere decir** (discutirlo)
- Se cumplen siempre tres propiedades fundamentales:

$$E\left(\bar{x}\right) = m$$

$$Var\left(\bar{x}\right) = \frac{\sigma^2}{N}$$

$$E\left(s^2\right) = \sigma^2$$

- Además la media muestral tiende a distribuirse normalmente (al aumentar el tamaño de la muestra) aunque la población muestreada no sea Normal

Muestreo de poblaciones normales

- Si la población muestreada es Normal, la media muestral se distribuye normalmente (sea cual sea N)
- En el muestreo de poblaciones normales aparecen tres nuevas distribuciones de gran importancia en Inferencia Estadística:
- La distribución Gi-2 de Pearson
- La distribución t de Student
- La distribución F de Fisher
- Estas distribuciones se irán presentando cuando aparezcan en las siguientes sesiones

Ejercicio sobre distribuciones muestrales

- En el archivo Excel *ejemplo_muestreo.xls* se recogen 480 valores generados al azar de una variable aleatoria de media 50 y desviación típica 5. La variable no se distribuye normalmente (comprobar sus coeficientes de asimetría y curtosis)
- Se supone que se han tomado al azar 30 muestras de tamaño 4 de dicha variable. Para ello la columna *muestra_4* recoge la muestra (de 1 a 30) a la que pertenece cada uno de los 120 primeros datos.
- Se supone también que se han tomado al azar 30 muestras de tamaño 16 de dicha variable. Para ello la columna *muestra_16* recoge la muestra (de 1 a 30) a la que pertenece cada uno de los 480 datos.
- Calcular las 30 medias y 30 varianzas de las muestras de tamaño 4 (llamarlas medias_4 y varianzas_4) y así mismo las 30 medias y 30 varianzas de las muestras de tamaño 16 (llamarlas medias_16 y varianzas_16)
- Comprobar que se cumplen aproximadamente las propiedades enunciadas sobre la distribución de la media y la varianza muestral, según se indica en la siguiente diapositiva

Ejercicio sobre distribuciones muestrales (continuación)

- Comprobar que la media de las medias muestrales coincide (aproximadamente) con la media de la población (50), tanto en las muestras pequeñas (tamaño 4) como en las grandes (tamaño 16)
- Comprobar que la desviación típica de las medias muestrales en torno a su media es mayor en las muestras pequeñas que en las grandes. ¿Cuánto mayor debería haber sido teóricamente en este caso?
- Comprobar que aunque la población muestreada no es normal, las medias muestrales tienden a distribuirse normalmente a medida que aumenta el tamaño de la muestra (analizar los coeficientes de asimetría y curtosis de los datos y de medias_4 y medias_16)
- ¿Se cumple también, no siendo los datos normales, que en promedio las varianzas muestrales coinciden con la varianza poblacional, tanto en muestras grandes como en pequeñas?
- La desviación típica de las varianzas muestrales ¿también disminuye al aumentar el tamaño de la muestra?
- ¿Tienden también a distribuirse normalmente las varianzas muestrales al aumentar el tamaño de la muestra?