Apuntes
IG23 Ampliación de Estadística

Ingeniería
técnica en
Informática de Gestión

Curso 2003/04
Irene Epifanio

LÉEME

El material de este curso lo componen:

-  Apuntes de teoría y problemas
-  Hojas de problemas
-  Hojas de sesiones de prácticas con el ordenador
-  Ejercicios de autoevaluación resueltos
-  Otro tipo de material: formulario, tablas, bibliografía, información complementaria
-  PERO SOBRE TODO: acudir y atender a las clases, y a las tutorías cuando se necesiten

- Los apuntes tienen como objetivo facilitar el seguimiento de las clases, sirviendo de apoyo a éstas, puesto que recogen los contenidos básicos del curso. Se trata de que la atención se centre en las explicaciones de clase y que el hecho de tomar notas no la dificulte. Se ha intentado presentar los conceptos de forma simple, siempre ayudándose de ejemplos-problemas, evitando las demostraciones que pueden encontrarse en la bibliografía o en tutorías. Se trata de que las herramientas estadísticas se entiendan, se sepan utilizar e interpretar, ya que su uso va a ser completamente aplicado. Estos apuntes van ligados a las clases, no se entienden como algo separado de ellas, pues **las clases** son, en realidad, **esenciales**.
- Los problemas aparecerán por un lado en las clases de teoría, donde se resolverán los marcados con  . Por otro lado, aparecen en las hojas de problemas de cada tema que se dejarán para que los trabajéis en casa: recordad que vale más resolver uno por vosotros mismos que copiar mil. También contaréis con una serie de problemas resueltos y de nuevo, podéis recurrir a la bibliografía o a tutorías si queréis afianzar más algún tema.
- En las prácticas, se trabajarán conceptos vistos en clase de teoría-problemas con la ayuda del ordenador (el papel del ordenador es fundamental, como se verá) y otros conceptos nuevos. Todas las prácticas están constituidas por una primera parte de explicaciones de

comandos y conceptos y otra parte con los problemas a resolver (con datos diferentes para cada persona). Es **muy recomendable** asistir a las prácticas, habiéndose leído con anterioridad las hojas de prácticas y habiendo repasado también los conceptos que se vayan a tratar en esa sesión.

- Las ejercicios de autoevaluación pretenden fomentar el estudio diario, ya que la continuidad es fundamental. Asimismo, los ejercicios de autoevaluación (que también son ejemplos del tipo de preguntas con las que se os evaluará) os permitirán comprobar vuestros progresos y detectar posibles errores, en cuyo caso ya sabéis que contáis con las tutorías.

- En mi página ( hhtp://www3.uji.es/∼epifanio, curs 2003-04, primer semestre) podréis encontrar diversa información y el material disponible en cada momento.

- Si habéis llegado a leer hasta aquí (thanks!), habréis comprobado que la palabra más repetida ha sido: TUTORÍAS. En cuanto tengáis cualquier duda, no *dudéis* en acudir a solucionarla, no dejéis que "se enrede la madeja". Mi deseo es ayudaros y las tutorías están para eso.

- Confío en que me digáis aquello que encontréis mejorable (cualquier sugerencia será bienvenida!), para ello también os pasaré un cuestionario sobre vuestra opinión en diversos aspectos.

- De verdad espero que todo esto os sirva de ayuda y sobre todo que no nos tengamos que ver las caras en septiembre próximo ... Ahora abrochémonos el cinturón y ¡al ataquerrrrr! ☺

Castellón, septiembre 2003

Irene Epifanio

Lista de Símbolos

-  : ¡Ojo!, observación
-  Definición (a leer)
-  ¡Alto! esto es muy importante
-  Ejemplo-problema hecho en clase (a escribir se ha dicho)
-  Información complementaria
-  Sesión de prácticas con el ordenador
-  Se verá más adelante
-  Ya se ha visto (a recordarlo)
-  Dificultad
-  Nota
-  Curvas peligrosas donde no debes estrellarte: ir con cuidado

Tema 0. Repaso previo.

Este es un tema de repaso que recapitula las ideas básicas que ya se vieron previamente el curso anterior en IG12 Estadística ⏮. Sólo pretende refrescar la memoria, centrar ideas y tener una visión global. ℹ Los libros de la bibliografía también consideran esta parte, además en mi página hhtp://www3.uji.es/∼epifanio, curs 2002-03, podéis encontrar material más detallado que el presente tema, aunque está orientado a la titulación de Enginyeria tècnica en Disseny Industrial.

Si tenéis dudas sobre este tema que os impidan continuar la marcha del presente curso, lo mejor es que acudáis a tutorías de inmediato!.

0.1. Introducción

A continuación, se presentan varios ejemplos del tipo de problemas que seremos capaces de resolver al final del curso (espero!).

Ejemplo 0.1. En una empresa, se realiza diariamente un control sobre el número de intentos de acceso fraudulentos a cuentas de los trabajadores de la empresa. El control se realiza a partir de una muestra de 500 intentos de acceso, seleccionados aleatoriamente del total de intentos de acceso diario. Los intentos de acceso se clasifican sencillamente en "buenos" o "malos" según si la contraseña escrita al intentar acceder es correcta o no. En teoría se considera que la tasa de intentos de acceso fraudulentos no ha de superar el 2 % del total de intentos. Supongamos que hoy, de los 500 intentos de acceso de la muestra, 12 han sido fraudulentos, es decir, un 2.4 %. ¿Tenemos motivos suficientes para sospechar que alguien está intentando acceder fraudulentamente al sistema o se debe únicamente al azar?

Ejemplo 0.2. Estamos interesados en comparar los tiempos de ejecución de 5 algoritmos de ordenación (*algoritmo de la burbuja, de selección, de inserción, quicksort, tree sort*) para un cierto tipo de datos de un tamaño determinado y con un cierto grado de desorden. Para ello, consideramos diversos conjuntos de entrada de entre los que estamos interesados y obtenemos el tiempo de CPU de ejecución con cada algoritmo. Algunas preguntas que querríamos contestar en base a los resultados obtenidos podrían ser:

¿Existe diferencia significativa entre los 5 algoritmos? ¿Hay un algoritmo mucho mejor que los otros? ¿Pueden clasificarse los algoritmos en diversos grupos homogéneos en el sentido que dentro de cada grupo no difieran significativamente?

El problema podría complicarse si, por ejemplo, el tamaño de los datos a ordenar o el grado

de desorden no son siempre los mismos, entonces deberíamos plantear un modelo adecuado al problema.

Ejemplo 0.3. Se pretende diseñar un ratón ergonómico para niños de 7 a 9 años. Hemos de conocer la forma de su mano derecha por lo que hemos de tomar distintos datos antropométricos de un conjunto de niños. Supongamos que estamos interesados en la longitud de su dedo índice. Realizamos un estudio piloto con 30 niños, de los que obtenemos una media de 6 cm y una desviación típica de 0.4 cm. Si deseamos poder afirmar con un 95 % de confianza que la media es imprecisa como mucho en 0.1 cm, ¿cuántos datos deberíamos tomar? Una vez tomados, podríamos calcular un intervalo de confianza al 95 % para la media.

Ejemplo 0.4. Este ejemplo se escapa de los objetivos del curso, pero muestra otro tipo de problemas que pueden resolverse utilizando la Estadística [1]. Desearíamos diseñar un detector automático de correo basura (*SPAM*), de forma que se filtrara este correo antes que colapsara los buzones de los usuarios. Utilizando la información de 4601 e-mails, se intentará predecir si un nuevo correo electrónico es correo basura o no, de manera automática. Por ejemplo, variables que podrían sernos útiles serían el porcentaje de aparición de determinadas palabras, como puede ser: "free", "our", "Irene", etc. Al final se podrían obtener (mediante métodos que no veremos) reglas como:

$$\begin{array}{lll} \text{si } (\,\%\text{ Irene} < 0.6) \;\&\; (\,\%\text{ our} > 1.5) & \text{entonces} & \textit{spam} \\ & \text{si no} & \textit{e-mail} \end{array}$$

Veamos ahora de qué se encarga la Estadística. La ciencia Estadística tiene un doble objetivo:

- La generación y recopilación de datos que contengan información relevante sobre un determinado problema (Muestreo).
- El análisis de dichos datos con el fin de extraer de ellos dicha información. El primer paso en el análisis de los datos consistirá en describirlos a través de ciertas medidas y gráficas, lo cual nos facilitará su comprensión (Estadística descriptiva). Sin embrago, buscamos ir más allá y poder sacar conclusiones basadas en dichos datos. Para ello, podremos recurrir a plantear un modelo matemático (teoría de la probabilidad) que nos permitirá después extraer las conclusiones que nos interesan (Inferencia estadística).

Por tanto, un modelo estadístico constará de varias partes: a) Muestreo (apartado 0.7), b) Estadística descriptiva (apartado 0.2), c) Confección de un modelo matemático (teoría probabilidad)(apartados 0.4,0.5,0.6), d) Inferencia estadística (este curso). Esta última parte (d), se considerará en este curso, mientras que las restantes se han tratado en la asignatura IG12 Estadística (o F04 para los procedentes del viejo plan de estudios).

En resumen, la **Estadística**: estudia los métodos científicos para recoger (hacer un muestreo), organizar, resumir y analizar datos (estadística descriptiva), así como para obtener conclusiones

[1]  "The Elements of Statistical Learning. Data mining, Inference and Prediction."Hastie, Tibshirani, Friedman.

válidas (inferencia estadística) y tomar decisiones razonables basadas en tal análisis.

Así, en el ejemplo 0.2, primero tomamos una muestra aleatoria de entre TODOS los archivos de ese tipo (tamaño y grado de desorden), obtenemos los tiempos de ejecución con cada algoritmo, después se describirían (medias, varianzas, gráficos, ...), se propondría un modelo adecuado y obtendríamos las conclusiones de interés (respuestas a las preguntas planteadas).

Repasemos ahora algunos conceptos básicos:


Población: conjunto de todos los individuos que son objeto de estudio y sobre los que queremos obtener ciertas conclusiones. **Ejemplos:**

- Todos los niños entre 7 y 9 años (ejemplo 0.3).
- Todos los mails recibidos y por recibir (ejemplo 0.4).

Como puede verse, a veces las poblaciones existen físicamente y son finitas aunque muy grandes, en cambio otras veces la población es de carácter abstracto. En general, en lugar de hacer un estudio de todos los elementos que componen la población (hacer un **censo**), se escoge un conjunto más reducido.


Muestra: es un subconjunto, una parte de la población que seleccionamos para un estudio.

Es deseable que la muestra extraída "se parezca" a la población, es decir, "que sea como la población pero en tamaño reducido". El objetivo es que la muestra sea representativa de la población. Notemos que si la muestra es mala, las conclusiones extraídas no serán válidas, podrían ser erróneas.


Ejemplo: si para obtener medidas para el ejemplo 0.3 acudiéramos a un entrenamiento de baloncesto de niños entre 10 a 11 años, ¿obtendríamos una muestra representativa de la población o sesgada?


Tamaño muestral: es el número de observaciones de la muestra, N.


Variable aleatoria: es una característica aleatoria que podemos expresar numéricamente, es la característica que estamos midiendo en cada individuo. Una característica aleatoria será una característica que tomará un valor para cada individuo.

Las variables aleatorias las denotaremos con letras mayúsculas: X, Y, ...

Las variables aleatorias pueden clasificarse en:

-  Cualitativas o categóricas: expresan una cualidad
-  Cuantitativas: tienen propiamente carácter numérico

Variables cualitativas. Las variables cualitativas a su vez se subdividen en: ordinales o no ordinales, según si las categorías pueden o no disponerse bajo un orden con sentido.

Ejemplos de variables cualitativas no ordinales:

- Distribución de linux: 1 = Red Hat, 2 = Suse, 3 = Debian, 4 = Otras
- Mail: 1 = SPAM, 0 = No SPAM
- Sexo de una persona: 1 = Mujer, 2 = Hombre
- Adicción al tabaco: 1 = Fuma, 2 = No fuma
- Tipo de defectos de un frigorífico defectuoso: 1 = Termostato, 2 = Compresor, 3 = Motor, 4 = Cableado, 5 = Revestimiento, 6 = Otros
- **Ejemplo 0.5:** los alumnos de 3^o ETIG quieren irse de viaje de fin de curso para celebrar que han aprobado y para sacarse unos euros deciden vender gorras. Quieren conocer el color preferido por los compradores potenciales, por tanto, les interesa la variable aleatoria: "Color de la gorra preferido por los miembros de la UJI", con posibles valores: 1 = Negro, 2 = Blanco, 3 = Rojo, 4 = Otros.

Ejemplos de variables cualitativas ordinales:

- Interés sobre una determinada materia: 1 = Bajo, 2 = Medio, 3 = Alto
- Cualquiera de las de la encuesta de evaluación: 1 = Muy desfavorable, 2 = Desfavorable, 3 = Indiferente, 4 = Favorable, 5 = Muy favorable

Las **variables cuantitativas** también se dividen en dos:

- Discretas: toman valores discretos, es decir, en un conjunto numerable (podemos contar los posibles valores que pueden adoptar). Existen "espacios"entre los posibles valores que puede adoptar la variable.
- Continuas: como indica su nombre, toman valores en un conjunto no numerable. "Los valores que adoptan estas variables, pueden estar tan cercanos como se quiera".

Ejemplos de variables discretas:

1. Número de piezas defectuosas en un lote de 100 piezas
2. Número de caras obtenidas al lanzar una moneda 20 veces
3. Número de $5's$ al lanzar un dado 60 veces

En los tres casos anteriores los valores que pueden adoptar son finitos: en a) de 0 a 100, en b) de 0 a 20, en c) de 0 a 60. Sin embargo, podría no ser así, podría adoptar valores discretos no limitados:

1. Número de errores en una superficie de grabación magnética
2. Número de mensajes que llegan a un servidor en una hora

3. Número de manchas de más de $1mm^2$ en una lámina
4. Número de defectos en 2m de cable
5. Número de veces al mes que va al cine un estudiante de ETIG

Ejemplos de variables continuas:

1. **Ejemplo 0.2:** tiempo de ejecución del algoritmo de la burbuja para el tipo de archivos considerado
2. **Ejemplo 0.3:** longitud de la mano de niños de 7 a 9 años
3. Peso de ciertas piezas
4. Tiempo de vida (duración) de ciertos motores
5. Dureza de cierto material
6. Resistencia de cierto producto
7. Notas de estudiantes de ETIG
8. Euros gastados con el móvil en un mes por un estudiante de la UJI

Observación: *La distinción entre variables continuas y discretas no es rígida. Las variables continuas anteriores corresponden a medidas físicas que siempre pueden ser redondeadas, por ejemplo, la longitud podemos medirla hasta el milímetro más cercano o el peso hasta el gramo más cercano. Aunque estrictamente hablando, la escala de dichas medidas sea discreta, las consideraremos continuas como una aproximación a la verdadera escala de medida.*

Resumiendo, las variables aleatorias pueden ser:

1. Categóricas o cualitativas
   a) No ordinales
   b) Ordinales
2. Cuantitativas
   a) Discretas
   b) Continuas

0.2. Descripción de una muestra

Para describir una muestra, podemos valernos de tablas de frecuencias, de métodos gráficos (histogramas, diagramas de cajas, etc.) y medidas descriptivas.

Recordémoslas brevemente, ayudándonos del ejemplo sobre las notas que ya hemos presentado con el programa.

Ejemplo 0.6. Tabla de frecuencias de las notas de Estadística de ETDI en junio de 2001.

Intervalo (límites de clase)	(Marca de clase)	Frecuencia absoluta	Frecuencia relativa	Frecuencia acumulada	Frecuencia relativa acumuluada
[0, 2.5)		9			
[2.5, 5)		21			
[5, 7.5)		63			
[7.5, 10]		46			

Frecuencia absoluta: número de observaciones en el intervalo

Frecuencia relativa: número de observaciones en el intervalo / tamaño muestral; suma 1; indica el porcentaje de observaciones en el intervalo

Frecuencia acumulada: suma de las frecuencias de los intervalos anteriores, incluyendo el actual. Indica el número de observaciones por debajo del extremo superior de la clase. Obviamente, el último valor es el tamaño muestral.

Frecuencia relativa acumulada: frecuencia acumulada / tamaño muestral. Indica el porcentaje muestral por debajo del extremo superior de la clase. El último valor será 1 (100 %).

Normalmente, las clases son de igual anchura, pero podrían no serlo:

Intervalo	Frec. abs.	Frec. rel.	Frec. acum.	Frec. rel. acum.
[0, 5)	30			
[5, 7)	44			
[7, 8.5)	39			
[8.5,10]	26			

Los gráficos nos permiten también ilustrar la distribución de los datos.

Histograma: pueden ser de frecuencias absolutas, relativas, acumuladas o relativas acumuladas, según que represente la altura de la barra.

Ejemplo 0.6.:

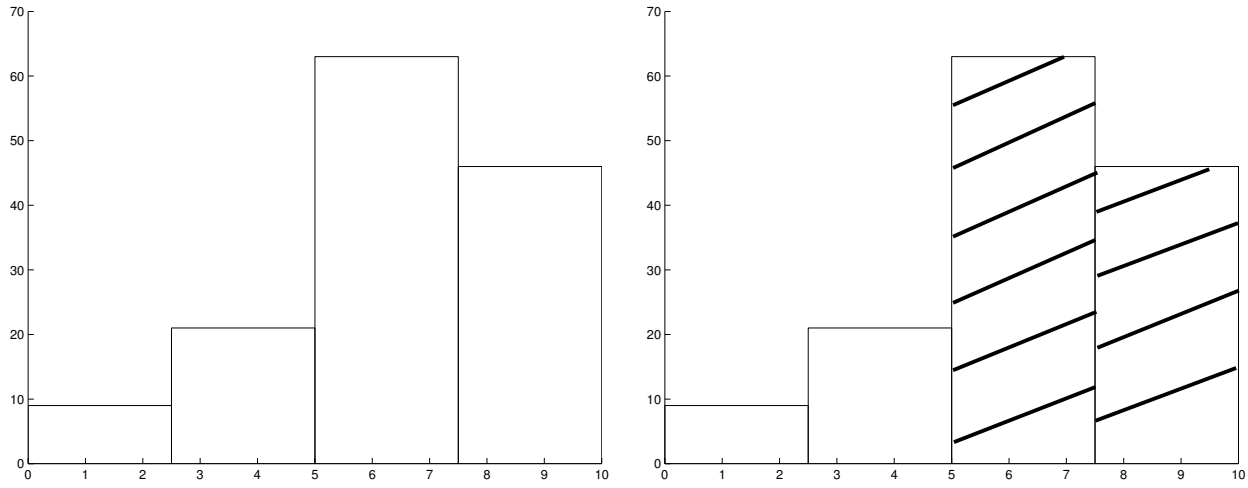


Figura 1: Histograma de frecuencias absolutas del ejemplo 0.6.

Los histogramas nos muestran como se distribuyen (como se reparten) los datos, las cimas de las barras indican la forma de la distribución. Además, el área de cada barra es proporcional a la correspondiente frecuencia. **Ejemplo:** el área rayada del ejemplo anterior es el 78.4 % del área total de todas las barras, por tanto, el 78.4 % de las notas están en las correspondientes clases, o sea, el 78.4 % de las notas están entre 5 (inclusive) y 10.

Hay muchos más métodos gráficos: diagramas de barra, de sectores, polígonos de frecuencias, diagrama de cajas (*boxplot*), Pareto,...

Además de las gráficas, otra forma de resumir los datos es mediante parámetros numéricos, que podemos dividir en:

- Medidas de posición o centrales: dan cuenta de la posición de las observaciones
- Medidas de dispersión: indican la dispersión (variabilidad) de los datos
- Medidas de forma: miden la forma de distribuirse los datos

Medidas de posición: media, mediana, moda y percentil.

Media: si tenemos una muestra $\{x_1, x_2, ..., x_N\}$,

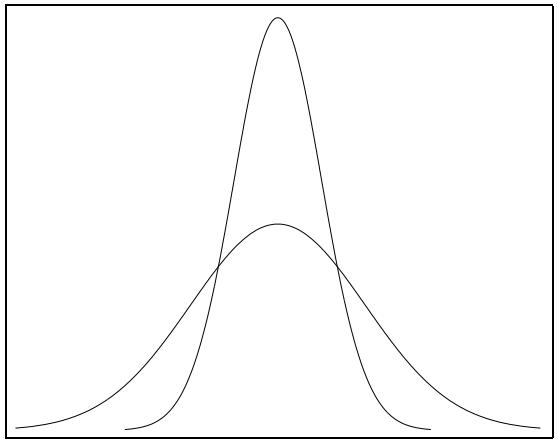
$$\overline{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}. \tag{1}$$

Calculadora: $\boxed{\overline{x}}$

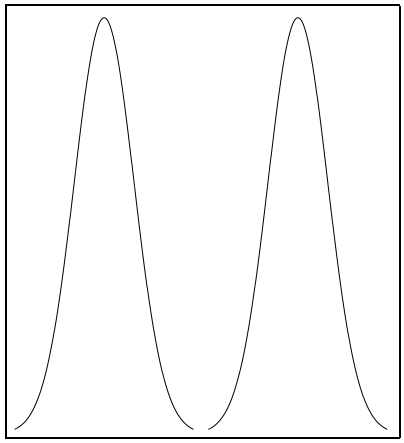
Ejemplo 0.7: Nota media de 5 prácticas: $\{10, 8, 9, 7, 9\}$

Una medida de posición no es suficiente para describir los datos, porque no informa acerca de la variabilidad de los datos. **Ejemplo 0.8.:** La nota media de prácticas es 5.2 tanto para $\{0, 2, 5, 9, 10\}$ como para $\{4, 5, 5, 6, 6\}$, sin embargo, claramente su dispersión es distinta.

Si representamos los histogramas mediante curvas continuas, apreciaremos la distinción entre posición y dispersión.



misma posición y diferente dispersión



distinta posición y misma dispersión

Medidas de dispersión: rango, rango intercuartílico, varianza, desviación típica o estándar, coeficiente de variación.

Varianza:

$$s^2 = \frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N-1} = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_N - \overline{x})^2}{N-1} \tag{2}$$

Fórmula alternativa:

$$s^2 = \frac{\sum_{i=1}^{N} {x_i}^2 - N \cdot \overline{x}^2}{N-1} = \frac{{x_1}^2 + {x_2}^2 + \cdots + {x_N}^2 - N \cdot \overline{x}^2}{N-1} \tag{3}$$

Calculadora: $\boxed{\sum x^2}$, $\boxed{\overline{x}}$ o bien $\boxed{\sigma_{N-1}}$, $\boxed{x^2}$

Observación: comprobación de la fórmula alternativa

$$\left(\sum_{i=1}^{N}(x_i - \overline{x})^2 = \sum_{i=1}^{N}(x_i^2 - 2\cdot\overline{x}\cdot x_i + \overline{x}^2) = \sum_{i=1}^{N} x_i^2 - 2\cdot\overline{x}\cdot\sum_{i=1}^{N} x_i + N\cdot\overline{x}^2 = \sum_{i=1}^{N} x_i^2 - 2\cdot N\cdot\overline{x}^2 + N\cdot\overline{x}^2 = \sum_{i=1}^{N} {x_i}^2 - N\cdot\overline{x}^2 \right)$$

Por la fórmula 2 puede apreciarse que a mayor varianza, mayor dispersión, pues calculamos desviaciones de la media al cuadrado. Por esto último (cuadrados), **la varianza siempre**

será mayor o igual que cero. [RECORDAD: NUNCA NEGATIVA, SIEMPRE POSITIVA ...].

Ejemplo 0.7:

¿Por qué dividir por N - 1, en lugar de por N? Por razones técnicas que ya se comentarán más adelante (tema 1); una justificación intuitiva sería considerar el caso en que N=1 (un único valor muestral). Si N es grande no habrá apenas diferencia.

Ejemplo 0.3: si sólo observáramos 1 niño (N=1) y nos diera como medida 7 cm, ¿cuál sería s^2? ¿Y si dividiéramos por N?

La varianza es muy apropiada por ciertas propiedades (si dos variables son independientes, la varianza de la suma es la suma de las varianzas), pero tiene un problema: cambia las unidades de los datos, ya que hacemos un cuadrado. Para resolverlo se usa la raíz cuadrada de la varianza:

Desviación típica o estándar:

$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N-1}} = \sqrt{s^2} \tag{4}$$

Calculadora: $\boxed{\sigma_{N-1}}$

Ejemplo 0.7:

0.3. Descripción de la población

Hasta ahora hemos examinado diversas formas de describir una muestra. Aunque la descripción de un conjunto de datos es a veces de interés por sí misma, normalmente lo que se pretende es generalizar y extender los resultados más allá de la limitación de la muestra. La población es realmente el foco de interés.

Como ya vimos (apartado 0.1), el proceso de sacar conclusiones sobre una población basándonos en las observaciones de una muestra de dicha población, es la Inferencia Estadística.

Puesto que las observaciones se realizan únicamente en la muestra, las características de la población nunca se conocerán exactamente. Para poder inferir ("deducir, concluir, tomar deci-

siones") de una muestra a la población, necesitaremos un lenguaje (paralelo al muestral) para describir la población.

Variables categóricas: podemos describir la población simplemente indicando la proporción de la población en cada categoría.

Ejemplo 0.5.:

Toda la población, todos los miembros de la UJI		La muestra de alumnos de 3^o ETIG	
Color	Frecuencia relativa (proporción)	Color	Frecuencia relativa (proporción)
1 = Negro	0.57	1 = Negro	0.52
2 = Blanco	0.14	2 = Blanco	0.07
3 = Rojo	0.09	3 = Rojo	0.13
4 = Otros	0.2	4 = Otros	0.28

La proporción muestral de una categoría es una *estimación* de la correspondiente proporción poblacional (en general desconocida). Puesto que no tienen porqué ser iguales (aunque sí que querríamos que fuesen cuanto más iguales mejor), las denotaremos con letras diferentes:

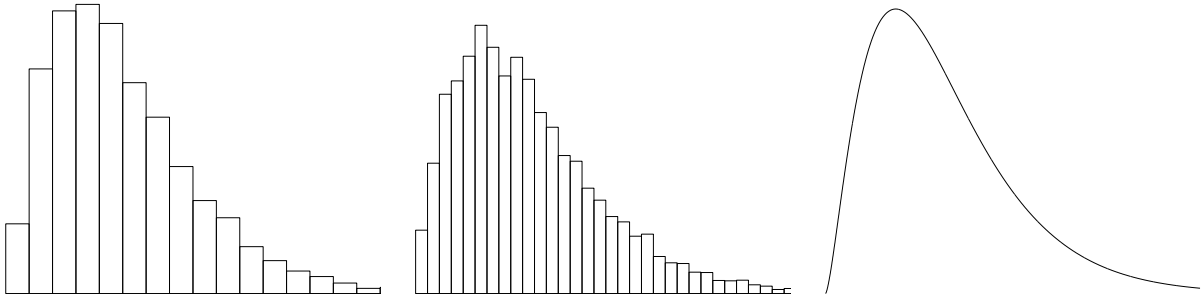
- p = proporción de la población
- $\hat{p}$ = proporción de la muestra

Variables cuantitativas: para variables cuantitativas, la media, varianza, desviación típica, etc. son descripciones de la población. Estas cantidades se calculan con los datos muestrales y constituyen una *estimación* de las correspondientes cantidades para la población. La media de la población la denotaremos mediante la letra μ, la varianza y desviación típica de la población con σ^2 y σ respectivamente. Recordemos que la media muestral era $\overline{x}$, la varianza muestral, s^2 y la desviación típica, s. Notemos que $\overline{x}$ es una *estimación* de μ (desconocida) y s es una estimación de σ (desconocida).

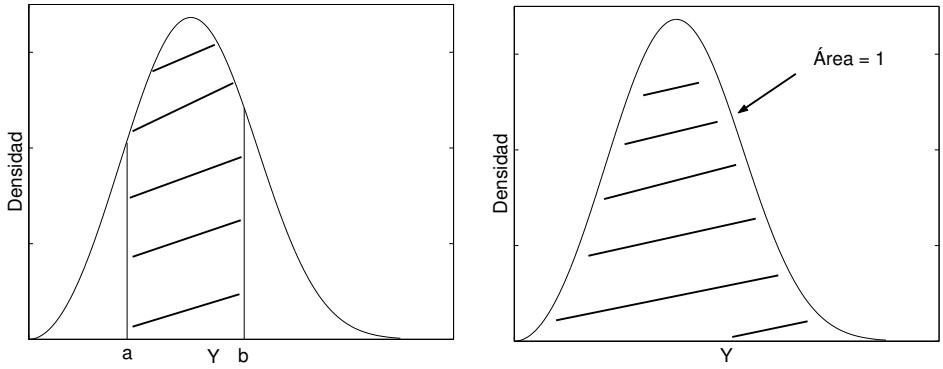
[♪ Nota: las cantidades poblacionales las denotamos con letras griegas que se corresponden con las respectivas letras latinas, para las cantidades muestrales].

Ejemplo 0.3.: con la muestra de 30 niños obtenemos $\overline{x} = 6$ y $s = 0.4$. La media de la población (todos los niños entre 7 y 9 años) la llamamos μ y no la conocemos. La desviación típica de la población (todos los niños entre 7 y 9 años) la llamamos σ y no la conocemos.

El histograma también es una buena herramienta que nos informa sobre la distribución de frecuencias de la población. Si, además, la variable es continua, podemos emplear una curva suave para describirla. Esta curva puede verse como una idealización del histograma con clases muy estrechas. Esta curva que representa la distribución de frecuencias, es la **curva de densidad**.

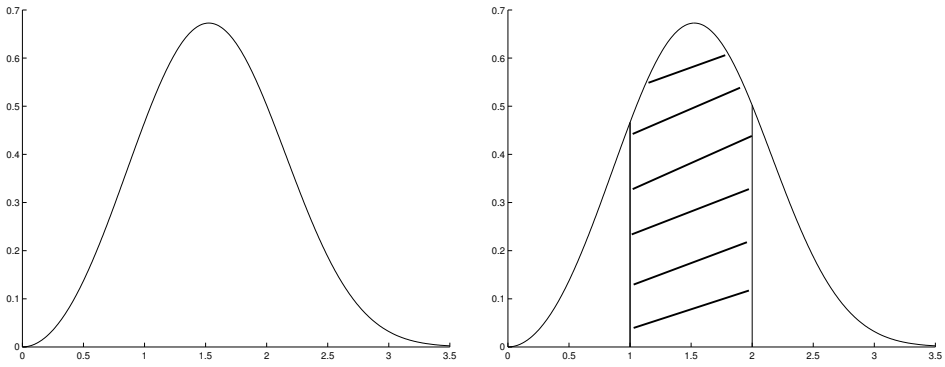


Interpretación de la densidad: el área bajo la curva de densidad entre los valores a y b equivale a la proporción de valores de la variable Y entre a y b.



Debido a la forma en que la curva es interpretada, el área bajo la curva entera debe ser igual a 1.

Ejemplo 0.9.: Supongamos que nos interesa la variable X = tiempo (en miles de horas) de vida de cierta clase de termostatos y que se distribuye según la siguiente curva de densidad:



Él área rayada es igual a 0.61, lo cual indica que el 61 % de los valores de la variable están entre 1 y 2.

Para calcular las áreas bajo las curvas de densidad, necesitaríamos integrar, pero ... usaremos tablas (formulario).

Observación: *¿Cuál sería la frecuencia relativa de un valor concreto, por ejemplo 6 cm, de la variable del ejemplo 0.3? La respuesta es cero (el área es cero). Aunque parezca extraño que la frecuencia relativa de una longitud igual a 6 cm sea cero, pensemos un poco. Si estamos midiendo hasta el milímetro más cercano, entonces, en realidad estamos preguntando la frecuencia relativa entre 5.95 cm y 6.05 cm, que no es cero. Pensemos en la longitud como una variable continua idealizada. Es similar al hecho de que una línea de 1 m, está compuesta de puntos, cada uno de ellos de longitud cero.*

En resumen, una medida numérica calculada a partir de los datos es un estadístico. La correspondiente medida numérica que describe la población es un parámetro. En la siguiente tabla se recogen las más importantes:

Medida	Valor muestral (estadístico)	Valor poblacional (parámetro)
Proporción	$\hat{p}$	p
Media	$\overline{x}$	μ
Desviación típica	s	σ

0.4. Probabilidad

¿Por qué hemos de estudiar la probabilidad? Las conclusiones de los análisis estadísticos de datos, vendrán generalmente dadas en términos probabilísticos (como ya se verá posteriormente en este curso ▶▶|, hasta ahora en el apartado 0.2 nos hemos limitado a describir los datos). La probabilidad entra en los análisis estadísticos, no únicamente porque el azar influya en los resultados de un experimento, sino también a causa de los modelos teóricos que se usarán en la parte de inferencia estadística. Para poder extraer conclusiones sobre la población, a partir de los datos de una muestra, será necesario recurrir a un modelo matemático (un esquema teórico de comportamiento) que nos determine las reglas de inferencia que es necesario utilizar. La probabilidad es el lenguaje y la fundamentación matemática de la estadística inferencial, al igual que las reglas de la gramática proporcionan las bases para organizar ideas a partir de las palabras que forman la lengua.

Espacio muestral y puntos muestrales: el espacio muestral S de una variable aleatoria X es el conjunto de valores que puede tomar dicha variable. Cada uno de los elementos de S se llama punto muestral. **Suceso:** es un subconjunto A de S.

STOP Una probabilidad es una cantidad numérica que expresa la verosimilitud de un cierto suceso A (*certidumbre de que el suceso A ocurra*), denotada como $P(A)$ (**probabilidad del suceso** A). Este número estará **SIEMPRE** entre 0 y 1 (ambos inclusive). Sólo tiene sentido hablar de probabilidad en el contexto de un **experimento aleatorio**, es decir, una operación (proceso) cuyo resultado viene determinado al menos parcialmente por el azar. De esta forma, cada vez que se lleva a cabo una operación, el suceso A puede ocurrir o no ocurrir. Dicho de otro modo, un experimento aleatorio es aquel que proporciona diferentes resultados aun cuando se

repita siempre de la misma manera.

La probabilidad podemos interpretarla en términos frecuenciales. Así, si un experimento aleatorio se pudiera repetir un número infinito de veces, la probabilidad de un suceso A, $P(A)$, se interpretaría como la frecuencia relativa de la ocurrencia del suceso A en una serie infinita de repeticiones de dicho experimento. O sea, si ese experimento se repitiera un número grande de veces y por cada repetición anotásemos la ocurrencia o no de A, se tendría:

$$\text{P(A)} \longleftrightarrow \frac{\text{número de veces que ocurre A}}{\text{número de veces que se repite el experimento}}$$

donde $\longleftrightarrow$ quiere decir: aproximadamente iguales si el experimento se repite muchas veces.

Ejemplo 0.10.: P("sacar cara") = 0.5, podéis lanzar una moneda muchas veces y comprobarla (si no está trucada, ¡claro!). De todas maneras, fíjate que para una realización concreta del experimento, quizá no obtengas exactamente la mitad de las veces cara. De hecho, cada vez que realices el experimento la frecuencia relativa seguramente cambiará, pero tras repetirlo muchísimas veces la frecuencia relativa (empírica o experimental) tenderá hacia la probabilidad teórica del suceso. La aproximación mejorará conforme más repeticiones se lleven a cabo. Las probabilidades de un experimento aleatorio a menudo se asignan sobre la base de un modelo razonable del sistema que se estudia. Otras veces, nos basaremos en los resultados de estudios realizados. Es decir, nosotros, para conocer las probabilidades de los sucesos, o bien nos basaremos en estudios realizados o bien, se asignarán siguiendo las especificaciones de un modelo teórico que plantearemos (en los dos próximos apartados) y que explicaría el fenómeno que se estudia.

Recuerda que siempre $0 \leq P(A) \leq 1$, siendo A un suceso.

0.5. Algunos modelos de distribuciones de probabilidad para variables discretas

Recordemos que una **variable aleatoria** es una variable cuyo valor depende del resultado de un experimento aleatorio. En el aparatado 0.1 ⏮ se vieron diversos ejemplos y se distinguió entre variables cualitativas (o categóricas) y cuantitativas. Dentro de éstas últimas, diferenciamos entre:

- **variables discretas**: toman un conjunto finito o infinito numerable (que se pueden contar) de valores
- **variables continuas**: su espacio muestral está formado por un conjunto infinito de valores que no podemos contar

A continuación repasaremos algunos modelos matemáticos concretos que nos darán la pauta de variabilidad asociada a una variable aleatoria. Estos modelos matemáticos se llaman distribuciones de probabilidad. 📄 Una **distribución de probabilidad** es un conjunto de probabilidades para los posibles distintos sucesos que pueden darse en un experimento aleatorio, en otras palabras, lo que nos proporciona es cómo se *distribuye* la probabilidad entre los sucesos que

pueden producirse.

[Nota: sólo visteis el caso univariante (una única variable), sin embargo también existen modelos que consideran varias variables conjuntamente].

Repasaremos 3 modelos (hay muchos más), que corresponderán a la consideración de experimentos con determinadas características. El fin de estos modelos teóricos es la descripción razonable de algunos fenómenos aleatorios. Son modelos aleatorios o estocásticos, que se diferencian de los modelos matemáticos determinísticos. Para los modelos determinísticos, los resultados se encuentran predeterminados por las condiciones bajo las cuales se verifica el experimento, es decir, dada una entrada, su salida (resultado) queda determinada. Por ejemplo, una fuente de alimentación (E) suministra corriente a circuito de resistencia eléctrica (R), el modelo matemático que nos describiría el flujo de corriente viene dado por la Ley de Ohm I=E/R. El modelo suministaría el valor de I tan pronto como se dieran los valores de E y R. Sin embargo, para los experimentos aleatorios, los resultados no pueden predecirse con certeza.

Los tres modelos que repasaremos son: la uniforme discreta, la Binomial y la Poisson. Tanto la distribución Binomial como la de Poisson tiene aplicación en fiabilidad y en control de calidad (cómo se verá en el tema 3 y en las prácticas con el Statgraphics). La fiabilidad estudia la probabilidad de funcionamiento de una unidad, entendida no sólo como parte indescomponible de un sistema, sino también como un sistema o subsistema en sí.

Distribución uniforme discreta: es la distribución que sigue una variable aleatoria X que toma n posibles valores x_1, x_2, ..., x_n con la misma probabilidad. Por tanto,

$$P(X = x_i) = \frac{1}{n} \qquad \text{i} = 1, ..., \text{n}$$

Ejemplo 0.11.: X="resultado al lanzar un dado no trucado"

0.5.1. Binomial

Esta distribución tiene una amplia gama de aplicaciones, sobre todo cuando se trata de realizar pruebas cuyo resultado sólo puede adoptar dos valores: "éxito" o "fracaso".

Supongamos que llevamos a cabo un **proceso de Bernoulli**, es decir, una serie de n pruebas. Cada prueba puede resultar en un "éxito" o en un "fracaso". La probabilidad de éxito es la misma cantidad, p, para cada prueba, sin importar los resultados de las otras pruebas, o sea, las pruebas son independientes.

La variable aleatoria X que representa el número de éxitos observados en un proceso de Bernoulli tiene una **distribución binomial**.

Ejemplo 0.12.: El ejemplo por excelencia de variable aleatoria distribuida como una binomial, sería X = "número de caras obtenidas al lanzar una moneda no trucada 5 (por ejemplo) veces", en este caso $n = 5$ y $p = 0.5$. O bien, X = "número de caras obtenidas al lanzar una

moneda trucada (de forma que la probabilidad de salir cara sea 0.7) 10 (por ejemplo) veces", en este caso $n = 10$ y $p = 0.7$.

Si la variable X sigue (se distribuye como) una distribución binomial de parámetros n y p (siendo n el número de pruebas y p la probabilidad de éxito), que representaremos como $X \sim Bi(n, p)$, las probabilidades se distribuyen de la siguiente manera (considerando combinatoria podría deducirse):

$$P(X = x) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x q^{n-x}, \qquad x = 0, 1, ..., n, \qquad q = 1 - p \qquad \text{, donde}$$

$$\left(\begin{array}{c} n \\ x \end{array} \right) = \frac{n!}{x! \cdot (n-x)!} \qquad \text{siendo} \qquad n! = n \cdot (n-1) \cdot (n-2) \cdot ... \cdot 2 \cdot 1$$

[ Fíjate que una variable $X \sim Bi(n, p)$, sólo puede tomar un número de valores FINITO, de 0 a n.]

Calculadora: para calcular $\left(\begin{array}{c} n \\ x \end{array} \right)$ puede emplearse las teclas $\boxed{nCr}$ o bien $\boxed{x!}$. Recuerda también que $\left(\begin{array}{c} n \\ 0 \end{array} \right) = \left(\begin{array}{c} n \\ n \end{array} \right) = 1$, $\left(\begin{array}{c} n \\ 1 \end{array} \right) = \left(\begin{array}{c} n \\ n-1 \end{array} \right) =$ n, 0! = 1 y 1! = 1.

En cada problema, debe especificarse qué quiere decir "éxito". "Éxito" puede ser "salir cara" como en el ejemplo 0.12., o bien, por ejemplo "ser defectuoso", "ser satisfactorio", o "cumplir las especificaciones", etc.

Más ejemplos de variables aleatorias con distribución Binomial son:

- Una máquina-herramienta desgastada produce 1 % de piezas defectuosas. La variable $X =$ "número de piezas defectuosas en las siguientes 25 piezas producidas" seguirá una distribución Binomial, con parámetros $n = 25$ y $p = 0.01$.
- De todos los bits transmitidos a través de un canal de transmisión digital, 10 % se reciben con error. La variable $X =$ "número de bits con error en los siguientes 5 bits transmitidos" se distribuye como una distribución Binomial(5,0.1).
- Un producto electrónico contiene 40 circuitos integrados. La probabilidad de que cualquiera de los circuitos integrados esté defectuoso es 0.01, y los circuitos integrados son independientes. La variable $X =$ "número de circuitos defectuosos de los 40" es Binomial(40,0.01).

Puesto que estamos estableciendo modelos teóricos que describan el *comportamiento* de ciertas variables aleatorias, también podremos establecer cuál sería la media poblacional, μ, y la varianza poblacional, σ^2 ( recuerda el apartado 0.3.), usando estos modelos.

Para una variable Binomial, $X \sim Bi(n, p)$, se tiene $\mu =$ n · p, y $\sigma^2 =$ n · p · q. La media, μ también se llama esperanza matemática.

0.5.2. Poisson

Consideremos ahora una serie de experimentos que consisten en observar el número de ocurrencias de un hecho en un intervalo de tiempo o espacio determinado. Por ejemplo:

Ejemplo: Número de errores en una superficie de grabación magnética.

Ejemplo: Número de mensajes que llegan a un servidor en una hora.

Ejemplo: Número de fallos de un equipo industrial durante 5 años.

Ejemplo: Número de defectos de fabricación por cada 1000 metros de cable.

 Una variable aleatoria X sigue una **distribución de Poisson**, si cuenta el número de ocurrencias por unidad de magnitud, cuando:

- el número de ocurrencias en un intervalo de tiempo o del espacio es independiente del número de ocurrencias en otro intervalo disjunto (proceso sin memoria).
- Además, la probabilidad de que haya una sola ocurrencia en un intervalo muy corto es proporcional a la amplitud del intervalo y
- la probabilidad de que haya más de una ocurrencia en un intervalo muy corto es despreciable.

Si la variable X sigue (se distribuye como) una distribución Poisson de parámetro λ ($X \sim \text{Po}(\lambda)$), donde λ indica el número medio de ocurrencias por unidad de magnitud y suele denominarse parámetro de intensidad, las probabilidades se distribuyen de la siguiente manera (también podemos probarla, aunque no lo haremos en clase):

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \qquad x = 0,\, 1,\, 2,\, 3,\, ... \qquad (x \in \mathbb{N})$$

[ Fíjate que una variable $X \sim \text{Po}(\lambda)$, puede tomar un número INFINITO NUMERABLE (contable) de valores.]

En el caso de una variable Poisson, $X \sim \text{Po}(\lambda)$, se tiene que $\mu = \lambda$ y $\sigma^2 = \lambda$.

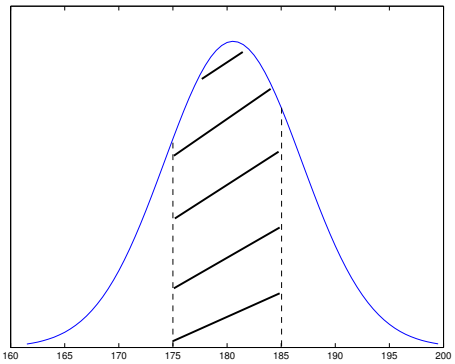
0.6. Algunos modelos de distribuciones de probabilidad para variables continuas

Se recordarán diversos modelos teóricos de distribuciones de probabilidad para variables continuas. En  el apartado 0.3., vimos como la distribución de la población de una variable aleatoria **continua** X podría describirse mediante una curva de densidad (como un histograma

idealizado), que representaba frecuencias relativas como áreas bajo la curva. Si en un histograma hacemos tender la amplitud del intervalo de clase a cero tendremos un número infinito de intervalos, convirtiéndose el histograma en un número infinito de barras de grosor infinitesimal, dispuestas de modo continuo (histograma idealizado). De esta forma, llegaríamos a la que llamamos en el apartado 0.3. **curva (o función) de densidad**, y que denotaremos como $f(x)$.

Cada uno de los modelos que repasaremos (y los que no repasaremos) tiene asociado su función de densidad y a través de ella podremos calcular probabilidades de distintos sucesos. La forma de calcular probabilidades para variables continuas difiere de la que se usa para variables discretas. Ahora para calcular la probabilidad de un suceso deberíamos calcular el área comprendida entre el eje x y la función de densidad (o sea, integrar), para los valores señalados por el suceso.

Ejemplo 0.13.: Si quisiéramos conocer la probabilidad de que un estudiante de la clase midiera entre 175 y 185 cm, P($175 \leq X \leq 185$), debemos calcular el área rallada, es decir, integrar la función de densidad entre 175 y 185 cm.



Según las reglas de probabilidad, tendremos que el **área total bajo la función de densidad es siempre 1**. Además, puesto que la integral de un punto al mismo punto vale cero (el área de una barra con grosor un punto es nula, recuerda también la última observación del punto 0.3.), se tiene que para **variables continuas, la probabilidad de que una variable aleatoria continua tome un valor puntual es cero**. Así, en el **ejemplo 0.13.**, P($X = 168.96$) = 0, por ejemplo. Por esta razón, para cualquier variable continua X se cumple: P(a $\leq$ X $\leq$ b) = P(a $<$ X $\leq$ b) = P(a $<$ X $<$ b) = P(a $\leq$ X $<$ b), o sea, PARA VARIABLES CONTINUAS ÚNICAMENTE, la probabilidad será la misma tanto si la desigualdad es o no estricta.

[Fíjate que esta última propiedad no se cumple para las variables discretas.]

Existe gran cantidad de modelos para variables continuas. Algunos modelos son: la Normal, uniforme, exponencial, Weibull, la t de Student, χ^2 Chi-cuadrado y F de Snedecor. Cada una de ellas tiene una curva de densidad y viene caracterizada por uno/s parámetros.

Como ya hemos dicho, para conocer la probabilidad de sucesos para variables continuas deberíamos integrar, sin embargo, para algunos modelos es posible expresar de forma analítica el valor de la integral mediante **la función de distribución acumulada** que denotaremos $F(x)$ y que nos proporcionará $P(X \leq x)$, es decir, para cada x, la función F nos devolverá la

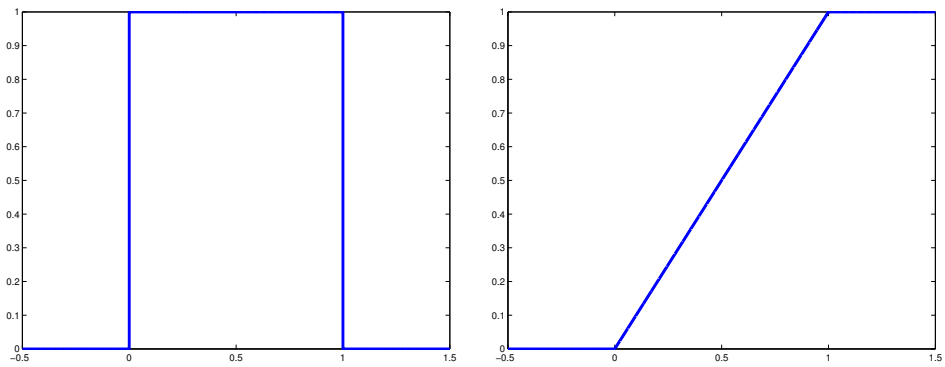
probabilidad de que la variable X tome un valor menor o igual que x. A veces, no existe tal expresión explícita y es preciso recurrir a tablas.

[A modo de resumen aclaratorio: cada modelo continuo viene determinado por su función de densidad, f. Hay que tener claro que la función de densidad, f, NO da probabilidades, sino el área bajo dicha función. Para calcular probabilidades hay que usar F, la función de distribución acumulada.]

.

0.6.1. Distribución uniforme(a,b)

Es la distribución que sigue una variable aleatoria X que toma valores en un intervalo [a,b] con la misma probabilidad. Por ejemplo, las calculadoras científicas con la tecla $\boxed{RAN\sharp}$ o $\boxed{Rnd}$ generan valores aleatorios de una variable uniforme entre 0 y 1. Su función de densidad y su función de distribución tienen la siguiente forma:



(a) (b)

Figura 2: (a) Función de densidad, f, de la Uniforme(0,1); (b) Función de distribución, F, de la Uniforme(0,1)

La función de densidad, de distribución acumulada, la media y varianza vienen dadas para una variable Uniforme(a,b) por:

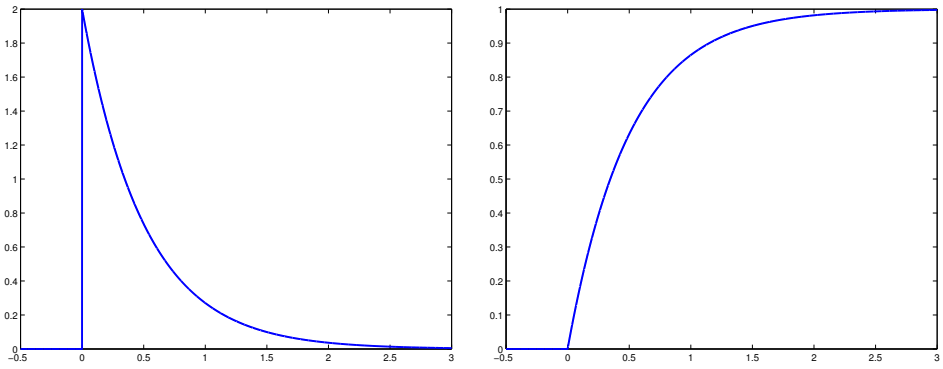
$$f(x;a,b) = \begin{cases} \frac{1}{b-a} & \text{si } a < x < b \\ 0 & \text{en otro caso} \end{cases} \; ; \; \mu = \frac{a+b}{2}, \; \sigma^2 = \frac{(b-a)^2}{12}$$

$$F(x;a,b) = \begin{cases} 0 & \text{si } x < a \\ \frac{x-a}{b-a} & \text{si } a < x < b \\ 1 & \text{si } x > b \end{cases}$$

0.6.2. STOP Distribución exponencial(λ)

Es usada muchas veces para modelizar el comportamiento de variables aleatorias del tipo "tiempo transcurrido hasta el fallo de un componente industrial" o "tiempo que se tarda en

completarse un proceso determinado". La función de densidad y función de distribución de una exponencial de parámetro λ tienen la siguiente forma:



(a) (b)

Figura 3: (a) Función de densidad, f, de la Exponencial(0.5); (b) Función de distribución, F, de la Exponencial(0.5)

La función de densidad, de distribución acumulada, la media y varianza vienen dadas para una variable Exponencial(λ) por:

$$f(x;\lambda) = \begin{cases} 0 & \text{si } x \leq 0 \\ \lambda e^{-\lambda x} & \text{si } x > 0 \end{cases} \; ; \; \mu = \frac{1}{\lambda}, \; \sigma^2 = \frac{1}{\lambda^2}$$

$$F(x;\lambda) = \begin{cases} 0 & \text{si } x \leq 0 \\ 1 - e^{-\lambda x} & \text{si } x > 0 \end{cases}$$

La distribución exponencial está relacionada con la Poisson de la siguiente forma: si el número de ocurrencias de un determinado fenómeno es una variable con distribución Poisson, el tiempo que pasa entre dos ocurrencias sucesivas es una variable con distribución exponencial.

La distribución Exponencial carece de memoria, se cumple $P(X > s+t|X > s) = P(X > t)$, en el contexto de "tiempos de vida" esto quiere decir que la probabilidad de fallar es independiente del pasado, el sistema no envejece. Aunque pueda parecer algo irreal, no es descabellado por ejemplo suponer que un fusible es "tan bueno como nuevo" mientras esté funcionado.

Más ejemplos de variables aleatorias exponenciales son:

- En una red de computadoras de una gran corporación, el acceso de usuarios al sistema puede modelarse como un proceso de Poisson con una media de 25 accesos por hora. La variable $X =$ "tiempo en horas desde el principio del intervalo hasta el primer acceso" tiene una distribución exponencial con $\lambda = 25$.

- El tiempo entre la entrada de correos electrónicos en una computadora podría modelizarse mediante una distribución exponencial.

- La CPU de un PC tiene un periodo de vida con una distribución exponencial con una vida media de 6 años.

0.6.3. Distribución Weibull(α,β)

Otra de las distribuciones que se aplica además de la Exponencial a problemas de fiabilidad y "tiempos de vida de componentes - equipos", es la Weibull(α,β). De hecho, para $\beta = 1$, la Weibull se reduce a la Exponencial. Esta distribución no se vio el curso pasado.

La función de densidad para Weibull(1,β) y distintos valores de β puede verse en el siguiente gráfico, $\beta > 0$ es un parámetro de forma y $\alpha > 0$ de escala.

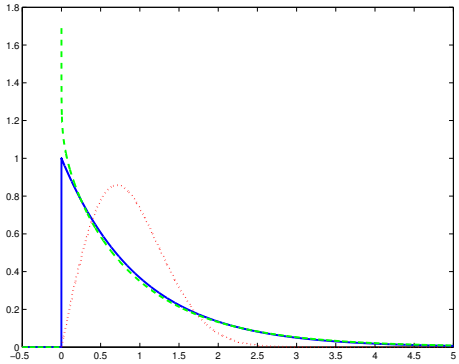


Figura 4: En azul y continua: Weibull(1,1), en rojo y puntos: Weibull(1,2), en verde y rayas: Weibull(1,0.95)

A continuación, aparece la expresión de su función de densidad:

$$f(x;\alpha,\beta) = \begin{cases} 0 & \text{si } x \leq 0 \\ \alpha\beta x^{\beta-1}e^{-\alpha x^{\beta}} & \text{si } x > 0 \end{cases}$$

Como ya se ha dicho, la distribución Weibull puede emplearse para modelar el tiempo hasta presentarse un fallo en muchos sistemas físicos diferentes. Los parámetros de esta distribución permiten gran flexibilidad para modelizar sistemas en los que el número de fallos aumenta con el tiempo (por ejemplo, el desgaste), disminuye con el tiempo (algunos semiconductores) o permanece constante (fallos provocados por causas externas al sistema). En la siguiente página *http* : *//www.itl.nist.gov/div*898/*handbook/apr/apr.htm* podréis encontrar un capítulo dedicado a la fiabilidad.

Más ejemplos de variables aleatorias Weibull son:

- Tiempo de vida (hasta el fallo) de un chip de memoria.
- Duración de cierto tipo de tubos al vacío.

0.6.4.  Distribución Normal(μ,σ^2)

La distribución Normal o Gaussiana es muy importante puesto que se utiliza para modelar muchísimos fenómenos aleatorios; además incluso se usa para aproximar otras distribuciones. La distribución Normal aproxima lo observado en muchos procesos de medición sin errores sistemáticos, por ejemplo medidas físicas del cuerpo humano (X = "altura de los jóvenes españoles"**ejemplo 0.13.**, X = "longitud del dedo índice de los niños"**ejemplo 0.3.**), medidas de calidad en muchos procesos industriales (**práctica 3** ), etc. Más ejemplos serían:

- En la detección de una señal digital, el ruido de fondo podría seguir una distribución normal (denominado ruido gaussiano) con media 0 volts y desviación típica de 0.45 volts.
- El diámetro de los puntos producidos por una impresora matricial seguiría una distribución normal con un diámetro promedio de 0.002 pulgadas y desviación típica de 0.0004 pulgadas.
- La vida de servicio efectiva de baterías usadas en un portátil.
- El diámetro de un eje en un propulsor de almacenamiento óptico, podría tener un distribución normal con media 0.2508 pulgadas y desviación típica de 0.0005 pulgadas.
- El volumen de llenado de una máquina automatizada usada para llenar latas de bebida carbonatada.
- La resistencia a la tensión del papel
- La vida de un componente electrónico bajo condiciones de alta temperatura para acelerar el mecanismo de fallo.
- Voltaje de ruptura de un diodo de un tipo particular
- Distribución de resistencia de resistores eléctricos, con media 40 ohmios y desviación típica de 2 ohmios

Una justificación de la frecuente aparición de la distribución Normal es el teorema central del límite: cuando los resultados de un experimento son debidos a un conjunto muy grande de causas independientes que actúan sumando sus efectos, cada uno de ellos de poca importancia respecto al conjunto, es esperable que los resultados sigan una distribución Normal.

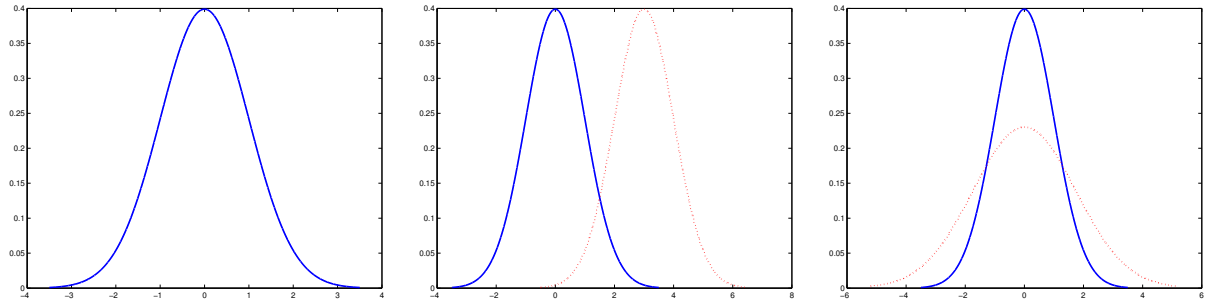
Ejemplo 0.3.: (ratón ergonómico). Este ejemplo, que nos ha ido *persiguiendo* durante todo el tema, nos va a permitir ver varios ejemplos más, de variables que podrían suponerse Normales.

Figura 5: Ratón óptico ergonómico 3M

Para comprobar científicamente las ventajas del ratón ergonómico frente al tradicional, se han realizado diversos estudios (en esta página, *http://www.animax.no/study*, puedes encontrar parte del trabajo). En esos estudios comparativos algunas de las variables empleadas y

que podemos suponer Normales son: tiempo de movimiento de cada ratón, actividad eléctrica de varios músculos del antebrazo durante la utilización de cada ratón, intensidad del dolor medida en una cierta escala (VAS).

La función de densidad de una Normal de parámetros μ (media de la población) y σ^2 (varianza de la población, siempre positiva), que denotaremos N(μ,σ^2) (a veces, como en el Statgraphics, se denota N(μ,σ)), tiene la forma siguiente:



(a) Normal (0,1) (b) (c)

Figura 6: (a) Normal(0,1); (b) Un cambio en la media, supone una traslación: Normal(0,1) en azul y continua, Normal(3,1) en rojo y punteada; (c) Un cambio en la varianza, supone un cambio en la variabilidad, pero el área bajo la curva sigue siendo 1, por ello tienen distinta altura: Normal(0,1) en azul y continua y Normal(0,3) en rojo y punteada

Como puede apreciarse, la Normal (campana de Gauss) es simétrica respecto de la media (que en este caso coincide con la mediana y la moda), o sea, el coeficiente de asimetría valdrá cero y además el coeficiente de curtosis es 3.

La función de densidad es:

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\ e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad , \quad x \in \mathbb{R}$$

La función de distribución acumulada es:

$$F(x;\mu,\sigma^2) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}}\ e^{-\frac{(y-\mu)^2}{2\sigma^2}}\, dy$$

La dejamos de esta forma, ya que un integrando de la forma e^{-z^2} no tiene primitiva. Por tanto, para calcularla o bien se emplea algún método numérico o se usan tablas, que es lo que haremos nosotros. Para ello necesitamos presentar la:

Distribución normal estándar: es aquella distribución normal con media 0 y varianza 1. La denotaremos mediante la letra Z.

Los valores que se recogen en las tablas (las tablas están en fotocopiadora) son para $N(0,1)$, además algunas calculadoras también permiten calcular probabilidades de una Normal estándar. La tabla nos proporciona:

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}} dx \quad , \quad Z \sim N(0,1)$$

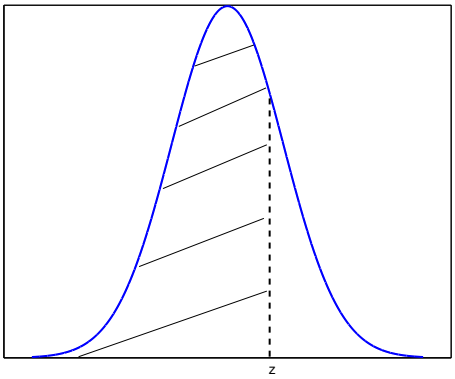
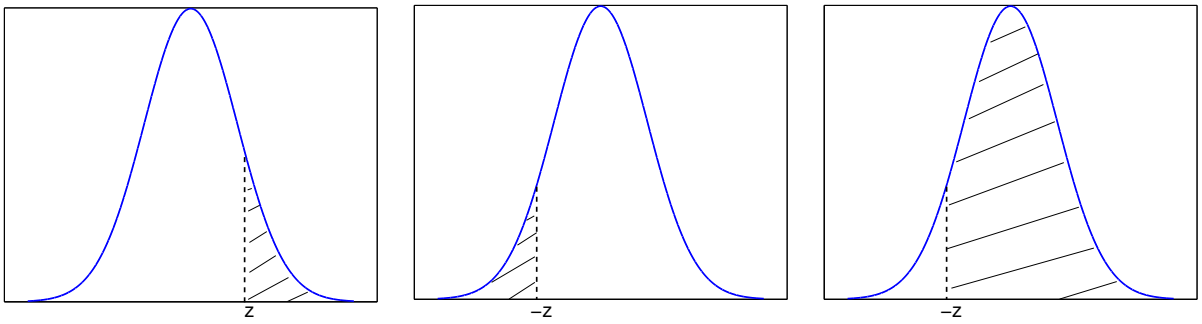


Figura 7: $\Phi(z)$

Durante el curso repasaremos el uso de la tabla.

Fíjate que, $P(Z \geq z) = 1$ - $P(Z \leq z)$, $P(Z \leq -z) = 1$ - $P(Z \leq z)$, $P(Z \geq -z) = P(Z \leq z)$. Ayúdate de un gráfico si lo necesitas.



(a) (b) (c)

Figura 8: (a) $P(Z \geq z)$; (b) $P(Z \leq -z)$; (c) $P(Z \geq -z)$

Con la tabla de la Normal(0,1) podemos calcular cualquier probabilidad de cualquier Normal, con cualquier media μ y varianza σ^2, no necesariamente N(0,1):

Estandarización: Sea $X \sim$ N(μ,σ^2), podemos estandarizarla (o tipificarla) y convertirla en una N(0,1) de la siguiente forma:

$$Z = \frac{X - \mu}{\sigma}$$

. O sea, si $X \sim$ N(μ,σ^2), $P(a < X < b) = P(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}) = P(Z < \frac{b-\mu}{\sigma})$ - $P(Z < \frac{a-\mu}{\sigma})$

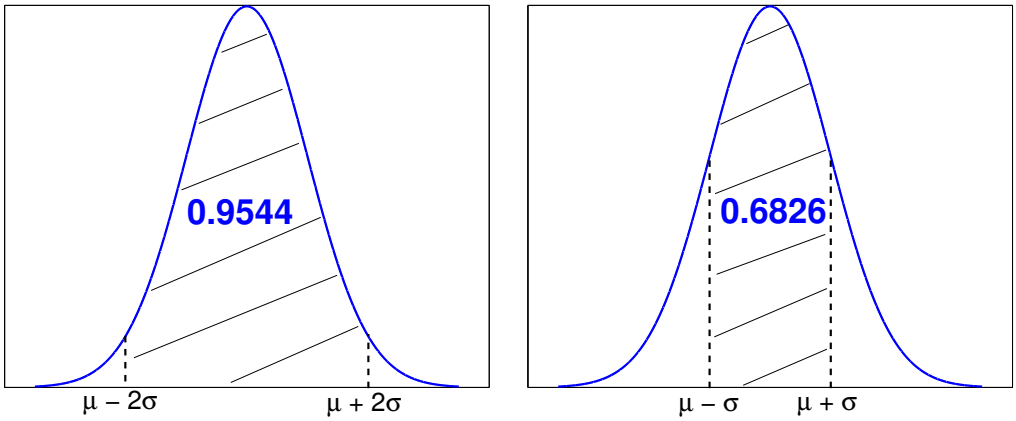
[Fíjate que para estandarizar, dividimos por la desviación típica σ, NO por la varianza σ^2].

[Fíjate que como la Normal es simétrica respecto su media μ, para $X \sim \mathrm{N}(\mu,\sigma^2)$: $P(X \leq \mu)$ $= P(X \geq \mu) = 0.5$. Además, si $x \geq \mu$, $P(X \leq x) \geq 0.5$. También, si $x \leq \mu$, $P(X \leq x) \leq 0.5$. Siempre que tengas dudas, recurre a hacer una representación gráfica].

Ejemplo 0.14.: Si $X \sim \mathrm{N}(\mu,\sigma^2)$, la fracción (proporción) de números que están a 3 desviaciones de la media es 0.9972, no importa el valor de μ, ni σ^2:

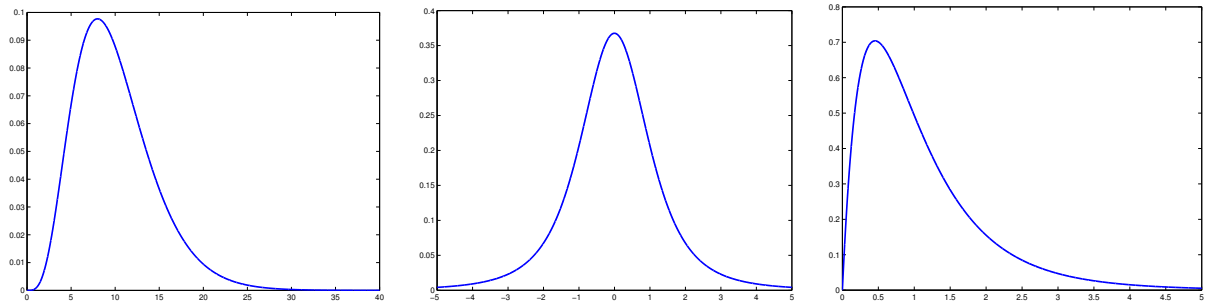
$P(\mu - 3\sigma < X < \mu + 3\sigma) = P(\frac{\mu-3\sigma-\mu}{\sigma} < Z < \frac{\mu+3\sigma-\mu}{\sigma}) = P(-3 < Z < 3) = P(Z < 3)$ - $P(Z < -3) = 0.9986$ - (1 - $P(Z < 3)$) = 0.9986 - (1 - 0.9986) = 0.9972

Puedes comprobar que la fracción de números que están a 2 desviaciones de la media es 0.9544 y la fracción de números que están a 1 desviación de la media es 0.6826.



Observación: *Aunque teóricamente la curva normal representa una distribución continua, a veces se usa para aproximadamente describir la distribución de una variable discreta. En esos casos, podría aplicarse una corrección de continuidad, para así obtener una mayor precisión.*

Otras distribuciones son la χ^2 Chi-cuadrado, t de Student y F de Snedecor, que usaremos y presentaremos este curso. Un ejemplo de ellas se muestra seguidamente.



(a) χ^2_{10} (b) t_3 (c) $F_{(4,20)}$

Figura 9: (a)χ^2 Chi-cuadrado; (b) t de Student; (c) F de Snedecor

0.7. Muestras aleatorias. Otros tipos de muestreo

Recordemos que nuestro objetivo es inferir sobre la POBLACIÓN. Nosotros sólo contamos con una muestra de la población. ¿Cómo generalizar más allá de un conjunto de datos particular? El primer paso para el desarrollo de una base para la inferencia estadística es encontrar un modelo probabilístico de las muestras que nos permita utilizarlas para inferir información sobre la población de la que se han extraído: el muestreo aleatorio simple.

Existen diversas técnicas de extracción de muestras de una población (como vereremos seguidamente). Nosotros nos centraremos en la más simple:

Muestreo aleatorio simple: se caracteriza por:

i) cada miembro de la población tiene la misma probabilidad de ser seleccionado

ii) las selecciones son independientes las unas de las otras.

Ejemplo 0.15.: Imaginemos que deseamos conocer el gasto en ocio (en un mes) de los jóvenes (18-30 años) españoles. Para ello extraemos una muestra de tamaño N (por ejemplo $N = 100$) por muestreo aleatorio simple (*pregunto el gasto a N jóvenes completamente al azar*). Si cada estudiante de la clase repitiera el experimento, tendríamos tantas muestras de tamaño N como estudiantes en la clase.

Por tanto, podemos considerar las variables aleatorias X_1, X_2, ..., X_N donde X_1 representa el valor (gasto) de la primera persona elegida (que variará de una muestra a otra), X_2 el valor de la segunda persona, ..., X_N el valor de la N-ésima persona.

Por la condición *i*), la distribución de cada X_i, $1 \leq i \leq N$, es la misma que la de la población (todas las variables X_i siguen la misma distribución). Por *ii*) X_1, X_2, ..., X_N son independientes (el conocimiento de una variable no aporta información acerca de los valores de la otra variable).

En consecuencia, X_1, X_2, ..., X_N, son independientes e idénticamente distribuidas (i.i.d) y constituyen una muestra aleatoria de tamaño N.

Estadístico: es cualquier función de las variables X_1, X_2, ..., X_N que constituyen una

muestra aleatoria. Algunos ejemplos son:

Media de muestreo:

$$\overline{X} = \frac{X_1 + X_2 + ... + X_N}{N}$$

Varianza de muestreo:

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

Un estadístico es una variable aleatoria por ser función de variables aleatorias, por lo cual tiene una distribución que se llama **distribución de muestreo**.

[♪ Nota: denotamos con mayúsculas los estadísticos de muestreo por ser variables aleatorias, de esta forma se distinguen de las cantidades muestrales ($\overline{x}$ y s^2, por ejemplo) que vimos en el apartado 0.2., que corresponden a una muestra concreta y tienen un valor numérico concreto.]

Aunque a lo largo de este curso siempre supondremos que nuestra muestra se ha obtenido por muestreo aleatorio simple, existen otros tipos de muestreo.

Un objetivo primordial de los procedimientos de muestreo es conseguir que la muestra sea representativa de la población ("como la población, pero en tamaño reducido"). Acabamos de presentar el muestreo aleatorio simple, que se usará cuando los elementos de la población sean homogéneos respecto a la característica a estudiar. Pero si disponemos de algún tipo de información sobre la población sería coveniente emplearla a la hora de seleccionar la muestra. Un ejemplo clásico son las encuesta de opinión, donde los elementos (personas) de la población son (o pueden serlo) heterogéneas en razón a su sexo, edad, profesión, etc. En estos casos interesaría que la muestra tuviera una composición análoga a la población, lo cual se conseguiría mediante muestreo estratificado.

Muestreo estratificado: los elementos de la población se dividen en clases o estratos. La muestra se toma asignando un número de miembros a cada estrato (pueden usarse distintos criterios: proporcional al tamaño relativo del estrato en la población, proporcional a la variabilidad del estrato, considerando costes, ...) y escogiendo los elementos por muestreo aleatorio simple dentro de cada estrato.

Ejemplo 0.15.: en este ejemplo, estaría bien dividir los elementos según su nivel económico, y por ejemplo dividirlos según la zona de la ciudad en que habiten: zona centro (clase alta), zona intermedia (clase media), barrios periféricos (clase baja).

Ejemplo 0.16.: queremos conocer la resistencia de los plásticos que hay en un almacén. Los plásticos provienen de dos fabricantes distintos. Sería mejor considerar dos estratos (cada fabricante), que los plásticos como un todo y muestrear sin distinción, porque puede que la distribución sea diferente en cada estrato.

Muestreo por conglomerados: se utiliza si la población se encuentra de manera natural agrupada en conglomerados, que podemos considerar como una muestra representativa de la población. La muestra se toma seleccionando algunos conglomerados al azar y dentro de ellos

analizando todos sus elementos o una muestra aleatoria simple.

Ejemplo 0.15.: siguiendo con este ejemplo, dentro de cada estrato (zona de la ciudad) podemos hacer divisiones en calles, las calles serían conglomerados ya que podemos considerarlas homogéneas respecto a la característica a estudiar.

Ejemplo 0.17.: supongamos que queremos analizar el diámetro de unas tuercas que tenemos almacendas en cajas. Sería más conveniente seleccionar al azar unas cajas y dentro de ellas realizar un muestreo aleatorio simple que llevar a cabo un muestreo aleatorio simple, pues esto implicaría seguramente abrir muchas más cajas.

Las ideas de estratificación y conglomerado son opuestas: la estratificación funciona tanto mejor cuanto mayor sean las diferencias entre los estratos y más homogéneos sean éstos internamente; los conglomerados funcionan si hay muy pocas diferencias entre ellos y son muy heterogéneos internamente.

Muestreo sistemático: cuando los elementos de la población están ordenados en listas, se usa el muestreo sistemático. Si la población es de tamaño N y la muestra deseamos que sea de tamaño n, tomaremos k como el entero más próximo a N/n, elegiremos un elemento al azar entre los k primeros, por ejemplo el n_1, después tomaremos los elementos $n_1 + k$, $n_1 + 2k$, etc, hasta completar la muestra.

Como se ha visto en el **ejemplo 0.15.**, los distintos tipos de muestreo pueden emplearse conjuntamente. Por ejemplo, en el análisis de diámetros de tuercas en cajas provenientes de dos fabricantes distintos (juntamos las ideas de los **ejemplos 0.16. y 0.17.**).

Observación: *el fin de esta aclaración es tratar de dar una visión general y localizar en que punto del temario nos encontramos, para no perder de vista el objetivo final, que trataremos en este curso. En el* **ejemplo 0.3** *(el del ratón ergonómico para niños), nos interesaba estudiar TODA la población de niños. Como eso es inviable, extraeremos una muestra (representativa) de la población, por ejemplo, $N = 100$ niños (muestreo aleatorio simple, aparatado 0.7.). A partir de esa muestra estudiaremos la variable X = "longitud del dedo índice" en la que estábamos interesados. Esta variable es cuantitativa y continua. (Podía habernos intersado más variables continuas como Y = "longitud entre dos puntos determinados de la mano", u otro tipo de variables, como Z = "satisfacción con un determinado juguete").*

Los datos (100 en este caso) que habríamos obtenido, primeramente los podríamos describir haciendo uso de las técnicas vistas en el aparatado 0.2.: tablas de frecuencias, gráficas (histogramas, diagramas de cajas, etc.) y medidas descriptivas: media ($\overline{x}$), mediana, varianza (s^2), desviación típica (s), percentiles, etc. Pero como ya sabemos, no estamos interesados en eso 100 niños concretos, sino en TODOS los niños, toda la POBLACIÓN. Para poder extraer conclusiones (INFERIR) acerca de la población (esto se verá en este curso), "necesitamos" asumir que nuestros datos provienen de una población que sigue un determinado modelo teórico (apartado 0.4, 0.5 y 0.6). A veces podría no asumirse un modelo parámetrico pero la estadística no paramétrica queda fuera de nuestro alcance. También existen tests para probar si nuestros datos provienen de un determinado modelo, que veremos en este curso.

Las conclusiones que obtendremos vendrán dadas en términos probabilísticos (por ejemplo,

el intervalo de confianza al 95 % para μ es ...) y serán conclusiones sobre descriptores de la población (apartado 0.3.): media (μ), varianza (σ^2), etc., que en realidad, muy difícilmente se conocen.