Muestreo y distribuciones muestrales.

Mathieu Kessler

Departamento de Matemática Aplicada y Estadística
Universidad Politécnica de Cartagena

Cartagena, Enero 2010

Guión

Guión

Introducción

El contexto

- Tenemos una pregunta acerca de un fenómeno aleatorio.
- Formulamos un modelo para la variable de interés X.
- Traducimos la pregunta de interés en términos de uno o varios parámetros del modelo.
- Repetimos el experimento varias veces, apuntamos los valores de X.
- ¿Cómo usar estos valores para sacar información sobre el parámetro?

Ejemplos

¿Está la moneda trucada?

- Experimento: tirar una moneda, v.a X=resultado obtenido. $\mathbb{P}(X=+)=p$, $P(X=c)=1-p$

$$¿p=1/2?$$

Sondeo sobre intención de participación en unas elecciones

- Queremos estimar la tasa de participación antes de unas elecciones generales.
- Formulamos un modelo: experimento: "escoger una persona al azar en el censo". X: variable dicotómica ("Sí", o "No"). $p=\mathbb{P}(X="Si")$.
- ¿Cuańto vale p?
- Censo: aprox. 32 000 000. Escogemos aprox. 3000 personas.

Ejemplos

Indice de audiencias

- ¿Cuánta gente vieron qué programas anoche?
- En rtve.es:
  - *TVE fue la cadena más vista para seguir las campanadas de 2010*
  - *En total más de 6.2 millones espectadores, el 45.1% del share*
- Taylor Nelson Sofres: aparatos de medición en 4500 hogares (sobre un total de 15 000 000 hogares españoles)

Determinación de la concentración de un producto

- Quiero determinar la concentración
- Formulo el modelo: exp="llevar a cabo una medición". X: "valor proporcionado por el aparato". $X \sim \mathcal{N}(\mu, \sigma^2)$.
- ¿Qué vale μ?

Surge una pregunta:

En todos estas situaciones, donde nos basamos en la repetición dun experimento simple...

- **¿Cómo sabemos que nuestra estimación es fiable?**
- **¿Qué confianza tenemos al extrapolar los resultados de una muestra de 3000 personas a una población de 32 millones de personas?**

Esbozo de respuesta: tasa de participación

Para convenceros, voy a construir un experimento de simulación.

- Voy a simular el proceso de extracción de una muestra de 3000 personas en una población de 32 millones de personas.

Construyo a mi antojo los distintos componentes:

- **La población:** Construyo en mi ordenador un fichero de 32 000 000 de ceros y unos. ($\Leftrightarrow$ el censo electoral)
    - "1" $\Leftrightarrow$ "la persona piensa ir a votar"
    - "0" $\Leftrightarrow$ "la persona NO piensa ir a votar"
- **La tasa de participación "real"**. Decido que en mi población el 70% piensa ir a votar$\Rightarrow$ 22 400 000 "1"s.
- **La extracción de una muestra**: construyo un pequeño programa que extrae al azar una muestra de 3000 números dentro del fichero grande.

Realización del experimento: conclusiones

- La enorme mayoría de las muestras de 3000 individuos proporcionan una tasa de participación muy próxima a la de la población.
- $\Rightarrow$ El riesgo de cometer un error superior a ± 2 puntos, al coger una muestra de 3000 individuos es muy pequeño (y asumible...)
- Si nos limitamos a muestras de 300 individuos, muchas muestras proporcionan una tasa de participación demasiado alejada de la de la población.

En la práctica

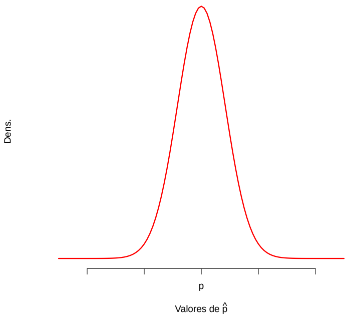
- Las empresas de sondeos no se basan en simulaciones sino en cálculos teóricos:
- Experimento aleatorio: escoger al azar una muestra de 3000 personas dentro de una población de 32 000 000, con una tasa de participación p.
- Llamamos a $\hat{p}$ la v.a.: proporción de "1"s en la muestra escogida.
- ¿Cuál es la distribución de valores de $\hat{p}$?

$$\boxed{\hat{p} \sim \mathcal{N}(p, \frac{p(1-p)}{n})}.$$

La distribución muestral de $\hat{p}$.

Uso de la distribución muestral

La distribución muestral de $\hat{p}$: la distribución esperada de los valores de $\hat{p}$ respecto a todas las muestras que podría extraer.



Antes de extraer una muestra:

- ¿Es suficiente el tamaño de la muestra para el riesgo asumible y la precisión requerida?

Una vez la muestra extraida:

- ¿Puedo dar un margen de error?
- ¿Puedo decidir si p poblacional es e.g. mayor que un valor dado?

Dos términos

Definición

- Cualquier cantidad calculada a partir de las observaciones de una muestra: estadístico.
- Experimento aleatorio: extraer una muestra. Consideramos un estadístico como v.a. La distribución del estadístico: distribución muestral.

Ejemplos de estadísticos

- Proporción muestral: $\hat{p}$
- Media muestral: $\bar{X}$.
- Desviación típica muestral: S_X.

Guión

La media muestral

Estudiamos una variable X cuantitativa

- Estamos interesados en μ, el centro de la distribución de X.
- Extraemos una muestra de tamaño n:

$$x_1, x_2, \ldots x_n$$

- Calculamos su media $\bar{x}$ para aproximar μ.
- ¿Cuál es la distribución muestral de $\bar{x}$?

Ejemplo

- Quiero medir una cantidad. Hay variabilidad en las mediciones.
- Introduzco una v.a. X="valor proporcionado por el aparato".
- μ representa el centro de los valores. Es mi objetivo
- Extraigo una muestra de tamaño 5 del valor de X. Estimo $\mu \simeq \bar{x}$.

Esperanza y varianza de la media muestral

Sea cual sea el modelo que hemos escogido para X.
Llamamos $\mu = \mathbb{E}[X]$ y $\sigma^2 = Var(X)$.

- Tenemos

$$\boxed{\mathbb{E}[\bar{X}] = \mu.}$$

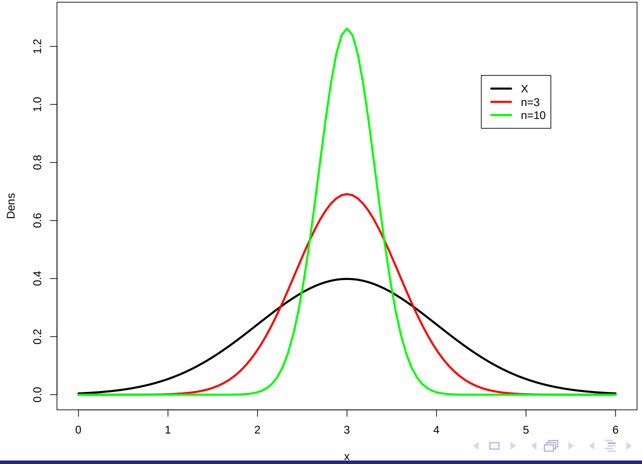
  El centro de la distribución muestral de $\bar{X}$ coincide con el centro de la distribución de X

- Tenemos

$$Var(\bar{X}) = \frac{\sigma^2}{n}, \qquad \Rightarrow \qquad \boxed{\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.}$$

  La dispersión de la distribución muestral de $\bar{X}$ es $\sqrt{n}$ veces más pequeña que la dispersión inicial de X

Ilustración, X inicial, $\bar{X}$ con $n = 3$, $\bar{X}$ con $n = 10$.

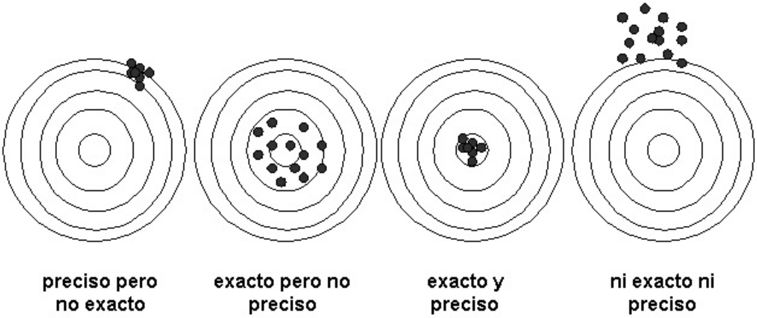


Consecuencia práctica

Aparato de medición

- Experimento: llevar a cabo una medición con un aparato.
- V.a X: valor proporcionado por el aparato.
- $\mathbb{E}[X]$: centro de la distribución de los valores proporcionados por el aparato.
  Lo deseable: $\mathbb{E}[X]=$ valor exacto de la cantidad que buscamos medir.
  En este caso, decimos: el aparato es exacto.
- σ_X: dispersión de la distribución de los valores proporcionados por el aparato.
  Lo deseable: σ_X pequeño.
  En este caso, decimos: el aparato es preciso.

Analogía con una diana



En resumen,

- Si el aparato de medición no es exacto, hay que calibrarlo.
- Si el aparato de medición no es preciso, podemos mejorar su precisión repitiendo las mediciones y proporcionando el valor promedio...

Distribución muestral de $\bar{X}$

Los resultados anteriores sobre $\mathbb{E}[\bar{X}]$ y $\sigma_{\bar{X}}$ son válidos sea cual sea el modelo escogido para la distribución de X.
Si queremos decir algo más preciso sobre la distribución de $\bar{X}$ (densidad etc..) necesitamos especificar la distribución de X.
Veremos dos casos:

1. Si la distribución de X es Normal.
2. Se la distribución de X no es Normal, pero n es grande.

Si la distribución de X es Normal

- Si la distribución de X es Normal

  $\Rightarrow$ la distribución de $\bar{X}$ también es Normal.
- Si $X \sim \mathcal{N}(\mu, \sigma^2)$,

$$\Rightarrow \quad \bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n}),$$

Si la distribución de X no es Normal

No podemos decir nada en general, excepto si n es grande...

Teorema Central del Límite

Si n es "suficientemente" grande, se puede aproximar la distribución de $\bar{X}$ por una Normal con media μ y varianza σ^2/n:

$$\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n}) \text{ aproximadamente.}$$

¿Cuándo considerar que n es grande?

- Depende de la forma de la distribución de X:
  - Si X casi Normal: n pequeño es suficiente
  - Si X muy asimétrico: n mucho más grande necesario.
- En general, se suele considerar $n \geq 30$ suficiente...

Guión

Distribución muestral de la proporción muestral

Contexto

- Hay situaciones donde X toma el valor 0 o 1, con probabilidades $1-p$ y p resp.
- E.g. exp: escoger una pieza en la producción. $X = 1$ si es defectuosa, $X = 0$ si es correcta.
- Repetimos n veces el experimento. Obtenemos

$$x_1, x_2, \ldots x_n$$

  Contamos N el número de "1"s.
- La proporción muestral es:

$$\hat{p} = \frac{N}{n}.$$

Distribución de la proporción muestral

Distribución exacta de $\hat{p}$

¿Cuál es la distribución de N?
Exp. simple con situación dicotómica, repetimos n veces...

$$N \sim \mathcal{B}(n, p).$$

Podemos usar esta distribución para hacer cálculos exactos...

Distribución aproximada de $\hat{p}$

Tenemos que $N = x_1 + x_2 + \ldots + x_n$
Por lo tanto

$$\hat{p} = \frac{N}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \bar{X}.$$

Por el Teorema Central del Límite:

$$\hat{p} \sim \mathcal{N}(p, p(1-p)/n) \quad \text{aprox.}$$

Podemos usar esta distribución para hacer cálculos aprox.

La obtención de las distribuciones muestrales exactas o aproximadas de los estadísticos que pueden resultar de interés para un problema dado es una tarea crucial en estadística...

Guión

Introducción a las gráficas de control

La obtención de las distribuciones muestrales $\Rightarrow$ propuesta de procedimientos de control estadístico de calidad en contextos industrial.
Shewhart (1920), Bell Labs...

¿Para qué sirve una gráfica de control?

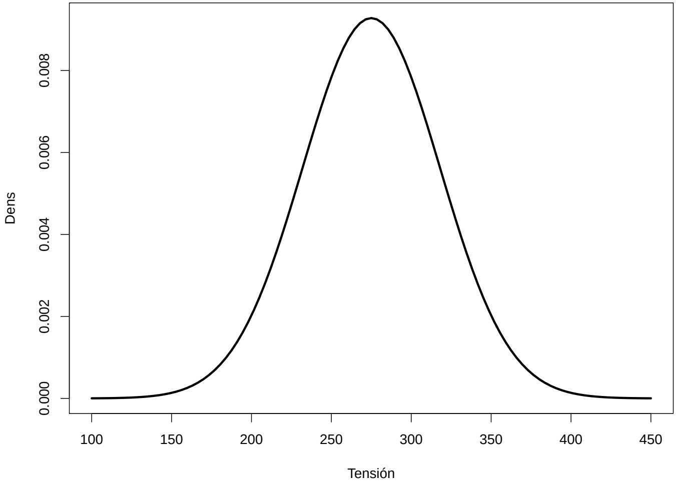
- Permiten comprobar de manera continua que se mantiene la calidad.
- Favorecen la intervención rápida.

La gráfica de control $\bar{X}$

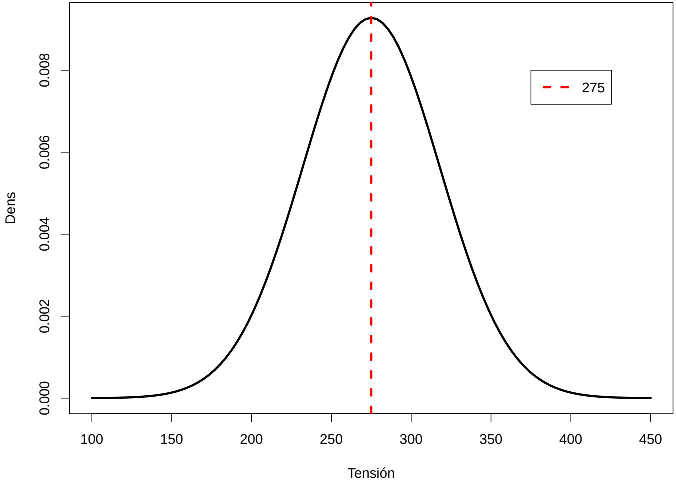
Ilustración con un ejemplo

- Una empresa produce monitores de ordenador. Una característica importante: la tensión en la rejilla de hilos de cobre.
- Tensión ideal: 275 mV.
- Es consciente de la variabilidad presente en la producción.
- Piensa que esta variabilidad es aprox. Normal, desviación típica $\sigma_X = 43mV$.

Distribución de X (Tensión)



Distribución de X (Tensión)

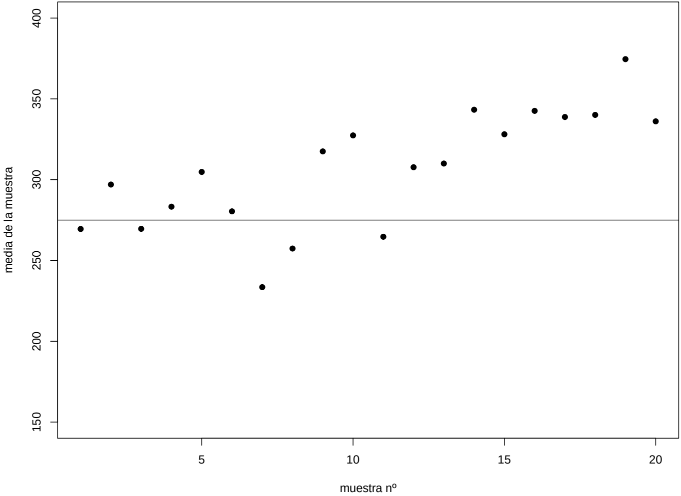


Procedimiento

La empresa escoge cada hora 4 pantallas, mide sus tensiones y apunta el promedio. Ejemplo de 20 muestras:

Muestra n^o	$\bar{x}$	Muestra n^o	$\bar{x}$
1	269.5	11	264.7
2	297.0	12	307.7
3	269.6	13	310.0
4	283.3	14	343.3
5	304.8	15	328.1
6	280.4	16	342.6
7	233.5	17	338.8
8	257.4	18	340.1
9	317.5	19	374.6
10	327.4	20	336.1

Gráfica



400
350
300
250
200
150
media de la muestra
5
10
15
20
muestra nº

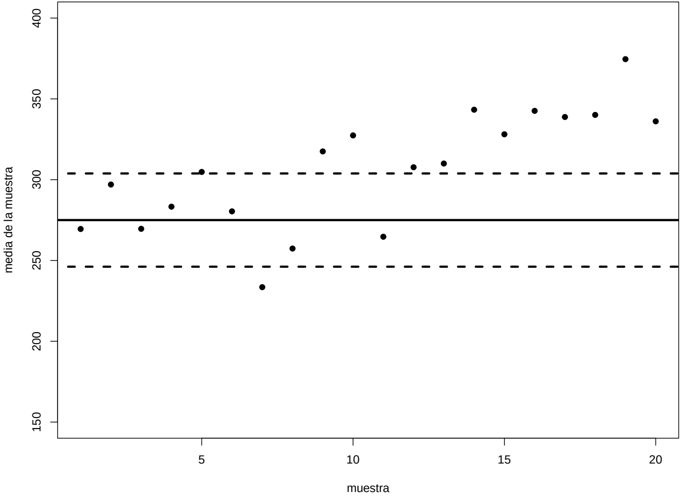
¿Cómo decidir cuándo el proceso se ha desajustado?

- Sabemos que la tensión en condiciones normales, $X \sim \mathcal{N}(275, 43^2)$.
- Al escoger 4 monitores y calcular la media $\bar{X}$, ¿qué distribución esperamos para los valores de $\bar{X}$?

$$\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n), \qquad \text{i.e } \bar{X} \sim \mathcal{N}(\mu, (\frac{43}{2})^2)$$

- ¿Entre qué valores debemos encontrar en condiciones normales el 99.7% de los valores de $\bar{X}$?
- Entre $\mu - 3\sigma_{\bar{X}}$ y $\mu + 3\sigma_{\bar{X}}$, es decir: entre $275 - 3 \times 21.5 = 210.5$ y $275 + 3 \times 21.5 = 339.5$.

Gráfica de control



Construcción de la gráfica

Representamos las medias muestrales obtenidas en una gráfica con:

- La línea objetivo μ.
- el límite de control superior en $\mu + 3\sigma/\sqrt{n}$
- el límite de control inferior en $\mu - 3\sigma/\sqrt{n}$.

Gráficas de control $\hat{p}$

En algunas situaciones, la calidad de la producción no se mide a tráves de una variable X sino a través de la proporción de defectuosos producidos. Gráfica de control $\hat{p}$.

Ejemplo

Una empresa fabrica placas base para ordenadores y se ha fijado como objetivo un 10% de defectuosos. Diariamente controla 400 placas.

16 días de control:

Día n^o	1	2	3	4	5	6	7	8
$\hat{p}$	0.1150	0.1600	0.1300	0.1225	0.1000	0.1225	0.1900	0.1150

Día n^o	9	10	11	12	13	14	15	16
$\hat{p}$	0.1000	0.1600	0.1675	0.1225	0.1375	0.1975	0.1525	0.1675

Construcción de la gráfica de control $\hat{p}$

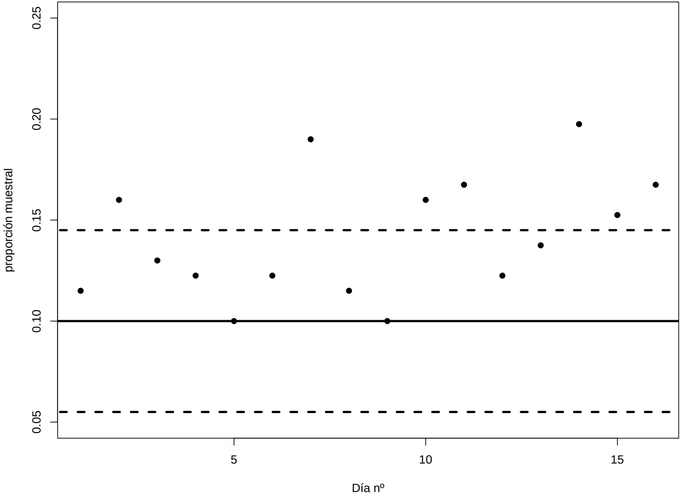
Sabemos

$$\hat{p} \sim \mathcal{N}(p, \frac{p(1-p)}{n}), \text{ aproximadamente.}$$

Procedimiento:

- la línea objetivo en p. $p = 0.1$
- el límite de control superior en $p + 3\frac{\sqrt{p(1-p)}}{\sqrt{n}}$. $= 0.145$
- el límite de control inferior en $p - 3\frac{\sqrt{p(1-p)}}{\sqrt{n}}$ $= 0.055$

Ejemplo



Otras señales de alarma

Existen otras señales de alarma. Por ejemplo:

- Se observan 9 puntos consecutivos por debajo (o por encima) de la línea objetivo.