Capítulo 3

Modelización de los datos. Inferencia estadística

1. Probabilidad. Repaso.

   Distribuciones de probabilidad: la binomial y la normal.

2. Inferencia estadística. Introducción y preliminares.

   Población y muestra. Parámetros y estadísticos.

   Objetivos de la inferencia: estimación, contraste y predicción.

   Obtención de datos: muestreo aleatorio simple.

3. Estadísticos. Distribución muestral.

   Parámetros, estadísticos y estimadores.

   Distribución muestral de un estadístico: medias y proporciones.

4. Decisiones que dependen de los parámetros del modelo: Estimación puntual.

   Introducción.

   Insesgadez y eficiencia.

   Comportamiento asintótico: consistencia.

   Propiedades de los estimadores hhabituales de la media, la varianza y la proporción.

5. Decisiones que dependen de los parámetros del modelo: Estimación por intervalos.

   Intervalos de confianza. Definición e interpretación.

   Métodos de construcción.

   Intervalos de confianza para medias y proporciones.

   Determinación del tamaño muestral.

6. Decisiones relacionadas con hipótesis sobre el modelo: Contrastes de hipótesis.

   Hipótesis estadísticas. El problema del contraste.

   Elementos de una contraste de hipótesis clásico.

   Contrastes para medias y proporciones. Comparación de medias.

   Dualidad entre contrastes de hipótesis y estimación por intervalos de confianza.

   Contraste de independencia en tablas de contingencia.

7. Inferencia en el modelo de regresión simple.

   Modelo de regresión lineal simple.

   Estimación del modelo.

   Inferencia bajo normalidad.

   Residuos.

   Extensiones del modelo.

3.0 Probabilidad. Repaso

REFERENCIA: Peña y Romo (1997). Capítulos 13 al 18

- **Variables aleatorias discretas y continuas.**
  - Modelización matemática de situaciones reales.
  - Función de masa. Función de densidad. Función de distribución.
  - Esperanza y varianza.
  - Distribución Binomial.
  - Distribución normal.
  - La normal como modelización de variables aleatorias que son *el efecto aditivo de muchas causas independientes de similar importancia.*
  - Aproximación mediante la normal de la binomial.
  - Cálculo de probabilidades en distribuciones normales.
- **¿Qué pasa cuando sumamos muchas variables aleatorias?.** Sea X v.a. con $E(X) = \mu$ y $V(X) = \sigma^2$. Sean $X_1, \ldots, X_n$ v.a.i.i.d. (variables aleatorias independientes e idénticamente distribuidas) con distribución igual a la de X. Definimos

$$S_n = \sum_{i=1}^{n} X_i, \quad \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Se tiene que

$$E(\bar{X}_n) = \mu, \quad V(\bar{X}_n) = \frac{\sigma^2}{n}$$

  - La media muestral $\bar{X}_n$ se acerca a la esperanza poblacional μ (Ley de los grandes números.)
  - La distribución de la media muestral tipificada se aproxima a la normal estándar (Teorema Central del Límite):

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \approx N(0,1)$$

3.1 Inferencia estadística. Introducción y preliminares

3.1.1 Población y muestra. Parámetros y estadísticos

REFERENCIA: Peña y Romo (1997). Capítulo 19

- La estadística descriptiva es útil para conjuntos de datos con un número *pequeño* de individuos.
- ¿Qué podemos hacer si el conjunto de individuos de interés es *demasiado grande*? *Grande* puede significar, por ejemplo, que examinar a todos los individuos sea costoso en tiempo y/o dinero, o incluso *de tamaño infinito*. Ejemplo: encuesta de presupuestos familiares (EPF), encuestas preelectorales, resultados de un experimento repetible infinitas veces.
- Lo que haremos será seleccionar algunos de los individuos y observar sólo estos.
- **Población.** Conjunto de individuos en cuyas características está interesado el investigador.
- **Muestra.** Subconjunto de individuos de la población que serán finalmente observados.
- Las herramientas de la estadística descriptiva aplicadas a una muestra, aportan mucha información sobre esa muestra. Pero el objetivo del investigador es llegar a hacer afirmaciones sobre toda la población, no sólo sobre la muestra concreta que observa.
- Si la muestra ha sido seleccionada *adecuadamente*, la teoría de probabilidad permite inferir características de toda la población a partir del estudio descriptivo de la muestra.
- En ocasiones, la característica poblacional de interés se identifica con un valor numérico desconocido (parámetro), y éste puede ser *inferido* a partir del cálculo de algunos estadísticos descriptivos en una muestra extraída de la población.

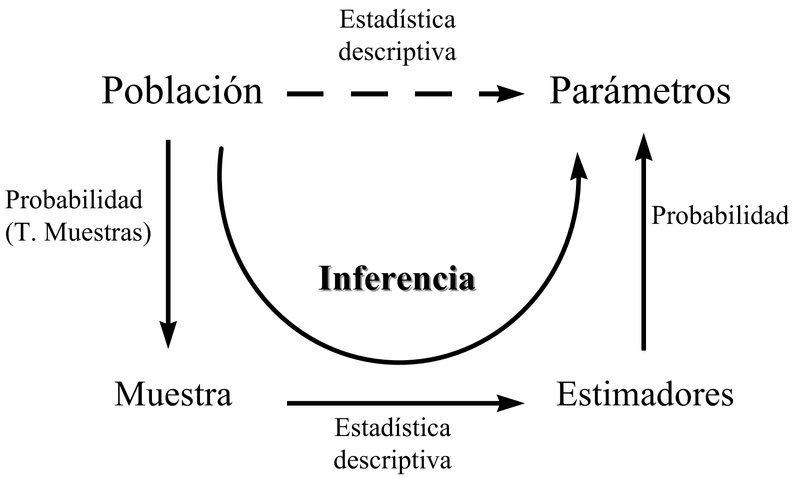


Figura 3.1: Esquema de la inferencia estadística.

- La inferencia estadística se vale de herramientas del cálculo de probabilidades y de la estadística descriptiva para construir afirmaciones válidas para toda la población a partir de la observación de una muestra representativa de ésta.

3.1.2 Objetivos de la inferencia: estimación, contraste y predicción

REFERENCIA: Peña y Romo (1997). Capítulo 19

- **Estimación.** Una característica de interés de la población se identifica con una constante numérica cuyo valor (desconocido) se infiere a partir de la información obtenida de una muestra.
- **Contraste.** Se realiza una afirmación sobre la población y se infiere su veracidad en base al estudio de dicha afirmación restringida a la muestra.
- **Predicción.** Se desea obtener información acerca del valor que toma una cierta característica en un individuo no observado de la población.

Para ello se observa dicha característica (y posiblemente otras) en los individuos de la muestra.

3.1.3 Obtención de datos: muestreo aleatorio simple

REFERENCIA: Peña y Romo (1997). Capítulo 19

- El control estadístico del proceso de obtención de datos es un requerimiento fundamental del posterior proceso de inferencia.
- El proceso de recogida de datos debe garantizar que la muestra sea representativa de la población de interés.
- **Muestreo aleatorio simple. Definición.** Se desea extraer una muestra de tamaño n de una población de tamaño N. El muestreo aleatorio simple es un procedimientos aleatorio de selección de la muestra que garantiza que todos los subconjuntos de n elementos de la población tienen la misma probabilidad de ser elegidos como muestra.
- Muestreos con y sin reemplazamiento.
- Normalmente se hará la hipótesis de que N es tan grande en relación a n que, en la práctica, se puede suponer que la población tiene tamaño infinito y que, por lo tanto, los muestreos con y sin reemplazamiento son equivalentes.
- En adelante se supondrá que las muestras de tamaño n son obtenidas mediante muestreo aleatorio simple de una población de tamaño N.
- **Estudio probabilístico de una m.a.s.** (Aquí estableceremos una conexión clara entre las tres partes en las que dividimos el temario de estadística y probabilidad: la estadística descriptiva, la probabilidad y la inferencia estadística.)
  - La población en cuya característica x estamos interesados puede representarse por los N valores que esta característica toma en los individuos que la integran:

$$\mathcal{X} = \{x^1, \dots, x^N\}.$$

– Supongamos, para simplificar la exposición, que la característica x es discreta y que puede tomar los k valores $c_1, \ldots, c_k$. Entonces, las técnicas de **estadística descriptiva** nos dicen que las características esenciales de la población $\mathcal{X}$ pueden expresarse (sin pérdida de información) en la tabla de frecuencias (absolutas N_i y relativas f_i)

Valores de x	N_i	$f_i = N_i/N$
c_1	N_1	f_1
$\vdots$	$\vdots$	$\vdots$
c_k	N_k	f_k
Total	N	1

– Si conociésemos esa distribución de frecuencias podríamos calcular cualquier característica (*parámetro*) de la población $\mathcal{X}$.

– Examinemos los elementos que forman la m.a.s. de la población $\mathcal{X}$. Empecemos por la primera observación:

> *Valor que toma la característica x en el individuo de la población que aleatoriamente es elegido como primer integrante de la m.a.s. extraída de la población $\mathcal{X}$.*

Este *valor* es una **variable aleatoria** (porque el primer individuo se elige aleatoriamente) que puede tomar cualquiera de los valores $x^1, \ldots, x^N$ (cualquier individuo puede ser elegido) con probabilidades $1/N$ (el m.a.s. garantiza equiprobabilidad).
Dado que la característica x es discreta, realmente ese *primer valor de la m.a.s.* sólo podrá tomar los k valores que puede tomar x, es decir, $c_1, \ldots, c_k$, y los tomará con probabilidades dadas por la regla de Laplace (*casos favorables partidos por casos posibles*):

$$P(\textit{primer valor en la m.a.s.} = c_j) = \frac{N_j}{N} = f_j, \quad j = 1, \ldots, k.$$

Así, la variable aleatoria *primer valor en la m.a.s.*, a la que llamaremos X_1, es una v.a. discreta cuya **distribución de probabilidades** viene dada por

X_1	c_1	$\ldots$	c_k
p_j	f_1	$\ldots$	f_k

– El resto de los valores de la m.a.s., a los que llamaremos $X_2, \ldots, X_n$ son también variables aleatorias con idéntica distribución que X_1

(suponemos muestreo con reemplazamiento). Además estas n variables aleatorias son mutuamente independientes (de nuevo, debido al esquema de muestreo con reemplazamiento).

– Podemos decir entonces que $X_1, \ldots, X_n$ *son n copias independientes de una variable aleatoria X cuya distribución de probabilidad coincide con la distribución de frecuencias relativas de la característica x en la población $\mathcal{X}$.*

– Identificaremos la población $\mathcal{X}$ con la variable aleatoria X, y diremos que $X_1, \ldots, X_n$ *es una muestra aleatoria simple de X.* Obsérvese que las *medidas descriptivas* de la población $\mathcal{X}$ (por ejemplo, la media aritmética de la característica x, $\bar{x} = \sum_{j=1}^{k} c_j f_j$) coinciden con los parámetros de la distribución de probabilidad de X (en el ejemplo, $E(X) = \sum_{j=1}^{k} c_j f_j$).

– En ocasiones no haremos referencia explícita a la población física en una de cuyas características estamos interesados, sino que directamente diremos que nos interesa estudiar la variable aleatoria X y para ello tomamos una m.a.s. de X.

– **Notación.**

$X_1, \ldots, X_n$ son variables aleatorias: los valores que tomará una cierta variable o característica X en los n individuos que acaben conformando una muestra aleatoria.

$x_1, \ldots, x_n$ son los valores concretos de la característica X en los n individuos que finalmente forman parte de la muestra. Son constantes, valores concretos, dado que una vez realizado el muestreo ya no hay incertidumbre alguna sobre qué individuos son observados ni, como consecuencia, sobre los n valores de x que son medidos.

3.2 Estadísticos. Distribución muestral

Un estadístico es cualquier función T_n que depende de la muestra. Adoptaremos la notación T_n para referirnos al valor del estadístico en una muestra aleatoria genérica $(X_1, \ldots, X_n)$, y llamaremos t_n al valor concreto del estadístico en una muestra $(x_1, \ldots, x_n)$ finalmente observada. Así, T_n es una variable aleatoria y t_n es una realización de T_n.

- Ejemplos.

La media muestral:

$$T_n(X_1, \dots, X_n) = \frac{X_1 + \cdots X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

El máximo de los datos observados:

$$T_n(X_1, \dots, X_n) = \max\{X_1, \dots, X_n\}$$

3.2.1 Parámetros, estadísticos y estimadores

REFERENCIA: Peña y Romo (1997). Capítulo 20

- Por regla general, la información que de una población se desea conocer puede expresarse como un número.
- Estas constantes cuyo valor es desconoce se denominan parámetros. Usaremos letras griegas para designar los parámetros.
- Cuando un estadístico T_n se usa para inferir el valor de un parámetro θ, diremos que T_n es un estimador de θ.

3.2.2 Distribución muestral de un estadístico: medias y proporciones

REFERENCIA: Peña y Romo (1997). Capítulo 20

- Un estadístico T_n es una variable aleatoria: es precisamente el hecho de que las muestras sean construidas aleatoriamente lo que confiere el carácter aleatorio a T_n.
- Por ello, la distribución de probabilidad de T_n se denomina distribución muestral de T_n o distribución de T_n en el muestreo. A la desviación típica de un estadístico T_n se la suele llamar **error estándar** de T_n: $\mathrm{EE}(T_n)$.

44	49	51	43	54	43	48	48	44	55	47	39
51	51	46	46	48	61	59	48	52	67	55	51
47	59	56	69	69	44	48	72	60	66	60	62
60	60	49	52	53	48	56	60	71	65	70	49
56	66	72	69	56	65	65	51	63	68	67	70
70	66	66	68	71	71	69	74	67	59	68	68
70	67	69	70	69	71	76	67	65	73	72	69
72	73	64	67	63	70	70	66	68	71	70	71
69	72	70	71	54	71	70	73	74	71	77	74
69	75	76	76	77	78	75	77	76	77	77	78
76	75	72	77	77	76	77	77	75	78	79	78

Tabla 3.1: Datos de la esperanza de vida para 132 países. Considerando este conjunto como población, se propone a grupos de alumnos que elijan 4 de esos valores *aleatoriamente*.

- Ejemplo 1: Considerar una población de tamaño $N = 6$, en cuyos individuos la característica X toma los valores $\{2, 4, 5, 6, 7, 8\}$. Se toman muestras sin reemplazamiento de tamaño $n = 2$. Llamar X_1 al posible valor de X en el primer individuo de la muestra y X_2 al segundo. Considerar como estadístico T_2 la media aritmética de las dos observaciones.
  En este ejemplo es posible deducir de forma exacta la distribución en el muestreo de T_n.
- Ejemplo 2: experimento en clase. Supongamos que tenemos una clase de 15 alumnos.
  - Mostrar una transparencia donde aparezca una lista de datos reales (por ejemplo, datos de una variable socio-económica observados en un conjunto de países como los que se muestran en la tabla 3.1; conviene suprimir de la lista los nombres de los países).
  - Pedir a un alumnos que elija 4 de los datos que está viendo y los anote en un papel. Tendremos así una muestra de tamaño $n = 4$. Discutir el carácter aleatorio del muestreo así planteado.
  - Pedir al alumno que calcule la media aritmética de los cuatro datos elegidos. Anotar el valor medio en la pizarra: una observación de la variable aleatoria "media aritmética de 4 datos".
  - ¿Cómo es la distribución en el muestreo de esa variable aleatoria? El número de posibles muestras de 4 elementos elegidas de la lista

de datos es elevadísimo (230.300 si hay 50 datos y no se permiten repeticiones). Para aproximar esa distribución, pedimos al resto de los alumnos que repitan lo que hizo su compañero. Con los quince datos de la media muestral dibujamos, por ejemplo, un diagrama de tallo y hojas en la pizarra. Este gráfico dará una idea aproximada de la distribución de la media muestral para muestras de tamaño 4. Podemos usar una tabla de dígitos aleatorios para aumentar el número de observaciones de la media muestral.

3.2.3 La hipótesis de normalidad

REFERENCIA: Peña y Romo (1997). Capítulo 18. Peña (1995), Vol. I. Capítulo 6.2.4

- Consideremos una variable estadística continua (Ej.: estatura) observada en una población de tamaño N (Ej.: si la población es el aula, $N = 15$). Pensemos en el histograma que describe esa variable.
- Aumentemos el tamaño de la población e imaginemos cómo evolucionan los histogramas que describen la variable estadística estudiada cuando N crece. Ej.: pasemos a considerar como poblaciones en primer lugar los alumnos de la EUM, después todos los alumnos universitarios de Catalunya, más tarde toda la población española con edades entre 18 y 25 años, etc.
- Una variable aleatoria continua puede servir como modelo matemático que aproxime la distribución de frecuencias de la variable estadística continua: cuando N crece, los histogrmas que describen las poblaciones sucesivas se estabilizan alrededor de la función de densidad de la variable aleatoria continua.
- En muchos casos, la distribución usada como modelo teórico es la distribución normal. Ej: la distribución de la estatura en una población suficientemenete grande es aproximadamente normal, en el sentido de que el histograma que la representa es simétrico y con una forma semejante a la campana de Gauss.
- Cuando se hace esta aproximación, no se fijan los valores de los parámetros.

- La distribución normal, $X \sim N(\mu, \sigma^2)$:

$$E(X) = \mu, \quad V(X) = \sigma^2, \quad DT(X) = \sqrt{V(X)} = \sigma$$

- Características que evidencian el carácter normal de una variable aleatoria: proceso de generación de los datos (TCL), simetría y apuntamiento.
- Contrastes gráficos de normalidad: histograma y gráfico cuantílico.

3.2.4 Distribución de la media muestral

REFERENCIA: Peña y Romo (1997). Capítulos 20 y 21

Sea $X_1, \ldots, X_n$ m.a.s. de $X \sim N(\mu, \sigma^2)$.

- Distribución de la media muestral en el caso de normalidad cuando σ es conocida:

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \Longleftrightarrow \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim N(0,1)$$

- Varianza desconocida. Estimación de σ^2:

$$s_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

- Comportamiento de los estimadores cuando n crece.

$$\frac{\sqrt{n}(\bar{X} - \mu)}{s_X} \approx N(0,1)$$

- Muestreo en poblaciones no normales. Teorema central del límite.

3.2.5 Estimación de una proporción

REFERENCIA: Peña y Romo (1997). Capítulo 20

Sea $X_1, \ldots, X_n$ m.a.s. de $X \sim B(1,p)$: X vale 1 con probabilidad p y 0 con probabilidad $q = 1-p$.

- Estimador de p: $\hat{P} = \bar{X}$ = proporción muestral de valores iguales a 1.
- Muestreo en la binomial. Aproximación de la binomial a la normal.

$$n\hat{P} \sim B(n,p) \approx N(np, np(1-p)) \Longrightarrow \frac{\sqrt{n}(\hat{P}-p)}{\sqrt{p(1-p)}} \approx N(0,1)$$

- Estimación de la varianza. Dos estimadores: s_X^2 y $\hat{P}(1-\hat{P})$. Están relacionados:

$$\hat{P}(1-\hat{P}) = \frac{n-1}{n} s_X^2$$

3.3 Decisiones que dependen de los parámetros del modelo: Estimación puntual

3.3.1 Introducción. Estimación de parámetros

REFERENCIA: Peña (1995), Vol. I. Capítulo 4.3

¿Por qué es importante estimar parámetros?

- En muchas ocasiones, los parámetros tienen interés en sí mismos.
  Ejemplos: Para una compañía de seguros médicos es importante saber cuál es el gasto anual medio de sus potenciales clientes. Una empresa de mensajería que quiere ofrecer un nuevo servicio necesita saber qué proporción de sus actuales clientes lo usarán.

- Modelización paramétrica:
  - Cuando se modeliza (se aproxima) una variable estadística de una población (ej.: estatura) mediante una variable aleatoria continua (ej.: normal) no se hace la aproximación a *una* variable concreta (ej.: la estatura es $N(\mu = 175, \sigma^2 = 10^2)$), sino que se elige una *familia paramétrica* (ej.: la estatura es normal $N(\mu, \sigma^2)$, con parámetros μ y σ^2 desconocidos).
  - Conocer el valor de unos pocos parámetros (Ej.: μ y σ) equivale a conocer toda la distribución de la variable de interés.
  - Esto hace que estimar esos parámetros sea un problema relevante.

3.3.2 Insesgadez y eficiencia

REFERENCIA: Peña (1995), Vol. I. Capítulo 4.4

Tenemos una población (una variable aleatoria) X y nos interesa un parámetro de esa población al que llamaremos θ.

- Sesgo de un estimador $\hat{\theta}$

$$B(\hat{\theta}) = E(\hat{\theta}) - \theta$$

- Estimadores insesgados. Precisión de estimadores insesgados. Eficiencia relativa de dos estimadores insesgados.

$$\text{Precisión}(\hat{\theta}) = \frac{1}{V(\hat{\theta})}; \quad \text{Ef.Rel}(\hat{\theta}_1, \hat{\theta}_2) = \frac{V(\hat{\theta}_2)}{V(\hat{\theta}_1)}$$

- Error cuadrático medio. Eficiencia relativa.

$$ECM(\hat{\theta}) = E[(\hat{\theta}-\theta)^2] = V(\hat{\theta}) + B^2(\hat{\theta})$$

- Estimadores eficientes.
- Importancia de disponer de técnicas que permitan aproximar la distribución en el muestreo de un estadístico.

- **Ejemplo:** Estimación del número de taxis que circulan por Barcelona.

N = número de taxis de Barcelona.

A. Primera parte.
Observamos n taxis y apuntamos el número de licencia l_i de cada uno.
Los taxis están numerados del 1 al N. Muestra de tamaño n.
l_i es la realización de una v.a. $L_i \sim \mathrm{UD}(1, N)$.
$L_1, \ldots, L_n$ v.a.i.i.d. $\mathrm{UD}(1, N)$.
N: parámetro de esa distribución.
Simplificación: $L_i \sim U(0, N)$.
Dos estimadores de N.
1. Estimador basado en la media.

$$E(L) = \frac{N}{2} \Longrightarrow N = 2E(L).$$

$$T_1(L_1, \ldots, L_n) = 2\bar{L}_n = 2(\sum_{i=1}^{n} L_i)/n \Longrightarrow E(T_1) = 2E(\bar{L}_n) = 2E(L) = N.$$

2. Estimador basado en el máximo. $M_n = \max_{i=1\ldots n}\{L_i\}$.

$$\text{Sea } x \in (0, N),\ F_{M_n}(x) = \mathrm{Prob}(M_n \leq x) =$$

$$= \mathrm{Prob}(L_i \leq x, i = 1 \ldots n) = \prod_{i=1}^{n} \mathrm{Prob}(L_i \leq x) = \frac{x^n}{N^n}.$$

$$f_{m_n}(x) = \frac{n x^{n-1}}{N^n} I_{(0,N)}(x).$$

$$E(M_n) = \int_0^N x f_{M_n}(x) dx = \frac{n}{n+1} N.$$

$$T_2(L_1, \ldots, L_n) = M_n(n+1)/n \Longrightarrow E(T_2) = N.$$

Muestra de $n = 49$ números de licencia de taxis de Barcelona.

315	616	671	1.079	1.301	1.503	1.601
1.629	1.941	2.227	3.007	3.176	3.826	4.064
4.181	4.135	4.327	4.413	4.661	4.797	4.880
5.899	6.018	6.111	6.206	6.219	6.556	7.274
7.595	8.165	8.222	8.325	8.424	8.258	8.830
9.156	9.399	9.413	9.442	9.870	9.897	9.897
9.928	10.276	10.361	10.794	10.888	10.947	10.975

$$\hat{N}_1 = t_1 = T_1(l_1, \ldots, l_{49}) = 12.314, \quad \hat{N}_2 = t_2 = T_2(l_1, \ldots, l_{49}) = 11.192$$

¿Cuál de las dos estimaciones de N es más fiable?
Distribución en el muestreo de los estimadores.
Precisión $\approx$ inverso de la varianza.

$$\mathrm{Var}\,(T_1) = 4\mathrm{Var}\,(\bar{L}_n) = 4\frac{\mathrm{Var}\,L_1}{n} = \frac{N^2}{3n}.$$

$$\mathrm{Var}\,(M_n) = \frac{n}{N^n}\int_0^N x^{n+1} dx - \left(\frac{n}{n+1}N\right)^2 = \frac{n}{(n+2)(n+1)^2} N^2,$$

$$\mathrm{Var}\,(T_2) = \frac{N^2}{n(n+2)}.$$

T_2 es más preciso que T_1.
Nota: $N = 11.000$

B. Segunda parte. T_2: distribución muestral completamente conocida.
T_1: sabemos su media y su varianza; su distribución asintótica es conocida, pero no hemos determinado totalmente su distribución en el muestreo.
Un tercer estimador basado en la mediana:

$$T_3 = T_3(L_1, \ldots, L_n) = 2\,\mathrm{med}(L_1, \ldots, L_n).$$

En la muestra de tamaño $n = 49$: $t_3 = 12.412$.
¿Como se comporta este estimador en relación a los otros dos?
$L \sim U(0, N)$ y conocemos el verdadero valor de N, que es 11.000.
Podemos utilizar métodos de simulación para aproximar la distribución de T_3.
$S = 2000$ simulaciones de muestras de licencias de tamaño $n = 49$. Cada dato se generó según una $U(0, 11.000)$.

Estimador	Valor Observado	Media teórica	Media en la simulación	Desv. típica teórica	Desv. típica en la simulación
T_1	12341	11000	10992	907	904
T_2	11199	11000	10993	220	229
T_3	12412	11000	11024	1571	1551

T_3, basado en la mediana, es menos preciso que los otros dos.
La simulación corrobora la relación teórica entre T_1 y T_2.
También podemos estimar correlaciones:

$$\mathrm{Corr}(T_1, T_2) = 0.252,\ \mathrm{Corr}(T_1, T_3) = 0.864,\ \mathrm{Corr}(T_2, T_3) = 0.160.$$

$\square$

3.3.3 Comportamiento en muestras grandes: consistencia

REFERENCIA: Peña (1995), Vol. I. Capítulo 4.4

- Estimadores consistentes.
- Insesgadez y eficiencia asintótica.

3.3.4 Propiedades de los estimadores usuales de la media, la varianza y la proporción

REFERENCIA: Peña (1995), Vol. I. Capítulo 4.4

La media muestral $\bar{X}_n$ como estimador de $\mu = E(X)$

- Insesgada, consistente.
- Su distribución es aproximadamente normal.
- Si además $X \sim N(\mu, \sigma^2)$, entonces $\bar{X}_n$ es eficiente: es el mejor entre los estimadores insesgados de μ.

La cuasivarianza muestral $\hat{\sigma}^2$ como estimador de $\sigma^2 = V(X)$.

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

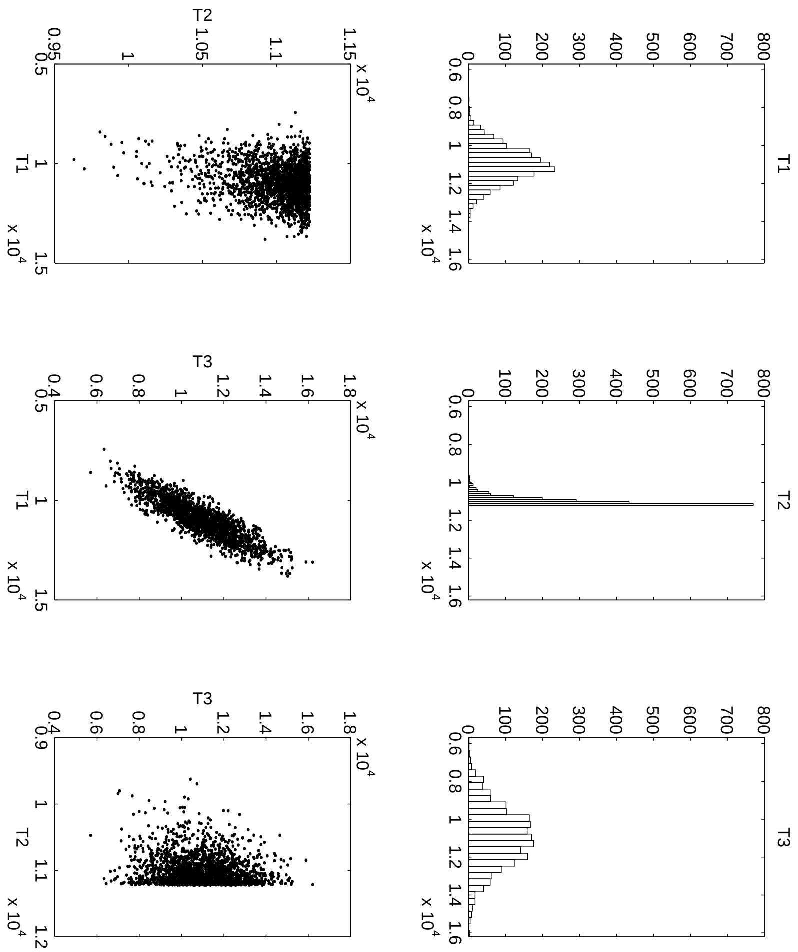


Figura 3.2: Comparación de tres estimadores en el problema del número de taxis

- Es insesgado.
- Es consistente.
- Es eficiente si X es normal.

La varianza muestral S_X^2 como estimador de $\sigma^2 = V(X)$.

$$S_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

- Es consistente.
- Es sesgado: $E(S_X^2) \neq \sigma^2$.
- Es tan eficiente como $\hat{\sigma}^2$ si el tamaño muestral tiende a infinito.

$\hat{\sigma}$ o S_X como estimadores de la desviación típica σ .

- Son sesgados.
- Son consistentes.

La proporción muestral $\hat{p}$ como estimador de $p = P(X = 1)$.
(Se supone que X puede valer 0 o 1).

- Es insesgado y consistente.
- Es eficiente, entre los estimadores insesgados.

3.4 Decisiones que dependen de los parámetros del modelo: Estimación por intervalos

REFERENCIA: Para esta parte mirar también Peña (1995), Vol. I. Capítulo 4.6

3.4.1 Intervalos de confianza. Definición e interpretación

REFERENCIA: Peña y Romo (1997). Capítulos 20, 21 y 22

Tenemos una población (una variable aleatoria) X, θ un parámetro de esa población que queremos estimar, y una m.a.s. de X: $X_1, \ldots, X_n$.

Supongamos que $\hat{\theta}_n$ es una estimación de θ. Nos podemos plantear algunas preguntas:

- ¿Cómo de fiable es esa estimación?
- El verdadero valor de θ no es $\hat{\theta}_n$. Aproximadamente, ¿a qué distancia de θ está $\hat{\theta}_n$?
- Si $\tilde{\theta}_n$ es otra estimación de θ, ¿cuál de las dos es mejor?

Estas preguntas ponen de manifiesto que una estimación debería aparecer siempre acompañada de alguna medida de su precisión. Por ejemplo, se podría dar siempre su error estándar.

Una forma habitual de inidcar la precisión de los estimadores es ofrecer como estimación un intervalo de posibles valores para el parámetro desconocido, en lugar de dar sólo un valor como estimación. Esos intervalos se llaman **intervalos de confianza**.

- Definición de intervalo de confianza. Nivel de confianza.

  Un estimador por intervalos de confianza $(1-\alpha)$ es un par de estadísticos (funciones de la muestra) $A_n = A(X_1, \ldots, X_n)$ y $B_n = B(X_1, \ldots, X_n)$ tales que

  $$P(A_n < \theta < B_n) = 1 - \alpha.$$

  Si la muestra observada es $x_1, \ldots, x_n$ y los valores de los estadísticos calculados para ella son $a_n = A(x_1, \ldots, x_n)$ y $b_n = B(x_1, \ldots, x_n)$, diremos que el intervalo (a_n, b_n) es una estimación por intervalos de confianza $(1 - \alpha)$ para θ, y lo escribiremos

  $$\mathrm{IC}_{1-\alpha}(\theta) \equiv (a_n, b_n).$$

- Ejemplo. Sea $X \sim N(\mu, 1)$. Queremos estimar el parámetro μ. Definimos las variables aleatorias siguientes:

  $$A_n = \bar{X}_n - \frac{1'96}{\sqrt{n}}, \; B_n = \bar{X}_n + \frac{1'96}{\sqrt{n}}.$$

Obsérvese que

$$P(A_n \leq \mu \leq B_n) = P(\bar{X}_n - \frac{1'96}{\sqrt{n}} \leq \mu \leq \bar{X}_n + \frac{1'96}{\sqrt{n}}) =$$

(restando $\bar{X}_n + \mu$ en los tres miembros)

$$P(-\mu - \frac{1'96}{\sqrt{n}} \leq -\bar{X}_n \leq -\mu + \frac{1'96}{\sqrt{n}}) =$$

(multiplicando por -1)

$$P(\mu + \frac{1'96}{\sqrt{n}} \geq \bar{X}_n \geq \mu - \frac{1'96}{\sqrt{n}}) = P(\mu - \frac{1'96}{\sqrt{n}} \leq \bar{X}_n \leq \mu + \frac{1'96}{\sqrt{n}}) =$$

(recordando que $\bar{X}_n \sim N(\mu, \frac{1}{\sqrt{n}})$)

$$P(\frac{\mu - \frac{1'96}{\sqrt{n}} - \mu}{\frac{1}{\sqrt{n}}} \leq \frac{\bar{X}_n - \mu}{\frac{1}{\sqrt{n}}} \leq \frac{\mu + \frac{1'96}{\sqrt{n}} - \mu}{\frac{1}{\sqrt{n}}}) =$$

$$P(-1,96 \leq Z \leq 1,96) = 0.95$$

(donde $Z \sim N(0,1)$)

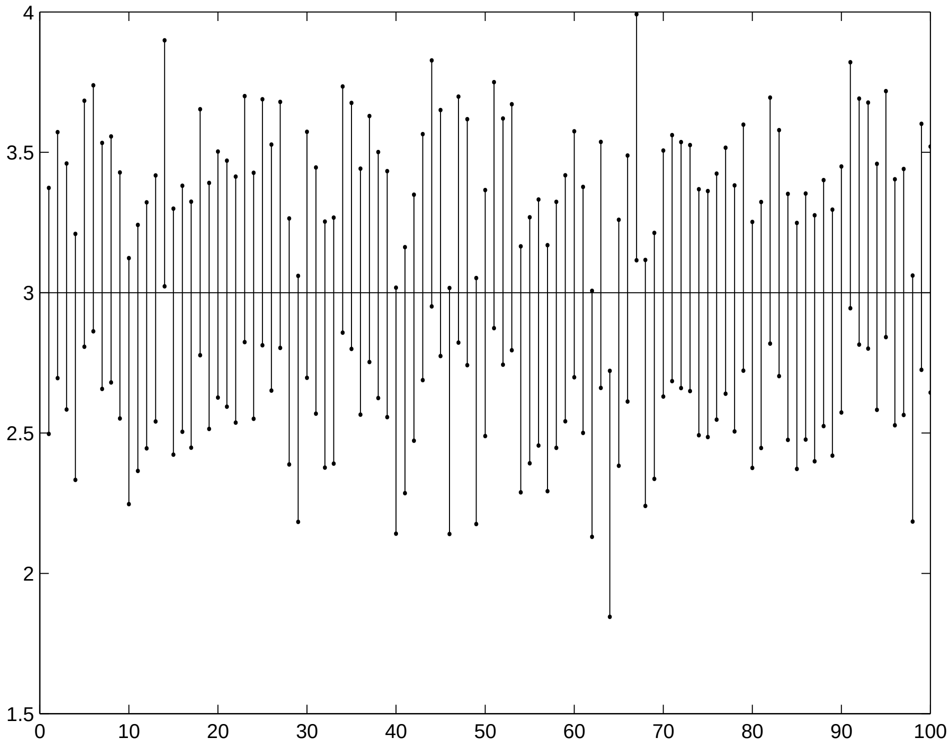
Hemos comporbado que $(A_n, B_n) = (\bar{X}_n - \frac{1'96}{\sqrt{n}},\ \bar{X}_n + \frac{1'96}{\sqrt{n}})$ es un estimador por intervalos de confianza 95%.

La estimación a que da lugar este estimador por intervalos es el intervalo de confianza

$$(\bar{x}_n - \frac{1'96}{\sqrt{n}},\ \bar{x}_n + \frac{1'96}{\sqrt{n}}) \equiv (\bar{x}_n \pm \frac{1'96}{\sqrt{n}})$$

- Interpretación frecuentista de un intervalo de confianza: Si repetimos la construcción de un intervalo de confianza $1-\alpha$ un número grande de veces, cada vez con una muestra distinta, la proporción de esos intervalos que contendrán el verdadero valor del parámetro es aproximadamente de $1-\alpha$.

- Ejemplo. En el ejemplo anterior, usamos el ordenador para construir 100 intervalo de confianza para la media de una población normal de varianza conocida: $X \sim N(\mu, 1)$. Para generar los datos se usó $\mu = 3$.



3.4.2 Intervalos de confianza para medias y proporciones

REFERENCIA: Peña y Romo (1997). Capítulos 20, 21 y 22

$X \sim N(\mu, \sigma)$, μ desconocida y σ conocida.

Intervalo de confianza $(1-\alpha)$ para la media μ basado en la $N(0,1)$.

$$IC_{1-\alpha}(\mu) \equiv \left(\bar{x}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{x}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

El valor $z_{\alpha/2}$ es tal que si Z es $N(0,1)$, entonces

$$P(Z > z_{\alpha/2}) = \frac{\alpha}{2}$$

$X \sim N(\mu, \sigma)$, mu y σ desconocidas.
Intervalo de confianza $(1-\alpha)$ para la media μ basado en la t de Student.

$$IC_{1-\alpha}(\mu) \equiv \left(\bar{x}_n - t_{\alpha/2}\frac{S_x}{\sqrt{n}},\ \bar{x}_n + t_{\alpha/2}\frac{S_x}{\sqrt{n}}\right)$$

S_x es la desviación típica muestral. El valor $t_{\alpha/2}$ es tal que si t es t de Student con $(n-1)$ grados de libertad, entonces

$$P(t > t_{\alpha/2}) = \frac{\alpha}{2}$$

Si n es grande: En este caso, si se sustituye σ por S_x en el primer intervalo, se obtiene prácticamente el mismo resultado que con el segundo método.

Si X no es normal: Como la media muestral es asintóticamente normal, se tiene que

$$IC_{1-\alpha}(\mu) \equiv \left(\bar{x}_n - z_{\alpha/2}\frac{S_x}{\sqrt{n}},\ \bar{x}_n + z_{\alpha/2}\frac{S_x}{\sqrt{n}}\right)$$

es un intervalo de confianza para μ.

Intervalos para proporciones Es un caso particular de lo anterior. Consideramos la variable $X \sim Bern(p)$. Entonces $\mu = E(X) = p$. Además, si observamos n valores de X, la media muestral $\bar{x}_n$ coincide en este caso con la proporción muestral $\hat{p}_n$. El intervalo de confianza para p queda así:

$$IC_{1-\alpha}(\mu) \equiv \left(\hat{p}_n - z_{\alpha/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}},\ \hat{p}_n + z_{\alpha/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}\right)$$

Intervalos para la diferencia de medias.

- Datos apareados. Se observan n pares de valores de las variables X e Y: $(x_1, y_1), \ldots, (x_n, y_n)$. Se quiere dar un intervalo de confianza para la diferencia $\mu_X - \mu_Y$, donde $\mu_X = E(X)$ y $\mu_Y = E(Y)$.

Se define la variable *diferencia*: $D = X - Y$. Entonces tenemos n observaciones de D: $d_i = x_i - y_i$, $i = 1, \ldots, n$. La esperanza de D es $\mu_D = E(D) = E(X - Y) = \mu_X - \mu_Y$.
Si hacemos un intervalo de confianza para μ_D tal como hemos indicado en los apartados anteriores, entonces habremos construido un intervalo para la diferencia de medias.

- Muestras independientes. Se observan n valores de la variable de interés en la población 1, y m valores en la población 2. Llamaremos $x_i, i = 1, \ldots, n$ a los primeros e $y_j, j = 1, \ldots, m$ a los segundos.
En este caso la expresión del intervalo es más compleja.
Se diferencian aquí dos casos: varianzas iguales en las dos poblaciones y varianzas distintas.

Intervalos para diferencias de proporciones. Caso particular de lo anterior.

Una regla (casi) universal: En muchos casos

$$(\theta \pm C_\alpha \mathrm{EE}(\theta))$$

es aproximadamente un $\mathrm{IC}_{1-\alpha}(\theta)$, donde C_α es un número que depende de α y del problema específico. En muchos casos, C_α es el cuantil $Z_{\alpha/2}$ de la $N(0,1)$:

$$\mathrm{IC}_{1-\alpha}(\theta) = (\theta - 1{,}96\mathrm{EE}(\theta), \theta + 1{,}96\mathrm{EE}(\theta)).$$

3.4.3 Determinación del tamaño muestral

REFERENCIA: Peña y Romo (1997). Capítulos 20, 21 y 22

- Queremos estimar un parámetro θ mediante un estimador $\hat{\theta}_n$.
- Si se fija un nivel de confianza $(1-\alpha)$ y un margen de error L, se puede calcular el tamaño muestral necesario para que, con una probabilidad de al menos $(1-\alpha)$, el error cometido al estimar θ por $\hat{\theta}_n$ no supere el margen de error L.

- Esto es posible porque la longitud de los intervalos de confianza decrece cuando n crece.
- **Ejemplo:** Tamaño muestral en la estimación de una proporción. Ficha técnica de las encuestas de opinión publicadas en la prensa.

3.5 Decisiones relacionadas con hipótesis sobre el modelo: Contrastes de hipótesis

REFERENCIA: Para esta parte mirar también Peña (1995), Vol. I. Capítulos 5.1 a 5.5

3.5.1 Hipótesis estadísticas. El problema del contraste.

REFERENCIA: Peña y Romo (1997). Capítulo 23

- Hipótesis sobre parámetros.
- Planteamiento del problema como un problema de toma de una decisión: hay que decidir si una hipótesis es cierta o falsa.
- Un contraste de hipótesis es una regla que asocia el valor 0 o el valor 1 a cada muestra que pudiera ser observada. Cuando se toma una muestra se le aplica el contraste y se observa si el resultado es 0 o 1. Si es 1 se considerará que la hipótesis es cierta. Si es 0, se considera que la hipótesis es falsa.
- Hipótesis nula y alternativa. Hipótesis simples y compuestas. Contrastes unilaterales y bilaterales.

- Paralelismo con la dinámica de un juicio: el acusado es inocente (esta es la hipótesis nula), mientras no se demuestre lo contrario (es decir, mientras que la evidencia aportada por la muestra no contradiga significativamente la hipótesis nula).
- Estructura general de un contraste de hipótesis: se mide la distancia entre lo que se observa en la muestra y lo que se esperaría observar si fuese cierta la afirmación que se sostiene bajo la hipótesis nula. Una discrepancia *demasiado grande* implica que la hipótesis nula está demasiado alejada de la realidad y que, por tanto, se debe rechazar la hipótesis nula en favor de la alternativa.
- Decir *demasiado grande* equivale a decir *muy poco probable si la hipótesis nula fuese cierta.*
- **Nivel crítico** o ***p*-valor**: probabilidad de observar una discrepancia igual o mayor que la que se ha observado, suponiendo que la hipótesis nula fuese cierta.

3.5.2 Elementos de una contraste de hipótesis clásico

REFERENCIA: Peña y Romo (1997). Capítulo 23

- Errores de tipo I y de tipo II.

	Realidad	
Resultado del test	H_0 cierta	H_0 falsa
Rechazar H_0	Error de tipo I	Decisión correcta
No rechazar H_0	Decisión correcta	Error de tipo II

- No es posible diseñar tests que tengan pocos errorres de tipo I y pocos errores de tipo II: cuando se reduce la probabilidad de cometer uno de los errores es a costa de aumentar la porbabilidad de cometer el otro tipo de error.
- Nivel del contraste = α = Probabilidad de rechazar la hipótesis nula cuando ésta es cierta:

$$\alpha = P(\text{Rechazar } H_0 \mid H_0 \text{ es cierta}).$$

Es la probabilidad del *Error de Tipo I*.

- Estadístico del test. Región crítica.
- Función de potencia. Probabilidad de rechazar la hipótesis nula cuando ésta es falsa:

  $P(\text{Rechazar } H_0 \mid H_0 \text{ es falsa}) = P(\text{Región crítica} \mid H_0 \text{ es falsa}) = 1-\beta$

  donde β es la probabilidad de *Error de Tipo II*.
- Relación entre el nivel de un test y el p-valor.

3.5.3 Contrastes para medias y proporciones. Comparación de medias

Referencia: Peña y Romo (1997). Capítulos 23 y 24

- Test para la media. Contrastes unilaterales y bilaterales. Ver la Tabla 3.2.
- Contrastes para dos muestras. Datos apareados (se reduce al problema de una muestra). Muestras independientes: igualdad de medias, igualdad de varianzas. Ver la Tabla 3.3.
- Contrastes para proporciones. Ver la Tabla 3.4.

3.5.4 Dualidad entre contrastes de hipótesis y estimación por intervalos de confianza

Referencia: Peña (1995), Vol. I. Capítulo 5.5.1

- Tenemos una población (variable aleatoria) X. Nos interesa el parámetro θ de esa población. En particular queremos contrastar $H_0 : \theta = \theta_0$.

Contraste	$X \sim N(\mu,\sigma^2)$, σ conocida	$X \sim N(\mu,\sigma^2)$, σ desconocida	X con distribución cualquiera n grande (usar σ si es conocida)
Estadístico	$D_n = \frac{\bar{X}_n-\mu_0}{\sigma/\sqrt{n}} \overset{H_0}{\sim} N(0,1)$	$D_n = \frac{\bar{X}_n-\mu_0}{S/\sqrt{n}} \overset{H_0}{\sim} t_{n-1}$	$D_n = \frac{\bar{X}_n-\mu_0}{S/\sqrt{n}} \overset{H_0}{\sim}_A N(0,1)$
$\left\{\begin{array}{ll} H_0: & \mu=\mu_0 \\ H_A: & \mu\neq\mu_0 \end{array}\right.$	Región crítica: $\left\{d_n \le -Z_{\alpha/2} \text{ ó } d_n \ge Z_{\alpha/2}\right\}$ p-val.$= P(\lvert N(0,1)\rvert > \lvert d_n\rvert)$	Región crítica: $\left\{d_n \le -t_{n-1,\alpha/2} \text{ ó } d_n \ge t_{n-1,\alpha/2}\right\}$ p-val.$= P(\lvert t_{n-1}\rvert > \lvert d_n\rvert)$	Región crítica: $\left\{d_n \le -Z_{\alpha/2} \text{ ó } d_n \ge Z_{\alpha/2}\right\}$ p-val.$= P(\lvert N(0,1)\rvert > \lvert d_n\rvert)$
$\left\{\begin{array}{ll} H_0: & \mu=\mu_0 \\ & (\text{ó } \mu\ge\mu_0) \\ H_A: & \mu<\mu_0 \end{array}\right.$	Región crítica: $\{d_n \le -Z_\alpha\}$ p-val.$= P(N(0,1) < d_n)$	Región crítica: $\{d_n \le -t_{n-1,\alpha}\}$ p-val.$= P(t_{n-1} < d_n)$	Región crítica: $\{d_n \le -Z_\alpha\}$ p-val.$= P(N(0,1) < d_n)$
$\left\{\begin{array}{ll} H_0: & \mu=\mu_0 \\ & (\text{ó } \mu\le\mu_0) \\ H_A: & \mu>\mu_0 \end{array}\right.$	Región crítica: $\{d_n \ge Z_\alpha\}$ p-val.$= P(N(0,1) > d_n)$	Región crítica: $\{d_n \ge t_{n-1,\alpha}\}$ p-val.$= P(t_{n-1} > d_n)$	Región crítica: $\{d_n \ge Z_\alpha\}$ p-val.$= P(N(0,1) > d_n)$

Tabla 3.2: **Contrastes para medias. Cuadro resumen.** Estadístico del test: D_n. Valor del estadístico del test en la muestra: d_n.

Situación	Estadístico del contraste
X, Y normales con varianzas conocidas	$$D = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_X^2}{n_x} + \frac{\sigma_Y^2}{n_y}}} \sim N(0,1)$$
Si los tamaños muestrales n_x, n_y, son *grandes*, el resultado es aproximadamente válido aunque X e Y no sean normales, o aunque las varianzas no sean conocidas (se sustituyen entonces por las varianzas muestrales S_X^2 y S_Y^2).	
X, Y normales con varianzas desconocidas y tamaños muestrales *pequeños*. Suponemos que podemos hacer la hipótesis adicional de igualdad de varianzas $$\sigma_X^2 = \sigma_Y^2$$	$$D = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{n_x+n_y}{n_x n_y}}} \sim t \text{ de Student}$$ $$S_p = \frac{(n_x-1)S_X^2 + (n_y-1)S_Y^2}{n_x+n_y-2}$$
Tamaños muestrales pequeños y varianzas distintas	No podemos decir nada en este caso, ni siquiera aunque X e Y sean normales.

Tabla 3.3: **Contrastes de igualdad de medias: muestras independientes.**

PROPORCIONES n grande	**Estadístico:** $D_n = \frac{\hat{p}_n - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \overset{H_0}{\sim}_A N(0,1)$
$\begin{cases} H_0 : p = p_0 \\ H_A : p \neq p_0 \end{cases}$	Región crítica: $\left\{d_n \leq -Z_{\alpha/2} \text{ ó } d_n \geq Z_{\alpha/2}\right\}$ p-val.$= P(\|N(0,1)\| > \|d_n\|)$
$\begin{cases} H_0 : p = p_0 (\text{ó } H_0 : p \geq p_0) \\ H_A : p < p_0 \end{cases}$	Región crítica: $\{d_n \leq -Z_\alpha\}$ p-val.$= P(N(0,1) < d_n)$
$\begin{cases} H_0 : p = p_0 (\text{ó } H_0 : p \leq p_0) \\ H_A : p > p_0 \end{cases}$	Región crítica: $\{d_n \geq Z_\alpha\}$ p-val.$= P(N(0,1) > d_n)$

Tabla 3.4: **Contrastes para proporciones. Cuadro resumen.** Estadístico del test: D_n. Valor del estadístico del test en la muestra: d_n.

- Para hacer ese contraste, tomamos una muestra de X: $x_1, \ldots, x_n$. Con ella podemos hacer dos cosas:
  1. Contrastar $H_0 : \theta = \theta_0$ al nivel α.
  2. Estimar θ mediante un intervalo de confianza $(1-\alpha)$:

$$(\hat{\theta} - L,\ \hat{\theta} + L)$$

- La relación que hay entre los dos procedimientos es ésta: rechazaremos H_0 si y sólo si θ_0 no está dentro del intervalo de confianza.
- En general se tiene que el conjunto de valores θ_0 que identifican hipótesis nulas $H_0 : \theta = \theta_0$ simples que no serían rechazadas a nivel α, forman un intervalo de confianza $(1-\alpha)$ para θ. El recíproco también es cierto, si θ_0 pertenece a un intervalo de confianza $(1-\alpha)$ para θ, entonces no se rechazará la hipótesis nula $H_0 : \theta = \theta_0$ si hacemos el contraste a nivel α.

3.5.5 Contraste de independencia en tablas de contingencia

Referencia: Peña y Romo (1997). Capítulo 24

- Ejemplo de un contraste con hipótesis no paramétricas.
- Modelización bajo la hipótesis nula: frecuencias esperadas si hay independencia E_{ij}. Frecuencias observadas O_{ij}.

$$O_{ij} = n_{ij} = n\ f_{ij}$$

$$E_{ij} = n\ f_{i.}\ f_{.j}$$

- Estadístico del contraste

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

3.6 Inferencia en el modelo de regresión simple

Referencia: Esta parte se sigue bien en Peña (1995), Vol. II. Capítulos 11.1 a 11.7, 11.9 y 11.11

3.6.1 Modelo de regresión lineal simple

El modelo. Dos variables X e Y siguen el modelo de regresión lineal simple si la variable Y se puede generar a partir de la variable X de la siguiente forma: si un individuo tienen un valor x_i de la primera variable, entonces el valor y_i que la segunda variable toma en ese individuo viene dado por

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

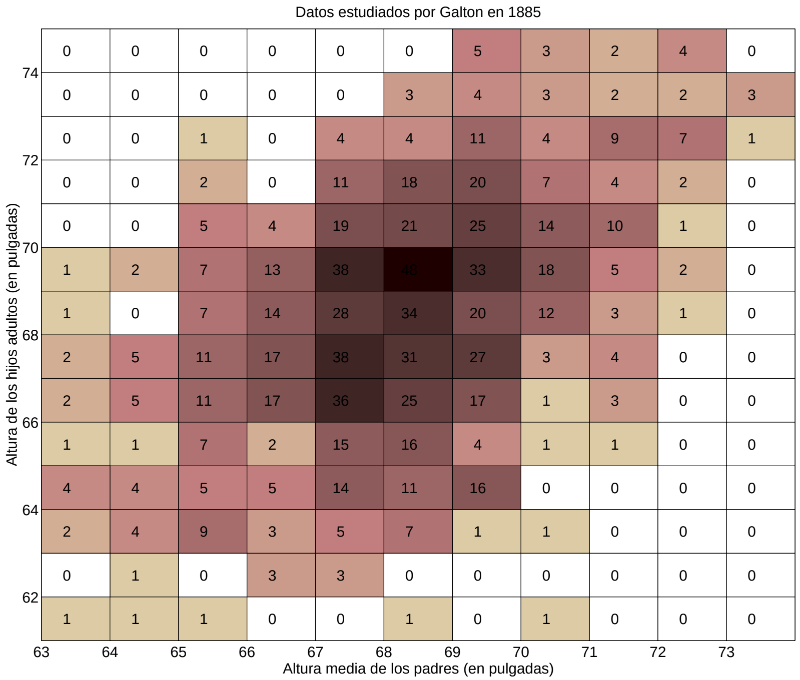


Figura 3.3: Frecuencias conjuntas de los datos recogidos por Sir Francis Galton (s. XIX)

donde β_0 y β_1 son parámetros desconocidos y e_i es una perturbación aleatoria que tiene esperanza 0 y varianza σ^2:

$$E(e_i) = 0,\ E(y_i/X = x_i) = \beta_0 + \beta_1 x_i,\ V(y_i/X = x_i) = V(e_i) = \sigma^2$$

- Variable dependiente y variable independiente.
- Diagrama de dispersión de dos variables. Relación lineal entre variables. Ejemplo clásico de las estaturas de padres e hijos estudiadas por Galton en el siglo XIX.
- La recta de regresión como recta de las esperanzas condicionadas.
- Residuos para valores dados de los coeficientes.

3.6.2 Estimación del modelo

- Estimación por mínimos cuadrados.

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

  Solución:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_x^2},\ \hat{\beta}_0 = \bar{y} - \beta_1 \bar{x}$$

- Residuos mínimo cuadráticos. Varianza residual.

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i,\ \hat{e}_i = y_i - \hat{y}_i,\ S_e^2 = \sum_{i=1}^{n} \hat{e}_i^2$$

- Variabilidad explicada y no explicada por el modelo.

$$VT = S_y^2,\ VE = S_{\hat{y}}^2,\ VNE = S_e^2$$

$$VT = VE + VNE \Longrightarrow 1 = \frac{VE}{VT} + \frac{VNE}{VT}$$

- Coeficiente de determinación R^2.

$$R^2 = \frac{VE}{VT}$$

- Propiedades de los estimadores de los coeficientes del modelo. $\hat{\beta}_0$ y $\hat{\beta}_1$ son insesgados.

- Propiedades de la varianza residual: s_e^2 es estimador insesgado de σ^2.

- Coeficiente de correlación: relación con los coeficientes del modelo.

$$r_{xy} = \frac{S_{xy}}{S_x S_y},\ r_{xy} = \frac{S_{xy}}{S_x^2}\frac{S_x}{S_y} = \hat{\beta}_1 \frac{S_x}{S_y},\ r_{xy}^2 = R^2$$

- Predicción de una nueva observación.
  Disponemos de una nueva observación x_h de la variable X. Queremos predecir el valor de la variable Y en el nuevo individuo. Sabemos que

$$y_h = \beta_0 + \beta_1 x_h + e_h$$

  Entonces el valor esperado de la nueva observación será

$$\mu_h = E(y_h) = \beta_0 + \beta_1 x_h$$

El valor del parámetro μ_h se puede estimar a partir de la estimación del modelo de regresión:

$$\hat{\mu}_h = \hat{\beta}_0 + \hat{\beta}_1 x_h$$

Si queremos predecir el valor de y_h, hemos de observar que

$$y_h = \mu_h + e_h$$

La predicción de y_h será la suma de la estimación de μ_h y de la predicción del valor del residuo e_h. Pero la mejor forma de predecir el residuo e_h es prediciéndolo con el valor 0, porque e_h es una perturbación de la que no conocemos nada, salvo que tiene valor esperado 0. Así, la predicción de y_h será

$$\hat{y}_h = \hat{\mu}_h = \hat{\beta}_0 + \hat{\beta}_1 x_h$$

Obsérvese que la estimación de μ_h será más precisa que la predicción de y_h, porque al predecir y_h hay dos fuentes de incertidumbre: priermo, la asociada a la estimación de μ_h y, segundo, el error que podemos cometer al predecir e_h por 0. Por lo tanto, los intervalos de confianza para μ_h serán más estrechos que los intervalos de confianza para y_h.

3.6.3 Inferencia bajo normalidad

- Hipótesis adicional: los residuos e_h son normales.

- Distribución en el muestreo de los estimadores de los coeficientes: $b\hat{e}ta_0$ y $b\hat{e}ta_1$ son normales.

$$\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2}{n}\left(1 + \left(\frac{\bar{x}}{s_x}\right)^2\right)\right)$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{n\ s_x^2}\right)$$

- Intervalos de confianza: $\hat{\theta} \mp z_\alpha EE(\hat{\theta})$.

- Intervalos de confianza para la predicción de una nueva observación.

- Contrastes de hipótesis.

```
 C1  C2
185  82
186  75
180  70
170  63
162  54

Regresión simple de  C2 sobre C1.

The regression equation is
C2 = - 108 + 1.00 C1

Predictor       Coef       Stdev    t-ratio        p
Constant     -107.97       30.77      -3.51    0.039
C1            1.0009      0.1740       5.75    0.010

s = 3.596       R-sq = 91.7%     R-sq(adj) = 88.9%
```

Regresión simple de y sobre x
`The regression equation is`
`y =` $\hat{\beta}_0$ `+` $\hat{\beta}_1 x$

Predictor (vars. indeps.)	Coef (estimadores)	Stdev (error estándar estimado)	t-ratio=Coef/Stdev (estadísticos t)	p (p-valores)
Constant	$\hat{\beta}_0$	$s_e(1/n+\bar{x}^2/S_{xx})$	t_{β_0}	$H_0:\beta_0=0$
x	$\hat{\beta}_1$	s_e/S_{xx}	t_{β_1}	$H_0:\beta_1=0$

`s =` $s_e=\hat{\sigma}=\sqrt{\frac{1}{n-2}\sum_{i=1}^n \hat{\epsilon}_i^2}$

`R-sq =` $R^2=\frac{VE}{VT}=1-\frac{VNE}{VT}$ (en porcentaje)

`R-sq(adj) =` R^2 ajustado $=1-\frac{VNE/(n-2)}{VT/(n-1)}$ (en porcentaje)

Tabla 3.5: Regresión simple: relación entre el Minitab y la presentación usual del modelo.

3.6.4 Residuos

- Propiedades elementales.
- Análisis gráfico: diagnosis del modelo.
- Observaciones atípicas e influyentes.

3.6.5 Extensiones del modelo

REFERENCIA: El capítulo 14 de Peña (1995), Vol. II, contine este material ampliado

- Heterocedasticidad. Estimación por mínimos cuadrados generalizados. Transformaciones de los datos para conseguir homocedasticidad.
- Relación no lineal entre las variables. Transformaciones para conseguir linealidad. Modelo de regresión polinómica.
- Regresión múltiple.