Muestras sesgadas

Jorge Navarro[1,2]

---

[1]Universidad de Murcia, Spain. E-mail: jorgenav@um.es
[2]Partially Supported by Ministerio de Ciencia y Tecnología under grant MTM2006-12834

Muestras sesgadas y muestras censuradas

- $(X_1, ..., X_n)$ m.a.s. de X

Muestras sesgadas y muestras censuradas

- $(X_1, ..., X_n)$ m.a.s. de X
- $(X_1, ..., X_n)$ v.a.i.i.d. con $\Pr(X_i \leq x) = \Pr(X \leq x)$

Muestras sesgadas y muestras censuradas

- $(X_1, ..., X_n)$ m.a.s. de X
- $(X_1, ..., X_n)$ v.a.i.i.d. con $\Pr(X_i \leq x) = \Pr(X \leq x)$
- Datos censurados: No todos X_i observables.

Muestras sesgadas y muestras censuradas

- $(X_1, ..., X_n)$ m.a.s. de X
- $(X_1, ..., X_n)$ v.a.i.i.d. con $\Pr(X_i \leq x) = \Pr(X \leq x)$
- Datos censurados: No todos X_i observables.
- Ejemplo: X =duración de...

$$2, 3, 5, 6, 7, ..., 1^+, 3^+, 4^+, ...$$

Muestras sesgadas y muestras censuradas

- $(X_1, ..., X_n)$ m.a.s. de X
- $(X_1, ..., X_n)$ v.a.i.i.d. con $\Pr(X_i \leq x) = \Pr(X \leq x)$
- Datos censurados: No todos X_i observables.
- Ejemplo: $X =$duración de...

$$2, 3, 5, 6, 7, ..., 1^+, 3^+, 4^+, ...$$

- 1^+ representa $X_i > 1$

Muestras sesgadas y muestras censuradas

- $(X_1,...,X_n)$ m.a.s. de X
- $(X_1,...,X_n)$ v.a.i.i.d. con $\Pr(X_i \leq x) = \Pr(X \leq x)$
- Datos censurados: No todos X_i observables.
- Ejemplo: X =duración de...

$$2, 3, 5, 6, 7, ..., 1^+, 3^+, 4^+, ...$$

- 1^+ representa $X_i > 1$
- Estos datos no se deben tirar

Muestras sesgadas y muestras censuradas

- $(X_1, ..., X_n)$ m.a.s. de X
- $(X_1, ..., X_n)$ v.a.i.i.d. con $\Pr(X_i \leq x) = \Pr(X \leq x)$
- Datos censurados: No todos X_i observables.
- Ejemplo: X =duración de...

$$2, 3, 5, 6, 7, ..., 1^+, 3^+, 4^+, ...$$

- 1^+ representa $X_i > 1$
- Estos datos no se deben tirar
- Datos sesgados: No todos tienen la misma probabilidad de aparecer en la muestra

Muestras sesgadas y muestras censuradas

- $(X_1, ..., X_n)$ m.a.s. de X
- $(X_1, ..., X_n)$ v.a.i.i.d. con $\Pr(X_i \leq x) = \Pr(X \leq x)$
- Datos censurados: No todos X_i observables.
- Ejemplo: X =duración de...

$$2, 3, 5, 6, 7, ..., 1^+, 3^+, 4^+, ...$$

- 1^+ representa $X_i > 1$
- Estos datos no se deben tirar
- Datos sesgados: No todos tienen la misma probabilidad de aparecer en la muestra
- Ejemplo: Muestrear a familias mediante sus hijos

Muestras sesgadas y muestras censuradas

- $(X_1, ..., X_n)$ m.a.s. de X
- $(X_1, ..., X_n)$ v.a.i.i.d. con $\Pr(X_i \le x) = \Pr(X \le x)$
- Datos censurados: No todos X_i observables.
- Ejemplo: X =duración de...

$$2, 3, 5, 6, 7, ..., 1^+, 3^+, 4^+, ...$$

- 1^+ representa $X_i > 1$
- Estos datos no se deben tirar
- Datos sesgados: No todos tienen la misma probabilidad de aparecer en la muestra
- Ejemplo: Muestrear a familias mediante sus hijos
- "Métodos de conocimiento" (de los valores muestrales)

Modelo Muestras sesgadas

- Ronald Aylmer, Fisher (1934). *Ann. Eugenics* 6, 13-25.

Modelo Muestras sesgadas

- Ronald Aylmer, Fisher (1934). *Ann. Eugenics* 6, 13-25.
- C.R. Rao (1965). *Sankhya Ser. A* 27, 311-324.

Modelo Muestras sesgadas

- Ronald Aylmer, Fisher (1934). *Ann. Eugenics* 6, 13-25.
- C.R. Rao (1965). *Sankhya Ser. A* 27, 311-324.
- Y distribución sesgada asociada a X y $w(t) \geq 0$ con

$$f_Y(t) = \frac{w(t) f_X(t)}{E(w(X))}$$

Modelo Muestras sesgadas

- Ronald Aylmer, Fisher (1934). *Ann. Eugenics* 6, 13-25.
- C.R. Rao (1965). *Sankhya Ser. A* 27, 311-324.
- Y distribución sesgada asociada a X y $w(t) \geq 0$ con

$$f_Y(t) = \frac{w(t) f_X(t)}{E(w(X))}$$

- La probabilidad de observar $X_i = t$ es proporcional a $w(t)$

Modelo Muestras sesgadas

- Ronald Aylmer, Fisher (1934). *Ann. Eugenics* 6, 13-25.
- C.R. Rao (1965). *Sankhya Ser. A* 27, 311-324.
- Y distribución sesgada asociada a X y $w(t) \geq 0$ con

$$f_Y(t) = \frac{w(t) f_X(t)}{E(w(X))}$$

- La probabilidad de observar $X_i = t$ es proporcional a $w(t)$
- Estudiar X mediante $(Y_1, ..., Y_n)$ m.a.s. de Y

Modelo Muestras sesgadas

- Ronald Aylmer, Fisher (1934). *Ann. Eugenics* 6, 13-25.
- C.R. Rao (1965). *Sankhya Ser. A* 27, 311-324.
- Y distribución sesgada asociada a X y $w(t) \geq 0$ con

$$f_Y(t) = \frac{w(t) f_X(t)}{E(w(X))}$$

- La probabilidad de observar $X_i = t$ es proporcional a $w(t)$
- Estudiar X mediante $(Y_1, ..., Y_n)$ m.a.s. de Y
- Datos censurados en A: $w(t) = 1$ si $t \in A$ (0 si no)

Modelo Muestras sesgadas

- Ronald Aylmer, Fisher (1934). *Ann. Eugenics* 6, 13-25.
- C.R. Rao (1965). *Sankhya Ser. A* 27, 311-324.
- Y distribución sesgada asociada a X y $w(t) \geq 0$ con

$$f_Y(t) = \frac{w(t) f_X(t)}{E(w(X))}$$

- La probabilidad de observar $X_i = t$ es proporcional a $w(t)$
- Estudiar X mediante $(Y_1, ..., Y_n)$ m.a.s. de Y
- Datos censurados en A: $w(t) = 1$ si $t \in A$ (0 si no)
- Datos sesgados: la probabilidad de observar X_i es proporcional a $w(X_i)$

Modelo Muestras sesgadas

- Ronald Aylmer, Fisher (1934). *Ann. Eugenics* 6, 13-25.
- C.R. Rao (1965). *Sankhya Ser. A* 27, 311-324.
- Y distribución sesgada asociada a X y $w(t) \geq 0$ con

$$f_Y(t) = \frac{w(t) f_X(t)}{E(w(X))}$$

- La probabilidad de observar $X_i = t$ es proporcional a $w(t)$
- Estudiar X mediante $(Y_1, ..., Y_n)$ m.a.s. de Y
- Datos censurados en A: $w(t) = 1$ si $t \in A$ (0 si no)
- Datos sesgados: la probabilidad de observar X_i es proporcional a $w(X_i)$
- Datos sesgados en longitud: $w(t) = t$

Modelo Muestras sesgadas

- Ronald Aylmer, Fisher (1934). *Ann. Eugenics* 6, 13-25.
- C.R. Rao (1965). *Sankhya Ser. A* 27, 311-324.
- Y distribución sesgada asociada a X y $w(t) \geq 0$ con

$$f_Y(t) = \frac{w(t) f_X(t)}{E(w(X))}$$

- La probabilidad de observar $X_i = t$ es proporcional a $w(t)$
- Estudiar X mediante $(Y_1, ..., Y_n)$ m.a.s. de Y
- Datos censurados en A: $w(t) = 1$ si $t \in A$ (0 si no)
- Datos sesgados: la probabilidad de observar X_i es proporcional a $w(X_i)$
- Datos sesgados en longitud: $w(t) = t$
- Estudiar X mediante $(X_1, ..., X_n)$ e $(Y_1, ..., Y_m)$

Modelo Muestras sesgadas

- Ronald Aylmer, Fisher (1934). *Ann. Eugenics* 6, 13-25.
- C.R. Rao (1965). *Sankhya Ser. A* 27, 311-324.
- Y distribución sesgada asociada a X y $w(t) \geq 0$ con

$$f_Y(t) = \frac{w(t) f_X(t)}{E(w(X))}$$

- La probabilidad de observar $X_i = t$ es proporcional a $w(t)$
- Estudiar X mediante $(Y_1, ..., Y_n)$ m.a.s. de Y
- Datos censurados en A: $w(t) = 1$ si $t \in A$ (0 si no)
- Datos sesgados: la probabilidad de observar X_i es proporcional a $w(X_i)$
- Datos sesgados en longitud: $w(t) = t$
- Estudiar X mediante $(X_1, ..., X_n)$ e $(Y_1, ..., Y_m)$
- ¡Los datos no concuerdan!

Modelo Muestras sesgadas

- Ronald Aylmer, Fisher (1934). *Ann. Eugenics* 6, 13-25.
- C.R. Rao (1965). *Sankhya Ser. A* 27, 311-324.
- Y distribución sesgada asociada a X y $w(t) \geq 0$ con

$$f_Y(t) = \frac{w(t)f_X(t)}{E(w(X))}$$

- La probabilidad de observar $X_i = t$ es proporcional a $w(t)$
- Estudiar X mediante $(Y_1, ..., Y_n)$ m.a.s. de Y
- Datos censurados en A: $w(t) = 1$ si $t \in A$ (0 si no)
- Datos sesgados: la probabilidad de observar X_i es proporcional a $w(X_i)$
- Datos sesgados en longitud: $w(t) = t$
- Estudiar X mediante $(X_1, ..., X_n)$ e $(Y_1, ..., Y_m)$
- ¡Los datos no concuerdan!
- ¿Cuál es mejor?

Ejemplo Rao

- Rao, C.R. (1977). A natural example of a weighted Binomial Distribution. The American Statistician 31, 24-26.
- Encuesta ($^{(*)}$incluyendose a uno mismo):

Sexo	Hermanos*	Hermanas*	Total
H	Y_i	X_i	$m_i = X_i + Y_i$
M	-	-	-

Ejemplo Rao

- Rao, C.R. (1977). A natural example of a weighted Binomial Distribution. The American Statistician 31, 24-26.
- Encuesta ($^{(*)}$incluyendose a uno mismo):

Sexo	Hermanos*	Hermanas*	Total
H	Y_i	X_i	$m_i = X_i + Y_i$
M	-	-	-

- Predicciones (sólo hombres)

Ejemplo Rao

- Rao, C.R. (1977). A natural example of a weighted Binomial Distribution. The American Statistician 31, 24-26.
- Encuesta ($^{(*)}$incluyendose a uno mismo):

Sexo	Hermanos*	Hermanas*	Total
H	Y_i	X_i	$m_i = X_i + Y_i$
M	-	-	-

- Predicciones (sólo hombres)
  1. $H = \sum Y_i >> M = \sum X_i$

Ejemplo Rao

- Rao, C.R. (1977). A natural example of a weighted Binomial Distribution. The American Statistician 31, 24-26.
- Encuesta ($^{(*)}$incluyendose a uno mismo):

Sexo	Hermanos*	Hermanas*	Total
H	Y_i	X_i	$m_i = X_i + Y_i$
M	-	-	-

- Predicciones (sólo hombres)
  1. $H = \sum Y_i >> M = \sum X_i$
  2. $H - M = \sum Y_i - \sum X_i \simeq k = n^o encuestados$

Ejemplo Rao

- Rao, C.R. (1977). A natural example of a weighted Binomial Distribution. The American Statistician 31, 24-26.
- Encuesta ($^{(*)}$incluyendose a uno mismo):

Sexo	Hermanos*	Hermanas*	Total
H	Y_i	X_i	$m_i = X_i + Y_i$
M	-	-	-

- Predicciones (sólo hombres)
  1. $H = \sum Y_i >> M = \sum X_i$
  2. $H - M = \sum Y_i - \sum X_i \simeq k = n^o encuestados$
  3. $H/N = (\sum Y_i)/(\sum m\ _i) >> 0.5,\ N = \sum m\ _i = H + M$

Ejemplo Rao

- Rao, C.R. (1977). A natural example of a weighted Binomial Distribution. The American Statistician 31, 24-26.
- Encuesta ($^{(*)}$incluyendose a uno mismo):

Sexo	Hermanos*	Hermanas*	Total
H	Y_i	X_i	$m_i = X_i + Y_i$
M	-	-	-

- Predicciones (sólo hombres)
  1. $H = \sum Y_i >> M = \sum X_i$
  2. $H - M = \sum Y_i - \sum X_i \simeq k = n^o encuestados$
  3. $H/N = (\sum Y_i)/(\sum m\ _i) >> 0.5,\ N = \sum m\ _i = H + M$
  4. $H/N = (\sum Y_i)/(\sum m\ _i) \simeq 0.5 + \dfrac{k}{2\sum m_i}$

Ejemplo Rao

- Rao, C.R. (1977). A natural example of a weighted Binomial Distribution. The American Statistician 31, 24-26.
- Encuesta ($^{(*)}$incluyendose a uno mismo):

Sexo	Hermanos*	Hermanas*	Total
H	Y_i	X_i	$m_i = X_i + Y_i$
M	-	-	-

- Predicciones (sólo hombres)
  1. $H = \sum Y_i >> M = \sum X_i$
  2. $H - M = \sum Y_i - \sum X_i \simeq k = n^o encuestados$
  3. $H/N = (\sum Y_i)/(\sum m\ _i) >> 0.5,\ N = \sum m\ _i = H + M$
  4. $H/N = (\sum Y_i)/(\sum m\ _i) \simeq 0.5 + \dfrac{k}{2\sum m_i}$
  5. $\dfrac{H-k}{N-k} = \dfrac{\sum Y_i - k}{\sum m_i - k} \simeq 0.5$

Resultados de Rao

Ciudad	N	H	M	H-M	k	H/N	$0.5+\frac{k}{2N}$	$\frac{H-k}{N-k}$
Tehran	105	65	40	25	21	0.619	0.600	0.524
Isphaha	77	45	32	13	11	0.584	0.571	0.515
Tokyo	124	90	34	56	50	0.726	0.701	0.540
Delhi	158	92	66	26	29	0.582	0.592	0.488
Calcutta	726	414	312	102	104	0.570	0.571	0.498
Waltair	211	123	88	35	39	0.583	0.592	0.488
Ahmedabad	133	84	49	35	29	0.632	0.609	0.529
Bangalore	307	180	127	53	55	0.586	0.589	0.496

Cuestiones

- ¿Cómo estimar p_H o p_M?

Cuestiones

- ¿Cómo estimar p_H o p_M?
- ¿Cómo estimar $E(m_i)$?

Cuestiones

- ¿Cómo estimar p_H o p_M?
- ¿Cómo estimar $E(m_i)$?
- ¿Cuál de las dos muestras (H o M) es mejor?

Cuestiones

- ¿Cómo estimar p_H o p_M?
- ¿Cómo estimar $E(m_i)$?
- ¿Cuál de las dos muestras (H o M) es mejor?
- ¿Se pueden usar ambas muestras de forma conjunta?

Cuestiones

- ¿Cómo estimar p_H o p_M?
- ¿Cómo estimar $E(m_i)$?
- ¿Cuál de las dos muestras (H o M) es mejor?
- ¿Se pueden usar ambas muestras de forma conjunta?
- ¿Cómo se obtinen las predicciones?

Soluciones

- X es Binomial $B(m, p_H)$, con $p_H \simeq 0.5$

$$\begin{aligned} p(x) &= \Pr(X = x) = \binom{m}{x} p_H^x \cdot {p_M}^{m-x} \\ E(X) &= mp_H \end{aligned}$$

Soluciones

- X es Binomial $B(m, p_H)$, con $p_H \simeq 0.5$

$$\begin{aligned} p(x) &= \Pr(X = x) = \binom{m}{x} p_H^x \cdot {p_M}^{m-x} \\ E(X) &= m p_H \end{aligned}$$

- La probabilidad de que Y_i aparezca en la muestra es proporcional a Y_i

Soluciones

- X es Binomial $B(m, p_H)$, con $p_H \simeq 0.5$

$$\begin{aligned} p(x) &= \Pr(X = x) = \binom{m}{x} p_H^x \cdot {p_M}^{m-x} \\ E(X) &= mp_H \end{aligned}$$

- La probabilidad de que Y_i aparezca en la muestra es proporcional a Y_i
- Si Y es una Binomial sesgada $Y \equiv B^*(m, p_H)$, con $w(t) = t$

$$\begin{aligned} p^*(x) = \frac{xp(x)}{E(X)} &= x\binom{m}{x} p_H^x \cdot {p_M}^{m-x}/(mp_H) \\ &= x\frac{xm!}{mx!(m-x)!} p_H^{x-1} p_M^{m-x} = \binom{m-1}{x-1} p_H^{x-1} p_M^{m-x}, \;\; x = 1, 2, .. \end{aligned}$$

Soluciones

- X es Binomial $B(m, p_H)$, con $p_H \simeq 0.5$

$$\begin{aligned} p(x) &= \Pr(X = x) = \binom{m}{x} p_H^x \cdot {p_M}^{m-x} \\ E(X) &= mp_H \end{aligned}$$

- La probabilidad de que Y_i aparezca en la muestra es proporcional a Y_i
- Si Y es una Binomial sesgada $Y \equiv B^*(m, p_H)$, con $w(t) = t$

$$p^*(x) = \frac{xp(x)}{E(X)} = x\binom{m}{x} p_H^x \cdot {p_M}^{m-x}/(mp_H)$$
$$= x\frac{xm!}{mx!(m-x)!} p_H^{x-1} p_M^{m-x} = \binom{m-1}{x-1} p_H^{x-1} p_M^{m-x}, \quad x = 1, 2, ..$$

- $Y - 1 \equiv B(m-1, p_H)$

Explicación predicciones.

- $Y_i - 1 \equiv B(m_i - 1, p_H)$

Explicación predicciones.

- $Y_i - 1 \equiv B(m_i - 1, p_H)$
- $E(Y_i) = 1 + (m_i - 1)p_H = 1 - p_H + m_i p_H$

Explicación predicciones.

- $Y_i - 1 \equiv B(m_i - 1, p_H)$
- $E(Y_i) = 1 + (m_i - 1)p_H = 1 - p_H + m_i p_H$
- $X_i \equiv B(m_i - 1, p_M)$

Explicación predicciones.

- $Y_i - 1 \equiv B(m_i - 1, p_H)$
- $E(Y_i) = 1 + (m_i - 1)p_H = 1 - p_H + m_i p_H$
- $X_i \equiv B(m_i - 1, p_M)$
- $E(X_i) = (m_i - 1)p_M$

Explicación predicciones.

- $Y_i - 1 \equiv B(m_i - 1, p_H)$
- $E(Y_i) = 1 + (m_i - 1)p_H = 1 - p_H + m_i p_H$
- $X_i \equiv B(m_i - 1, p_M)$
- $E(X_i) = (m_i - 1)p_M$
- $E(\sum Y_i) = \sum E(Y_i) = \sum(1 - p_H + m_i p_H) = k(1 - p_H) + p_H \sum m_i$

Explicación predicciones.

- $Y_i - 1 \equiv B(m_i - 1, p_H)$
- $E(Y_i) = 1 + (m_i - 1)p_H = 1 - p_H + m_i p_H$
- $X_i \equiv B(m_i - 1, p_M)$
- $E(X_i) = (m_i - 1)p_M$
- $E(\sum Y_i) = \sum E(Y_i) = \sum(1 - p_H + m_i p_H) = k(1 - p_H) + p_H \sum m_i$
- $E(\sum X_i) = \sum E(X_i) = \sum(m_i - 1)p_M = p_M \sum m_i - kp_M$

Explicación predicciones.

- $Y_i - 1 \equiv B(m_i - 1, p_H)$
- $E(Y_i) = 1 + (m_i - 1)p_H = 1 - p_H + m_i p_H$
- $X_i \equiv B(m_i - 1, p_M)$
- $E(X_i) = (m_i - 1)p_M$
- $E(\sum Y_i) = \sum E(Y_i) = \sum(1 - p_H + m_i p_H) = k(1 - p_H) + p_H \sum m_i$
- $E(\sum X_i) = \sum E(X_i) = \sum(m_i - 1)p_M = p_M \sum m_i - kp_M$
- $E(\sum Y_i - \sum X_i) = 2kp_M \simeq k$

Explicación predicciones.

- $Y_i - 1 \equiv B(m_i - 1, p_H)$
- $E(Y_i) = 1 + (m_i - 1)p_H = 1 - p_H + m_i p_H$
- $X_i \equiv B(m_i - 1, p_M)$
- $E(X_i) = (m_i - 1)p_M$
- $E(\sum Y_i) = \sum E(Y_i) = \sum(1 - p_H + m_i p_H) = k(1 - p_H) + p_H \sum m_i$
- $E(\sum X_i) = \sum E(X_i) = \sum(m_i - 1)p_M = p_M \sum m_i - kp_M$
- $E(\sum Y_i - \sum X_i) = 2kp_M \simeq k$
- $E\left(\frac{\sum Y_i}{\sum m_i}\right) = \frac{kp_M + p_H \sum m_i}{\sum m_i} = p_H + \frac{kp_M}{\sum m_i} \simeq 0.5 + \frac{k}{2\sum m_i}$

Explicación predicciones.

- $Y_i - 1 \equiv B(m_i - 1, p_H)$
- $E(Y_i) = 1 + (m_i - 1)p_H = 1 - p_H + m_i p_H$
- $X_i \equiv B(m_i - 1, p_M)$
- $E(X_i) = (m_i - 1)p_M$
- $E(\sum Y_i) = \sum E(Y_i) = \sum(1 - p_H + m_i p_H) = k(1 - p_H) + p_H \sum m_i$
- $E(\sum X_i) = \sum E(X_i) = \sum(m_i - 1)p_M = p_M \sum m_i - kp_M$
- $E(\sum Y_i - \sum X_i) = 2kp_M \simeq k$
- $E\left(\frac{\sum Y_i}{\sum m_i}\right) = \frac{kp_M + p_H \sum m_i}{\sum m_i} = p_H + \frac{kp_M}{\sum m_i} \simeq 0.5 + \frac{k}{2\sum m_i}$
- $E\left(\frac{\sum Y_i - k}{\sum m_i - k}\right) = \frac{k(1 - p_H) + p_H \sum m_i - k}{\sum m_i - k} = p_H \simeq 0.5$

Explicación de las cuestiones

- ¿Cómo estimar p_H?

Explicación de las cuestiones

- ¿Cómo estimar p_H?
- Podemos usar

$$\begin{aligned}
T &= \frac{\sum Y_i - k}{\sum m_i - k} \\
E(T) &= E\left(\frac{\sum Y_i - k}{\sum m_i - k}\right) = p_H \\
Vat(T) &= p_H p_M / \left(\sum m_i - k\right) \to 0 \\
\sum Y_i - k &\equiv B\left(\sum m_i - k, p_H\right) \\
T &\approxeq Normal \\
T & \quad \text{es un EIMV}
\end{aligned}$$

Explicación de las cuestiones

- ¿Cómo usar ambas muestras de forma conjunta?

Explicación de las cuestiones

- ¿Cómo usar ambas muestras de forma conjunta?
- $X_1 ... X_n$ m.a.s. insesgada (de familias) $X_i \equiv B(n_i, p)$

Explicación de las cuestiones

- ¿Cómo usar ambas muestras de forma conjunta?
- $X_1 ... X_n$ m.a.s. insesgada (de familias) $X_i \equiv B(n_i, p)$
- $Y_1 ... Y_m$ m.a.s. sesgada (sólo de hombres)

Explicación de las cuestiones

- ¿Cómo usar ambas muestras de forma conjunta?
- $X_1 \dots X_n$ m.a.s. insesgada (de familias) $X_i \equiv B(n_i, p)$
- $Y_1 \dots Y_m$ m.a.s. sesgada (sólo de hombres)
- $Y_j - 1 \equiv B(m_j - 1, p)$

$$\begin{aligned}
T &= \frac{\sum X_i + \sum (Y_j - 1)}{\sum n_i + \sum (m_j - 1)} \\
E(T) &= E(\frac{\sum X_i + \sum (Y_j - 1)}{\sum n_i + \sum (m_j - 1)}) = p \\
Vat(T) &= p(1-p)/(\sum n_i - \sum (m_j - 1)) \\
\sum X_i + \sum (Y_j - 1) &\equiv B(\sum n_i - \sum (m_j - 1), p_H) \\
T &\approxeq Normal \\
T & \quad \text{es un EIMV}
\end{aligned}$$

Explicación de las cuestiones

- ¿Cuál de las dos muestras es mejor?

Explicación de las cuestiones

- ¿Cuál de las dos muestras es mejor?
- Nótese que si $Y_j = 1$, entonces no aporta ninguna información

Explicación de las cuestiones

- ¿Cuál de las dos muestras es mejor?
- Nótese que si $Y_j = 1$, entonces no aporta ninguna información
- X_i aporta más información que Y_j si $n_i > m_j - 1$

Explicación de las cuestiones

- ¿Cuál de las dos muestras es mejor?
- Nótese que si $Y_j = 1$, entonces no aporta ninguna información
- X_i aporta más información que Y_j si $n_i > m_j - 1$
- Las informaciones de Fisher valen:

$$\begin{aligned} I_{X_i}(p) &= \frac{n_i}{pq} \\ I_{Y_j}(p) &= \frac{m_j - 1}{pq} \end{aligned}$$

Explicación de las cuestiones

- ¿Cuál de las dos muestras es mejor?
- Nótese que si $Y_j = 1$, entonces no aporta ninguna información
- X_i aporta más información que Y_j si $n_i > m_j - 1$
- Las informaciones de Fisher valen:

$$\begin{aligned} I_{X_i}(p) &= \frac{n_i}{pq} \\ I_{Y_j}(p) &= \frac{m_j - 1}{pq} \end{aligned}$$

- $E(n_i) =?, E(m_j) =?$ $(m_j \geq 1)$

Explicación de las cuestiones

- ¿Cuál de las dos muestras es mejor?
- Nótese que si $Y_j = 1$, entonces no aporta ninguna información
- X_i aporta más información que Y_j si $n_i > m_j - 1$
- Las informaciones de Fisher valen:

$$
\begin{aligned}
I_{X_i}(p) &= \frac{n_i}{pq} \\
I_{Y_j}(p) &= \frac{m_j - 1}{pq}
\end{aligned}
$$

- $E(n_i) =?, E(m_j) =?$ $(m_j \geq 1)$
- En nuestra "encuesta" son iguales (sería mejor incluir a todos si no hay sesgo)

Explicación de las cuestiones

- ¿Cuál de las dos muestras es mejor?
- Nótese que si $Y_j = 1$, entonces no aporta ninguna información
- X_i aporta más información que Y_j si $n_i > m_j - 1$
- Las informaciones de Fisher valen:

$$
\begin{aligned}
I_{X_i}(p) &= \frac{n_i}{pq} \\
I_{Y_j}(p) &= \frac{m_j - 1}{pq}
\end{aligned}
$$

- $E(n_i) = ?, E(m_j) = ?$ $(m_j \geq 1)$
- En nuestra "encuesta" son iguales (sería mejor incluir a todos si no hay sesgo)
- Pero, si sospechamos sesgo en nuestros estudiantes (el sexo influye en la elección de la carrera), es mejor usar la muestra sesgada (cuyo sesgo es conocido)

Explicación de las cuestiones

- ¿Cómo estimar el tamaño familiar n?

Explicación de las cuestiones

- ¿Cómo estimar el tamaño familiar n?
- ¿Podemos usar $\overline{m} = \frac{1}{k} \sum m_i$?

Explicación de las cuestiones

- ¿Cómo estimar el tamaño familiar n?
- ¿Podemos usar $\overline{m} = \frac{1}{k} \sum m_i$?
- Si usamos hombres y mujeres, la probabilidad de que una familia de tamaño m_i aparezca en la muestra es proporcional a m_i

Explicación de las cuestiones

- ¿Cómo estimar el tamaño familiar n?
- ¿Podemos usar $\overline{m} = \frac{1}{k} \sum m_i$?
- Si usamos hombres y mujeres, la probabilidad de que una familia de tamaño m_i aparezca en la muestra es proporcional a m_i
- Si sólo usamos hombres, será proporcional a $E(X_i) = m_i p_H$

Explicación de las cuestiones

- ¿Cómo estimar el tamaño familiar n?
- ¿Podemos usar $\overline{m} = \frac{1}{k} \sum m_i$?
- Si usamos hombres y mujeres, la probabilidad de que una familia de tamaño m_i aparezca en la muestra es proporcional a m_i
- Si sólo usamos hombres, será proporcional a $E(X_i) = m_i p_H$
- Luego $m_1, \ldots, m_k$ será una muestra sesgada en longitud de n

Explicación de las cuestiones

- ¿Cómo estimar el tamaño familiar n?
- ¿Podemos usar $\overline{m} = \frac{1}{k} \sum m_i$?
- Si usamos hombres y mujeres, la probabilidad de que una familia de tamaño m_i aparezca en la muestra es proporcional a m_i
- Si sólo usamos hombres, será proporcional a $E(X_i) = m_i p_H$
- Luego $m_1, \ldots, m_k$ será una muestra sesgada en longitud de n
- ¿Cómo estimar $E(n)$ usando $m_1, \ldots, m_k$?

Explicación de las cuestiones

- Si $n \equiv Poisson(\mu), \mu =$número medio de hijos por familia

$$p(x) = \mu^x e^{-\mu}/x!$$

Explicación de las cuestiones

- Si $n \equiv Poisson(\mu), \mu =$número medio de hijos por familia

$$p(x) = \mu^x e^{-\mu}/x!$$

- Luego $m_j \equiv Poisson\ sesgada$,

$$p^*(x) = \frac{xp(x)}{\mu} = \frac{\mu^x e^{-\mu}}{\mu x!} = \frac{\mu^{x-1} e^{-\mu}}{x!}, x = 1, 2, \ldots$$

Explicación de las cuestiones

- Si $n \equiv Poisson(\mu), \mu =$número medio de hijos por familia

$$p(x) = \mu^x e^{-\mu}/x!$$

- Luego $m_j \equiv Poisson\ sesgada$,

$$p^*(x) = \frac{xp(x)}{\mu} = \frac{\mu^x e^{-\mu}}{\mu x!} = \frac{\mu^{x-1} e^{-\mu}}{x!}, x = 1, 2, \ldots$$

- Es decir: $m_j - 1 \equiv Poisson(\mu)$

Explicación de las cuestiones

- Si $n \equiv Poisson(\mu), \mu =$número medio de hijos por familia

$$p(x) = \mu^x e^{-\mu}/x!$$

- Luego $m_j \equiv Poisson\ sesgada$,

$$p^*(x) = \frac{xp(x)}{\mu} = \frac{\mu^x e^{-\mu}}{\mu x!} = \frac{\mu^{x-1} e^{-\mu}}{x!}, x = 1, 2, \ldots$$

- Es decir: $m_j - 1 \equiv Poisson(\mu)$
- $E(\overline{m}) = \frac{1}{k} \sum E(m_i) = \frac{1}{k} \sum (\mu + 1) = \mu + 1$

Explicación de las cuestiones

- Si $n \equiv Poisson(\mu), \mu =$número medio de hijos por familia

$$p(x) = \mu^x e^{-\mu}/x!$$

- Luego $m_j \equiv Poisson\ sesgada$,

$$p^*(x) = \frac{xp(x)}{\mu} = \frac{\mu^x e^{-\mu}}{\mu x!} = \frac{\mu^{x-1} e^{-\mu}}{x!}, x = 1, 2, \ldots$$

- Es decir: $m_j - 1 \equiv Poisson(\mu)$
- $E(\overline{m}) = \frac{1}{k} \sum E(m_i) = \frac{1}{k} \sum(\mu + 1) = \mu + 1$
- $T = \overline{m} - 1 = \frac{1}{k} \sum(m_i - 1)$

Explicación de las cuestiones

- Si $n \equiv Poisson(\mu), \mu =$número medio de hijos por familia

$$p(x) = \mu^x e^{-\mu}/x!$$

- Luego $m_j \equiv Poisson\ sesgada$,

$$p^*(x) = \frac{xp(x)}{\mu} = \frac{\mu^x e^{-\mu}}{\mu x!} = \frac{\mu^{x-1} e^{-\mu}}{x!}, x = 1, 2, \ldots$$

- Es decir: $m_j - 1 \equiv Poisson(\mu)$
- $E(\overline{m}) = \frac{1}{k} \sum E(m_i) = \frac{1}{k} \sum (\mu + 1) = \mu + 1$
- $T = \overline{m} - 1 = \frac{1}{k} \sum (m_i - 1)$
- $E(T) = \mu$

Explicación de las cuestiones

- Si $n \equiv Poisson(\mu), \mu =$número medio de hijos por familia

$$p(x) = \mu^x e^{-\mu}/x!$$

- Luego $m_j \equiv Poisson\ sesgada$,

$$p^*(x) = \frac{xp(x)}{\mu} = \frac{\mu^x e^{-\mu}}{\mu x!} = \frac{\mu^{x-1} e^{-\mu}}{x!}, x = 1, 2, \ldots$$

- Es decir: $m_j - 1 \equiv Poisson(\mu)$
- $E(\overline{m}) = \frac{1}{k} \sum E(m_i) = \frac{1}{k} \sum(\mu + 1) = \mu + 1$
- $T = \overline{m} - 1 = \frac{1}{k} \sum(m_i - 1)$
- $E(T) = \mu$
- $Var(T) = \mu/k \to 0$

Explicación de las cuestiones

- Si $n \equiv Poisson(\mu), \mu =$número medio de hijos por familia

$$p(x) = \mu^x e^{-\mu}/x!$$

- Luego $m_j \equiv Poisson\ sesgada$,

$$p^*(x) = \frac{xp(x)}{\mu} = \frac{\mu^x e^{-\mu}}{\mu x!} = \frac{\mu^{x-1} e^{-\mu}}{x!}, x = 1, 2, \ldots$$

- Es decir: $m_j - 1 \equiv Poisson(\mu)$
- $E(\overline{m}) = \frac{1}{k}\sum E(m_i) = \frac{1}{k}\sum(\mu + 1) = \mu + 1$
- $T = \overline{m} - 1 = \frac{1}{k}\sum(m_i - 1)$
- $E(T) = \mu$
- $Var(T) = \mu/k \to 0$
- $\sum(m_i - 1) \equiv Poisson(k\mu)$

Explicación de las cuestiones

- Si $n \equiv Poisson(\mu), \mu =$número medio de hijos por familia

$$p(x) = \mu^x e^{-\mu}/x!$$

- Luego $m_j \equiv Poisson\ sesgada$,

$$p^*(x) = \frac{xp(x)}{\mu} = \frac{\mu^x e^{-\mu}}{\mu x!} = \frac{\mu^{x-1} e^{-\mu}}{x!}, x = 1, 2, \ldots$$

- Es decir: $m_j - 1 \equiv Poisson(\mu)$
- $E(\overline{m}) = \frac{1}{k} \sum E(m_i) = \frac{1}{k} \sum(\mu + 1) = \mu + 1$
- $T = \overline{m} - 1 = \frac{1}{k} \sum(m_i - 1)$
- $E(T) = \mu$
- $Var(T) = \mu/k \to 0$
- $\sum(m_i - 1) \equiv Poisson(k\mu)$
- $T \approxeq Normal$

Explicación de las cuestiones

- Nuestro resultado

Explicación de las cuestiones

- Nuestro resultado
- $\overline{m} = \frac{1}{k} \sum m_i = N/k$

Explicación de las cuestiones

- Nuestro resultado
- $\overline{m} = \frac{1}{k} \sum m_i = N/k$
- $T = \overline{m} - 1 = (N - k)/k$

Explicación de las cuestiones

- Nuestro resultado
- $\overline{m} = \frac{1}{k}\sum m_i = N/k$
- $T = \overline{m} - 1 = (N-k)/k$
- Resultados de Rao

Ciudad	N	H	M	k	$\overline{m} = N/k$	$T = \overline{m} - 1$
Tehran	105	65	40	21	5.000	4
Isphahan	77	45	32	11	7.000	6
Tokyo	124	90	34	50	2.480	1.480
Delhi	158	92	66	29	5.448	4.448
Calcutta	726	414	312	104	6.980	5.980
Waltair	211	123	88	39	5.410	4.410
Ahmedabad	133	84	49	29	4.580	3.580
Bangalore	307	180	127	55	5.582	4.582

Ejemplo de Fisher

- R. A., Fisher (1934). The effect of methods of ascertainment upon the estimation of frequencies. *Annals Eugenics* **6**, 13-25.

Ejemplo de Fisher

- R. A., Fisher (1934). The effect of methods of ascertainment upon the estimation of frequencies. *Annals Eugenics* **6**, 13-25.
- Objetivo: Estudiar la proporción p de niños albinos

Ejemplo de Fisher

- R. A., Fisher (1934). The effect of methods of ascertainment upon the estimation of frequencies. *Annals Eugenics* **6**, 13-25.
- Objetivo: Estudiar la proporción p de niños albinos
- Las leyes de Medel aseguran que si es un gen recesivo, entonces, en las familias sin padres albinos pero capaces de tener niños albinos, p debe ser $1/4$

Ejemplo de Fisher

- R. A., Fisher (1934). The effect of methods of ascertainment upon the estimation of frequencies. *Annals Eugenics* **6**, 13-25.
- Objetivo: Estudiar la proporción p de niños albinos
- Las leyes de Medel aseguran que si es un gen recesivo, entonces, en las familias sin padres albinos pero capaces de tener niños albinos, p debe ser $1/4$
- ¡En la práctica no podemos saber cuándo una pareja es capaz de tener un niño albino!

Ejemplo de Fisher

- R. A., Fisher (1934). The effect of methods of ascertainment upon the estimation of frequencies. *Annals Eugenics* **6**, 13-25.
- Objetivo: Estudiar la proporción p de niños albinos
- Las leyes de Medel aseguran que si es un gen recesivo, entonces, en las familias sin padres albinos pero capaces de tener niños albinos, p debe ser $1/4$
- ¡En la práctica no podemos saber cuándo una pareja es capaz de tener un niño albino!
- Por eso, Fisher sólo consideró familias con hijos albinos.

Ejemplo de Fisher

- R. A., Fisher (1934). The effect of methods of ascertainment upon the estimation of frequencies. *Annals Eugenics* **6**, 13-25.
- Objetivo: Estudiar la proporción p de niños albinos
- Las leyes de Medel aseguran que si es un gen recesivo, entonces, en las familias sin padres albinos pero capaces de tener niños albinos, p debe ser $1/4$
- ¡En la práctica no podemos saber cuándo una pareja es capaz de tener un niño albino!
- Por eso, Fisher sólo consideró familias con hijos albinos.
- Como sospechaba que este hecho influiría en la proporción de albinos, sólo considero familias con 5 hijos, obteniendo los datos siguientes:

Ejemplo de Fisher

	Número de niños albinos en la familia					
N	1	2	3	4	5	Total
1	140	80	35	4	0	259
2	-	52	12	7	1	72
3	-	-	7	0	0	7
4	-	-	-	2	0	2
5	-	-	-	-	0	0
Total	140	132	54	13	1	340

- N=Número de niños albinos conocidos de forma independiente

Ejemplo de Fisher

	Número de niños albinos en la familia					
N	1	2	3	4	5	Total
1	140	80	35	4	0	259
2	-	52	12	7	1	72
3	-	-	7	0	0	7
4	-	-	-	2	0	2
5	-	-	-	-	0	0
Total	140	132	54	13	1	340

- N=Número de niños albinos conocidos de forma independiente
- Nótese que tenemos 340 familias obtenidas de 432 niños albinos obtenidos de forma independiente

Solución 1

- ¿Qué hacer con esos datos?

Solución 1

- ¿Qué hacer con esos datos?
- Si $X_1, ..., X_n$ es una m.a.s. de tamaño $n =$340 de una Binomial $B(k = 5, p = 1/4)$

Solución 1

- ¿Qué hacer con esos datos?
- Si $X_1, ..., X_n$ es una m.a.s. de tamaño n =340 de una Binomial $B(k = 5, p = 1/4)$
- La estimación de p es

$$\widehat{p}_1 = \frac{\sum_{i=1}^{n} X_i}{5n} = \frac{140 + 2 \cdot 132 + ...}{5 \cdot 340} = \frac{623}{1700} = 0.3665$$

Solución 1

- ¿Qué hacer con esos datos?
- Si $X_1, ..., X_n$ es una m.a.s. de tamaño $n = 340$ de una Binomial $B(k = 5, p = 1/4)$
- La estimación de p es

$$\widehat{p}_1 = \frac{\sum_{i=1}^{n} X_i}{5n} = \frac{140 + 2 \cdot 132 + ...}{5 \cdot 340} = \frac{623}{1700} = 0.3665$$

- Con varianza

$$\sigma^2(\widehat{p}_1) = \frac{p(1-p)}{5n} = \frac{0.25 \cdot 0.75}{1700} = 0.0001$$

Solución 1

- ¿Qué hacer con esos datos?
- Si $X_1, \ldots, X_n$ es una m.a.s. de tamaño $n =$340 de una Binomial $B(k = 5, p = 1/4)$
- La estimación de p es

$$\widehat{p}_1 = \frac{\sum_{i=1}^{n} X_i}{5n} = \frac{140 + 2 \cdot 132 + \ldots}{5 \cdot 340} = \frac{623}{1700} = 0.3665$$

- Con varianza

$$\sigma^2(\widehat{p}_1) = \frac{p(1-p)}{5n} = \frac{0.25 \cdot 0.75}{1700} = 0.0001$$

- Esto da $2\sigma(\widehat{p}_1) \simeq 0.021$ lo que nos conduce a rechazar $p = 0.25$

Solución 1bis

- Si usamos las familias repetidas (es decir, si más de un niño albino ha sido conocido de forma independiente, entonces la familia se cuenta repetidas veces)

Solución 1bis

- Si usamos las familias repetidas (es decir, si más de un niño albino ha sido conocido de forma independiente, entonces la familia se cuenta repetidas veces)
- Así, tendríamos $n = 432$ y la estimación de p sería

$$\widehat{p}_1 = \frac{\sum_{i=1}^{n} X_i}{5n} = \frac{140 + 2 \cdot 184 + \ldots}{5 \cdot 432} = 0.399$$

Solución 1bis

- Si usamos las familias repetidas (es decir, si más de un niño albino ha sido conocido de forma independiente, entonces la familia se cuenta repetidas veces)
- Así, tendríamos $n = 432$ y la estimación de p sería

$$\widehat{p}_1 = \frac{\sum_{i=1}^{n} X_i}{5n} = \frac{140 + 2 \cdot 184 + \ldots}{5 \cdot 432} = 0.399$$

- Tampoco concuerda con $p = 0.25$

Solución 2

- Se podría decir que las familias con 0 niños albinos no aparecerán nunca en la muestra

Solución 2

- Se podría decir que las familias con 0 niños albinos no aparecerán nunca en la muestra
- Así, tendríamos una muestra censurada con $w(x) = 1$ si $x \neq 0$ y $w(0) = 0$.

Solución 2

- Se podría decir que las familias con 0 niños albinos no aparecerán nunca en la muestra
- Así, tendríamos una muestra censurada con $w(x) = 1$ si $x \neq 0$ y $w(0) = 0$.
- Su f.p.p. sería $p^*(x) = p(x)/(1 - q^5)$, con $p(x)$ la de una Binomial $B(5, 1/4)$

Solución 2

- Se podría decir que las familias con 0 niños albinos no aparecerán nunca en la muestra
- Así, tendríamos una muestra censurada con $w(x) = 1$ si $x \neq 0$ y $w(0) = 0$.
- Su f.p.p. sería $p^*(x) = p(x)/(1 - q^5)$, con $p(x)$ la de una Binomial $B(5, 1/4)$
- Entonces el estimador máximo verosimil se obtendría de la igualdad

$$\frac{p}{1 - q^5} = \frac{\sum_{i=1}^{n} X_i}{5n}$$

Solución 2

- Se podría decir que las familias con 0 niños albinos no aparecerán nunca en la muestra
- Así, tendríamos una muestra censurada con $w(x) = 1$ si $x \neq 0$ y $w(0) = 0$.
- Su f.p.p. sería $p^*(x) = p(x)/(1 - q^5)$, con $p(x)$ la de una Binomial $B(5, 1/4)$
- Entonces el estimador máximo verosimil se obtendría de la igualdad

$$\frac{p}{1 - q^5} = \frac{\sum_{i=1}^{n} X_i}{5n}$$

- Lo que da $\widehat{p}_2 = 0.3085$ ($\widehat{p}_2 = 0.35$ si usamos las familias repetidas).

Solución 2

- Se podría decir que las familias con 0 niños albinos no aparecerán nunca en la muestra
- Así, tendríamos una muestra censurada con $w(x) = 1$ si $x \neq 0$ y $w(0) = 0$.
- Su f.p.p. sería $p^*(x) = p(x)/(1 - q^5)$, con $p(x)$ la de una Binomial $B(5, 1/4)$
- Entonces el estimador máximo verosimil se obtendría de la igualdad

$$\frac{p}{1 - q^5} = \frac{\sum_{i=1}^{n} X_i}{5n}$$

- Lo que da $\widehat{p}_2 = 0.3085$ ($\widehat{p}_2 = 0.35$ si usamos las familias repetidas).
- En ambos casos, rechazaríamos $p = 1/4$.

Solución 3

- Note que la probabilidad de que una familia con x niños albinos aparezca en la muestra es proporcional a x

Solución 3

- Note que la probabilidad de que una familia con x niños albinos aparezca en la muestra es proporcional a x
- La solución correcta es que X_i siguen una Binomial sesgada en longitud X^*

Solución 3

- Note que la probabilidad de que una familia con x niños albinos aparezca en la muestra es proporcional a x
- La solución correcta es que X_i siguen una Binomial sesgada en longitud X^*
- Es decir, $X_i - 1$ sigue una Binomial $B(4, 1/4)$

Solución 3

- Note que la probabilidad de que una familia con x niños albinos aparezca en la muestra es proporcional a x
- La solución correcta es que X_i siguen una Binomial sesgada en longitud X^*
- Es decir, $X_i - 1$ sigue una Binomial $B(4, 1/4)$
- Entonces, usando las familias repetidas, la estimación sería

$$\widehat{p}_3 = \frac{\sum_{i=1}^{n}(X_i - 1)}{4n} = \frac{1 \cdot 184 + 2 \cdot 80 + \dots}{4 \cdot 432} = 0.2488$$

Solución 3

- Note que la probabilidad de que una familia con x niños albinos aparezca en la muestra es proporcional a x
- La solución correcta es que X_i siguen una Binomial sesgada en longitud X^*
- Es decir, $X_i - 1$ sigue una Binomial $B(4, 1/4)$
- Entonces, usando las familias repetidas, la estimación sería

$$\widehat{p}_3 = \frac{\sum_{i=1}^{n}(X_i - 1)}{4n} = \frac{1 \cdot 184 + 2 \cdot 80 + \ldots}{4 \cdot 432} = 0.2488$$

- La varianza verifica $2\sigma(\widehat{p}_3) \simeq 0.0208$, lo que es consistente con $p = 1/4$

Solución 3

- Note que la probabilidad de que una familia con x niños albinos aparezca en la muestra es proporcional a x
- La solución correcta es que X_i siguen una Binomial sesgada en longitud X^*
- Es decir, $X_i - 1$ sigue una Binomial $B(4, 1/4)$
- Entonces, usando las familias repetidas, la estimación sería

$$\widehat{p}_3 = \frac{\sum_{i=1}^{n}(X_i - 1)}{4n} = \frac{1 \cdot 184 + 2 \cdot 80 + \ldots}{4 \cdot 432} = 0.2488$$

- La varianza verifica $2\sigma(\widehat{p}_3) \simeq 0.0208$, lo que es consistente con $p = 1/4$
- Notese que si no usamos las familias repetidas, entonces p es subestimada

$$\widehat{p}_4 = \frac{\sum_{i=1}^{n}(X_i - 1)}{4n} = \frac{1 \cdot 132 + 2 \cdot 54 + \ldots}{4 \cdot 340} = 0.2080$$

Otros ejemplos

- Estudios familiares y genéticos

Otros ejemplos

- Estudios familiares y genéticos
- Tiempos de vida: edades, duraciones, etc.

Otros ejemplos

- Estudios familiares y genéticos
- Tiempos de vida: edades, duraciones, etc.
- Tiempos de espera: autobús, servicios de atención, etc.

Otros ejemplos

- Estudios familiares y genéticos
- Tiempos de vida: edades, duraciones, etc.
- Tiempos de espera: autobús, servicios de atención, etc.
- Tiempos de estancia: turismo, paro, operaciones, etc.

Otros ejemplos

- Estudios familiares y genéticos
- Tiempos de vida: edades, duraciones, etc.
- Tiempos de espera: autobús, servicios de atención, etc.
- Tiempos de estancia: turismo, paro, operaciones, etc.
- Economía: detección de rentas, tiempos de venta de acciones, etc.

Otros ejemplos

- Estudios familiares y genéticos
- Tiempos de vida: edades, duraciones, etc.
- Tiempos de espera: autobús, servicios de atención, etc.
- Tiempos de estancia: turismo, paro, operaciones, etc.
- Economía: detección de rentas, tiempos de venta de acciones, etc.
- Industriales: longitud de fibras textiles, ópticas, etc.

Otros ejemplos

- Estudios familiares y genéticos
- Tiempos de vida: edades, duraciones, etc.
- Tiempos de espera: autobús, servicios de atención, etc.
- Tiempos de estancia: turismo, paro, operaciones, etc.
- Economía: detección de rentas, tiempos de venta de acciones, etc.
- Industriales: longitud de fibras textiles, ópticas, etc.
- Ecología: muestreos mediante trayectos, especies en peligro de estinción, pesca, etc.

Otros ejemplos

- Estudios familiares y genéticos
- Tiempos de vida: edades, duraciones, etc.
- Tiempos de espera: autobús, servicios de atención, etc.
- Tiempos de estancia: turismo, paro, operaciones, etc.
- Economía: detección de rentas, tiempos de venta de acciones, etc.
- Industriales: longitud de fibras textiles, ópticas, etc.
- Ecología: muestreos mediante trayectos, especies en peligro de estinción, pesca, etc.
- Medicina: detección de tumores, tiempos de curación o supervivencia, etc.

Otros ejemplos

- Estudios familiares y genéticos
- Tiempos de vida: edades, duraciones, etc.
- Tiempos de espera: autobús, servicios de atención, etc.
- Tiempos de estancia: turismo, paro, operaciones, etc.
- Economía: detección de rentas, tiempos de venta de acciones, etc.
- Industriales: longitud de fibras textiles, ópticas, etc.
- Ecología: muestreos mediante trayectos, especies en peligro de estinción, pesca, etc.
- Medicina: detección de tumores, tiempos de curación o supervivencia, etc.
- Física/Química: distribución de gases, estudio de radiaciones espaciales, etc.

Otros ejemplos

- Estudios familiares y genéticos
- Tiempos de vida: edades, duraciones, etc.
- Tiempos de espera: autobús, servicios de atención, etc.
- Tiempos de estancia: turismo, paro, operaciones, etc.
- Economía: detección de rentas, tiempos de venta de acciones, etc.
- Industriales: longitud de fibras textiles, ópticas, etc.
- Ecología: muestreos mediante trayectos, especies en peligro de estinción, pesca, etc.
- Medicina: detección de tumores, tiempos de curación o supervivencia, etc.
- Física/Química: distribución de gases, estudio de radiaciones espaciales, etc.
- Arqueología: Longitudes de huesos, etc.

¿Por qué tarda tanto el autobús?

- R.C. Gupta 1979. Waiting time paradox and size biased sampling. *Communications in Statistics, Theory and Methods* **A8** (6), 601-607.

¿Por qué tarda tanto el autobús?

- R.C. Gupta 1979. Waiting time paradox and size biased sampling. *Communications in Statistics, Theory and Methods* **A8** (6), 601-607.
- Supongamos que los autobuses pasan exactamente cada 20 min. y que no conocemos su horario:

|− − − −|− − − −|− − − −|− − − −|− − − −|− − − −|

¿Por qué tarda tanto el autobús?

- R.C. Gupta 1979. Waiting time paradox and size biased sampling. *Communications in Statistics, Theory and Methods* **A8** (6), 601-607.
- Supongamos que los autobuses pasan exactamente cada 20 min. y que no conocemos su horario:

$$|----|----|----|----|----|----|$$

- Con ésto el tiempo de espera T sería Uniforme $(0, 20)$

¿Por qué tarda tanto el autobús?

- R.C. Gupta 1979. Waiting time paradox and size biased sampling. *Communications in Statistics, Theory and Methods* **A8** (6), 601-607.
- Supongamos que los autobuses pasan exactamente cada 20 min. y que no conocemos su horario:

$$|----|----|----|----|----|----|$$

- Con ésto el tiempo de espera T sería Uniforme $(0, 20)$
- El tiempo de espera medio será $E(T) = 20/2 = 10\,\text{min}$.

¿Por qué tarda tanto el autobús?

- R.C. Gupta 1979. Waiting time paradox and size biased sampling. *Communications in Statistics, Theory and Methods* **A8** (6), 601-607.
- Supongamos que los autobuses pasan exactamente cada 20 min. y que no conocemos su horario:

$$|----|----|----|----|----|----|$$

- Con ésto el tiempo de espera T sería Uniforme $(0, 20)$
- El tiempo de espera medio será $E(T) = 20/2 = 10\,\mathrm{min}$.
- ¡Todos sabemos que ésto no es lo que realmente ocurre!

¿Por qué tarda tanto el autobús?

- Lo que realmente ocurre es:

$| - - - | - - - - - | - - - - - | - - | - - - - - - | - - - |$

¿Por qué tarda tanto el autobús?

- Lo que realmente ocurre es:

  |— — —|— — — — —|— — — — —|— —|— — — — — —|— — —|

- El tiempo entre autobuses es una v.a. X

¿Por qué tarda tanto el autobús?

- Lo que realmente ocurre es:

$$|---|------|------|--|------|---|$$

- El tiempo entre autobuses es una v.a. X
- Supongamos que sí es cierto que $\mu = E(X) = 20\,\text{min}$.

¿Por qué tarda tanto el autobús?

- Lo que realmente ocurre es:

$$|---|------|------|--|-------|---|$$

- El tiempo entre autobuses es una v.a. X
- Supongamos que sí es cierto que $\mu = E(X) = 20\,\text{min}$.
- El tiempo de espera T sería Uniforme $(0, X)$

¿Por qué tarda tanto el autobús?

- Lo que realmente ocurre es:

$$|---|------|------|--|-------|---|$$

- El tiempo entre autobuses es una v.a. X
- Supongamos que sí es cierto que $\mu = E(X) = 20\,\text{min}$.
- El tiempo de espera T sería Uniforme $(0, X)$
- El tiempo de espera $T = UX$, donde U sería Uniforme $(0, 1)$

¿Por qué tarda tanto el autobús?

- Lo que realmente ocurre es:

$$| - - - | - - - - - - | - - - - - - | - - | - - - - - - - | - - - |$$

- El tiempo entre autobuses es una v.a. X
- Supongamos que sí es cierto que $\mu = E(X) = 20\,\text{min}$.
- El tiempo de espera T sería Uniforme $(0, X)$
- El tiempo de espera $T = UX$, donde U sería Uniforme $(0, 1)$
- El tiempo de espera medio será
  $E(T) = E(UX) = E(X)/2 = 10\,\text{min}$

¿Por qué tarda tanto el autobús?

- Lo que realmente ocurre es:

$$|---|------|------|--|-------|---|$$

- El tiempo entre autobuses es una v.a. X
- Supongamos que sí es cierto que $\mu = E(X) = 20\,\text{min}$.
- El tiempo de espera T sería Uniforme $(0, X)$
- El tiempo de espera $T = UX$, donde U sería Uniforme $(0, 1)$
- El tiempo de espera medio será
  $E(T) = E(UX) = E(X)/2 = 10\,\text{min}$
- ¡Todos sabemos que ésto no es lo que realmente ocurre!

¿Por qué tarda tanto el autobús?

- En realidad lo que ocurre es:

$$|---|-----|-----|--|------|---|$$

¿Por qué tarda tanto el autobús?

- En realidad lo que ocurre es:

$$|---|-----|-----|--|------|---|$$

- Si llamamos $X_1, X_2, ..., X_n$ a los tiempos entre autobuses, la probabilidad de que nos toque un determinado tiempo X_i es proporcional a X_i

¿Por qué tarda tanto el autobús?

- En realidad lo que ocurre es:

$$|---|------|------|--|-------|---|$$

- Si llamamos $X_1, X_2, ..., X_n$ a los tiempos entre autobuses, la probabilidad de que nos toque un determinado tiempo X_i es proporcional a X_i
- Luego en realidad T es uniforme $(0, X^*)$ donde X^* es la sesgada en longitud

$$E(X^*) = \int_0^\infty x f^*(x) dx = \int_0^\infty x \frac{xf(x)}{\mu} dx = \frac{E(X^2)}{E(X)} = \mu + \frac{\sigma^2}{\mu}$$

¿Por qué tarda tanto el autobús?

- En realidad lo que ocurre es:

$$|---|------|------|--|-------|---|$$

- Si llamamos $X_1, X_2, ..., X_n$ a los tiempos entre autobuses, la probabilidad de que nos toque un determinado tiempo X_i es proporcional a X_i
- Luego en realidad T es uniforme $(0, X^*)$ donde X^* es la sesgada en longitud

$$E(X^*) = \int_0^\infty x f^*(x) dx = \int_0^\infty x \frac{xf(x)}{\mu} dx = \frac{E(X^2)}{E(X)} = \mu + \frac{\sigma^2}{\mu}$$

- Es decir $E(T) = E(X^*/2) = 10 + \sigma^2/(20) > 10$

¿Por qué tarda tanto el autobús?

- En realidad lo que ocurre es:

$$|---|------|------|--|-------|---|$$

- Si llamamos $X_1, X_2, ..., X_n$ a los tiempos entre autobuses, la probabilidad de que nos toque un determinado tiempo X_i es proporcional a X_i
- Luego en realidad T es uniforme $(0, X^*)$ donde X^* es la sesgada en longitud

$$E(X^*) = \int_0^\infty x f^*(x) dx = \int_0^\infty x \frac{x f(x)}{\mu} dx = \frac{E(X^2)}{E(X)} = \mu + \frac{\sigma^2}{\mu}$$

- Es decir $E(T) = E(X^*/2) = 10 + \sigma^2/(20) > 10$
- ¡En el primer caso suponíamos $\sigma^2 = 0$!

¿Por qué tarda tanto el autobús?

- En realidad lo que ocurre es:

$$|---|------|------|--|-------|---|$$

- Si llamamos $X_1, X_2, ..., X_n$ a los tiempos entre autobuses, la probabilidad de que nos toque un determinado tiempo X_i es proporcional a X_i
- Luego en realidad T es uniforme $(0, X^*)$ donde X^* es la sesgada en longitud

$$E(X^*) = \int_0^\infty x f^*(x) dx = \int_0^\infty x \frac{x f(x)}{\mu} dx = \frac{E(X^2)}{E(X)} = \mu + \frac{\sigma^2}{\mu}$$

- Es decir $E(T) = E(X^*/2) = 10 + \sigma^2/(20) > 10$
- ¡En el primer caso suponíamos $\sigma^2 = 0$!
- Consecuencia: ¡Es muy importante la puntualidad!

¿Por qué tarda tanto el autobús?

- En particular, si $X \equiv Exp(\mu = 20\,\mathrm{min})$

¿Por qué tarda tanto el autobús?

- En particular, si $X \equiv Exp(\mu = 20\,\text{min})$
- $E(X^*) = \mu + \sigma^2/\mu = \mu + \mu^2/\mu = 2\mu$

¿Por qué tarda tanto el autobús?

- En particular, si $X \equiv Exp(\mu = 20\,\text{min})$
- $E(X^*) = \mu + \sigma^2/\mu = \mu + \mu^2/\mu = 2\mu$
- $E(T) = E(X^*/2) = E(X)$!!

¿Por qué tarda tanto el autobús?

- En particular, si $X \equiv Exp(\mu = 20\,\text{min})$
- $E(X^*) = \mu + \sigma^2/\mu = \mu + \mu^2/\mu = 2\mu$
- $E(T) = E(X^*/2) = E(X)$!!
- Paradoja: ¡Aunque los autobuses pasen cada 20 min, nosotros debemos esperar 20 min!

¿Por qué tarda tanto el autobús?

- En particular, si $X \equiv Exp(\mu = 20\,\text{min})$
- $E(X^*) = \mu + \sigma^2/\mu = \mu + \mu^2/\mu = 2\mu$
- $E(T) = E(X^*/2) = E(X)$!!
- Paradoja: ¡Aunque los autobuses pasen cada 20 min, nosotros debemos esperar 20 min!
- Esquemas similares:

¿Por qué tarda tanto el autobús?

- En particular, si $X \equiv Exp(\mu = 20\,\text{min})$
- $E(X^*) = \mu + \sigma^2/\mu = \mu + \mu^2/\mu = 2\mu$
- $E(T) = E(X^*/2) = E(X)$!!
- Paradoja: ¡Aunque los autobuses pasen cada 20 min, nosotros debemos esperar 20 min!
- Esquemas similares:
- Duración de componentes industriales en inspecciones aleatorias en procesos de renovación.

¿Por qué tarda tanto el autobús?

- En particular, si $X \equiv Exp(\mu = 20\,\text{min})$
- $E(X^*) = \mu + \sigma^2/\mu = \mu + \mu^2/\mu = 2\mu$
- $E(T) = E(X^*/2) = E(X)$!!
- Paradoja: ¡Aunque los autobuses pasen cada 20 min, nosotros debemos esperar 20 min!
- Esquemas similares:
- Duración de componentes industriales en inspecciones aleatorias en procesos de renovación.
- Tiempos de estancia o de espera, etc...

Solución general: distribución de equilibrio

- Una pieza es sustituida al fallar por otra de características similares

$$| - - - | - - - - - - | - - - - - - | - - | - - - - - - - | - - - |$$

Solución general: distribución de equilibrio

- Una pieza es sustituida al fallar por otra de características similares

$$|---|------|------|--|-------|---|$$

- El tiempo que dura cada pieza X_i es una v.a. con la misma distribución que X

Solución general: distribución de equilibrio

- Una pieza es sustituida al fallar por otra de características similares

$$|---|------|------|--|------|---|$$

- El tiempo que dura cada pieza X_i es una v.a. con la misma distribución que X
- Hacemos inspecciones aleatorias

Solución general: distribución de equilibrio

- Una pieza es sustituida al fallar por otra de características similares

$$|---|------|------|--|-------|---|$$

- El tiempo que dura cada pieza X_i es una v.a. con la misma distribución que X
- Hacemos inspecciones aleatorias
- El tiempo restante de funcionamiento es $T = UX$, donde U sería Uniforme $(0, 1)$

Solución general: distribución de equilibrio

- Una pieza es sustituida al fallar por otra de características similares

$$|---|------|------|--|-------|---|$$

- El tiempo que dura cada pieza X_i es una v.a. con la misma distribución que X
- Hacemos inspecciones aleatorias
- El tiempo restante de funcionamiento es $T = UX$, donde U sería Uniforme $(0,1)$
- El tiempo de funcionamiento hasta ese momento es $T = UX$, donde U sería Uniforme $(0,1)$

Solución general: distribución de equilibrio

- En realidad es $T = UX^*$, con lo que:

$$f(x,u) \quad = \quad f^*(x) = \frac{xf(x)}{\mu};\ 0 < u < 1, x > 0$$

Solución general: distribución de equilibrio

- En realidad es $T = UX^*$, con lo que:

$$f(x,u) \quad = \quad f^*(x) = \frac{xf(x)}{\mu};\ 0 < u < 1, x > 0$$

- Si $\overline{F}_T(t) = \Pr(T > t) = \Pr(UX^* > t)$, $t > 0$,

$$\overline{F}_T(t) = \int_t^\infty \int_{t/x}^1 \frac{xf(x)}{\mu} dudx = \int_t^\infty \frac{(x-t)f(x)}{\mu} dx = \int_t^\infty \frac{\overline{F}(x)}{\mu} dx$$

Solución general: distribución de equilibrio

- En realidad es $T = UX^*$, con lo que:

$$f(x,u) = f^*(x) = \frac{xf(x)}{\mu};\ 0 < u < 1, x > 0$$

- Si $\overline{F}_T(t) = \Pr(T > t) = \Pr(UX^* > t)$, $t > 0$,

$$\overline{F}_T(t) = \int_t^\infty \int_{t/x}^1 \frac{xf(x)}{\mu} du dx = \int_t^\infty \frac{(x-t)f(x)}{\mu} dx = \int_t^\infty \frac{\overline{F}(x)}{\mu} dx$$

- Es decir,

$$f_T(t) = \overline{F}'_T(t) = \frac{\overline{F}(t)}{\mu} = \frac{1-F(t)}{f(t)} \frac{f(t)}{\mu} = w(t)\frac{f(t)}{\mu};\ t > 0$$

$$w(t) = \frac{1-F(t)}{f(t)} = \frac{1}{h(t)};\ \text{con } h(t) = \frac{f(t)}{1-F(t)} \text{ razón de fallo}$$

Solución paradoja autobuses

- Si X es Exponencial, $T = UX^*$, $t > 0$,

$$\overline{F}_T(t) = \int_t^\infty \frac{\overline{F}(x)}{\mu} dx = \int_t^\infty \frac{\exp(-x/\mu)}{\mu} dx = \exp(-t/\mu)$$

Solución paradoja autobuses

- Si X es Exponencial, $T = UX^*$, $t > 0$,

$$\overline{F}_T(t) = \int_t^\infty \frac{\overline{F}(x)}{\mu} dx = \int_t^\infty \frac{\exp(-x/\mu)}{\mu} dx = \exp(-t/\mu)$$

- Es decir, T es Exponencial con la misma media: $X =_d UX^*$.

Solución paradoja autobuses

- Si X es Exponencial, $T = UX^*$, $t > 0$,

$$\overline{F}_T(t) = \int_t^\infty \frac{\overline{F}(x)}{\mu} dx = \int_t^\infty \frac{\exp(-x/\mu)}{\mu} dx = \exp(-t/\mu)$$

- Es decir, T es Exponencial con la misma media: $X =_d UX^*$.
- En realidad el modelo exponencial es el único modelo que verifica $X =_d UX^*$ (ó $X =_d T$).

Solución paradoja autobuses

- Si X es Exponencial, $T = UX^*$, $t > 0$,

$$\overline{F}_T(t) = \int_t^\infty \frac{\overline{F}(x)}{\mu} dx = \int_t^\infty \frac{\exp(-x/\mu)}{\mu} dx = \exp(-t/\mu)$$

- Es decir, T es Exponencial con la misma media: $X =_d UX^*$.
- En realidad el modelo exponencial es el único modelo que verifica $X =_d UX^*$ (ó $X =_d T$).
- La distribución de T (distribución de equilibrio) tiene propiedades muy interesantes.

Solución paradoja autobuses

- Si X es Exponencial, $T = UX^*$, $t > 0$,

$$\overline{F}_T(t) = \int_t^\infty \frac{\overline{F}(x)}{\mu} dx = \int_t^\infty \frac{\exp(-x/\mu)}{\mu} dx = \exp(-t/\mu)$$

- Es decir, T es Exponencial con la misma media: $X =_d UX^*$.
- En realidad el modelo exponencial es el único modelo que verifica $X =_d UX^*$ (ó $X =_d T$).
- La distribución de T (distribución de equilibrio) tiene propiedades muy interesantes.
- Por ejemplo,

$$h_T(t) = \frac{f_T(t)}{\overline{F}_T(t)} = \frac{\overline{F}_T(t)}{\int_t^\infty \overline{F}_T(x)dx} = \frac{1}{m(t)}$$

con $m(t) = E(X - t|X > t)$ (función vida media residual).

¿Cómo se detectan las muestras sesgadas?

- En el ejemplo de Fisher (1934) porque no coincidían con los resultados esperados

¿Cómo se detectan las muestras sesgadas?

- En el ejemplo de Fisher (1934) porque no coincidían con los resultados esperados
- En el ejemplo de los autobuses porque los resultados eran 'paradójicos'

¿Cómo se detectan las muestras sesgadas?

- En el ejemplo de Fisher (1934) porque no coincidían con los resultados esperados
- En el ejemplo de los autobuses porque los resultados eran 'paradójicos'
- Estudio de la duración de las estancias de turistas en Marruecos (INSEA, 1966)
- G.P. Patil 1981. Proceedings of the Indian Statistical Institute Jubilee International Conference on Statistics: Applications and New Directions, 478-503)

¿Cómo se detectan las muestras sesgadas?

- En el ejemplo de Fisher (1934) porque no coincidían con los resultados esperados
- En el ejemplo de los autobuses porque los resultados eran ‘paradójicos’
- Estudio de la duración de las estancias de turistas en Marruecos (INSEA, 1966)
- G.P. Patil 1981. Proceedings of the Indian Statistical Institute Jubilee International Conference on Statistics: Applications and New Directions, 478-503)
- Muestra en la frontera: n=3000, media=9.0 días

¿Cómo se detectan las muestras sesgadas?

- En el ejemplo de Fisher (1934) porque no coincidían con los resultados esperados
- En el ejemplo de los autobuses porque los resultados eran 'paradójicos'
- Estudio de la duración de las estancias de turistas en Marruecos (INSEA, 1966)
- G.P. Patil 1981. Proceedings of the Indian Statistical Institute Jubilee International Conference on Statistics: Applications and New Directions, 478-503)
- Muestra en la frontera: n=3000, media=9.0 días
- Muestra en los hoteles: n=12321, media=17.8 días

¿Cómo se detectan las muestras sesgadas?

- En el ejemplo de Fisher (1934) porque no coincidían con los resultados esperados
- En el ejemplo de los autobuses porque los resultados eran 'paradójicos'
- Estudio de la duración de las estancias de turistas en Marruecos (INSEA, 1966)
- G.P. Patil 1981. Proceedings of the Indian Statistical Institute Jubilee International Conference on Statistics: Applications and New Directions, 478-503)
- Muestra en la frontera: n=3000, media=9.0 días
- Muestra en los hoteles: n=12321, media=17.8 días
- ¡Los datos no coinciden!

¿Cómo se detectan las muestras sesgadas?

- En el ejemplo de Fisher (1934) porque no coincidían con los resultados esperados
- En el ejemplo de los autobuses porque los resultados eran 'paradójicos'
- Estudio de la duración de las estancias de turistas en Marruecos (INSEA, 1966)
- G.P. Patil 1981. Proceedings of the Indian Statistical Institute Jubilee International Conference on Statistics: Applications and New Directions, 478-503)
- Muestra en la frontera: n=3000, media=9.0 días
- Muestra en los hoteles: n=12321, media=17.8 días
- ¡Los datos no coinciden!
- La estimación segunda fue descartada

Solución Patil

- La muestra en los hoteles es sesgada en longitud (X^*)

Solución Patil

- La muestra en los hoteles es sesgada en longitud (X^*)
- Un turista que está 6 días tiene doble probabilidad de aparecer en la muestra que uno que está 3 días

Solución Patil

- La muestra en los hoteles es sesgada en longitud (X^*)
- Un turista que está 6 días tiene doble probabilidad de aparecer en la muestra que uno que está 3 días
- Por eso, $E(X) \simeq 9.0 > E(X^*) \simeq 17.8$

Solución Patil

- La muestra en los hoteles es sesgada en longitud (X^*)
- Un turista que está 6 días tiene doble probabilidad de aparecer en la muestra que uno que está 3 días
- Por eso, $E(X) \simeq 9.0 > E(X^*) \simeq 17.8$
- El hecho de que $E(X^*) \simeq 2E(X)$ puede indicar que X es Exponencial

Solución Patil

- La muestra en los hoteles es sesgada en longitud (X^*)
- Un turista que está 6 días tiene doble probabilidad de aparecer en la muestra que uno que está 3 días
- Por eso, $E(X) \simeq 9.0 > E(X^*) \simeq 17.8$
- El hecho de que $E(X^*) \simeq 2E(X)$ puede indicar que X es Exponencial
- Es decir $E(X) \simeq 17.8/2 = 8.9$

Solución Patil

- La muestra en los hoteles es sesgada en longitud (X^*)
- Un turista que está 6 días tiene doble probabilidad de aparecer en la muestra que uno que está 3 días
- Por eso, $E(X) \simeq 9.0 > E(X^*) \simeq 17.8$
- El hecho de que $E(X^*) \simeq 2E(X)$ puede indicar que X es Exponencial
- Es decir $E(X) \simeq 17.8/2 = 8.9$
- Ejemplos similares en ecología, en estudios de Fiabilidad de componentes, etc.

¿Qué hacer?

- Solución óptima (usar toda la información disponible)

¿Qué hacer?

- Solución óptima (usar toda la información disponible)
- $X_1, \ldots., X_n$ es m.a.s. de $X \equiv Exp(\mu)$

¿Qué hacer?

- Solución óptima (usar toda la información disponible)
- $X_1, ..., X_n$ es m.a.s. de $X \equiv Exp(\mu)$
- $Y_1, ..., Y_m$ es m.a.s. de $X^* \equiv Exp^*(\mu)$

¿Qué hacer?

- Solución óptima (usar toda la información disponible)
- $X_1, ..., X_n$ es m.a.s. de $X \equiv Exp(\mu)$
- $Y_1, ..., Y_m$ es m.a.s. de $X^* \equiv Exp^*(\mu)$
- El estimador máximo verosímil es:

$$\widehat{\mu} = \frac{1}{n+2m}\left(\sum_{i=1}^{n} X_i + \sum_{j=1}^{m} Y_j\right)$$

¿Qué hacer?

- Solución óptima (usar toda la información disponible)
- $X_1, ..., X_n$ es m.a.s. de $X \equiv Exp(\mu)$
- $Y_1, ..., Y_m$ es m.a.s. de $X^* \equiv Exp^*(\mu)$
- El estimador máximo verosímil es:

$$\widehat{\mu} = \frac{1}{n+2m}\left(\sum_{i=1}^{n} X_i + \sum_{j=1}^{m} Y_j\right)$$

- Es insesgado ya que $E(X_i) = \mu$ y $E(Y_j) = 2\mu$

¿Qué hacer?

- Solución óptima (usar toda la información disponible)
- $X_1, ..., X_n$ es m.a.s. de $X \equiv Exp(\mu)$
- $Y_1, ..., Y_m$ es m.a.s. de $X^* \equiv Exp^*(\mu)$
- El estimador máximo verosímil es:

$$\widehat{\mu} = \frac{1}{n+2m}\left(\sum_{i=1}^{n} X_i + \sum_{j=1}^{m} Y_j\right)$$

- Es insesgado ya que $E(X_i) = \mu$ y $E(Y_j) = 2\mu$
- Su varianza es

$$Var(\widehat{\mu}) = \frac{\mu^2}{n+2m}$$

¿Qué hacer?

- Solución óptima (usar toda la información disponible)
- $X_1, ..., X_n$ es m.a.s. de $X \equiv Exp(\mu)$
- $Y_1, ..., Y_m$ es m.a.s. de $X^* \equiv Exp^*(\mu)$
- El estimador máximo verosímil es:

$$\widehat{\mu} = \frac{1}{n + 2m} \left(\sum_{i=1}^{n} X_i + \sum_{j=1}^{m} Y_j \right)$$

- Es insesgado ya que $E(X_i) = \mu$ y $E(Y_j) = 2\mu$
- Su varianza es

$$Var(\widehat{\mu}) = \frac{\mu^2}{n + 2m}$$

- Es un EIMV

Solución ejemplo Patil

- En nuestro ejemplo, la mejor estimación posible es

$$\widehat{\mu} = \frac{1}{3000 + 2 \cdot 12321}(3000 \cdot 9 + 12321 \cdot 17.8) = 8.91$$

Solución ejemplo Patil

- En nuestro ejemplo, la mejor estimación posible es

$$\widehat{\mu} = \frac{1}{3000 + 2 \cdot 12321}(3000 \cdot 9 + 12321 \cdot 17.8) = 8.91$$

- Con varianza verificando $2\sigma(\widehat{\mu}) \simeq 0.11$

Solución ejemplo Patil

- En nuestro ejemplo, la mejor estimación posible es

$$\widehat{\mu} = \frac{1}{3000 + 2 \cdot 12321}(3000 \cdot 9 + 12321 \cdot 17.8) = 8.91$$

- Con varianza verificando $2\sigma(\widehat{\mu}) \simeq 0.11$
- Sólo frontera $\widehat{\mu} = 9.0$

Solución ejemplo Patil

- En nuestro ejemplo, la mejor estimación posible es

$$\widehat{\mu} = \frac{1}{3000 + 2 \cdot 12321}(3000 \cdot 9 + 12321 \cdot 17.8) = 8.91$$

- Con varianza verificando $2\sigma(\widehat{\mu}) \simeq 0.11$
- Sólo frontera $\widehat{\mu} = 9.0$
- Sólo hoteles $\widehat{\mu} = 8.9$

Solución general: exponencial

- ¿Qué hacer la próxima vez?

Solución general: exponencial

- ¿Qué hacer la próxima vez?
- ¿Cuál de los dos métodos de muestreo es mejor?

Solución general: exponencial

- ¿Qué hacer la próxima vez?
- ¿Cuál de los dos métodos de muestreo es mejor?
- Nótese que si $n = m$, entonces el estimador basado en la muestra sesgada ($n = 0$) es mejor (tiene menos varianza) que el estimador basado en la muestra insesgada

Solución general: exponencial

- ¿Qué hacer la próxima vez?
- ¿Cuál de los dos métodos de muestreo es mejor?
- Nótese que si $n = m$, entonces el estimador basado en la muestra sesgada ($n = 0$) es mejor (tiene menos varianza) que el estimador basado en la muestra insesgada
- Las informaciones de Fisher para $n = m = 1$ son

$$\begin{aligned} I^*(\mu) &= 2/\mu^2 \\ I(\mu) &= 1/\mu^2 \end{aligned}$$

Solución general: exponencial

- ¿Qué hacer la próxima vez?
- ¿Cuál de los dos métodos de muestreo es mejor?
- Nótese que si $n = m$, entonces el estimador basado en la muestra sesgada ($n = 0$) es mejor (tiene menos varianza) que el estimador basado en la muestra insesgada
- Las informaciones de Fisher para $n = m = 1$ son

$$\begin{aligned} I^*(\mu) &= 2/\mu^2 \\ I(\mu) &= 1/\mu^2 \end{aligned}$$

- ¡Cada dato de la muestra sesgada tiene doble información!

Solución general: exponencial

- ¿Qué hacer la próxima vez?
- ¿Cuál de los dos métodos de muestreo es mejor?
- Nótese que si $n = m$, entonces el estimador basado en la muestra sesgada ($n = 0$) es mejor (tiene menos varianza) que el estimador basado en la muestra insesgada
- Las informaciones de Fisher para $n = m = 1$ son

$$\begin{aligned} I^*(\mu) &= 2/\mu^2 \\ I(\mu) &= 1/\mu^2 \end{aligned}$$

- ¡Cada dato de la muestra sesgada tiene doble información!
- En este caso, si el sesgo es $w(x) = x^k$, se tiene más información conforme aumenta k

Ejemplo de DeGroot

- G.P. Patil (1991). Encountered data, Statistical Ecology Environmental Statistics and weighted distribution methods. Environmetrics 2(4), 377-423.

Ejemplo de DeGroot

- G.P. Patil (1991). Encountered data, Statistical Ecology Environmental Statistics and weighted distribution methods. Environmetrics 2(4), 377-423.
- Discusión de Morris H. DeGroot (fue profesor de Carnegie Mellon University, Pittsburgh).

Ejemplo de DeGroot

- G.P. Patil (1991). Encountered data, Statistical Ecology Environmental Statistics and weighted distribution methods. Environmetrics 2(4), 377-423.
- Discusión de Morris H. DeGroot (fue profesor de Carnegie Mellon University, Pittsburgh).
- ¿Cómo hacer predicciones seguras en la bolsa?

Ejemplo de DeGroot

- G.P. Patil (1991). Encountered data, Statistical Ecology Environmental Statistics and weighted distribution methods. Environmetrics 2(4), 377-423.
- Discusión de Morris H. DeGroot (fue profesor de Carnegie Mellon University, Pittsburgh).
- ¿Cómo hacer predicciones seguras en la bolsa?
- Enviaremos unas 128 cartas, la mitad de ellas dirán:

Ejemplo de DeGroot

- G.P. Patil (1991). Encountered data, Statistical Ecology Environmental Statistics and weighted distribution methods. Environmetrics 2(4), 377-423.
- Discusión de Morris H. DeGroot (fue profesor de Carnegie Mellon University, Pittsburgh).
- ¿Cómo hacer predicciones seguras en la bolsa?
- Enviaremos unas 128 cartas, la mitad de ellas dirán:
- "Usted no me conoce pero soy un experto analista de bolsa y tengo un sistema para precedir cómo se comportará la bolsa y, como prueba de amistad le envío hoy el siguiente consejo gratis: Las acciones de TIM van a SUBIR esta semana".

Ejemplo de DeGroot

- G.P. Patil (1991). Encountered data, Statistical Ecology Environmental Statistics and weighted distribution methods. Environmetrics 2(4), 377-423.
- Discusión de Morris H. DeGroot (fue profesor de Carnegie Mellon University, Pittsburgh).
- ¿Cómo hacer predicciones seguras en la bolsa?
- Enviaremos unas 128 cartas, la mitad de ellas dirán:
- "Usted no me conoce pero soy un experto analista de bolsa y tengo un sistema para precedir cómo se comportará la bolsa y, como prueba de amistad le envío hoy el siguiente consejo gratis: Las acciones de TIM van a SUBIR esta semana".
- La otra mitad de ellas dirán: "... BAJAR esta semana".

Ejemplo de DeGroot

- La semana siguiente escribiremos la siguiente carta sólo a los que enviamos una predicción correcta (64):

Ejemplo de DeGroot

- La semana siguiente escribiremos la siguiente carta sólo a los que enviamos una predicción correcta (64):
- "Recuerda usted que la semana pasada le hice una predicción que, como usted sabe ha sido correcta, y como prueba de amistad le envío de nuevo el siguiente consejo gratis: Las acciones de TIM van a SUBIR (BAJAR) esta semana".

Ejemplo de DeGroot

- La semana siguiente escribiremos la siguiente carta sólo a los que enviamos una predicción correcta (64):
- "Recuerda usted que la semana pasada le hice una predicción que, como usted sabe ha sido correcta, y como prueba de amistad le envío de nuevo el siguiente consejo gratis: Las acciones de TIM van a SUBIR (BAJAR) esta semana".
- Repetiremos este proceso durante 7 semanas.

Ejemplo de DeGroot

- La semana siguiente escribiremos la siguiente carta sólo a los que enviamos una predicción correcta (64):
- "Recuerda usted que la semana pasada le hice una predicción que, como usted sabe ha sido correcta, y como prueba de amistad le envío de nuevo el siguiente consejo gratis: Las acciones de TIM van a SUBIR (BAJAR) esta semana".
- Repetiremos este proceso durante 7 semanas.
- Finalmente enviaremos esta carta a la persona a la que le hemos enviado 7 predicciones correctas:

Ejemplo de DeGroot

- La semana siguiente escribiremos la siguiente carta sólo a los que enviamos una predicción correcta (64):
- "Recuerda usted que la semana pasada le hice una predicción que, como usted sabe ha sido correcta, y como prueba de amistad le envío de nuevo el siguiente consejo gratis: Las acciones de TIM van a SUBIR (BAJAR) esta semana".
- Repetiremos este proceso durante 7 semanas.
- Finalmente enviaremos esta carta a la persona a la que le hemos enviado 7 predicciones correctas:
- "Bueno, creo que ya le he probado la eficiencia de mi método de predicción bursatil, por lo que, si usted quiere recibir mi predicción para esta semana, deberá pagarme 10.000$"

Solución Ejemplo de DeGroot

- Desde el punto de vista del inversor bursatil, se tiene una m.a.s. $X_1, ..., X_7$ de una *Bernoulli* $B(p)$ con p probabilidad de acertar en las predicciones con $X_i = 1$ ($\widehat{p} = 7/7 = 1$).

Solución Ejemplo de DeGroot

- Desde el punto de vista del inversor bursatil, se tiene una m.a.s. $X_1, \dots, X_7$ de una *Bernoulli* $B(p)$ con p probabilidad de acertar en las predicciones con $X_i = 1$ ($\widehat{p} = 7/7 = 1$).
- Pero debe preguntarse, ¿cuál es la probabilidad de que una observación aparezca en la muestra?

Solución Ejemplo de DeGroot

- Desde el punto de vista del inversor bursatil, se tiene una m.a.s. $X_1, \ldots, X_7$ de una *Bernoulli* $B(p)$ con p probabilidad de acertar en las predicciones con $X_i = 1$ ($\widehat{p} = 7/7 = 1$).
- Pero debe preguntarse, ¿cuál es la probabilidad de que una observación aparezca en la muestra?
- ¡Ciertamente la probabilidad de que X_i aparezca en la muestra es proporcional a X_i!

Solución Ejemplo de DeGroot

- Desde el punto de vista del inversor bursatil, se tiene una m.a.s. $X_1, ..., X_7$ de una *Bernoulli* $B(p)$ con p probabilidad de acertar en las predicciones con $X_i = 1$ ($\widehat{p} = 7/7 = 1$).
- Pero debe preguntarse, ¿cuál es la probabilidad de que una observación aparezca en la muestra?
- ¡Ciertamente la probabilidad de que X_i aparezca en la muestra es proporcional a X_i!
- En realidad tenemos un m.a.s. de X^* con $p^*(x) = xp(x)/\mu$, es decir $X^* = 1$.

Conclusiones

- ¡¡Todos hemos sospechado alguna vez sobre los métodos de selección de una muestra!!

Conclusiones

- ¡¡Todos hemos sospechado alguna vez sobre los métodos de selección de una muestra!!
- Importancia de conocer los mecanimos de selección de la muestra

Conclusiones

- ¡¡Todos hemos sospechado alguna vez sobre los métodos de selección de una muestra!!
- Importancia de conocer los mecanimos de selección de la muestra
- Si tenemos dos muestras y conocemos su sesgo, la mejor solución es usar las dos muestras

Conclusiones

- ¡¡Todos hemos sospechado alguna vez sobre los métodos de selección de una muestra!!
- Importancia de conocer los mecanimos de selección de la muestra
- Si tenemos dos muestras y conocemos su sesgo, la mejor solución es usar las dos muestras
- Con sólo una muestra sesgada, podemos obtener resultados tan buenos e incluso mejores a los que se obtienen con muestras insesgadas

Conclusiones

- ¡¡Todos hemos sospechado alguna vez sobre los métodos de selección de una muestra!!
- Importancia de conocer los mecanimos de selección de la muestra
- Si tenemos dos muestras y conocemos su sesgo, la mejor solución es usar las dos muestras
- Con sólo una muestra sesgada, podemos obtener resultados tan buenos e incluso mejores a los que se obtienen con muestras insesgadas
- Si podemos elegir, optaremos por el método de muestreo que proporcione más información sobre el parámetro que se quiera estudiar (sea sesgado o no)

- Rao, C.R. (1994). Estadística y verdad. PPU.

- Rao, C.R. (1994). Estadística y verdad. PPU.
- del Aguila, Y. (2001). Caracterizaciones e inferencia a partir de muestras sesgadas, Tesis Doctoral, Universidad de Almería.

- Guillamon, Navarro and Ruiz (1998). Kernel density estimation using weighted data. *Comm. Statist. Theory and Methods* **27**, 2123-2135.

- Guillamon, Navarro and Ruiz (1998). Kernel density estimation using weighted data. *Comm. Statist. Theory and Methods* **27**, 2123-2135.
- Navarro, Ruiz and del Aguila (2001). Parametric estimation from weighted samples. Biometrical Journal 43, 297-311.

- Guillamon, Navarro and Ruiz (1998). Kernel density estimation using weighted data. *Comm. Statist. Theory and Methods* **27**, 2123-2135.
- Navarro, Ruiz and del Aguila (2001). Parametric estimation from weighted samples. Biometrical Journal 43, 297-311.
- Navarro, Ruiz and del Aguila (2003). How to detect biased samples?. Biometrical Journal 44, 742-763.

- Guillamon, Navarro and Ruiz (1998). Kernel density estimation using weighted data. *Comm. Statist. Theory and Methods* **27**, 2123-2135.
- Navarro, Ruiz and del Aguila (2001). Parametric estimation from weighted samples. Biometrical Journal 43, 297-311.
- Navarro, Ruiz and del Aguila (2003). How to detect biased samples?. Biometrical Journal 44, 742-763.
- Pakes, Navarro, Ruiz and del Aguila (2003). Characterizations using weighted distributions. Journal of Statistical Planning and Inference 116, 389-420.

- Guillamon, Navarro and Ruiz (1998). Kernel density estimation using weighted data. *Comm. Statist. Theory and Methods* **27**, 2123-2135.
- Navarro, Ruiz and del Aguila (2001). Parametric estimation from weighted samples. Biometrical Journal 43, 297-311.
- Navarro, Ruiz and del Aguila (2003). How to detect biased samples?. Biometrical Journal 44, 742-763.
- Pakes, Navarro, Ruiz and del Aguila (2003). Characterizations using weighted distributions. Journal of Statistical Planning and Inference 116, 389-420.
- Navarro, Ruiz and del Aguila (2006). Multivariate weighted distributions. Statistics 40 (1), 51-54.

- Guillamon, Navarro and Ruiz (1998). Kernel density estimation using weighted data. *Comm. Statist. Theory and Methods* **27**, 2123-2135.
- Navarro, Ruiz and del Aguila (2001). Parametric estimation from weighted samples. Biometrical Journal 43, 297-311.
- Navarro, Ruiz and del Aguila (2003). How to detect biased samples?. Biometrical Journal 44, 742-763.
- Pakes, Navarro, Ruiz and del Aguila (2003). Characterizations using weighted distributions. Journal of Statistical Planning and Inference 116, 389-420.
- Navarro, Ruiz and del Aguila (2006). Multivariate weighted distributions. Statistics 40 (1), 51-54.
- Pakes and Navarro (2007). Distributional characterizations through scaling relations. Australian and New Zealand Journal of Statistics 49 (2), 115-135.

- Bayarri, M. J. and DeGroot, M. H. (1986). Information in selection models. Technical report.

- Bayarri, M. J. and DeGroot, M. H. (1986). Information in selection models. Technical report.
- Arnold, B. C.; Castillo, E.; Sarabia, J. M. (2005). Distributions with conditionals in truncated weighted families. Statistics 39, no. 2, 133–147.

- Bayarri, M. J. and DeGroot, M. H. (1986). Information in selection models. Technical report.
- Arnold, B. C.; Castillo, E.; Sarabia, J. M. (2005). Distributions with conditionals in truncated weighted families. Statistics 39, no. 2, 133–147.
- Cristóbal, J. A.; Alcalá, J. T. (2000). Nonparametric regression estimators for length biased data. J. Statist. Plann. Inference 89 (2000), no. 1-2, 145–168.

- Bayarri, M. J. and DeGroot, M. H. (1986). Information in selection models. Technical report.
- Arnold, B. C.; Castillo, E.; Sarabia, J. M. (2005). Distributions with conditionals in truncated weighted families. Statistics 39, no. 2, 133–147.
- Cristóbal, J. A.; Alcalá, J. T. (2000). Nonparametric regression estimators for length biased data. J. Statist. Plann. Inference 89 (2000), no. 1-2, 145–168.
- Cristóbal, José A.; Alcalá, José T. (2001). An overview of nonparametric contributions to the problem of functional estimation from biased data. Test 10, no. 2, 309–332.

- Bayarri, M. J. and DeGroot, M. H. (1986). Information in selection models. Technical report.
- Arnold, B. C.; Castillo, E.; Sarabia, J. M. (2005). Distributions with conditionals in truncated weighted families. Statistics 39, no. 2, 133–147.
- Cristóbal, J. A.; Alcalá, J. T. (2000). Nonparametric regression estimators for length biased data. J. Statist. Plann. Inference 89 (2000), no. 1-2, 145–168.
- Cristóbal, José A.; Alcalá, José T. (2001). An overview of nonparametric contributions to the problem of functional estimation from biased data. Test 10, no. 2, 309–332.
- Cristóbal, J. A.; Ojeda, J. L.; Alcalá, J. T. (2004). Confidence bands in nonparametric regression with length biased data. Ann. Inst. Statist. Math. 56, no. 3, 475–496.

- Marron, J. S.; de Uña-Álvarez, J. (2004). SiZer for length biased, censored density and hazard estimation. J. Statist. Plann. Inference 121, no. 1, 149–161.

- Marron, J. S.; de Uña-Álvarez, J. (2004). SiZer for length biased, censored density and hazard estimation. J. Statist. Plann. Inference 121, no. 1, 149–161.
- de Uña-Álvarez, J.; Rodríguez-Casal, A. (2007). Nonparametric estimation from length-biased data under competing risks. Comput. Statist. Data Anal. 51, no. 5, 2653–2669.

- Marron, J. S.; de Uña-Álvarez, J. (2004). SiZer for length biased, censored density and hazard estimation. J. Statist. Plann. Inference 121, no. 1, 149–161.
- de Uña-Álvarez, J.; Rodríguez-Casal, A. (2007). Nonparametric estimation from length-biased data under competing risks. Comput. Statist. Data Anal. 51, no. 5, 2653–2669.
- de Uña-Álvarez, J.; Saavedra, A. (2004). Bias and variance of the nonparametric MLE under length-biased censored sampling: a simulation study. Comm. Statist. Simulation Comput. 33, no. 2, 397–413

- Marron, J. S.; de Uña-Álvarez, J. (2004). SiZer for length biased, censored density and hazard estimation. J. Statist. Plann. Inference 121, no. 1, 149–161.
- de Uña-Álvarez, J.; Rodríguez-Casal, A. (2007). Nonparametric estimation from length-biased data under competing risks. Comput. Statist. Data Anal. 51, no. 5, 2653–2669.
- de Uña-Álvarez, J.; Saavedra, A. (2004). Bias and variance of the nonparametric MLE under length-biased censored sampling: a simulation study. Comm. Statist. Simulation Comput. 33, no. 2, 397–413
- de Uña-Álvarez, J.; Rodríguez-Casal, A. (2006). Comparing nonparametric estimators for length-biased data. Comm. Statist. Theory Methods 35, no. 4-6, 905–919.