Biased samples (in honor of Prof. C.R. Rao)

Jorge Navarro[1,2]

[1]Universidad de Murcia, Spain. E-mail: jorgenav@um.es
[2]JN is partially supported by Ministerio de Ciencia y Tecnología de España under grant MTM2009-08311 and Fundación Séneca (C.A.R.M.) under grant 08627/PI/08.

Biased and censored samples

- $X_1, \ldots, X_n$ sample from X

Biased and censored samples

- $X_1, ..., X_n$ sample from X
- $X_1, ..., X_n$ i.i.d. $\Pr(X_i \leq x) = \Pr(X \leq x)$

Biased and censored samples

- $X_1, ..., X_n$ sample from X
- $X_1, ..., X_n$ i.i.d. $\Pr(X_i \leq x) = \Pr(X \leq x)$
- Censored sample: Some X_i are unknown.

Biased and censored samples

- $X_1, ..., X_n$ sample from X
- $X_1, ..., X_n$ i.i.d. $\Pr(X_i \leq x) = \Pr(X \leq x)$
- Censored sample: Some X_i are unknown.
- Example: X =lifetime of...

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

Biased and censored samples

- $X_1, ..., X_n$ sample from X
- $X_1, ..., X_n$ i.i.d. $\Pr(X_i \le x) = \Pr(X \le x)$
- Censored sample: Some X_i are unknown.
- Example: X =lifetime of...

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

- 1^+ means $X_i > 1$

Biased and censored samples

- $X_1, ..., X_n$ sample from X
- $X_1, ..., X_n$ i.i.d. $\Pr(X_i \leq x) = \Pr(X \leq x)$
- Censored sample: Some X_i are unknown.
- Example: X =lifetime of...

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

- 1^+ means $X_i > 1$
- Biased sample: the sample probability of X_i depends on X_i.

Biased and censored samples

- $X_1, ..., X_n$ sample from X
- $X_1, ..., X_n$ i.i.d. $\Pr(X_i \leq x) = \Pr(X \leq x)$
- Censored sample: Some X_i are unknown.
- Example: X =lifetime of...

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

- 1^+ means $X_i > 1$
- Biased sample: the sample probability of X_i depends on X_i.
- Example: A sample from families recover from their children.

Biased and censored samples

- $X_1, \ldots, X_n$ sample from X
- $X_1, \ldots, X_n$ i.i.d. $\Pr(X_i \leq x) = \Pr(X \leq x)$
- Censored sample: Some X_i are unknown.
- Example: X =lifetime of...

$$2, 3, 5, 6, 7, \ldots, 1^+, 3^+, 4^+, \ldots$$

- 1^+ means $X_i > 1$
- Biased sample: the sample probability of X_i depends on X_i.
- Example: A sample from families recover from their children.
- Censored samples are a particular case.

A model for biased samples

- First example: Fisher (1934, Ann. Eugenics 6, 13-25).

A model for biased samples

- First example: Fisher (1934, Ann. Eugenics 6, 13-25).
- Model: C.R. Rao (1965, Sankhya Ser. A 27, 311-324).

A model for biased samples

- First example: Fisher (1934, Ann. Eugenics 6, 13-25).
- Model: C.R. Rao (1965, Sankhya Ser. A 27, 311-324).
- Y has the biased (or weighted) distribution associated to X and $w(t) \geq 0$ if

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

A model for biased samples

- First example: Fisher (1934, Ann. Eugenics 6, 13-25).
- Model: C.R. Rao (1965, Sankhya Ser. A 27, 311-324).
- Y has the biased (or weighted) distribution associated to X and $w(t) \geq 0$ if

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

- With this model the probability of observe $X_i = t$ is proportional to $w(t)$.

A model for biased samples

- First example: Fisher (1934, Ann. Eugenics 6, 13-25).
- Model: C.R. Rao (1965, Sankhya Ser. A 27, 311-324).
- Y has the biased (or weighted) distribution associated to X and $w(t) \geq 0$ if

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

- With this model the probability of observe $X_i = t$ is proportional to $w(t)$.
- How to study X from a sample $Y_1, ..., Y_n$ from Y?

A model for biased samples

- First example: Fisher (1934, Ann. Eugenics 6, 13-25).
- Model: C.R. Rao (1965, Sankhya Ser. A 27, 311-324).
- Y has the biased (or weighted) distribution associated to X and $w(t) \geq 0$ if

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

- With this model the probability of observe $X_i = t$ is proportional to $w(t)$.
- How to study X from a sample $Y_1, ..., Y_n$ from Y?
- Censored data in A: $w(t) = 1$ if $t \in A$ (0 elsewhere).

A model for biased samples

- First example: Fisher (1934, Ann. Eugenics 6, 13-25).
- Model: C.R. Rao (1965, Sankhya Ser. A 27, 311-324).
- Y has the biased (or weighted) distribution associated to X and $w(t) \geq 0$ if
$$f_Y(t) = \frac{w(t) f_X(t)}{E(w(X))}$$
- With this model the probability of observe $X_i = t$ is proportional to $w(t)$.
- How to study X from a sample $Y_1, \ldots, Y_n$ from Y?
- Censored data in A: $w(t) = 1$ if $t \in A$ (0 elsewhere).
- Biased data: the probability of observe X_i is proportional to $w(X_i)$.

Rao's example

- Rao (1977, American Statistician 31, 24-26).

Rao's example

- Rao (1977, American Statistician 31, 24-26).
- In a survey we ask for the number of brother and sisters ($^{(*)}$including yourself):

Sex	Brothers*	Sisters*	Total
M or W	Y_i	X_i	$m_i = X_i + Y_i$
-	-	-	-

Rao's example

- Rao (1977, American Statistician 31, 24-26).
- In a survey we ask for the number of brother and sisters ($^{(*)}$including yourself):

Sex	Brothers*	Sisters*	Total
M or W	Y_i	X_i	$m_i = X_i + Y_i$
-	-	-	-

- Predictions (sample from men)

Rao's example

- Rao (1977, American Statistician 31, 24-26).
- In a survey we ask for the number of brother and sisters ($^{(*)}$including yourself):

Sex	Brothers*	Sisters*	Total
M or W	Y_i	X_i	$m_i = X_i + Y_i$
-	-	-	-

- Predictions (sample from men)
  1. $M = \sum Y_i >> W = \sum X_i$

Rao's example

- Rao (1977, American Statistician 31, 24-26).
- In a survey we ask for the number of brother and sisters ($^{(*)}$including yourself):

Sex	Brothers*	Sisters*	Total
M or W	Y_i	X_i	$m_i = X_i + Y_i$
-	-	-	-

- Predictions (sample from men)
  1. $M = \sum Y_i >> W = \sum X_i$
  2. $M - W = \sum Y_i - \sum X_i \simeq k = n$ (sample size)

Rao's example

- Rao (1977, American Statistician 31, 24-26).
- In a survey we ask for the number of brother and sisters ($^{(*)}$including yourself):

Sex	Brothers*	Sisters*	Total
M or W	Y_i	X_i	$m_i = X_i + Y_i$
-	-	-	-

- Predictions (sample from men)
  1. $M = \sum Y_i >> W = \sum X_i$
  2. $M - W = \sum Y_i - \sum X_i \simeq k = n$ (sample size)
  3. $M/N = (\sum Y_i)/(\sum m\ _i) >> 0.5,\ N = \sum m_i = M + W$

Rao's example

- Rao (1977, American Statistician 31, 24-26).
- In a survey we ask for the number of brother and sisters ($^{(*)}$including yourself):

Sex	Brothers*	Sisters*	Total
M or W	Y_i	X_i	$m_i = X_i + Y_i$
-	-	-	-

- Predictions (sample from men)
  1. $M = \sum Y_i >> W = \sum X_i$
  2. $M - W = \sum Y_i - \sum X_i \simeq k = n$ (sample size)
  3. $M/N = (\sum Y_i)/(\sum m\ _i) >> 0.5,\ N = \sum m_i = M + W$
  4. $M/N = (\sum Y_i)/(\sum m\ _i) \simeq 0.5 + \dfrac{k}{2\sum m_i}$

Rao's example

- Rao (1977, American Statistician 31, 24-26).
- In a survey we ask for the number of brother and sisters ($^{(*)}$including yourself):

Sex	Brothers*	Sisters*	Total
M or W	Y_i	X_i	$m_i = X_i + Y_i$
-	-	-	-

- Predictions (sample from men)
  1. $M = \sum Y_i >> W = \sum X_i$
  2. $M - W = \sum Y_i - \sum X_i \simeq k = n$ (sample size)
  3. $M/N = (\sum Y_i)/(\sum m_i) >> 0.5, \ N = \sum m_i = M + W$
  4. $M/N = (\sum Y_i)/(\sum m_i) \simeq 0.5 + \dfrac{k}{2\sum m_i}$
  5. $\dfrac{M-k}{N-k} = \dfrac{\sum Y_i - k}{\sum m_i - k} \simeq 0.5$

Rao's results

City	N	M	W	M-W	k	M/N	$\frac{1}{2}+\frac{k}{2N}$	$\frac{M-k}{N-k}$
Tehran	105	65	40	25	21	0.619	0.600	0.524
Isphahan	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
Ahmed.	133	84	49	35	29	0.632	0.609	0.529
Bangalore	307	180	127	53	55	0.586	0.589	0.496

Questions

- How to estimate p_H or p_M?

Questions

- How to estimate p_H or p_M?
- How to estimate $E(m_i)$?

Questions

- How to estimate p_H or p_M?
- How to estimate $E(m_i)$?
- Which sample is the best one?

Questions

- How to estimate p_H or p_M?
- How to estimate $E(m_i)$?
- Which sample is the best one?
- Can we use both samples together?

Questions

- How to estimate p_H or p_M?
- How to estimate $E(m_i)$?
- Which sample is the best one?
- Can we use both samples together?
- How can we obtain the best results?

Solutions

- The number of brothers is a Binomial $B(m, p_M)$, with $p_M \simeq 0.5$

$$p(x) = \Pr(X = x) = \binom{m}{x} p_M^x \cdot p_W^{m-x}$$
$$E(X) = m p_M$$

Solutions

- The number of brothers is a Binomial $B(m, p_M)$, with $p_M \simeq 0.5$

$$p(x) = \Pr(X = x) = \binom{m}{x} p_M^x \cdot p_W^{m-x}$$
$$E(X) = m p_M$$

- The sampling probability of Y_i is proportional to Y_i.

Solutions

- The number of brothers is a Binomial $B(m, p_M)$, with $p_M \simeq 0.5$

$$p(x) = \Pr(X = x) = \binom{m}{x} p_M^x \cdot p_W^{m-x}$$
$$E(X) = mp_M$$

- The sampling probability of Y_i is proportional to Y_i.
- Hence Y is a length biased Binomial $Y \equiv B^*(m_i, p_M)$

$$p^*(x) = \frac{xp(x)}{E(X)} = x\binom{m_i}{x} p_M^x \cdot p_W^{m_i-x}/(m_i p_M)$$
$$= x\frac{xm_i!}{m_i x!(m_i - x)!} p_M^{x-1} p_W^{m_i-x} = \binom{m_i - 1}{x - 1} p_M^{x-1} p_W^{m-x}; x = 1, 2,$$

Solutions

- The number of brothers is a Binomial $B(m, p_M)$, with $p_M \simeq 0.5$

$$p(x) = \Pr(X = x) = \binom{m}{x} p_M^x \cdot p_W^{m-x}$$
$$E(X) = mp_M$$

- The sampling probability of Y_i is proportional to Y_i.
- Hence Y is a length biased Binomial $Y \equiv B^*(m_i, p_M)$

$$p^*(x) = \frac{xp(x)}{E(X)} = x\binom{m_i}{x} p_M^x \cdot p_W^{m_i-x}/(m_i p_M)$$
$$= x\frac{xm_i!}{m_i x!(m_i-x)!} p_M^{x-1} p_W^{m_i-x} = \binom{m_i-1}{x-1} p_M^{x-1} p_W^{m-x}; x = 1, 2,$$

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

Predictions:

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

Predictions:

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

Predictions:

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

Predictions:

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

Predictions:

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

Predictions:

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

Predictions:

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

Predictions:

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

Predictions:

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

Questions

- How to estimate p_M?

Questions

- How to estimate p_M?
- We can use:

$$
\begin{aligned}
T &= \frac{\sum Y_i - k}{\sum m_i - k} \\
E(T) &= E(\frac{\sum Y_i - k}{\sum m_i - k}) = p_M \\
Vat(T) &= p_M p_W / (\sum m_i - k) \to 0 \\
\sum Y_i - k &\equiv B(\sum m_i - k, p_M) \\
T &\approxeq Normal \\
T & \quad \text{is an UMVUE}
\end{aligned}
$$

Questions

- How to use both samples?

Questions

- How to use both samples?
- Let $X_1, ..., X_n$ be an unbiased sample from $X_i \equiv B(n_i, p)$.

Questions

- How to use both samples?
- Let $X_1, \ldots, X_n$ be an unbiased sample from $X_i \equiv B(n_i, p)$.
- Let $Y_1, \ldots, Y_m$ be a length biased sample.

Questions

- How to use both samples?
- Let $X_1, \ldots, X_n$ be an unbiased sample from $X_i \equiv B(n_i, p)$.
- Let $Y_1, \ldots, Y_m$ be a length biased sample.
- Then $Y_j - 1 \equiv B(m_j - 1, p)$ and

$$
\begin{aligned}
T &= \frac{\sum X_i + \sum (Y_j - 1)}{\sum n_i + \sum (m_j - 1)} \\
E(T) &= E\left(\frac{\sum X_i + \sum (Y_j - 1)}{\sum n_i + \sum (m_j - 1)}\right) = p \\
Vat(T) &= p(1-p)/(\sum n_i - \sum (m_j - 1)) \\
\sum X_i + \sum (Y_j - 1) &\equiv B(\sum n_i - \sum (m_j - 1), p_M) \\
T &\approxeq Normal \\
T & \quad \text{is UMVUE}
\end{aligned}
$$

Questions

- What is the best sample?

Questions

- What is the best sample?
- If $Y_j = 1$, then the information is null.

Questions

- What is the best sample?
- If $Y_j = 1$, then the information is null.
- X_i has more information than Y_j if $n_i > m_j - 1$

Questions

- What is the best sample?
- If $Y_j = 1$, then the information is null.
- X_i has more information than Y_j if $n_i > m_j - 1$
- The Fisher's information ($I_1 = E[(\frac{\partial}{\partial p} p(x))^2]$) are:

$$I_{X_i}(p) = \frac{n_i}{pq}$$

$$I_{Y_j}(p) = \frac{m_j - 1}{pq}$$

Questions

- What is the best sample?
- If $Y_j = 1$, then the information is null.
- X_i has more information than Y_j if $n_i > m_j - 1$
- The Fisher's information ($I_1 = E[(\frac{\partial}{\partial p} p(x))^2]$) are:

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

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

Questions

- What is the best sample?
- If $Y_j = 1$, then the information is null.
- X_i has more information than Y_j if $n_i > m_j - 1$
- The Fisher's information ($I_1 = E[(\frac{\partial}{\partial p} p(x))^2]$) are:

$$I_{X_i}(p) = \frac{n_i}{pq}$$

$$I_{Y_j}(p) = \frac{m_j - 1}{pq}$$

- $E(n_i) =?, E(m_j) =?$ $(m_j \geq 1)$
- In our survey $m_j - 1 = n_j$, so both samples have the same information (in each data).

Questions

- What is the best sample?
- If $Y_j = 1$, then the information is null.
- X_i has more information than Y_j if $n_i > m_j - 1$
- The Fisher's information ($I_1 = E[(\frac{\partial}{\partial p} p(x))^2]$) are:

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

- $E(n_i) = ?, E(m_j) = ?$ $(m_j \geq 1)$
- In our survey $m_j - 1 = n_j$, so both samples have the same information (in each data).
- The best option is to use both samples together!

Additional questions

- How to estimate the number of children m?

Additional questions

- How to estimate the number of children m?
- Can we use $\overline{m} = \frac{1}{k} \sum m_i$?

Additional questions

- How to estimate the number of children m?
- Can we use $\overline{m} = \frac{1}{k}\sum m_i$?
- If we use men and women, the sampling probability of a family with m_i children is proportional to m_i

Additional questions

- How to estimate the number of children m?
- Can we use $\overline{m} = \frac{1}{k} \sum m_i$?
- If we use men and women, the sampling probability of a family with m_i children is proportional to m_i
- If we only use men, it is proportional to $E(X_i) = m_i p_M$

Additional questions

- How to estimate the number of children m?
- Can we use $\overline{m} = \frac{1}{k} \sum m_i$?
- If we use men and women, the sampling probability of a family with m_i children is proportional to m_i
- If we only use men, it is proportional to $E(X_i) = m_i p_M$
- Then $m_1, ..., m_k$ is a length biased sample from m.

Additional questions

- How to estimate the number of children m?
- Can we use $\overline{m} = \frac{1}{k} \sum m_i$?
- If we use men and women, the sampling probability of a family with m_i children is proportional to m_i
- If we only use men, it is proportional to $E(X_i) = m_i p_M$
- Then $m_1, \ldots, m_k$ is a length biased sample from m.
- How to estimate $E(m)$ using $m_1, \ldots, m_k$?

Additional questions

- If $m \equiv Poisson(\mu)$, μ =mean number of children

$$p(x) = \mu^x e^{-\mu}/x!; x = 0, 1, \ldots$$

Additional questions

- If $m \equiv Poisson(\mu), \mu =$mean number of children

$$p(x) = \mu^x e^{-\mu}/x!; x = 0, 1, \ldots$$

- Hence $m_j \equiv$ size biased Poisson with

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

Additional questions

- If $m \equiv Poisson(\mu), \mu =$mean number of children

$$p(x) = \mu^x e^{-\mu}/x!; x = 0, 1, \ldots$$

- Hence $m_j \equiv$ size biased Poisson with

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

- Then $m_j - 1 \equiv Poisson(\mu)$

Additional questions

- If $m \equiv Poisson(\mu), \mu =$mean number of children

$$p(x) = \mu^x e^{-\mu}/x!; x = 0, 1, \dots$$

- Hence $m_j \equiv$ size biased Poisson with

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

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

Additional questions

- If $m \equiv Poisson(\mu), \mu =$mean number of children

$$p(x) = \mu^x e^{-\mu}/x!; x = 0, 1, \ldots$$

- Hence $m_j \equiv$ size biased Poisson with

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

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

Additional questions

- If $m \equiv Poisson(\mu), \mu =$mean number of children

$$p(x) = \mu^x e^{-\mu}/x!; x = 0, 1, \ldots$$

- Hence $m_j \equiv$ size biased Poisson with

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

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

Additional questions

- If $m \equiv Poisson(\mu), \mu =$mean number of children

$$p(x) = \mu^x e^{-\mu}/x!; x = 0, 1, \ldots$$

- Hence $m_j \equiv$ size biased Poisson with

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

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

Additional questions

- If $m \equiv Poisson(\mu), \mu =$mean number of children

$$p(x) = \mu^x e^{-\mu}/x!; x = 0, 1, \ldots$$

- Hence $m_j \equiv$ size biased Poisson with

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

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

Additional questions

- If $m \equiv Poisson(\mu), \mu =$mean number of children

$$p(x) = \mu^x e^{-\mu}/x!; x = 0, 1, \dots$$

- Hence $m_j \equiv$ size biased Poisson with

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

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

Additional questions

- Results

Additional questions

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

Additional questions

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

Additional questions

- Results
- $\overline{m} = \frac{1}{k}\sum m_i = N/k$
- $T = \overline{m} - 1 = (N-k)/k$
- Rao's results

City	N	M	W	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

Fisher's example

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

Fisher's example

- R. A., Fisher (1934). The effect of methods of ascertainment upon the estimation of frequencies. *Annals Eugenics* **6**, 13-25.
- Purpose: to study the proportion p of albino children from non-albino parents (which can have albino children).

Fisher's example

- R. A., Fisher (1934). The effect of methods of ascertainment upon the estimation of frequencies. *Annals Eugenics* **6**, 13-25.
- Purpose: to study the proportion p of albino children from non-albino parents (which can have albino children).
- From Medel's laws, p should be $1/4$

Fisher's example

- R. A., Fisher (1934). The effect of methods of ascertainment upon the estimation of frequencies. *Annals Eugenics* **6**, 13-25.
- Purpose: to study the proportion p of albino children from non-albino parents (which can have albino children).
- From Medel's laws, p should be $1/4$
- We do not know if two non-albino parents can have albino children!

Fisher's example

- R. A., Fisher (1934). The effect of methods of ascertainment upon the estimation of frequencies. *Annals Eugenics* **6**, 13-25.
- Purpose: to study the proportion p of albino children from non-albino parents (which can have albino children).
- From Medel's laws, p should be $1/4$
- We do not know if two non-albino parents can have albino children!
- So Fisher only consider families with albino children.

Fisher's example

- R. A., Fisher (1934). The effect of methods of ascertainment upon the estimation of frequencies. *Annals Eugenics* **6**, 13-25.
- Purpose: to study the proportion p of albino children from non-albino parents (which can have albino children).
- From Medel's laws, p should be $1/4$
- We do not know if two non-albino parents can have albino children!
- So Fisher only consider families with albino children.
- He only consider families with 5 children, obtaining the following data:

Fisher data

	Number of albino children in the family					
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=Number of albino children in the sample.

Fisher data

	Number of albino children in the family					
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=Number of albino children in the sample.
- Nótice that we have 340 families sampled from 432 different albino children.

Solution 1

- What to do with these data?

Solution 1

- What to do with these data?
- If $X_1, \ldots, X_n$ is a sampe of size $n = 340$ from a Binomial$B(k = 5, p = 1/4)$,

Solution 1

- What to do with these data?
- If $X_1, \ldots, X_n$ is a sampe of size $n = 340$ from a Binomial $B(k = 5, p = 1/4)$,
- p can be estimated as

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

Solution 1

- What to do with these data?
- If $X_1, \ldots, X_n$ is a sampe of size $n = 340$ from a Binomial $B(k = 5, p = 1/4)$,
- p can be estimated as

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

- with variance

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

Solution 1

- What to do with these data?
- If $X_1, \ldots, X_n$ is a sampe of size $n = 340$ from a Binomial $B(k = 5, p = 1/4)$,
- p can be estimated as

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

- with variance

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

- This gives $2\sigma(\widehat{p}_1) \simeq 0.021$ and we reject $p = 0.25$.

Solution 1bis

- If we use the families several times then

Solution 1bis

- If we use the families several times then
- the sample size is $n = 432$ and p is estimated as

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

Solution 1bis

- If we use the families several times then
- the sample size is $n = 432$ and p is estimated as

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

- This also leads to reject $p = 0.25$.

Solution 2

- The families with 0 albino children cannot appear in the sample.

Solution 2

- The families with 0 albino children cannot appear in the sample.
- Thus, we might think in a censored sample with $w(x) = 1$ for $x \neq 0$ and $w(0) = 0$.

Solution 2

- The families with 0 albino children cannot appear in the sample.
- Thus, we might think in a censored sample with $w(x) = 1$ for $x \neq 0$ and $w(0) = 0$.
- Then $p^*(x) = p(x)/(1 - q^5)$, where $p(x) \equiv$ Binomial $B(5, 1/4)$

Solution 2

- The families with 0 albino children cannot appear in the sample.
- Thus, we might think in a censored sample with $w(x) = 1$ for $x \neq 0$ and $w(0) = 0$.
- Then $p^*(x) = p(x)/(1 - q^5)$, where $p(x) \equiv$ Binomial $B(5, 1/4)$
- Then the MLE satisfies

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

Solution 2

- The families with 0 albino children cannot appear in the sample.
- Thus, we might think in a censored sample with $w(x) = 1$ for $x \neq 0$ and $w(0) = 0$.
- Then $p^*(x) = p(x)/(1 - q^5)$, where $p(x) \equiv$ Binomial $B(5, 1/4)$
- Then the MLE satisfies

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

- which gives $\widehat{p}_2 = 0.3085$ ($\widehat{p}_2 = 0.35$ with the repeated families).

Solution 2

- The families with 0 albino children cannot appear in the sample.
- Thus, we might think in a censored sample with $w(x) = 1$ for $x \neq 0$ and $w(0) = 0$.
- Then $p^*(x) = p(x)/(1 - q^5)$, where $p(x) \equiv$ Binomial $B(5, 1/4)$
- Then the MLE satisfies

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

- which gives $\widehat{p}_2 = 0.3085$ ($\widehat{p}_2 = 0.35$ with the repeated families).
- In both cases we reject $p = 1/4$.

Solution 3 (the correct one)

- Note that the sampling probability of a family with x albino children is proportional to x.

Solution 3 (the correct one)

- Note that the sampling probability of a family with x albino children is proportional to x.
- Then $X_i \equiv$ length biased Binomial.

Solution 3 (the correct one)

- Note that the sampling probability of a family with x albino children is proportional to x.
- Then $X_i \equiv$ length biased Binomial.
- That is, $X_i - 1 \equiv$Binomial $B(4, 1/4)$

Solution 3 (the correct one)

- Note that the sampling probability of a family with x albino children is proportional to x.
- Then $X_i \equiv$ length biased Binomial.
- That is, $X_i - 1 \equiv$Binomial $B(4, 1/4)$
- Then, using the repeated families p is estimated as

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

Solution 3 (the correct one)

- Note that the sampling probability of a family with x albino children is proportional to x.
- Then $X_i \equiv$ length biased Binomial.
- That is, $X_i - 1 \equiv$Binomial $B(4, 1/4)$
- Then, using the repeated families p is estimated as

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

- The variance satisfies $2\sigma(\widehat{p}_3) \simeq 0.0208$, which is consistent with $p = 1/4$.

Solution 3 (the correct one)

- Note that the sampling probability of a family with x albino children is proportional to x.
- Then $X_i \equiv$ length biased Binomial.
- That is, $X_i - 1 \equiv$Binomial $B(4, 1/4)$
- Then, using the repeated families p is estimated as

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

- The variance satisfies $2\sigma(\widehat{p}_3) \simeq 0.0208$, which is consistent with $p = 1/4$.
- Notice that if we do not use the repeated families the p is underestimated as

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

Waiting time paradox

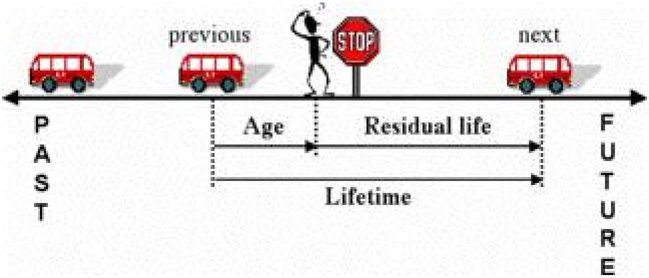

Figure: If a passenger arrives at a bus-stop at some random point and the interval time between the buses is 20 min, what is the mean waiting time until the next bus?

Waiting time paradox

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

Waiting time paradox

- R.C. Gupta 1979. Waiting time paradox and size biased sampling. *Communications in Statistics, Theory and Methods* **A8** (6), 601-607.
- Let us assume that the buses pass every 20 min. and that we do not know the time table:

|– – – –|– – – –|– – – –|– – – –|– – – –|– – – –|

Waiting time paradox

- R.C. Gupta 1979. Waiting time paradox and size biased sampling. *Communications in Statistics, Theory and Methods* **A8** (6), 601-607.
- Let us assume that the buses pass every 20 min. and that we do not know the time table:

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

- Then the waiting time T should be Uniform $(0, 20)$

Waiting time paradox

- R.C. Gupta 1979. Waiting time paradox and size biased sampling. *Communications in Statistics, Theory and Methods* **A8** (6), 601-607.
- Let us assume that the buses pass every 20 min. and that we do not know the time table:

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

- Then the waiting time T should be Uniform $(0, 20)$
- Then the expected waiting time should be $E(T) = 20/2 = 10\,\text{min}$.

Waiting time paradox

- R.C. Gupta 1979. Waiting time paradox and size biased sampling. *Communications in Statistics, Theory and Methods* **A8** (6), 601-607.
- Let us assume that the buses pass every 20 min. and that we do not know the time table:

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

- Then the waiting time T should be Uniform $(0, 20)$
- Then the expected waiting time should be $E(T) = 20/2 = 10\,\text{min}$.
- We know that this is not true!

Waiting time paradox

- The real times are:

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

Waiting time paradox

- The real times are:

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

- Then the time between buses is a random variable X.

Waiting time paradox

- The real times are:

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

- Then the time between buses is a random variable X.
- Let us assume that $\mu = E(X) = 20\,\text{min}$.

Waiting time paradox

- The real times are:

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

- Then the time between buses is a random variable X.
- Let us assume that $\mu = E(X) = 20\,\text{min}$.
- Then $T \equiv$Uniforme $(0, X)$

Waiting time paradox

- The real times are:

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

- Then the time between buses is a random variable X.
- Let us assume that $\mu = E(X) = 20\,\text{min}$.
- Then $T \equiv$Uniforme $(0, X)$
- The waiting time should be $T = UX$, where $U \equiv$Uniform $(0, 1)$.

Waiting time paradox

- The real times are:

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

- Then the time between buses is a random variable X.
- Let us assume that $\mu = E(X) = 20\,\text{min}$.
- Then $T \equiv$Uniforme $(0, X)$
- The waiting time should be $T = UX$, where $U \equiv$Uniform $(0, 1)$.
- Then the expected waiting time should be $E(T) = E(UX) = E(X)/2 = 10\,\text{min}$

Waiting time paradox

- The real times are:

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

- Then the time between buses is a random variable X.
- Let us assume that $\mu = E(X) = 20\,\text{min}$.
- Then T ≡Uniforme $(0, X)$
- The waiting time should be $T = UX$, where U ≡Uniform $(0, 1)$.
- Then the expected waiting time should be $E(T) = E(UX) = E(X)/2 = 10\,\text{min}$
- We know that this is not true!

Waiting time paradox

- The real times are:

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

Waiting time paradox

- The real times are:

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

- If $X_1, X_2, ..., X_n$ are the times between buses, the the probability of have a time X_i is proportional to X_i.

Waiting time paradox

- The real times are:

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

- If $X_1, X_2, ..., X_n$ are the times between buses, the the probability of have a time X_i is proportional to X_i.
- Then $T \equiv$Uniform $(0, X^*)$, where X^* is the length biased r.v.

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

Waiting time paradox

- The real times are:

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

- If $X_1, X_2, ..., X_n$ are the times between buses, the the probability of have a time X_i is proportional to X_i.
- Then $T \equiv$Uniform $(0, X^*)$, where X^* is the length biased r.v.

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

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

Waiting time paradox

- The real times are:

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

- If $X_1, X_2, ..., X_n$ are the times between buses, the the probability of have a time X_i is proportional to X_i.
- Then $T \equiv$Uniform $(0, X^*)$, where X^* is the length biased r.v.

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

- That is $E(T) = E(X^*/2) = 10 + \sigma^2/(20) > 10$
- We only have $E(T) = 10$ if $\sigma^2 = 0$!

Waiting time paradox

- The real times are:

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

- If $X_1, X_2, ..., X_n$ are the times between buses, the the probability of have a time X_i is proportional to X_i.
- Then $T \equiv$Uniform $(0, X^*)$, where X^* is the length biased r.v.

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

- That is $E(T) = E(X^*/2) = 10 + \sigma^2/(20) > 10$
- We only have $E(T) = 10$ if $\sigma^2 = 0$!
- It is very important the regularity!

Exponential case

- In particular, if $X \equiv Exp(\mu = 20\,\text{min})$

Exponential case

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

Exponential case

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

Exponential case

- In particular, if $X \equiv Exp(\mu = 20\,\text{min})$
- $E(X^*) = \mu + \sigma^2/\mu = \mu + \mu^2/\mu = 2\mu$
- $E(T) = E(X^*/2) = E(X)$!!
- Paradox: If the expected time between buses is 20 min., we have to wait 20 min.!

Exponential case

- In particular, if $X \equiv Exp(\mu = 20\,\text{min})$
- $E(X^*) = \mu + \sigma^2/\mu = \mu + \mu^2/\mu = 2\mu$
- $E(T) = E(X^*/2) = E(X)!!$
- Paradox: If the expected time between buses is 20 min., we have to wait 20 min.!
- Similar results are obtained in renewal processes (with random inspections).

General solution

- When a unit fails, it is replaced by a similar one

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

General solution

- When a unit fails, it is replaced by a similar one

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

- The unit lifetimes $X_1, X_2, \ldots$ are i.i.d. from X

General solution

- When a unit fails, it is replaced by a similar one

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

- The unit lifetimes $X_1, X_2, \ldots$ are i.i.d. from X
- We do random inspections.

General solution

- When a unit fails, it is replaced by a similar one

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

- The unit lifetimes $X_1, X_2, \ldots$ are i.i.d. from X
- We do random inspections.
- The forward (or backward) time from a sample point is $T = UX$, where $U \equiv$Uniform $(0, 1)$ (X and U are independent).

General solution

- When a unit fails, it is replaced by a similar one

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

- The unit lifetimes $X_1, X_2, \ldots$ are i.i.d. from X
- We do random inspections.
- The forward (or backward) time from a sample point is $T = UX$, where $U \equiv$Uniform $(0, 1)$ (X and U are independent).
- This is not true!

General solution

- The correct solution is $T = UX^*$, and hence

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

General solution

- The correct solution is $T = UX^*$, and hence

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

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

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

General solution

- The correct solution is $T = UX^*$, and hence

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

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

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

- Thus,

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

$$w(t) = \frac{1-F(t)}{f(t)} = \frac{1}{h(t)};\ \text{where } h(t) = \frac{f(t)}{1-F(t)} \text{ is the hazard r}$$

Waiting time solution

- If $X \equiv$, $T = UX^*$, $t > 0$,

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

Waiting time solution

- If $X \equiv$, $T = UX^*$, $t > 0$,

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

- That is, $T \equiv$Exponential: $X =_d UX^*$.

Waiting time solution

- If $X \equiv$, $T = UX^*$, $t > 0$,

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

- That is, $T \equiv$Exponential: $X =_d UX^*$.
- Actually, the exponential model is the unique model such that $X =_d UX^*$ (or $X =_d T$).

Waiting time solution

- If $X \equiv$, $T = UX^*$, $t > 0$,

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

- That is, $T \equiv$Exponential: $X =_d UX^*$.
- Actually, the exponential model is the unique model such that $X =_d UX^*$ (or $X =_d T$).
- The distribution of T is called the equilibrium distribution and very interesting properties.

Waiting time solution

- If $X \equiv$, $T = UX^*$, $t > 0$,

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

- That is, $T \equiv$Exponential: $X =_d UX^*$.
- Actually, the exponential model is the unique model such that $X =_d UX^*$ (or $X =_d T$).
- The distribution of T is called the equilibrium distribution and very interesting properties.
- For example,

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

where $m(t) = E(X - t|X > t)$ in the mean residual life.

How to detect biased samples?

- In Fisher and Rao examples the results do not fit to the expected values.

How to detect biased samples?

- In Fisher and Rao examples the results do not fit to the expected values.
- In the waiting time paradox the results do not coindice with our experience.

How to detect biased samples?

- In Fisher and Rao examples the results do not fit to the expected values.
- In the waiting time paradox the results do not coindice with our experience.
- This example is based on two surveys to study the mean sojourn time per tourist in Morocco (INSEA, 1966).
- G.P. Patil 1981. Proceedings of the Indian Statistical Institute Jubilee International Conference on Statistics: Applications and New Directions, 478-503)

How to detect biased samples?

- In Fisher and Rao examples the results do not fit to the expected values.
- In the waiting time paradox the results do not coindice with our experience.
- This example is based on two surveys to study the mean sojourn time per tourist in Morocco (INSEA, 1966).
- G.P. Patil 1981. Proceedings of the Indian Statistical Institute Jubilee International Conference on Statistics: Applications and New Directions, 478-503)
- A sample at the border stations: n=3000, mean=9.0 days

How to detect biased samples?

- In Fisher and Rao examples the results do not fit to the expected values.
- In the waiting time paradox the results do not coindice with our experience.
- This example is based on two surveys to study the mean sojourn time per tourist in Morocco (INSEA, 1966).
- G.P. Patil 1981. Proceedings of the Indian Statistical Institute Jubilee International Conference on Statistics: Applications and New Directions, 478-503)
- A sample at the border stations: n=3000, mean=9.0 days
- A sample at the hotels: n=12321, mean=17.8 days

How to detect biased samples?

- In Fisher and Rao examples the results do not fit to the expected values.
- In the waiting time paradox the results do not coindice with our experience.
- This example is based on two surveys to study the mean sojourn time per tourist in Morocco (INSEA, 1966).
- G.P. Patil 1981. Proceedings of the Indian Statistical Institute Jubilee International Conference on Statistics: Applications and New Directions, 478-503)
- A sample at the border stations: n=3000, mean=9.0 days
- A sample at the hotels: n=12321, mean=17.8 days
- The results are very different!

How to detect biased samples?

- In Fisher and Rao examples the results do not fit to the expected values.
- In the waiting time paradox the results do not coindice with our experience.
- This example is based on two surveys to study the mean sojourn time per tourist in Morocco (INSEA, 1966).
- G.P. Patil 1981. Proceedings of the Indian Statistical Institute Jubilee International Conference on Statistics: Applications and New Directions, 478-503)
- A sample at the border stations: n=3000, mean=9.0 days
- A sample at the hotels: n=12321, mean=17.8 days
- The results are very different!
- The second sample was discarded.

Patil's Solution

- The sample at the hotels is length biased !

Patil's Solution

- The sample at the hotels is length biased !
- A tourist staying 6 days has double sampling probability than a tourist staying 3 days.

Patil's Solution

- The sample at the hotels is length biased !
- A tourist staying 6 days has double sampling probability than a tourist staying 3 days.
- Hence $E(X) \simeq 9.0 < E(X^*) \simeq 17.8$

Patil's Solution

- The sample at the hotels is length biased !
- A tourist staying 6 days has double sampling probability than a tourist staying 3 days.
- Hence $E(X) \simeq 9.0 < E(X^*) \simeq 17.8$
- Actually $E(X^*) \simeq 2E(X)$ might indicate that X is Exponential.

Patil's Solution

- The sample at the hotels is length biased !
- A tourist staying 6 days has double sampling probability than a tourist staying 3 days.
- Hence $E(X) \simeq 9.0 < E(X^*) \simeq 17.8$
- Actually $E(X^*) \simeq 2E(X)$ might indicate that X is Exponential.
- Correct estimation $E(X) \simeq 17.8/2 = 8.9$

Patil's Solution

- The sample at the hotels is length biased !
- A tourist staying 6 days has double sampling probability than a tourist staying 3 days.
- Hence $E(X) \simeq 9.0 < E(X^*) \simeq 17.8$
- Actually $E(X^*) \simeq 2E(X)$ might indicate that X is Exponential.
- Correct estimation $E(X) \simeq 17.8/2 = 8.9$
- Similar examples in other fields.

What to do?

- The best solution is to use all the information available!

What to do?

- The best solution is to use all the information available!
- $X_1, ..., X_n$ is a sample from $X \equiv Exp(\mu)$.

What to do?

- The best solution is to use all the information available!
- $X_1, ..., X_n$ is a sample from $X \equiv Exp(\mu)$.
- $Y_1, ..., Y_m$ is a sample from $X^* \equiv Exp^*(\mu)$.

What to do?

- The best solution is to use all the information available!
- $X_1, ..., X_n$ is a sample from $X \equiv Exp(\mu)$.
- $Y_1, ..., Y_m$ is a sample from $X^* \equiv Exp^*(\mu)$.
- The MLE (exponential) is:

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

What to do?

- The best solution is to use all the information available!
- $X_1, ..., X_n$ is a sample from $X \equiv Exp(\mu)$.
- $Y_1, ..., Y_m$ is a sample from $X^* \equiv Exp^*(\mu)$.
- The MLE (exponential) is:

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

- It is unbiased since $E(X_i) = \mu$ y $E(Y_j) = 2\mu$

What to do?

- The best solution is to use all the information available!
- $X_1, ..., X_n$ is a sample from $X \equiv Exp(\mu)$.
- $Y_1, ..., Y_m$ is a sample from $X^* \equiv Exp^*(\mu)$.
- The MLE (exponential) is:

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

- It is unbiased since $E(X_i) = \mu$ y $E(Y_j) = 2\mu$
- With variance

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

What to do?

- The best solution is to use all the information available!
- $X_1, ..., X_n$ is a sample from $X \equiv Exp(\mu)$.
- $Y_1, ..., Y_m$ is a sample from $X^* \equiv Exp^*(\mu)$.
- The MLE (exponential) is:

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

- It is unbiased since $E(X_i) = \mu$ y $E(Y_j) = 2\mu$
- With variance

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

- It is the UMVUE.

The best estimation

- The best estimation is

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

The best estimation

- The best estimation is

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

- with variance $2\sigma(\widehat{\mu}) \simeq 0.11$

The best estimation

- The best estimation is

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

- with variance $2\sigma(\widehat{\mu}) \simeq 0.11$
- First sample $\widehat{\mu} = 9.0$

The best estimation

- The best estimation is

$$\widehat{\mu} = \frac{1}{3000 + 2 \cdot 12321}(3000 \cdot 9 + 12321 \cdot 17.8) = 8.\,91,$$

- with variance $2\sigma(\widehat{\mu}) \simeq 0.11$
- First sample $\widehat{\mu} = 9.0$
- Second sample $\widehat{\mu} = 8.9$

General solution in the exponencial case

- What to do the next time ?

General solution in the exponencial case

- What to do the next time ?
- Which sample is the best one?

General solution in the exponencial case

- What to do the next time ?
- Which sample is the best one?
- Notice that the estimator from the second sample has less variance.

General solution in the exponencial case

- What to do the next time ?
- Which sample is the best one?
- Notice that the estimator from the second sample has less variance.
- The Fisher information for $n = m = 1$ are

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

General solution in the exponencial case

- What to do the next time ?
- Which sample is the best one?
- Notice that the estimator from the second sample has less variance.
- The Fisher information for $n = m = 1$ are

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

- Each data in the second sample has double information!

General solution in the exponencial case

- What to do the next time ?
- Which sample is the best one?
- Notice that the estimator from the second sample has less variance.
- The Fisher information for $n = m = 1$ are

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

- Each data in the second sample has double information!
- If the bias is $w(x) = x^k$, the Fisher information is increasing in k.

General solution in the exponencial case

- What to do the next time ?
- Which sample is the best one?
- Notice that the estimator from the second sample has less variance.
- The Fisher information for $n = m = 1$ are

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

- Each data in the second sample has double information!
- If the bias is $w(x) = x^k$, the Fisher information is increasing in k.
- Other models, see Navarro et al. (2001, Biom. J. 43).

DeGroot's example

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

DeGroot's example

- G.P. Patil (1991). Encountered data, Statistical Ecology Environmental Statistics and weighted distribution methods. Environmetrics 2(4), 377-423.
- Discussion by M.H. DeGroot (Profesor Carnegie Mellon University, Pittsburgh).

DeGroot's example

- G.P. Patil (1991). Encountered data, Statistical Ecology Environmental Statistics and weighted distribution methods. Environmetrics 2(4), 377-423.
- Discussion by M.H. DeGroot (Profesor Carnegie Mellon University, Pittsburgh).
- How to predict the stock market behavior?

DeGroot's example

- G.P. Patil (1991). Encountered data, Statistical Ecology Environmental Statistics and weighted distribution methods. Environmetrics 2(4), 377-423.
- Discussion by M.H. DeGroot (Profesor Carnegie Mellon University, Pittsburgh).
- How to predict the stock market behavior?
- We send 128 letters, 64 saying:

DeGroot's example

- G.P. Patil (1991). Encountered data, Statistical Ecology Environmental Statistics and weighted distribution methods. Environmetrics 2(4), 377-423.
- Discussion by M.H. DeGroot (Profesor Carnegie Mellon University, Pittsburgh).
- How to predict the stock market behavior?
- We send 128 letters, 64 saying:
- "I'm an expert analyst and I have a model to predict the the stock market behavior. To show that I inform you fro FREE that the stocks of the company SOME are going to go UP this week.".

DeGroot's example

- G.P. Patil (1991). Encountered data, Statistical Ecology Environmental Statistics and weighted distribution methods. Environmetrics 2(4), 377-423.
- Discussion by M.H. DeGroot (Profesor Carnegie Mellon University, Pittsburgh).
- How to predict the stock market behavior?
- We send 128 letters, 64 saying:
- "I'm an expert analyst and I have a model to predict the the stock market behavior. To show that I inform you fro FREE that the stocks of the company SOME are going to go UP this week.".
- The other 64 letters say: "... to go DOWN this week".

DeGroot's example

- The next week we will send similar letter but only to the people (64) with the correct predictions saying:

DeGroot's example

- The next week we will send similar letter but only to the people (64) with the correct predictions saying:
- "Last week I sent you a correct predictions. To show you that my model does not fail I send you another correct prediction for FREE this week: the stocks of the company SOME are going to go UP (DOWN)".

DeGroot's example

- The next week we will send similar letter but only to the people (64) with the correct predictions saying:
- "Last week I sent you a correct predictions. To show you that my model does not fail I send you another correct prediction for FREE this week: the stocks of the company SOME are going to go UP (DOWN)".
- We repeat this process 7 weeks.

DeGroot's example

- The next week we will send similar letter but only to the people (64) with the correct predictions saying:
- "Last week I sent you a correct predictions. To show you that my model does not fail I send you another correct prediction for FREE this week: the stocks of the company SOME are going to go UP (DOWN)".
- We repeat this process 7 weeks.
- Finally we sent the following letter:

DeGroot's example

- The next week we will send similar letter but only to the people (64) with the correct predictions saying:
- "Last week I sent you a correct predictions. To show you that my model does not fail I send you another correct prediction for FREE this week: the stocks of the company SOME are going to go UP (DOWN)".
- We repeat this process 7 weeks.
- Finally we sent the following letter:
- "Well I think that I have show you that my model does not fail. Now if you want to know the next prediction you have to pay 10.000$".

Solution of DeGroot's example

- We have a sample $X_1, \ldots, X_7$ from a Bernoulli $B(p)$ with a probability p of a correct prediction $X_i = 1$ and a estimation $\widehat{p} = 7/7 = 1$.

Solution of DeGroot's example

- We have a sample $X_1, ..., X_7$ from a Bernoulli $B(p)$ with a probability p of a correct prediction $X_i = 1$ and a estimation $\widehat{p} = 7/7 = 1$.
- But, what is the probability of a value X_i appear in the sample?

Solution of DeGroot's example

- We have a sample $X_1, ..., X_7$ from a Bernoulli $B(p)$ with a probability p of a correct prediction $X_i = 1$ and a estimation $\widehat{p} = 7/7 = 1$.
- But, what is the probability of a value X_i appear in the sample?
- Clearly, it is proportional to X_i!

Solution of DeGroot's example

- We have a sample $X_1, ..., X_7$ from a Bernoulli $B(p)$ with a probability p of a correct prediction $X_i = 1$ and a estimation $\widehat{p} = 7/7 = 1$.
- But, what is the probability of a value X_i appear in the sample?
- Clearly, it is proportional to X_i!
- That is we have a sample from the length biased r.v. X^* with $p^*(x) = xp(x)/\mu$, $x = 0, 1$, that is, $X^* = 1$.

Conclusions

- We have to think about the selection methods in a sample!

Conclusions

- We have to think about the selection methods in a sample!
- They can be biased but with a known bias.

Conclusions

- We have to think about the selection methods in a sample!
- They can be biased but with a known bias.
- If we have two different samples (with a known bias), the best solution is always to use both together. We need to change the classical estimators.

Conclusions

- We have to think about the selection methods in a sample!
- They can be biased but with a known bias.
- If we have two different samples (with a known bias), the best solution is always to use both together. We need to change the classical estimators.
- With a biased sample, we can obtain results as good as (or even better) that an unbiased sample. We need to change the classical estimators.

Conclusions

- We have to think about the selection methods in a sample!
- They can be biased but with a known bias.
- If we have two different samples (with a known bias), the best solution is always to use both together. We need to change the classical estimators.
- With a biased sample, we can obtain results as good as (or even better) that an unbiased sample. We need to change the classical estimators.
- If we have to choose, we should use the sample (biased or not) with the highest information about the parameter.

- 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 J. 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 J. 43, 297-311.
- Navarro, Ruiz and del Aguila (2003). How to detect biased samples?. Biometrical J. 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 J. 43, 297-311.
- Navarro, Ruiz and del Aguila (2003). How to detect biased samples?. Biometrical J. 44, 742-763.
- Pakes, Navarro, Ruiz and del Aguila (2003). Characterizations using weighted distributions. JSPI 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 J. 43, 297-311.
- Navarro, Ruiz and del Aguila (2003). How to detect biased samples?. Biometrical J. 44, 742-763.
- Pakes, Navarro, Ruiz and del Aguila (2003). Characterizations using weighted distributions. JSPI 116, 389-420.
- Navarro, Ruiz and del Aguila (2006). Multivariate weighted distributions. Statist. 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 J. 43, 297-311.
- Navarro, Ruiz and del Aguila (2003). How to detect biased samples?. Biometrical J. 44, 742-763.
- Pakes, Navarro, Ruiz and del Aguila (2003). Characterizations using weighted distributions. JSPI 116, 389-420.
- Navarro, Ruiz and del Aguila (2006). Multivariate weighted distributions. Statist. 40 (1), 51-54.
- Pakes and Navarro (2007). Distributional characterizations through scaling relations. Australian and New Zealand J. Statist. 49 (2), 115-135.

- 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 J. 43, 297-311.
- Navarro, Ruiz and del Aguila (2003). How to detect biased samples?. Biometrical J. 44, 742-763.
- Pakes, Navarro, Ruiz and del Aguila (2003). Characterizations using weighted distributions. JSPI 116, 389-420.
- Navarro, Ruiz and del Aguila (2006). Multivariate weighted distributions. Statist. 40 (1), 51-54.
- Pakes and Navarro (2007). Distributional characterizations through scaling relations. Australian and New Zealand J. Statist. 49 (2), 115-135.
- Navarro and Sarabia (2010). Alternative definitions of bivariate equilibrium distributions. JSPI 140, 2046-2056.

- 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.