What Population does your sample represent ?

Barry C. Arnold

Statistics Department, UCR.

Gradquant, May 20, 2015

Summary: Not infrequently data are collected to study a particular distribution or population but, because of the sampling mechanism used, the sample is not representative of the desired target distribution. For example, size biasing occurs when large items are more likely to be included in the sample than are small ones (a relatively frequent occurrence). Hidden truncation occurs when observations are only made subject to constraints on covariables (also more frequent than one might suspect).

We’ll begin with some examples.

(1)Hospital sojourn times.

We wish to determine the average length of stay in a large hospital.

AccuMax
Hill-Rom

Strategy: Pick a particular time of day, say

Noon on July 13.

Randomly select 50 of the occupied beds in the hospital.

Identify the patient in each bed.

- For each patient, say patient “i”, determine how long he/she has been in the hospital say Y(i), and also track the patient to determine Z(i), the additional time until that patient is released.

- The total time that patient “i” spent in the hospital is then
-
- $X(i)=Y(i)+Z(i)$

Our estimate of the average time that a patient spends in the hospital is then:

$$\hat{\mu} = \frac{1}{50} \sum_{i=1}^{50} X_i .$$

Our estimate of the average time that a patient spends in the hospital is then:

$$\hat{\mu} = \frac{1}{50} \sum_{i=1}^{50} X_i .$$

Sounds good ?

Our estimate of the average time that a patient spends in the hospital is then:

$$\hat{\mu} = \frac{1}{50}\sum_{i=1}^{50} X_i.$$

Sounds good ?

Any flaws ?

We’ll come back to the hospital in a while, but before we do, let’s look at another example.

We’ll come back to the hospital in a while, but before we do, let’s look at another example.

(2) What proportion of the children in the families of UCR students are female ?

We’ll come back to the hospital in a while, but before we do, let’s look at another example.

(2) What proportion of the children in the families of UCR students are female ?

To get a quick estimate, assuming my class is not atypical, I’ll do the following:

- In my class there are 35 female students.
- For each female student , I ask how many brothers she has and how many sisters she has. Assume that all 35 are from different families.

- For female student “i” we get $Y(i)$ brothers and $Z(i)$ sisters.

A TYPICAL FAMILY

Our estimate of the proportion of females at UCR will then be:

$$\hat{p} = \frac{\sum_{i=1}^{35} [1 + Y(i)]}{\sum_{i=1}^{35} [1 + Y(i) + X(i)]}$$

Our estimate of the proportion of females at UCR will then be:

$$\hat{p} = \frac{\sum_{i=1}^{35} [1 + Y(i)]}{\sum_{i=1}^{35} [1 + Y(i) + X(i)]}$$

Sounds good ?

Our estimate of the proportion of females at UCR will then be:

$$\hat{p} = \frac{\sum_{i=1}^{35} [1 + Y(i)]}{\sum_{i=1}^{35} [1 + Y(i) + X(i)]}$$

Sounds good ?

Any flaws ?

Size bias !!!!

Both examples are instances in which size bias is present.

Size bias !!!!

Both examples are instances in which size bias is present.

This occurs when big items are more likely to be included in the sample than are small items.

Size bias !!!!

Both examples are instances in which size bias is present.

This occurs when big items are more likely to be included in the sample than are small items.

And, of course, it typically leads to overestimation of the population mean.

One more example:Estimating the average size of reindeer herds in Lapland.

One more example:Estimating the average size of reindeer herds in Lapland.

We fly around in a helicopter and when we see a herd, we take a picture so that we can count the number of animals in the herd, when we land after a day of searching.

© ExclusivePix

One more example:Estimating the average size of reindeer herds in Lapland.

We fly around in a helicopter and when we see a herd, we take a picture so that we can count the number of animals in the herd, when we land after a day of searching.

The resulting data will be N(1),N(2),…,N(m), where m denotes the number of herds that we observed, and N(i) is the number of animals in the i”th herd.

One more example:Estimating the average size of reindeer herds in Lapland.

Our estimate of the average herd size will then be:

$$\hat{\mu} = \frac{1}{m} \sum_{i=1}^{m} N(i).$$

One more example:Estimating the average size of reindeer herds in Lapland.

Our estimate of the average herd size will then be:

$$\hat{\mu} = \frac{1}{m}\sum_{i=1}^{m} N(i).$$

Sounds good ?

One more example:Estimating the average size of reindeer herds in Lapland.

Our estimate of the average herd size will then be:

$$\hat{\mu} = \frac{1}{m} \sum_{i=1}^{m} N(i).$$

Sounds good ? Any flaws ?

One more example:Estimating the average size of reindeer herds in Lapland.

This example is one in which size-bias is quite evident.

One more example:Estimating the average size of reindeer herds in Lapland.

This example is one in which size-bias is quite evident.

It will hard to spot small herds from the helicopter, big herds will be hard to miss.

How serious is the problem ?

How serious is the problem ?

With the reindeer herds, it will depend on how sharp-eyed our spotter is. It may be hard to get a good estimate from such a potentially flawed sampling method.

How serious is the problem ?

With the reindeer herds, it will depend on how sharp-eyed our spotter is. It may be hard to get a good estimate from such a potentially flawed sampling method.

Concerning the proportion of females using data from my class, the estimate is bad, but we can correct it.

How serious is the problem ?

With the reindeer herds, it will depend on how sharp-eyed our spotter is. It may be hard to get a good estimate from such a potentially flawed sampling method.

Concerning the proportion of females using data from my class, the estimate is bad, but we can correct it.

Concerning the average stay in the hospital; the estimate can be off by almost a factor of 2   !!

On the proportion of females

Our naïve estimate was

$$\hat{p} = \frac{\sum_{i=1}^{35} [1 + Y(i)]}{\sum_{i=1}^{35} [1 + Y(i) + X(i)]}$$

On the proportion of females

Our naïve estimate was

$$\hat{p} = \frac{\sum_{i=1}^{35} [1 + Y(i)]}{\sum_{i=1}^{35} [1 + Y(i) + X(i)]}$$

A better estimate would be

$$\hat{\hat{p}} = \frac{\sum_{i=1}^{35} [Y(i)]}{\sum_{i=1}^{35} [Y(i) + X(i)]}$$

On the proportion of females

Our naïve estimate was

$$\hat{p} = \frac{\sum_{i=1}^{35}[1 + Y(i)]}{\sum_{i=1}^{35}[1 + Y(i) + X(i)]}$$

A better estimate would be

$$\hat{\hat{p}} = \frac{\sum_{i=1}^{35}[Y(i)]}{\sum_{i=1}^{35}[Y(i) + X(i)]}$$

Why ?

Time spent in the hospital.

Time spent in the hospital.

Here the bias can be considerable.

Time spent in the hospital.

Here the bias can be considerable.

We’ll look at a simple special case.

Time spent in the hospital.

Here the bias can be considerable.
We’ll look at a simple special case.

But, the argument will be a bit technical.

Time spent in the hospital.

Here the bias can be considerable.
We’ll look at a simple special case.

But, the argument will be a bit technical.

Perhaps a good time to check your e-mail on your phone, or your Apple watch.

Time spent in the hospital.

Here the bias can be considerable.
We’ll look at a simple special case.

But, the argument will be a bit technical.

Perhaps a good time to check your e-mail on your phone, or your Apple watch.

If you haven’t already done so.

Time spent in the hospital

Suppose that hospital stays are exponentially distributed, i.e., we have

$$f_X(x|\lambda) = \lambda e^{-\lambda x}, \quad x > 0$$

Assume that when a patient vacates a bed, it is immediately occupied by the next patient.

Time spent in the hospital

Under these asumptions the counting process tracking the number of occupants of a particular bed over time will be a Poisson process {N(t):t>0} with intensity \lambda.

Thus:

$$N(t) \sim Poisson(\lambda t)$$

$$P(N(t) = k) = \frac{(\lambda t)^k e^{-\lambda t}}{k!}$$

Such a process has stationary independent increments.

Time spent in the hospital

In this case a hospital stay will have an exponential distribution with mean $1/\lambda$.

But, curiously if we fix a time t*, and observe the expected wait until the next event (the next patient in the bed), it also will have an exponential distribution with mean $1/\lambda$ and, provided the hospital has been in operation for quite a while, the the time that the bed occupant has been in place has mean approximately equal to $1/\lambda$

Time spent in the hospital

So our estimate will have mean approximately equal to $2/\lambda$, a serious over-estimate of the true mean hospital stay, namely $1/\lambda$.

Time spent in the hospital

So our estimate will have mean approximately equal to $2/\lambda$, a serious over-estimate of the true mean hospital stay, namely $1/\lambda$.

The problem with our estimation strategy is that when we arrive at a bed at noon on July 13, we are much more likely to encounter a long term patient in the bed than a short term patient.

Time spent in the hospital

So our estimate will have mean approximately equal to $2/\lambda$, a serious over-estimate of the true mean hospital stay, namely $1/\lambda$.

The problem with our estimation strategy is that when we arrive at a bed at noon on July 13, we are much more likely to encounter a long term patient in the bed than a short term patient. This leads to consideration of weighted distributions.

Weighted distributions.

We’ll begin with a discrete example.

Weighted distributions.

We’ll begin with a discrete example.

A random variable is discrete if it has only a finite or a countably infinite number of possible values.

Weighted distributions.

We’ll begin with a discrete example.

A random variable is discrete if it has only a finite or a countably infinite number of possible values.

E.G, X=the number that shows on top when we roll a die.

E.G., Y=the number of tosses of a coin until a head appears

Weighted distributions.

Consider a random variable X with possible values {x(1),x(2),….} and associated probabilities {p(1),p(2),….}.

Weighted distributions.

Consider a random variable X with possible values {x(1),x(2),….} and associated probabilities {p(1),p(2),….}.

But now assume that if the random variable takes the value x(i), it is only observed with probability w(x(i)). Thus the probability of observing a realized value of X depends on the value assumed by X, according to a weight function w(.).

Weighted distributions

A simple, and often reasonable, choice for the weight function is w(x)=x. This corresponds to size-biasing.

Weighted distributions

A simple, and often reasonable, choice for the weight function is $w(x)=x$. This corresponds to size-biasing.

Big items are more likely to be observed than are small items. Remember the reindeer herds.

More notation

If X has discrete density $f_X(x)$ then the weighted version of X, denoted by X^w has density

$$f_{X^w}(x) = \frac{w(x) f_X(x)}{E(w(X))}$$

More notation

If X has discrete density $f_X(x)$ then the

weighted version of X, denoted by X^w

has density $$f_{X^w}(x) = \frac{w(x)f_X(x)}{E(w(X))}$$

We will use the same notation with essentially the same interpretation in the case in which X is a continuous variable.

- Of course, w(.) can be any non-negative function
- Subject to the requirement that E(w(X)) exists.
- As we mentioned, for non-negative X’s , popular choices for w(x) include:

$$w(x) = x$$

$$w(x) = x^k$$

$$w(x) = I(x > c)$$

If X can take on both positive and negative values, then popular choices for w(x) include:

$$w(x) = |x|$$

$$w(x) = |x|^k$$

$$w(x) = I(x > c)$$

What if we suspect the presence of size biasing.

- Suppose we have non-negative observations, supposed to have come from the density $f(x;\theta)$

- But we suspect size biasing has occurred i.e., instead our density might be

$$f^w(x;\theta) \propto x f(x;\theta)$$

What should we do?

What might we do ?

We could check which of the two models best fits the data using goodness of fit statistics.

Another possibility would be to consider the more general model

$$f(x, k, \theta) \propto x^k f(x; \theta)$$

and test the hypothesis $H : k = 0$

Back to the hospital beds with exponential occupancy times.

Back to the hospital beds with exponential occupancy times.

Here our unweighted density is

$$f(x;\lambda) = \lambda e^{-\lambda x} I(x > 0)$$

Back to the hospital beds with exponential occupancy times.

Here our unweighted density is

$$f(x;\lambda) = \lambda e^{-\lambda x} I(x > 0)$$

And the simple weighted, size biased density is

$$f^{w}(x) \propto x\lambda e^{-\lambda x} I(x > 0)$$

Back to the hospital beds with exponential occupancy times.

Here our unweighted density is

$$f(x;\lambda) = \lambda e^{-\lambda x} I(x > 0)$$

And the simple weighted, size biased density is

$$f^{w}(x) \propto x\lambda e^{-\lambda x} I(x > 0)$$

$$= \lambda^2 x e^{-\lambda x} I(x > 0)$$

Back to the hospital beds with exponential occupancy times.

Here our unweighted density is

$$f(x;\lambda) = \lambda e^{-\lambda x} I(x > 0)$$

And the simple weighted, size biased density is

$$f^w(x) \propto x\lambda e^{-\lambda x} I(x > 0)$$

$$= \lambda^2 x e^{-\lambda x} I(x > 0)$$

i.e.,

$$X^w \sim \Gamma(2, 1/\lambda)$$

Thus in this case we have

$$X^w =^d X_1 + X_2$$

where the X’s are i.i.d. exponential variables.

Exponentials in the hospital

This agrees with our earlier observation that, in this case, the data values that we collected look like sums of two independent exponential variables, instead of one.

Exponentials in the hospital

This agrees with our earlier observation that, in this case, the data values that we collected look like sums of two independent exponential variables, instead of one.

What would have been a better data collecting strategy ?

Better data

After we identify the 50 beds that we will observe, wait in each case until a new occupant is installed and observe how long the new occupant stays.

This will indeed give us i.i.d. data and will avoid size-bias.

Enough of size-biasing.

Biasing mechanisms can involve covariables and weight functions can also.

Let’s see an example.

We wish to study the weight distribution of applicants to become Riverside police officers.

Applicant distribution

A plausible model is that the weights of the applicants follow a normal distribution with mean μ and variance σ^2.

Applicant distribution

A plausible model is that the weights of the applicants follow a normal distribution with mean μ and variance σ^2 .

But we only have reliable weight data for the officers that have been hired.

Applicant distribution

A plausible model is that the weights of the applicants follow a normal distribution with mean μ and variance σ^2 .

But we only have reliable weight data for the officers that have been hired.

Is it reasonable to use this data to estimate μ and σ^2 ?

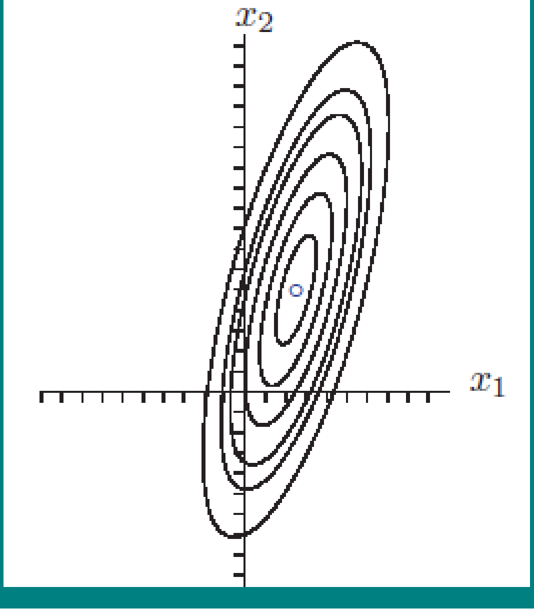
Weights and heights

- A plausible model for the two dimensional variable (height,weight) is a clasical bivariate normal model with normal marginal distributions and elliptical contours

Weights and heights

- A plausible model for the two dimensional variable (height,weight) is a clasical bivariate normal model with normal marginal distributions and elliptical contours

Like this:



Selection (discrimination !)

Selection (discrimination !)

If you are too tall, you won’t be hired.

Selection (discrimination !)

If you are too tall, you won’t be hired.

You wouldn’t fit in a police car.

Selection (discrimination !)

If you are too tall, you won’t be hired.

You wouldn’t fit in a police car.

Also if you are too short, you won’t be hired.

Selection (discrimination !)

If you are too tall, you won’t be hired.

You wouldn’t fit in a police car.

Also if you are too short, you won’t be hired.

You couldn’t reach the brake pedal.

We have “hidden truncation”.

We have “hidden truncation”.

Let’s, to simplify matters, just consider truncation from below.

We have “hidden truncation”.

Let’s, to simplify matters, just consider truncation from below.

i.e.; we observe X only if a covariable Y exceeds some threshold c.

We have “hidden truncation”.

Let’s, to simplify matters, just consider truncation from below.

i.e.; we observe X only if a covariable Y exceeds some threshold c.

The distribution of observed X’s will not be normal even if (X,Y) has a bivariate normal distribution.

We have “hidden truncation”.

Let’s, to simplify matters, just consider truncation from below.

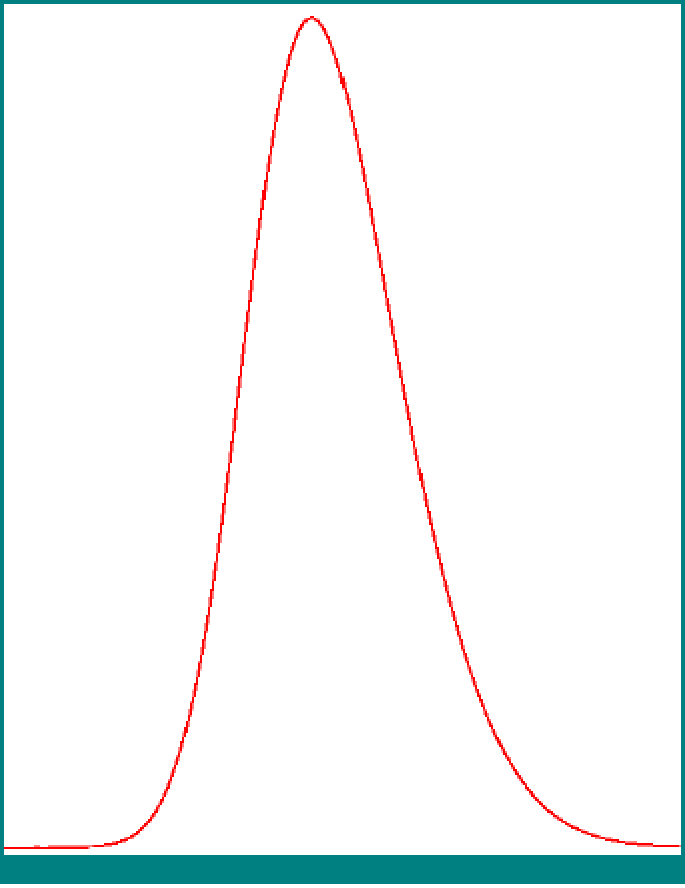
i.e.; we observe X only if a covariable Y exceeds some threshold c.

The distribution of observed X’s will not be normal even if (X,Y) has a bivariate normal distribution.

(Unless X and Y are independent.)

Hidden truncation

- The distribution of the observed X’s will be skewed. It will look something like:



Hidden truncation:theory.

Begin with the Azzalini (1985) skew-normal density:

$$f(x;\lambda) = 2\phi(x)\Phi(\lambda x), \quad -\infty < x < \infty,$$

Or the two-parameter extension

From Azzalini to hiddden truncation

We add location and scale parameters to get a 4-parameter model which represents the family of all possible densities arising as marginal densities of X given that X is the first coordinate of a correlated bivariate normal variable that is only observed if the second coordinate exceeds a particular value.

i.e., this is the hidden truncation model

Hidden truncation model

The density is thus:

$$f(x:\mu,\sigma,\lambda_0,\lambda_1)=\frac{\phi(\frac{x-\mu}{\sigma})\Phi(\lambda_0+\lambda_1(\frac{x-\mu}{\sigma}))}{\sigma\Phi(\frac{\lambda_0}{\sqrt{1+\lambda_1^2}})}$$

Hidden truncation model

The density is thus:

$$f(x:\mu,\sigma,\lambda_0,\lambda_1)=\frac{\phi(\frac{x-\mu}{\sigma})\Phi(\lambda_0+\lambda_1(\frac{x-\mu}{\sigma}))}{\sigma\Phi(\frac{\lambda_0}{\sqrt{1+\lambda_1^2}})}$$

Note that this includes the normal model as a special case when $\lambda_1 = 0$.

Note that the model

$$f(x:\mu,\sigma,\lambda_0,\lambda_1)=\frac{\phi(\frac{x-\mu}{\sigma})\Phi(\lambda_0+\lambda_1(\frac{x-\mu}{\sigma}))}{\sigma\Phi(\frac{\lambda_0}{\sqrt{1+\lambda_1^2}})}$$

is a weighted distribution, with a weight function which depends indirectly on the correlation and the truncation parameter.

Inference

Inference

To see whether hidden truncation is present we can test the hypothesis $H : \lambda_1 = 0$.

Inference

To see whether hidden truncation is present we can test the hypothesis $H : \lambda_1 = 0$.

The 4 parameters in the model can be estimated using maximum likelihood or the method of moments.

Inference

To see whether hidden truncation is present we can test the hypothesis $H : \lambda_1 = 0$.

The 4 parameters in the model can be estimated using maximum likelihood or the method of moments.

Warning: Easier said than done.

The likelihood surface may have no achievable maximum.

The likelihood surface may have no achievable maximum.

Reparameterization is often suggested.

The likelihood surface may have no achievable maximum.

Reparameterization is often suggested.

Careful choice of moment equations is required.

- Note, hidden truncation may occur without our knowledge of the source.

- Note, hidden truncation may occur without our knowledge of the source.
- However, the data can sometimes indicate its presence.

- Note, hidden truncation may occur without our knowledge of the source.

- However, the data can sometimes indicate its presence.

- It may be much more prevalent than we suspect.

Multivariate hidden truncation

A univariate model may begin with (X,Y) having a bivariate normal distribution but with X being observed only if $Y<c$.

Both X and Y can instead be multidimensional.

Multivariate hidden truncation

e.g., Students seeking admission to Prestigious University take four tests:

An English test

A Math test

An IQ test

A physical fitness test.

Multivariate hidden truncation

Students are admitted to Prestigious University on the basis of their scores on

The English test and the math test.

Multivariate hidden truncation

Students are admitted to Prestigious University on the basis of their scores on

The English test and the math test.

The distribution of IQ scores and physical fitness scores of the admitted students will involve hidden truncation.

Multivariate hidden truncation

For a hidden truncation model in higher dimensions. Begin with ($\underline{X}$,$\underline{Y}$) a random vector of dimension k+m and consider:

$$f_{\underline{X}|\underline{Y}>\underline{y}_0}(\underline{x}) = f_{\underline{X}}(\underline{x}) \frac{P(\underline{Y}>\underline{y}_0|\underline{X}=\underline{x})}{P(\underline{Y}>\underline{y}_0)}$$

If $(\underline{X},\underline{Y})$ has a k+m dimensional normal distribution this construction yields what is known as the closed skew normal distribution for $\underline{X}$.

Closed, since the model has marginals and conditionals of the same type.

If $(\underline{X},\underline{Y})$ has a k+m dimensional normal distribution, i.e.

$$\begin{pmatrix} \underline{X} \\ \underline{Y} \end{pmatrix} \sim N^{(k+m)}\left(\begin{pmatrix} \underline{\mu} \\ \underline{\nu} \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right)$$

Then

$$\underline{X} \sim N^{(k)}(\underline{\mu}, \Sigma_{11})$$

$$\underline{Y}|\underline{X} = \underline{x} \sim N^{(m)}(\underline{\nu} + \Sigma_{21}\Sigma_{11}^{-1}(\underline{x} - \underline{\mu}), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}).$$

A little algebraic bookkeeping yields the standard closed skew normal model in the form:

$$f_{\underline{y}_0}(\underline{z}) = \phi^{(k)}(\underline{z}) \frac{\Phi^{(m)}(\underline{\lambda}_0 + \Lambda \underline{z}; \underline{0}, \Delta)}{\Phi^{(m)}(\underline{\lambda}_0; \underline{0}, \Delta + \Lambda^T \Lambda)}$$

for suitably defined $\underline{\lambda}_0, \Delta$ and

Λ (which will depend on the choice of $\underline{y}_0$).

As can be imagined, estimation of the parameters in this model

$$\lambda_0, \Delta \text{ and } \Lambda$$

can be expected to be challenging.

And life will be more complicated when we introduce location parameters, etc.

So our message is: Be alert for the possibility that hidden selection mechanisms have been at work causing our target model to be inappropriate, and that a weighted version will better fit the data. Two particularly common cases involve hidden truncation and size biasing; but there are many others that might be encountered.

Look closely at the data !!!

Remember the reindeer and the policemen.

Some references

Arellano-Valle, R.B., Branco, M. and Genton, M. (2006). A unified view of skewed distributions arising from selections. Canadian Journal of Statistics, 34, 581-601.

Arnold, B.C. and Beaver, R.J. (2000). Hidden truncation models. Sankhya, series A, 62(1), 22-35.

Arnold, B.C. and Beaver, R.J. (2002). Skewed multivariate models related to hidden truncation and/or selective reporting. TEST, 11(1), 7-54.

Bluementhal, S. (1967). Proportional sampling in life length studies. Technometrics, 9, 205-218.

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

Gupta, R.C. and Kirmani, S.N.U.A. (1990). The role of weighted distributions in stochastic modeling. Communications in Statistics, Theory and Methods, 19(9), 3147-3162.

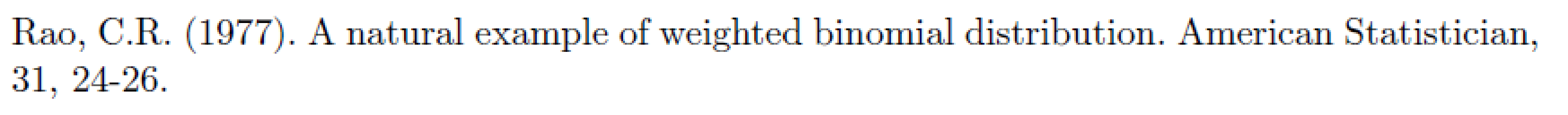
Patil, G.P. and Rao, C.R. (1976). On size biased sampling and related form -invariant weighted distributions. Sankhya, series B, 38, 48-61.

Rao, C.R. (1965).On discrete distributions arising out of methods of ascertainment: In Classical and Contagious Discrete Distributions. G.P.Patil, Ed., Pergamom Press and Statistical Publishing Society, Calcutta, 320-332.

Two more

Two more

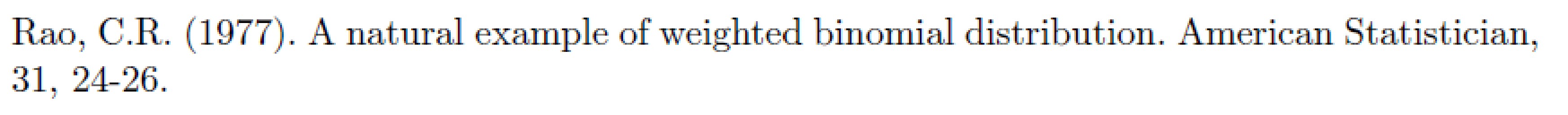
One which includes the male/female ratio example.



Rao, C.R. (1977). A natural example of weighted binomial distribution. American Statistician, 31, 24-26.

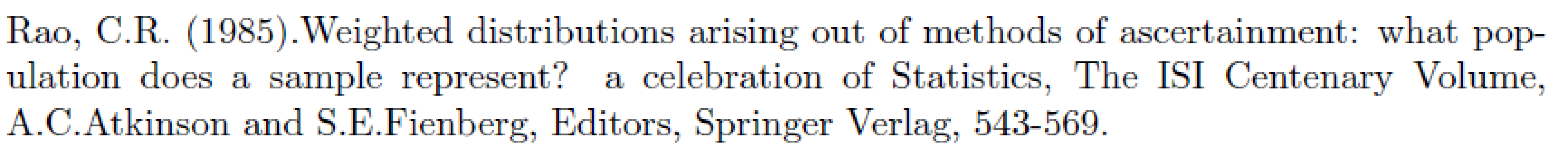
Two more

One which includes the male/female ratio example.



Rao, C.R. (1977). A natural example of weighted binomial distribution. American Statistician, 31, 24-26.

And one from which I stole my title !!



Rao, C.R. (1985).Weighted distributions arising out of methods of ascertainment: what population does a sample represent? a celebration of Statistics, The ISI Centenary Volume, A.C.Atkinson and S.E.Fienberg, Editors, Springer Verlag, 543-569.

Thank you for your attention.