Ranked Set Sampling

Christof Seiler

Stanford University, Spring 2016, Stats 205

Introduction

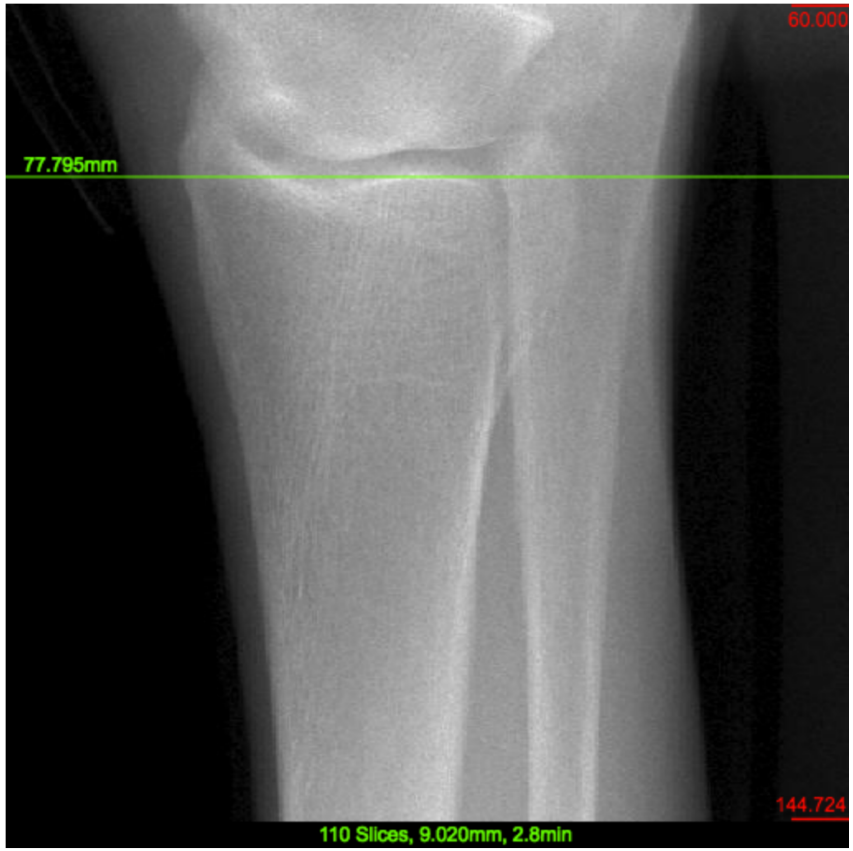
- Keys to any statistical inference most commonly is random sample from the population in question
- Goal is to get the best possible representation of the population for the quantity of interest
- How to collect sample data?
- Are there alternative to a random sample?
- Usually, once the sample items have been chosen, the desired measurements are collected from each of the selected items
- **Ranked set sampling**: It is not a sampling technique; it is a data measurement techniques

Introduction

- First introduced by McIntyre in 1952 for situations where
  - actual measurements for one sample observations is expensive
  - ranking observations is cheap
  - collecting sample units is cheap and reliable
- Example: Bone Mineral Density (BMD) in a human population
  - Subjects for such a study are plentiful
  - Measurements of BMD via Dual X-ray Absorptiometry (DXA) on selected subject is expensive
  - Because we need medical experts (e.g. orthopedic surgeons or anatomists) to manually segment images
  - Thus it is important to minimize the number of measurements required for the study without sacrificing information about the BMD makeup of the population

Introduction

- Radiograph of the radius bone



Source: `http://webapps.radiology.ucsf.edu/refline/`

Introduction

- Consider random sample $X_1, \ldots, X_n$
- Additionally to simple random sampling, there is stratified sampling, proportional sampling, and concomitant variable based sampling
- These are different kinds of sampling techniques defined **prior** to data collection
- Ranked set sampling is a techniques **during** data collection
- It helps to direct attention toward measurements of more representative units
- As a result, we get measurements that are more likely to span the range of values in the population

Collecting a Ranked Set Sample

- To obtain k observations from a population
- First, use simple random sampling to select k units from a population
- Then, rank order them according to a predefined attribute
- A variety of mechanisms are:
  - Visual comparison
  - Expert opinion
  - Auxiliary variable
  - Cannot involve your quantity of interest!
- The unit that is judged the smallest is included in your ranked set sample
- This first unit is called the **first judgement order statistics** and denoted by $X_{[1]}$
- Square brackets are used because this may not coincide with true order statistics usually denoted by $X_{(1)}$

Collecting a Ranked Set Sample

- The remaining $k-1$ units are not considered further (their role was only to assist selecting the smallest ranked unit)
- Then we repeat the same procedure to select the second $X_{[2]}$ by picking the second smallest judgement order statistic
- And continue this procedure until $X_{[k]}$
- One such run sequence $X_{[1]}, \ldots, X_{[k]}$ is called a **cycle**
- This is called a **balanced** ranked set sample since we collect one unit for each rank
- To obtain $n = k \times m$ observations we repeat the entire process

Cycle 1	$X_{[1]1}$	$\ldots$	$X_{[k]1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
Cycle m	$X_{[1]m}$	$\ldots$	$X_{[k]m}$

Collecting a Ranked Set Sample (Example)

- Unburned hydrocarbons emitted from automobile tailpipes and via evaporation are among the primary contributions to ozone and smog in large cities
- One way to reduce pollution is to use reformulated gasoline, designed to reduce volatility measured through Reid Vapor Pressure (RVP) value
- To enforce this, regular gasoline samples are taken from pumps at stations and RVP is measured
- There is crude field measure and a more refined lab measure
- The goal is to use the cruder and cheaper measure as a surrogate for the more expansive lab measure in order to reduce the actual amount of lab measures

Collecting a Ranked Set Sample (Example)

Sample number	Field RVP value	Sample number	Field RVP value
1	7.60	19	7.85
2	9.25	20	7.86
3	7.73	21	7.92
4	7.88	22	7.95
5	8.89	23	7.85
6	8.88	24	7.95
7	9.14	25	7.98
8	9.15	26	7.80
9	8.25	27	7.80
10	8.98	28	8.01
11	8.63	29	7.96
12	8.62	30	7.86
13	7.90	31	8.89
14	8.01	32	7.89
15	8.28	33	7.73
16	8.25	34	9.21
17	8.17	35	8.01
18	10.72	36	8.32

Source: Hollander and Wolfe, and Chicken (2013)

Collecting a Ranked Set Sample (Example)

- Group of triples $k = 3$ in four cycles $m = 4$
- Random triples from all 36 original observations

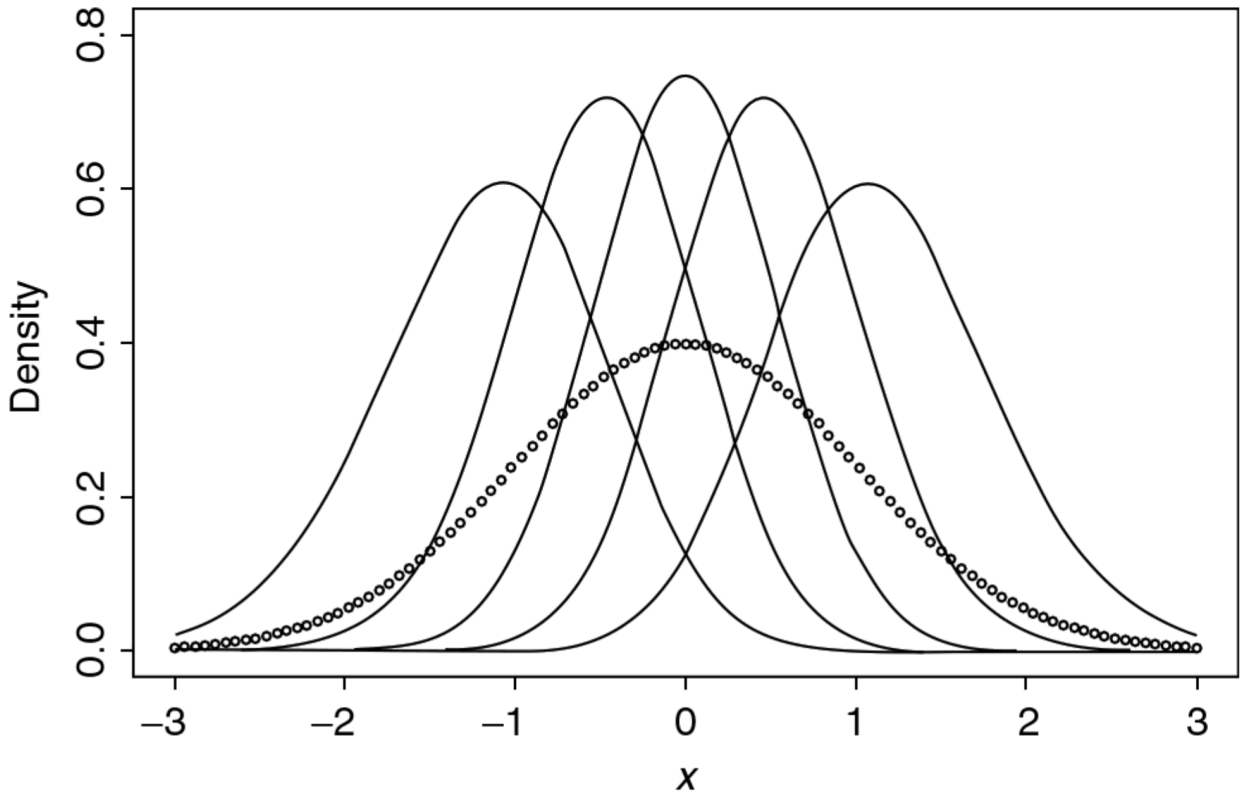
8.98	7.90	7.85	8.63	8.62	8.17	8.25	9.25	7.92
8.28	10.72	8.32	9.21	7.80	8.89	7.95	8.89	8.01
7.95	8.01	7.85	7.86	8.25	7.88	8.01	7.89	9.15
7.73	7.80	7.96	9.14	7.60	8.88	7.73	7.86	7.98

- Circled values are the ranked set sample (these are now send to the lab for further analysis)

8.98	7.90	(7.85)	8.63	(8.62)	8.17	8.25	(9.25)	7.92
(8.28)	10.72	8.32	9.21	7.80	(8.89)	7.95	(8.89)	8.01
7.95	8.01	(7.85)	7.86	8.25	(7.88)	8.01	7.89	(9.15)
(7.73)	7.80	7.96	9.14	7.60	(8.88)	7.73	7.86	(7.98)

Collecting a Ranked Set Sample

- Marginal densities of the order statistics $X_{(1)}, \ldots, X_{(5)}$ for random sample of size 5 from the standard normal



Source: Hollander and Wolfe, and Chicken (2013)

Ranked Set Sample Estimation of Population Mean

- Two sets of n observations each from a population
- One set $X_1, \ldots, X_n$ is collected using simple random sampling
- The second using balanced ranked set sampling with set size k and cycles m, $n = k \times m$
- Assumptions:
  - Population is continuous with distribution F and density f, finite mean μ, and finite variance σ^2
  - All $2n$ are mutually independent
- Ranked Set Sample (RSS) mean

$$\mu_{\text{RSS}} = \overline{X}_{\text{RSS}} = \sum_{j=1}^{m} \sum_{i=1}^{k} \frac{X_{[i]j}}{km}$$

Ranked Set Sample Estimation of Population Mean

- The μ_{RSS} is an unbiased estimator for the population mean μ regardless of ranking quality
- This is valid in general, but we'll show it under assumption of perfect rankings
- Meaning that RSS rankings are order statistics: $X_{[i]} = X_{(i)}$
- Consider only one cycle $m = 1$
- Under perfect rankings assumptions

$$\mathrm{E}(\widehat{\mu}_{\text{RSS}}) = \mathrm{E}(\overline{X}_{\text{RSS}}) = \frac{1}{k}\sum_{i=1}^{k}\mathrm{E}(X^*_{(i)})$$

Ranked Set Sample Estimation of Population Mean

- Since $\mathrm{E}(X^*_{(i)})$ is distributed like the ith order statistics for random sample of size k from a continuous distribution $F(x)$ and density $f(x)$ under perfect ranking (for $i = 1, \ldots, k$):

$$\mathrm{E}(X^*_{(i)}) = \int_{-\infty}^{\infty} x \, \frac{k!}{(i-1)!(k-i)!} F(x)^{i-1}(1-F(x))^{k-i} f(x) dx$$

- This can be written as (with $q = i - 1$)

$$\mathrm{E}(\overline{X}_{\mathrm{RSS}}) = \int_{-\infty}^{\infty} x \, f(x) \left\{ \sum_{q=0}^{k-1} \binom{k-1}{q} F(x)^q (1-F(x))^{(k-i)-q} \right\} dx$$

- The expression $\{\cdot\}$ is sum over the entire sample space of the probabilities for a binomial random variable with parameter $k-1$ and $p = F(x)$, thus $\{\cdot\} = 1$

Ranked Set Sample Estimation of Population Mean

- Therefore $\widehat{\mu}_{\mathrm{RSS}}$ is an unbiased estimator for μ

$$\mathrm{E}(\widehat{\mu}_{\mathrm{RSS}}) = \mathrm{E}(\overline{X}_{\mathrm{RSS}}) = \int_{-\infty}^{\infty} x\, f(x) dx = \mu$$

- To obtain the variance of $\widehat{\mu}_{\mathrm{RSS}}$, mutual independence allows us to write

$$\mathrm{Var}(\overline{X}_{\mathrm{RSS}}) = \frac{1}{k^2} \sum_{i=1}^{k} \mathrm{Var}(X^*_{(i)})$$

- Recall bias and variance decomposition

$$\mathrm{E}((X^*_{(i)} - \mu)^2) = (\mathrm{E}(X^*_{(i)}) - \mu))^2 + \mathrm{Var}(X^*_{(i)})$$

Ranked Set Sample Estimation of Population Mean

- Combine

$$\mathrm{Var}(\overline{X}_{\mathrm{RSS}}) = \frac{1}{k^2}\sum_{i=1}^{k}\mathrm{E}((X^*_{(i)} - \mu)^2) - \frac{1}{k^2}\sum_{i=1}^{k}(\mathrm{E}(X^*_{(i)}) - \mu)^2$$

- We can proceed like with the expectation and use the "binomial distribution trick", which results in the first term being $k\sigma^2$

$$\mathrm{Var}(\overline{X}_{\mathrm{RSS}}) = \frac{\sigma^2}{k} - \frac{1}{k^2}\sum_{i=1}^{k}(\mathrm{E}(X^*_{(i)}) - \mu)^2$$

Ranked Set Sample Estimation of Population Mean

- Comparison of Simple Random Sampling (SRS) and Ranked Set Sampling (RSS) estimators
- For SRS, we have $\mathrm{E}(\mu_{\mathrm{SRS}}) = \mathrm{E}(\overline{X}) = \mu$ and $\mathrm{Var}(\mu_{\mathrm{SRS}}) = \frac{\sigma^2}{k}$
- Thus both SRS and RSS are unbiased estimators
- And the variance

$$\mathrm{Var}(\overline{X}_{\mathrm{RSS}}) = \frac{\sigma^2}{k} - \frac{1}{k^2}\sum_{i=1}^{k}(\mathrm{E}(X^*_{(i)}) - \mu)^2$$

$$= \mathrm{Var}(\overline{X}) - \frac{1}{k^2}\sum_{i=1}^{k}(\mathrm{E}(X^*_{(i)}) - \mu)^2 \leq \mathrm{Var}(\overline{X})$$

since

$$\frac{1}{k^2}\sum_{i=1}^{k}(\mathrm{E}(X^*_{(i)}) - \mu)^2 \geq 0$$

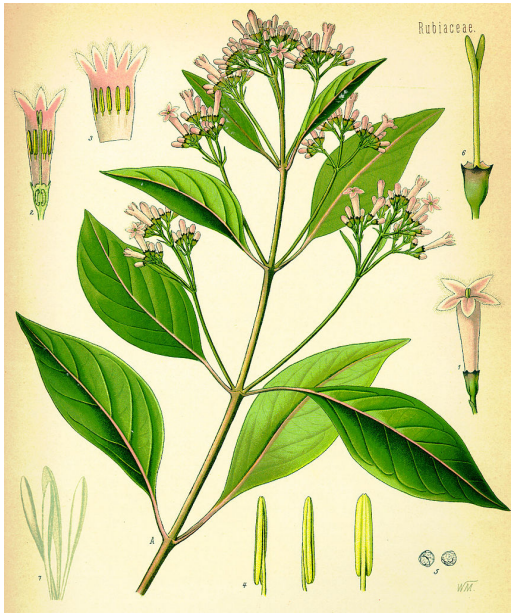
Ranked Set Sample Estimation of Population Mean

- So since
$$\mathrm{Var}(\overline{X}_{\mathrm{RSS}}) \leq \mathrm{Var}(\overline{X})$$
- In the case of perfect rankings, not only is $\overline{X}_{\mathrm{RSS}}$ an unbiased estimator but its variance is also never larger than the variance of the SRS estimator $\overline{X}$ based on the same number of measured observations
- In fact, this is a strict inequality unless
$$\mathrm{E}(X^*_{(i)}) = \mu$$
- Which is the case only if the rankings are purely random

Ranked Set Sample Estimation of Population Mean

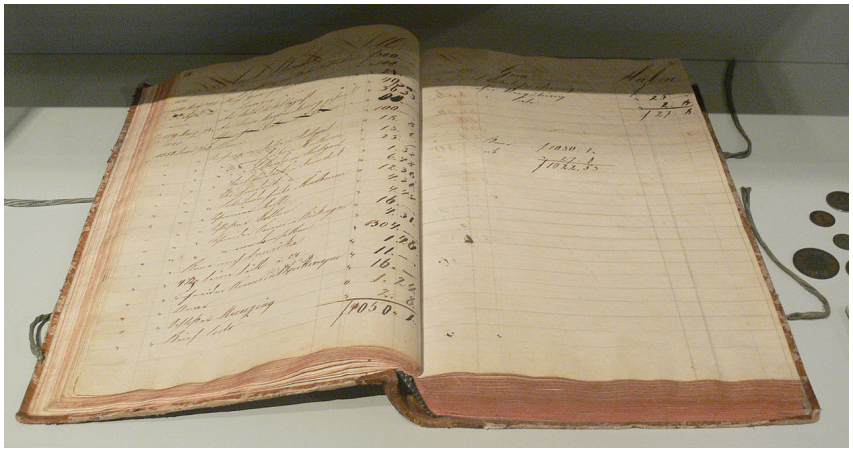
- The k components of the **SRS** estimator are **mutually independent and identically distributed** and each is itself an unbiased estimator
- While the k components of the **RSS** estimator are also mutually independent, they are **not identically distributed** and **none** of them (except for the middle order statistic when k is odd and the underlying distribution is symmetric about μ) are **individually unbiased**
- Yet the **averaging** process leaves $\widehat{\mu}_{\mathsf{RSS}}$ **unbiased**
- Interestingly, it is the additional structure associated with the nonidentical nature of the distributions for the terms in $\widehat{\mu}_{\mathsf{RSS}}$ that leads to the improvement in precision for $\widehat{\mu}_{\mathsf{RSS}}$ relative to $\widehat{\mu}_{\mathsf{SRS}}$

Quinine Content in Cinchona Plants



- Primary source of quinine for use in treatment of malaria
- One source of these plants are in steep hills of southern India
- Indian government wanted to estimate dry bark and quinine content of these plants
- Full measurement requires uprooting plant, stripping the bark, and drying
- Luckily, easy measurements such as volume, height, etc. of plant correlate with quinine yield

Auditing to Detect Fraud



- Assessing the true value of an account can be expansive
- Goal is to check whether sales invoices are fraudulent (over the value that an auditor would assign)
- Instead of auditing all sales invoices, we can use the book value (readily available in the company's electronic ledgers) as the auxiliary ranking variable to select a subset
- Then the audit is done only on this subset of invoices
- To estimate the percentage of fraudulent sales or total amount of fraud

Other Important Issues

- Set size k
- Imperfect rankings
- Unbalanced ranked set sampling
- Cost considerations

References

- Nahhas, Wolfe, and Chen (2002). Ranked Set Sampling: Cost and Optimal Set Size
- Wolfe (2004). Ranked Set Sampling: An Approach to More Efficient Data Collection
- Hollander, Wolfe, and Chicken (2013). Nonparametric Statistical Methods