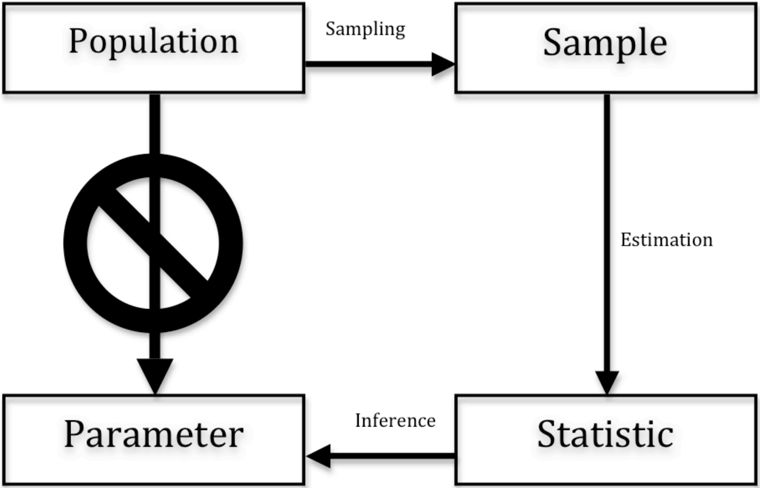
PLSC 502 – Autumn 2024 Randomization, Sampling, and Sampling Distributions

October 14, 2024

What We Do



Population

Sampling

Sample

Estimation

Statistic

Inference

Parameter

Some Terminology

- **Population**: All of the units of analysis; there are $\mathfrak{N}$ units in the population.
- **Units of analysis**: The "things" that make up the population.
- **Sample**: A subgroup of units from some larger population.
- **Sampling frame**: The pool of units of analysis available to be sampled.
- **Primary sampling units**: The "things" being sampled.
- **Sample size**: The number of units sampled from the population. Denoted N.
- **Stratum** (plural: strata): A subgroup of the population sharing a common trait or traits.

Two Problems With Samples

Bias

- *Systematic* differences between the sample and the population.
- Usually due to the sampling (or research) design.

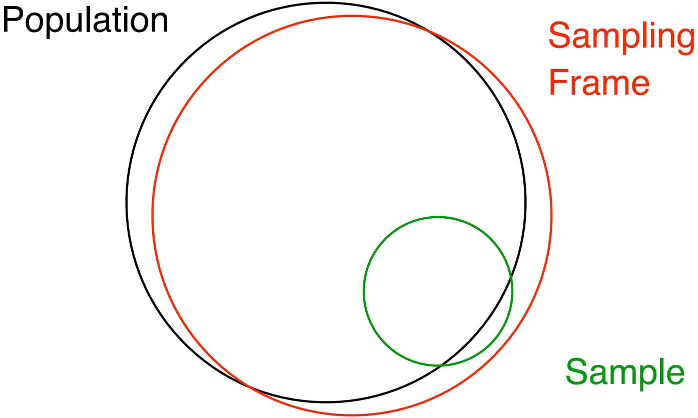
Sampling Error

- Differences between the sample and the population that are *nonsystematic*.
- Due to the randomness inherent to the sampling design.

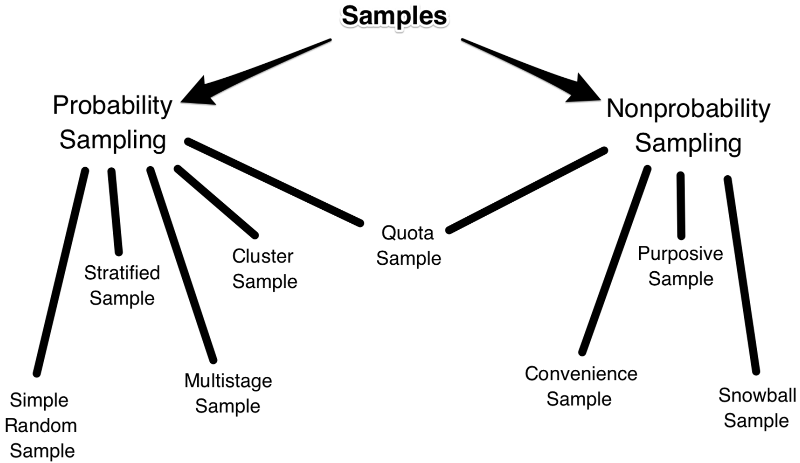
In general:

Bias is a much bigger problem than sampling error.

Population vs. Sampling Frame



Sample Types



Simple Random Sampling

Any sampling design where $\Pr(\text{Unit } i \text{ is sampled}) = \frac{1}{\mathfrak{N}} \ \forall\, i.$

or

Any sampling design where the probability of any given unit being selected into the sample is the same as any other unit in the population.

Simple Random Sampling: Pros and Cons

The Good:

- Mathematically easy to understand and implement
- Leads to the simplest / most straightforward methods of inference

The Bad:

- Difficult to define and draw; requires that you...
  - ...*know* every unit in the population, and
  - ...be able to *include* all selected units in the sample drawn.
- Can yield poor results for small subpopulations / strata.

Stratified Sampling

Steps:

1. Divide the sample into strata based on predefined characteristics.
2. Conduct simple random sampling within each stratum.

For two groups A and B with populations $\mathfrak{N}_A$ and $\mathfrak{N}_B$ $(\mathfrak{N}_A + \mathfrak{N}_B = \mathfrak{N})$ respectively:

- If $\Pr(\text{Unit } i_{A,B} \text{ is sampled}) = \frac{1}{\mathfrak{N}_{A,B}} \ \forall \ i_{A,B}$, then we have a *proportional stratified sample*.
- If (say) $\Pr(\text{Unit } i_A \text{ is sampled}) > \frac{1}{\mathfrak{N}_A}$, then we have *oversampled* from A (and *undersampled* from B).

Cluster Sampling

Steps:

1. Divide the sample into clusters based on predefined characteristics.
2. Draw a simple random sample of the clusters.
3. Include all units in each selected cluster in the final sample.

Cluster sampling:

- Changes the primary sampling unit from the unit of analysis to the cluster...
- Makes Pr(sample unit i) nonconstant / undefined
- *Most major media polls are done via cluster sampling*

Multistage Sampling

Steps:

1. Select a “cluster,” identify subclusters of units within the cluster, etc. until we get to the “lowest” level cluster.
2. Select – randomly or in a stratified way – some number of top-level clusters.
3. Within each selected cluster, select – again, randomly or stratifying – some number of subclusters.
4. Within subclusters, select sub-subclusters, etc.
5. At the “lowest” subcluster level, select some number of units from each sub-cluster.

Multistage Sampling

Example (from Agresti): sample survey respondents by first selecting blocks, then selecting houses within blocks, then selecting residents within each (selected) house.

- Blocks are *clusters*, houses are *subclusters*, and the individuals are the units finally sampled.
- Allows for probability samples without knowing identities of every unit sampled, via sampling rules (e.g., "select one person from among those in each house with equal probability.")
- *Most large, national surveys are conducted using multistage sampling.*

Nonprobability Samples

A sample where probability that every unit is in the sample is not (or cannot be) known.

Flavors:

- **Convenience Sampling**: What the name suggests.
- **Purposive Sampling**: The researcher selects units on the basis of whether s/he believes they ought to be in the sample.
- **Snowball Sampling**: Selects a unit, and then sample other units with some relationship to that first unit.

Quota Sampling

Researcher samples units within strata up to some quota, and then stops.

- E.g., a survey researcher might question 100 men and 100 women.
- Used a great deal in pre-WWII studies.
- Combined with (say) convenience sampling $\rightarrow$ nonprobability sample.
- Combined with probability (e.g., stratified) sampling $\rightarrow$ better.

Key Points

- Probability samples yield sampling error.
  - Smallest = (generally) simple random sampling
  - Stratified *can* be smaller
  - Multi-stage = complex...
- Nonprobability samples can lead to bias; also have (complex) sampling error.

The Margin of Error (MOE)

Sampling error is the (random) difference between the value you want to know in the population and its respective value in the sample.

Characteristics:

- Intuition: "Repeated samples"
- A function of:
  - The sample size,
  - The sampling design, and
  - The size of the population.

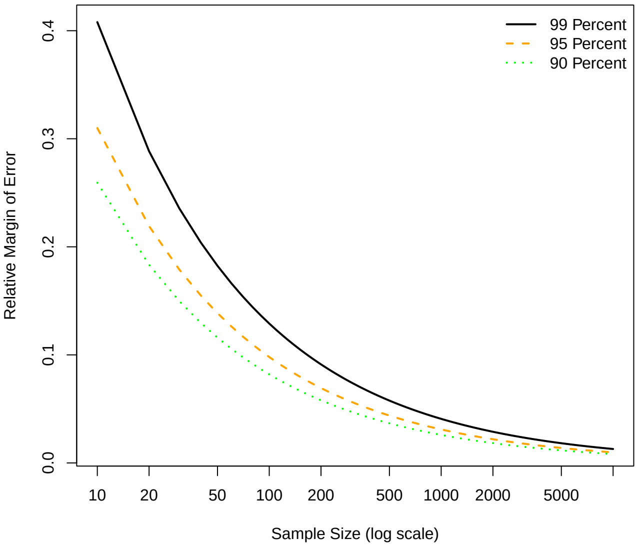
MOE Example

Consider the proportion P of observations in the population that have some (binary) trait. For a simple random sample of size N, the margin of error (sampling error) for the sample proportion p is:

$$\text{Standard error} = \sqrt{\frac{p(1-p)}{N}}$$

We typically calculate <u>relative</u> sampling error for a given *level of confidence*...

MOE and Sample Size



Sampling Distributions

Anything that is a function of random variables is, itself, a random variable.

Bitterville, pop. 1000

Bitterville voters:

$$\begin{aligned} \mathfrak{N}_D = 500 &\quad \rightarrow \quad X_i = 1 \\ \mathfrak{N}_R = 500 &\quad \rightarrow \quad X_i = 0 \end{aligned}$$

so that μ (the population mean) $= 0.5$.

For a sample with $N = 10$:

$$\begin{aligned} \bar{X} &= \frac{\sum_{i=1}^{10} X_i}{10} \\ &= \left(\frac{1}{10}\right) X_1 + \left(\frac{1}{10}\right) X_2 + \ldots + \left(\frac{1}{10}\right) X_{10} \\ &= aX_1 + aX_2 + \ldots + aX_{10} \end{aligned}$$

where $a = 0.1 \, \forall \, i$.

Bitterville, continued

Because

$$\mathsf{E}(aX + b) = a\mathsf{E}X + b,$$

then:

$$\begin{aligned}
\mathsf{E}(\bar{X}) &= \sum_{i=1}^{10} a\mathsf{E}(X_i) \\
&= \sum_{i=1}^{10} a\mu \\
&= \mu \sum_{i=1}^{10} a \\
&= \mu \sum_{i=1}^{10} \frac{1}{10} \\
&= \mu
\end{aligned}$$

The Variance of the Mean

Similarly:

$$\begin{aligned}
\mathrm{Var}(\bar{X}) &= \sum_{i=1}^{10} a^2 \mathrm{Var}(X_i) \\
&= \sum_{i=1}^{10} \left(\frac{1}{10}\right)^2 \sigma_i^2 \\
&= \left(\frac{1}{100}\right) \sum_{i=1}^{10} \sigma_i^2 \\
&= \left(\frac{1}{100}\right) 10\sigma^2 \\
&= \frac{\sigma^2}{10}
\end{aligned}$$

Means and Variances of $\bar{X}$, Generally

In general,

$$\mathsf{E}(\bar{X}) = \mu$$

and

$$\mathsf{Var}(\bar{X}) = \frac{\sigma^2}{N},$$

and so

$$\sqrt{\mathsf{Var}(\bar{X})} \equiv \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}.$$

A Rule of Thumb

Roughly speaking, under simple random sampling,

One must quadruple the sample size to halve the sampling error.

Example: For a binary X with $\bar{X} = 0.5$, we know that $\sigma^2 = 0.5(1 - 0.5) = 0.25$.

- When $N = 100 \rightarrow \sigma_{\bar{X}} = \frac{0.5}{\sqrt{100}} = 0.05$
- To get to $\sigma_{\bar{X}} = 0.025$, we'd need:

$$
\begin{aligned}
0.025 &= \frac{0.50}{\sqrt{N}} \\
0.025\sqrt{N} &= 0.50 \\
\sqrt{N} &= 20 \\
N &= 400.
\end{aligned}
$$

An Illustration

To illustrate, we'll:

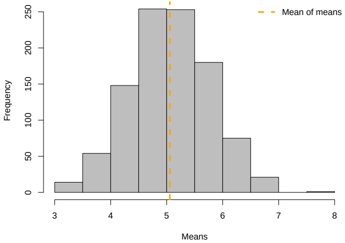
1. Generate a "population" of $\mathfrak{N} = 100000$ individuals,
2. ...where a variable of interest X has $\mu = 5$ and $\sigma = 5$; then
3. ...draw 1000 samples of size $N = 50$ from that population,
4. ...calculate the mean in each sample, and
5. ...examine the distribution of those sample means...

Note that, in this example, we should expect the sampling standard deviation of the mean $\sigma_{\bar{X}}$ to be:

$$\begin{aligned} \sigma_{\bar{X}} &= \frac{\sigma}{\sqrt{N}} \\ &= \frac{5}{\sqrt{50}} \\ &= \frac{5}{7.07} \\ &= 0.707 \end{aligned}$$

Illustration (continued)

```
> PopN<-100000          # Population is 100,000
> N <- 50               # Sample size N=50
> set.seed(7222009)
> Pop<-rnorm(PopN,5,5) # Population with mu=5 and sigma=5
>
> Nsims<-1000
> Means<-numeric(Nsims)
> set.seed(7222009)
> for(i in 1:Nsims){
+   Means[i]<-mean(sample(Pop,N,replace=FALSE))
+ }

> psych::describe(Means)
   vars    n mean   sd median trimmed  mad  min max range skew kurtosis   se
X1    1 1000 5.06 0.71   5.05    5.05 0.73 3.02 7.7  4.68 0.06    -0.14 0.02
```



Big Samples / Small Populations

Typically, sampling occurs *without replacement*...

- Sampling *with* replacement → $\text{Cov}(X_i, X_j) = 0 \, \forall \, i \neq j$.
- When we sample *without* replacement,

$$\text{Cov}(X_i, X_j) = -\frac{\sigma^2}{\mathfrak{N} - 1} \, \forall \, i \neq j$$

- Obviously, this number goes to 0 as $\mathfrak{N}$ gets very large...

In general, then, the “standard” formula for $\sigma_{\bar{X}}$ assumes $\mathfrak{N} >> N$...

N and $\mathfrak{N}$

Now suppose population $\mathfrak{N}$ is small; or, equivalently, the sample N is a large proportion of the population...

- Simple random sampling without replacement →
- (Relatively) high *negative* covariance among observations in the sample $\{X_1, X_2, ...X_N\}$
- → the "usual" estimate of $\sigma_{\bar{X}}$ will *overestimate* the variability of the sample mean $\bar{X}$.

The solution? The *Finite Population Correction*:

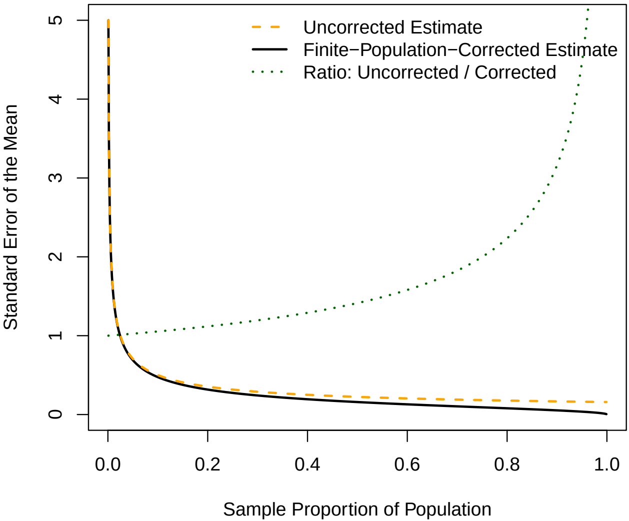
$$\begin{aligned}
\sigma_{\bar{X}} &= \frac{\sigma}{\sqrt{N}} \times \sqrt{\frac{\mathfrak{N} - N}{\mathfrak{N} - 1}} \\
&= \sqrt{\frac{\sigma^2}{N}\left(1 - \frac{N}{\mathfrak{N}}\right)}
\end{aligned}$$

FPC Example

For example, suppose we have $\mathfrak{N} = 1000$ and the standard deviation of X is $\sigma = 5$...

- "Uncorrected": $\sigma_{\bar{X}} = \frac{5}{\sqrt{N}}$
- "Corrected": $\sigma_{\bar{X}} = \frac{5}{\sqrt{N}} \times \sqrt{\frac{1000-N}{1000-1}}$
- Note that $0 \leq \sqrt{\frac{1000-N}{1000-1}} \leq 1$ for $N \in [1, 1000]$, so
- "Corrected" $\leq$ "Uncorrected"

FPC Example (continued)



Sampling *Distributions*: The Mean

For $X_i \sim$ i.i.d. $\mathcal{N}(\mu_i, \sigma_i^2)$,

$$\sum_{i=1}^{N} X_i \sim \mathcal{N}\left(\sum_N \mu_i, \sum_N \sigma_i^2\right)$$

which means that

$$\begin{aligned} \frac{1}{N}\sum_{i=1}^{N} X_i &\sim \mathcal{N}\left[\frac{1}{N}\sum_N \mu_i, \left(\frac{1}{N^2}\right)\sum_N \sigma_i^2\right] \\ &\sim \mathcal{N}\left(\mu, \frac{\sigma^2}{N}\right). \end{aligned}$$

Sampling Distribution of the Variance

The sample variance is:

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

which means that:

$$\begin{aligned}
\mathsf{E}(s^2) &= \frac{1}{N-1}\left\{\mathsf{E}\left[\sum_{i=1}^{N}(X_i - \bar{X})^2\right]\right\} \\
&= \frac{1}{N-1}\left\{\mathsf{E}\left[\sum_{i=1}^{N}(X_i - \mu)^2 - N(\bar{X} - \mu)^2\right]\right\} \\
&= \frac{1}{N-1}\left[\sum_{i=1}^{N}\mathsf{E}(X_i - \mu)^2 - N\mathsf{E}(\bar{X} - \mu)^2\right] \\
&= \frac{1}{N-1}\left(N\sigma^2 - N\frac{\sigma^2}{N}\right) \\
&= \sigma^2
\end{aligned}$$

Sampling Distribution: Variance

A transformation:

$$\mathfrak{s}^2 = \frac{(N-1)s^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

We can then show that:

$$\mathfrak{s}^2 \sim \chi^2_{N-1}$$

Variances Are Chi-Square...

For $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$, $\bar{X} = \frac{X_1+X_2}{2}$, and so:

$$s^2 = \frac{(X_1 - X_2)^2}{2}.$$

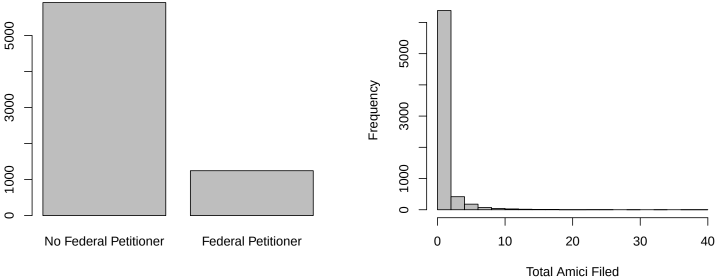
From this:

$$\begin{aligned} \mathfrak{s}^2 = \frac{(N-1)s^2}{\sigma^2} &= \frac{(X_1 - X_2)^2}{2\sigma^2} \\ &= \left(\frac{X_1 - X_2}{\sqrt{2\sigma^2}}\right)^2 \sim \chi_1^2 \end{aligned}$$

Data Example: The Warren & Burger Courts

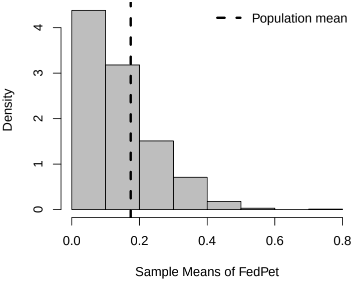
```
> psych::describe(WB)

        vars    n    mean      sd median trimmed  mad min  max range skew kurtosis    se
us*        1 7161 3036.64 1778.56   3008 3032.19 2274   1 6141  6140 0.02    -1.21 21.02
id         2 7161 3581.00 2067.35   3581 3581.00 2654   1 7161  7160 0.00    -1.20 24.43
amrev      3 7161    0.43    1.34      0    0.13    0   0   33    33 8.41   125.01  0.02
amaff      4 7161    0.41    1.30      0    0.11    0   0   37    37 7.59   117.94  0.02
sumam      5 7161    0.84    2.19      0    0.32    0   0   39    39 5.70    54.21  0.03
fedpet     6 7161    0.17    0.38      0    0.09    0   0    1     1 1.72     0.96  0.00
constit    7 7161    0.25    0.44      0    0.19    0   0    1     1 1.13    -0.72  0.01
sgam       8 7161    0.08    0.27      0    0.00    0   0    1     1 3.13     7.80  0.00
```

Frequencies for `fedpet` and `sumam`

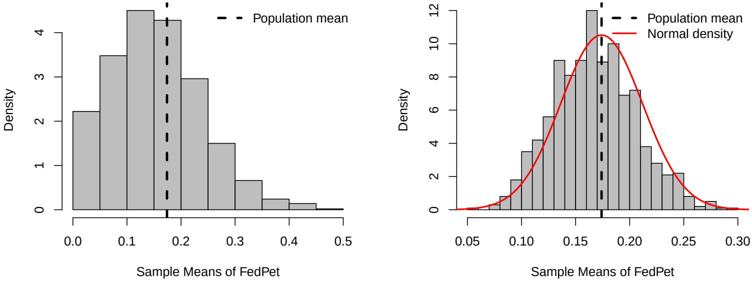


1000 Sample Means ($N = 10$)

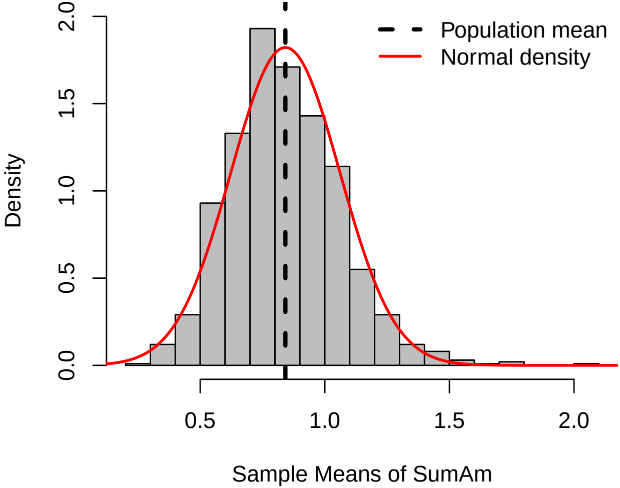
```
set.seed(7222009)
MFP10<-numeric(1000)
for (i in 1:1000){
  MFP10[i]<- with(WB, mean(sample(fedpet,10,replace=F)))
   }
```



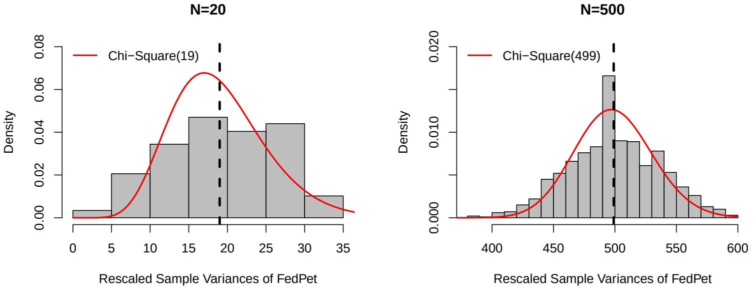
1000 Sample Means ($N = 20$ and 100)



Also For `sumam` ($N = 100$)



Sample Variances ($N = 20$ and 500)



Stratified Sampling using `sampling`

Constitutional decisions:

```
> table(WB$constit)

   0    1
5345 1816
```

Task: Draw a single stratified random sample ($N = 20$), with 10 observations from `constit=0` and 10 from `constit=1`.

Stratified Sampling using sampling

```
> library(sampling)         # package
> set.seed(7222009)         # set seed
> sample<-strata(WB,stratanames=c("constit"),
+                         size=c(10,10),method="srswor")
> sample.data<-getdata(WB,sample)
>
> summary(sample.data)

     us                  id           amrev            amaff            sumam            fedpet
 Length:20          Min.   : 122   Min.   : 0.0   Min.   : 0.00   Min.   : 0.00   Min.   :0.00
 Class :character   1st Qu.:1902   1st Qu.: 0.0   1st Qu.: 0.00   1st Qu.: 0.00   1st Qu.:0.00
 Mode  :character   Median :3776   Median : 0.0   Median : 0.00   Median : 0.00   Median :0.00
                    Mean   :3402   Mean   : 1.5   Mean   : 0.85   Mean   : 2.35   Mean   :0.05
                    3rd Qu.:4925   3rd Qu.: 0.0   3rd Qu.: 0.25   3rd Qu.: 1.00   3rd Qu.:0.00
                    Max.   :6178   Max.   :27.0   Max.   :12.00   Max.   :39.00   Max.   :1.00
     sgam         constit       ID_unit          Prob             Stratum
 Min.   :0.00   Min.   :0.0   Min.   : 121   Min.   :0.00187   Min.   :1.0
 1st Qu.:0.00   1st Qu.:0.0   1st Qu.:1901   1st Qu.:0.00187   1st Qu.:1.0
 Median :0.00   Median :0.5   Median :3775   Median :0.00369   Median :1.5
 Mean   :0.05   Mean   :0.5   Mean   :3401   Mean   :0.00369   Mean   :1.5
 3rd Qu.:0.00   3rd Qu.:1.0   3rd Qu.:4924   3rd Qu.:0.00551   3rd Qu.:2.0
 Max.   :1.00   Max.   :1.0   Max.   :6176   Max.   :0.00551   Max.   :2.0
```

Sampling Bias And “Big Data”

Thought experiment: If we wanted to estimate some population’s quantity from a sample of cases, are we (on-average) better off:

- Using a very small, simple random sample, or
- Using a much larger, but very slightly biased sample?

Meng (2018)

Suppose, in a population of size $\mathfrak{N}$, we are interested in learning the average of G (call it $\bar{G}_{\mathfrak{N}}$) using the mean $\bar{G}_N$ from a sample of size N.

Define:

- the "R-mechanism" R_J as the indicator / mechanism by which individuals are included in the sample,
- the sampling rate $f = \frac{N}{\mathfrak{N}}$, and
- the standard deviation of $G = \sigma_G$.

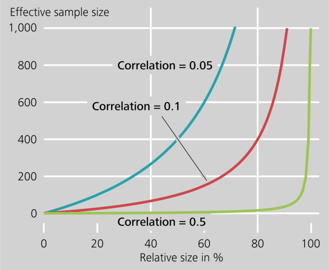
A key insight is that:

$$\bar{G}_N - \bar{G}_{\mathfrak{N}} = \underbrace{\rho(R, G)}_{\text{“Data Quality”}} \times \underbrace{\sqrt{\frac{1-f}{f}}}_{\text{“Data Quantity”}} \times \underbrace{\sigma_G}_{\text{“Problem Difficulty”}}$$

Meng (2018) (continued)

Some key points:

- Under simple random sampling, $E[\rho(R,G)] = 0$...
- The "Law of Large Populations": for $\rho(R,G) \neq 0$, estimation error increases in $\sqrt{N}$
- This means that when $\rho(R,G) \neq 0$, the "effective sample size" is dramatically reduced...



Back To The Warren/Burger Court Data

We know that, for all $\mathfrak{N} = 7161$ cases in the population, $\overline{\texttt{sumam}} = 0.84192$.

Let's compare:

- A small, simple random sample with $N = 179$ (that is, $f \equiv \frac{N}{\mathfrak{N}} = 0.025$), vs.
- A large sample with $N = 3580$ (that is, $f \equiv \frac{N}{\mathfrak{N}} = 0.50$) where Pr(sample inclusion) is a function of `sumam`

More specifically, Pr(sample inclusion) is such that:

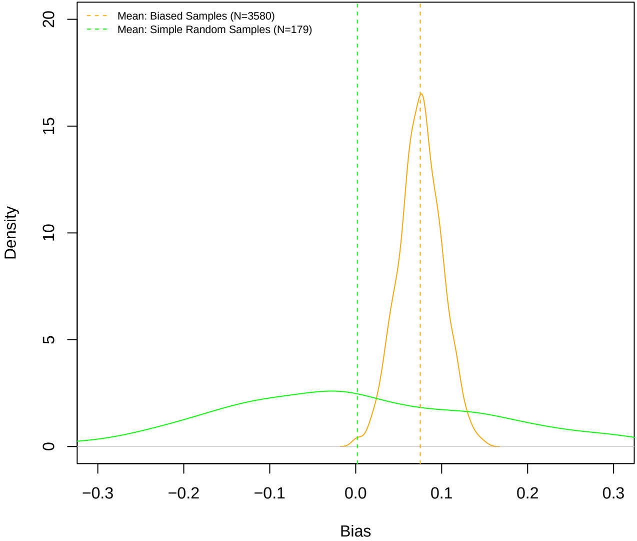
$$\frac{\Pr(\text{sample inclusion}|\text{maximum } \texttt{sumapp})}{\Pr(\text{sample inclusion}|\text{minimum } \texttt{sumapp})} = 2$$

What We're Doing

Process:

1. Draw a sample of N (either 179 or 3580) observations;
2. Calculate $\overline{\texttt{sumapp}}$ for that sample;
3. Calculate the *bias* in that sample, defined as $B = \overline{\texttt{sumapp}} - 0.84192$;
4. Repeat steps 1-3 many (say, 1000) times, and summarize the bias over those repetitions.

Big Data Won't Save You

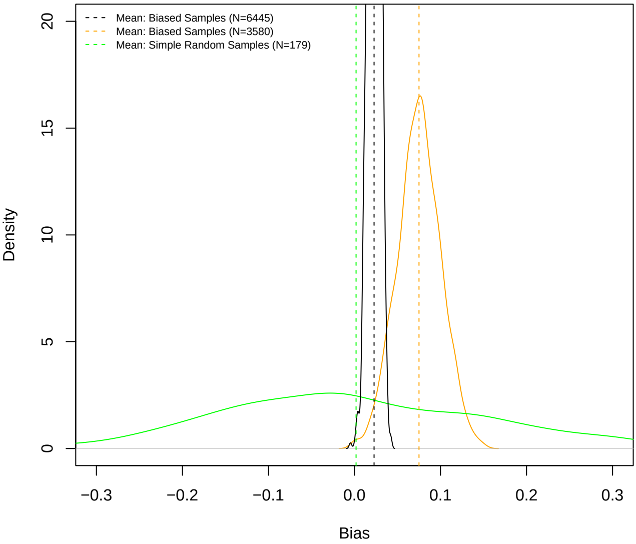


What We're Doing, Again

Process:

1. Draw a sample of N (either 179 or 3580) observations;
2. Calculate $\overline{\texttt{sumapp}}$ for that sample;
3. Calculate the *bias* in that sample, defined as $B = \overline{\texttt{sumapp}} - 0.84192$;
4. Repeat steps 1-3 many (say, 1000) times, and summarize the bias over those repetitions.
5. Repeat steps 1-4 for a 90-percent sample – that is, one with $N = 6445$.

Big Data Won't Save You, Part II


Mean: Biased Samples (N=6445)
Mean: Biased Samples (N=3580)
Mean: Simple Random Samples (N=179)
Density
Bias