BIOMED SCI 552:

STATISTICAL THINKING

LECTURE 5: SAMPLING

QUESTIONS FROM THURSDAY?

A NOTE ON THE NORMAL DISTRIBUTION

- σ: standard deviation
- σ^2: variance
- We use both of these functionally

SAMPLING SO FAR

- We've alluded to sampling for several lectures now
- In principle: We can't (usually) measure the population we're interested in, so we have to take a sample
- This is both critically important and non-trivial
- A bad sample is a hole you may not be able to dig yourself out of
  - And even if you can, it will likely be much harder than if you got a good sample in the first place

DATA GENERATING PROCESS

- …a process that generates data
- More helpfully – this is the process by which the real world "generates" the data you are interested in
- For the laboratory sciences, this is often quite direct
- For the population health sciences…
  - There's some underlying infection process. Some number of infected individuals experience symptoms, and then seek care. Some of those are tested, and some of those tests are reported…

SAMPLING PROCESS

- The sampling process is the part of the data generating process where we go from what exists (unknowably) in reality to a sample
- We take a sample, which has its own distribution, mean and variance
- As we discussed in an earlier lecture, there's inherently sampling error that means this sample's underlying distribution will be *different* from the true population distribution
  - And this is okay

SAMPLING WITH REPLACEMENT

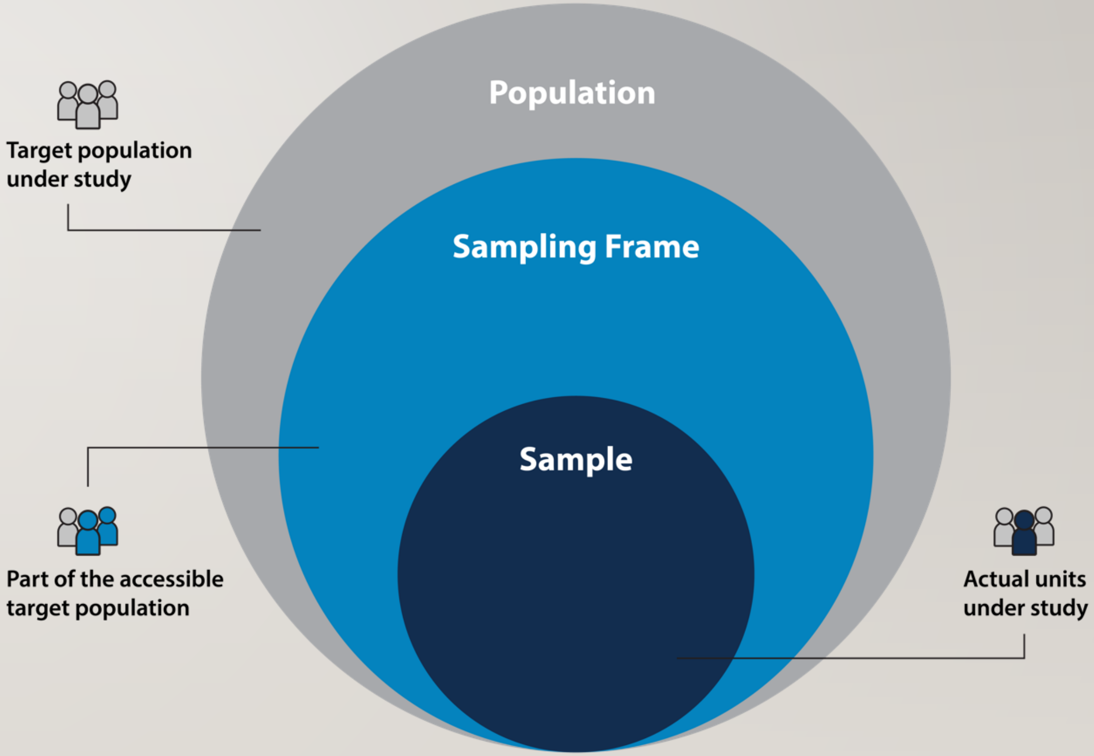
- There are some circumstances where we sample *with* replacement – we draw a sample from the population, and it is possible, with some probability, that we draw that sample again
- Capture-Recapture methods, for example, can be used to estimate population size by asking what population is most probable given we've captured the same bat *N* times
- There are some methods known as resampling methods that also use sampling with replacement, but they are beyond the scope of this class

SAMPLING WITHOUT REPLACEMENT

- Every time we draw a sample from the population, that sample is not eligible to be sampled again
- This does mean that every time you sample, your population decreases by 1
- Most of the time, you have a large enough population that this doesn't practically matter
    - In small populations, it potentially does if there are also consequences to being sampled
- Most of the time, this is the sampling we do in the biomedical sciences

SAMPLING FRAME

- This is the actual list of individuals who can be drawn to make your sample
- In a perfect world, this is everyone in your target population, and no one outside it
  - We do not live in a perfect world
  - Our sampling frame is itself a subset of the population, and may not be random
    - Hard to reach populations may not be in the sampling frame even if they are in the population
- A biased sampling frame, unsurprisingly, results in a biased sample

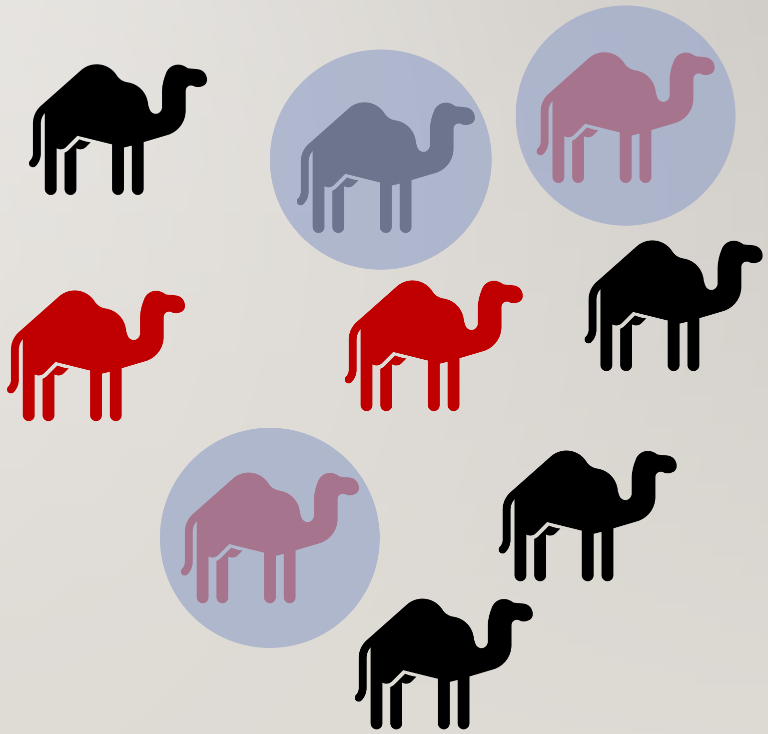


TYPES OF RANDOM SAMPLES

- Simple random sampling
- Systematic sampling
- Stratified sampling
- Cluster sampling

SIMPLE RANDOM SAMPLING

- The most basic, and potentially appealing type of sample
- Every member of the population has an equal probability of being included in the sample
- This is most easily done if you have a complete roster of the population in some form
  - A registry, patient records, a census, a population cohort, etc.



WHAT MIGHT BE SOME DRAWBACKS TO A SIMPLE RANDOM SAMPLE?

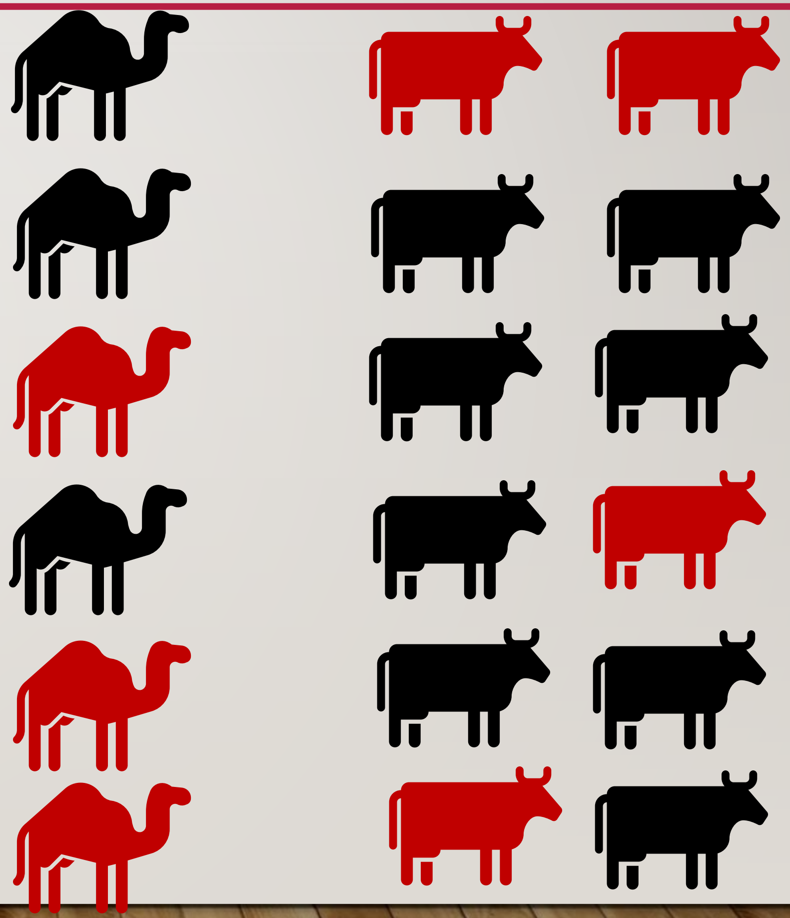
SYSTEMATIC SAMPLING

- Similar to a simple random sample
- The individuals in the sampling frame are arrayed in some order, and then every N^{th} element of that array is included in the sample
- This is often easier to implement, you control the sample size, etc.

WHAT ARE WE ASSUMING?

STRATIFIED SAMPLING

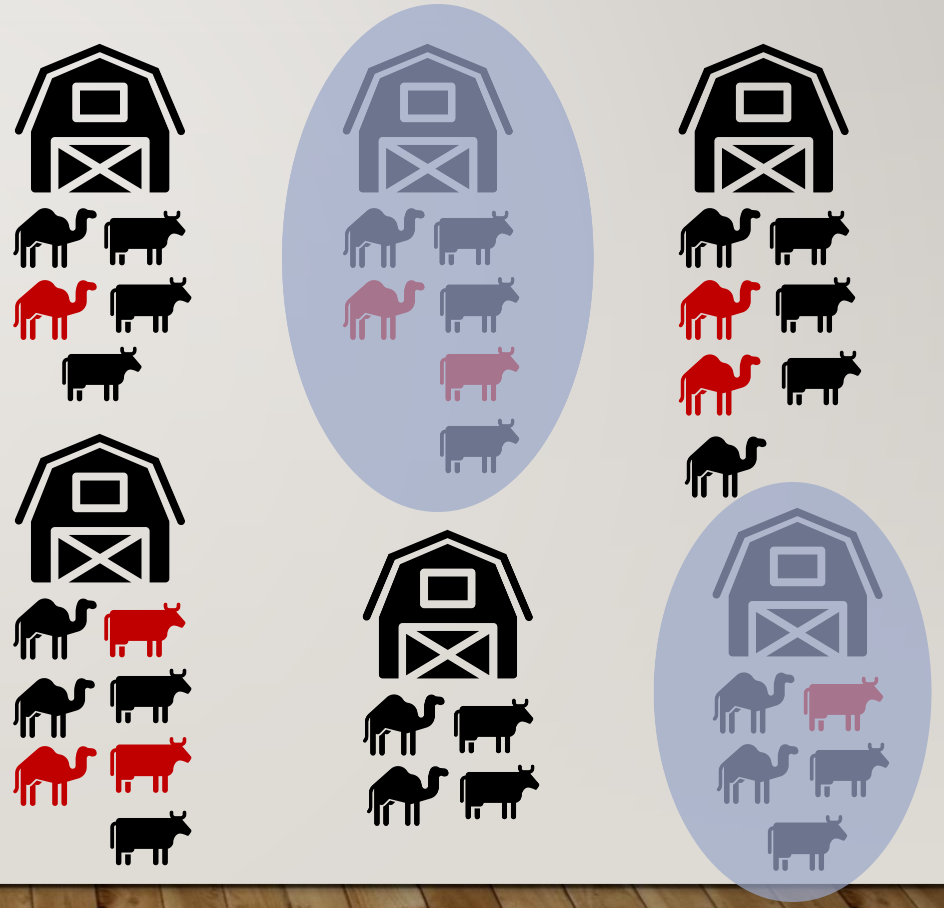
- Divide the population into some logical subgroups (called strata)
  - Age range, job role, species, phenotype, etc.
  - Take the proportion of the population in each strata, and sample that proportion of your total sample from the strata
    - i.e. there should be two cow samples for every one camel sample
    - *Within* the strata, sampling should be random
- Beneficial because it ensures a representative sample *among strata*



WHAT MIGHT BE SOME DRAWBACKS?

CLUSTER SAMPLING

- Your population is divided up into some logical grouping
  - Farms, classrooms, hospitals (or hospital units), etc.
- Which *groups* are chosen is then randomly selected
- This can be convenient, is often appropriate for non-independence, and can show group-level dynamics
- But your sample size just got *very* small



A NOTE ON CLUSTERING

- It has been my observation that lab-based researchers doing field sampling *love* adding clustering to this data
  - We're going to sample by household, out of a sample of villages, in selected districts, in particular seasons…
  - This is often out of necessity – sample collection periods are inherently pulsed, you can't pop back and forth between villages easily, etc.
    - Basically, this is *okay*
  - But…this can swiftly mean that the number of individuals in any given combination of strata can become very small
  - You should consult with a statistician beforehand to make sure you have adequately powered your study for the level of clustering you're about to add

RANDOM NUMBER GENERATION

- Usually, randomization is done by a computer these days
- How do random number generators work?
- What is a “seed” and why do I care?

```
int getRandomNumber()
{
    return 4;  // chosen by fair dice roll.
               // guaranteed to be random.
}
```

NONPROBABILISTIC SAMPLING

- All of the methods we've discussed have some sort of probabilistic aspect to them
- There are *nonprobabilistic* sampling methods that…unsurprisingly…don't involve randomization
- What's one example we've already talked about?

NONPROBABILISTIC SAMPLING

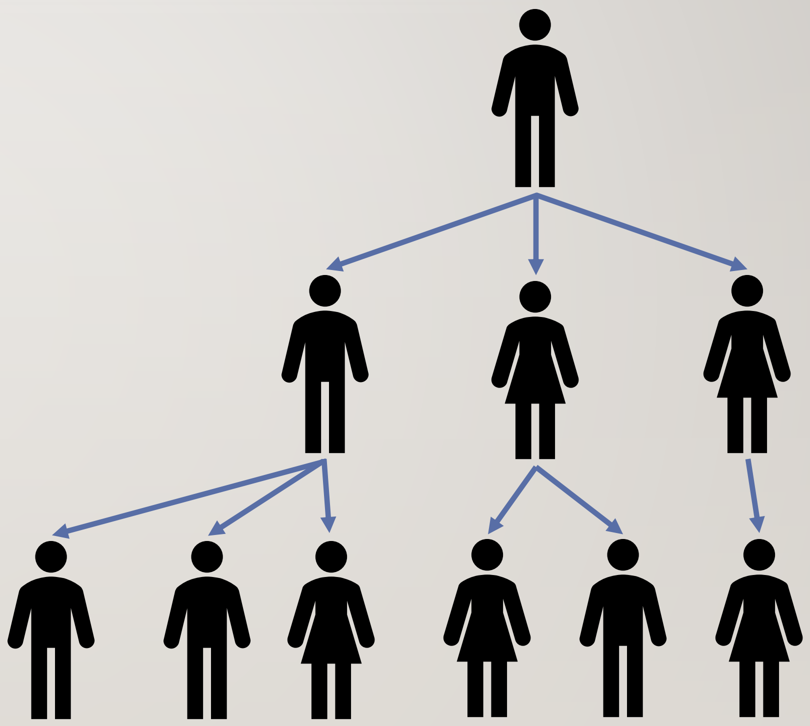
- All of the methods we've discussed have some sort of probabilistic aspect to them
- There are *nonprobabilistic* sampling methods that…unsurprisingly…don't involve randomization
- What's one example we've already talked about?
- Types
  - Convenience samples
  - Purposive samples
  - Snowball samples
  - Quota samples

PURPOSIVE SAMPLING

- Sometimes called "Judgement Sampling", involves the researcher using their expertise to select a sample
- This is used in qualitative research to get details on specific phenomena, etc.
- Makes statistical inference hard if not impossible
- It's important to be *very* clear about how these choices are being made
- There's a risk of observer bias

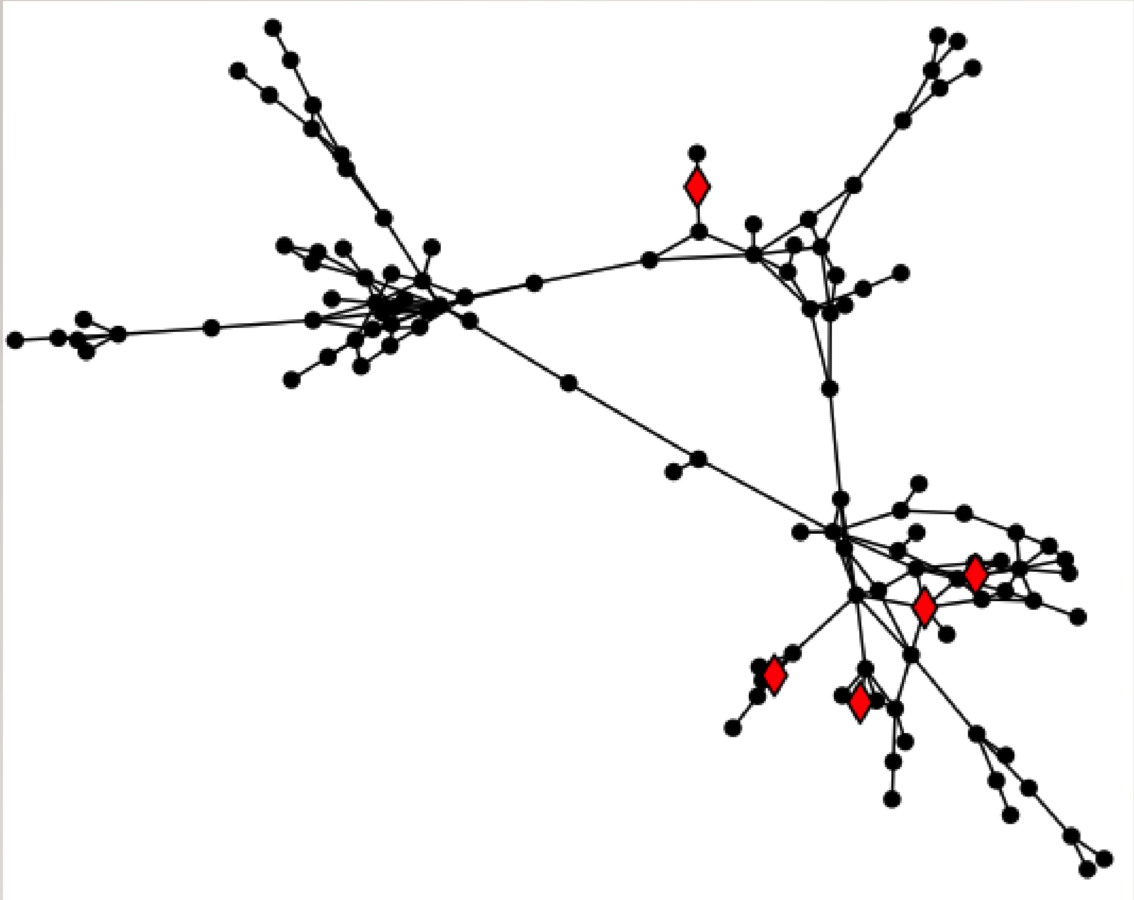
SNOWBALL SAMPLING

- Also known as "Respondent Driven Sampling"
- You recruit someone, they nominate one or more potential recruits, who in turn nominate more…
- Friends, sexual partners, etc.
- This is obviously a non-random sample
- Can be very useful for getting information from hard to reach groups

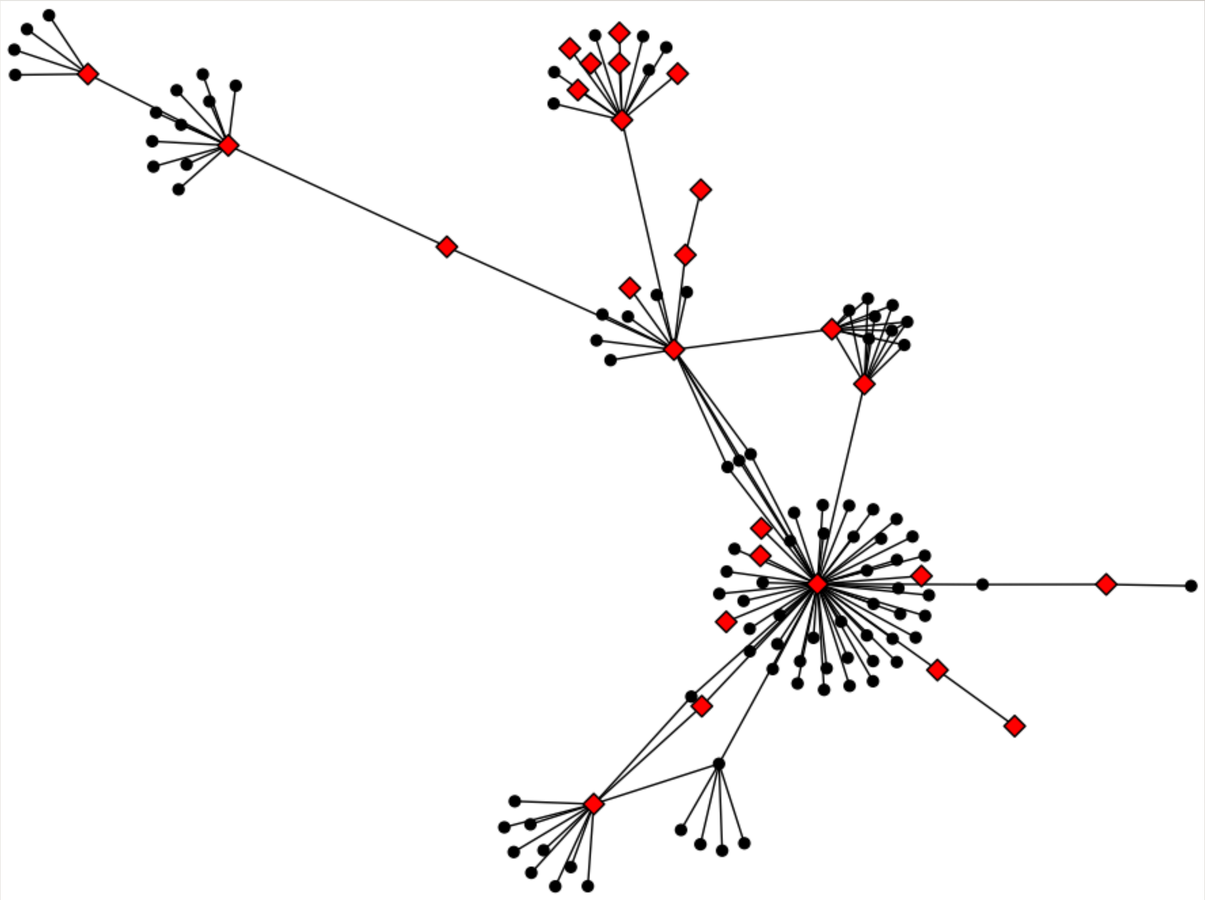


THE FRIENDSHIP PARADOX

- On average, your friends have more friends than you do
- Why?



Friendship Network



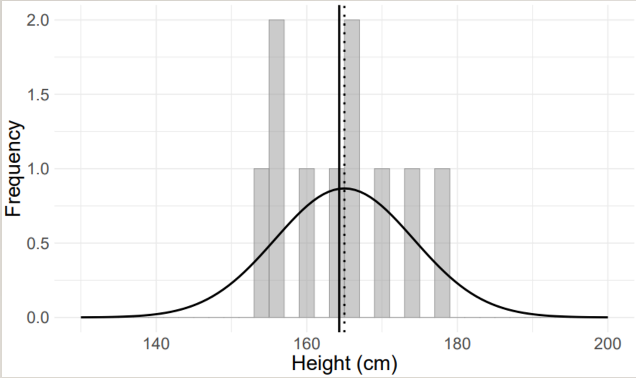
Sexual Contact Network

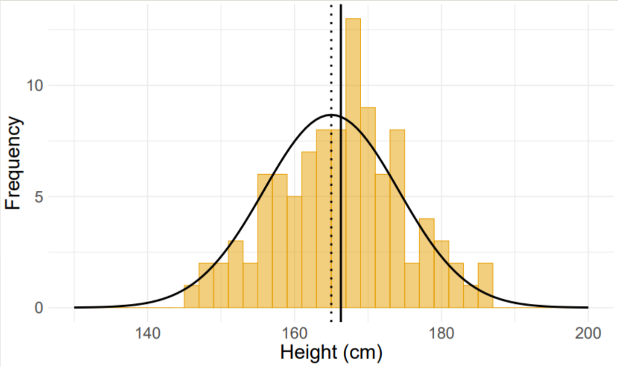
QUOTA SAMPLING

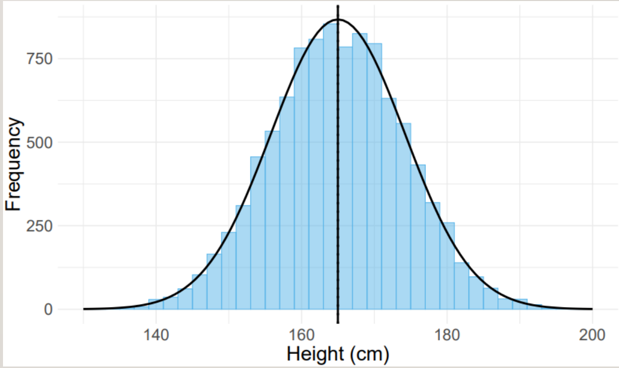
- Non-random selection of a predetermined number of individuals of a given type
    - Again, the population is divided into strata
    - A fixed number of people from each strata are then selected
- This ensures you get a broad swathe of your population, but a non-random one
- Again, this can be used heavily in qualitative research

SAMPLING FROM A NORMAL DISTRIBUTION

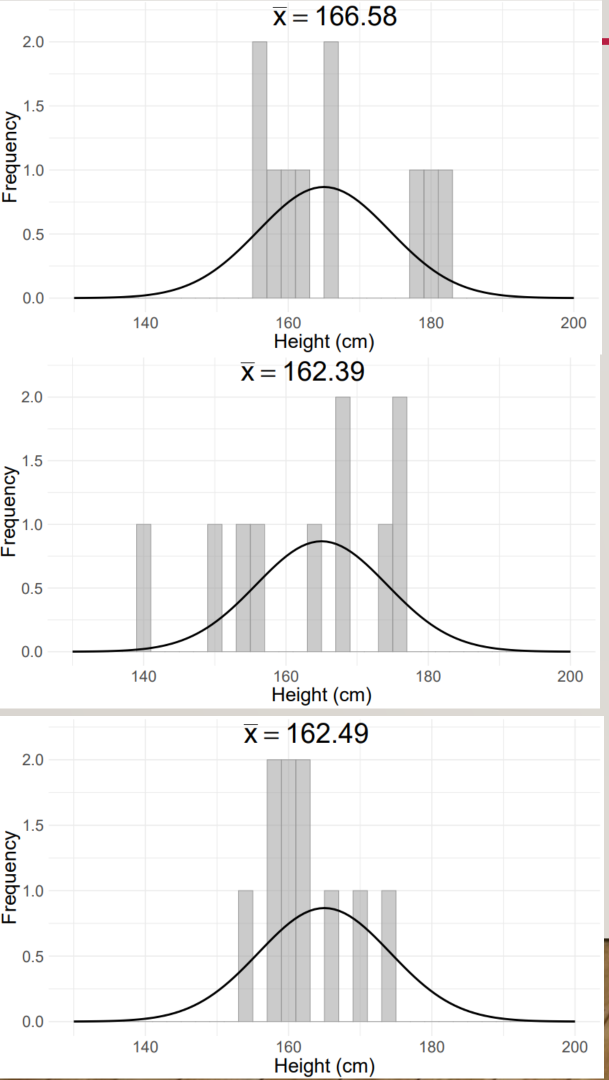
- The true population mean is $\boldsymbol{\mu}$ and the true population SD is $\boldsymbol{\sigma}$.
- Each time we sample a population, we get a different subset purely by chance with mean $\bar{\boldsymbol{x}}$ and SD $\boldsymbol{s}$.
- Larger sample sizes give us more certainty about the true population distribution.
- Note the true population distribution doesn't *change*





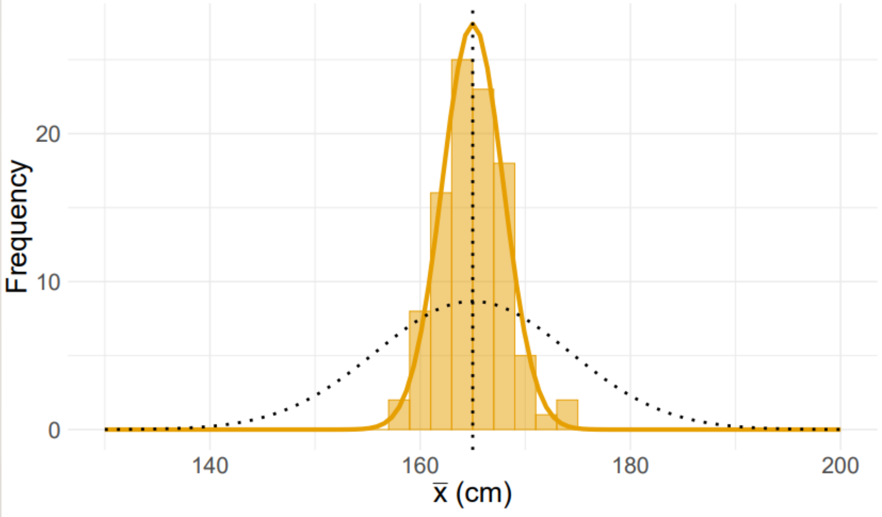


SAMPLING DISTRIBUTIONS



- If we repeatedly take a sample of 10 heights and calculate the mean of each sample, we generate a distribution of mean heights.
- A distribution of sample means is an example of a ***sampling distribution***.
- We use *sampling distributions* to quantify the uncertainty of estimates.

Distribution of Sample Means

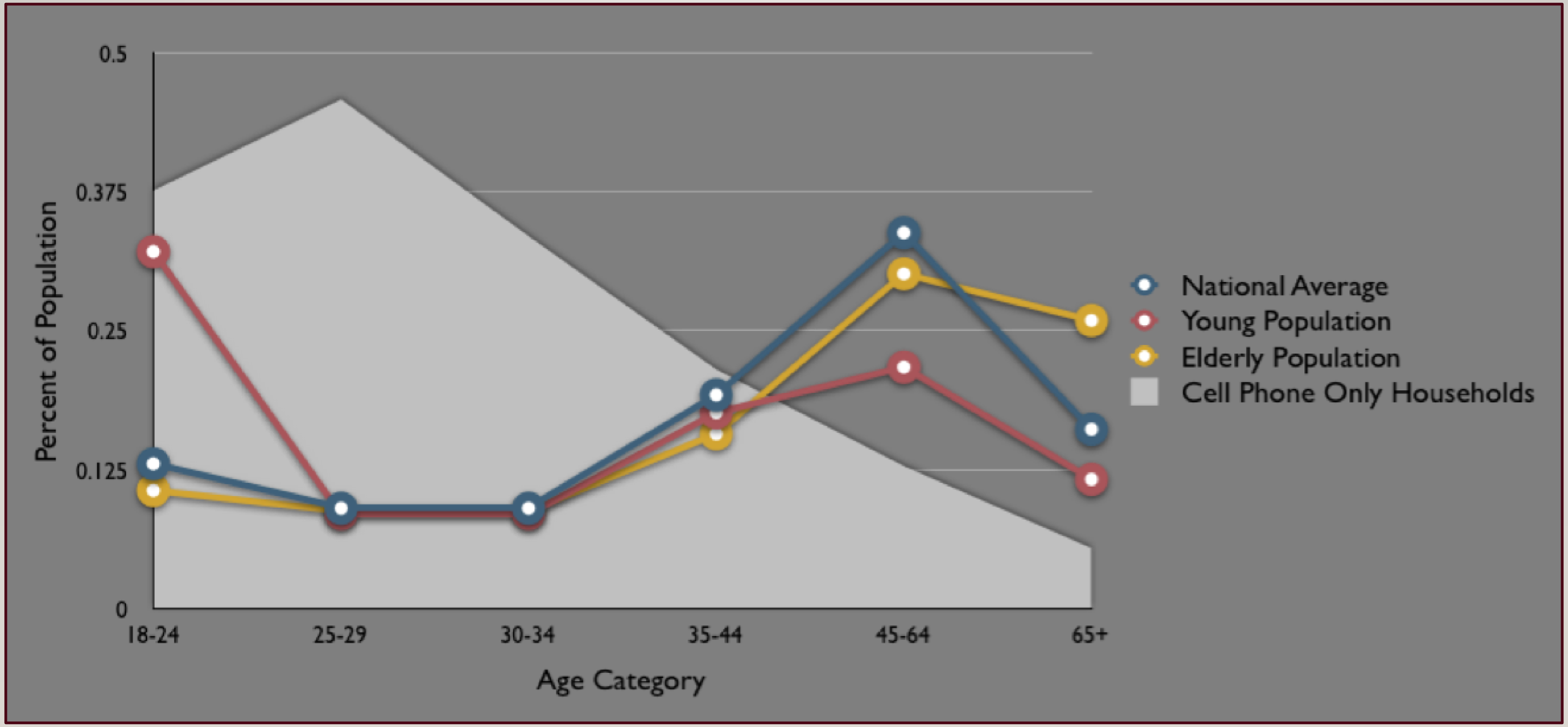


The ***Standard Error*** of an estimate is the standard deviation of its *sampling distribution*:

$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

SAMPLING PROBLEMS – SOME EXAMPLES

- Case-Control studies are trying to sample both cases (people with an outcome) and controls (those without it)
- Cases are often from a known source – diagnostic cases, credit card records from a restaurant, etc.
- Controls are somewhat more difficult to recruit
  - We *used* to be able to do this via random digit dialing for a specific geographic area
- What's the problem with this?


0.5
0.375
0.25
0.125
0
Percent of Population
18-24
25-29
30-34
35-44
45-64
65+
Age Category
National Average
Young Population
Elderly Population
Cell Phone Only Households

VOLUNTEER BIAS

- People who volunteer, consent to studies, etc. can be systematically different than those who don’t
- There’s potentially very legitimate reasons for this

HEALTHY WORKER BIAS

- People who are employed are, on average, healthier than those who aren't
- Why might this be?

- The result is that occupational cohorts are inherently healthier than the population as a whole
- Similarly, many hospital based populations, while being made up of sick people, are made up of *sick people with access to care*

TIME-BASED BIAS

- This often occurs in infectious disease and outbreak research
- Early estimates of case-fatality rates, etc. are often biased (and were in COVID-19 in Italy, for example) because they are drawing from hospitalizations or severe cases, rather than broad diagnostic testing
  - This is a problem if you extrapolate to the whole population
  - Italy CFR: 7.2%   Korea CFR: 1.0%
- For zoonotic disease, this may also involve a heavier proportion of primary cases (those with direct animal contact) vs. secondary cases (human-to-human transmission)

[2]Onder G, Rezza G, Brusaferro S. Case-Fatality Rate and Characteristics of Patients Dying in Relation to COVID-19 in Italy. *JAMA.* 2020;323(18):1775–1776.

HOW CAN WE ADDRESS THIS BIAS?