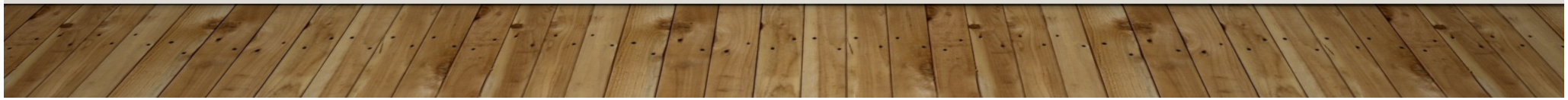
BIOMED SCI 552:

STATISTICAL THINKING

LECTURE 3: SAMPLING AND ESTIMATION

QUESTIONS FROM LAST CLASS?



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

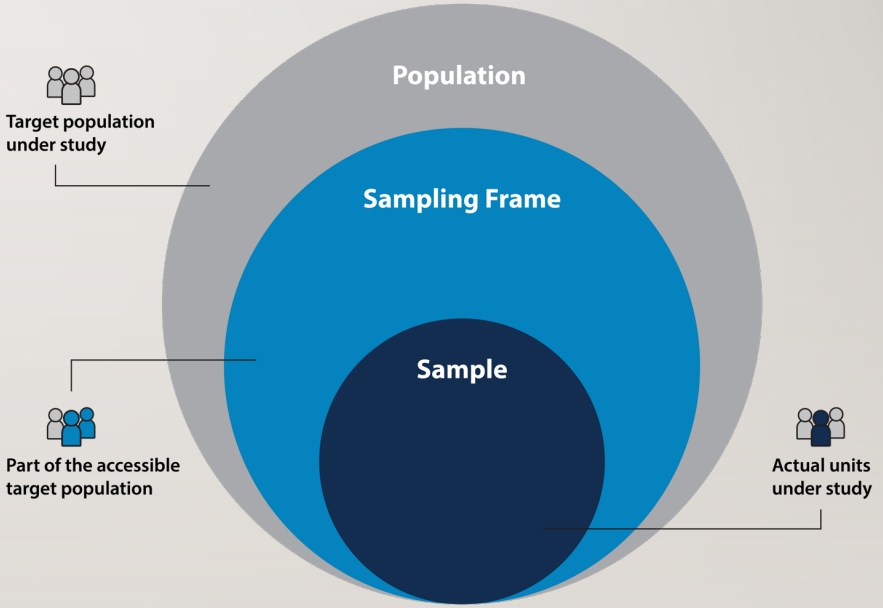
- There are some circumstances where we sample *with* replacement – we draw a sample from the population, and it is possible, with some probability, that we draw that sample again
- Capture-Recapture methods, for example, can be used to estimate population size by asking what population is most probable 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

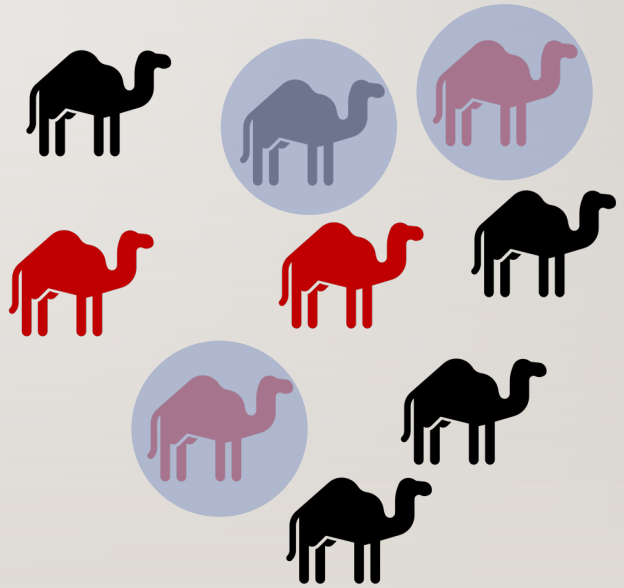


TYPES OF RANDOM SAMPLES

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

SIMPLE RANDOM SAMPLING

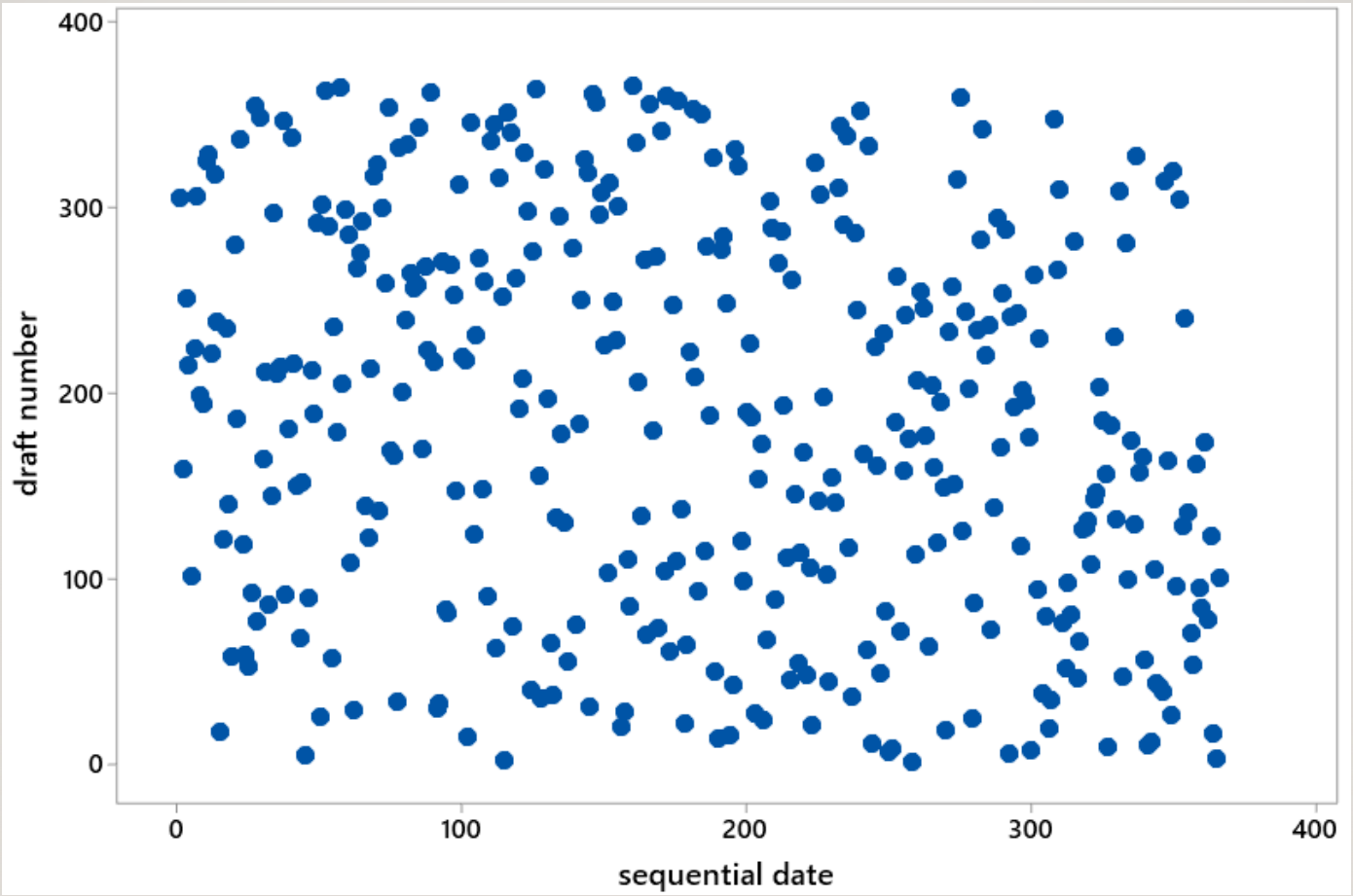
- The most basic, and potentially appealing type of sample
- Every member of the population has an equal probability of being included in the sample
- This is most easily done if you have a complete roster of the population in some form
  - A registry, patient records, a census, a population cohort,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

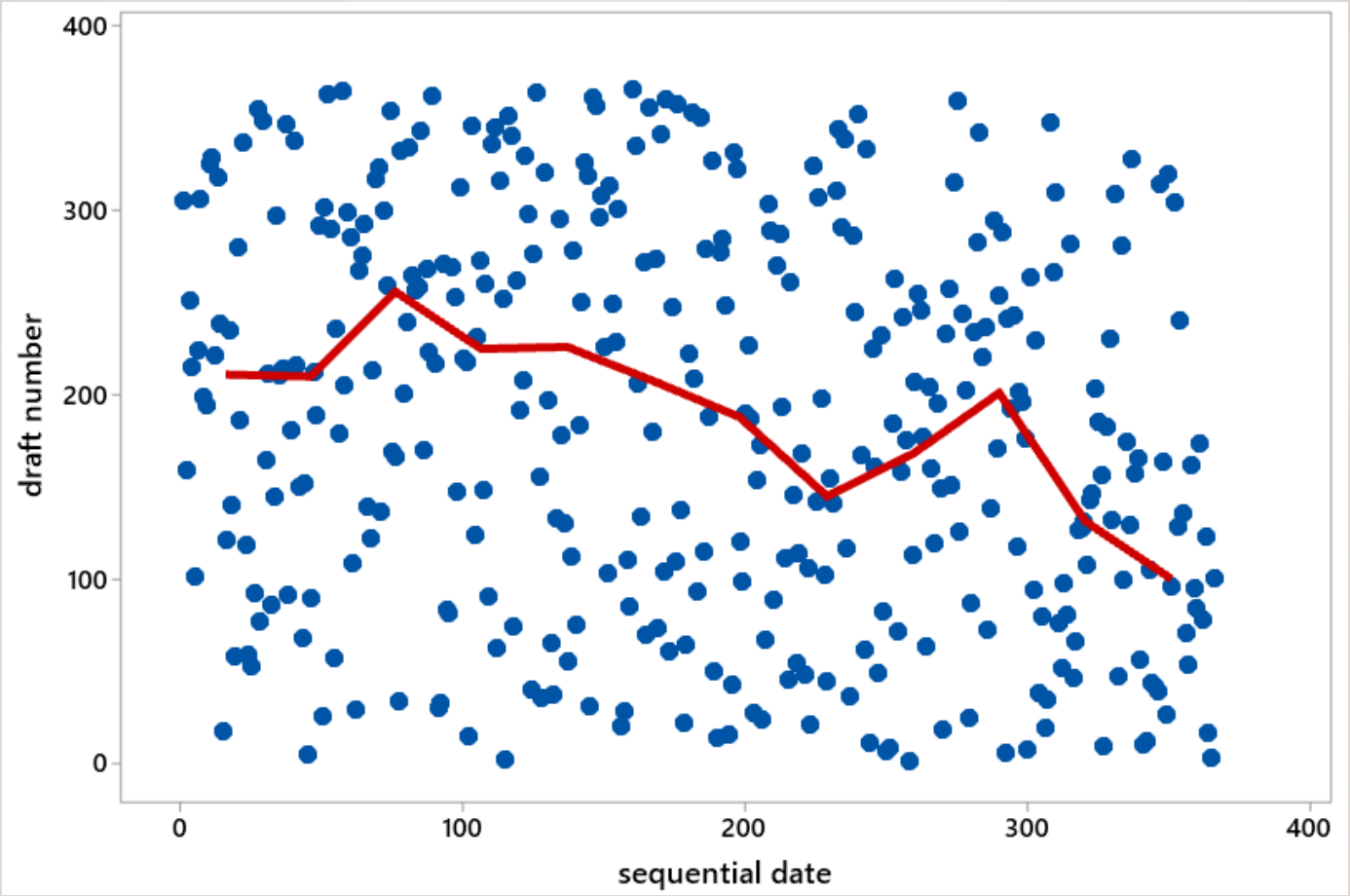


WHAT MIGHT BE SOME DRAWBACKS TO A SIMPLE RANDOM SAMPLE?

THE VIETNAM DRAFT

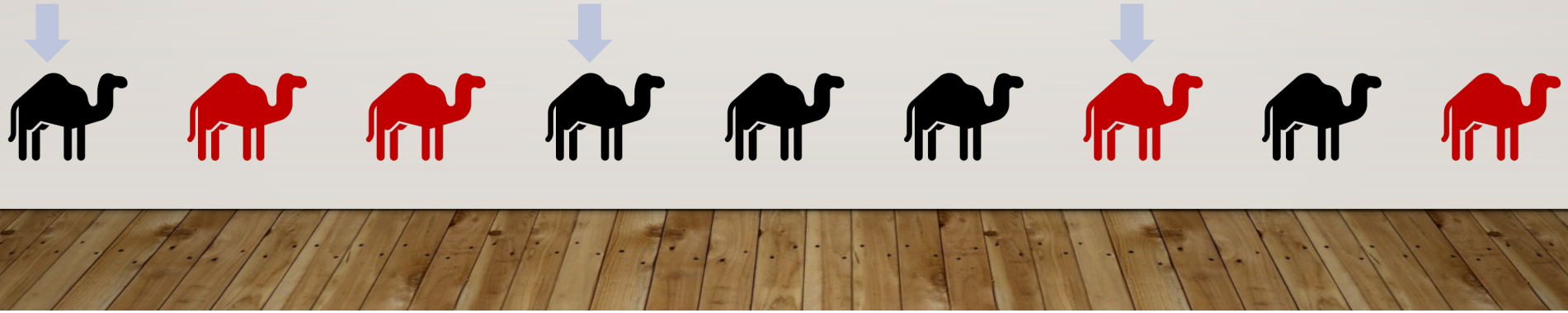
- The Vietnam Draft Lottery of 1969 was intended (rightly) as a simple random sample of birth dates (so as to hopefully not disadvantage any one group)
  - What is the assumption here?
- 366 capsules, each with a birth date, were placed in a bin, and then drawn in order to assign draft numbers
  - Number one, called first, was Sept. 14th


400
300
200
100
0
draft number
0
100
200
300
400
sequential date


400
300
200
100
0
draft number
0
100
200
300
400
sequential date

SYSTEMATIC SAMPLING

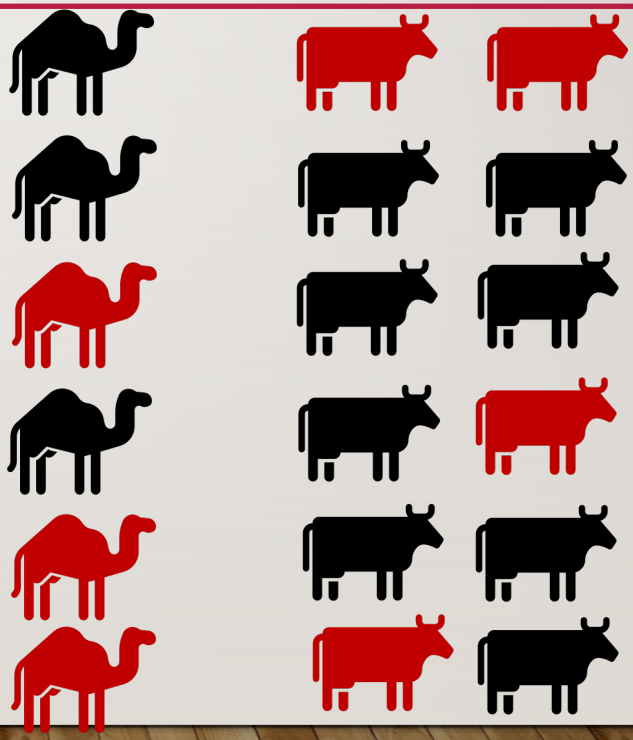
- Like a simple random sample
- The individuals in the sampling frame are arrayed in some order, and then every N^{th} element of that array is included in the sample
- This is often easier to implement, you control the sample size, etc.



WHAT ARE WE ASSUMING?

STRATIFIED SAMPLING

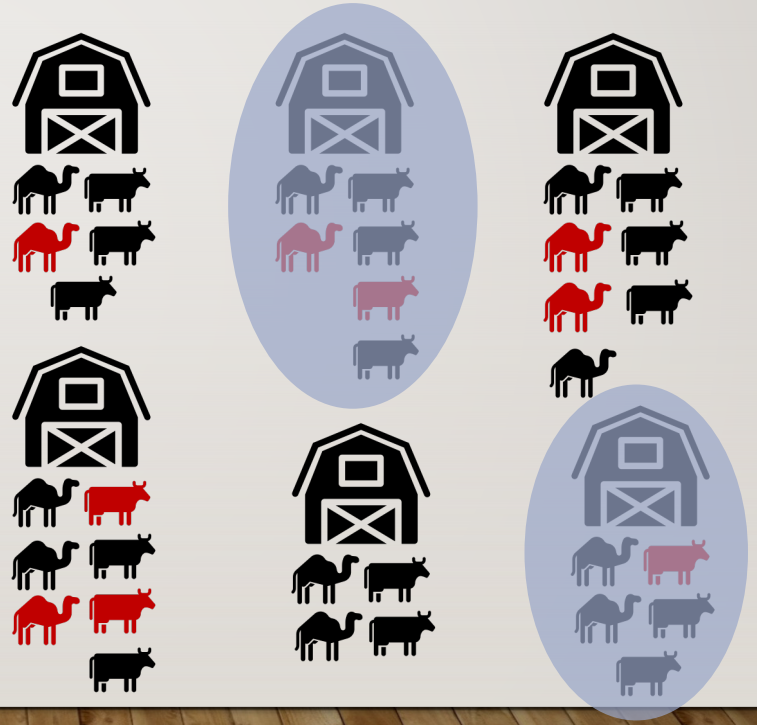
- Divide the population into some logical subgroups (called strata)
  - Age range, job role, species, phenotype, etc.
  - Take the proportion of the population in each strata, and sample that proportion of your total sample from the strata
    - i.e. there should be two cow samples for every one camel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

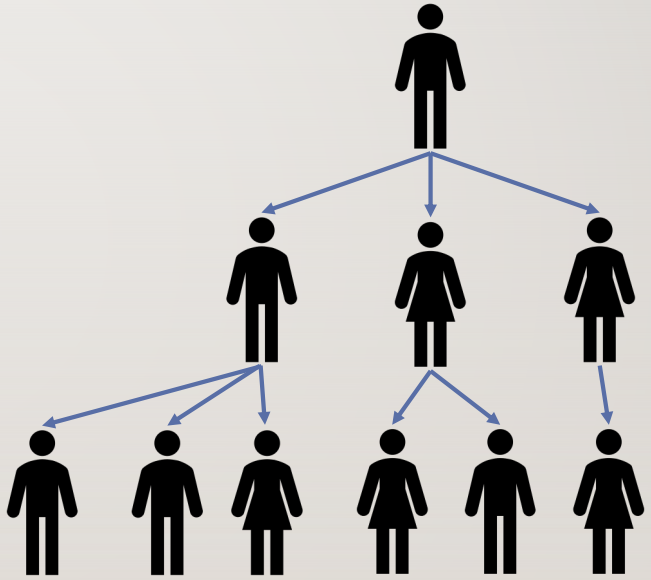
- All of the methods we've discussed have some sort of probabilistic aspect to them
- There are *nonprobabilistic* sampling methods that…unsurprisingly…don't involve randomization
- What's one example we've already talked about?
- Types
  - Convenience samples
  - Purposive samples
  - Snowball samples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

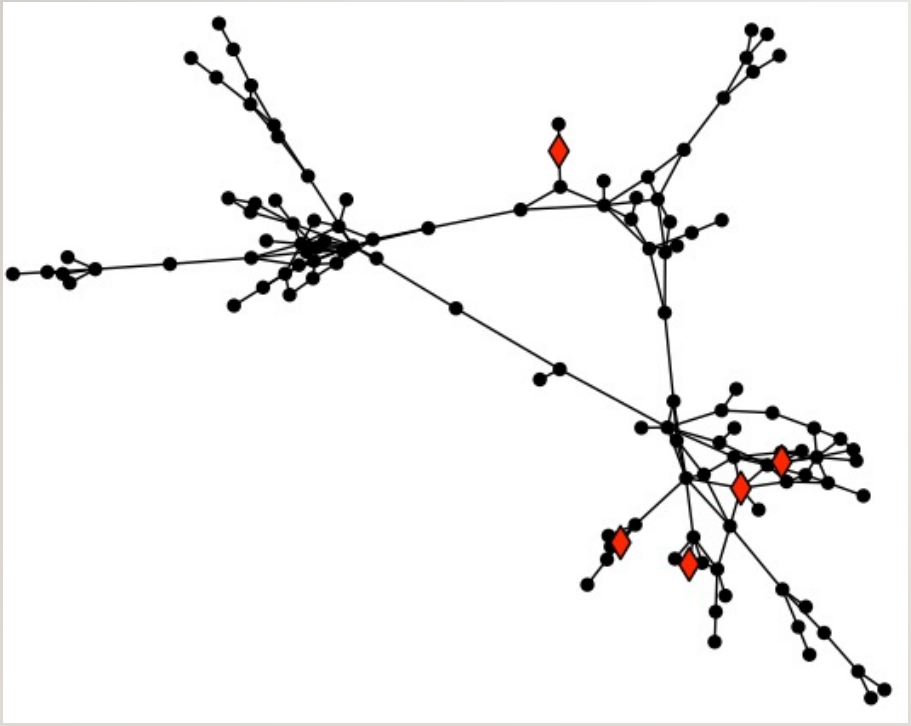
SNOWBALL SAMPLING

- 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

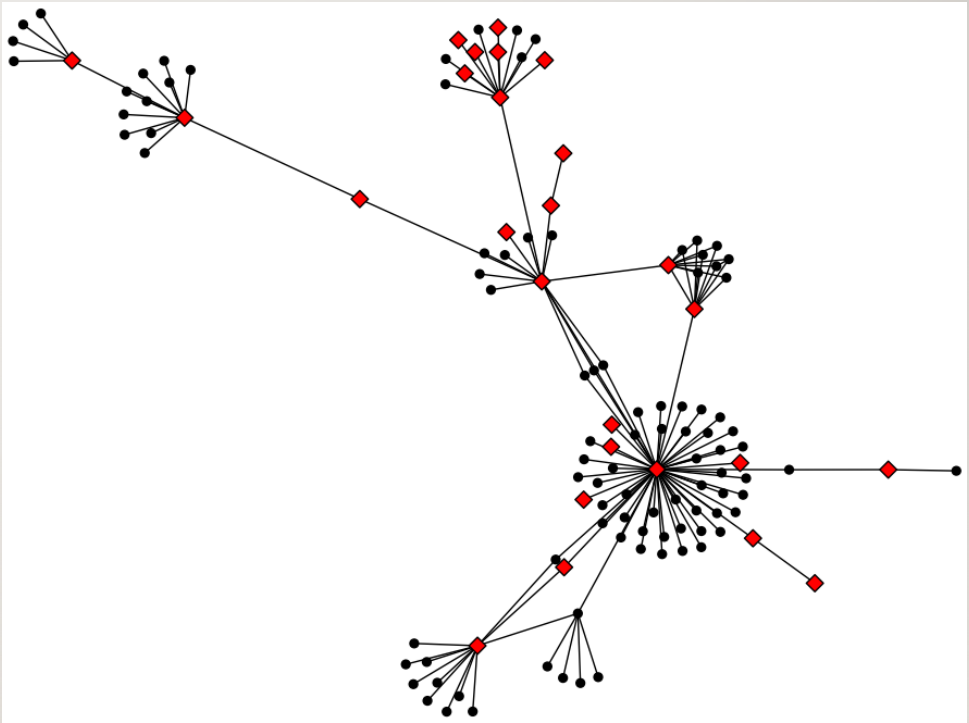


THE FRIENDSHIP PARADOX

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



Friendship Network



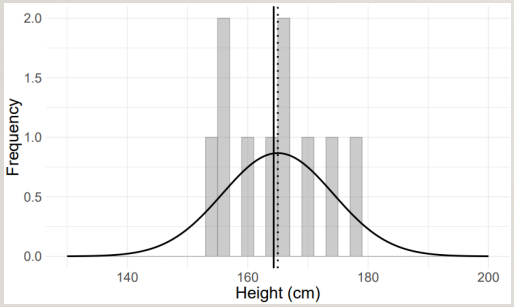
Sexual Contact Network

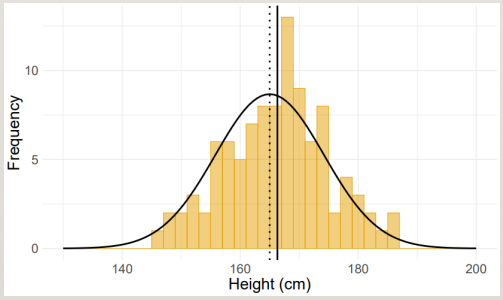
QUOTA SAMPLING

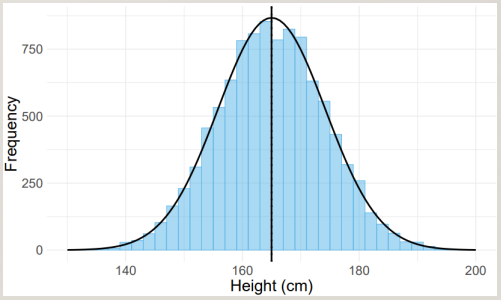
- Non-random selection of a predetermined number of individuals of a given type
  - Again, the population is divided into strata
  - A fixed number of people from each strata are then selected
- This ensures you get a broad swathe of your population, but a non-random one
- Again, this can be used heavily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

SAMPLING FROM A NORMAL DISTRIBUTION

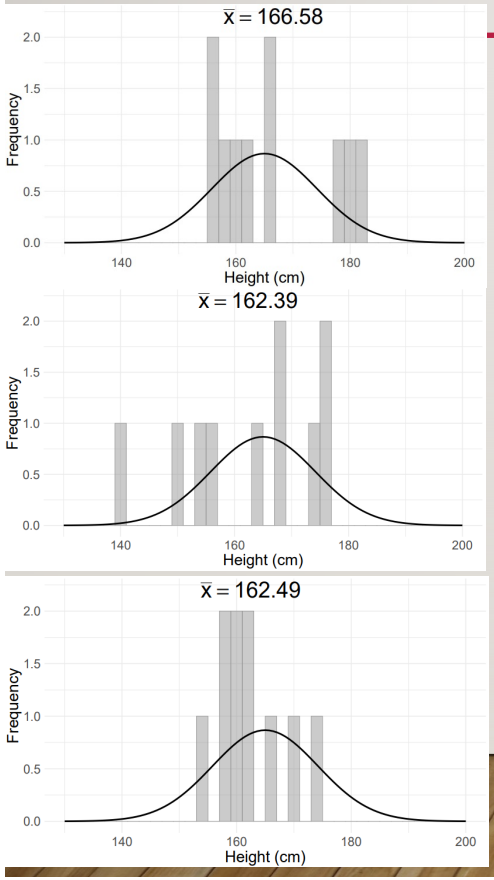
- The true population mean is $\boldsymbol{\mu}$ and the true population SD is $\boldsymbol{\sigma}$.
- Each time we sample a population, we get a different subset purely by chance with mean $\bar{\boldsymbol{x}}$ and SD $\boldsymbol{s}$.
- Larger sample sizes give us more certainty about the true population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





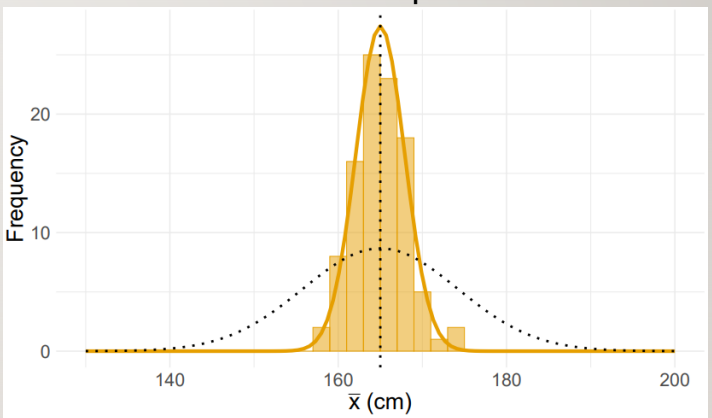


SAMPLING DISTRIBUTIONS



- 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

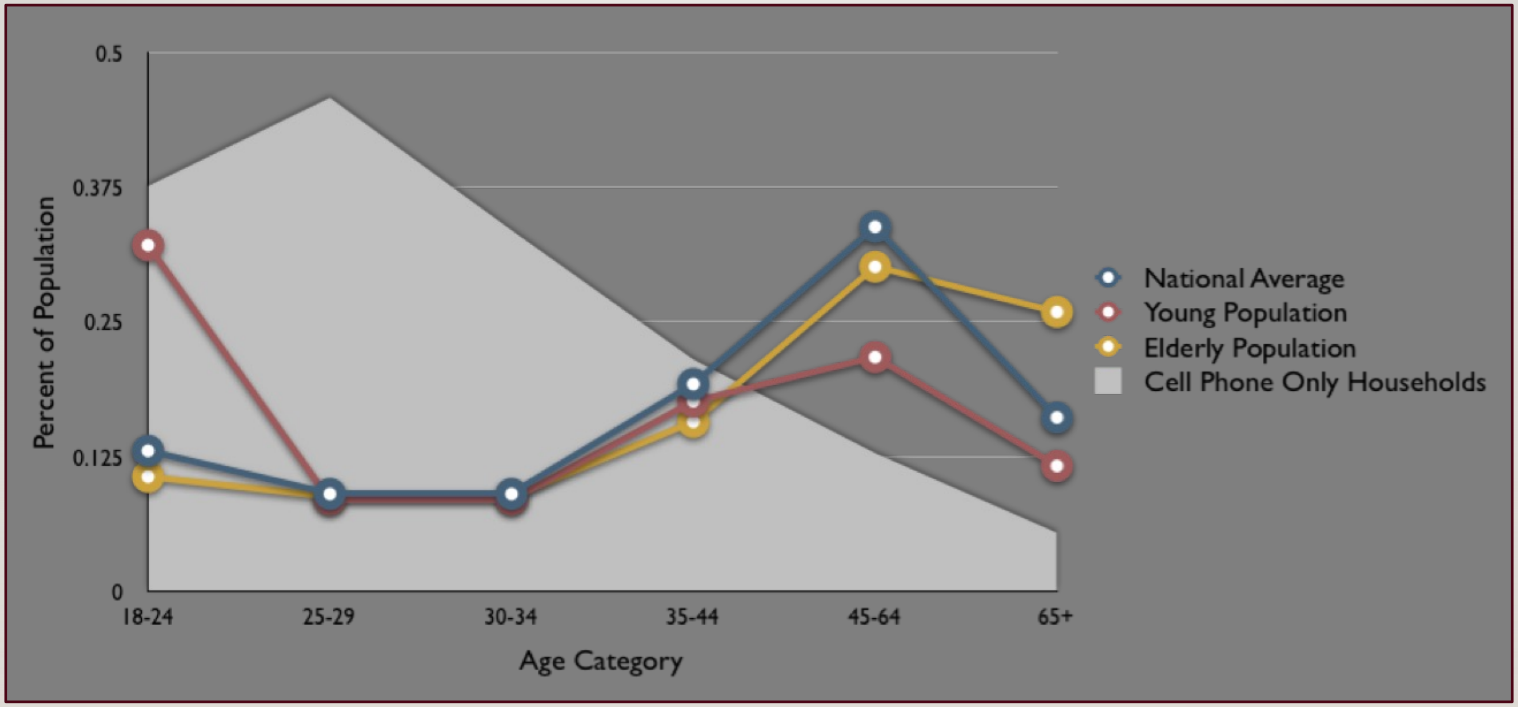


The ***Standard Error*** of an estimate is the standard deviation of its *sampling distribution*:

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

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?

QUESTIONS?

WHAT IS ESTIMATION?

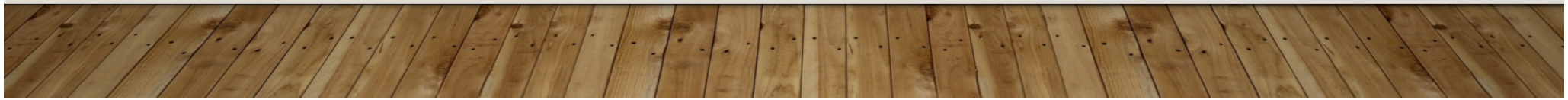
- We've alluded to estimation a lot in this class so far, because all these concepts are somewhat intertwined
- The core of much of what we want to *do* with statistics is called estimation
- Estimation is the process of inferring an unknown (and unknowable) quantity from the population using data from the same
- It is, in effect, our best guess using data

SO WHAT'S AN ESTIMATOR?

- An estimator is the method by which we obtain an estimate
- These can be quite sophisticated, or quite simple
- One of the ones we've already used in this class a lot is the estimator of the sample mean, $\bar{X}$, which is an estimate of the population mean
- $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$

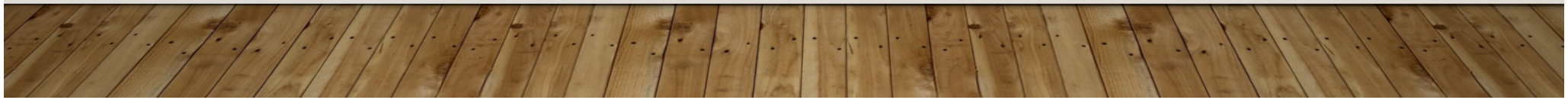
CONSISTENT ESTIMATORS

- A consistent estimator (sometimes called an asymptotically consistent estimator) is one where as sample size increases, you will eventually converge on the true estimate
- An example of an inconsistent estimator would be an estimate of the population mean that has the first value you sample
  - Technically, this is an unbiased estimator
  - But it does not get closer to the population estimate because it is fixed regardless of sample size
- Encountering an inconsistent estimator in the wild is *rare*



EFFICIENT ESTIMATORS

- This just means that there's no *other* estimator that would arrive at the same estimate with a smaller accompanying variance



HATS AND BARS

- $\hat{X}$ and $\bar{X}$ are both going to appear today
- They feel like they often mean the same thing, but they don't
- $\bar{X}$ is the sample mean of something
- $\hat{X}$ is an estimate
- We can use $\hat{X}$ as an estimate of $\bar{X}$, but it's not the only one

BIASED VS UNBIASED ESTIMATORS

- The bias of an estimator is the difference between the estimator's expected value and the true value of the parameter being estimated
- For an unbiased estimator, this difference is 0
- For a biased estimator, the difference is either <0 or >0

- Why would we ever use a biased estimator?

SOMETIMES AN UNBIASED ESTIMATOR DOESN'T EXIST

- The is, for example, no unbiased estimator for $\frac{1}{\mu}$ with observations from a Poisson distribution with a mean of μ

BIAS-VARIANCE TRADE OFF

- It is possible, and indeed quite common, that an *unbiased* estimator may also be considerably less precise
- If we think about the true population value we're estimating as the target, we can measure how "off" we are by looking at "Mean Square Error"
- $MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$
- It can be shown that this is equal to:
- $MSE(\hat{\theta}) = (Bias(\hat{\theta}, \theta)^2 + Variance(\hat{\theta})$

BIAS-VARIANCE TRADEOFF

- This means that there are often situations where it might be desirable to trade a little bit of bias for a large reduction in variance, and still end up close to your answer
- While we said that high bias, low variance is a bad place to potentially be, smallish bias, low variance may be preferable to unbiased and high variance

AN EXAMPLE…

- Lets say I tell you I'm thinking of a number, and the closer you get to the number, the greater your prize
- I tell you that to pick your number, you have to choose between one of two methods to randomly draw a guess (i.e. I let you choose between two estimators):
  - $\hat{X} \sim Uniform(X - 5, X)$
  - $\hat{X} \sim Uniform(X - 100, X + 100)$
- Which would you prefer?

IN A PERFECT WORLD

- Obviously, in a perfect world, we would prefer an unbiased estimate with low variance, but we cannot always promise this will be true
- Sometimes, known bias in an estimator can be corrected, etc.
- Many, if not most, estimators are biased
- A deeper dive into this gets very complicated very quickly
  - For the most part, the methods we use are at the very least ones with known properties

HOW THIS COMES UP IN THE BIOMEDICAL SCIENCES

- There are two major ways this comes up in the biomedical sciences
- Pure estimation
  - A colleague of mine: "Epidemiologists just…*measure things*."
  - Very commonly estimation is part of a larger chain of research
  - Examples:
    - Estimating the average patient length of stay to model hospital bed availability
    - Estimating the prevalence of a particular condition for allocating resources for it
    - Estimating a bacterial growth rate under certain conditions
  - Here, the goal is to minimize MSE (or some other measure of overall error), *not* just to minimize bias

"CONTROLLING FOR VARIABLES"

- When we adjust for things, why don't we just adjust for everything we can possibly think of?

"CONTROLLING FOR VARIABLES"

- When we adjust for things, why don't we just adjust for everything we can possibly think of?

- The more variables we add, the more variance increases
- We are essentially "spending" variance to protect us from bias
- At some point there's a diminishing payoff for this
  - There are formal ways to assess this

THE BAYESIAN PERSPECTIVE

- Gelman et al., *Bayesian Data Analysis*
- "From a Bayesian perspective, the principle of unbiasedness is reasonable in the limit of large samples but otherwise is potentially misleading. The major difficulties arise when there are many parameters to be estimated and our knowledge or partial knowledge of some of these parameters is clearly relevant to the estimation of others. Requiring unbiased estimates will often lead to relevant information being ignored... In sampling theory terms, minimizing bias will often lead to counterproductive increases in variance."

IS A PRIOR BIAS?

CONNECTING THIS TO SAMPLING

- All of this falls under what's called "sampling theory" because it's centered on what inference we can draw from a sample
- Our estimator itself cannot be ensured to be unbiased
- There is often a notion that "Why can't we just make a method that doesn't have *X* undesirably property?"
- Sadly, math doesn't work like that

CONNECTING THIS TO SAMPLING

- All of this falls under what's called "sampling theory" because it's centered on what inference we can draw from a sample
- Our estimator itself cannot be ensured to be unbiased
- There is often a notion that "Why can't we just make a method that doesn't have *X* undesirably property?"
- Sadly, math doesn't work like that

CONNECTING THIS TO SAMPLING

- Given we cannot realistically control the properties of the estimators we use…
- We should endeavor to minimize the bias that comes from things we *can* control
- Our goal is to ensure that we do not add any more bias into a sample than what is inevitable due to problems with sampling frames, estimators, etc.