Advanced Statistics for Social Sciences 1

Statistical Thinking and Data Analysis with R

Instructor: Rachit Dubey
Email: rdubey@ucla.edu

Summary of last 2 weeks

Week 3:

- Probability basics
- Independent and dependent events
- Bayes' Rule
- Counting: Permutations and Combinations

Week 4:

- Random variables
- Discrete random variables: Bernoulli, Binomial, and Geometric
- Expectation and variance
- Probability Mass Function
- Continuous Random Variables: PDF and CDF

Lecture 5: Sampling

Normal Distribution

- Normal Distribution and sufficient statistics
- Standard normal distribution

Sampling

- Samples, populations and sampling
- Law of large numbers
- Sampling distributions
- Central Limit Theorem

Normal distribution

(Worked on whiteboard)

Demo: https://seeing-theory.brown.edu/probability-distributions/index.html

Sampling

Inferential Statistics

Until now, we have primarily dealt with Descriptive Statistics

This is only a small part of statistics!

The bigger, more useful part of statistics is that it lets us make inferences about data

Probability & Statistics i.e., why did we learn probability?

Probability Theory: “Doctrine of chances”

- What are the chances of a fair coin coming up heads 10 times in a row?
- If I roll two six sided dice, how likely is it that I will roll two sixes?

Here: Truth of the world is known e.g., P(heads) = 0.5 or P(rolling 6) = 1/6.

We use probability to ask “what kinds of events” will happen (given the truth).

In other words, in probability theory, we know the ***model*** but the data is unknown.

Probability & Statistics i.e., why did we learn probability?

Statistics: In statistics, we do not know the truth of the world.

- If my friend flips a coin 10 times and gets 10 heads, are they playing a trick on us?
- If five cards off the top of the card deck are all hearts, how likely is it that the deck was shuffled?

Here: All we have is data, and we want to *learn* the truth about the world from the data.

Goal of statistics: Use the data to ***infer*** whether or not we should arrive at certain conclusions.

The statistical inference problem is to figure out which of the possible probability models is right

Sampling Theory

Sampling theory

- Our first task is usually to come up with some general assumptions about data that make sense.
- This is where **sampling theory** comes in.
- If probability theory is the foundation upon which all statistical theory builds, sampling theory is the frame around which you can build the rest of the house.
- **Before that:** We need to be a bit more explicit about what we are drawing inferences from (*sample*) and what is it that we are drawing inferences about (*population*).

Population

Population

- **All** possible observations we care about
- An abstract idea: refers to **all possible observations** we want to draw conclusions about
- Sometimes it is easy to define a population
    - E.g., polling — we want to draw conclusions about all voters in America
- Most times, really hard to define a population

Example: I am running a psychology experiment to understand how memory influences curiosity. I run my experiments on 100 undergraduate students at UCLA.

What is "the population"?

- All undergrad students at UCLA?
- All undergrad students, anywhere in the world?
- All Californians currently alive?
- Anyone who is alive?
- Any human being?
- Any biological being with reasonable intelligence?
- Any intelligent being?

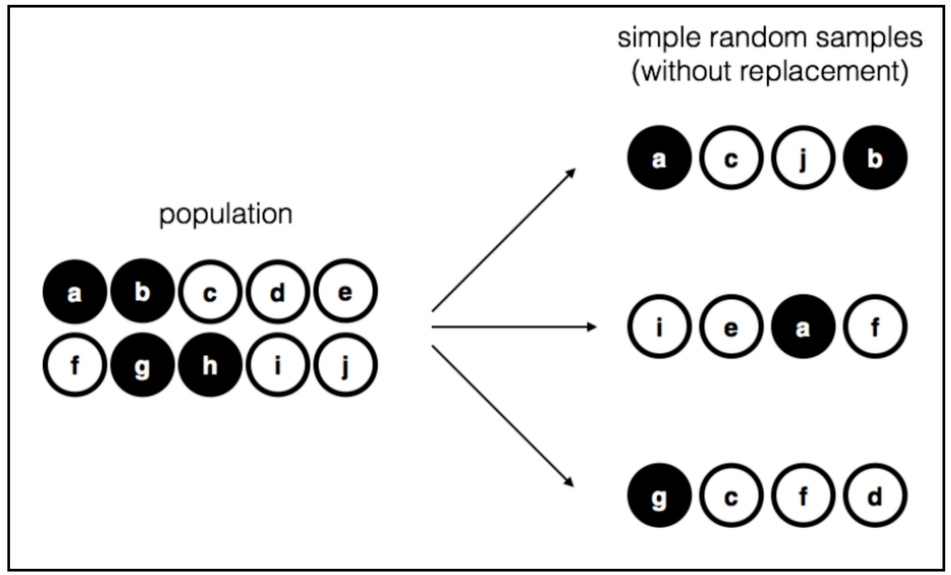
Sample

Sample

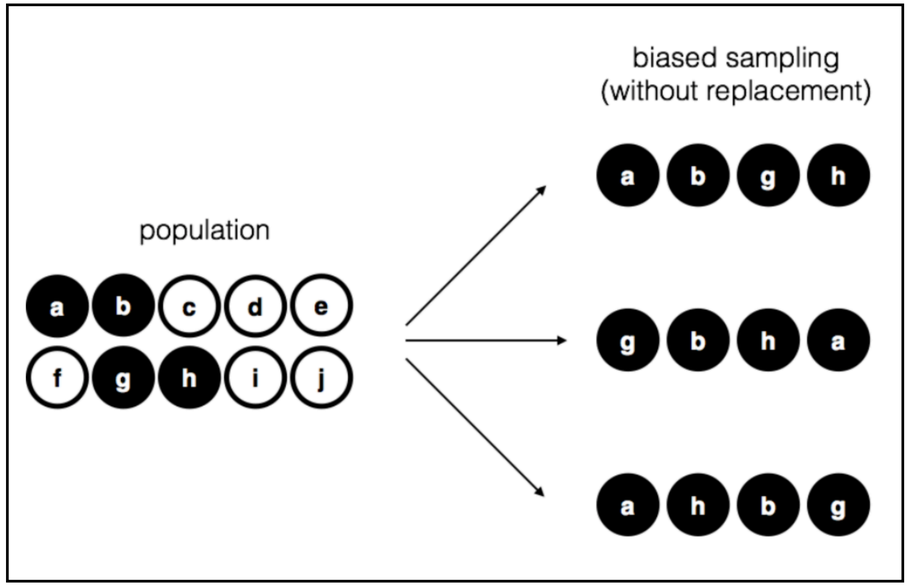
- The data we actually collected.
- **Goal:** Use knowledge of the sample to draw inferences about the properties of the population.
- **Statistical Inference:** The process of making claims about a population based on information from a sample.
- Life would be easy if we were able to observe the whole population — we could simply do descriptive analyses!

- **Key question:** What can we infer about the population from our sample?
- The relationship between the two depends on the ***sampling method*** i.e., the procedure via which the sample was selected.

Simple random sampling

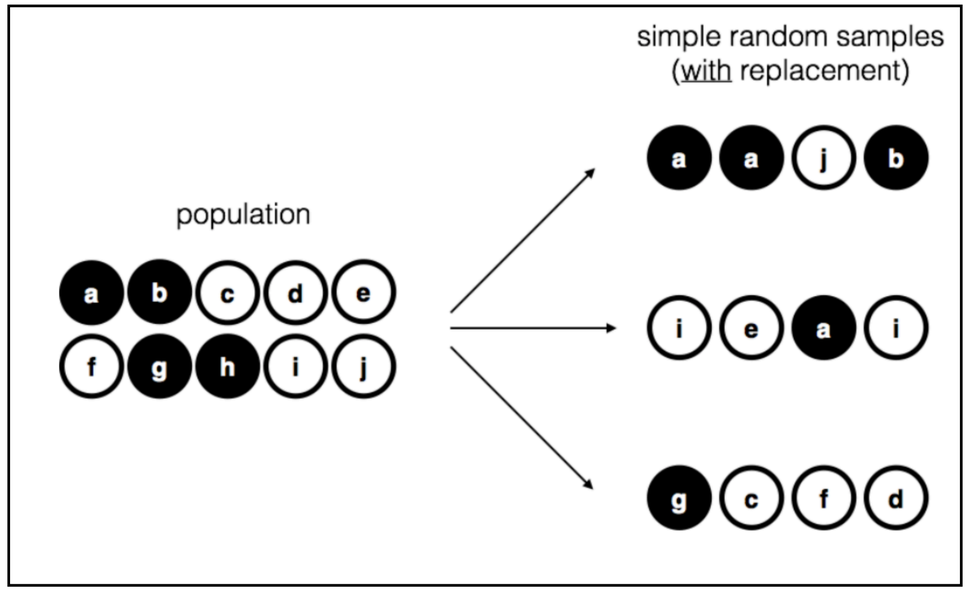


Simple random sampling **without replacement** from a finite population



Biased sampling without replacement from a finite population

Simple random sampling



Simple random sampling ***with* replacement** from a finite population

Usually rare for psychology experiments (we usually don't allow a participant to take part in our experiments twice), but can be helpful in certain contexts (e.g., pilot testing or bootstrapping, more on this next week).

Note: if our sample size is large, this starts mattering less..

Most samples are NOT simple random samples

Would be a miracle if my participants (in the previous experiment) turned out to be a truly random sample of the undergraduate population at UCLA.

Some other important sampling schemes:

- **Stratified Sampling:** Divide population into subpopulations (*strata*) and collect separate random samples from each.
    - Easier and more efficient than simple random sampling, esp for rare subpopulations
    - E.g., Schizophrenia research — would be hard to get participants if you randomly sampled from the population. Instead, sample equal numbers from schizophrenic and non-schizophrenic groups.
    - *Oversampling* deliberately over-represents rare groups to ensure adequate representation
- **Snowball Sampling:** Useful for hard-to-access or “hidden” populations.
    - Process: Start with initial contacts, ask them to refer others, continue until sufficient data is collected.
    - Disadvantages: Highly non-random sample, potential ethical concerns (consent, unintended outing).
- **Convenience Sampling:** Samples chosen based on researcher convenience rather than random selection.
    - Non-random by population restriction and self-selection of participants.
    - Easy but can introduce significant bias.

Non-random sampling

Does non-random sampling matter?

- Yes, it can matter—but not always in the ways you might think
- Stratified sampling introduces deliberate bias, but this is unproblematic because you know the bias and can statistically adjust for it

More important question: Is the bias relevant?

- Random sampling is a means to an end, not the end itself
- A biased sample is only problematic if it causes you to draw wrong conclusions
- You only need randomness with respect to the phenomenon of interest, not every characteristic

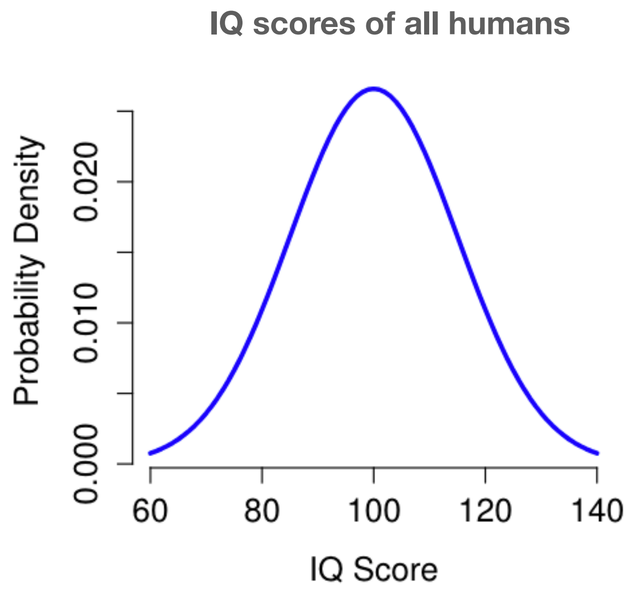
Example: An experiment on effects of working memory on curiosity

- Option 1: Random sample of all humans, except these people are all born on Tuesdays
- Option 2: Random sample of all college-educated Americans

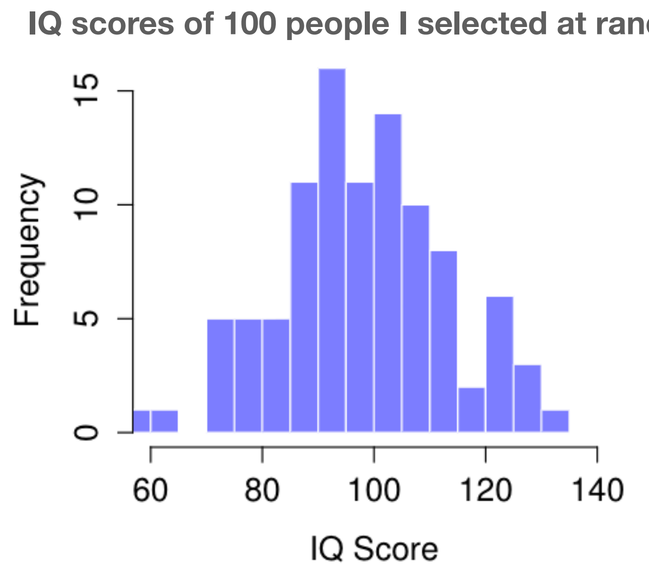
Practical Implication

- Think carefully about which population you want to generalize to
- Consider what characteristics of your convenience sample might actually affect your results
- Try to sample in ways appropriate to your target population, even if forced to use convenient samples

Population parameters & sample statistics



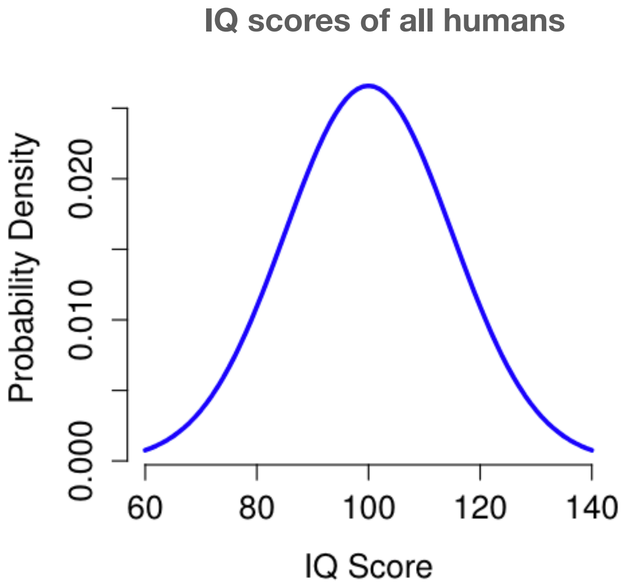
IQ scores are defined such that the average IQ is 100, and standard deviation is 15.
These values are the ***population parameters***



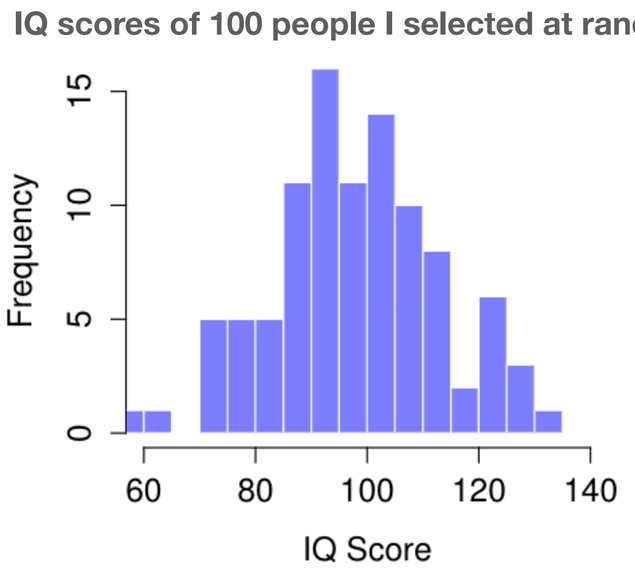
Average IQ = 98.5, and Standard deviation =15.9
These values are the ***sample statistics***

Law of large numbers

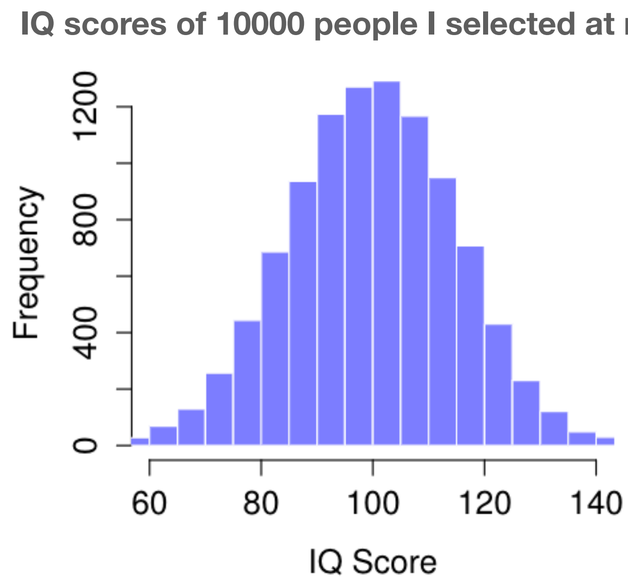
As the sample gets larger, the sample mean tends to get closer to the true population mean (we have seen this before)



Mean IQ = 100, and Standard deviation =15



Average IQ = 98.5, and Standard deviation =15.9



Average IQ = 99.9, and Standard deviation =15.1

Sampling Distributions

Limits of law of large numbers

[The] long run is a misleading guide to current affairs. In the long run we are all dead. Economists set themselves too easy, too useless a task, if in tempestuous seasons they can only tell us, that when the storm is long past, the ocean is flat again. (Keynes, 1923, p. 80)

It is not enough to know that we will *eventually* arrive at the right answer when calculating sample mean.

Knowing that an infinitely large data set will tell me the exact value of the population mean is cold comfort when my ***actual*** data set has a sample size of $N = 100$.

Sampling distribution of the mean

Let's abandon the idea we have sample sizes of 100,000 and let's start with a much more modest experiment.

- We sample $N = 5$ people and measure their IQ scores: [90, 82, 94, 99, 100]. We find Mean = 95.
- Now let's **replicate** this experiment:
  - Randomly sample 5 people again and measure their IQ: [78,88,111,111,117]. Mean = 101.
- Let's say we run this replication 10 times

	Person 1	Person 2	Person 3	Person 4	Person 5	Sample Mean
Replication 1	90	82	94	99	110	95.0
Replication 2	78	88	111	111	117	101.0
Replication 3	111	122	91	98	86	101.6
Replication 4	98	96	119	99	107	103.8
Replication 5	105	113	103	103	98	104.4
Replication 6	81	89	93	85	114	92.4
Replication 7	100	93	108	98	133	106.4
Replication 8	107	100	105	117	85	102.8
Replication 9	86	119	108	73	116	100.4
Replication 10	95	126	112	120	76	105.8

Sampling distribution of the mean

Let's abandon the idea we have sample sizes of 100,000 and let's start with a much more modest experiment.

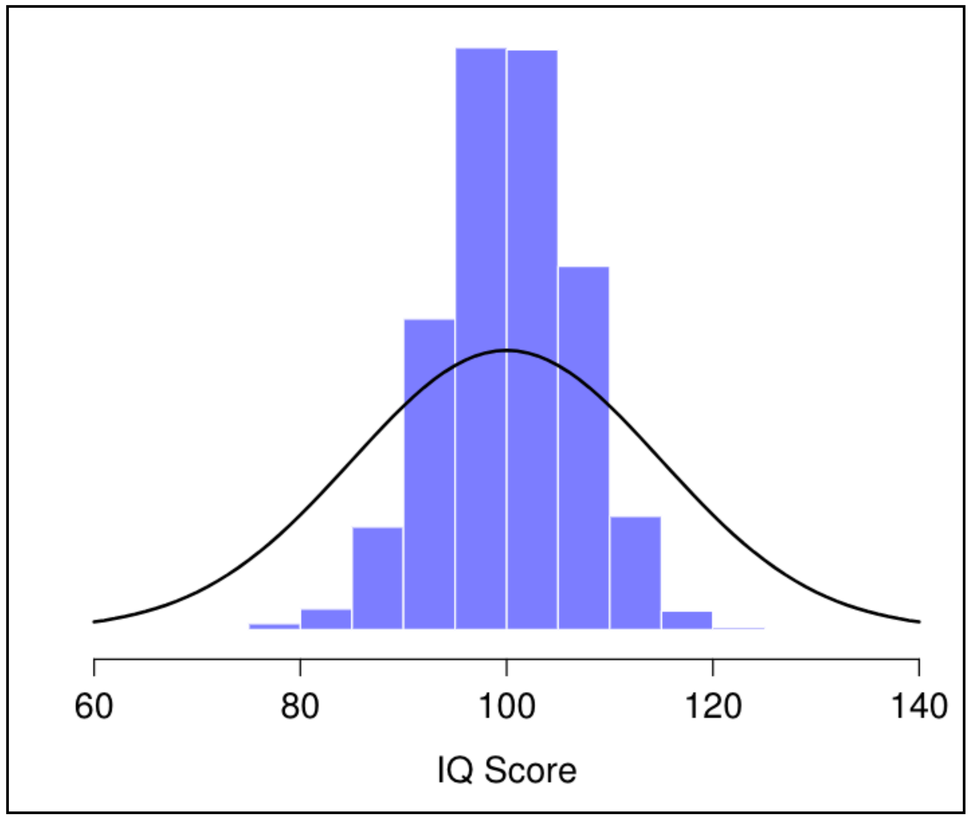
- We sample $N = 5$ people and measure their IQ scores: [90, 82, 94, 99, 100]. We find Mean = 95.
- Now let's **replicate** this experiment:
  - Randomly sample 5 people again and measure their IQ: [78,88,111,111,117]. Mean = 101.
- Let's say we run this replication 10 times

	Person 1	Person 2	Person 3	Person 4	Person 5	Sample Mean
Replication 1	90	82	94	99	110	95.0
Replication 2	78	88	111	111	117	101.0
Replication 3	111	122	91	98	86	101.6
Replication 4	98	96	119	99	107	103.8
Replication 5	105	113	103	103	98	104.4
Replication 6	81	89	93	85	114	92.4
Replication 7	100	93	108	98	133	106.4
Replication 8	107	100	105	117	85	102.8
Replication 9	86	119	108	73	116	100.4
Replication 10	95	126	112	120	76	105.8

Sampling distribution of the mean

What if we continued like this for 10,000 replications?

We end up with a *distribution* of sample means. This is the **sampling distribution of the mean**.

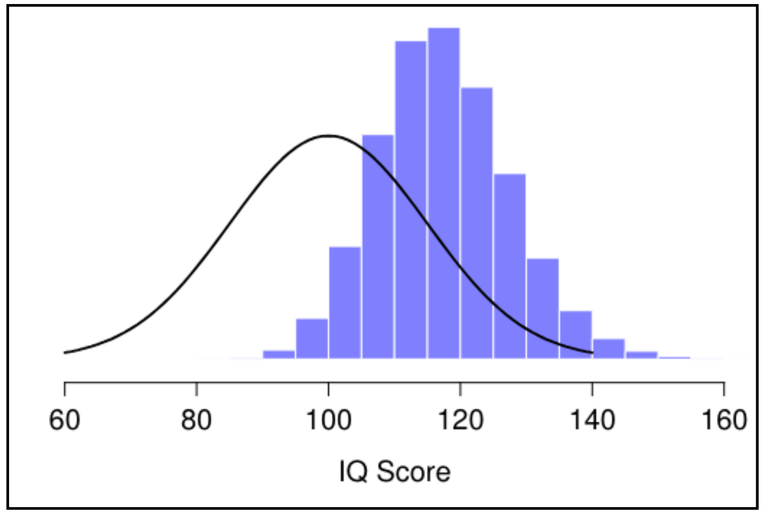


If you sample 5 people at random and calculate their average IQ, you'll almost certainly get a number between 80 and 120, even though there are quite a lot of individuals who have IQs above 120 or below 80. For comparison, the black line plots the population distribution of IQ scores.

Sampling distribution

Side note: Sampling distribution exists for any sample statistic!

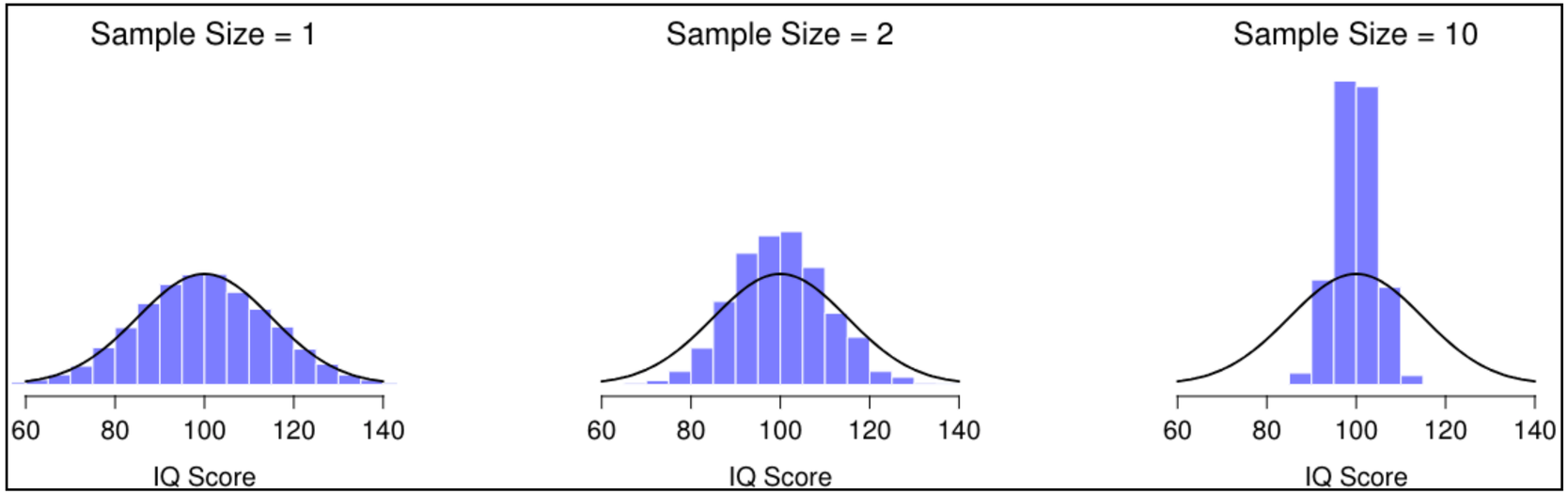
E.g., sampling distribution of the *maximum*.



If you sample 5 people at random and select the one with the largest IQ, you'll probably see someone with an IQ between 100 and 140.
For comparison, the black line plots the population distribution of IQ scores.

Sampling distribution of the mean

Side note 2: Sampling distribution of the mean depends on the sample size.



The bigger the sample size, the narrower the sampling distribution of the mean.

In other words, the sample mean gets more accurate as our sample size increases.

Standard error of the mean

Obviously, our sample distribution of the mean is going to have some errors.. we can quantify this!

Standard error of the mean tells us how much variation should we expect between the means of different samples.

How do we calculate this?

Recap: Standard deviation tells us how well the mean summarizes the data.

We can use this to calculate standard deviation of the sampling distribution!

$$SEM = \frac{\hat{\sigma}}{\sqrt{n}}$$

Standard error of the mean

Sample Size and Standard Error of the Mean (SEM)

- Larger sample size yields smaller SEM when population variability is constant
- We cannot control population variability, but we can *control* sample size

Practical Strategy

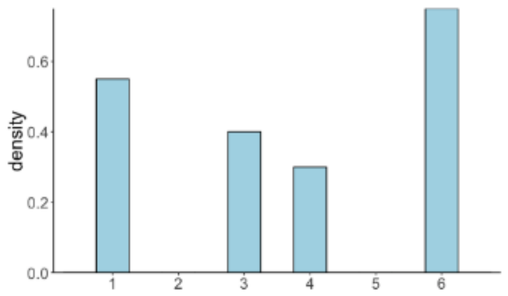
- To improve sample statistics and reduce sampling variability, use larger samples

The Law of Diminishing Returns

- The utility of larger samples diminishes with the square root of the sample size
- Doubling sample size does not halve the SEM; improvement follows a square root relationship
- Implication: Each additional participant has less impact on reducing variability as sample size grows

Standard error of the mean

our sample

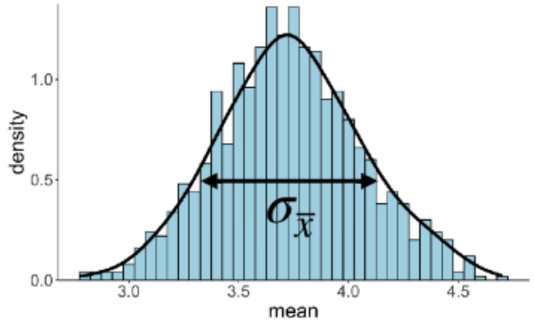


standard deviation

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N - 1}}$$

gives a sense for how well the mean summarizes the data

sampling distribution



standard error

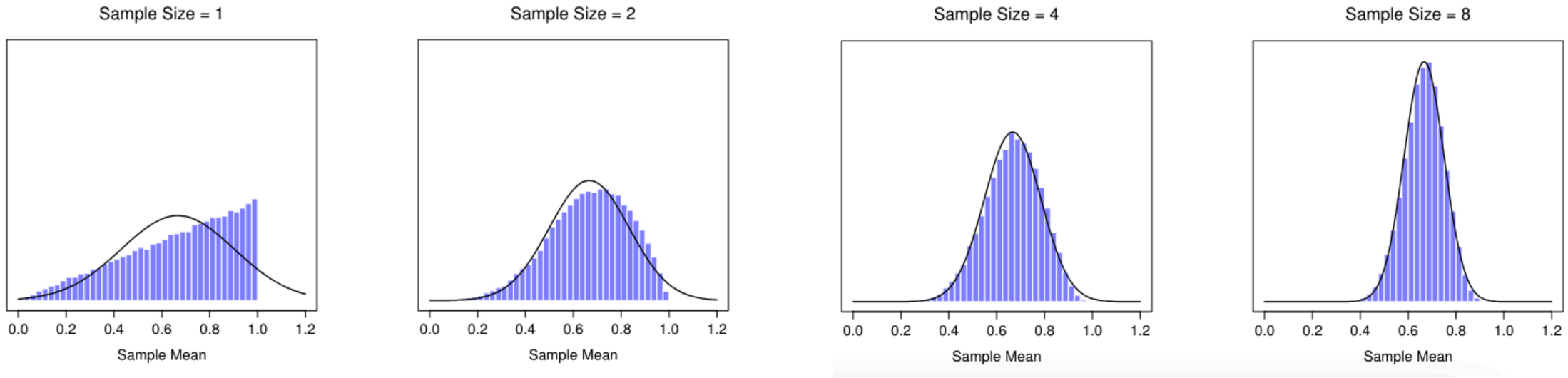
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$$

the standard deviation of the sampling distribution

how much variation would we expect between the means of different samples

how likely is it that our sample mean is representative of the population mean?

Sampling distribution of the mean



Central Limit Theorem: As long as your sample size isn't tiny, the sampling distribution of the mean will be approximately ***normal*** no matter what your population distribution looks like!

Central Limit Theorem

Central Limit Theorem: As long as your sample size isn't tiny, the sampling distribution of the mean will be approximately ***normal*** no matter what your population distribution looks like!

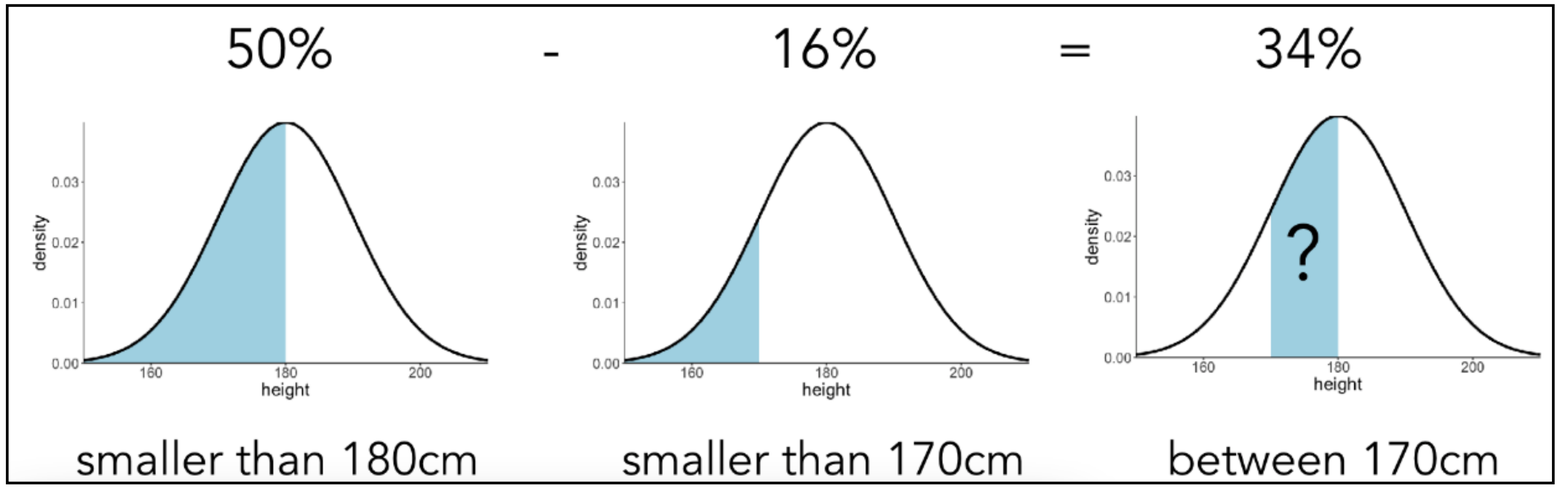
- The mean of the sampling distribution is the same as the mean of the population
- The standard deviation of the sampling distribution gets smaller as sample size increases
- The shape of the sampling distribution becomes normal as the sample size increases

Subsequent weeks: Our goal is to infer something about the population from the samples we have, CLT will be a very important foundation for these goals!

Demo: https://seeing-theory.brown.edu/probability-distributions/index.html

Sidenote: mathematical benefits of sampling

What proportion of people are between 170cm and 180cm?

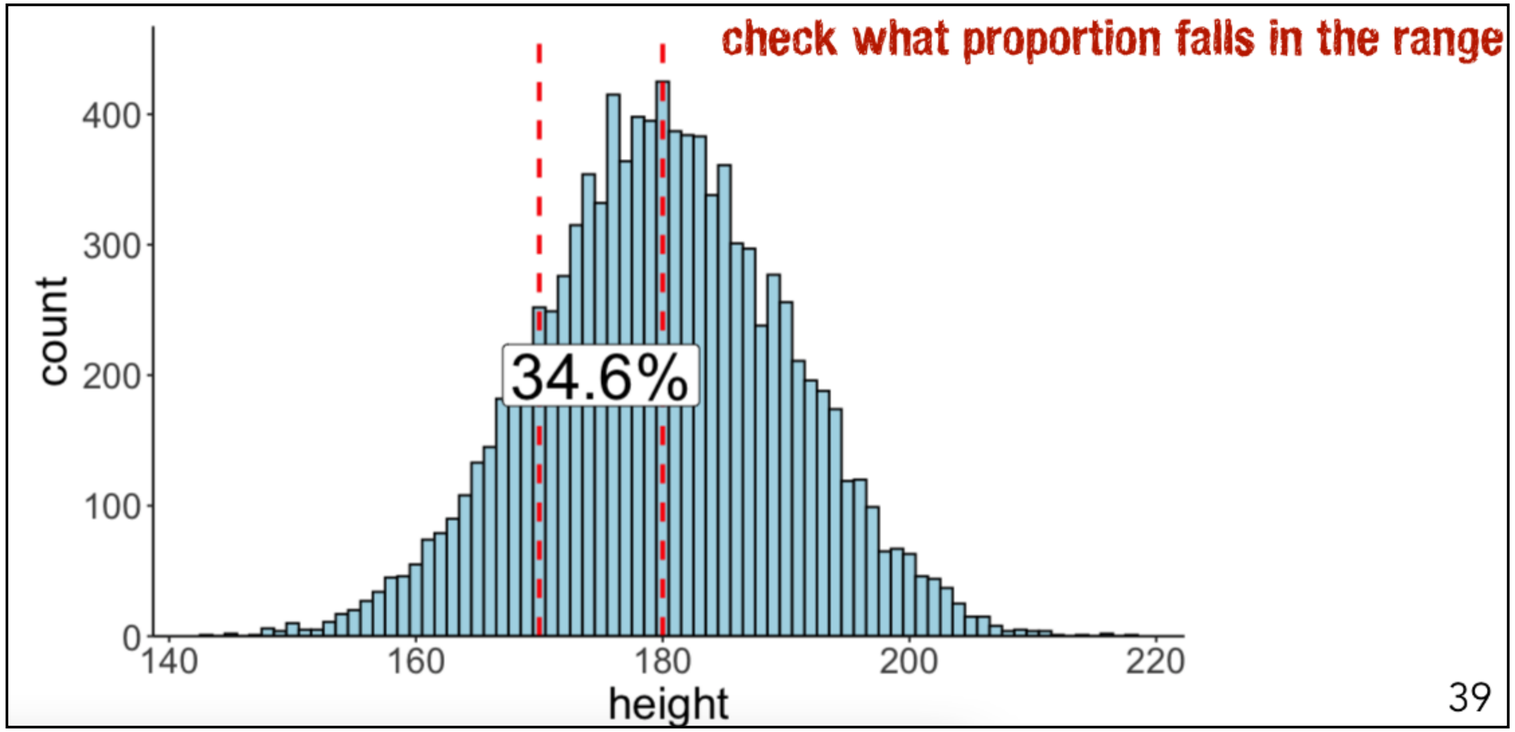


Analytical solution can be a quite nasty and involved..

Sidenote: mathematical benefits of sampling

What proportion of people are between 170cm and 180cm?

Sampling solution is relatively straightforward…



Next week: Drawing inferences about the population