Sampling Distributions

What do we want?

We want to make inferences about a population

- But the population is too large to measure directly
- So we need to estimate the population parameters

Population

- The population distribution is a *theoretical probability distribution* that has some mathematical form
- We want to know how the population distribution influences the distribution of **sample** statistics
- Ultimately we want to use a sample distribution to understand the population distribution

What is the *point* of inferential stats?

Point estimation: we use our sample statistics to take our best guess of the population parameter

We know that our estimates will vary from sample to sample

We're using our sample as an estimate

- Sample mean $\bar{X}$ is an ***estimator***
- A specific sample is an ***estimate***

Population vs. Sample

	Population Distribution	**Sample Distribution**
Distribution consists of:	Individual observations x	Individual observations x
Central tendency	μ	$\bar{x}$
Dispersion	σ^2	s^2
	σ	s
Type	Parameter	Statistic
T vs. O	Theoretical	Observed

Sampling Distribution

- The major goal that we have in statistical inference is to make confident claims about the *population* based on a small representation of it, the *sample*.
- Any sample will be off the mark in how well it captures the important features of a population. The **sampling distribution** tells us how far off the mark we can expect a sample statistic to be.

Sampling Distribution

- We use features of the sample (*statistics*) to inform us about features of the population (*parameters*).
- The quality of this information goes up as sample size goes up -- **the Law of Large Numbers**.

All sample statistics are wrong, but they become more useful as sample size increases.

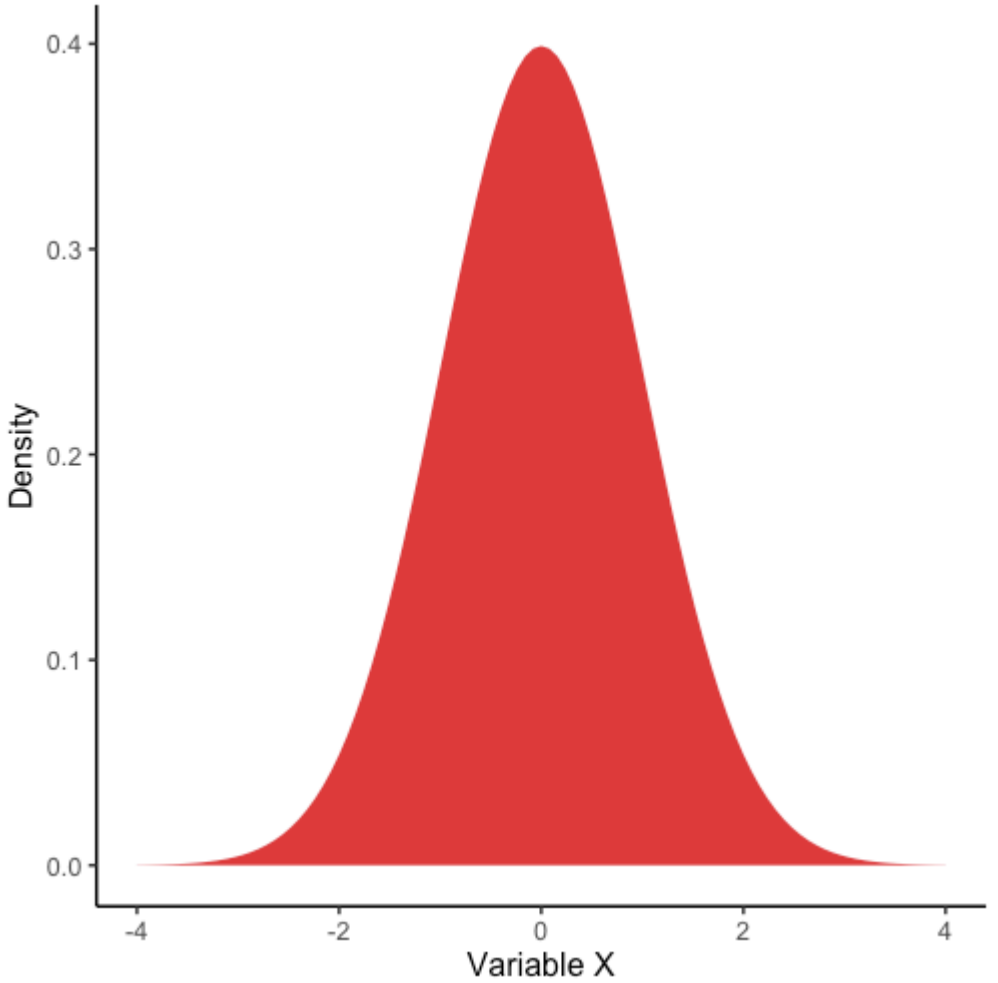
So...

- how big does the sample need to be?
- for a given sample size, how precise is a sample statistic as a representation of the population?

Population distribution

The parameters of this distribution are unknown.

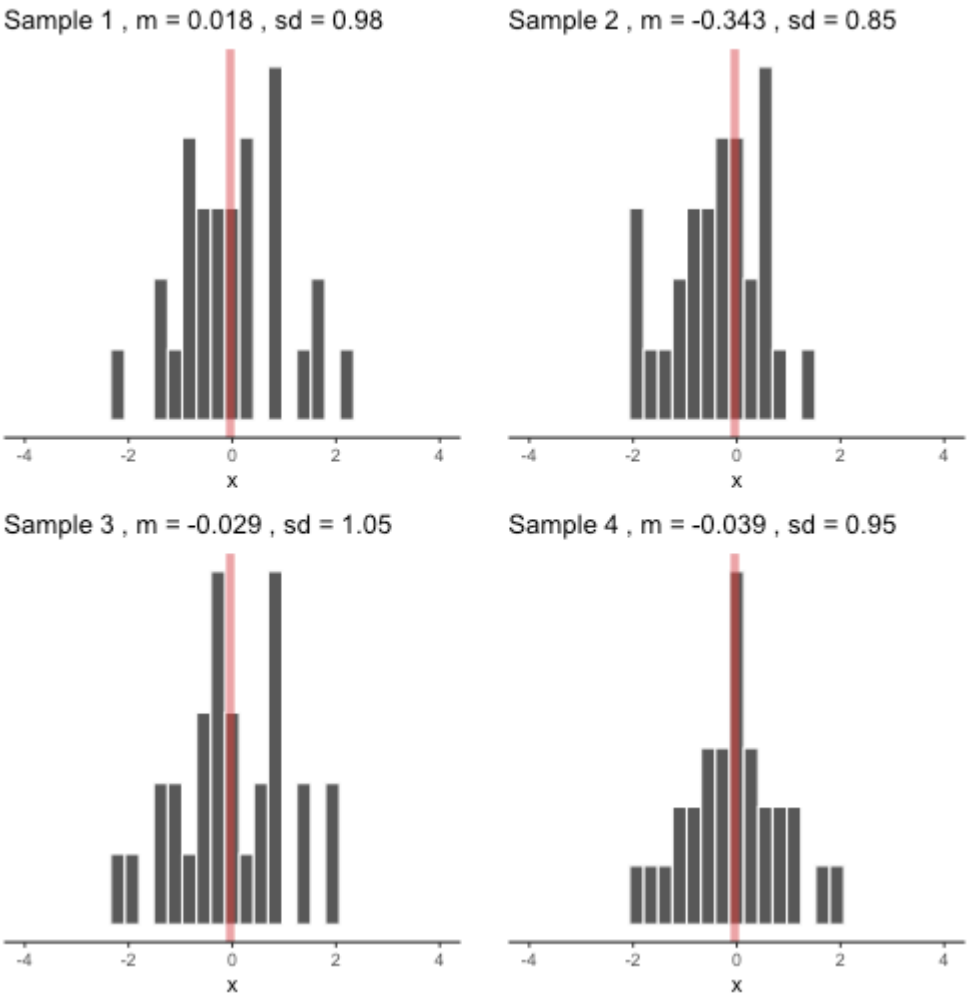
We use the sample to inform us about the likely characteristics of the population.



Samples from the population

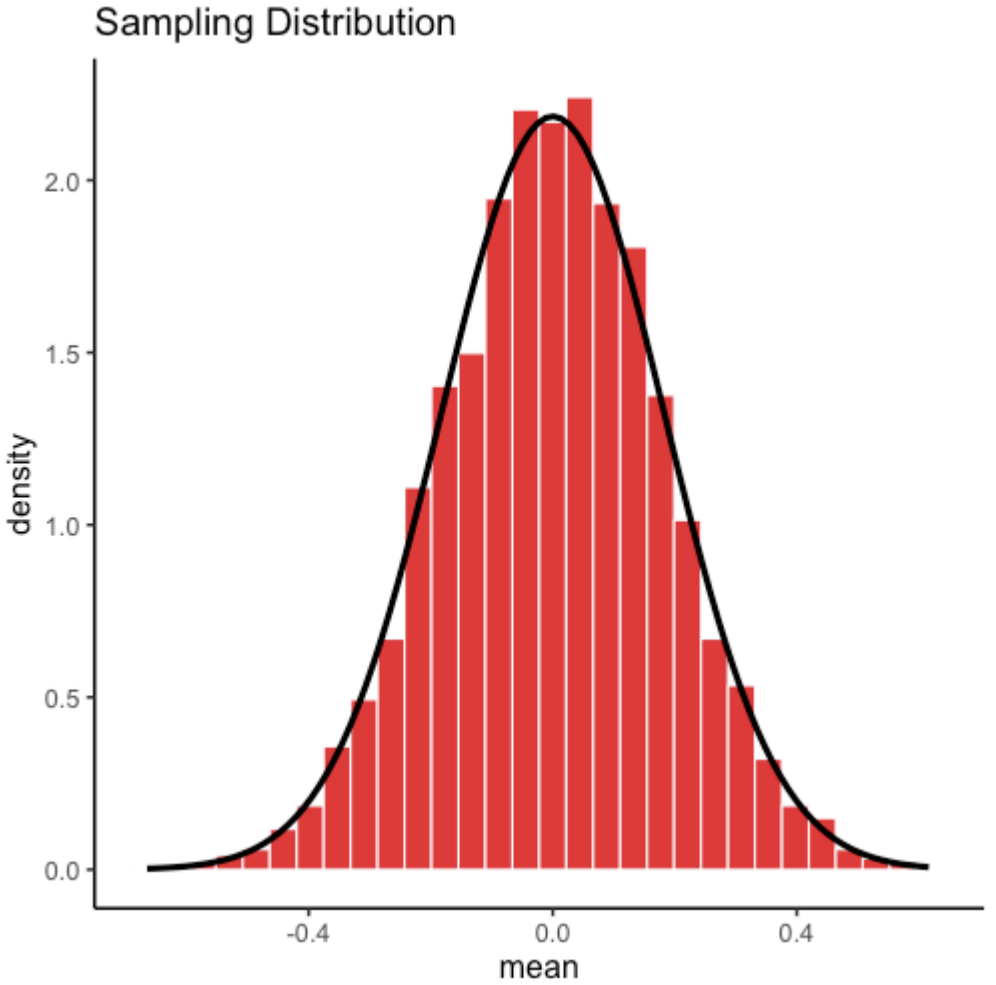
Each *sample distribution* will resemble the population. That resemblance will be better as sample size increases: The Law of Large Numbers.

Statistics (e.g., mean) can be calculated for any sample.



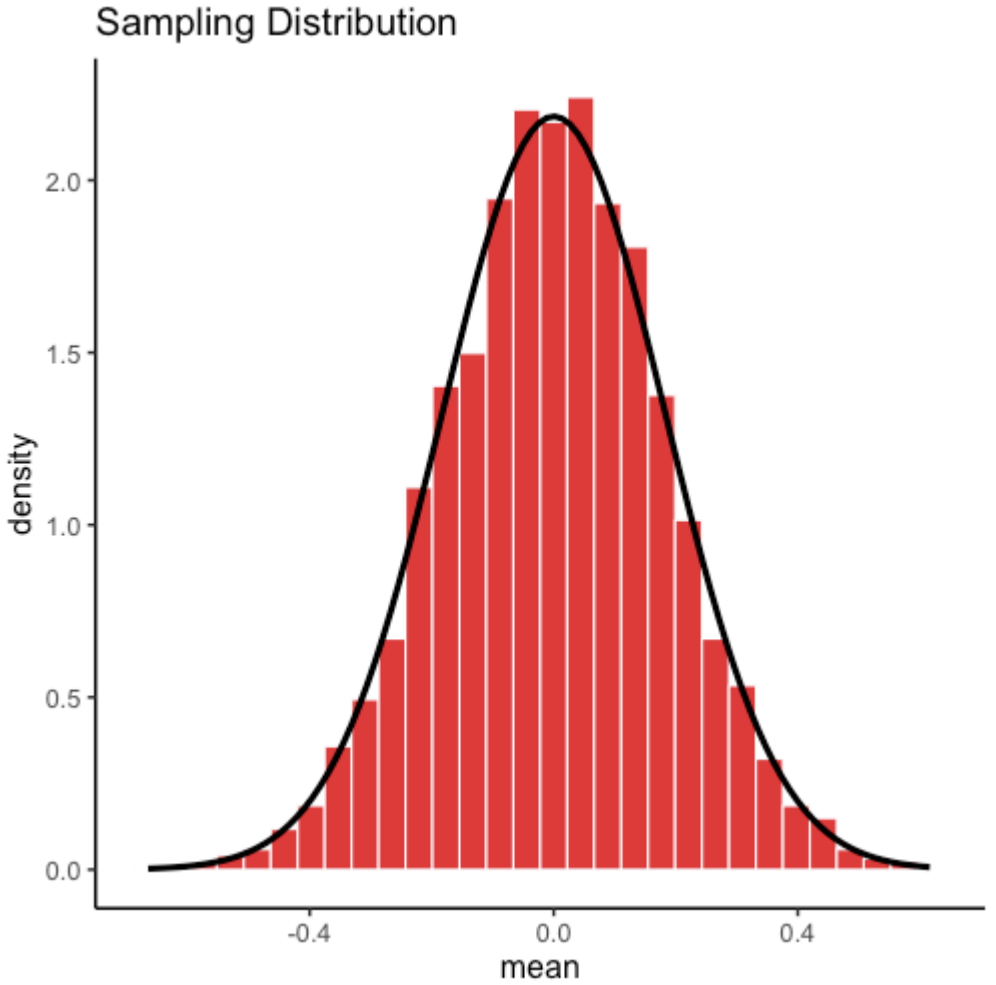
A statistic (e.g., mean) from a large number of samples also has a distribution: the **sampling distribution**.

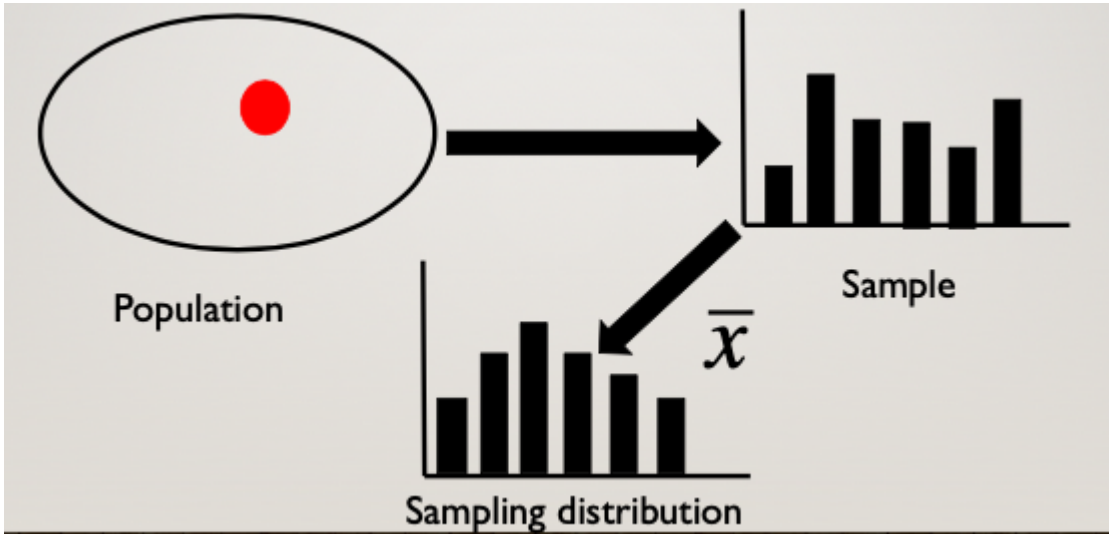
The mean of the **sampling distribution** converges on μ



This distribution has a standard deviation that tells us how typical or rare values of the sample statistic are likely to be.

The sampling distribution of the mean is of particular interest, it's called the **standard error of the mean** (SEM).




Population
Sample
$\overline{x}$
Sampling distribution

Sampling Distributions

- Distribution of values of a particular statistic ($\bar{x}$, s^2, s) across all possible samples of N observations
  - To keep it simple, let's just focus on the mean
  - Statistic will be our *estimator* of the population
- **Sampling distribution $\neq$ sample distribution**

Sampling Distributions

A theoretical probability distribution of all possible values of some statistic, computed from samples of the same size randomly drawn from the same population

Provides the frequency/probability with which values of statistics are observed or are expected to be observed when random samples of size N are drawn from a given population

Interactive Example

PLAY WITH THIS!

Sampling distribution example

Sampling

Part 2

Statistical Inference

We want to learn from incomplete or imperfect data.

1. *Sampling model*: learn something about the population, which we need to estimate from a sample
2. *Measurement error model*: learn about some underlying pattern, but the there is measurement error
3. *Model error*: "inevitable imperfections of models that we apply to real data"

Regression and Other Stories, Gelman, Hill, & Vehtari

Sampling Distribution

We can have a sampling distribution of data. All of the possible datasets that could have been observed if you ran the study a gajillion times.

This is **generative**. If you knew the underlying process, you could create (or generate) data.

We could have collected the mean for each of those gajillion times.

If we plot those gajillion means, then we have a sampling distribution of an **estimate**.

Sampling Distribution Approximates the Normal

One of the most important discoveries in statistics is that the sampling distributions of many statistics are approximately **normal** even when the sample (and population) distributions are not.

- For example, the mean of a random sample will not precisely equal the population mean. But, how far off will it be? And what distribution shape will these possible sample mean values have?

The error represented by how far off a sample mean is from the population mean is called **sampling error**.

According to the **central limit theorem**, as sample size increases, the *sampling* distribution of the mean approaches normality, even when the OG data are not normally distributed.

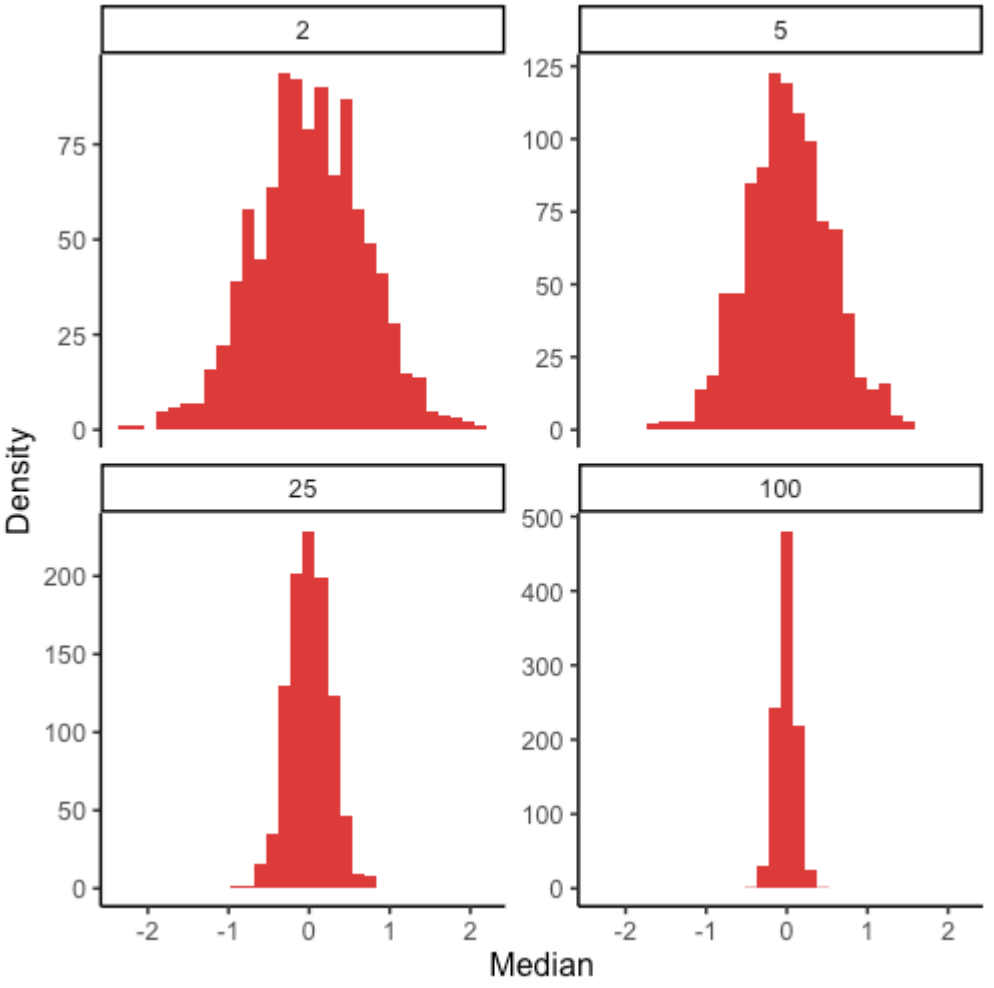
The sample size necessary to be "approximately normal" depends on the nature of the underlying data. The less normal it is, the larger the sample size necessary in order for the sampling distribution of the means to become normal.

"Around sample size of 30" is a common rule of thumb.

- Note, however, that this rule of thumb is sufficiently only to assume that the sampling distribution is normal. It doesn't magically mean your sample will be normally distributed

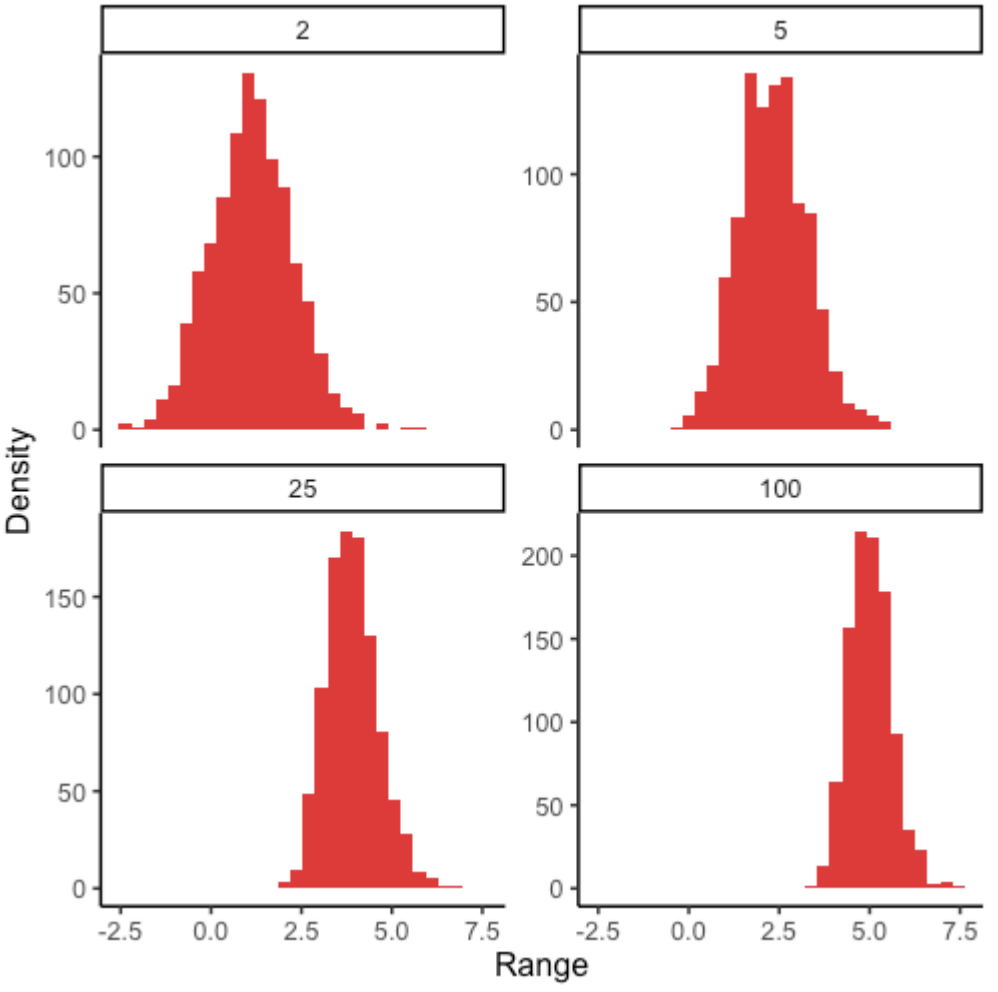
Quite a few sample statistics approach normality as sample size increases.

Here is the sample median from a normal distribution with $\sigma = 1$.



And the range.

Note that the sampling distribution of σ^2 is not normally distributed. It is χ^2 distributed. More on this later!



Relationship Between Population & Sampling

- If the population is normally distributed, the sampling distribution of the mean will be normally distributed
- If the population distribution is not normally distributed, the sampling distribution of the mean will become increasingly normally distributed as sample size increases
- We can use the normal distribution to make inferences about the unknown population mean, based on the sample mean and sample standard deviation

Sampling Mean = Population Mean

$$E(\bar{x}) = \mu$$

Expected value is the long run average

Sampling Mean = Population Mean

$$\bar{x} = \frac{(x_1 + x_2 + \ldots x_N)}{N}$$

$$E(\bar{x}) = E\left(\frac{(x_1 + x_2 + \ldots x_N)}{N}\right)$$

$$E(\bar{x}) = \frac{E(x_1) + E(x_2) + \ldots E(x_N)}{E(N)}$$

$$E(\bar{x}) = \frac{\mu + \mu + \ldots \mu}{N}$$

$$E(\bar{x}) = \frac{N\mu}{N}$$

$$E(\bar{x}) = \mu$$

Expected Value Variance

You can also have a "long run" or expected variance, too. Equivalent to the population variance.

$$\sigma = E(\bar{x}^2) - (E(\bar{x}))^2$$

First term:

- square your x (each score is squared)
- multiply each value by probability (if 5 scores, 1/5)
- take the sum

Second term:

- get the mean of x
- square it

Sampling SD $\neq$ Population SD

$$\sigma_M^2 = E(\bar{x}^2) - (E(\bar{x}))^2$$

[insert long proof here...]

$$E(\bar{x}^2) = \frac{\sigma_M^2}{N} + \mu^2$$

Variability in the Sampling Distribution

Standard Error of the Mean

- The standard deviation of the sampling distribution
- Gives us a sense of **uncertainty** about mean*
- Directly related to the variability of the underlying data:

$$\sigma_m = \frac{\sigma_X}{\sqrt{N}}$$

- The smaller the SEM, the more likely it is that your sample estimate of the mean will be closer to the population estimate of the mean
- SEM is a function of sample size! The more accurate your sample mean, the closer you are to approximating the population, and the smaller the standard error

More SEM

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

This is an unbiased estimate of σ (why?) and relies on the sample mean, which is an unbiased estimate of μ.

This is different from the uncorrected sample standard deviation, which divides the sum of squares by N rather than $N - 1$.

$$SEM = \sigma_M = \frac{\hat{\sigma}}{\sqrt{N}} = \frac{\text{Estimate of pop SD}}{\sqrt{N}}$$

(Most methods of calculating standard deviation assume you're estimating the population from a sample and correct for bias.)

Making Statements

The sampling distribution of means can be used to make probabilistic statements about means in the same way that the standard normal distribution is used to make probabilistic statements about scores.

For example, we can determine the range within which the population mean is likely to be with a particular level of confidence.

Or, we can propose different values for the population mean and ask how typical or rare the sample mean would be if that population value were true. We can then compare the plausibility of different such “models” of the population.

Confidence Intervals

The sampling distribution of the mean has variability, represented by the SEM, **reflecting uncertainty in the sample mean as an estimate of the population mean**.

The assumption of normality allows us to construct an interval within which we have good reason to believe a sample mean will fall if it comes from a particular population:

$$\mu - (1.96 \times SEM) \leq \bar{X} \leq \mu + (1.96 \times SEM)$$

We can reorganize this to allow statements about the population mean, which is usually not known:

$$\bar{X} - (1.96 \times SEM) \leq \mu \leq \bar{X} + (1.96 \times SEM)$$

Confidence Intervals

$$\bar{X} - (1.96 \times SEM) \leq \mu \leq \bar{X} + (1.96 \times SEM)$$

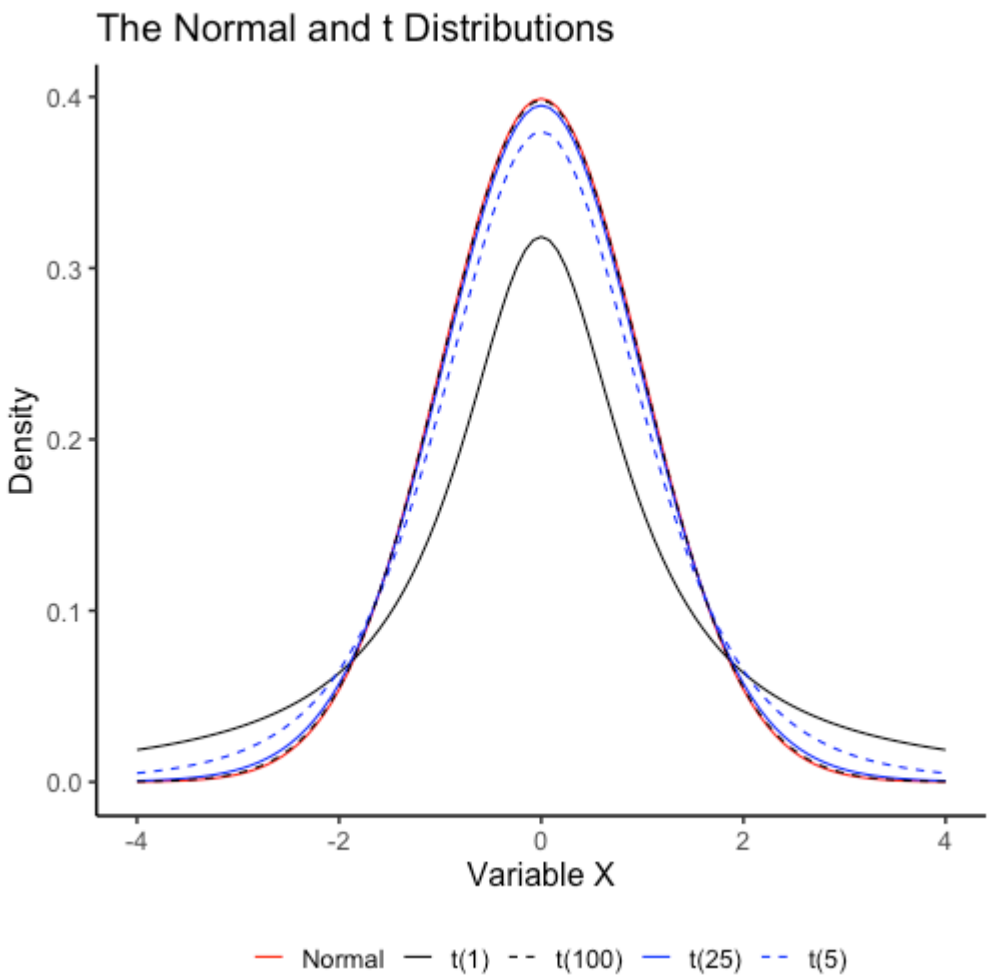
- This is referred to as the **95% confidence interval (CI)**
- Note the assumption of normality, which should hold by the Central Limit Theorem, if N is sufficiently large.

The 95% CI is sometimes represented as:

$$CI_{95} = \bar{X} \pm [1.96 \frac{\hat{\sigma}}{\sqrt{N}}]$$

The normal assumes we know the μ and σ. But we don't. We only know $\bar{x}$ and s, and those have some uncertainty about them.

This is reduced with large samples, so that the normal is "close enough." In small samples, the t dist has a better approximation.



t distribution

- The primary difference between the normal distribution and the t distribution is the fatter tails
  - This produces wider confidence intervals
  - The penalty we have to pay for our ignorance about the population
- The form of the confidence interval remains the same. We simply substitute a corresponding value from the t distribution (using df = $N - 1$).

$$CI_{95} = \bar{X} \pm [1.96 \frac{\hat{\sigma}}{\sqrt{N}}]$$

$$CI_{95} = \bar{X} \pm [t_{.975, df=N-1} \frac{\hat{\sigma}}{\sqrt{N}}]$$

Confidence Intervals

What does it NOT mean?

- there is a 95% probability that the true mean lies inside the confidence interval

What it *actually* means:

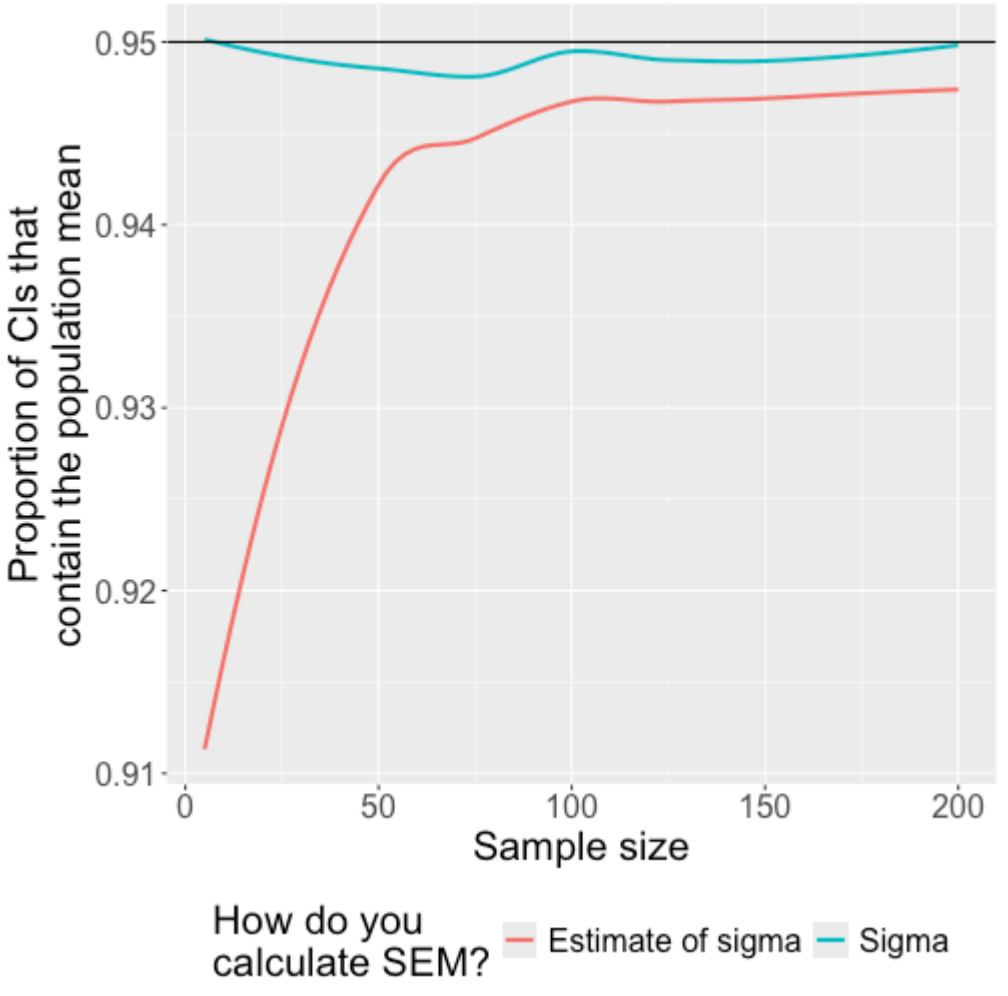
- If we carried out random sampling from the population a large number of times...
- and calculated the 95% confidence interval each time...
- then 95% of those intervals can be expected to contain the population mean.

Simulation

At each sample size, draw 5000 samples from known population ($\mu = 0$, $\sigma = 1$).

Calculate CI for each sample using s and record whether or not 0 was in that interval.

Calculate CI using for each sample using σ.



A better simulation...

https://rpsychologist.com/d3/ci/

In the past, my classroom exams (aggregating over many classes) have a mean of 90 and a standard deviation of 8.

My next class will have 100 students. What range of exam means would be plausible if this class is similar to past classes (comes from the same population)?

```
M = 90
SD = 8
N = 100

sem = SD/sqrt(N)

ci_lb_z = M - sem * qnorm(p = .975)
ci_ub_z = M + sem * qnorm(p = .975)
print(c(ci_lb_z, ci_ub_z))
```

```
## [1] 88.43203 91.56797
```

```
ci_lb_z = M - sem * qt(p = .975, df = N-1)
ci_ub_z = M + sem * qt(p = .975, df = N-1)
print(c(ci_lb_z, ci_ub_z))
```

```
## [1] 88.41263 91.58737
```

I give a classroom exam that produces a mean of 83.4 and a standard deviation of 10.6. A total of 26 students took the exam.

What is the 95% confidence interval around the mean?

```
M = 83.4
SD = 10.6
N = 26

sem = SD/sqrt(N)

ci_lb_z = M - sem * qnorm(p = .975)
ci_ub_z = M + sem * qnorm(p = .975)
print(c(ci_lb_z, ci_ub_z))
```

```
## [1] 79.32557 87.47443
```

```
ci_lb_z = M - sem * qt(p = .975, df = N-1)
ci_ub_z = M + sem * qt(p = .975, df = N-1)
print(c(ci_lb_z, ci_ub_z))
```

```
## [1] 79.11857 87.68143
```

	Population Distribution	Sample Distribution	Sampling Distribution
Distribution consists of:	Individual observations x	Individual observations x	Statistics $\bar{x}, s, s^2$
Central tendency	μ	$\bar{x}$	μ_M
Dispersion	σ^2	s^2	σ_M^2
	σ	s	SEM σ_M
Type	Parameter	Statistic	Statistic of statistics
T vs. O	Theoretical	Observed	Theoretical

Next time...

Exam 1 Review with Ran

Exam 1

(ahhhhhhhhhhh!)