Sampling Distributions

Recap

Normal distributions all have well-characterized properties

- AUC = 1
- ~68% fall within 1 σ, ~95% within 2 σ, and ~99.7% within 3 σ

The standard normal distribution is a particular type of normal distribution

- distribution of z-scores
- $\mu = 0$
- $\sigma = 1$

Using the standard normal & these cool properties, we can make probability statements

- What is the probability of getting a z-score or more extreme?

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
- 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*** of μ
- A specific sample is an ***estiamte***

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 tell us about features of the population (*parameters*).

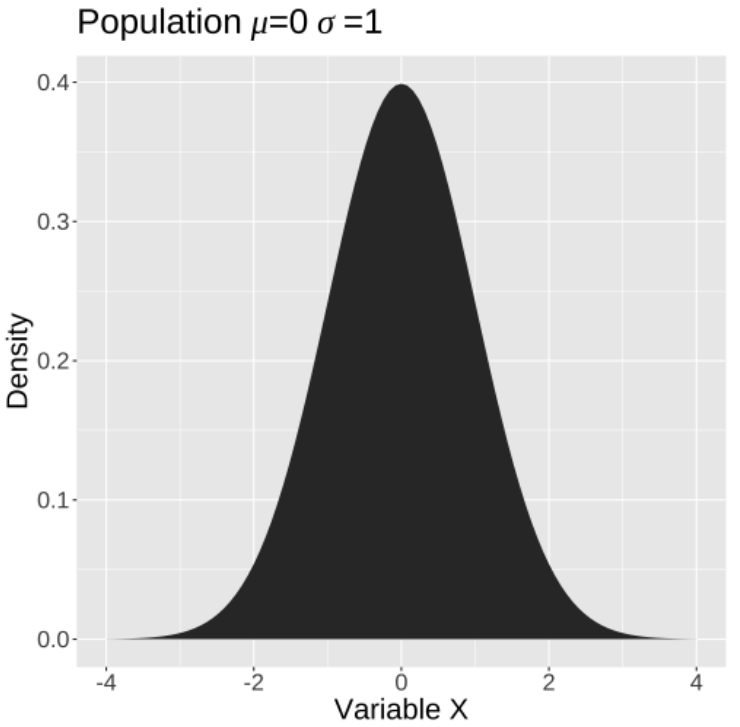
The quality of this information goes up as sample size goes up

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

Sampling Distribution

The parameters of this distribution are unknown.

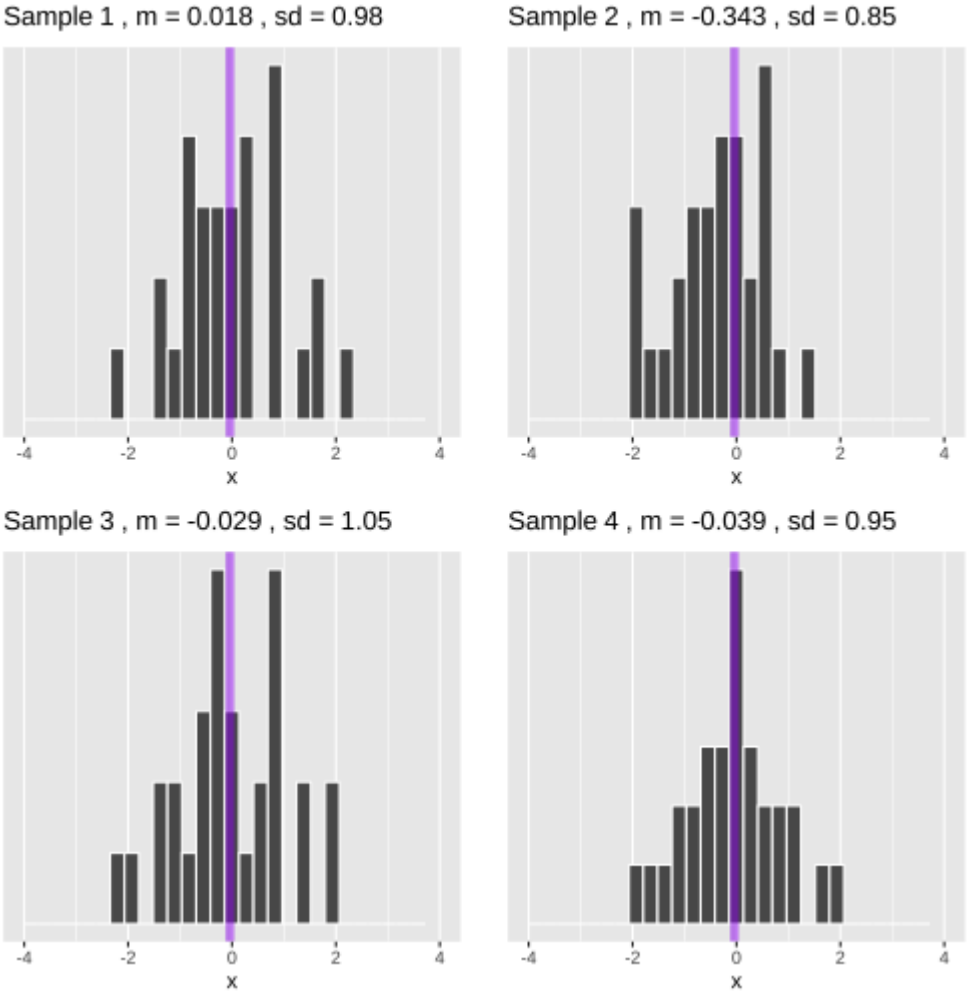
What can we do? We can use the sample to inform us about the likely characteristics of the population.



Samples from the population

Each *sample distribution* will resemble the population. That resemblance will be better as sample size increases.

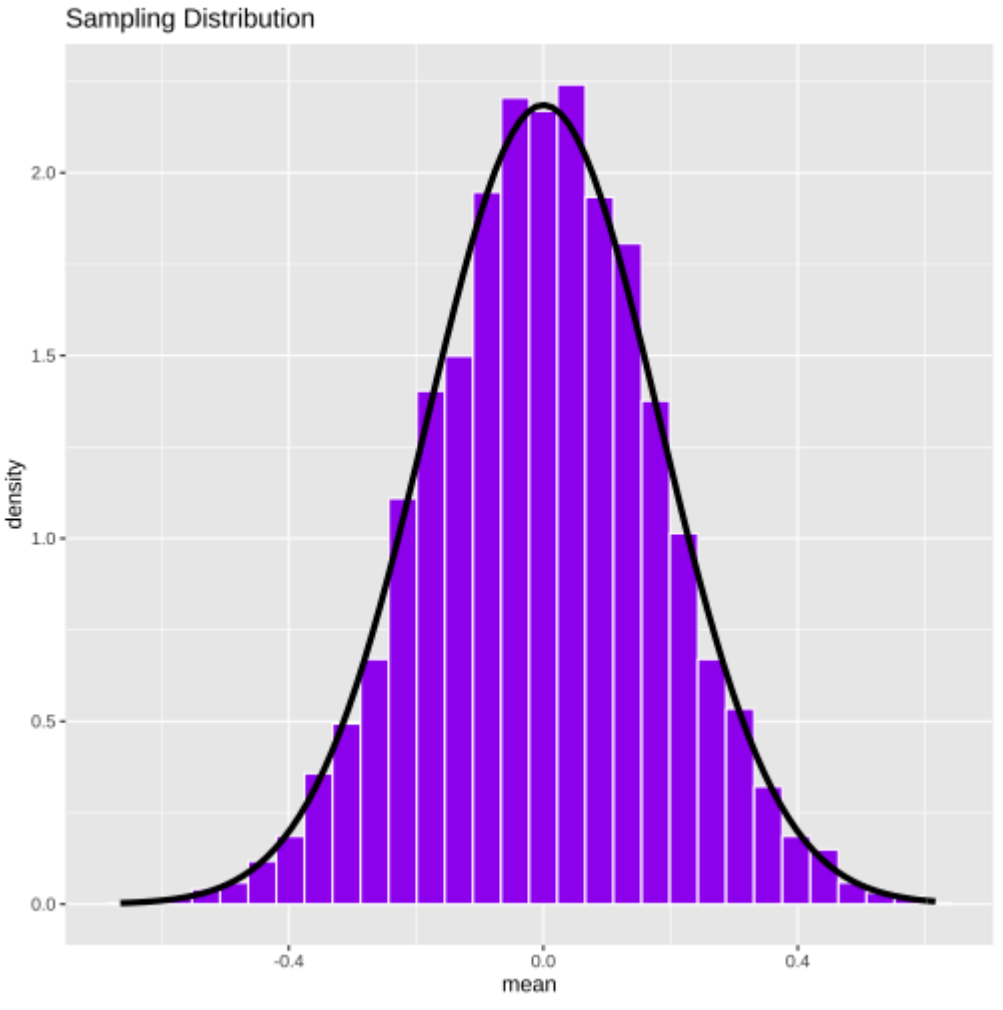
Statistics (e.g., mean) can be calculated for any sample.



Say you repeated an experiment 100 times, each time using a new sample. Each sample has it's own mean (and other statistics).

That's 100 means. You can have a distribution of means rather than of scores. That's a sampling distribution! **Distribution of statistics**

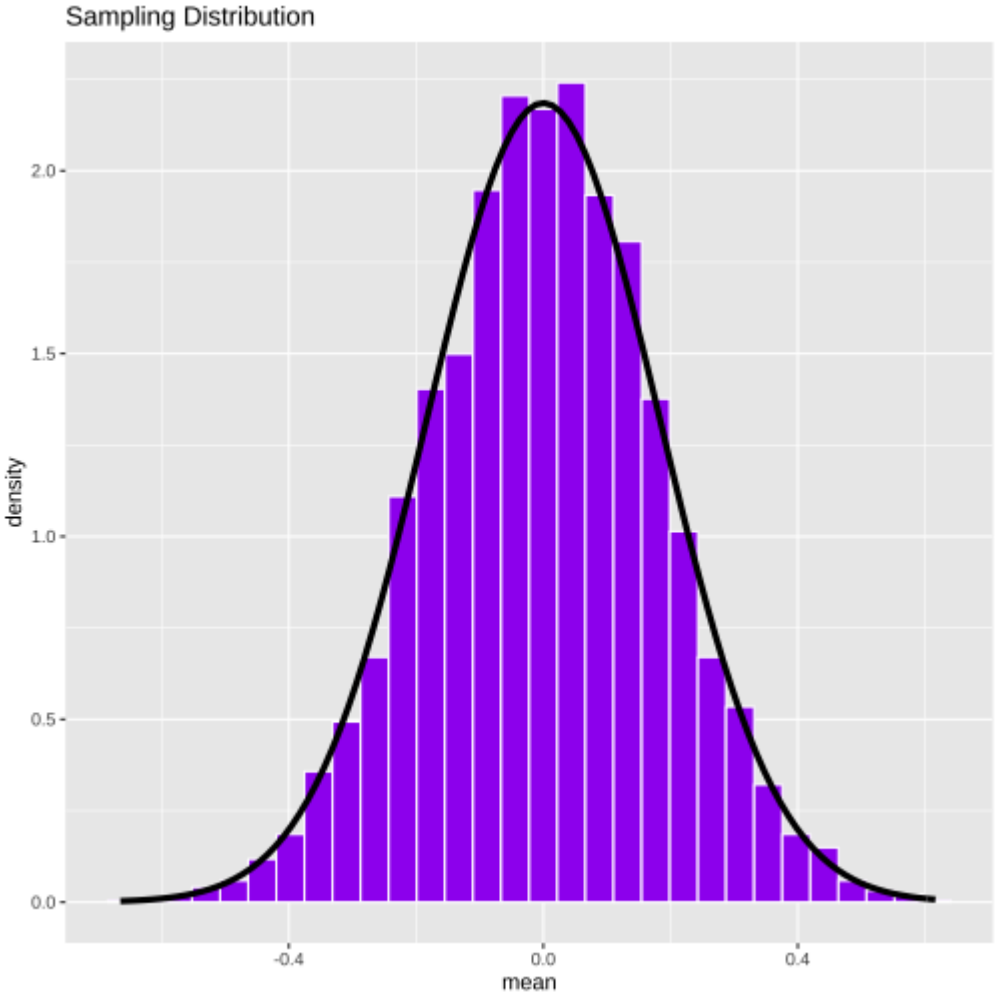
The mean of the **sampling distribution** converges on population mean, μ



Sampling Distribution

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).



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**
- 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 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.

The mean of a random sample will not precisely equal the population mean. But, how far off will it be?

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

Central Limit Theorem

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

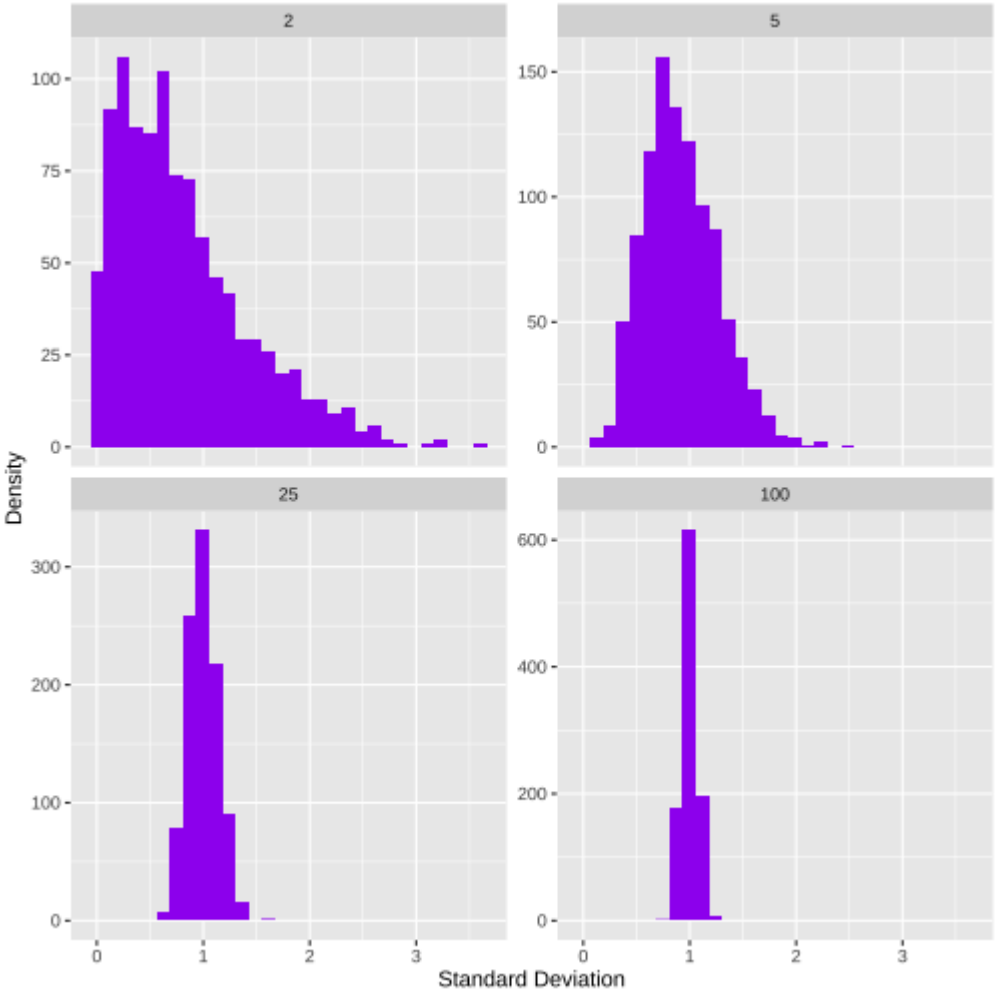
The sample size necessary to be "approximately normal" depends on the nature of the underlying data. The less normal it is, the larger the sample size necessary in order for the sampling distribution of the means to become normal.

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

Central Limit Theorem

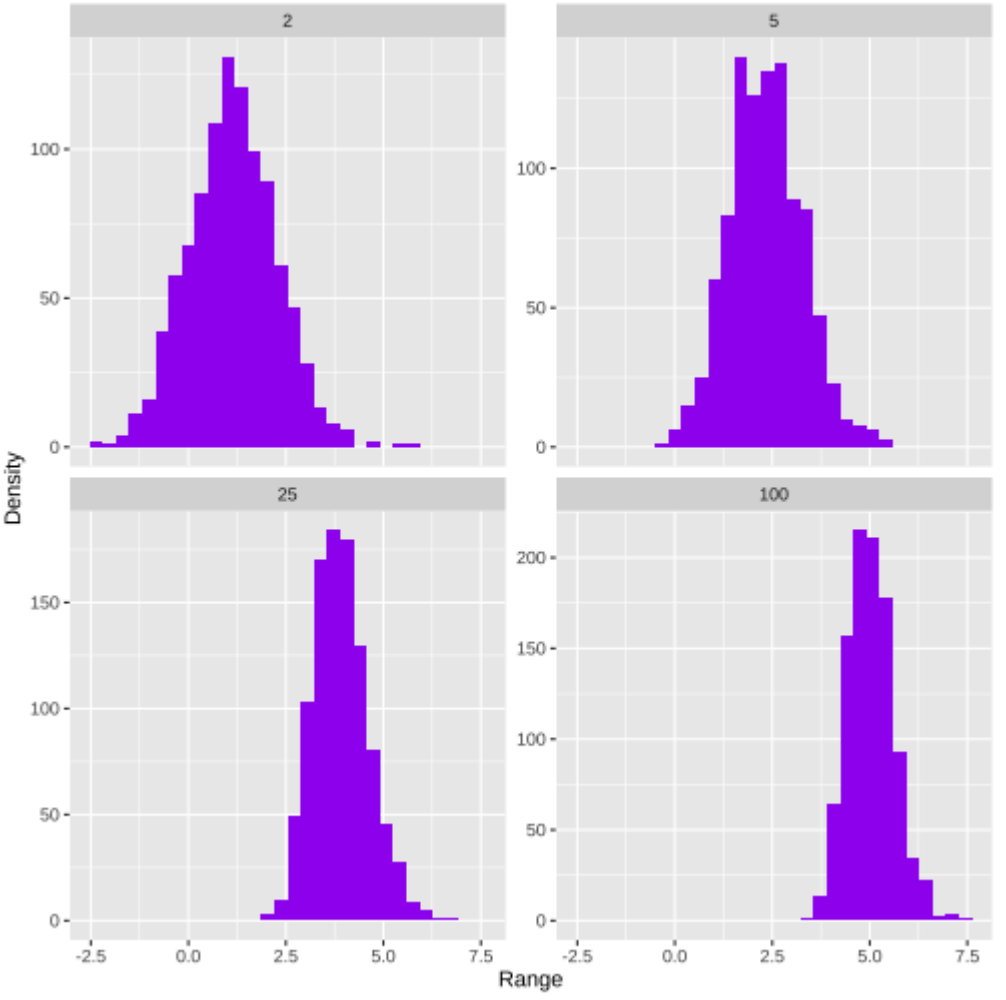
Ends up that quite a few sample statistics approach normality as sample size increases.

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



Central Limit Theorem

And the range.



Relationship Between Population & Sampling

- If the population is normally distributed, the sampling distribution of the mean will be normally disributed
- If the population distribution is not normally distributed, the sampling distribution of the mean will become increasingly normally disributed 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

Mean of the Sampling Distribution

The mean of the sampling distribution converges on the population mean, μ

Our sample mean is not biased; we'll be a little wrong, but it'll be OK

Variability of the Sampling Distribution

Standard Error of the Mean

- The standard deviation of the sampling distribution
- 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 the sample estimate of the population standard deviation. This is an unbiased estimate of σ and relies on the sample mean, which is an unbiased estimate of μ.

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

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

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 population mean will fall:

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

Confidence Intervals

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

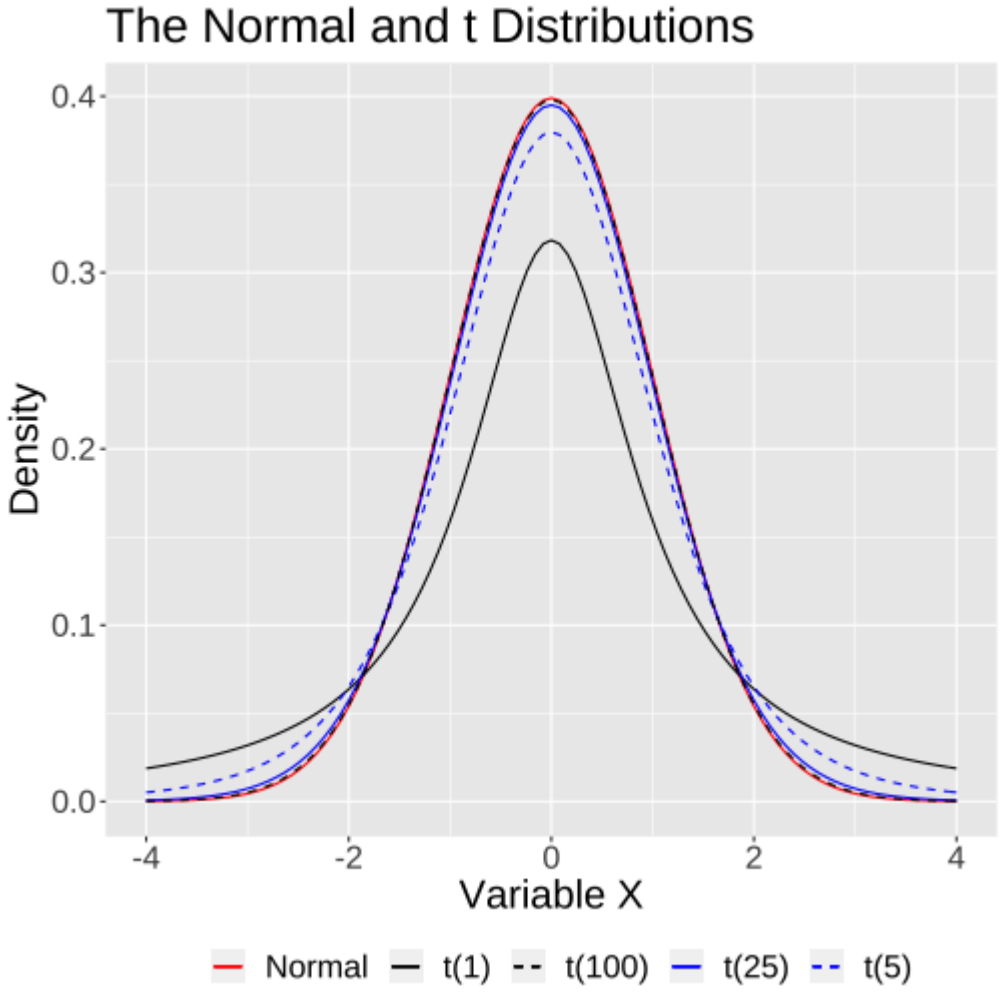
- This is referred to as the **95% confidence interval (CI)**
- The 95% CI is sometimes represented as:

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

The t

The normal distribution assumes we know the population mean and standard deviation. But we don't. We only know the *sample* mean and standard deviation, and those have some uncertainty about them.

That uncertainty is reduced with large samples, so that it's "close enough" to the normal. In small samples, the t distribution is better.



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:

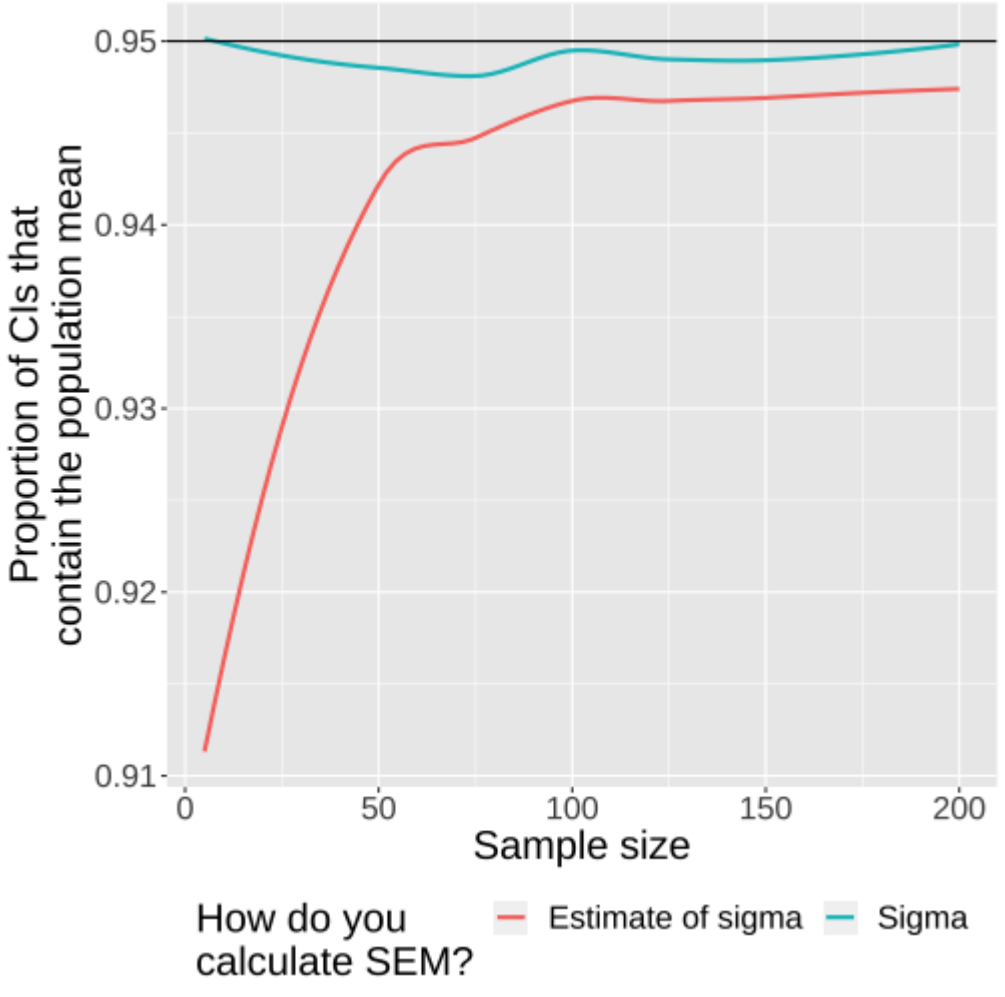
- If we carried out random sampling from the population a large number of times...
- and calculated the 95% confidence interval each time...
- then 95% of those intervals can be expected to contain the population mean.

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.



Examples

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
```

Examples

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
```

All Together

	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