Sampling Distributions

The most important distinction in statistics

sample
vs
population

When analyzing data (or reading the literature), think about whether you want to discuss the sample that you observed or want to make statements that are more generally true.

Statistics is the only field that gives us the correct framework to generalize from our sample to the population.

Sample vs. Population

Example: T cell counts from 40 women with triple negative breast cancer were observed. What can we do with this information?

Option 1: Discuss the 40 women. What was the mean T cell count? What was its variation?

Option 2: Generalize the information about the 40 women to make statements about all women with triple negative breast cancer.

These are 2 different approaches to using the same information.

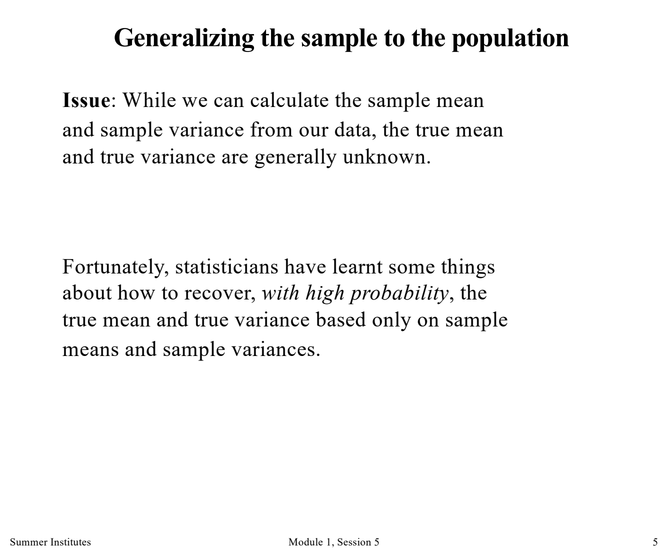
Language for making these distinctions

Sample

- Size = n
- Sample Mean $\overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$
- Sample variance = $s^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(X_j - \overline{X}\right)^2$

Population

- Size = N (usually ∞)
- Mean $\mu = \sum p_j X_j \quad or \quad \int\ldots$
- Variance $\sigma^2 = \sum p_j \left(X_j - \mu\right)^2 \quad or \quad \int\ldots$

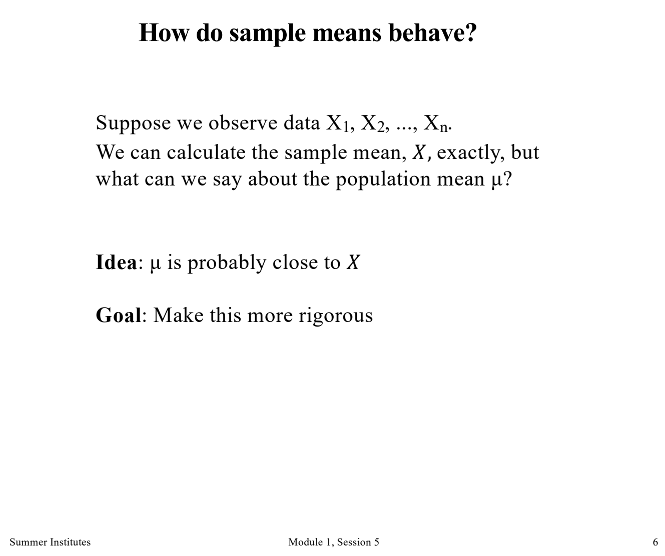


Generalizing the sample to the population

Issue: While we can calculate the sample mean and sample variance from our data, the true mean and true variance are generally unknown.

Fortunately, statisticians have learnt some things about how to recover, *with high probability*, the true mean and true variance based only on sample means and sample variances.

5



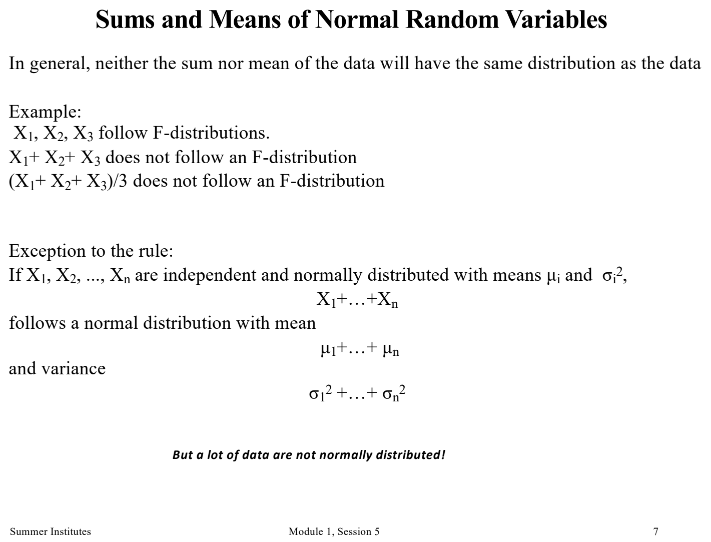
How do sample means behave?

Suppose we observe data $X_1, X_2, ..., X_n$.
We can calculate the sample mean, X, exactly, but what can we say about the population mean μ?

Idea: μ is probably close to X

Goal: Make this more rigorous

6



Sums and Means of Normal Random Variables

In general, neither the sum nor mean of the data will have the same distribution as the data

Example:
X_1, X_2, X_3 follow F-distributions.
$X_1+ X_2+ X_3$ does not follow an F-distribution
$(X_1+ X_2+ X_3)/3$ does not follow an F-distribution

Exception to the rule:
If $X_1, X_2, ..., X_n$ are independent and normally distributed with means μ_i and σ_i^2,

$$X_1+\ldots+X_n$$

follows a normal distribution with mean

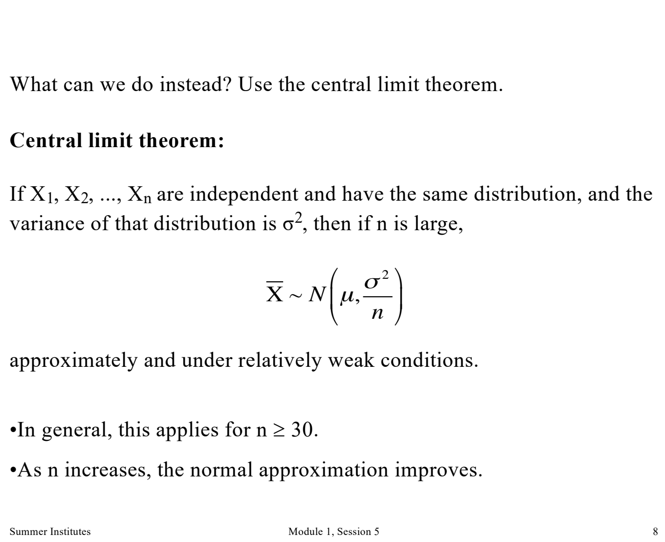
$$\mu_1+\ldots+\mu_n$$

and variance

$$\sigma_1^2+\ldots+\sigma_n^2$$

But a lot of data are not normally distributed!

7



What can we do instead? Use the central limit theorem.

Central limit theorem:

If $X_1, X_2, ..., X_n$ are independent and have the same distribution, and the variance of that distribution is σ^2, then if n is large,

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

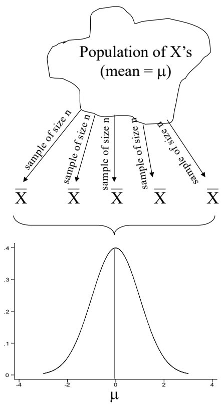
approximately and under relatively weak conditions.

•In general, this applies for $n \geq 30$.

•As n increases, the normal approximation improves.

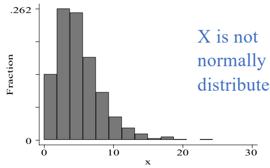
8

Distribution of the Sample Mean

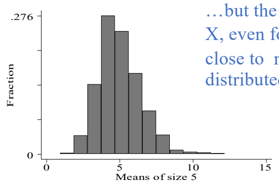


Central Limit Theorem Illustration

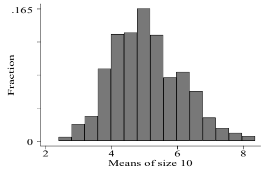
Population

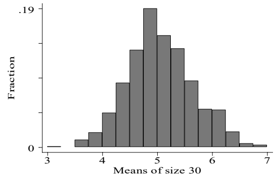


X is not normally distributed…



…but the means of X, even for n=5, are close to normally distributed





…and the means of X for n=10 and 30 become closer and closer to normally distributed

Central limit theorem

The central limit theorem allows us to use the sample ($X_1 \ldots X_n$) to discuss the population (μ)

We do not need to know the distribution of the data to make statements about the true mean of the population!

Distribution of the Sample Mean

Example:

Suppose that for sixth grade students in Seattle, the mean number of missed school days is 5.4 days with a standard deviation of 2.8 days.

What is the probability that a random sample of size 49 will have a mean number of missed days greater than 6 days?

Find the probability that a random sample of size 49 from the population of Seattle sixth graders will have a mean greater than 6 days.

μ = 5.4 days

σ = 2.8 days

n = 49

$$\sigma_{\overline{X}} = \sigma / \sqrt{n} = 2.8 / \sqrt{49} = 0.4$$

$$\mu_{\overline{X}} = 5.4$$

$$P(\overline{X} > 6) = P\left(\frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} > \frac{6 - 5.4}{0.4} \right)$$

$$= P(Z > 1.5) = 0.0668$$

$$= 0.00668$$

Pause- break time then work on exercises

Question 1

What is the probability that a random sample (size 49) from this population of Seattle sixth graders has a mean between 4 and 6 days?

Recall: μ = 5.4 days, σ = 2.8 days

Confidence Intervals

Confidence Intervals

"(L, U) is a 95% confidence interval for a parameter θ"

means that

- The true parameter θ has probability at least 0.95 of being between L and U.
- In repeated samples, 95% of the resulting confidence intervals would contain θ."

We calculate L and U from our data to get an interval estimate of θ, an idea of its plausible values.

Note: Confidence intervals are about parameters. Prediction intervals (different!) are intervals about random variables.

95% Confidence Interval for the Mean

Because $\overline{\mathrm{X}} \sim N\left(\mu, \frac{\sigma^2}{n}\right),$

we know that $P\left[-1.96 \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le +1.96\right] = 0.95.$

Rearranging gives us that $\left(\overline{X} - 1.96 \times \frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96 \times \frac{\sigma}{\sqrt{n}}\right)$

is a 95% confidence interval for the true mean μ

(1- α) Confidence Interval for the Mean

If we want a (1 - α) confidence interval we can derive it based on the statement

$$P\left[Q_Z^{\left(\frac{\alpha}{2}\right)} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < Q_Z^{\left(1-\frac{\alpha}{2}\right)}\right] = 1-\alpha$$

That is, we find constants $Q_Z^{\left(\frac{\alpha}{2}\right)}$ and $Q_Z^{\left(1-\frac{\alpha}{2}\right)}$ that have exactly (1 - α) probability between them.

A (1 - α) Confidence Interval for the Population Mean

$$\left(\overline{X} + Q_Z^{\left(\frac{\alpha}{2}\right)} \times \frac{\sigma}{\sqrt{n}}, \overline{X} + Q_Z^{\left(1-\frac{\alpha}{2}\right)} \times \frac{\sigma}{\sqrt{n}}\right)$$

Confidence Intervals Example

Suppose gestational times are normally distributed with a standard deviation of 6 days. A sample of 30 second-time mothers have a mean pregnancy length of 279.5 days. Construct a 95% confidence interval for the mean length of second pregnancies based on this sample.

$$279.5 \pm Q_Z^{0.975} \times \frac{6}{\sqrt{30}}$$

$$279.5 \pm 1.96 \times \frac{6}{\sqrt{30}}$$

(277.35, 281.65)

Confidence intervals when σ unknown:

use t distribution

When σ is unknown we replace it with the estimate, s, and use the t-distribution. The statistic

$$\frac{\overline{X}-\mu}{s/\sqrt{n}}$$

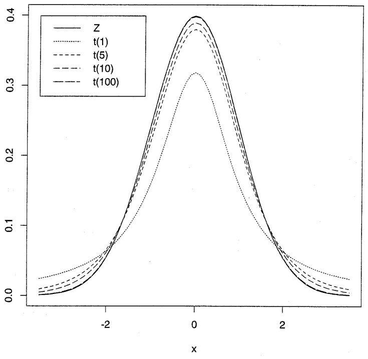
has a t-distribution with n-1 degrees of freedom.

We can use this distribution to obtain a confidence interval for μ even when σ is not known.

A (1-α) Confidence Interval for the Population Mean when σ is unknown

$$\left(\overline{X}+t_{(n-1)}^{\left(\frac{\alpha}{2}\right)}\times s/\sqrt{n},\ \overline{X}+t_{(n-1)}^{\left(1-\frac{\alpha}{2}\right)}\times s/\sqrt{n}\right)$$

Normal and t distributions



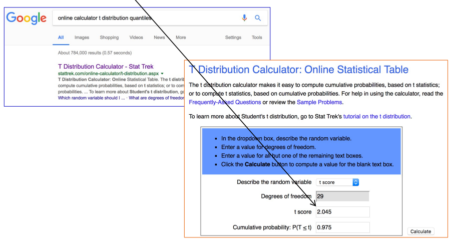
Confidence Intervals for σ² unknown

Example

Given our 30 mothers with a mean gestation of 279.5 days and a variance of 28.3 days², we can compute a 95% confidence interval for the mean length of pregnancies for second-time mothers using the t-distribution:

$$279.5 \pm t_{29}^{0.975} \times \frac{\sqrt{28.3}}{\sqrt{30}}$$

e.g., https://stattrek.com/online-calculator/t-distribution.aspx



online calculator t distribution quantiles

About 784,000 results (0.57 seconds)

T Distribution Calculator - Stat Trek
stattrek.com/online-calculator/t-distribution.aspx
T Distribution Calculator: Online Statistical Table. The t distrib
compute cumulative probabilities, based on t statistics; or to com
probabilities. ... To learn more about Student's t distribution, go
Which random variable should I ... · What are degrees of freedor

T Distribution Calculator: Online Statistical Table

The t distribution calculator makes it easy to compute cumulative probabilities, based on t statistics; or to compute t statistics, based on cumulative probabilities. For help in using the calculator, read the Frequently-Asked Questions or review the Sample Problems.

To learn more about Student's t distribution, go to Stat Trek's tutorial on the t distribution.

- In the dropdown box, describe the random variable.
- Enter a value for degrees of freedom.
- Enter a value for all but one of the remaining text boxes.
- Click the **Calculate** button to compute a value for the blank text box.

Describe the random variable: t score

Degrees of freedom: 29

t score: 2.045

Cumulative probability: P(T ≤ t): 0.975

Calculate

Take Home Points

- General (1 - α) Confidence Intervals:
  - Confidence intervals are only for parameters
  - Greater confidence → wider interval
  - Larger sample size → narrower interval
- CI for true population mean μ when σ assumed known → use a standard normal, Z.
- CI for μ, σ unknown → use a t-distribution.

Pause- break time then work on exercises

Questions 2, 3

Suppose we take a larger sample of second-time mothers. Instead of only sampling n=30, we sample n=45 and the sample mean is the same, 279.5 days.

2: What is a 95% confidence interval for the mean length of second pregnancies based on this larger sample?

3: How does it compare to the 95% CI calculated from the n=30 second-time mothers?