Sampling Distributions

Session 5

Module 1 Probability & Statistical Inference

The Summer Institutes
DEPARTMENT OF BIOSTATISTICS
SCHOOL OF PUBLIC HEALTH
UNIVERSITY of WASHINGTON

The most important distinction in Statistics:

sample vs. **population**

When analyzing data, think about whether you want make statements about the sample or statements that are hold more generally (i.e., for the population).

The field of **Statistics** provides the correct framework to generalize from sample to population.

Sample vs. Population

Example: T cell counts from 40 women with triple negative breast cancer were observed.

Option 1: Summarize the data for these 40 women- report mean T cell count and variance.

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

2 different approaches to using the same information.

Notation for distinguishing between Sample vs. Population

	sample	population
Size	n	N (usually ∞)
Mean	$\overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$	$\mu = \sum p_j X_j \quad or \quad \int \ldots$
Variance	$s^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(X_j - \overline{X}\right)^2$	$\sigma^2 = \sum p_j \left(X_j - \mu\right)^2 \quad or \quad \int \ldots$

Generalizing the sample to the population

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

Statistics allows us to estimate, *with high probability*, the true mean and true variance based only on the sample mean and sample variance.

How do sample means behave?

Suppose we observe data $X_1, X_2, \ldots, X_n$.

We can calculate the sample mean $\bar{X}$ exactly, but what can we say about the population mean μ?

Idea: μ is probably close to $\bar{X}$

Goal: Make this more rigorous

Central Limit Theorem:

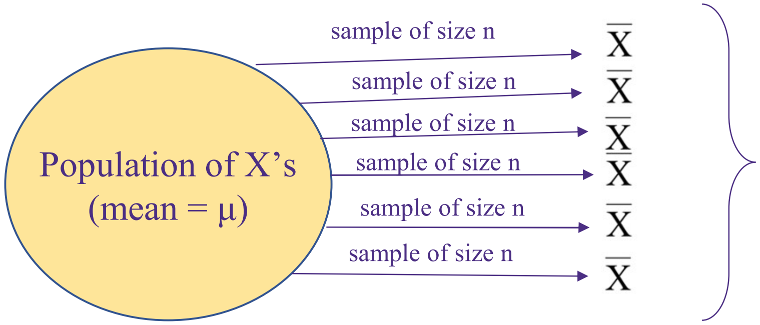
If $X_1, X_2, \ldots, X_n$ are independent and have the same distribution with variance σ^2, then if n is large ($n \geq 30$), the sample is approximately normally distributed.

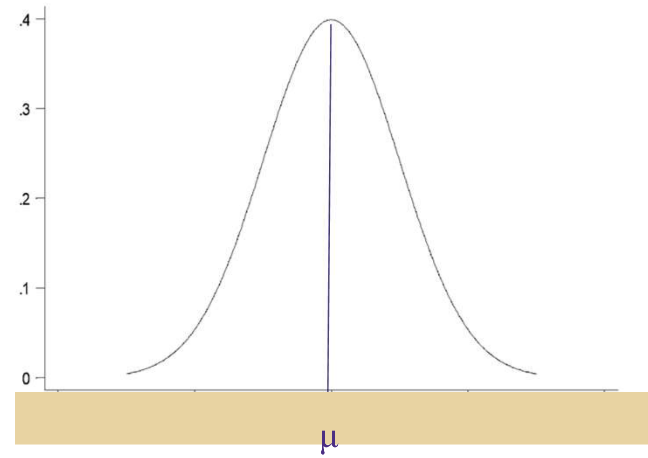
$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

As n increases, the normal approximation improves.

This is incredibly powerful and helpful!

Distribution of the Sample Mean is Normal regardless of underlying distribution of the data

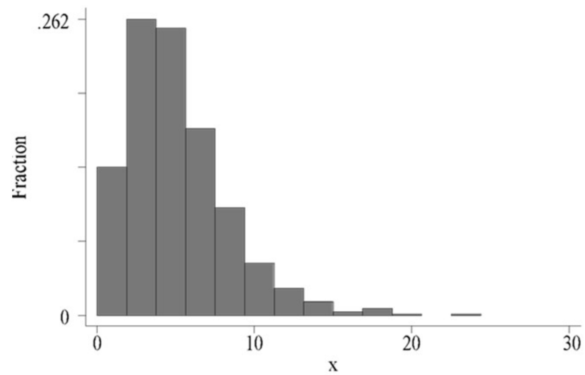




… provided n is large

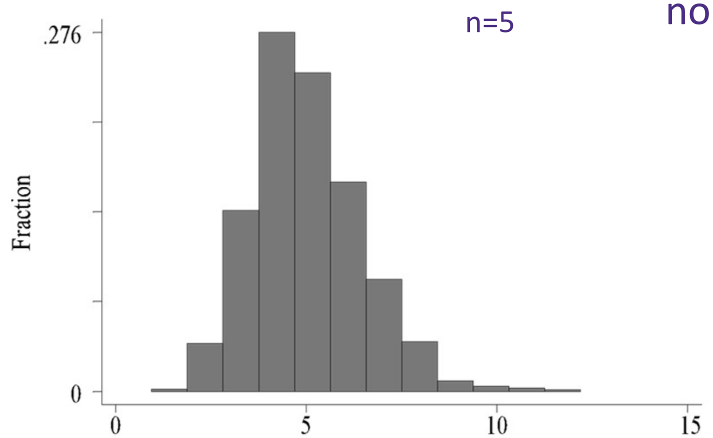
Central Limit Theorem Illustration

Distribution of X (not normally distributed…)



Distribution of $\overline{X}$

n=5

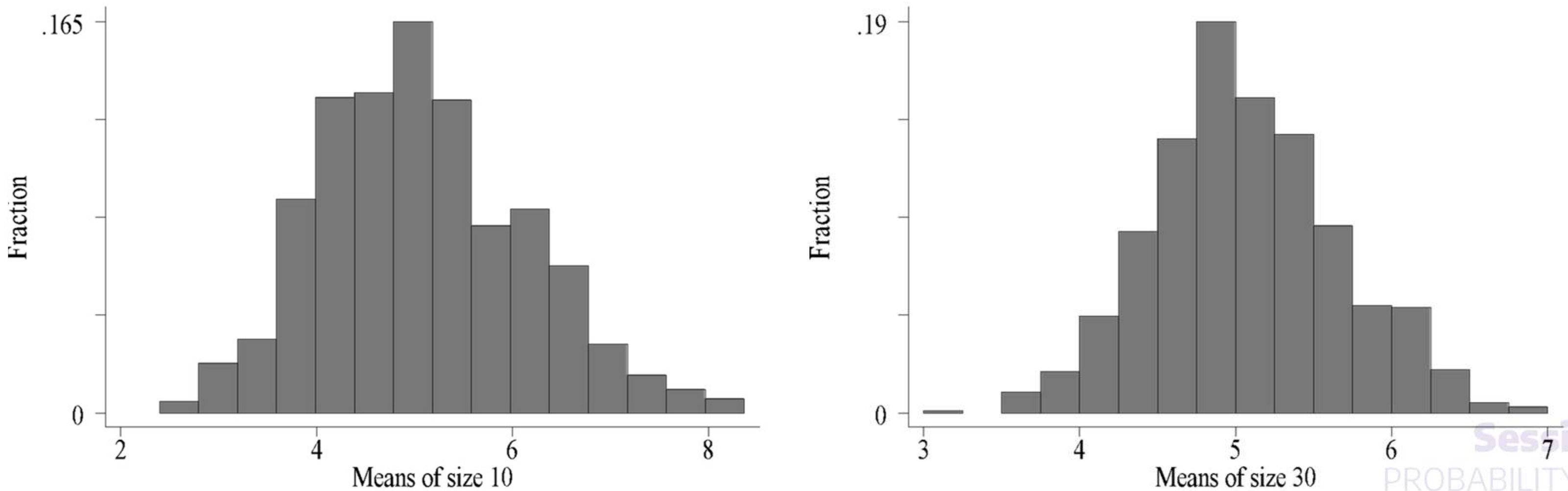


…but the means of X, even for n=5, are close to normally distributed

Session 5
PROBABILITY AND
INFERENTIAL STATISTICS

UNIVERSITY of WASHINGTON

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



Session 5
PROBABILITY AND
INFERENTIAL STATISTICS

UNIVERSITY of WASHINGTON

Central Limit Theorem

The central limit theorem allows us to use the sample ($X_{1,}$…, X_n) to discuss the population mean, 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?

Calculate 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

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

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

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

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

Pause- break time then work on exercise 1

14

Confidence Intervals

Session 5
PROBABILITY AND
INFERENTIAL STATISTICS

UNIVERSITY of WASHINGTON

Confidence Intervals

"(LL, UL) is a 95% confidence interval for a parameter θ" means that

- In repeated samples, 95% of the resulting confidence intervals would contain θ.

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

Note: Confidence intervals are about observed data. Prediction intervals (different) are intervals about new observations.

95% Confidence Interval for the Mean

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

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: Normal Distribution

Suppose gestational times are normally distributed with a standard deviation of 6 days. A sample of n=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

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

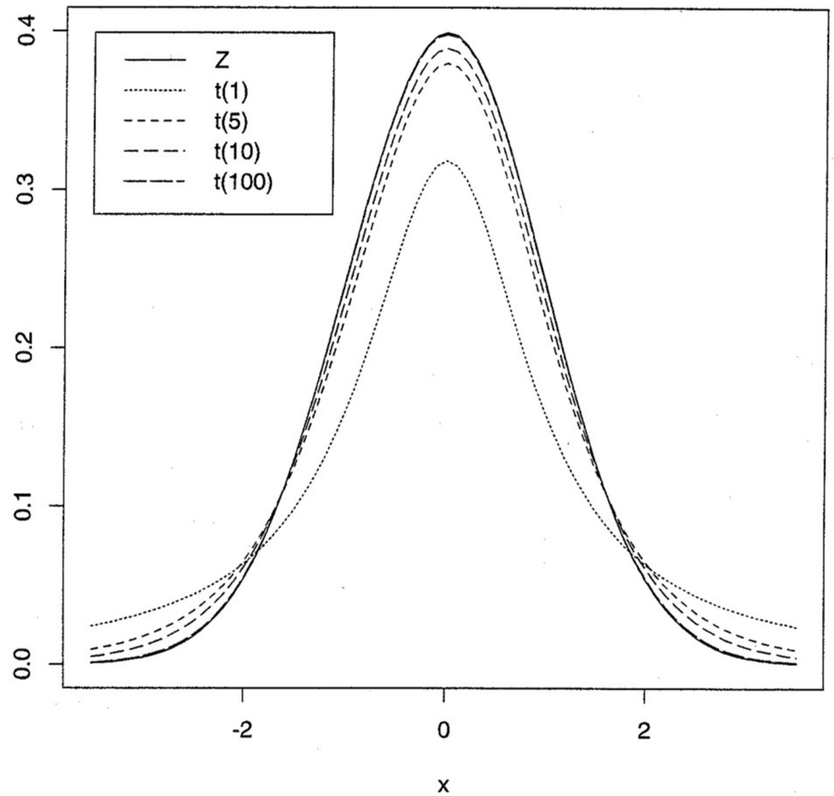
has a t-distribution with n-1 degrees of freedom.

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



Session 5
PROBABILITY AND
INFERENTIAL STATISTICS

UNIVERSITY of WASHINGTON

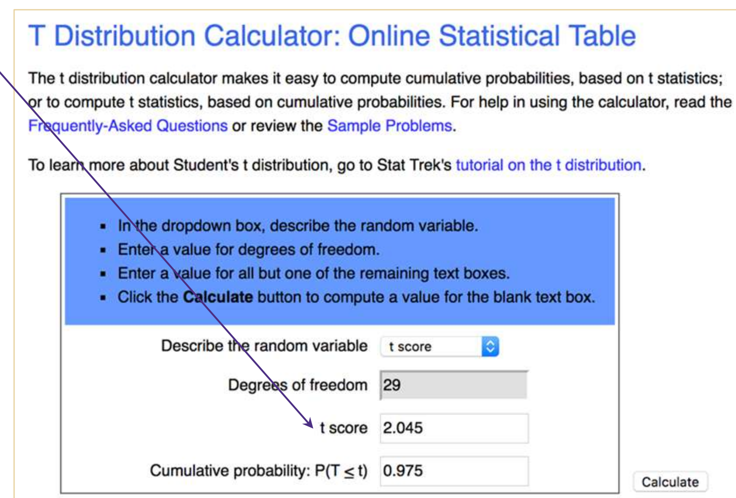
Confidence Intervals for σ^2 unknown

Example

Given 30 mothers with a mean gestation of 279.5 days and a variance of 28.3 days^2, 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



Take Home Points

- General (1 - α) Confidence Intervals:
  - Confidence intervals apply to 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 2-3