

450A Political Methodology I

Lecture 2: Univariate Statistical Inference

2. Interval Estimation

Stanford University

Interval Estimation

Mean of Normal Population with Known Variance

Mean of Normal Population with Unknown Variance

Mean of Population with Unknown Distribution

Summary

What is Interval Estimation?

- A **point estimator** $\hat{\theta}$ estimates a scalar population parameter θ with a single number.
- However, because we are dealing with a random sample, we might also want to report **uncertainty** in our estimate.
- An **interval estimator** for θ takes the following form:

$$[\hat{\theta}_{lower}, \hat{\theta}_{upper}]$$

where $\hat{\theta}_{lower}$ and $\hat{\theta}_{upper}$ are random quantities that vary from sample to sample.

- The interval represents the range of possible values within which we estimate the true value of θ to fall.
- An **interval estimate** is a realized value from an interval estimator. The estimated interval typically forms what we call a **confidence interval**, which we will define shortly.

Interval Estimation

Mean of Normal Population with Known Variance

Mean of Normal Population with Unknown Variance

Mean of Population with Unknown Distribution

Summary

Normal Population with Known σ^2

Suppose we have an i.i.d. random sample of size n , $X_1, \dots, X_n$, from $X \stackrel{iid}{\sim} N(\mu, 1)$.

Because a sum of independent normal variables is normal, the sampling distribution of the sample average is exactly (for every n):

$$\bar{X}_n \sim N(\mu, \sigma^2/n) \equiv N(\mu, 1/n)$$

Therefore, the standardized sample average is distributed as follows:

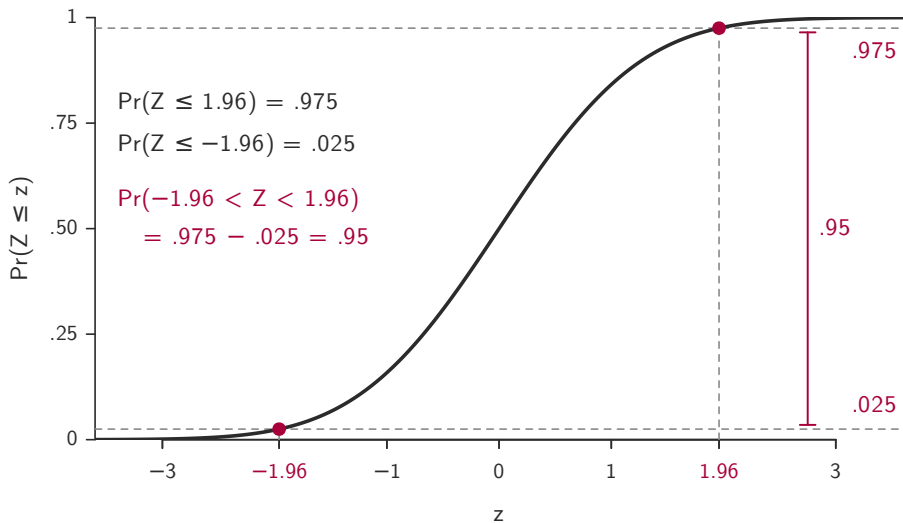
$$\frac{\bar{X}_n - \mu}{1/\sqrt{n}} \sim N(0, 1)$$

This implies

$$\Pr\left(-1.96 < \frac{\bar{X}_n - \mu}{1/\sqrt{n}} < 1.96\right) = .95$$

Why?

CDF of the Standard Normal Distribution



Constructing a Confidence Interval with Known σ^2

So we know that:

$$\Pr\left(-1.96 < \frac{\bar{X}_n - \mu}{1/\sqrt{n}} < 1.96\right) = .95$$

Rearranging yields:

$$\Pr\left(\bar{X}_n - 1.96/\sqrt{n} < \mu < \bar{X}_n + 1.96/\sqrt{n}\right) = .95$$

This implies that the following interval estimator

$$\left[\bar{X}_n - 1.96/\sqrt{n}, \bar{X}_n + 1.96/\sqrt{n}\right]$$

contains the true population mean μ with probability 0.95.

We call this estimator a 95% **confidence interval** for μ .

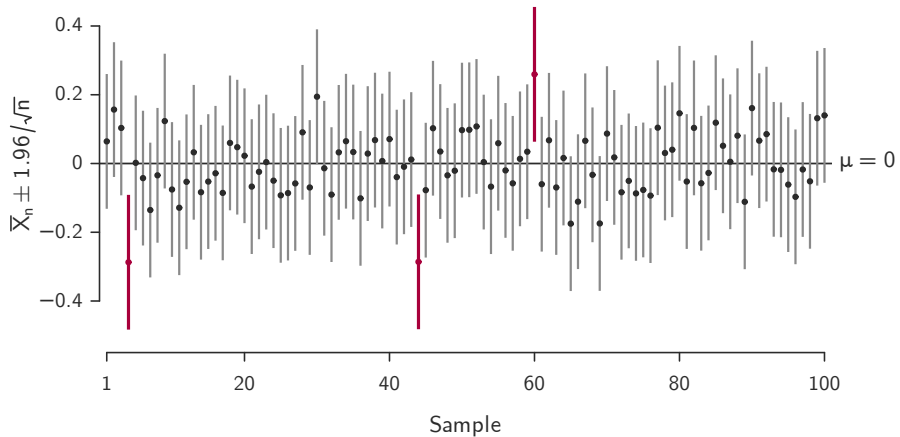
Interpreting a Confidence Interval

- Imagine we implement the interval estimator $\bar{X}_n \pm 1.96/\sqrt{n}$ for a particular sample and obtain the estimate of $[2.5, 3.3]$.
- Does this mean that there is a .95 probability that the true parameter value μ lies between these two particular numbers? No!
- Confidence intervals are easy to construct, but difficult to interpret:
 - Each confidence interval estimate from a particular sample either contains μ or not
 - However, if we were to repeatedly calculate the interval estimator over many random samples from the same population, 95% of the time the constructed confidence intervals would cover μ
 - Therefore, we refer to .95 as the **coverage probability**

Simulating Coverage Probability of Confidence Intervals

100 samples of size 100 from $N(0, 1)$:

97 of 100 intervals contain $\mu = 0$; the 3 that miss are in red



Constructing a Confidence Interval with Known σ^2

- The same principle easily extends to normal populations with any known variance.
- If $X \stackrel{iid}{\sim} N(\mu, \sigma^2)$ and we draw an i.i.d. random sample of size n , $X_1, \dots, X_n$, a 95% confidence interval for the population mean μ is:

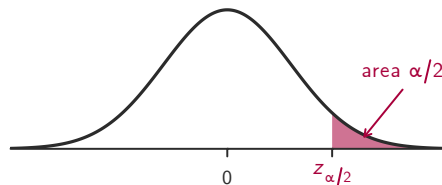
$$[\bar{X}_n - 1.96\sigma/\sqrt{n}, \bar{X}_n + 1.96\sigma/\sqrt{n}]$$

- More generally, a $100(1 - \alpha)\%$ confidence interval for μ is

$$[\bar{X}_n - z_{\alpha/2} \sigma/\sqrt{n}, \bar{X}_n + z_{\alpha/2} \sigma/\sqrt{n}]$$

where $z_{\alpha/2}$ is the $100(1 - \alpha/2)$ percentile of the standard normal distribution.

Confidence level	α	$z_{\alpha/2}$
90%	.1	1.645
95%	.05	1.960
99%	.01	2.576



Example: Feeling Thermometer Scores in ANES 2020

```
=====
INTRO PRE-ELECTION POLITICAL FIGURE THERMOMETERS
=====
```

I'd like to get your feelings toward some of our political leaders and other people who are in the news these days. I'll read the name of a person and I'd like you to rate that person using something we call the feeling thermometer. Ratings between 50 degrees and 100 degrees mean that you feel favorable and warm toward the person. Ratings between 0 degrees and 50 degrees mean that you don't feel favorable toward the person and that you don't care too much for that person. You would rate the person at the 50 degree mark if you don't feel particularly warm or cold toward the person.

How would you rate: Joe Biden (V201151)

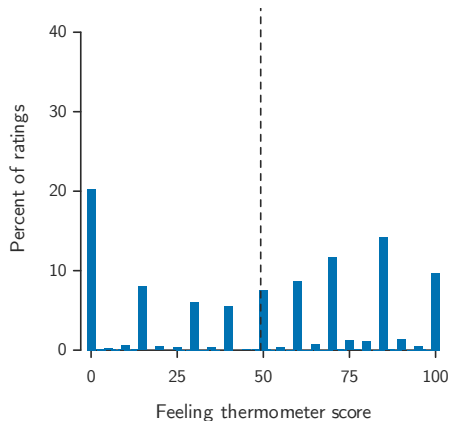
How would you rate: Donald Trump (V201152)

Source: American National Election Studies. 2021. ANES 2020 Time Series Study Full Release [dataset and documentation]. February 10, 2022 version.
electionstudies.org/data-center/2020-time-series-study

Empirical Distribution of FT Scores

Joe Biden

n = 8,060 mean = 49.2 SD = 34.5



Donald Trump

n = 8,048 mean = 40.4 SD = 40.3



Dashed line: mean. 99% of all ratings are multiples of 5. Data: 2020 ANES, pre-election.

Feeling Thermometer Example: Normal & Known Variance

Assume (for now) that Biden's FT scores are (1) normally distributed and (2) we (somehow) know that $\sigma = 34.5$.

We treat the ANES respondents as an i.i.d. sample, ignoring the survey weights and sampling design: $n = 8,060$ respondents rated Biden, with sample mean FT score $\bar{X}_n = 49.24$.

Q. What is the 95% confidence interval for the true mean FT score for Biden?

A. Plug $n = 8,060$, $\sigma = 34.5$ and $\bar{X}_n = 49.24$ in:

$$\begin{aligned} & [\bar{X}_n - 1.96\sigma/\sqrt{n}, \bar{X}_n + 1.96\sigma/\sqrt{n}] \\ &= \left[49.24 - 1.96 \cdot 34.5/\sqrt{8,060}, 49.24 + 1.96 \cdot 34.5/\sqrt{8,060} \right] \\ &= [48.49, 49.99] \end{aligned}$$

Example: FT Scores Interval Estimates

95% intervals $\bar{X}_n \pm 1.96\sigma/\sqrt{n}$, with σ set to each sample SD. Data: 2020 ANES, pre-election.

Joe Biden

49.2 [48.5, 50.0]



$n = 8,060$, $SD = 34.5$

Donald Trump

40.4 [39.6, 41.3]



$n = 8,048$, $SD = 40.3$



Mean feeling thermometer score

Interval Estimation

Mean of Normal Population with Known Variance

Mean of Normal Population with Unknown Variance

Mean of Population with Unknown Distribution

Summary

Normal Population with Unknown σ^2

In practice, it is rarely the case that we somehow know the true value of σ^2 .

Suppose now that we have an i.i.d. random sample $X_1, \dots, X_n$ from $X \stackrel{iid}{\sim} N(\mu, \sigma^2)$, where σ^2 is **unknown**.

Then, as before,

$$\bar{X}_n \sim N(\mu, \sigma^2/n) \quad \text{and so} \quad \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

Previously, we then constructed the interval:

$$[\bar{X}_n - z_{\alpha/2}\sigma/\sqrt{n}, \bar{X}_n + z_{\alpha/2}\sigma/\sqrt{n}]$$

But we can **not** directly use this now because σ^2 is unknown.

Instead, we need an **estimator** of σ^2 , $\hat{\sigma}^2$.

Estimators for the Population Variance

Two possible estimators of population variance:

$$S_{0n}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \quad S_{1n}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

Which do we prefer? Let's check properties of these estimators. **Show (Worksheet Q1):**

① Unbiasedness:

$$E[S_{0n}^2] = \frac{n-1}{n} \sigma^2 \quad \text{and} \quad E[S_{1n}^2] = \sigma^2$$

② Consistency:

$$S_{0n}^2 \xrightarrow{P} \sigma^2 \quad \text{and} \quad S_{1n}^2 \xrightarrow{P} \sigma^2$$

S_{1n}^2 (unbiased and consistent) is commonly called the **sample variance**.

Estimating σ^2 and the Standard Error

We use the sample variance $S_n^2 \equiv S_{1n}^2$ to estimate the population variance σ^2 . (The plug-in (analogy-principle) estimator of the variance is S_{0n}^2 ; dividing by $n - 1$ instead of n corrects its bias.)

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

Our estimated standard deviation of the sampling distribution (i.e. estimated **standard error**) of $\bar{X}_n$ is therefore $S_n/\sqrt{n}$.

Does $\bar{X}_n \sim N(\mu, S_n^2/n)$, which would imply $\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \sim N(0, 1)$?

No, because S_n is a random variable instead of a parameter (like σ).

Thus, we need to derive the sampling distribution of the new random variable

$$T_n \equiv \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \sim ?$$

Sampling Distribution with the Estimated Variance

It turns out that T_n follows **Student's t -distribution** with $n - 1$ **degrees of freedom**, written τ_{n-1} .

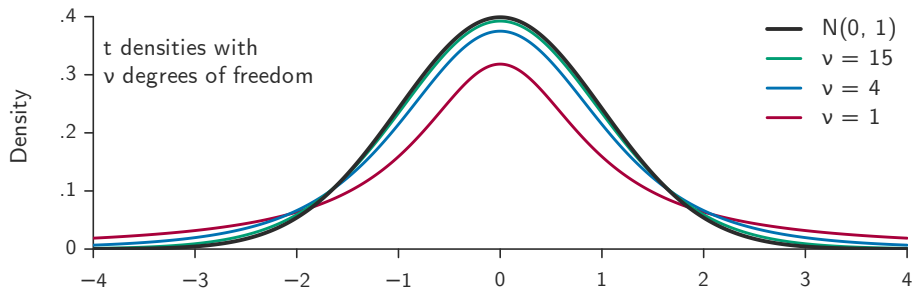
Theorem (Distribution of t -Value from a Normal Population)

Suppose we have an i.i.d. random sample of size n from $N(\mu, \sigma^2)$. Then, the sample mean $\bar{X}_n$ standardized with the estimated standard error $S_n/\sqrt{n}$ satisfies,

$$T_n \equiv \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \sim \tau_{n-1}$$

- **Prove (Worksheet Q2)**
- The quantity T_n is commonly called the **t -value**. It depends on the unknown μ , but for a normal population its distribution does not depend on μ or σ^2 .
- Student's t -distribution has well known properties, and so can be used to construct confidence intervals like $N(0, 1)$.

Properties of Student's t -Distribution



- The shape of the t -distribution is determined by one parameter, ν (degrees of freedom)
- If $X \sim \tau_\nu$, then $E[X] = 0$ (for $\nu > 1$) and $V[X] = \nu/(\nu - 2)$ (for $\nu > 2$)
- Bell-shaped like the normal distribution, but has fatter tails for $\nu < \infty$
- $X \xrightarrow{d} N(0, 1)$ as $\nu \rightarrow \infty$
- R functions: `dt(x, df)` (PDF), `pt(q, df)` (CDF)
`qt(p, df)` (quantiles), `rt(n, df)` (random draws)

Constructing a Confidence Interval with Unknown σ^2

To calculate confidence intervals, we follow the same procedure as the case of known variance, except that we now use percentiles of t instead of z : $t_{\alpha/2, \nu}$ is the $100(1 - \alpha/2)$ percentile of τ_ν .

Confidence level	α	$t_{\alpha/2, 1}$	$t_{\alpha/2, 4}$	$t_{\alpha/2, 15}$	$\cdots$	$z_{\alpha/2}$
90%	.1	6.314	2.132	1.753	$\cdots$	1.645
95%	.05	12.71	2.776	2.131	$\cdots$	1.960
99%	.01	63.66	4.604	2.947	$\cdots$	2.576

That is, we have

$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \sim \tau_{n-1}$$

So

$$\begin{aligned} \Pr\left(-t_{\alpha/2, (n-1)} < \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} < t_{\alpha/2, (n-1)}\right) &= 1 - \alpha \\ \iff \Pr\left(\bar{X}_n - t_{\alpha/2, (n-1)}S_n/\sqrt{n} < \mu < \bar{X}_n + t_{\alpha/2, (n-1)}S_n/\sqrt{n}\right) &= 1 - \alpha \end{aligned}$$

Therefore, the $100(1 - \alpha)\%$ confidence interval for μ with unknown σ^2 is

$$\left[\bar{X}_n - t_{\alpha/2, (n-1)}S_n/\sqrt{n}, \bar{X}_n + t_{\alpha/2, (n-1)}S_n/\sqrt{n}\right]$$

A good rule of thumb for a 95% C.I. for a typical sample size is $t_{\alpha/2, (n-1)} \approx 2$.

Suppose we have an i.i.d. random sample of size n , $X_1, \dots, X_n$, from $X \stackrel{iid}{\sim} N(\mu, \sigma^2)$. For constants a, b with $b \neq 0$, derive a $100(1 - \alpha)\%$ confidence interval for $(a + b\mu)$ under

- (i) known σ^2
- (ii) unknown σ^2

Interval Estimation

Mean of Normal Population with Known Variance

Mean of Normal Population with Unknown Variance

Mean of Population with Unknown Distribution

Summary

Asymptotic Confidence Intervals for Unknown Distribution

Often, observed distribution of X is so far from normal that we do not want to assume $X \sim N(\mu, \sigma^2)$ (e.g. Trump's FT scores).

How can we proceed? *The central limit theorem!*

If $X_1, \dots, X_n$ are i.i.d. with finite variance $\sigma^2 > 0$, the CLT gives

$$Z_n \equiv \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0, 1)$$

and we also know that

$$S_n^2 \xrightarrow{p} \sigma^2$$

as $n \rightarrow \infty$.

This therefore implies

$$T_n \equiv \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \xrightarrow{d} N(0, 1)$$

Result:

If the sample is i.i.d. with finite variance, the normal interval $\bar{X}_n \pm z_{\alpha/2} S_n/\sqrt{n}$ covers μ with probability close to $1 - \alpha$ when n is large, even if the population is not normal and σ^2 is unknown.

Using the t Interval for Robust Inference

The asymptotic normal confidence interval is convenient, but we must assume that n is large enough.

If n is small, the normal intervals are typically **overconfident**, i.e. too narrow.

To address this concern, some propose that we use the t interval instead of normal.

The logic:

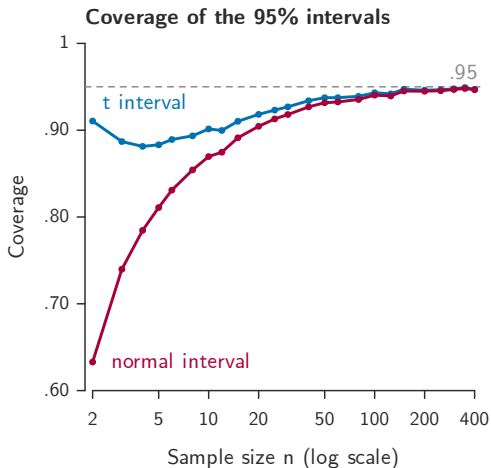
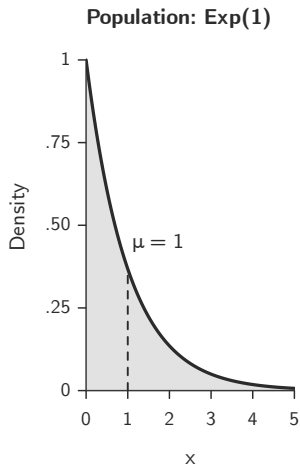
- If n is large, $t_{\alpha/2, (n-1)} \approx z_{\alpha/2}$ (1.984 vs. 1.960 at $n = 100$), so the t and normal intervals are nearly the same, and the CLT makes both valid.
- If n is small, the t interval is wider than the normal interval, so it is less overconfident. It can still cover μ less than 95% of the time when the population is skewed (next slide).

Therefore, it is sensible to use the t interval even if we do not assume that the population distribution is normal, especially if n is small.

Simulation Study: 95% Asymptotic Confidence Interval

Population: $X \sim \text{Exp}(\mu = 1)$ (Exponential Distribution with Mean of 1)

Coverage of the asymptotic confidence intervals as n varies from 2 to 400



Interval Estimation

Mean of Normal Population with Known Variance

Mean of Normal Population with Unknown Variance

Mean of Population with Unknown Distribution

Summary

Summary:

To calculate a confidence interval for the population mean:

- If we can safely assume that the population is normal (rather unlikely),
 - Use the **normal interval** if the variance is also known (even less likely)
 - Use the **t interval** if the variance has to be estimated from data.
- If we do not want to assume a normal population distribution,
 - Use either the **normal interval** or **t interval** (choice doesn't matter) if sample size is large ($n > 100$ or so)
 - Use the **t interval** if sample size is medium (about $20 \leq n \leq 100$)
 - If sample size is smaller than that ($n < 20$) for now also use **t interval**—more on this later in methods sequence

Keep in mind that it is almost always better to be conservative when it comes to statistics.