Confidence Intervals and Hypothesis Test

STAT-UB.0001 Statistics for Business Control

Ningshan Zhang

IOMS Department
nzhang@stern.nyu.edu

Jul 31, 2018

* Office Hours: Wed & Fri 10:00 - 11:30 AM, KMC 8-174

Review: CLT

Suppose $X_1, X_2, \ldots, X_n$ are sampled independently from a population with mean μ and standard deviation σ. Let $\bar{X}$ be the sample mean,

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Then,

- $\mu_{\bar{X}} = \mathbb{E}(\bar{X}) = \mu$, *(holds for any n)*
- $\sigma_{\bar{X}} = \text{sd}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$, *(holds for any n)*
- If n is sufficiently large *($n \geq 30$)*, then $\bar{X}$ is approximately normal.

Review: Distribution of $\bar{X}$

Suppose $X_1, X_2, \ldots, X_n$ are sampled independently from the same population. Let $\bar{X}$ be the sample mean,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Table: Relationship between population and sample mean.

Population	**Sample size n**	**Sample mean $\bar{X}$**
Normal	Any $n \geq 1$	Normal
Any distribution	$n \geq 30$	Approximately normal

* Demo: `http://onlinestatbook.com/stat_sim/sampling_dist`

Review: CI for the Mean with Known Variance

Setup: assume the population variance σ^2 is known, build a CI for the population mean μ using a sample of n observations.

- By CLT, when $n \geq 30$, $\bar{X}$ is (roughly) normally distributed with mean μ and standard deviation $\sigma/\sqrt{n}$.
- Let $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$, then Z is a standard normal (roughly).
- In particular, (one can use 2 instead of 1.96)

$$\mathbb{P}(-1.96 < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < 1.96) = 0.95.$$

- $(\bar{X} - \frac{1.96\sigma}{\sqrt{n}}, \bar{X} + \frac{1.96\sigma}{\sqrt{n}})$ is a CI for μ with confidence level 0.95.

Review: Confidence Intervals

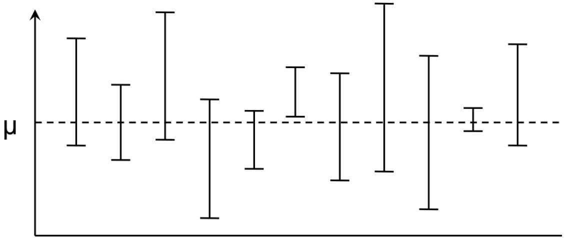
When is the confidence interval valid?

- The observations $X_1, \cdots, X_n$ are drawn independently from the population.
- Either the population is normal, or sample size $n \geq 30$.

Review: Interpretations of Confidence Intervals

What does "a CI for μ with confidence level 0.95" mean?

- If we repeat this process of drawing a random sample and constructing a confidence interval many many times,
- Then the proportion of these intervals that contain μ is equal to 0.95.



CI for the Mean: Known Variance

$(\bar{X} - \frac{1.96\sigma}{\sqrt{n}}, \bar{X} + \frac{1.96\sigma}{\sqrt{n}})$ is a CI for μ with confidence level 0.95.

- Problem: need to know the population variance σ^2.
- Unfortunately, the assumption that σ^2 is known is unrealistic in many situations.
- Solution: in practice, we typically use S^2, the sample variance, to estimate it.

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

CI for the Mean: *Unknown* Variance

Consider the random variable

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}.$$

Why it is a random variable?

- The observations $X_1, \cdots, X_n$ are random.
- Thus, the sample mean $\bar{X}$ and sample variance S^2 are random;
- Thus, the ratio T is random.

CI for the Mean: Unknown Variance

Distribution of T

If $\bar{X}$ is normally distributed, then

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

follows a *Student's t-distribution* with $n - 1$ degrees of freedom.

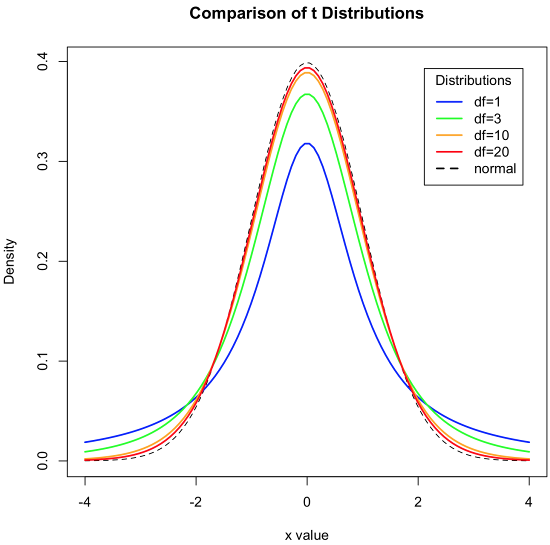
When is $\bar{X}$ is normally distributed?

- The observations $X_1, \cdots, X_n$ are drawn independently from the population.
- Either the population is normal, or sample size $n \geq 30$.

The t-Distribution

- The t-distribution is "similar" to the standard normal distribution.
- It is continuous, bell-shaped, and symmetric around zero, but it has fatter tails than the standard normal distribution.
- The t-distribution has one parameter, the degrees of freedom (df).

The t-Distribution



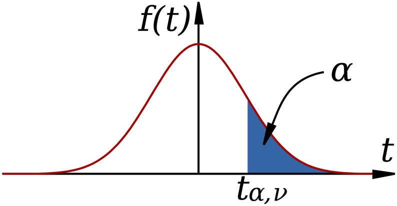
The t- Distribution

- When the df is large (df ≥ 30), the t-distribution is close to the standard normal distribution Z.
- As df $\to \infty$, the t-distribution converges to the standard normal distribution Z.

The t-Distribution

Notation $t_{\alpha,\nu}$: the point that the area to its right under the t-distribution curve with df$=\nu$ is α, thus

$$\mathbb{P}(-t_{\alpha,\nu} \leq T \leq t_{\alpha,\nu}) = 1 - 2\alpha.$$



- Use t-table to find $t_{\alpha,\nu}$, for different α,ν.
- Example: What is $t_{0.05,19}$? What is $t_{0.025,9}$?

CI for the Mean: Unknown Variance

A $1 - \alpha$ CI for μ is ...

- When n is large ($n \geq 30$), approximate T with Z:

$$\mathbb{P}(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq z_{\alpha/2}) = 1 - \alpha$$

$$\Rightarrow \mathbb{P}(\bar{X} - z_{\alpha/2}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}) = 1 - \alpha.$$

- When n is small and the population is normal:

$$\mathbb{P}(-t_{\alpha/2,n-1} \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq t_{\alpha/2,n-1}) = 1 - \alpha$$

$$\Rightarrow \mathbb{P}(\bar{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}) = 1 - \alpha.$$

CI for the Mean: Summary

	σ known	σ unknown
$n \geq 30$	$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$	$\bar{X} \pm z_{\alpha/2} \frac{S}{\sqrt{n}}$
$n < 30$, pop. is normal	$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$	$\bar{X} \pm t_{\alpha/2,n-1} \frac{S}{\sqrt{n}}$
$n < 30$, pop. isn't normal	N.A.	N.A.

Degrees of Freedom (df)

The random variable S computed with n observations has degrees of freedom $n-1$.

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

Proof sketch: define random variables $Y_i = X_i - \bar{X}$. Then, $S^2 = \frac{1}{n-1}\sum_{i=1}^{n} Y_i^2$. However, Y_i must satisfy one restriction:

$$\sum_{i=1}^{n} Y_i = (\sum_{i=1}^{n} X_i) - n\bar{X} = 0.$$

Thus S^2 loses one degree of "freedom". $\square$

* Note: the definition of df is not required for this class.

CI for the Proportion

Often interested to know the proportion of the population that satisfies a condition, e.g.

- Proportion of NYU undergrads owning an iPhone;
- Proportion of voters supporting candidate A.

Notation

- p = proportion in population (population parameter)
- $\widehat{p}$ = proportion in sample (sample statistic)

Goal: construct a confidence interval for p.

CI for the Proportion

Assume $X_1, \cdots, X_n$ are drawn indepedently from the population. Each X_i is a random variable, with

$$X_i = \begin{cases} 0, & \text{with prob. } 1-p \\ 1, & \text{with prob. } p \end{cases}$$

Then by definition,

$$\begin{aligned} \mathbb{E}(X_i) &= p \\ \text{var}(X_i) &= (1-p)(0-p)^2 + p(1-p)^2 = p(1-p) \end{aligned}$$

CI for the Proportion

The sample proportion is also a sample mean:

$$\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

By CLT,

- $\mu_{\widehat{p}} := \mathbb{E}(\widehat{p}) = p$.
- $\sigma^2_{\widehat{p}} := \mathrm{var}(\widehat{p}) = \frac{p(1-p)}{n}$.
- When n is large, $\widehat{p}$ is normally distributed.
  - Here we need $np \geq 15$ and $n(1-p) \geq 15$.

* ":=" means "is defined as".

CI for the Proportion

When $np \geq 15$ and $n(1-p) \geq 15$,

$$\mathbb{P}(-z_{\alpha/2} \leq \frac{\widehat{p}-p}{\sigma_{\widehat{p}}} \leq z_{\alpha/2}) = 1-\alpha$$
$$\Rightarrow \mathbb{P}(\widehat{p} - z_{\alpha/2}\,\sigma_{\widehat{p}} \leq p \leq \widehat{p} + z_{\alpha/2}\,\sigma_{\widehat{p}}) = 1-\alpha$$

- $\sigma_{\widehat{p}} = \sqrt{\frac{p(1-p)}{n}}$, use $\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$ as an approximation.
- Thus, a $1-\alpha$ confidence interval for population proportion p is
$$\left(\widehat{p} - z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}},\ \widehat{p} + z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right)$$

Hypothesis Testing

Often, someone makes a claim about the world, such as

- A Mini Cooper achieve 37 highway miles per gallon.
- AT&T is the nation's fastest 4G LTE network.
- A Subway footlong sub is 12 inches long.

We collect some data, and we want to evaluate the plausibility of that claim in the face of data.

Hypothesis Testing

Common use cases for hypothesis testing:

- Check stated claims, e.g. model is realistic.
- Check if possible that something happens by chance alone, e.g. 10 heads in a roll by a fair coin.
- Check for effects of an intervention, e.g. A/B testing.

Hypothesis Testing

Null Hypothesis (H_0):

- The hypothesis that will be accepted unless the data provide convincing evidence that it is false.
- Example: stated claim is true; event occurs by chance; the intervention has no effect.

Alternative Hypothesis (H_A):

- The hypothesis that will be accepted when the null hypothesis is rejected.
- Example: stated claim is false; event doesn't occurs by chance alone; the intervention has effect.