STA13: Elementary Statistics

Lecture 14
Book Sections 5.2-5.3

Dmitriy Izyumin

February 14 2018

CI for the Population Mean - Large Sample

Population: Mean μ, standard deviation σ

Sample: Size $n \geq 30$, sample mean $\bar{x}$, sample s.d. s.

- point estimator of μ is $\bar{x}$
- $(1-\alpha)100\%$ confidence interval for μ is $\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$
  (If σ is known, use it instead of s)
- margin of error is $z_{\alpha/2}\frac{s}{\sqrt{n}}$

Interpreting a Confidence Interval

We say:

We are ___1___% confident that the true mean ___2___ is between ___3___ and ___4___ ___5___.

The blanks are:

1. Confidence level $100(1-\alpha)\%$
2. Parameter of interest (in terms of problem)
3. Lower bound of interval
4. Upper bound of interval
5. Units in which the population mean is measured

Example

In a random sample of 64 UCD graduates, the average annual earning within a year of graduation is \$35,000, and the standard deviation is \$4,600. Use this data to construct a 90% confidence interval for μ, the average earning of all UCD students within a year after graduation.

Example

Given: $\bar{x} = 35,000$, $s = 4600$, $n = 64$

Use this to construct the interval:

- $SE(\bar{x}) = \frac{s}{\sqrt{n}} = \frac{4600}{\sqrt{64}} = 575$
- Since $n > 30$, we know $\bar{x}$ is approximately normal (CLT)
- 90% confidence $\rightarrow \alpha = 1 - 0.90 = 0.10 \rightarrow \alpha/2 = 0.05$
- $z_{\alpha/2} = z_{0.05} = 1.645$

The confidence interval is: $35000 \pm 1.645(575)$

Equivalently: $(34013, 35987)$

We are 90% confident that the true mean annual earning of UCD graduates is between \$34054.13 and \$35945.88.

Understanding Estimators

Setting: The population has mean μ and s.d. σ.
We are curious about μ.

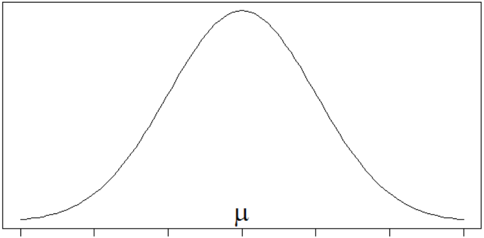
- Obtain a random sample of size n .
- Use the sample mean $\bar{x}$ to make inferences about the population mean μ.
- Suppose n is large ($n > 30$), so we can use the CLT.

Understanding Estimators

Here's what we know about $\bar{x}$ *before* taking the sample:

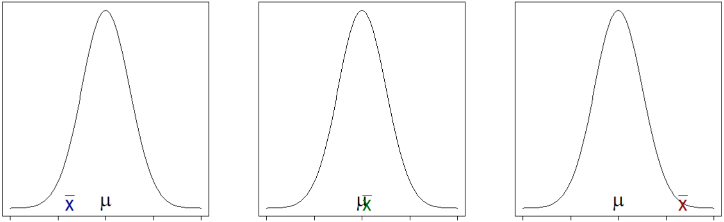
- $\bar{x}$ has mean μ and s.d. $\frac{\sigma}{\sqrt{n}}$
- $\bar{x}$ is distributed Normal$(\mu, \frac{\sigma}{\sqrt{n}})$. (Thanks to the CLT)

We don't know what μ is, but we know the pdf of $\bar{x}$ looks like this:



Understanding Estimators

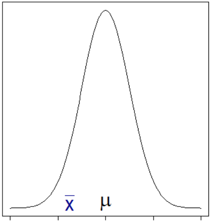
- Take a sample of size n, and compute $\bar{x}$.
- We know the value of $\bar{x}$ (ex. $\bar{x} = 5.67$).
- We don't know where it falls on the x-axis (because we don't know μ).



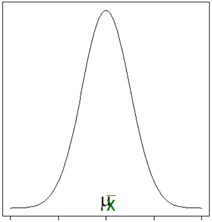
Is it here? ...Or here? ...Or here?

Understanding Estimators

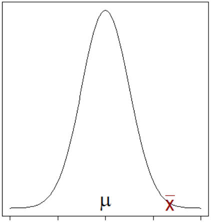
- The point estimator for μ is $\bar{x}$.
- If we have to estimate μ by a single number, we would use $\bar{x}$ (ex. 5.67).
- But we don't know how far our value of $\bar{x}$ is from μ.



Is it here?



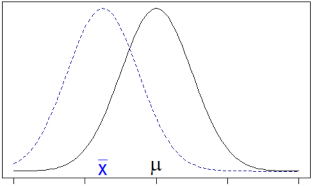
...Or here?

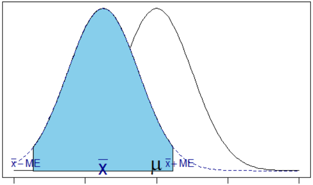


...Or here?

Understanding Estimators

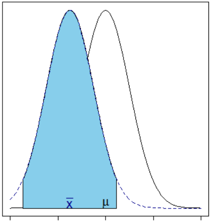
- Imagine an identical curve centered around $\bar{x}$.
- Choose the interval that contains the middle 95% of the area under that curve.
- This is the 95% confidence interval for μ.
- We can do this for whichever value of $\bar{x}$ we get.



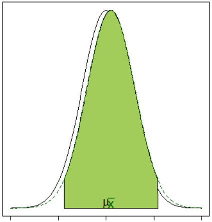


Understanding Estimators

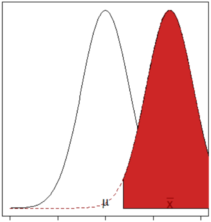
- Each interval either contains the true parameter or not.
- There's no way to know for a particular interval.
- If we repeat the procedure for many samples of size n, then 95% of the intervals will contain the parameter.



Contains ✓



Contains ✓



Doesn't contain ✗

Interpreting a Confidence Interval

Each individual confidence interval either contains the true parameter, or does not.

But...

If we could repeat the procedure many times (take many samples of size n from the same population, and construct a confidence interval from each), then $(1-\alpha)100\%$ of the intervals would contain the true value of the parameter.

What if the Sample is Small?

We have a way to compute a $(1 - \alpha)100\%$ confidence interval for μ when the sample is large ($n \geq 30$).

What happens if the sample is small? ($n < 30$)

1. Distribution of $\bar{x}$ depends on the population distribution. (Can't rely on CLT here)

2. The estimate s might not be very accurate.

Dealing with Problem 1

1. Distribution of $\bar{x}$ depends on the population distribution. (Can't rely on CLT here)

Two possibilities:

- If the population is normally distributed, then $\bar{x}$ is normally distributed even with small n.
- If the population distribution is unknown, we can *assume* that the population distribution is normal, and carry out the rest of the procedure. In this case, it is necessary to point out that we are making an extra assumption, and our estimation procedure may not be accurate in this case.

Dealing with Problem 2

2 The estimate s might not be very accurate.

Use the t distribution instead of $N(0, 1)$.

The small-sample confidence interval for μ becomes

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

The t Distribution

The t distributions are a class of continuous distributions.

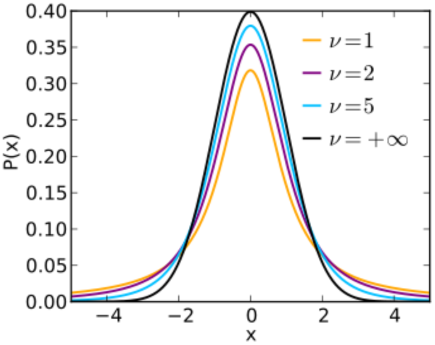
- Similar to the standard normal $N(0, 1)$
- It arises when estimating the population mean, but the sample size is small, and the population s.d. is unknown.
- Each t distribution is determined by one number called the degrees of freedom.

The t Distribution

The pdf of a t distribution...

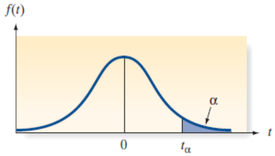
- is centered around 0,
- is bell-shaped,
- has wider tails than the pdf of $N(0,1)$,
- depends on the degrees of freedom,
- gets closer to the pdf of $N(0,1)$ for larger degrees of freedom.

The t Distribution



- Image from Wikipedia.
- Here ν denotes the degrees of freedom.
- The black curve ($\nu = \infty$) is $N(0, 1)$.

Using the t-table



Degrees of Freedom	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$	$t_{.001}$
1	3.078	6.314	12.706	31.821	63.657	318.31
2	1.886	2.920	4.303	6.965	9.925	22.326
3	1.638	2.353	3.182	4.541	5.841	10.213
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930

- Figure out what α you need to use.
- Figure out what degrees of freedom you need to use.
- Get t_α from table.

Using the t-table

Degrees of Freedom	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$
1	3.078	6.314	12.706	31.821
2	1.886	2.920	4.303	6.965
3	1.638	2.353	3.182	4.541
4	1.533	2.132	2.776	3.747
5	1.476	2.015	2.571	3.365
6	1.440	1.943	2.447	3.143
7	1.415	1.895	2.365	2.998
8	1.397	1.860	2.306	2.896
9	1.383	1.833	2.262	2.821
10	1.372	1.812	2.228	2.764
11	1.363	1.796	2.201	2.718
12	1.356	1.782	2.179	2.681

What if the desired α is not in the table?

- Suppose we want $t_{0.02}$ for degrees of freedom 10.
- Table only has $t_{0.025} = 2.228$ and $t_{0.010} = 2.764$.
- The best we can say is $2.228 < t_{0.02} < 2.764$.

Using the t-table

8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.767
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

What if the desired degrees of freedom are not in the table?

- ALWAYS ROUND DOWN
- Need df 32 $\rightarrow$ use df 30
- Need df 58 $\rightarrow$ use df 40

CI for the Population Mean - Small Sample

Population: Mean μ, standard deviation σ
Either population is normal (stated in problem), or we make the assumption that it is normal (you need to state this).

Sample: Size $n < 30$, sample mean $\bar{x}$, sample s.d. s.

- point estimator of μ is $\bar{x}$
- We use t distribution with $n-1$ degrees of freedom
- $(1-\alpha)100\%$ confidence interval for μ is $\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$
- margin of error is $t_{\alpha/2}\frac{s}{\sqrt{n}}$

Example 2

In a random sample of 14 UCD graduates, the average annual earning within a year of graduation is \$35,000, and the standard deviation is \$4,600. Use this data to construct a 90% confidence interval for μ, the average earning of all UCD students within a year after graduation. Assume that the annual earnings of recent UCD graduates are normally distributed.

Example 2

Given: $\bar{x} = 35,000$, $s = 4600$, $n = 14$

Use this to construct the interval:

- $SE(\bar{x}) \approx \frac{s}{\sqrt{n}} = \frac{4600}{\sqrt{14}} = 1229.40$
- $n < 30$, but $\bar{x}$ is normal (since population is normal)
- 90% confidence $\rightarrow \alpha = 1 - 0.90 = 0.10 \rightarrow \alpha/2 = 0.05$
- Since $n < 30$, use t distribution with df=14-1=13
- $t_{\alpha/2} = t_{0.05} = 1.771$

The confidence interval is: $35000 \pm 1.771(1229.4)$

Equivalently: (32822.73, 37177.27)

We are 90% confident that the true mean annual earning of UCD graduates is between \$32822.73 and \$37177.27.

Example 3

In a random sample of 14 UCD graduates, the average annual earning within a year of graduation is $35,000, and the standard deviation is $4,600. Use this data to construct a 90% confidence interval for μ, the average earning of all UCD students within a year after graduation.

Example 3

In a random sample of 14 UCD graduates, the average annual earning within a year of graduation is \$35,000, and the standard deviation is \$4,600. Use this data to construct a 90% confidence interval for μ, the average earning of all UCD students within a year after graduation.

We will assume that the annual earnings of recent UCD graduates are normally distributed. Keep in mind that our estimates may not be accurate if this assumption is violated.

Example 3

Given: $\bar{x} = 35,000$, $s = 4600$, $n = 14$

Use this to construct the interval:

- $SE(\bar{x}) \approx \frac{s}{\sqrt{n}} = \frac{4600}{\sqrt{14}} = 1229.40$
- $n < 30$, but $\bar{x}$ is normal (because of our assumption)
- 90% confidence $\rightarrow \alpha = 1 - 0.90 = 0.10 \rightarrow \alpha/2 = 0.05$
- Since $n < 30$, use t distribution with df=14-1=13
- $t_{\alpha/2} = t_{0.05} = 1.771$

The confidence interval is: $35000 \pm 1.771(1229.4)$

Equivalently: (32822.73, 37177.27)

We are 90% confident that the true mean annual earning of UCD graduates is between \$32822.73 and \$37177.27.