Recitation, Week 8

Ye Wang

New York University

POL-850

Spring 2018

Outline

1. Confidence Intervals and its interpretation
2. Student's t distribution

Confidence Intervals (CI)

- A subset of the parameter space
- A region where we can be fairly confident that the true parameter μ (the population mean, for example) lies
- How much confident? We can choose this!
- The larger the region is, the more confident we will be
- In practice, people usually choose 95%
- How should we construct this interval?

Confidence Intervals (CI)

- We focus on the unidimensional parameter space (the real axis)
- We want an interval on the real axis, $[a, b]$, so that $P[a \leq \mu \leq b] = 0.95$
- It is known that $\frac{\bar{X}-\mu}{SE} = Z$, so

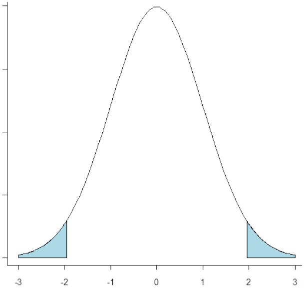
Confidence Intervals (CI)

- We focus on the unidimensional parameter space (the real axis)
- We want an interval on the real axis, $[a, b]$, so that $P[a \leq \mu \leq b] = 0.95$
- It is known that $\frac{\bar{X}-\mu}{SE} = Z$, so

$$\begin{aligned} P[a \leq \mu \leq b] &= P[\frac{\bar{X}-b}{SE} \leq \frac{\bar{X}-\mu}{SE} \leq \frac{\bar{X}-a}{SE}] \\ &= P[\frac{\bar{X}-b}{SE} \leq Z \leq \frac{\bar{X}-a}{SE}] \\ &= 0.95 \end{aligned}$$

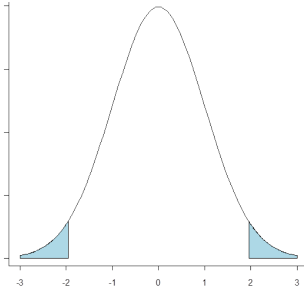
Confidence Intervals (CI)

- Usually we want this inverval around Z to be symmetric, so we choose a number $z_{0.025}$, such that $P[-z_{0.025} \leq Z \leq z_{0.025}] = 0.95$



Confidence Intervals (CI)

- Usually we want this inverval around Z to be symmetric, so we choose a number $z_{0.025}$, such that $P[-z_{0.025} \leq Z \leq z_{0.025}] = 0.95$



- How to know this $z_{0.025}$? We check the z score table!

Table 6-3 Proportions of the Normal Curve above the Absolute Value of Z

First digit and first decimal of Z	Second decimal of Z									
	.00	*.01*	*.02*	*.03*	*.04*	*.05*	*.06*	*.07*	*.08*	*.09*
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	**.0250**	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010

Table 6-3 Proportions of the Normal Curve above the Absolute Value of Z

First digit and first decimal of Z	Second decimal of Z									
	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233

- First we look for the probability 0.025
- Then we can read the first digit and first decimal of Z value 1.9
- Lastly the second decimal of Z value 0.06
- Therefore, $z_{0.025} = 1.96$

Confidence Intervals (CI)

- Clearly, $\frac{\bar{X}-b}{SE} = -1.96$, and $\frac{\bar{X}-a}{SE} = 1.96$
- So, $a = \bar{X} - 1.96 * SE$, $b = \bar{X} + 1.96 * SE$
- Remember that $SE = \frac{\sigma}{\sqrt{n}}$
- 95% Confidence Interval is:

$$[\overline{X} - z_{0.025} \ \frac{\sigma}{\sqrt{n}}, \overline{X} + z_{0.025} \ \frac{\sigma}{\sqrt{n}}]$$

- A 95% confidence interval means that were you to perform repeated random sampling of your data, 95% of the confidence intervals drawn with each of those samples with a z-score of 1.96 will contain μ.

Confidence Intervals (CI)

- Clearly, $\frac{\bar{X}-b}{SE} = -1.96$, and $\frac{\bar{X}-a}{SE} = 1.96$
- So, $a = \bar{X} - 1.96 * SE$, $b = \bar{X} + 1.96 * SE$
- Remember that $SE = \frac{\sigma}{\sqrt{n}}$
- 95% Confidence Interval is:

$$[\overline{X} - z_{0.025}\ \frac{\sigma}{\sqrt{n}}, \overline{X} + z_{0.025}\ \frac{\sigma}{\sqrt{n}}]$$

- A 95% confidence interval means that were you to perform repeated random sampling of your data, 95% of the confidence intervals drawn with each of those samples with a z-score of 1.96 will contain μ.

  i.e., 95% of the time it will contain the true parameter μ.

Confidence Intervals (CI)

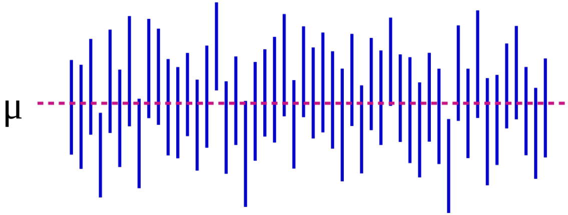
- Clearly, $\frac{\bar{X}-b}{SE} = -1.96$, and $\frac{\bar{X}-a}{SE} = 1.96$
- So, $a = \bar{X} - 1.96 * SE$, $b = \bar{X} + 1.96 * SE$
- Remember that $SE = \frac{\sigma}{\sqrt{n}}$
- 95% Confidence Interval is:

$$[\overline{X} - z_{0.025}\ \frac{\sigma}{\sqrt{n}}, \overline{X} + z_{0.025}\ \frac{\sigma}{\sqrt{n}}]$$

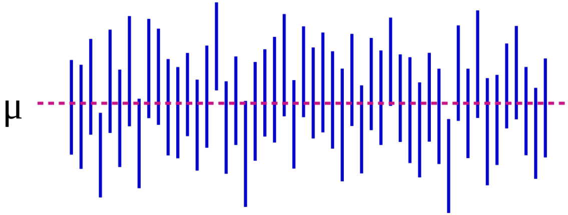
- A 95% confidence interval means that were you to perform repeated random sampling of your data, 95% of the confidence intervals drawn with each of those samples with a z-score of 1.96 will contain μ.

  i.e., 95% of the time it will contain the true parameter μ.

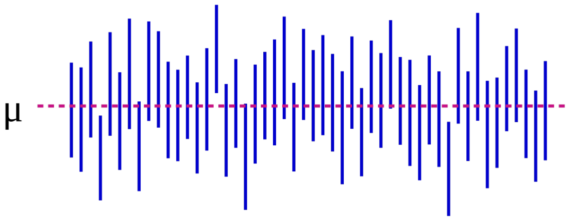
  What does this mean exactly?



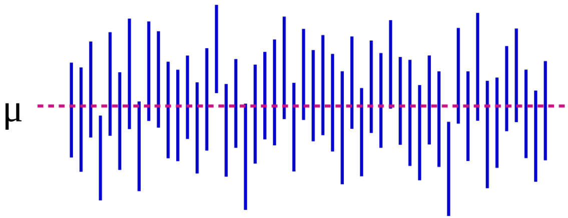
- Note that μ is a fixed number (unknown)!



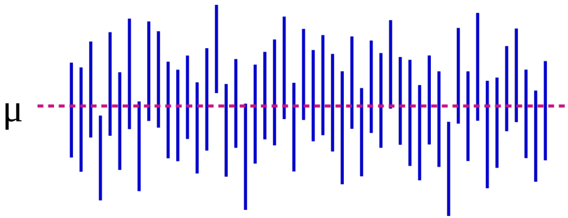
- Note that μ is a fixed number (unknown)!
- And our confidence interval varies every time we draw a sample; Why?



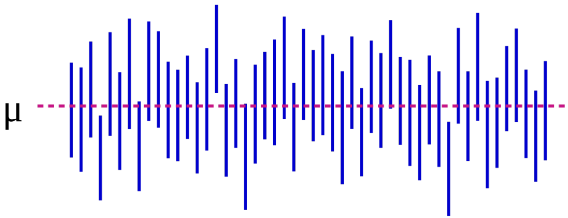
- Note that μ is a fixed number (unknown)!
- And our confidence interval varies every time we draw a sample; Why?
- Because $\overline{X}$ (cf. SE does not change) varies every time!



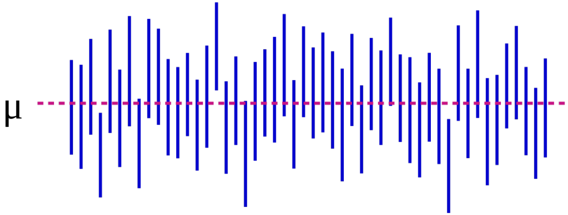
- Note that μ is a fixed number (unknown)!
- And our confidence interval varies every time we draw a sample; Why?
- Because $\overline{X}$ (cf. SE does not change) varies every time!
- In this case, we draw 50 samples and calculate 50 CIs.



- Note that μ is a fixed number (unknown)!
- And our confidence interval varies every time we draw a sample; Why?
- Because $\overline{X}$ (cf. SE does not change) varies every time!
- In this case, we draw 50 samples and calculate 50 CIs.
- How many times did these CIs capture μ?



- Note that μ is a fixed number (unknown)!
- And our confidence interval varies every time we draw a sample; Why?
- Because $\overline{X}$ (cf. SE does not change) varies every time!
- In this case, we draw 50 samples and calculate 50 CIs.
- How many times did these CIs capture μ? **47 times**!



- Note that μ is a fixed number (unknown)!
- And our confidence interval varies every time we draw a sample; Why?
- Because $\overline{X}$ (cf. SE does not change) varies every time!
- In this case, we draw 50 samples and calculate 50 CIs.
- How many times did these CIs capture μ? **47 times**!
- Therefore, $\frac{47}{50} \times 100 = 94\%$. This could be a 94% confidence interval for μ

Confidence Intervals (CI)

95% Confidence Interval is:

$$[\overline{X} - 1.96\ \frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96\ \frac{\sigma}{\sqrt{n}}]$$

As long as we know

- $\overline{X}$: the sample mean
- σ: the population standard deviation
- n: the sample size
- 1.96: z-score for the 95% interval

We can construct a confidence interval for **the true but unknown parameter** μ!

Individual Exercise 1

A quick example:

A sample of size n $= 100$ produced the sample mean of $\overline{X} = 16$. Assuming the population standard deviation $\sigma = 5$, compute a 95% confidence interval for the population mean μ

Individual Exercise 1

A quick example:

A sample of size n = 100 produced the sample mean of $\overline{X} = 16$. Assuming the population standard deviation $\sigma = 5$, compute a 95% confidence interval for the population mean μ

$$
\begin{aligned}
&[\overline{X} - 1.96\ \frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96\ \frac{\sigma}{\sqrt{n}}] \\
&= [16 - 1.96 \times \frac{5}{\sqrt{100}}, 16 + 1.96 \times \frac{5}{\sqrt{100}}] \\
&= [15.02, 16.98]
\end{aligned}
$$

95% of confidence intervals generated in this way will contain μ
i.e. The statement that [15.02, 16.98] covers μ is true with 95%

Source: http://www.utdallas.edu/ mbaron/3341/Practice12.pdf

Going back to the CI

95% Confidence Interval is:

$$[\overline{X} - z_{0.025}\ \frac{\sigma}{\sqrt{n}}, \overline{X} + z_{0.025}\ \frac{\sigma}{\sqrt{n}}]$$

- $\overline{X}$: the sample mean
- σ: the population standard deviation
- n: the sample size
- 1.96: z-score for the 95% interval

Going back to the CI

95% Confidence Interval is:

$$[\overline{X} - z_{0.025}\ \frac{\sigma}{\sqrt{n}}, \overline{X} + z_{0.025}\ \frac{\sigma}{\sqrt{n}}]$$

- $\overline{X}$: the sample mean
- σ: the population standard deviation
- n: the sample size
- 1.96: z-score for the 95% interval

In reality, we cannot use this CI... why?

Going back to the CI

95% Confidence Interval is:

$$[\overline{X} - z_{0.025}\ \frac{\sigma}{\sqrt{n}}, \overline{X} + z_{0.025}\ \frac{\sigma}{\sqrt{n}}]$$

- $\overline{X}$: the sample mean
- σ: the population standard deviation
- n: the sample size
- 1.96: z-score for the 95% interval

In reality, we cannot use this CI... why?
Because we don't know σ!

The problem of unknown σ

So we use **s** (sample standard deviation) instead of σ (population standard deviation) for constructing a 95% Confidence Interval:

$$[\overline{X} - z_{0.025}\ \frac{\mathbf{s}}{\sqrt{n}}, \overline{X} + z_{0.025}\ \frac{\mathbf{s}}{\sqrt{n}}]$$

But this is wrong! Why?

The problem of unknown σ

So we use **s** (sample standard deviation) instead of σ (population standard deviation) for constructing a 95% Confidence Interval:

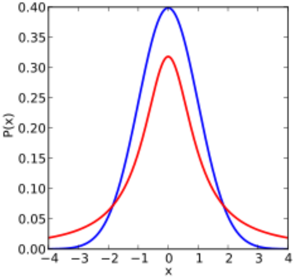
$$[\overline{X} - z_{0.025}\ \frac{\mathbf{s}}{\sqrt{n}}, \overline{X} + z_{0.025}\ \frac{\mathbf{s}}{\sqrt{n}}]$$

But this is wrong! Why?

$\overline{X}$ is not normally distributed any more since **s** is an estimate and it varies every time we draw a sample (therefore random sampling error).

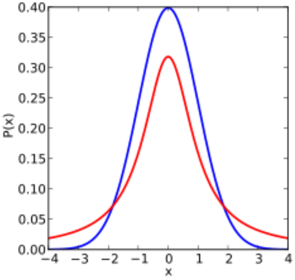
In fact, $\overline{X}$ has a **Student's t-distribution** now.

Student's t-distribution



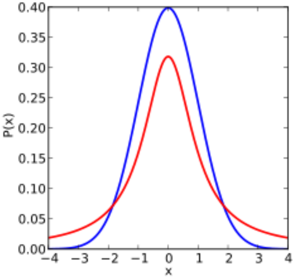
- It is very similar to a normal distribution

Student's t-distribution



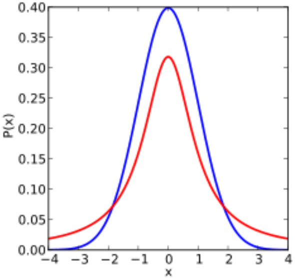
- It is very similar to a normal distribution
- It has a shorter peak and fatter tails than a normal curve

Student's t-distribution



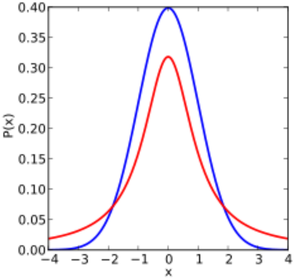
- It is very similar to a normal distribution
- It has a shorter peak and fatter tails than a normal curve
- This curve requires **only one** parameters (cf. a normal: only two)

Student's t-distribution



- It is very similar to a normal distribution
- It has a shorter peak and fatter tails than a normal curve
- This curve requires **only one** parameters (cf. a normal: only two)
- The degree of freedom $(n-1)$

Student's t-distribution



- It is very similar to a normal distribution
- It has a shorter peak and fatter tails than a normal curve
- This curve requires **only one** parameters (cf. a normal: only two)
- The degree of freedom $(n-1)$
- As n grows, t distribution looks more and more like normal

Student's t-distribution

Student's t-distribution

Mr. "Student", William Sealy Gosset

The correct CI with unknown σ

Therefore, the correct 95% CI with unknown σ is as below:

$$[\overline{X} - t_{0.025,n-1}\ \frac{\mathbf{s}}{\sqrt{n}}, \overline{X} + t_{0.025,n-1}\ \frac{\mathbf{s}}{\sqrt{n}}]$$

The previous 95% CI with known σ (an assumption)

$$[\overline{X} - z_{0.025}\ \frac{\sigma}{\sqrt{n}}, \overline{X} + z_{0.025}\ \frac{\sigma}{\sqrt{n}}]$$

But then how can we get the t score ($t_{0.025,n-1}$) for a 95% CI?

t table

- We should read the t table (Pollock p. 137)!

Table 6-4 The Student's *t*-Distribution

Degrees of freedom	Area under the curve .10	.05	.025	.01
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
13	1.350	1.771	2.160	2.650
14	1.345	1.761	2.145	2.624
15	1.341	1.753	2.131	2.602
16	1.337	1.746	2.120	2.583
17	1.333	1.740	2.110	2.567
18	1.330	1.734	2.101	2.552
19	1.328	1.729	2.093	2.539
20	1.325	1.725	2.086	2.528
21	1.323	1.721	2.080	2.518
22	1.321	1.717	2.074	2.508
23	1.319	1.714	2.069	2.500
24	1.318	1.711	2.064	2.492
25	1.316	1.708	2.060	2.485
26	1.315	1.706	2.056	2.479
27	1.314	1.703	2.052	2.473
28	1.313	1.701	2.048	2.467
29	1.311	1.699	2.045	2.462
30	1.310	1.697	2.042	2.457
40	1.303	1.684	2.021	2.423
60	1.296	1.671	2.000	2.390
90	1.291	1.662	**1.987**	2.368
100	1.290	1.660	1.984	2.364
120	1.289	1.658	1.980	2.358
1,000	1.282	1.646	1.962	2.330
Normal (Z)	1.282	1.645	1.960	2.326

Table 6-4 The Student's *t*-Distribution

Degrees of freedom	Area under the curve .10	.05	.025	.01
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

- Suppose we have a sample size $n = 10$

Table 6-4 The Student's *t*-Distribution

Degrees of freedom	Area under the curve			
	.10	*.05*	*.025*	*.01*
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

- Suppose we have a sample size $n = 10$
- Then, what's the **degree of freedom?**

Table 6-4 The Student's *t*-Distribution

Degrees of freedom	Area under the curve .10	.05	.025	.01
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

- Suppose we have a sample size $n = 10$
- Then, what's the **degree of freedom?** $n - 1 = 9$
- 95% CI means we want 0.025 in each tail (0.025 "under the curve")
  cf. $p = 0.95$ and $\frac{1-p}{2} = \frac{0.05}{2} = 0.025$
- Thus, the $t_{0.025,9}$ is **2.262**

t table

Therefore, the correct 95% CI with unknown σ when $n = 10$

$$[\overline{X} - t_{0.025,9} \frac{\mathbf{s}}{\sqrt{n}}, \overline{X} + t_{0.025,9} \frac{\mathbf{s}}{\sqrt{n}}]$$

Putting the value of $t_{0.025,9} = 2.262$ yields:

$$[\overline{X} - 2.262 \frac{\mathbf{s}}{\sqrt{n}}, \overline{X} + 2.262 \frac{\mathbf{s}}{\sqrt{n}}]$$

Individual Exercise 2

Table 6-4 The Student's *t*-Distribution

Degrees of freedom	Area under the curve			
	.10	*.05*	*.025*	*.01*
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

- Find the t score for the 90% CI when $n = 8$.
- Find the t score for the 98% CI when $n = 5$.
- Find the t score for the 80% CI when $n = 10$.

Individual Exercise 3

To assess the accuracy of a scale, a standard weight that is known to weigh 10 kg is repeatedly weighed 9 times.

The mean of the resulting measurements (in kg) is 11. And its standard deviation (in kg) is 3.

Construct a 95% confidence interval for the true parameter.

(Assume that the weighings by the scale when the true weight is 10 kg are normally distributed with mean μ)

Individual Exercise 3

1. $\overline{X} = 11$, $s = 3$, and $n = 9$
2. $\overline{X}$ (the sample mean) follows t distribution because σ is unknown
3. Let's find $t_{0.025,8}$ (the degree of freedom is $n - 1 = 8$)
4. $t_{0.025,8} = 2.306 \approx 2.3$
5. Almost done: $[\overline{X} - t_{0.025,8}\ \frac{s}{\sqrt{n}}, \overline{X} + t_{0.025,8}\ \frac{s}{\sqrt{n}}]$
6. $[11 - 2.3 \times \frac{3}{\sqrt{9}}, 11 + 2.3 \times \frac{3}{\sqrt{9}}]$
7. The 95% Confidence Interval for μ is $[8.7, 13.3]$

Key takeaways

- Confidence interval? (conceptual and mechanical)

Key takeaways

- Confidence interval? (conceptual and mechanical)
  Our conjecture about μ with $\overline{X}$

Key takeaways

- Confidence interval? (conceptual and mechanical)
  Our conjecture about μ with $\overline{X}$
  $[\overline{X} - 1.96\ \frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96\ \frac{\sigma}{\sqrt{n}}]$ (e.g. 95% CI)

Key takeaways

- Confidence interval? (conceptual and mechanical)
  Our conjecture about μ with $\overline{X}$
  $[\overline{X} - 1.96\ \frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96\ \frac{\sigma}{\sqrt{n}}]$ (e.g. 95% CI)
- When σ is unknown, we should make two changes in CI:

Key takeaways

- Confidence interval? (conceptual and mechanical)
  Our conjecture about μ with $\overline{X}$
  $[\overline{X} - 1.96\ \frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96\ \frac{\sigma}{\sqrt{n}}]$ (e.g. 95% CI)
- When σ is unknown, we should make two changes in CI:
  First, we should use s (the sample variance)

Key takeaways

- Confidence interval? (conceptual and mechanical)
  Our conjecture about μ with $\overline{X}$
  $[\overline{X} - 1.96\ \frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96\ \frac{\sigma}{\sqrt{n}}]$ (e.g. 95% CI)
- When σ is unknown, we should make two changes in CI:
  First, we should use s (the sample variance)
  Second, we should use t score instead of z score