Statistical and Mathematical Methods for Data Analysis

Dr. Faisal Bukhari

Punjab University College of Information Technology (PUCIT)

Textbooks

- **Probability & Statistics for Engineers & Scientists**, Ninth Edition, Ronald E. Walpole, Raymond H. Myer

- **Elementary Statistics: Picturing the World,** 6th Edition, Ron Larson and Betsy Farber

- **Elementary Statistics,** 13th Edition, Mario F. Triola

Reference books

- **Probability and Statistical Inference, Ninth Edition,** Robert V. Hogg, Elliot A. Tanis, Dale L. Zimmerman
- **Probability Demystified**, Allan G. Bluman
- **Practical Statistics for Data Scientists: 50 Essential Concepts,** Peter Bruce and Andrew Bruce
- **Schaum's Outline of Probability,** Second Edition, Seymour Lipschutz, Marc Lipson
- **Python for Probability, Statistics, and Machine Learning,** José Unpingco

References

Readings for these lecture notes:

- Probability & Statistics for Engineers & Scientists, Ninth edition, Ronald E. Walpole, Raymond H. Myer

- Elementary Statistics, Tenth Edition, Mario F. Triola

These notes contain material from the above resources.

Point Estimate vs. Confidence Interval

Definition

A **point estimate** is a single value (or point) used to approximate a population parameter.

Definition

A **confidence interval** (or **interval estimate**) is a range (or an interval) of values used to estimate the true value of a population parameter. A confidence interval is sometimes abbreviated as **CI**.

Confidence Level [1]

- **Confidence Level:** It is the probability **1 - α** (often expressed as the equivalent percentage value) that is the **proportion of times** that the **confidence interval** actually does contain the **population parameter**, assuming that **the estimation process is repeated** a large number of times.

- The confidence level is also called the **degree of confidence**, or the **confidence coefficient**.

Confidence Level [2]

- The most common choices for the confidence level are 90% (**with α = 0.10**), 95% (**with α = 0.05**), and 99% (**with α = 0. 01**).

- The **choice of 95%** is most common because it provides a good balance between precision (as reflected in the width of the confidence interval) and reliability (as expressed by the confidence level).

Area under the Normal Curve [2]

Table A.3 (continued) Areas under the Normal Curve

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

1. When $\alpha = 0.05$

$z_{\alpha/2} = z_{0.0250} = 1.96$ $\because 1 - \alpha/2 = 1 - 0.250 = 0.9750$

2. When $\alpha = 0.01$

$z_{\alpha/2} = z_{0.005} = 2.575$ $\because 1 - \alpha/2 = 1 - 0.005 = 0.9950$

3. When $\alpha = 0.10$

$z_{\alpha/2} = z_{0.05} = 1.645$ $\because 1 - \alpha/2 = 1 - 0.05 = 0.9500$

Why Do We Need Confidence Intervals?

- We have **no indication** of just how ***good*** our best (points) estimate is.

- Because a **point estimate** has the **serious flaw of not revealing anything** about **how good it is**, statisticians have cleverly developed another type of estimate.

Interpreting a Confidence Interval[1]

- Here's an example of a confidence interval based on the sample data of **280 trials** of touch therapists, with 44% of the trials resulting in correct identification sof the hand that was selected:

The 0.95 (or 95%) confidence interval estimate of the population proportion *p* is

$$0.381 < p < 0.497$$

Interpreting a Confidence Interval[2]

- We must be careful to **interpret confidence intervals** correctly.

- There is a correct interpretation and many different and creative wrong interpretations of the confidence interval

$0.381 < p < 0.497$.

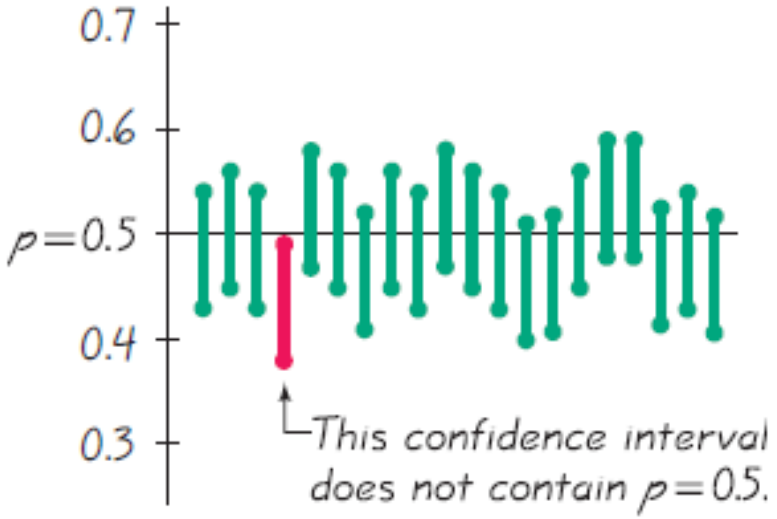
- ***Correct:*** "We are **95%** confident that the interval from **0.381 to 0.497** actually **does contain the true value of *p***." This means that if we were to select many different **samples of size 280** and construct the corresponding confidence intervals, **95% of them** would actually contain the **value of the population proportion *p*.** (Note that in this correct interpretation, the level of 95% refers to the success rate of the *process* being used to estimate the proportion, and it does not refer to the population proportion itself.)
- ***Wrong:*** "There is a 95% chance that the true value of *p* will fall **between 0.381 and 0.497**."

- At **any specific point in time**, a population has a fixed and **constant value *p***, and a confidence interval constructed from a sample **either includes *p* or does not**.
- Similarly, if a **baby** has **just been born** and the doctor is about to announce **its gender**, it's **wrong** to say that there is **probability of 0.5** that the **baby** is a **girl**; the baby is a **girl or is not**, and there's no probability involved.
- A **population proportion *p*** is like the baby that has been born—the value of ***p* is fixed**, so the **confidence interval limits** either **contain p or do not**, and that is why it's wrong to say that there is a **95% chance** that *p* will fall between values such as **0.381 and 0.497**.

A **confidence level of 95%** tells us that the *process* we are using will, in the long run, result in confidence interval limits that contain the **true population proportion 95%** of the time. Suppose that the true proportion of all correct hand identifications made by touch therapists is $p = 0.5$.

Then the confidence interval obtained from the given sample data would not contain the population proportion, because the true population proportion **0.5** is **not between 0.381 and 0.497**. this with 19 of the confidence intervals containing *p*, while one confidence interval does not contain *p*.

This is illustrated in figure below. It shows typical confidence intervals resulting from **20 different samples**. With **95% confidence**, we expect that 1**9 out of 20** samples should result in confidence intervals that do contain the true value of *p*, and the figure illustrates



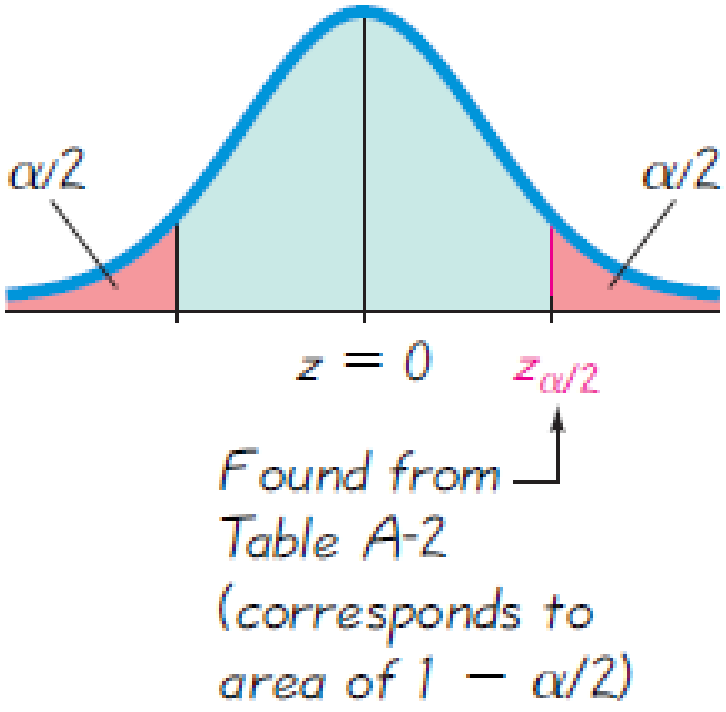
Confidence Intervals from 20 Different Samples

Critical Values [1]

- A **critical value** is the number on the borderline separating sample statistics that **are likely** to occur from those that are **unlikely to occur.**

- The number $z_{\alpha/2}$ is a **critical value** that is a ***z* score** with the property that it separates an area of **$\alpha/2$** in the right tail of the standard normal distribution.

Critical Values[2]



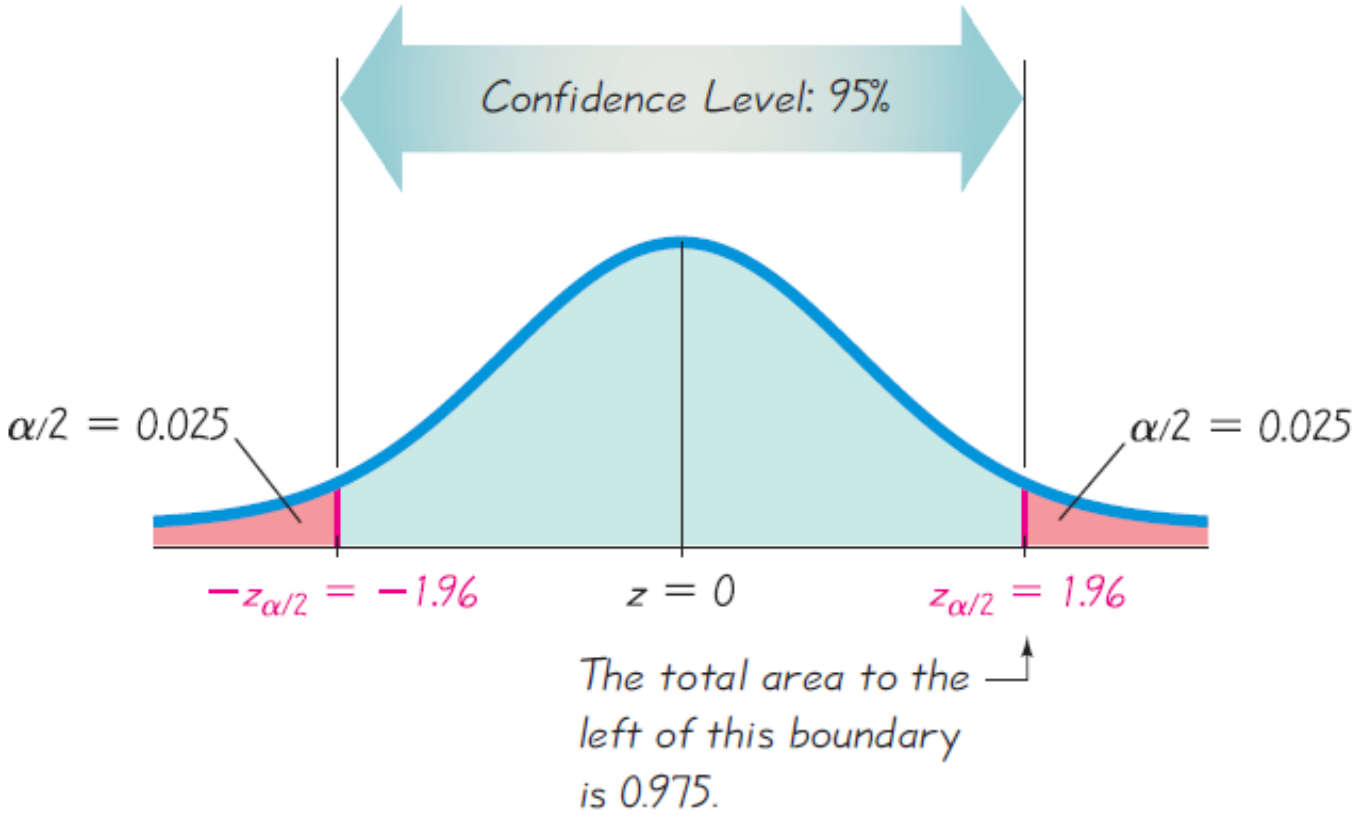
Critical Value $z_{\alpha/2}$ in the Standard Normal Distribution

EXAMPLE Finding a Critical Value Find the critical value $z_{\alpha/2}$ corresponding to a 95% confidence level.

Solution

When $\alpha = 0.05$

$z_{\alpha/2} = z_{0.0250} = 1.96$ $\quad \because 1 - \alpha/2 = 1 - 0.250 = 0.9750$



Confidence Level	α	Critical values, $z_{\alpha/2}$
90%	0.10	1.645
95%	0.05	1.96
99%	0.01	2.575

Margin of Error

- When data from a simple random sample are used to estimate a population proportion *p*, the **margin of error**, denoted by ***E or e***, is the maximum likely (with probability **1 - α**) difference between the observed sample proportion $\hat{p}$ and the true value of the **population proportion *p***.
- The **margin of error *E*** is also called the ***maximum error of the estimate*** and can be found by multiplying the critical value and the standard deviation of sample proportions, as shown

$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Confidence Interval (or Interval Estimate) for the Population Proportion *p*

$\hat{p} - E < p < \hat{p} + E$ *where* $E = z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$

The confidence interval is often expressed in the following equivalent formats:

$\hat{p} \pm E$

or

$(\hat{p} - E, \hat{p} + E)$

Determining Sample Size

- Suppose we want to **collect sample data** with the objective of **estimating some population proportion**. How do we know *how* ***many*** **sample items** must be obtained?
- If we take the expression for the **margin of error**

$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$, then **solve for *n***, we get

$$\boldsymbol{n = \frac{\hat{p}\hat{q}\, z^2_{\alpha/2}}{E^2}}$$

- If no such estimate is known (as is often the case), we replace $\hat{p}$ by 0.5 and replace $\hat{q}$ by 0.5

$$\mathbf{n = \frac{0.25\, z^2_{\alpha/2}}{E^2}}$$

Sample Size for Estimating Proportion *p*

When an estimate $\hat{p}$ is known:

$$n = \frac{\hat{p}\hat{q}\, z^2_{\alpha/2}}{E^2}$$

When no estimate $\hat{p}$ is known:

$$n = \frac{0.25\, z^2_{\alpha/2}}{E^2}$$

Round-Off Rule for Determining Sample Size

- In order to ensure that the required **sample size** is at least as large as it should be, if the computed sample size is **not a whole number**, round it up to the **next *higher* whole number**.

Population Size

- Many people **incorrectly believe** that the sample size should be **some percentage of the population**, but $\mathbf{n = \frac{0.25\ z^2_{\alpha/2}}{E^2}}$, shows that the population size is irrelevant.

- In reality, the **population size** is sometimes used, but only in cases in which we **sample without replacement** from a **relatively small population**.

- Polls commonly use **sample sizes** in the range of **1000 to 2000** and, even though such polls may involve a very small percentage of the **total population**, they can provide results that are quite good.

S^2(Biased sample variance) vs s^2(Unbiased sample variance)

$$1.\ S^2 = \frac{\sum(x-\bar{x})^2}{n} \text{ or } S^2 = \frac{(n\sum_{i=1}^{n} x^2 - (\sum_{i=1}^{n} x)^2}{n^2}$$

$$2.\ s^2 = \frac{\sum(x-\bar{x})^2}{n-1} \text{ or } s^2 = \frac{1}{n(n-1)}\{n\sum_{i=1}^{n} x^2 - (\sum_{i=1}^{n} x)^2\}$$

Where $\sum(x-\bar{x})^2 = \sum_{i=1}^{n} x^2 - \frac{(\sum_{i=1}^{n} x)^2}{n} = \frac{(n\sum_{i=1}^{n} x^2 - (\sum_{i=1}^{n} x)^2}{n}$

The Case of σ Unknown [1]

Frequently, we must attempt to estimate the mean of a population when the **variance is unknown**. If we have a random sample from a normal distribution, then the random variable

$$t = \frac{\bar{X} - \mu}{s / \sqrt{n}}$$

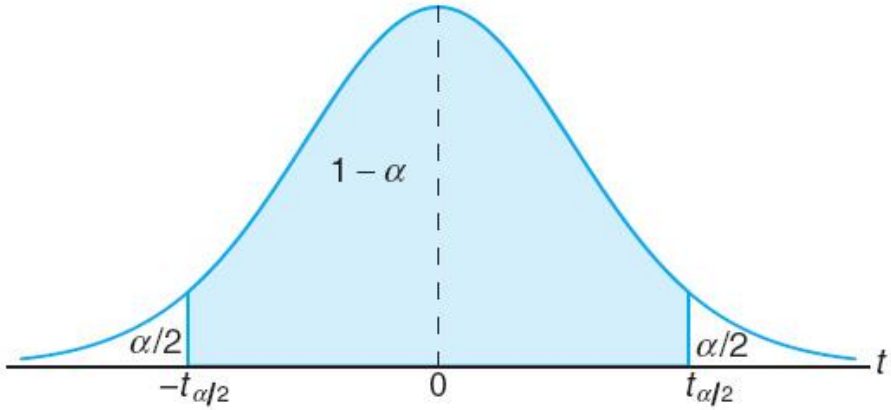
has a **Student t-distribution** with **n – 1 degrees of freedom**. Here **s** is the sample standard deviation. In this situation, with **σ unknown**, T can be used to construct a confidence interval on μ.

The Case of σ Unknown [2]

$$P(-t_{\alpha/2} < T < t_{\alpha/2}) = 1 - \alpha, \text{ where } T = \frac{\bar{x}-\mu}{s/\sqrt{n}}$$

$$\Longrightarrow P(-t_{\alpha/2} < \frac{\bar{x}-\mu}{s/\sqrt{n}} < t_{\alpha/2}) = 1 - \alpha$$

$$\Longrightarrow P(\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}) = 1 - \alpha$$



The Case of σ Unknown [3]

If $\bar{x}$ and s are the mean and standard deviation of a random sample from a normal population with unknown variance σ^2, a 100(1−α)% confidence interval for μ is

$$\bar{x} - t_{(\alpha/2,n-1)} \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{(\alpha/2,\, n-1)} \frac{s}{\sqrt{n}}$$

OR

$$C.I = \bar{x} \pm t_{(\alpha/2,n-1)} \frac{s}{\sqrt{n}}$$

$$\text{where } s^2 = \frac{\sum (x-\bar{x})^2}{n-1}$$

The Case of σ Unknown [4]

$$\sum(x-\bar{x})^2 = \sum_{i=1}^{n} x^2 - \frac{(\sum_{i=1}^{n} x)^2}{n} = \frac{n\sum_{i=1}^{n} x^2 - (\sum_{i=1}^{n} x)^2}{n}$$

$$\text{or } s^2 = \frac{1}{n(n-1)}\{n\sum_{i=1}^{n} x^2 - (\sum_{i=1}^{n} x)^2\}$$

where $t_{\alpha/2}$ is the t- value with n – 1 degrees of freedom, leaving an area of $\alpha/2$ to the right.

The Case of σ Unknown [5]

We have made a distinction between the cases of ***σ* known** and ***σ* unknown** in computing confidence interval estimates. We should emphasize that for ***σ* known** we exploited the **Central Limit Theorem**, whereas for ***σ* unknown** we made use of the **sampling distribution** of the **random variable *T***.

However, the use of the *t* distribution is based on the premise that the **sampling** is from a **normal distribution**. As long as the distribution is approximately bell shaped, confidence intervals can be computed when σ^2 is unknown by using the *t*-distribution and we may expect very good results.

One-Sided Confidence Bounds on μ, σ^2 unknown [1]

If $\overline{X}$ is the mean of a random sample of size n from a population with unknown variance σ^2, the one-sided $100(1 - \alpha)\%$ confidence bounds for μ are given by

upper one-sided bound: $\overline{x} + t_{(\alpha, n-1)} \frac{s}{\sqrt{n}}$

lower one-sided bound: $\overline{x} - t_{(\alpha, n-1)} \frac{s}{\sqrt{n}}$

Critical Values of the t-Distribution

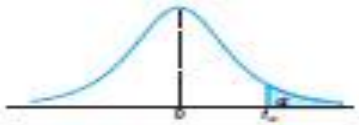


Table A.4 Critical Values of the t-Distribution

	α						
v	0.40	0.30	0.20	0.15	0.10	0.05	0.025
1	0.325	0.727	1.376	1.963	3.078	6.314	12.706
2	0.289	0.617	1.061	1.386	1.886	2.920	4.303
3	0.277	0.584	0.978	1.250	1.638	2.353	3.182
4	0.271	0.569	0.941	1.190	1.533	2.132	2.776
5	0.267	0.559	0.920	1.156	1.476	2.015	2.571
6	0.265	0.553	0.906	1.134	1.440	1.943	2.447
7	0.263	0.549	0.896	1.119	1.415	1.895	2.365
8	0.262	0.546	0.889	1.108	1.397	1.860	2.306
9	0.261	0.543	0.883	1.100	1.383	1.833	2.262
10	0.260	0.542	0.879	1.093	1.372	1.812	2.228
11	0.260	0.540	0.876	1.088	1.363	1.796	2.201
12	0.259	0.539	0.873	1.083	1.356	1.782	2.179
13	0.259	0.538	0.870	1.079	1.350	1.771	2.160
14	0.258	0.537	0.868	1.076	1.345	1.761	2.145
15	0.258	0.536	0.866	1.074	1.341	1.753	2.131
16	0.258	0.535	0.865	1.071	1.337	1.746	2.120
17	0.257	0.534	0.863	1.069	1.333	1.740	2.110
18	0.257	0.534	0.862	1.067	1.330	1.734	2.101
19	0.257	0.533	0.861	1.066	1.328	1.729	2.093
20	0.257	0.533	0.860	1.064	1.325	1.725	2.086
21	0.257	0.532	0.859	1.063	1.323	1.721	2.080
22	0.256	0.532	0.858	1.061	1.321	1.717	2.074
23	0.256	0.532	0.858	1.060	1.319	1.714	2.069
24	0.256	0.531	0.857	1.059	1.318	1.711	2.064
25	0.256	0.531	0.856	1.058	1.316	1.708	2.060
26	0.256	0.531	0.856	1.058	1.315	1.706	2.056
27	0.256	0.531	0.855	1.057	1.314	1.703	2.052
28	0.256	0.530	0.855	1.056	1.313	1.701	2.048
29	0.256	0.530	0.854	1.055	1.311	1.699	2.045
30	0.256	0.530	0.854	1.055	1.310	1.697	2.042
40	0.255	0.529	0.851	1.050	1.303	1.684	2.021
60	0.254	0.527	0.848	1.045	1.296	1.671	2.000
120	0.254	0.526	0.845	1.041	1.289	1.658	1.980
∞	0.253	0.524	0.842	1.036	1.282	1.645	1.960

Critical Values of the t-Distribution

	α						
v	0.02	0.015	0.01	0.0075	0.005	0.0025	0.0005
1	15.894	21.205	31.821	42.433	63.656	127.321	636.578
2	4.849	5.643	6.965	8.073	9.925	14.089	31.600
3	3.482	3.896	4.541	5.047	5.841	7.453	12.924
4	2.999	3.298	3.747	4.088	4.604	5.598	8.610
5	2.757	3.003	3.365	3.634	4.032	4.773	6.869
6	2.612	2.829	3.143	3.372	3.707	4.317	5.959
7	2.517	2.715	2.998	3.203	3.499	4.029	5.408
8	2.449	2.634	2.896	3.085	3.355	3.833	5.041
9	2.398	2.574	2.821	2.998	3.250	3.690	4.781
10	2.359	2.527	2.764	2.932	3.169	3.581	4.587
11	2.328	2.491	2.718	2.879	3.106	3.497	4.437
12	2.303	2.461	2.681	2.836	3.055	3.428	4.318
13	2.282	2.436	2.650	2.801	3.012	3.372	4.221
14	2.264	2.415	2.624	2.771	2.977	3.326	4.140
15	2.249	2.397	2.602	2.746	2.947	3.286	4.073
16	2.235	2.382	2.583	2.724	2.921	3.252	4.015
17	2.224	2.368	2.567	2.706	2.898	3.222	3.965
18	2.214	2.356	2.552	2.689	2.878	3.197	3.922
19	2.205	2.346	2.539	2.674	2.861	3.174	3.883
20	2.197	2.336	2.528	2.661	2.845	3.153	3.850
21	2.189	2.328	2.518	2.649	2.831	3.135	3.819
22	2.183	2.320	2.508	2.639	2.819	3.119	3.792
23	2.177	2.313	2.500	2.629	2.807	3.104	3.768
24	2.172	2.307	2.492	2.620	2.797	3.091	3.745
25	2.167	2.301	2.485	2.612	2.787	3.078	3.725
26	2.162	2.296	2.479	2.605	2.779	3.067	3.707
27	2.158	2.291	2.473	2.598	2.771	3.057	3.689
28	2.154	2.286	2.467	2.592	2.763	3.047	3.674
29	2.150	2.282	2.462	2.586	2.756	3.038	3.660
30	2.147	2.278	2.457	2.581	2.750	3.030	3.646
40	2.123	2.250	2.423	2.542	2.704	2.971	3.551
60	2.099	2.223	2.390	2.504	2.660	2.915	3.460
120	2.076	2.196	2.358	2.468	2.617	2.860	3.373
∞	2.054	2.170	2.326	2.432	2.576	2.807	3.290

Example: The contents of seven similar containers of sulfuric acid are **9.8, 10.2, 10.4, 9.8, 10.0, 10.2**, and **9.6** liters. Find a **95%** confidence interval for the mean contents of all such containers, assuming an approximately normal distribution.

Solution:

$n = 7$

$\sum x = 70$

$\Rightarrow \bar{x} = \sum x \ / \ n = 70 / 7 = 10$

$\sum (x - \bar{x})^2 = 0.4800,$

$s^2 = \frac{\sum (x - \bar{x})^2}{n-1} = .48 / 6 = 0.0800$

$\Rightarrow$ **s = 0.28,** v = n – 1 = 6 degrees of freedom

$t_{(0.025,\ 6)} = 2.447$

95% confidence interval for μ is

$$\bar{x} - t_{(\alpha/2,\, n-1)} \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{(\alpha/2,\, n-1)} \frac{s}{\sqrt{n}}$$

$$\Rightarrow 10.0 - \frac{(2.447)(0.283)}{\sqrt{7}} < \mu < 10.0 + \frac{(2.447)(0.283)}{\sqrt{7}}$$

$$\Rightarrow 9.74 < \mu < 10.26.$$