Stat 401, section 7.3 Intervals Based on a Normal Population Distribution

notes by Tim Pilachowski

In sections 7.1 and 7.2, we relied on the Central Limit Theorem for cases where n was "large enough" for the shape of sampling distribution to be close enough to a normal distribution to be able to use the normal distribution table to answer questions about population parameters. [Recall that our rules of thumb are $n > 30$ for situations where σ is known, and $n > 40$ for situations where σ is not known.]

But, what if the sample size n is not large enough?

Back in section 5.4, we (and the text) noted that, given a population which has a normal distribution, the resulting sampling distributions for any sample size n will retain the symmetry of the population distribution. We will retain this idea as our basic assumption: "The population of interest is normal, so that $X_1, \ldots, X_n$, constitutes a random sample from a normal distribution with both μ and σ unknown."

[The text notes that we could begin with other types of populations, but that, in practice, researchers assume a normally distributed population more often than any other type.]

When n is large enough, the random variable $Z = \dfrac{\overline{X} - \mu}{S/\sqrt{n}}$ has approximately a standard normal distribution.

That is, the size of n helps to ameliorate the effect of having two random variables present in our transformation formula.

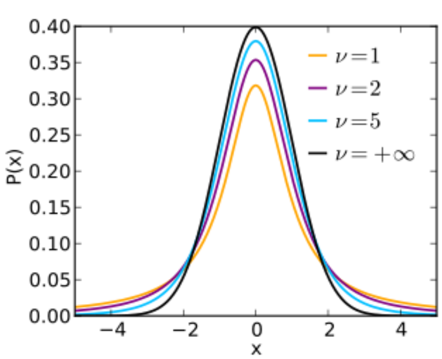
When n is small, the additional variability resulting from using S in the denominator means that the sampling distribution will be more spread out than a normal distribution is.

At this juncture, we get to rely on work done by others to present a Theorem.

Theorem: When $\overline{X}$ is the mean of a random sample of size n from a normal distribution with mean μ, the random variable $T = \dfrac{\overline{X} - \mu}{S/\sqrt{n}}$ has a probability distribution called a t distribution with $n - 1$ degrees of freedom.

In 1908, a worker at the Guinness Brewery in Dublin, Ireland, William Sealy Gosset, published a paper in *Biometrika* using the pseudonym "Student", in which he discusses the "frequency distribution of standard deviations of samples drawn from a normal population". Gosset was addressing the issue of small samples, for example, the chemical properties of barley where sample sizes might be as low as 3. Gosset's work was developed by Ronald A. Fisher, who called the distribution "Student's distribution" and gave the random variable the designation t.

The probability density function of Student's t-distribution is $\dfrac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\dfrac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$ where ν is the number of degrees of freedom (often abbreviated df) and Γ is the gamma function. (See section 4.4.)



The probability density function is symmetric, centered at 0, with an overall shape that is similar to the bell shape of a standard normal distribution, except that it is a bit lower and wider. As the sample size, and number of degrees of freedom, gets larger, the t-distribution approaches the standard normal distribution with mean 0 and variance 1. The z curve is often called the t-curve with df = ∞.

The t-distribution can be used with any statistic having a bell-shaped distribution (i.e., approximately normal).

Why use df = $\nu = n - 1$? The random variable S is calculated using the n deviations $\left(X_i - \overline{X}\right)$. But since $\sum_{i=1}^{n}\left(X_i - \overline{X}\right) = 0$, only $n - 1$ of these are "freely determined".

(An analogy: Think about pulling names from a hat.)

When calculating large sample confidence intervals, we only needed $\alpha/2$ or α to determine critical values for z. When using the t-distribution, we also need to know the degrees of freedom. We'll use the notations $t_{\alpha/2,\nu}$ and $t_{\alpha,\nu}$ for our needed critical values.

If we use our basic assumption, that a population has a normal distribution, we can use random variable T to develop a confidence interval formula for small sample situations.

notes on the proof:

$$
\begin{aligned}
& P\left(\overline{X} - t_{\alpha/2,\nu}\left(\tfrac{S}{\sqrt{n}}\right) < \mu, \quad \mu < \overline{X} + t_{\alpha/2,\nu}\left(\tfrac{S}{\sqrt{n}}\right)\right) \\
&= P\left(\overline{X} - \mu < t_{\alpha/2,\nu}\left(\tfrac{S}{\sqrt{n}}\right), \quad -t_{\alpha/2,\nu}\left(\tfrac{S}{\sqrt{n}}\right) < \overline{X} - \mu \right) \\
&= P\left(\overline{X} - \mu < t_{\alpha/2,\nu}\left(\tfrac{S}{\sqrt{n}}\right), \quad \overline{X} - \mu > -t_{\alpha/2,\nu}\left(\tfrac{S}{\sqrt{n}}\right) \right) \\
&= P\left(\frac{\overline{X} - \mu}{S/\sqrt{n}} < t_{\alpha/2,\nu}, \quad \frac{\overline{X} - \mu}{S/\sqrt{n}} > -t_{\alpha/2,\nu} \right) \\
&= P\left(T < t_{\alpha/2,\nu}, \quad T > -t_{\alpha/2,\nu} \right) \\
&= P\left(-t_{\alpha/2,\nu} < T < t_{\alpha/2,\nu} \right) \\
&= 1 - \alpha
\end{aligned}
$$

Note also the similarity of the small-sample confidence interval formula to the large-sample formula. The only difference is the source of the critical value.

Example A: A sample, taken from a population with a normal distribution, has mean = 150 and standard deviation = 22. For a random sample of size 28, find the 90% two-sided confidence interval.

work:

interpretation:

Example B. In their 1992 study of human internal body temperature, Mackowiak, Wasserman and Levine, their sample mean, $\bar{x}$ = 98.25 with s = 0.73, led them to a hypothesis that human internal body temperature is lower than the conventional 98.6º F. There is reason to believe that human internal body temperature has a normal probability distribution. Suppose that their sample size was only 38. Does a one-sided 95% confidence interval based on their sample data include 98.6º F?

work:

interpretation:

In some applications, the goal is to be able to predict a future value rather than to estimate the mean of the population.

Proposition: A **prediction interval** for a single observation to be selected from a normal population distribution, with prediction level 100(1 – α)% is $\bar{x} \pm t_{\alpha/2,\nu} * s\sqrt{1+\frac{1}{n}}$.

To prove our proposition, we get to rely on work done by others to present another Theorem.

Theorem: The random variable $T = \dfrac{\bar{X} - X_{n+1}}{S * \sqrt{1+\frac{1}{n}}}$ has a *t*-distribution with n – 1 degrees of freedom.

notes on the proof:

$$
\begin{aligned}
& P\left(\bar{X} - t_{\alpha/2,\nu} * S\sqrt{1+\tfrac{1}{n}} < X_{n+1} ,\ X_{n+1} < \bar{X} + t_{\alpha/2,\nu} * S\sqrt{1+\tfrac{1}{n}} \right) \\
&= P\left(\bar{X} - X_{n+1} < t_{\alpha/2,\nu} * S\sqrt{1+\tfrac{1}{n}} ,\ -t_{\alpha/2,\nu} * S\sqrt{1+\tfrac{1}{n}} < \bar{X} - X_{n+1} \right) \\
&= P\left(\bar{X} - X_{n+1} < t_{\alpha/2,\nu} * S\sqrt{1+\tfrac{1}{n}} ,\ \bar{X} - X_{n+1} > -t_{\alpha/2,\nu} * S\sqrt{1+\tfrac{1}{n}} \right) \\
&= P\left(\frac{\bar{X} - X_{n+1}}{S\sqrt{1+\frac{1}{n}}} < t_{\alpha/2,\nu} ,\ \frac{\bar{X} - X_{n+1}}{S\sqrt{1+\frac{1}{n}}} > -t_{\alpha/2,\nu} \right) \\
&= P\left(T < t_{\alpha/2,\nu} ,\ T > -t_{\alpha/2,\nu} \right) \\
&= P\left(-t_{\alpha/2,\nu} < T < t_{\alpha/2,\nu} \right) \\
&= 1 - \alpha
\end{aligned}
$$

Example A revisited: A sample, taken from a population with a normal distribution, has mean = 150 and standard deviation = 22. For a random sample of size 28, find the 90% two-sided prediction interval.

work:

interpretation:

comparison:

Three things:

1) A one-sided prediction interval can be calculated by using critical value $t_{\alpha,\nu}$ and only "+" [upper bound] or only "–" [lower bound] in place of "+/–" in the prediction interval formula used above.

2) A large-sample prediction interval can be calculated by using the formula above with critical *z*-value in place of the critical *t*-value in the prediction interval formula used above.

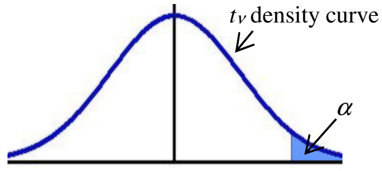
3) How do you know when to use small sample or large sample formulas?

If the sample size is greater than 40, then the sampling distribution of a statistic will be normal or nearly normal, and will be considered a large sample case.

If the population distribution is normal and the sample size is less than 40, then the sampling distribution of a statistic will be a *t*-distribution, and will be considered a small sample case.

Also, in practice, if the sampling distribution is symmetric, unimodal, and without outliers, then the sampling distribution of a statistic will be close to a *t*-distribution, and may be considered a small sample case. (Some investigation and statistical justification would be necessary in this case.)

Appendix Table A.5
Critical Values for t Distributions



	α						
ν	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
1	3.078	6.314	12.706	31.821	63.657	318.31	636.62
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
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
32	1.309	1.694	3.037	2.449	2.738	3.365	3.622
34	1.307	1.691	2.032	2.441	2.728	3.348	3.601
36	1.306	1.688	2.028	2.434	2.719	3.333	3.582
38	1.304	1.686	2.024	2.429	2.712	3.319	3.566
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
50	1.299	1.676	2.009	2.403	2.678	3.262	3.496
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