Recap.

- $X \sim N[\mu, \sigma^2]$, σ^2 known.

  $100(1-\alpha)\%$ CI for μ: $\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

  - <u>Pivot:</u> $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$.

- - sample size formula: $n \geq \left[\frac{2 Z_{\alpha/2} \sigma}{W} \right]^2$ (Round up).

$100(1-\alpha)\%$. $X \sim N[\mu, \sigma^2]$, σ^2 unknown.

Result: CI for μ: $\overline{X} \pm t_{\alpha/2,n-1} \underbrace{S/\sqrt{n}}_{\widehat{se}(\overline{X})}$

Proof:

$t_{\alpha/2,n-1}$ = $(1-\alpha/2)$th quantile of t_{n-1} dist.

$P[t_{n-1} \geq t_{\alpha/2,n-1}] = \alpha/2$

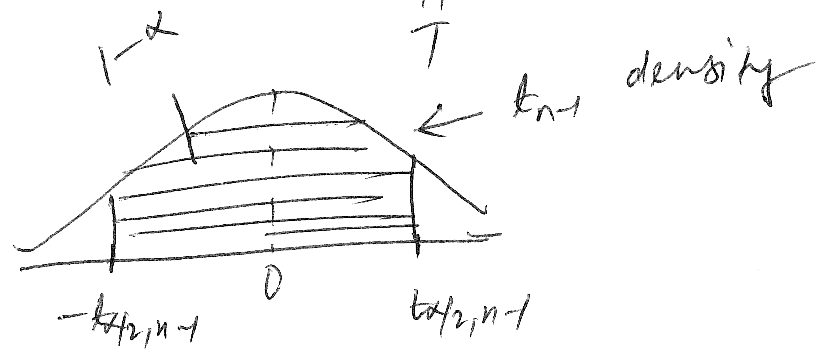
<u>Pivot:</u> $T = \dfrac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$

Need to verify that: coverage prob. of this CI $= 1-\alpha$.

$$P\left[\bar{X} - t_{\alpha/2, n-1}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right]$$

$$= P\left[-t_{\alpha/2,n-1} \leq \underbrace{\frac{\bar{X}-\mu}{S/\sqrt{n}}}_{=T} \leq t_{\alpha/2,n-1}\right].$$

$=$



$= 1-\alpha$. for all (μ, σ^2).

Note: If n is large: $t_{\alpha/2,n-1} \approx z_{\alpha/2}$

- The t critical points are tabulated in the t-table. Alternatively, we can use `qt` function in `R`.
- Sample size calculation now becomes complicated than before because S needs to be known before data are collected.
- One option is to make an intelligent guess about S and be conservative (guess a larger value of S so that n larger than necessary is chosen).

Ex: If an unauthorized person accesses a computer account with the correct username and password (stolen or cracked), can this intrusion be detected? One way to do this is to compare mean time between keystrokes of the user trying to log in with that of the account owner. The intrusion is detected if there is a noticeable difference. The following times between keystrokes (in seconds) were recorded when a user typed the username and password:

$$0.46, 0.38, 0.31, 0.24, 0.20, 0.31, 0.34, 0.42, 0.09, 0.18, 0.46, 0.21$$

Find a 95% CI for mean time between keystrokes for the user trying to log in. Assume a normal distribution for the times.

$X \sim N[\mu, \sigma^2] \to$ unknown $\Rightarrow$ t-interval.

X – time b/w keystroke for user trying to log in

$\bar{X} \pm t_{\alpha/2, n-1} \cdot \frac{S}{\sqrt{n}}$

```
x <- c(0.46, 0.38, 0.31, 0.24, 0.20, 0.31, 0.34,
0.42, 0.09, 0.18, 0.46, 0.21)
#> mean(x)
#[1] 0.3
#> sd(x)
#[1] 0.1183216
#>qt(0.975, 11)
#[1] 2.200985
```

$n=12$

$\bar{x}$

s

d.f.

$1-\alpha/2$ for 95% confidence

95% CI: $0.3 \pm 2.2 \frac{0.118}{\sqrt{12}} = 0.3 \pm 0.07 = [0.23, 0.37]$ s.

plausible values of μ.

(pop) mean time b/w keystrokes for ~~the~~ user trying to log in.

suppose μ account owner $= 0.4$ sec

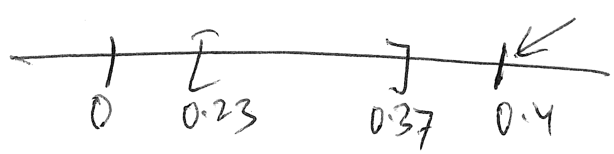


Table A5. Table of Student's T-distribution

t_α; critical values, such that $P\{t > t_\alpha\} = \alpha$ Think of this as '$\alpha/2$' for constructing CI.

ν (d.f.)	α, the right-tail probability									
	.10	.05	.025	.02	.01	.005	.0025	.001	.0005	.0001
1	3.078	6.314	12.706	15.89	31.82	63.66	127.3	318.3	636.6	3185
2	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60	70.71
3	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92	22.20
4	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610	13.04
5	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.894	6.869	9.676
6	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959	8.023
7	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408	7.064
8	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041	6.442
9	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781	6.009
10	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587	5.694
11	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437	5.453
12	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318	5.263
13	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221	5.111
14	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140	4.985
15	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073	4.880
16	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015	4.790
17	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965	4.715
18	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.610	3.922	4.648
19	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883	4.590
20	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850	4.539
21	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819	4.492
22	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792	4.452
23	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768	4.416
24	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745	4.382
25	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725	4.352
26	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707	4.324
27	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.689	4.299
28	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674	4.276
29	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.660	4.254
30	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646	4.234
32	1.309	1.694	2.037	2.141	2.449	2.738	3.015	3.365	3.622	4.198
34	1.307	1.691	2.032	2.136	2.441	2.728	3.002	3.348	3.601	4.168
36	1.306	1.688	2.028	2.131	2.434	2.719	2.990	3.333	3.582	4.140
38	1.304	1.686	2.024	2.127	2.429	2.712	2.980	3.319	3.566	4.115
40	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551	4.094
45	1.301	1.679	2.014	2.115	2.412	2.690	2.952	3.281	3.520	4.049
50	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496	4.014
55	1.297	1.673	2.004	2.104	2.396	2.668	2.925	3.245	3.476	3.985
60	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460	3.962
70	1.294	1.667	1.994	2.093	2.381	2.648	2.899	3.211	3.435	3.926
80	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416	3.899
90	1.291	1.662	1.987	2.084	2.368	2.632	2.878	3.183	3.402	3.878
100	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390	3.861
200	1.286	1.653	1.972	2.067	2.345	2.601	2.838	3.131	3.340	3.789
∞	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.090	3.290	3.719

Here: $X \sim$ some distribution with mean μ.

Large sample CI for mean μ

Recall: When n is large, an approximate $100(1-\alpha)\%$ CI for mean μ of any population is : $\bar{X} \pm z_{\alpha/2}\, s/\sqrt{n}$

Pivot: $Z = \frac{\bar{X}-\mu}{s/\sqrt{n}} \approx N(0,1)$ if n is large.

Ex: We wish to estimate the mean execution time of a program. The program was run 35 times on randomly selected inputs, and the sample mean and the sample standard deviation of the execution times were evaluated as 230 ms ($\bar{x}$) and 14 ms, (s) respectively. Find a 95% CI for the true mean execution time μ.

X - don't know the distn. but $n = 35 \Rightarrow$ large-sample CI.

'' Execution time of the program

$$\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}} = 230 \pm 1.96 \frac{14}{\sqrt{35}}$$

$$= [\quad , \quad] \text{ ms.}$$

Large sample CI for success proportion p

Population: $X \sim$ Bernoulli (p), where $p =$ proportion of successes in population; $p = E(X)$.

Sample data: $X_1, \ldots, X_n$. (Note: they are 0s and 1s).

Recall: Estimator for $p = \hat{p} =$ proportion of successes in the sample.

Also: Estimated var(X) = estimate of $p(1-p) =$

Result: An approximate CI for p: $\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

large sample interval:

$\bar{X} \pm z_{\alpha/2} \, \hat{SE}(\bar{X})$.

~~*Proof:*~~

Recall:

$\hat{SE}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

$\Rightarrow \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Ex: From a large population of RAM chips, a random sample of 50 is taken and a test carried out on each to see whether they perform correctly. In the test, only 20 chips are found to perform correctly. Find a 95% CI for p, the true proportion of chips that perform correctly.

n

$X \sim \text{Bernoulli}(p)$, p

Indicator of chip performance ('1' or '0').

$\hat{p} = \frac{20}{50} = 0.4$

95% CI for p: $\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

$= 0.4 \pm 1.96 \sqrt{\frac{(0.4)(0.6)}{50}}$

$= [\_\_, \_\_]$.

Choosing the sample size n

- Width of CI = $2 z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = w$
- Let w = desired CI width for $1-\alpha$ confidence.
- Margin of error = $w/2$
- Set CI width = desired width and solve for n to get

$$n \geq \left[\frac{2 z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})}}{w}\right]^2$$

$\uparrow \hat{p}$ is unknown.

saw this during "monte carlo"

- This formula involves $\hat{p}$, which is not known before the experiment.
- One alternative: take $\hat{p}$= 0.5 because $\hat{p}(1-\hat{p})$ is maximum when $\hat{p}=0.5$. This strategy will yield a conservative values of n. (The sample size will be larger than necessary.)

$$\Rightarrow \quad n \geq \left[\frac{\cancel{2} z_{\alpha/2}\cdot \frac{1}{\cancel{2}}}{w}\right]^2 = \left[\frac{z_{\alpha/2}}{w}\right]^2$$

Ex: Suppose we are planning a survey to estimate the proportion of American who approve of President Obama's job. We would like our estimate to be within 3% of the true proportion with 95% confidence. How much sample size should we take?

desired margin of error.

$1-\alpha$

$$n \geq \left[\frac{z_{\alpha/2}}{w}\right]^2 = \left[\frac{1.96}{0.06}\right]^2 \approx 1068.$$

width

So far: One-sample problem.

Now: Two-sample problem.

$$\boxed{X \sim f_{\theta_1}(x)} \qquad\qquad \boxed{Y \sim f_{\theta_2}(x)}$$

$$\downarrow \qquad\qquad\qquad \downarrow$$

$$X_1, X_2, \ldots, X_{n_1} \qquad\qquad X_1, X_2, \ldots, X_{n_2}$$

First thing: Are these samples "independent" or "paired"?

Two designs:

Design 1: (Two indep. samples)

indep. subjects

Group 1	Group 2
X_1	Y_1
X_2	Y_2
$\vdots$	$\vdots$
X_{n_1}	Y_{n_2}

The two groups have different subjects.

Design 2: Paired design

Subject	X	Y
1	X_1	Y_1
2	X_2	Y_2
⋮		
n	X_n	Y_n

"paired data"

~~two~~ the X and the Y samples are not indep.