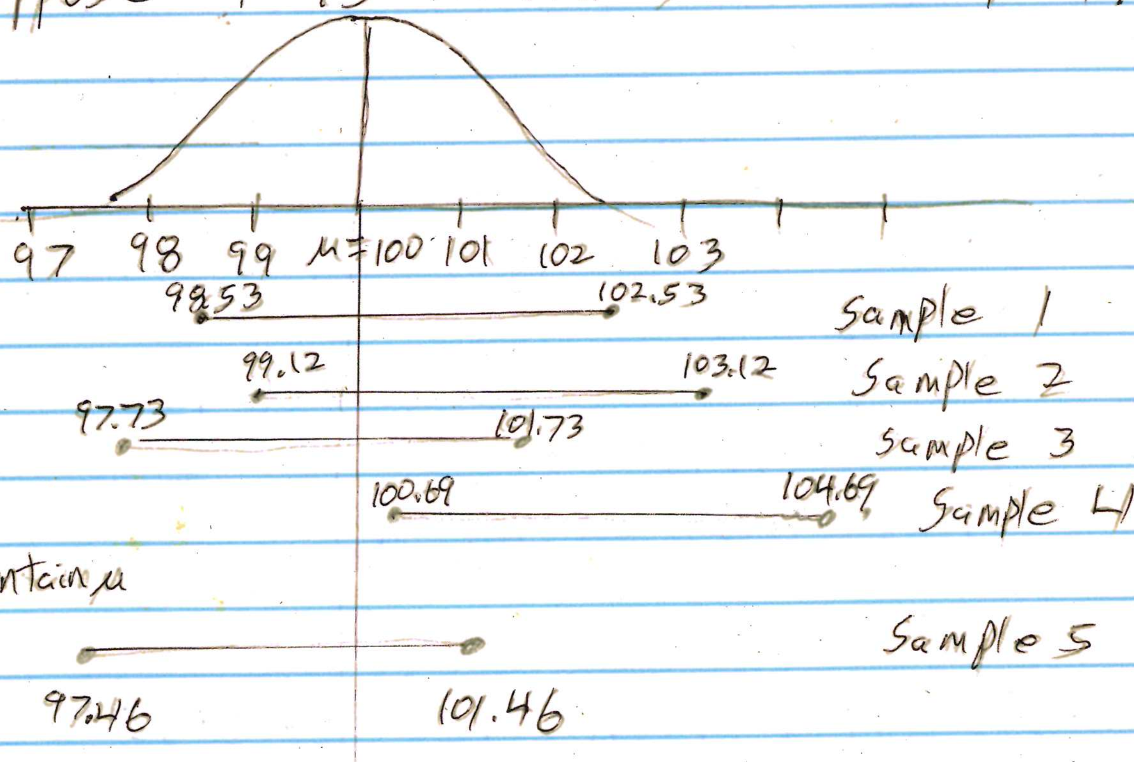
experiment, all probabilities are 0 or 1.

so $P(\mu \in CI) = \begin{cases} 0 & \text{if } \mu \notin CI \\ 1 & \text{if } \mu \in CI. \end{cases}$

ex] Flip coin, before the experiment $P(H) = P(T) = \frac{1}{2}$. If you flip coin and it is T, then after the experiment, $P(T) = 1$, $P(H) = 0$.

ex] Suppose $\bar{Y}$ is normal $\mu = 100$ $\sigma_{\bar{Y}} = 1$.




"bad sample" CI does not contain $\mu$

13] P414-5 Another interpretation of a k% CI: Suppose 100 independent samples are used to make 100 different CI's. Then about k% will contain $\theta$ and $(100-k)\%$ will not contain $\theta$. Ie. about 95 of 100 samples produce 95% CI's that contain $\theta$.

See Fig 8.8 p.414

Note: only 2 sided CIs will be used in this class.

14) p 407 The <u>pivotal method</u> for constructing a CI uses a <u>pivotal quantity</u> that

i) is a function of $Y_1, \dots, Y_n$ and $\theta$ where $\theta$ is the only unknown (so not a statistic)

ii) the probability distribution of the pivotal quantity does not depend on $\theta$.

ex) $Y_1, \dots, Y_n$ are iid $N(\mu, \sigma^2)$, $\sigma^2$ known, $\mu$ unknown. The pivotal quantity $\dfrac{\bar{Y} - \mu}{\sigma/\sqrt{n}}$ is $N(0,1)$ regardless of $\mu$.

ex) $Y_1, \dots, Y_n$ iid $N(\mu, \sigma^2)$

$$\frac{(n-1)\,S^2}{\sigma^2} \sim \chi^2_{n-1}$$

ex) $Y_1, \dots, Y_n$ iid $N(\mu, \sigma^2)$

$$\sqrt{n}\,\frac{\bar{Y} - \mu}{S} \sim t_{n-1}$$

Skip ex 8.4 and ex 8.5. 48353

§8.6 15] $Z$ is an approximate pivotal quantity if

$$Z = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}}, \quad \sigma_{\hat{\theta}} \text{ is known}$$

and $Z \approx N(0,1)$ if $n$ is large.

16] Then for large $n$

$$P\left(-z_{\frac{\alpha}{2}} \le Z \le z_{\frac{\alpha}{2}}\right) \approx 1-\alpha$$

where $P\left(Z \le z_{\alpha/2}\right) = 1 - \frac{\alpha}{2}$

or $P\left(Z > z_{\alpha/2}\right) = \alpha/2$.

So $P\left(-z_{\alpha/2} \le \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} \le z_{\alpha/2}\right) \approx 1-\alpha$

or $P\left(-z_{\alpha/2}\,\sigma_{\hat{\theta}} \le \hat{\theta} - \theta \le z_{\alpha/2}\,\sigma_{\hat{\theta}}\right) \approx 1-\alpha$

or $P\left(\hat{\theta} - z_{\alpha/2}\,\sigma_{\hat{\theta}} \le \theta \le \hat{\theta} + z_{\alpha/2}\,\sigma_{\hat{\theta}}\right) \approx 1-\alpha$.

So $\left[\hat{\theta} - z_{\alpha/2}\,\sigma_{\hat{\theta}},\ \hat{\theta} + z_{\alpha/2}\,\sigma_{\hat{\theta}}\right]$ is

a (large sample approximate)

$100(1-\alpha)\%$ CI for $\theta$.

(big assumption)

17] *p 412 If $\sigma$ is known,

$Y_1, \ldots, Y_n$ iid from a population with mean $\mu$ and standard deviation $\sigma$, and if the CLT holds, then

$$\bar{Y} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = \left[\bar{Y} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}},\ \bar{Y} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right]$$

is a $100(1-\alpha)\%$ CI for $\mu$.

18] know p413 If $n-1 > 29$ or $n > 30$ and $\sigma$ is unknown, $Y_1, \ldots, Y_n$ iid & CLT applies, use $\bar{Y} \pm z_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$ as a $100(1-\alpha)\%$ CI for $\mu$,

where $P(Z > z_{\alpha/2}) = \alpha/2$

or $P(-z_{\alpha/2} \le Z \le z_{\frac{\alpha}{2}}) = 1-\alpha$

and $Z$ is $N(0,1)$.

19]

| | $z_{.05}$ | $z_{.025}$ | $z_{.005}$ |
|---|---|---|---|
| $z_{\alpha/2}$ | 1.645 | 1.96 | 2.576 |
| $(1-\alpha)100\%$ | 90% | 95% | 99% |

ex] 90% CI, $1-\alpha = .9$, so $\alpha = .1$

and $\frac{\alpha}{2} = .05$. $P(Z \le z_{.05}) = 1 - \frac{\alpha}{2} = .95$

$P(Z > z_{.05}) = .05$,

| | 04 | 05 |
|---|---|---|
| 1.6 | .0495 | .0505 |

$$z_{.05} = \frac{1.64 + 1.65}{2} = 1.645$$

20] Don't use CI $\bar{y} \pm z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$

if $n \le 30$ or if the CLT does not apply.

ex] MBA salaries in 1994 if person got MBA in 70's and are sole source of household income.
$n = 91$ $\bar{x} = 124510$, $s = 180000$, assume CLT holds. Find a 90% CI for $\mu$.

soln $z_{\frac{\alpha}{2}} = 1.645$

$$\bar{x} \pm z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} = 124510 \pm 1.645 \frac{180000}{\sqrt{91}}$$

$$= 124510 \pm 3103.97$$

$$= [121406.03, 127613.97].$$

21] $\bar{Y} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$

CI length increases if $\underbrace{1-\alpha \text{ increases}}_{\text{so } z_{\frac{\alpha}{2}} \text{ increases}}$
if $\sigma$ increases
if $n$ decreases

22) Know the sample proportion

$\hat{p} = \dfrac{\text{count of "successes"}}{n}$. That is

let sample have size $n$
and $W_1, \ldots, W_n$ be iid
with $W_i = \begin{cases} 1 & \text{"success"} \\ 0 & \text{"failure"} \end{cases}$ what you count / what you don't count.

Then $\hat{p} = \frac{1}{n} \underbrace{\sum W_i}_{\text{binomial}(n,p)} = \bar{W} = \frac{\#\text{ successes}}{n}$

23] $E\hat{p} = p$ = population proportion of successes

$V(\hat{p}) = \frac{p(1-p)}{n}$, $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$.

If $n > 9\frac{p}{1-p}$ and $n > 9\frac{1-p}{p}$,

then $\dfrac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \approx N(0,1)$

and $\dfrac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \approx N(0,1)$.

p 401, 415

24] Know A $100(1-\alpha)\%$ CI for $p$ is

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

where $P\left(-z_{\frac{\alpha}{2}} \le Z \le z_{\frac{\alpha}{2}}\right) = 1-\alpha$

or $P(Z > z_{\alpha/2}) = \frac{\alpha}{2}$ and $Z$ is $N(0,1)$.

ex] Random sample of $n = 1711$ from records of people who died in bicycle crashes between 1987 and 1991. 386 of the 1711 had blood alcohol levels above 0.10% (were drunk). Find a 95% CI for $p$.

Soln] $\hat{p} = \frac{386}{1711} = 0.2256$

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{.2256\,(1-.2256)}{1711}} = .0101$$

$z_{\alpha/2} = 1.96$, 95% CI is $\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

$= .2256 \pm 1.96\,(.0101) = .2256 \pm .0198$

$= [.2058, .2454]$.

Estimate that between 20% and 25% of people killed in bycycle crashes were drunk.

25] Let $X_1, \dots, X_{n_1}$ be iid from a pop with mean $\mu_1$ and variance $\sigma_1^2$. Let $Y_1, \dots, Y_{n_2}$ be iid from a pop with mean $\mu_2$ and variance $\sigma_2^2$. Assume the 2 samples are independent and that the CLT holds for $\bar{X}$ and $\bar{Y}$. Then $Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \approx N(0,1)$.

If also $n_1 > 30$ and $n_2 > 30$, then

$$\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \approx N(0,1).$$

26] Know under the conditions of 25]), an approximate $100(1-\alpha)\%$ CI for $\mu_1-\mu_2$ is

$$(\bar{X}-\bar{Y}) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

where $P(Z > z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$ and $Z$ is $N(0,1)$.

ex] Gas mileage for two types of engines

| | | mean | var or $s^2$ | $n$ | SD or $s$ |
|---|---|---|---|---|---|
| A | X | 42 | 64 | 75 | 8 |
| B | Y | 36 | 36 | 50 | 6 |

95% CI $\quad z_{\alpha/2} = 1.96$

$$\bar{X}-\bar{Y} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} =$$

$$(42-36) \pm 1.96 \sqrt{\frac{64}{75}+\frac{36}{50}}$$

$$= 6 \pm 1.96\sqrt{1.5733} = 6 \pm 1.96(1.25426)$$

$$= 6 \pm 2.458 = [3.542, 8.458]$$

Engine A gets between 3.5 and 8.5 mpg better than engine B on average.

27] Two independent samples

| $p$ | $\hat{p}$ | $n$ |
|---|---|---|
| $p_1$ | $\hat{p}_1$ | $n_1$ |
| $p_2$ | $\hat{p}_2$ | $n_2$ |

$$n_1 > 9 \frac{\hat{p}_1}{1-\hat{p}_1}, \quad n_1 > 9 \frac{1-\hat{p}_1}{\hat{p}_1}$$

$$n_2 > 9 \frac{\hat{p}_2}{1-\hat{p}_2} \quad n_2 > 9 \frac{1-\hat{p}_2}{1-\hat{p}_2}.$$

Then $$\frac{\hat{p}_1-\hat{p}_2 - (p_1-p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \approx N(0,1)$$

and $$\frac{\hat{p}_1-\hat{p}_2 - (p_1-p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \approx N(0,1)$$

p4 13

28] Know Under the conditions of 27], a $100(1-\alpha)\%$ CI for $p_1 - p_2$ is

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

where $P(Z > z_{\alpha/2}) = \alpha/2$ and $Z$ is $N(0,1)$.

Text uses $\hat{q}_1 = 1-\hat{p}_1$, $\hat{q}_2 = (1-\hat{p}_2)$.

ex] 

| | | $\hat{p}$ |
|---|---|---|
| 100 men | 31 | .31 |
| 100 women | 67 | .67 |

(≃2000?)

$\underbrace{\phantom{67}}$ # who say women are safer drivers than men

www.usatoday.com  Find a 90% CI for $p_1 - p_2$.

$9\,\frac{.31}{.69} < 9\,\frac{.69}{.31} = 20.03 < 100$

$9\,\frac{.33}{.67} < 9\,\frac{.67}{.33} = 18.27 < 100$

$z_{\alpha/2} = 1.645$

$$\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = \sqrt{\frac{(.31)(.69)}{100} + \frac{(.67)(.33)}{100}}$$

$$= \sqrt{.002139 + .002211} = \sqrt{.00435} = 0.06595.$$

So the 90% CI $\hat{p}_1 - \hat{p}_2 \pm z^* (.06595)$

$$= (.31 - .67) \pm 1.645 (.06595) =$$

$$-0.36 \pm 0.108 = [-0.468, -.252]$$

The proportion of men who say that women are safer is 25% to 47% smaller than the proportion of women.

§ 8.7 29] The <u>error of estimation</u> or <u>margin of error</u> $= \frac{1}{2}$ CI width

$= B$ is a measure of the accuracy of the interval estimate of $\theta$.

483 57

30] Often want to select the sample size n <u>before</u> setting up the experiment so that the CI width $\le 2B$.

31] idea: CI is $\hat{\theta} \pm z_{\frac{\alpha}{2}} \sigma_{\hat{\theta}}$

so $\frac{1}{2}$ CI width $= B = z_{\alpha/2}\, \sigma_{\hat{\theta}} = h(n)$.
Solve $n = h^{-1}(B)$ where $n$ is the sample size needed for $\hat{\theta}$ to be within $B$ of $\theta$ with $100(1-\alpha)\%$ confidence

32] * Convention: round $n$ <u>UP</u> to the nearest integer

33] * $100(1-\alpha)\%$ CI for $\mu$ if

p421 $\sigma$ is known, data are iid $Y_1, \ldots, Y_n$ and CLT holds, is $\bar{Y} \pm z_{\alpha/2}\, \sigma/\sqrt{n}$.
The sample size $n$ for $\bar{Y}$ to be within $B$ of $\mu$ with $100(1-\alpha)\%$ confidence is $n = \left(\frac{z_{\frac{\alpha}{2}}\sigma}{B}\right)^2$, round up

(set $B = z_{\alpha/2}\, \sigma/\sqrt{n}$).

ex] Want to estimate average weight $\mu$ to <u>within</u> 0.1 oz with 95% confidence and $\sigma = 1$ oz. Find $n$.

soln] Key word "within" means $B = 0.1$.

so $n = \left(\frac{z_{\alpha/2}\,\sigma}{B}\right)^2 = \left(\frac{(1.96)\,1}{0.1}\right)^2$

$= (19.6)^2 = 384.16$ so
$n = 385$ (round up so level

is at least 95%).

57.5

Note: Often the cost of the experiment is proportional to the sample size $n$. want $n$ as small as possible subject to being sufficiently accurate ( width $\leq 2B$).

skip if behind

34] * $Y_1, \ldots, Y_n$ iid, $\mu$ and $\sigma$ unknown, $n$ large enough for CLT to hold, $s^*$ a guess for $\sigma$, has CI $\bar{Y} \pm z_{\frac{\alpha}{2}} s/\sqrt{n}$.

(S is unknown)

Sample size $n$ needed for $\bar{Y}$ to be within $B$ of $\mu$ with $100(1-\alpha)\%$ confidence is $n = \left(\frac{z_{\frac{\alpha}{2}} s^*}{B}\right)^2$, round up

(Since $B = z_{\alpha/2}\, s/\sqrt{n}$).

Note: actually $n = \max\left\{30, \left(\frac{z_{\alpha/2} s^*}{B}\right)^2\right\}$ if $\sigma$ is unknown.

skip

Note: if $Y_i \in [a,b]$, $b-a =$ range, $s^* =$ range/4 is often used

ex] want to estimate $\mu$ to within 0.1 with 95% confidence and $s^* = 1$. Data will be a random sample from some pop. Find $n$.

$s^*$ instead of $\sigma$

soln] $n = \left(\frac{z_{\alpha/2}\, s^*}{B}\right)^2 = \left(\frac{1.96(1)}{0.1}\right)^2$

$= 384.16$ so $n = 385$.

Know

35] * Random sample $Y_1, \dots, Y_n$ 0's and 1's

$100(1-\alpha)\%$ CI for $p$ is

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

but $p$ is unknown.

Suppose that $p^*$ is a good guess for $p$ based on previous experiments. Then the sample size $n$ for $\hat{p}$ to be within $B$ of $p$ with $100(1-\alpha)\%$ confidence is

$$n = \left(\frac{z_{\frac{\alpha}{2}}}{B}\right)^2 p^*(1-p^*), \quad \text{round up.}$$

Proof] Set $B = z_{\frac{\alpha}{2}} \sqrt{\frac{p^*(1-p^*)}{n}}$.

Then $\left(\frac{B}{z_{\alpha/2}}\right)^2 = \frac{p^*(1-p^*)}{n}$, or

$$n = \left(\frac{z_{\alpha/2}}{B}\right)^2 p^*(1-p^*).$$

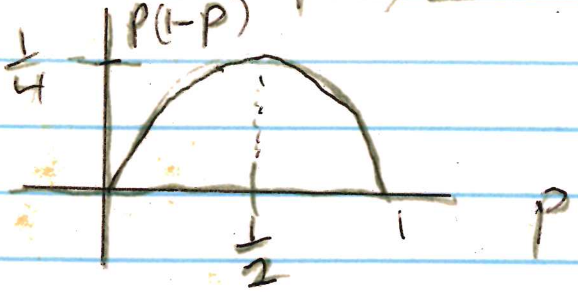
Know

36] If no good guess for $p$ is available, use $p^* = 1/2$.

So $n = \left(\frac{z_{\alpha/2}}{B}\right)^2 \frac{1}{2}\left(1-\frac{1}{2}\right) = \left(\frac{z_{\alpha/2}}{2B}\right)^2.$

Note $p = 1/2$ maximizes $p(1-p)$ if $0 \le p \le 1$

58.5




Note: If $p$ is far from 0.5, then using $p^* = \frac{1}{2}$ will result in a much larger sample size $n$ than actually needed.

E4 problem

ex] a) Find $n$ to estimate $p$ to within 0.01 with 90% confidence if $p^* = .2$.

soln] $n = \left(\frac{z_{\alpha/2}}{B}\right)^2 p^*(1-p^*)$

$= \left(\frac{1.645}{.01}\right)^2 .2(1-.2) = 4329.64$

$\boxed{n = 4330}$

b) Same but no good guess for $p$ is available.

soln] use $p^* = 0.5$

$n = \left(\frac{1.645}{.01}\right)^2 \frac{1}{2}\left(1-\frac{1}{2}\right) = (164.5)^2 \frac{1}{4} = 6765.06$

use $n = \boxed{6766}$

Note: $n$ is much larger so the expt is much more expensive.

37] With 2 samples, the procedures usually work best with $n_1 = n_2 = n$.

Need guesses $s_1^{*2}$, $s_2^{*2}$

or $p_1^*$, $p_2^*$

then solve for $n$.

See ex 8.10 on p 423 – 424.

§8.8] 38] * Let $Y_1, \ldots, Y_n$ be iid from a normal population with $\mu$ and $\sigma^2$ unknown and $n-1 \le 29$ or $n \le 30$. Then a (small sample) $100(1-\alpha)\%$ CI for $\mu$ is the $t$-interval

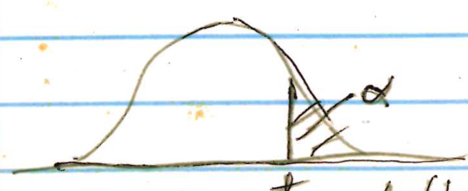
$$\bar{Y} \pm t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$$

where $P(T > t_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$

and $T \sim t_{n-1}$, a $t$ distribution with $df = n-1$.

Note] * If $n > 30$ then the CI is $\bar{Y} \pm z_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$.

39] Find $t_{\alpha/2}$ with the $t$ table in the front of the book.




can only find $t_{\alpha/2}$ for 80%, 90%, 95%, 98% and 99% CI's.

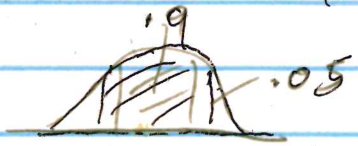
left tail $t$ tables good for 90%, 95%, 99%

ex] i) $n=8$ want 95% CI $df=n-1=7$


| $t_{.025}$ | df |
|---|---|
| 2.365 | 7 |
| 1.96 | inf |

$t_{\alpha/2} = 2.365$

ii) $n=8$ want 90% CI




| $t_{.05}$ | df |
|---|---|
| 1.895 | 7 |
| 1.645 | inf |

iii) $n=8$ want 99% CI


| $t_{.005}$ | df |
|---|---|
| 3.499 | 7 |
| 2.576 | inf |

40] * 2 common mistakes
i) plug in $z_{\alpha/2}$ instead of $t_{\alpha/2}$.
ii) forget $df = n-1$, not $n$.

41] When can the $t$ interval be used?
i) data needs to be a random sample (eg data measurements from an experiment done $n$ times)
ii) If $2 \le n \le 30$ use if CLT holds: if data is normal or approximately normal.
iii) If $n \le 30$ and the data come from a highly skewed pop, do not use the $t$ interval.