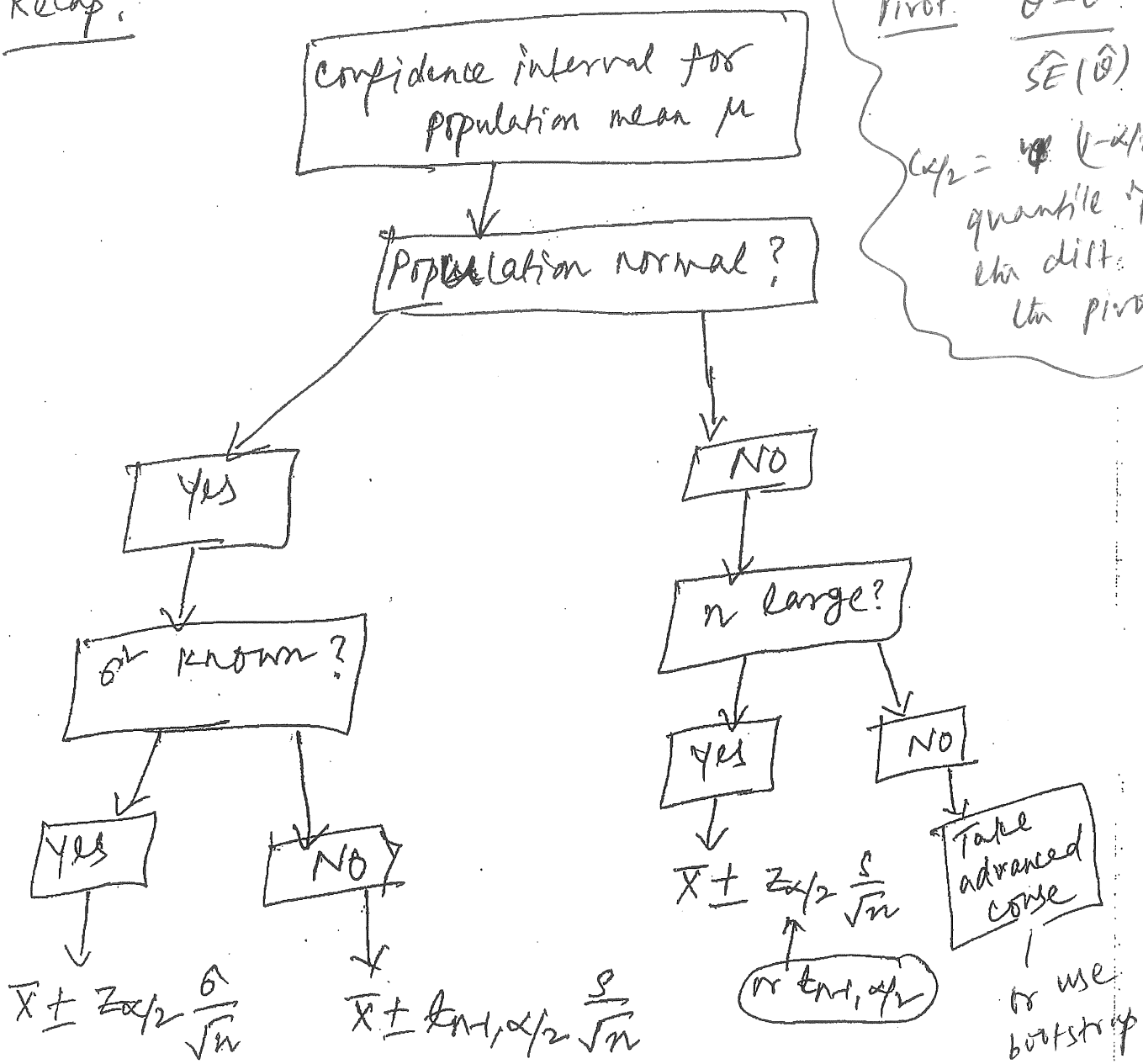
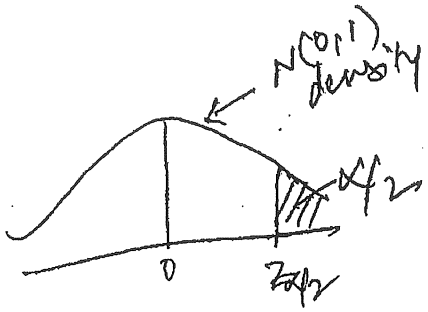
Recall: Generally: A $100(1-\alpha)\%$ CI for θ: $\hat{\theta} \pm c_{\alpha/2}\, \hat{SE}(\hat{\theta})$,

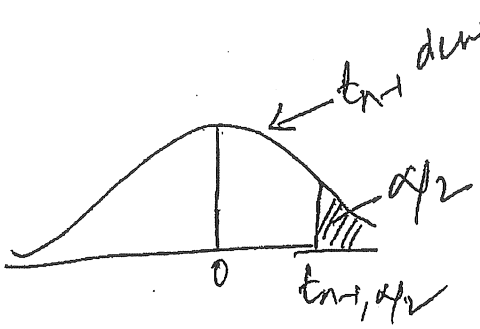
Pivot: $\dfrac{\hat{\theta}-\theta}{\hat{SE}(\hat{\theta})}$

($c_{\alpha/2}$ = $(1-\alpha/2)$-th quantile of the dist. of the pivot)

Recap:







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:

x_1 x_2 x_3 — $n = 12$

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 = (typical) time b/w two keystrokes for the person trying to log in.

known: $X \sim N[\mu, \sigma^2]$

Goal: 95% CI for μ!

Note: σ^2 is unknown and needs to be estimated using data.

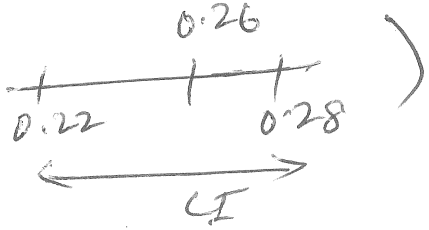
t-interval: $\boxed{\bar{x} \pm t_{11,\,0.025}\,\frac{s}{\sqrt{n}}}$

$= [0.22, 0.28]$ → plausible values of μ.

↑ using R.

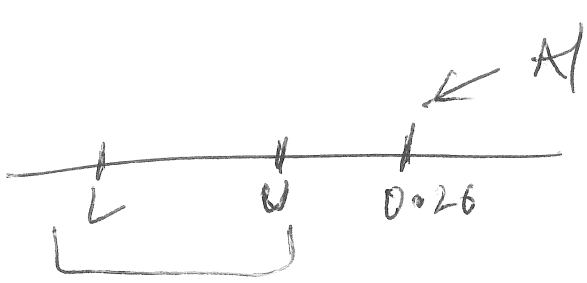
```
x <- c(0.46, 0.38, 0.31, 0.24, 0.20, 0.31, 0.34,
0.42, 0.09, 0.18, 0.46, 0.21)
#> mean(x)
#[1] 0.3
#> sd(x)
#[1] 0.1183216
#>qt(0.975, 11)
#[1] 2.200985
```

Suppose: Mean for A/c owner is : 0.26



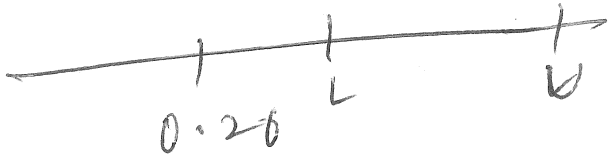
Since 0.26 is a plausible value (because for the person who is trying to login, access should be granted.

Suppose:



person who is trying to log in is typing faster than the a/c owner ⇒ deny access

Similarly:



⇒ deny access.

Large sample CI for mean μ

Recall: When n is large, an approximate $100(1-\alpha)\%$ CI for mean μ of any population is

Recall: $\bar{X} \overset{\cdot}{\sim} N\left[\mu, \frac{\sigma^2}{n}\right] \Rightarrow Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \overset{\cdot}{\sim} N[0,1]$

$\Rightarrow \bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ is an approx. CI $\left[\bar{X} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}\right]$

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 and 14 ms, respectively. Find a 95% CI for the true mean execution time μ.

X = (typical) execution time (ms)

$\bar{X}$ s

KNOW: $n = 35$,

95% CI for μ: use large-sample CI: $\bar{X} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$

$= [225.4, 234.6]$ ← using R.

↑ plausible values for μ.

100(1−α)% CI for θ: $[L, U]$ is such that

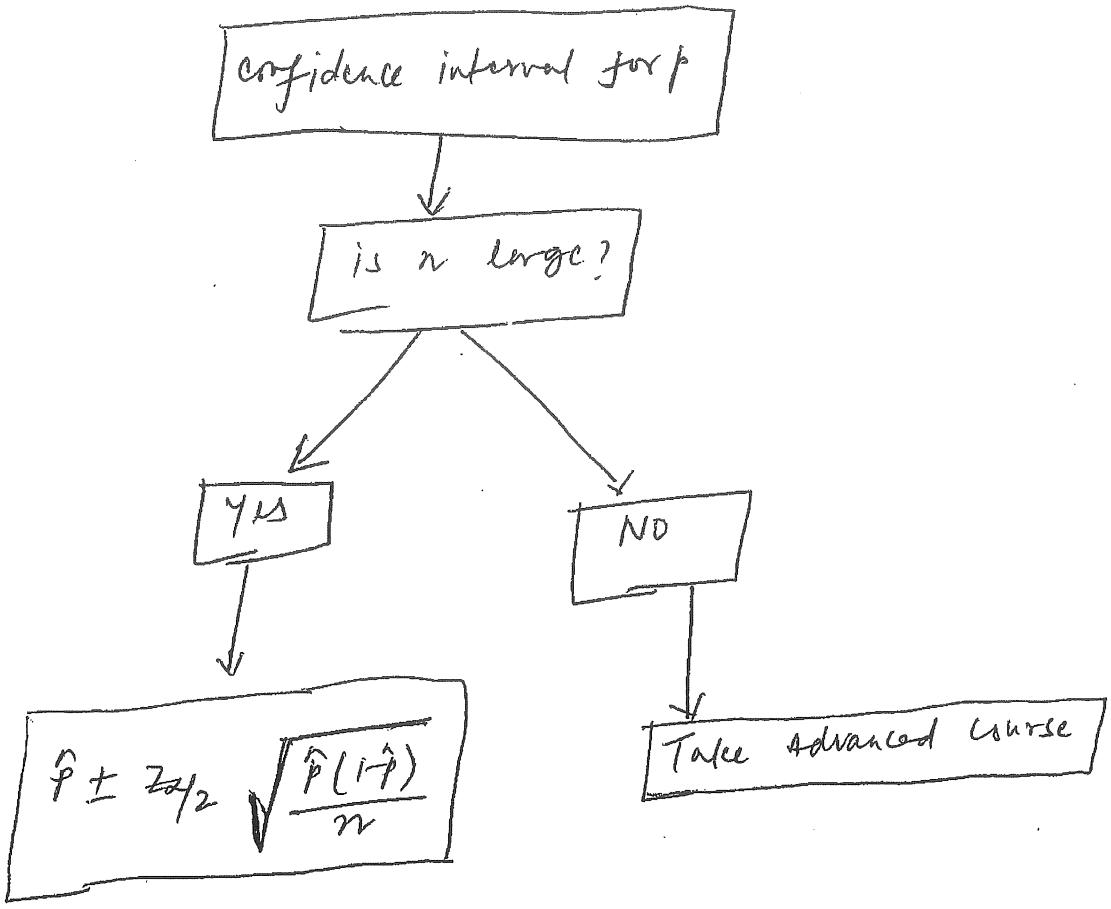
$$P[L \le \theta \le U] = 1-\alpha \quad (\text{for all } \theta)$$

— "Exact"

Approx. CI:

$$P[L \le \theta \le U] \approx 1-\alpha \text{ when } n \text{ is large}$$

[As $n \to \infty$, $P[L \le \theta \le U] \to 1-\alpha$]


confidence interval for p
is n large?
YES
NO
$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
Take advanced course

Large sample CI for success proportion p

Population: $X \sim$ Bernoulli (p), where p = proportion of successes in population; $p = E(X) = P[X=1]$.

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

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

Also: Estimated $\text{var}(X)$ = estimate of $p(1-p) =$

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

only if n is large.

Proof:

Formula: $\bar{X} \pm z_{\alpha/2} \frac{(\sigma \text{ or } s)}{\sqrt{n}}$

$\bar{X} = \hat{p}$

$\text{var}[\hat{p}] = \frac{p(1-p)}{n}$

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

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.

$X = \begin{cases} 1, & \text{if a chip performs correctly} \\ 0, & \text{o/w.} \end{cases}$

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

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

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

$= [0.26, 0.54]$ ← using R

plausible values for p.

Choosing the sample size n

- Width of CI = $U - L = 2 z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- 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

Solve for n:

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

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

- 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.)

Replacing $\hat{p}$ by $\frac{1}{2}$ gives:

$$n \approx \left[\frac{2 z_{\alpha/2}}{w} \right]^2$$

more than what is really needed.

Ex: Suppose we are planning a survey to estimate the proportion of American who approve of President Trump'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?

$X = \begin{cases} 1, & \text{if person approves the job} \\ 0, & \text{otherwise} \end{cases}$

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

p = approval rating.

$1-\alpha = 0.95$

Want: $|\hat{p} - p| \le 0.03$ ← desired margin of error

$\Rightarrow w = 2(0.03) = 0.06$

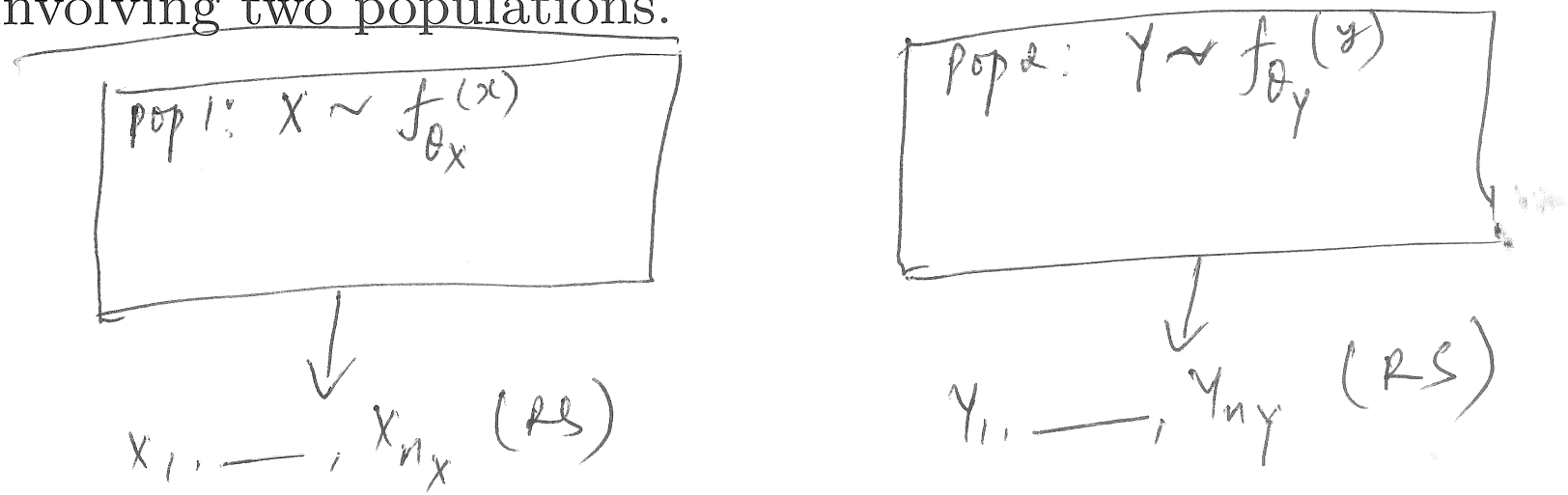
$$n \approx \left\lceil \frac{1.96}{0.06} \right\rceil^2 \approx 1068.$$

So far: One-sample problems

Two-sample problems

So far: One-sample problems — inference on parameter(s) of a single population.

Now: Two-sample problems — inference on parameters involving two populations.



Often: $\mu_X - \mu_Y = E[X] - E[Y]$

Independent or paired samples?

Design 1: (Two independent samples)

group 1: $X_1, \ldots, X_{n_X}$ group 2: $Y_1, Y_2, \ldots, Y_{n_Y}$

independent.

Design 2: (Paired samples)

ID	(X, Y)
1	(X_1, Y_1)
2	(X_2, Y_2)
⋮	
⋮	
n	(X_n, Y_n)

Generally: Assume different subjects are indep.

Given a choice, if feasible, a paired design is better than the two-sample (indep) design. This is because in case of the former, any diff. in X and Y is almost due to the effect of the intervention in study. In case of the latter, the effect of the intervention gets confounded with the difference in the groups.