STAT 515 Lec 12 slides

Confidence interval for the mean when variance unknown

Karl Gregory

University of South Carolina

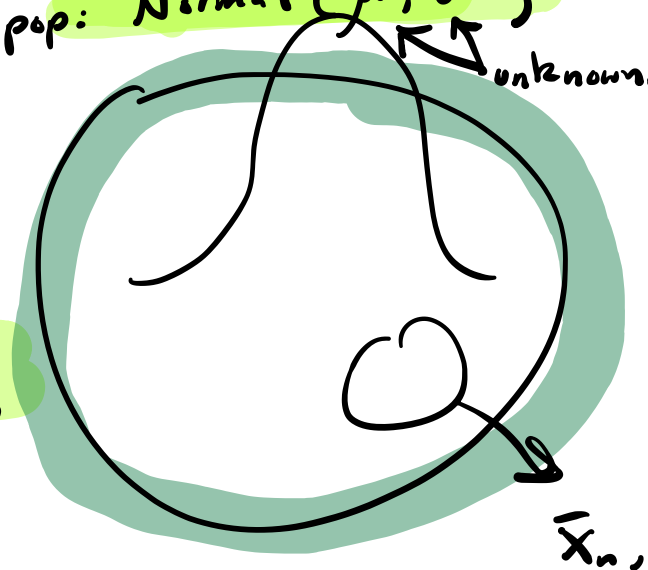
These slides are an instructional aid; their sole purpose is to display, during the lecture, definitions, plots, results, etc. which take too much time to write by hand on the blackboard. They are not intended to explain or expound on any material.

Recall: If $X_1, \ldots, X_n \overset{\text{ind}}{\sim} \text{Normal}(\mu, \sigma^2)$, then

unknown

$$\bar{X}_n \pm \underbrace{z_{\alpha/2} \frac{\sigma}{\sqrt{n}}}_{\text{margin of error}}$$

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



But what if we don't know σ?

$$Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

sample standard deviation

Using $\bar{X}_n \pm z_{\alpha/2} S_n/\sqrt{n}$ is okay if n is large, but not if n is small...

Replace σ with S_n?

Can we use $\bar{X}_n \pm z_{\alpha/2} \dfrac{S_n}{\sqrt{n}}$

$$Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

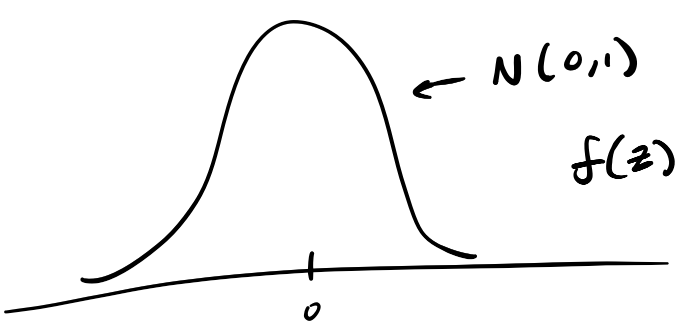
"standardize" $\bar{X}_n$

replace with S_n

Sampling distribution of "studentized" mean

If $X_1, \dots, X_n \overset{\text{ind}}{\sim} \text{Normal}(\mu, \sigma^2)$, then $\boxed{\dfrac{\bar{X}_n - \mu}{S_n/\sqrt{n}}} \sim t_{n-1}$.

In the above, t_{n-1} represents the t-distribution with $n-1$ degrees of freedom...



$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2}$$

The t-distributions

The probability distribution with pdf given by

$$f(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\nu\pi\Gamma(\frac{\nu}{2})}\left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \quad \text{where } \Gamma(z) = \int_0^\infty u^{z-1}e^{-u}du,$$

for $\nu > 0$ is called the *t-distribution* with *degrees of freedom ν*.

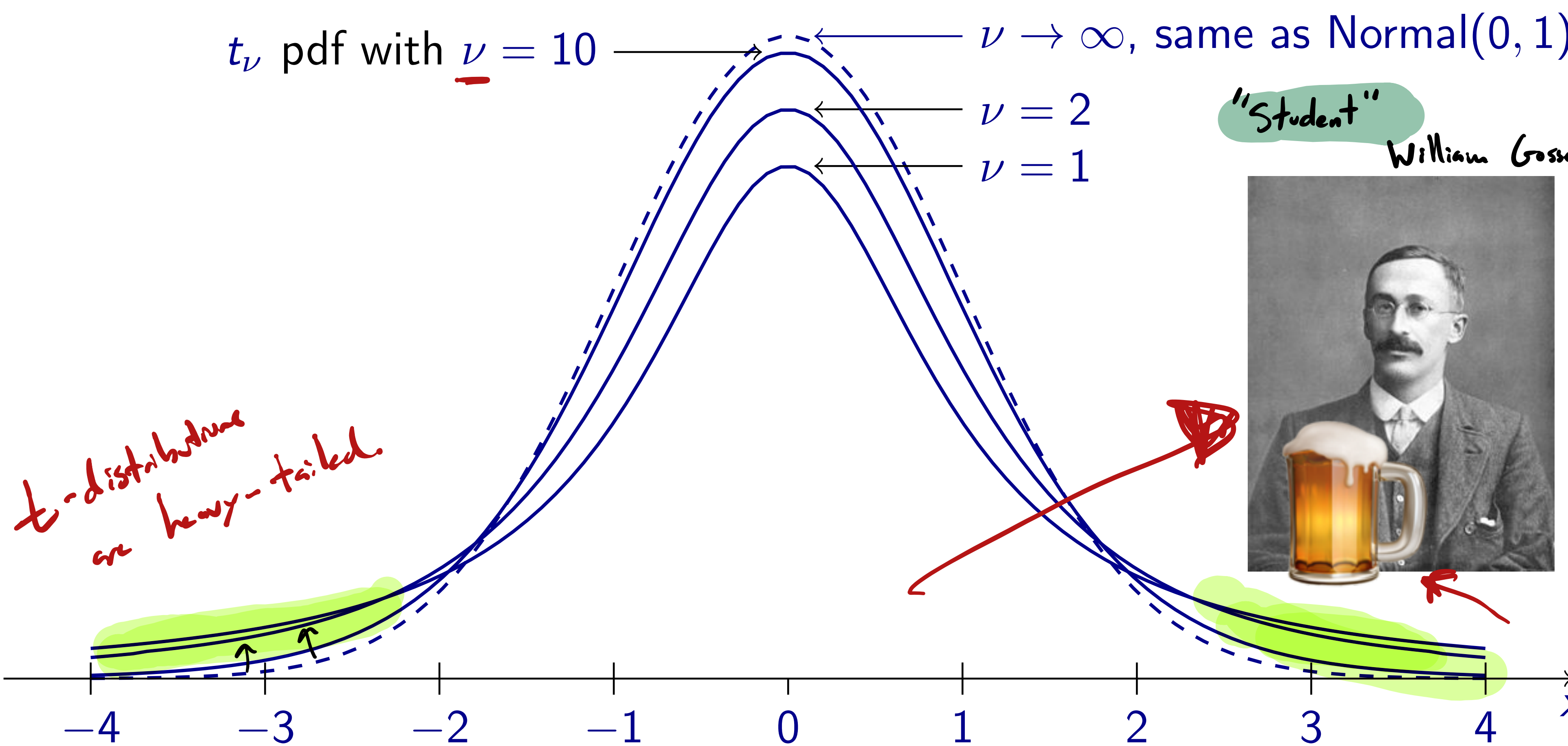
For a random variable T with this distribution we write $T \sim t_\nu$.

How to build a t_ν random variable

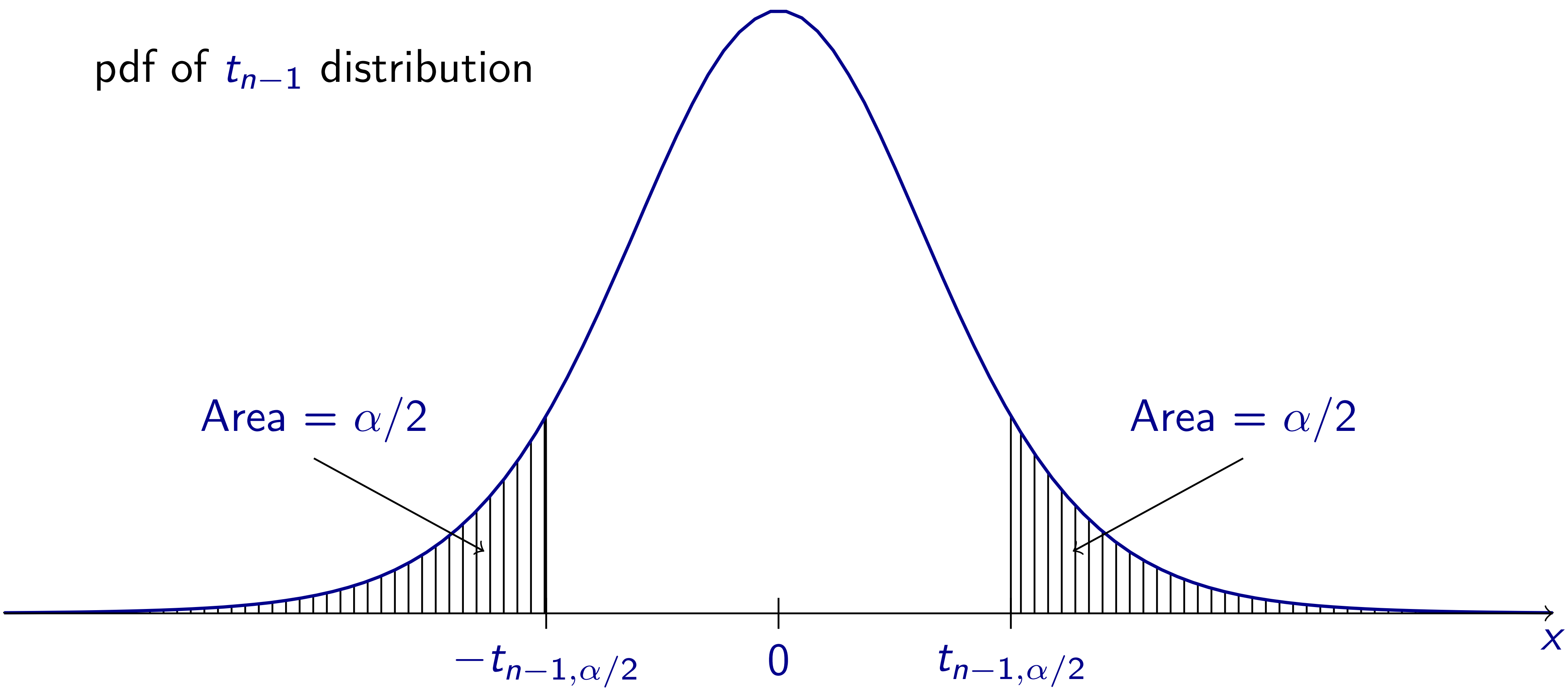
If $Z \sim \text{Normal}(0,1)$ and $W \sim \chi^2_\nu$ are independent rvs, then

$$T = \frac{Z}{\sqrt{W/\nu}} \sim t_\nu.$$

$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \sim t_{n-1}$$



"Student"
William Gossett



Can use function `qt()` or a t-table to look up the values, e.g.

$$t_{19,0.025} = \texttt{qt(.975,19)} = 2.093024$$
$$t_{19,0.005} = \texttt{qt(.995,19)} = 2.860935.$$

If σ is known: $\bar{X}_n \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

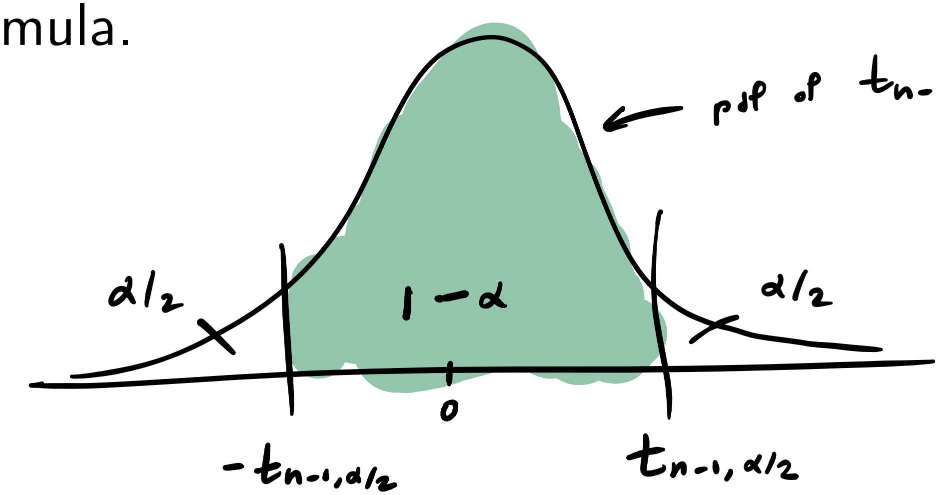
Confidence interval for mean of a Normal population with σ unknown

Let $X_1, \dots, X_n \overset{\text{ind}}{\sim} \text{Normal}(\mu, \sigma^2)$. Then a $(1-\alpha) \times 100\%$ CI for μ is

$$\bar{X}_n \pm \underbrace{t_{n-1,\alpha/2}}_{\text{a little bit bigger than } z_{\alpha/2}.} \frac{S_n}{\sqrt{n}}.$$

Discuss: How to obtain the above formula.

$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \sim t_{n-1}$$



$$P\left(-t_{n-1,\alpha/2} < \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} < t_{n-1,\alpha/2}\right) = 1-\alpha$$

Rearrange

$\iff$

$$P\left(-t_{n-1,\alpha/2}\frac{S_n}{\sqrt{n}} < \bar{X}_n - \mu < t_{n-1,\alpha/2}\frac{S_n}{\sqrt{n}}\right) = 1-\alpha$$

$$\iff P\left(-\bar{X}_n - t_{n-1,\alpha/2}\frac{S_n}{\sqrt{n}} < -\mu < -\bar{X}_n + t_{n-1,\alpha/2}\frac{S_n}{\sqrt{n}}\right) = 1-\alpha$$

$\iff$

$$P\left(\underbrace{\bar{X}_n + t_{n-1,\alpha/2}\frac{S_n}{\sqrt{n}}}_{\text{upper}} > \mu > \underbrace{\bar{X}_n - t_{n-1,\alpha/2}\frac{S_n}{\sqrt{n}}}_{\text{lower}}\right) = 1-\alpha$$

So a $(1-\alpha)\cdot 100\%$ C.I. for μ is

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

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

Exercise: These are the commute times (sec) to class of a sample of students.

1832	1440	1620	1362	577	934	928	998	1062	900
1380	913	654	878	172	773	1171	1574	900	900

1. Make a Q-Q plot to check Normality of the population.
2. Construct a 95% confidence interval for the mean commute time of all students.
3. Construct a 99% confidence interval for the mean commute time of all students.
4. What if the intervals $\bar{X}_n \pm z_{\alpha/2} \cdot S_n/\sqrt{n}$ are used? How are they different?

$n = 20$

$t_{20-1,\frac{0.05}{2}} = t_{19,0.025}$

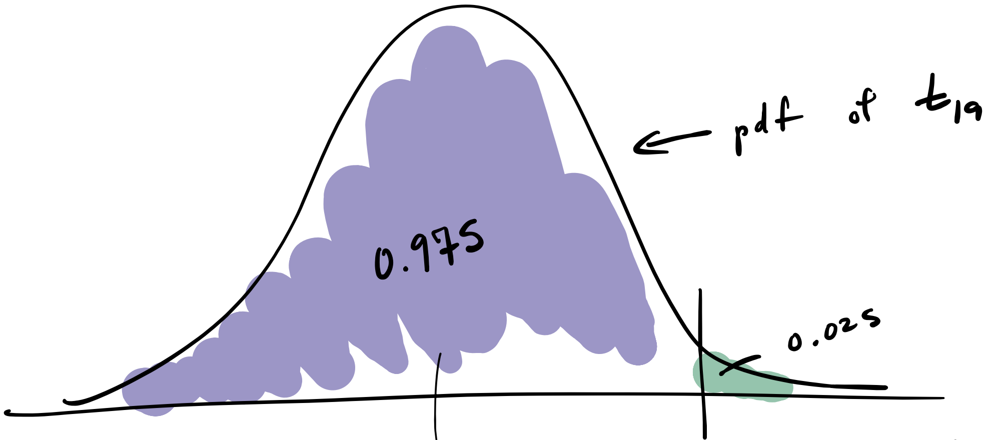
(2)

$\bar{X}_n = 1048.4$

$S_n = 394.6439$

$\alpha = 0.05$

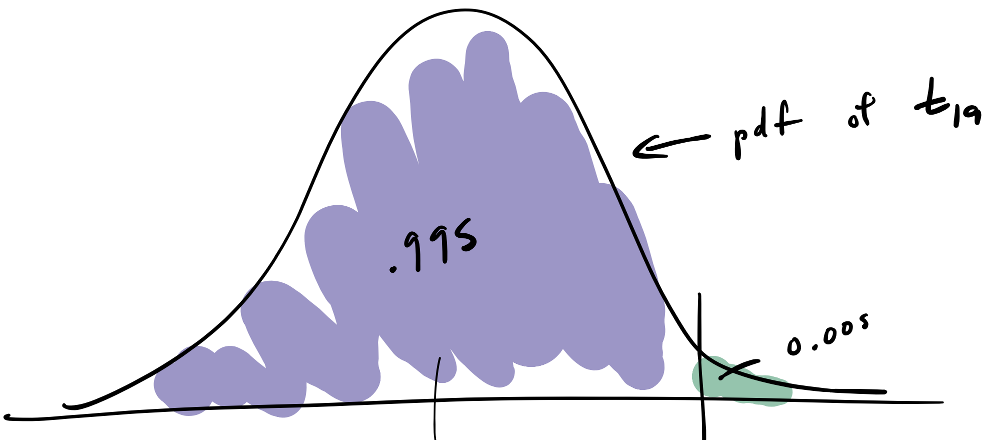
$$1048.4 \pm \underbrace{t_{19,\,0.025}}_{2.093} \frac{394.6489}{\sqrt{20}} = (863.7,\ 1233.097)$$



$$t_{19,\,0.025} = qt(0.975,\ 19) = 2.093$$

(3) $\alpha = 0.01$ $\quad \alpha/2 = 0.01/2 = 0.005$

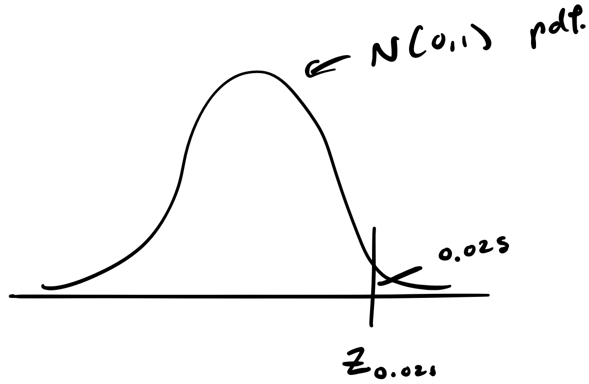
$$1048.4 \pm \underbrace{t_{19,\,0.005}}_{2.861} \frac{394.6439}{\sqrt{20}} = (795.93,\ 1300.87)$$



$$t_{19,\,0.005} = qt(.995,\ 19) = 2.861$$

(4) What if we use

$$\bar{X}_n \pm z_{\alpha/2} \frac{S_n}{\sqrt{n}}$$



$\alpha = 0.05 \Rightarrow z_{\frac{0.05}{2}} = z_{0.025} = 1.96$

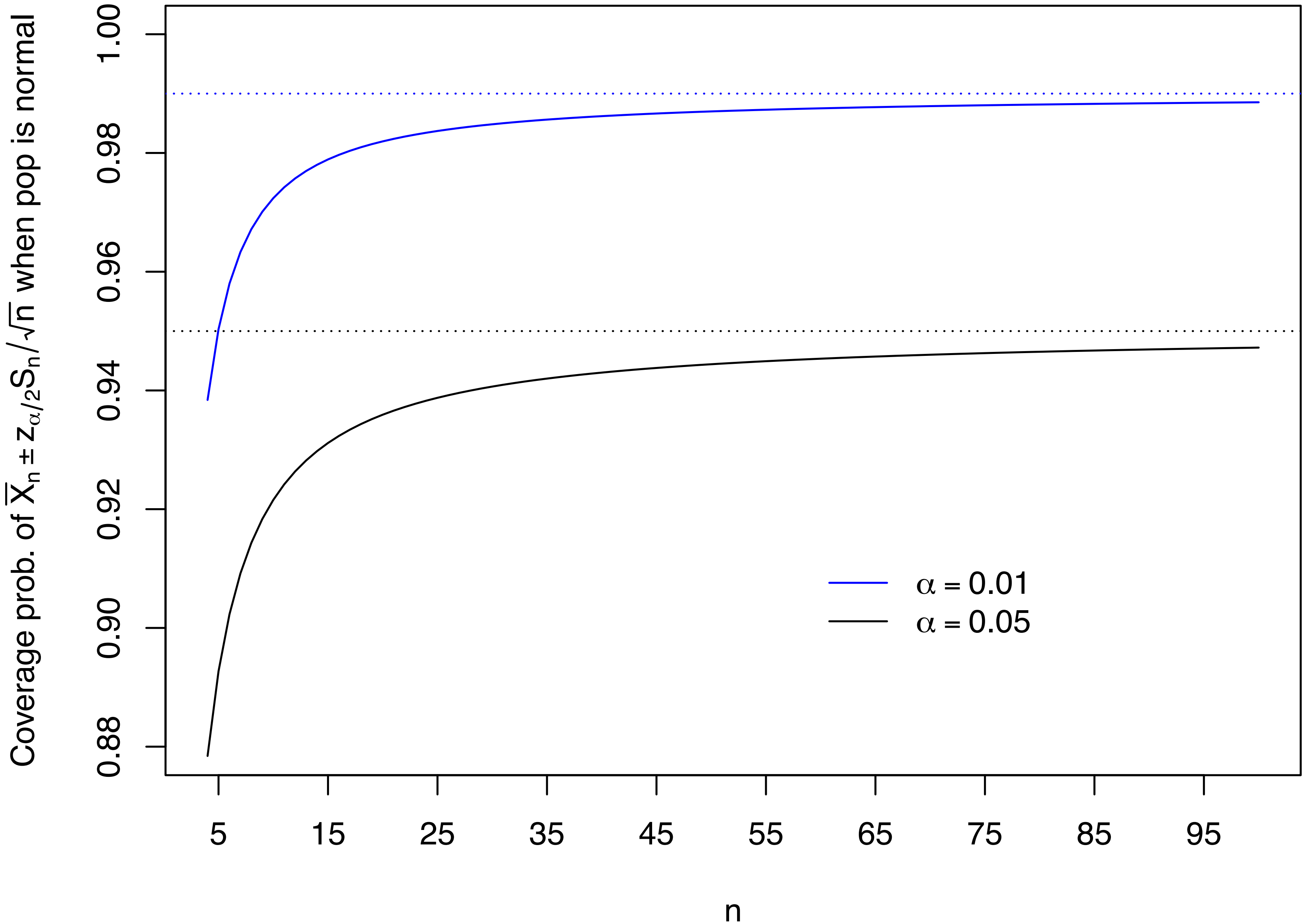
95% Naive interval:

$$1048.4 \pm \underbrace{1.96}_{\text{should not use}} \frac{394.6487}{\sqrt{20}}$$

95% "Student's" interval:

$$1048.4 \pm \underbrace{2.093}_{\text{right number}} \frac{394.6487}{\sqrt{20}}$$

```
X <- c(1832,1440,1620,1362,577,934,928,998,1062,900,
       1380,913,654,878,172,773,1171,1574,900,900)
qqnorm(X)
Xbar <- mean(X)
Sn <- sd(X)
n <- length(X)
alpha <- 0.05
tval <- qt(1-alpha/2,n-1)
lo <- Xbar - tval * Sn / sqrt(n)
up <- Xbar + tval * Sn / sqrt(n)
```


Coverage prob. of $\overline{X}_n \pm z_{\alpha/2} S_n/\sqrt{n}$ when pop is normal
1.00
0.98
0.96
0.94
0.92
0.90
0.88
$\alpha = 0.01$
$\alpha = 0.05$
5
15
25
35
45
55
65
75
85
95
n

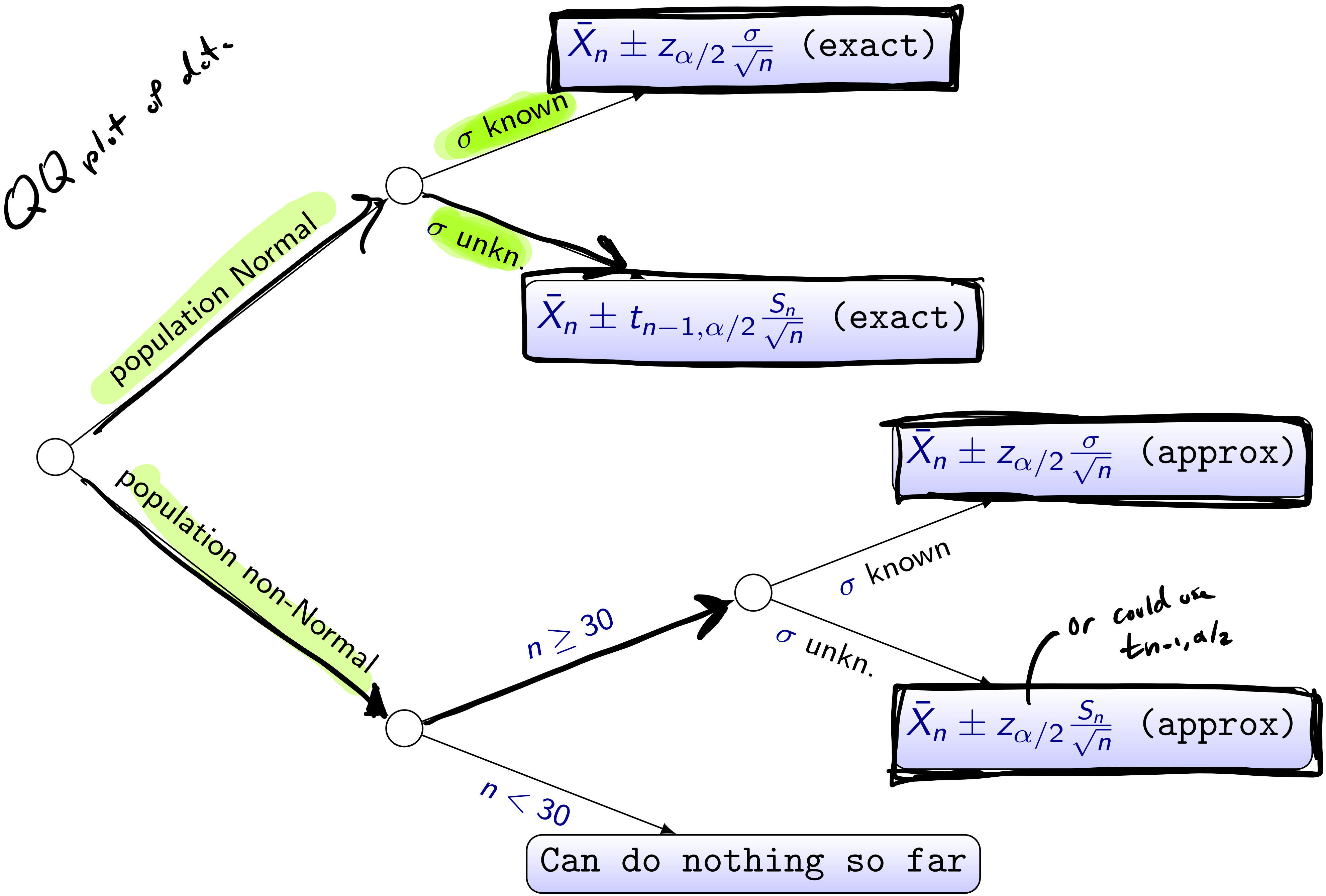
CI for mean of non-Normal population with σ unknown

Let $X_1, \dots, X_n$ be a rs from a pop. with mean μ, and with $\mu_4 < \infty$, then

$$\bar{X}_n \pm z_{\alpha/2} \cdot \frac{S_n}{\sqrt{n}}$$

is an approximate $(1-\alpha) \times 100\%$ CI for μ when n is large (≥ 30, say).

In the above $\mu_4 = \mathbb{E}|X_1|^4$. This limits the heavy-tailedness of the population.


QQ plot of data
population Normal
σ known
$\bar{X}_n \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ (exact)
σ unkn.
$\bar{X}_n \pm t_{n-1,\alpha/2}\frac{S_n}{\sqrt{n}}$ (exact)
population non-Normal
$n \geq 30$
σ known
$\bar{X}_n \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ (approx)
σ unkn.
or could use $t_{n-1,\alpha/2}$
$\bar{X}_n \pm z_{\alpha/2}\frac{S_n}{\sqrt{n}}$ (approx)
$n < 30$
Can do nothing so far