Recap:

- If $\hat{\theta} \sim N[\theta, \underbrace{var(\hat{\theta})}_{\text{known}}]$, then $\hat{\theta} \pm z_{\alpha/2}\, SE(\hat{\theta})$ is a $100(1-\alpha)\%$ CI for θ.

- Pivot: $\underbrace{\dfrac{\hat{\theta}-\theta}{SE(\hat{\theta})}}_{\text{A variable that has a completely known distribution}} \sim N(0,1)$

- CI is approximate if the normal distribution is an approximation to the $\underbrace{\text{dist. of } \hat{\theta}}_{\text{sampling distribution of } \hat{\theta}}$

Confidence interval for population mean μ

Recall:

Case 1: The sample comes from a normal distribution with known variance. In this case,

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

Use $\bar{X}$ as an estimator of μ.

Know: $\bar{X} \sim N\left[\mu, \frac{\sigma^2}{n}\right]$. $\Rightarrow SE(\bar{X}) = \frac{\sigma}{\sqrt{n}}$

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

Case 2: The sample comes from a any distribution, but n is large. In this case,

Know: $\bar{X} \overset{.}{\sim} N\left[\mu, \frac{\sigma^2}{n}\right]$

$\Rightarrow$ Approx. CI for μ: $\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ — if σ is known

$\bar{X} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$ — if σ is unknown

s → sample SD.

σ^2 n

Ex: Suppose that an observed sample of size 20 from a $\underbrace{N(\mu, 10)}_{\text{Dist. of } X}$ population gives $\overline{x} = 2.45$ Find the 95% CI for μ.

$1-\alpha = 0.95 \Rightarrow \alpha = 1-0.95 = 0.05$

$\Rightarrow \alpha/2 = 0.025$.

we need: $z_{0.025} = 1.96$ (R)

$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 2.45 \pm 1.96 \frac{\sqrt{10}}{\sqrt{20}} = \underbrace{[1.06, 3.84]}_{\text{plausible values for } \mu}$

Notice that this interval is *fixed* — it's a numerical interval. There is nothing random about it.

Q: Can we say that this observed interval contains the true value of μ with 95% probability?

No. don't know $\overset{?}{\underset{1}{\overset{0}{\text{or}}}} > \underbrace{P[\underset{\text{fixed (unknown)}}{\mu} \in \underset{\text{fixed}}{[1.06, 3.84]}]}$ $\overset{?}{=} 0.95$

So, how do we really interpret a CI?

Know: $P\Big[\underbrace{\overline{X} - 1.96\frac{\sigma}{\sqrt{n}}}_{\text{random}} \le \mu \le \underbrace{\overline{X} + 1.96\frac{\sigma}{\sqrt{n}}}_{\text{random}}\Big] = 0.95$

$\overline{X}$ is random.

Interpretation of a CI

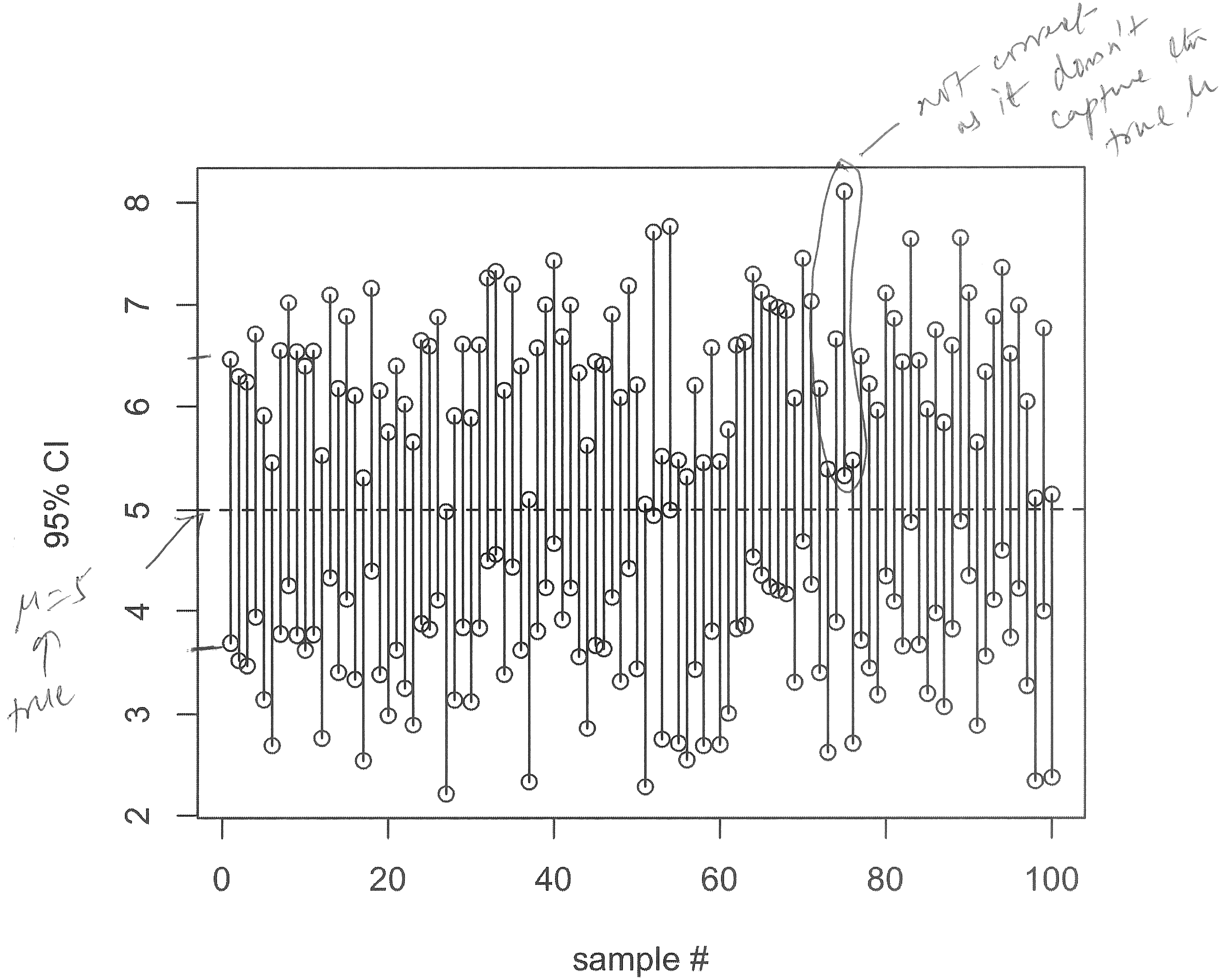
Recall:

- Usual long-term proportion interpretation of probability — i.e., if we repeat a large number of times the process of taking a random sample of size n from $N(\mu, \sigma^2)$ population and construct the CI using the above formula, then roughly 95% of times the observed CIs will be correct, i.e., it will capture the true value of μ.
- This CI formula gives an incorrect interval 5% (small) of the times.
- It is **wrong** to say that the observed interval contains the true value of μ with 95% probability. The CI either contains the true value of μ or it does not — we don't know what the case is.
- Thus, in a sense, we have 95% confidence in the CI formula — it gives the correct answer 95% of the times.

Dist. of X μ

Lets use simulation to verify this interpretation.

- Draw a random sample of size 20 from a $N(5, 10)$ σ^2 distribution and use the observed sample to construct a 95% CI for μ using the above formula.
- Repeat this procedure 10,000 times. The figure on the next page plots the constructed CIs for the first 100 samples.
- Find the proportion of times the CI captures the true value.



95% CI
sample #
2
3
4
5
6
7
8
0
20
40
60
80
100
$\mu = 5$ true
not correct as it doesn't capture the true μ

```r
# A function to simulate data from a N(mu, sigma^2)
# distribution and computing CI

conf.int <- function(mu, sigma, n, alpha){
x <- rnorm(n, mu, sigma)
ci <- mean(x) + c(-1,1) * qnorm(1-(alpha/2)) *
sigma/sqrt(n)
return(ci)
}

# Get one CI

mu <- 5
sigma <- sqrt(10)
n <- 20
alpha <- 0.05
```

$\bar{x}$ $\sigma/\sqrt{n}$ $z_{\alpha/2}$

```
# > conf.int(mu, sigma, n, alpha)
# [1] 3.520961 6.292768

# Repeat the process nsim times

nsim <- 10000
ci.mat <- replicate(nsim, conf.int(mu, sigma, n, alpha))

# > dim(ci.mat)
# [1]     2 10000

# The first 5 intervals

# > ci.mat[, 1:5]
          # [,1]     [,2]     [,3]     [,4]     [,5]
# [1,] 3.689654 3.519999 3.466402 3.937424 3.140117
# [2,] 6.461462 6.291807 6.238210 6.709231 5.911925
# >
```

2x 10,000

#1

#2

(l u)

(l u)

```
# Graphing the first 100 intervals

plot(1:100, ci.mat[1, 1:100],
ylim=c(min(ci.mat[,1:100]), max(ci.mat[,1:100])),
xlab="sample #", ylab="95% CI", type="p")
points(1:100, ci.mat[2, 1:100])
for (i in 1:100) {
segments(i, ci.mat[1, i], i, ci.mat[2,i], lty=1)
}
abline(h=5, lty=2)
```

see R handout on cleaning for doing this in one line

```
# Proportion of times the interval is correct
```

$l \le \mu \le u$

```
# > mean( (mu >= ci.mat[1,])*(mu <= ci.mat[2,]) )
# [1] 0.9502
# >
```

$l \le \mu$

$u \ge \mu$

close to the true prob 0.95 that we expect.

Confidence interval for a normal mean (known variance, cont'd)

Q: Given a random sample, which CI for μ would you prefer — a 95% CI or a 99% CI? (Note: `qnorm(0.975) = 1.959964`, `qnorm(0.995) = 2.575829`.)

$$\boxed{\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}}$$

Higher confidence $\Rightarrow$ wider ~~confident~~ interval.

Note the tradeoff:

- The **precision** of a CI is given by its **width**. The **accuracy** of a CI is given its **confidence level**.
- Higher confidence = lower precision (wider).
- The width of a 100% CI is: $(-\infty, \infty)$ It is a useless interval — extremely "accurate" but extremely imprecise!

CI : $[L, U]$

How much cash do I have?

<u>Guess 1:</u> $ [50, 150]. — $\underbrace{\text{more useful}}_{\to\ \text{more precise}}$ but may be less accurate

<u>Guess 2:</u> $ [0, 10K] — $\underbrace{\text{less useful}}_{\to\ \text{less precise}}$ but may be more accurate

Q: What can we do to get a narrower CI without lowering the confidence?

$\hat{\theta} \pm \boxed{z_{\alpha/2}\, SE(\hat{\theta})}$

"margin of error"

- Width = $U - L = 2 z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$
- Increase n to make CI more precise.

Choosing the sample size n:

- Let w = desired CI width for $1 - \alpha$ confidence.
- Margin of error = $w/2$
- Set the CI width to the desired width and solve for n to get

$$2 z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = w$$

$$\Rightarrow \sqrt{n} = \frac{2 z_{\alpha/2}\, \sigma}{w}$$

$$\Rightarrow n = \left[\frac{2 z_{\alpha/2} \cdot \sigma}{w}\right]^2 \longrightarrow$$ round up to make it an integer.

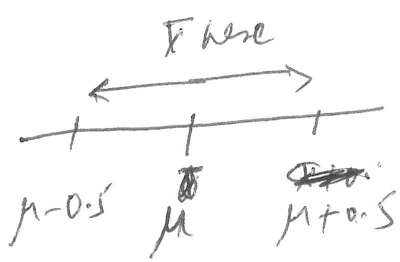
$\bar{X}$

Ex: Suppose that we wish to estimate the mean CPU service time of a job and we wish to assert with 99% confidence that the estimated value is within less than 0.5 sec of the true value. Suppose that the past experience suggests that CPU service time is normally distributed with standard deviation $\sigma = 1.5$ sec. How many observations should we take?

$X =$ CPU time of a $\underbrace{\text{randomly selected}}_{\text{(typical)}}$ job (sec).

Know: $X \sim N[\mu, \sigma^2 = 1.5^2]$

$1 - \alpha = 0.99$



$0.5 =$ desired margin of error

$\Rightarrow$ $W =$ desired width $= 2(0.5) = 1$ sec.

$$n = \left[\frac{2\, z_{0.005}\, 1.5}{1}\right]^2 \approx 60 \quad \text{(using } N\text{)}.$$

So far: Assumed that $X \sim N[\mu, \sigma^2]$, where σ^2 is known.

Confidence interval for a normal mean (unknown variance)

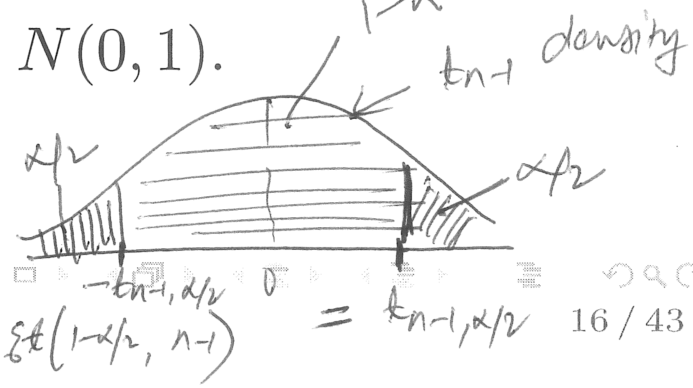
- Unknown variance σ^2 is more realistic.
- Estimate σ^2 by sample variance, $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

Pivot:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

Result: $T \sim t_{n-1}$, i.e., a t-distribution $(n-1)$ degrees of freedom, instead of the $N(0,1)$ distribution.

- A t_{n-1}-distribution looks like a $N(0,1)$ but it has heavier tails. A heavier tail accounts for the fact that there is there is more uncertainty in T when S is used in place of σ
- When n is large, a t_{n-1}-distribution $\approx N(0,1)$.
- As $n \to \infty$, t_{n-1} dist. $\to N(0,1)$.
- $t_{n-1,\alpha/2} \gg z_{\alpha/2}$
- As $n \to \infty$, $t_{n-1,\alpha/2} \to z_{\alpha/2}$



Result: CI for μ: $\overline{X} \pm t_{\alpha/2,n-1} S/\sqrt{n}$

Proof:

By def: Know: $P\left[-t_{n-1,\alpha/2} \le \underbrace{\cancel{T}}_{=T} \le t_{n-1,\alpha/2}\right] = 1-\alpha$ ← t_{n-1} r.v.

Want: $1-\alpha \overset{?}{=} P\left[\bar{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} \le \mu \le \bar{X} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right]$

subtract $\bar{X}$

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

Divide by $S/\sqrt{n}$

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

multiply by -1

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

$= 1-\alpha.$

- For large n, $t_{n-1,\alpha/2} \approx z_{\alpha/2}$.

- 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).