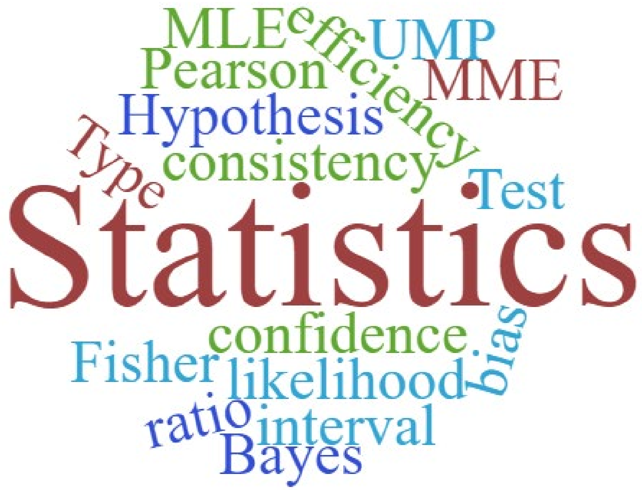
Math 181A: Mathematical Statistics



David Quarfoot
Associate Teaching Professor
Email: dquarfoot@ucsd.edu
Office: APM 5220

Learning Goals:

- Recognize why an interval estimator might feel natural (compared to a point estimator)
- Understand the design of a confidence procedure (CP)
- Develop a 95% CP for the mean of a normal distribution with known variance
- Define quantiles and see how they can be used to change the confidence level of a CP
- Develop an X% CP for the mean of a normal distribution with known variance

Confidence Intervals (CIs), Day 1

As you've seen on homework, an estimator (be it an MME or MLE) is a function of your random data: $\widehat{\theta} = \widehat{\theta}(X_1, X_2, \ldots, X_n)$. As such, an estimator is itself a RV and has a distribution with some center and spread.

Important Question, If $\widehat{\theta}$ was built to estimate the parameter θ (a fixed value from a continuous RV), then what is $P(\widehat{\theta} = \theta)$ and why?

Recall: If X is a continuous RV, then $P(X = a) = P(a \leq X \leq a) = \int_a^a f_X(x)\, dx = 0$.

Because of this issue, some people prefer to create an *interval estimate* for θ (a **confidence interval**), rather than a *point estimate* (an expression/value for $\widehat{\theta}$)

Enter the Confidence Procedure (CP)

Neyman's (1937) idea for a CP:

- Given a random sample $X_1, \ldots, X_n$, use this to generate an interval by creating a lower bound L and an upper bound U.
- Note that $L = f(X_1, \ldots, X_n)$ and $U = g(X_1, \ldots, X_n)$ are both RVs, and hence, the CP creates intervals whose endpoints are RVs.
- Goal: Design f and g so that $P(L < \theta < U) = 1 - \alpha$, where α is small (often 0.05) and decided by the user.

The idea: Make a process that gives intervals that *usually* capture (contain) the true value we are looking for. Some intervals may not contain θ, but we have high confidence in *the process*. $1 - \alpha$, or $(1 - \alpha) \cdot 100\%$, is known as the **confidence level** and is often 90%, 95%, or 99% ($\alpha = 0.1, 0.05, 0.01$).

CP How-To Manual: $X \sim N(\mu, \sigma^2)$, μ Unknown, σ^2 Known

For a normal population with unknown center and known variance, the MLE is $\widehat{\mu} = \overline{X}$. We now design a 95% CP for capturing μ.

Notice that $\widehat{\mu} = \overline{X} = \dfrac{1}{n} \sum X_i$ (sum of a bunch of normal RVs!)

Your Turn 1: Find the mean and variance of $\overline{X}$ in terms of μ and σ^2.

$$E[\overline{X}] \overset{\text{HW}}{=} E[X] = \mu. \qquad Var[\overline{X}] \overset{\text{HW}}{=} \frac{Var[X]}{n} = \frac{\sigma^2}{n}.$$

Since $\overline{X} = \dfrac{1}{n} \sum X_i$ is a linear combination of independent normals, it is too (proved via mgfs). Hence, $\overline{X} \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$.

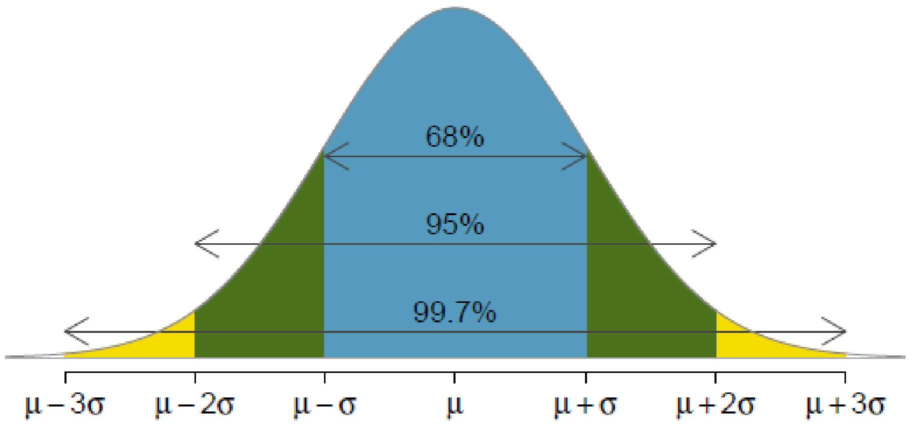
Memorize: $\boxed{X \sim N(\mu, \sigma^2) \implies \overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)}$

Your Turn 2: What is the distribution of $\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} = \dfrac{\sqrt{n}(\overline{X}-\mu)}{\sigma}$?

Take any normal distribution and subtract its mean and divide by its SD and you get the standard normal distribution: $Z \sim N(0,1)$.

Your Turn 3: What does $P(-1.96 < Z < 1.96)$ approximately equal?

The 68-95-99.7 rule says about 95% of the area under a normal is within 2 SDs (or more precisely, 1.96 SDs) of the mean. So, $P(-1.96 < Z < 1.96) \approx 0.95$.



A "Pivot"al Moment

$$\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu$$

Write: $0.95 \approx P\left(-1.96 < \dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} < 1.96\right) = P\left(-1.96\dfrac{\sigma}{\sqrt{n}} < \overline{X}-\mu < 1.96\dfrac{\sigma}{\sqrt{n}}\right)$

$$\mu < \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}$$

$$= P\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

😱 Inequality pivoted to get μ in center! 😱

This says if $L = \overline{X} - 1.96\dfrac{\sigma}{\sqrt{n}}$ and $U = \overline{X} + 1.96\dfrac{\sigma}{\sqrt{n}}$, then $P(L < \mu < U) = 0.95$, or μ has a 95% chance of being in the random interval (L, U).

Your Turn: Sleep Easy 😴

How much does the average college student sleep per night? Assume $X \sim N(\mu, 1)$ is the sleep time in hours of a random college student. Build a 95% CI for μ assuming you gather this data: $x_1 = 7, x_2 = 6.5, x_3 = 7.5, x_4 = 9$.

Note that $\overline{x} = \dfrac{7 + 6.5 + 7.5 + 9}{4} = 7.5.$

Thus, our CI is $\left(\overline{x} - 1.96\dfrac{\sigma}{\sqrt{n}}, \overline{x} + 1.96\dfrac{\sigma}{\sqrt{n}}\right) = \left(7.5 - 1.96\dfrac{1}{\sqrt{4}}, 7.5 + 1.96\dfrac{1}{\sqrt{4}}\right)$

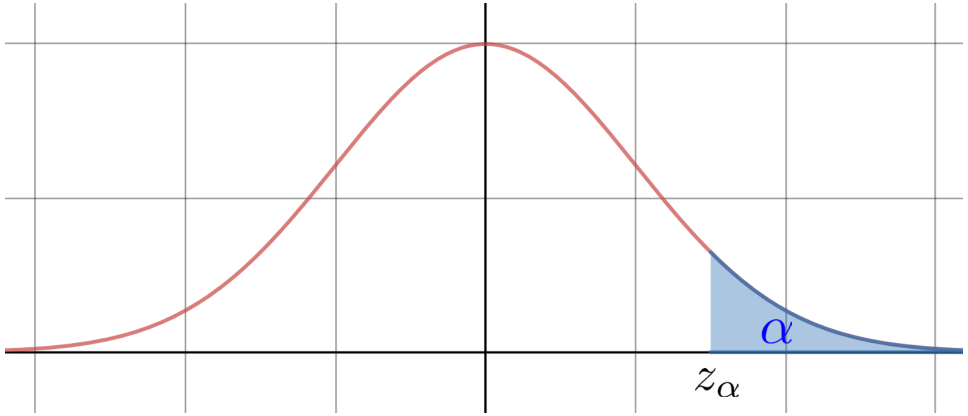
$$= \boxed{(6.52 \text{ hours}, 8.48 \text{ hours})}.$$

Remember:

- Put units on CIs. The units match those on μ.
- For a 95% CP, only 95% of the random CIs contain the parameter, so it is possible that $\mu \notin (6.52, 8.48)$.

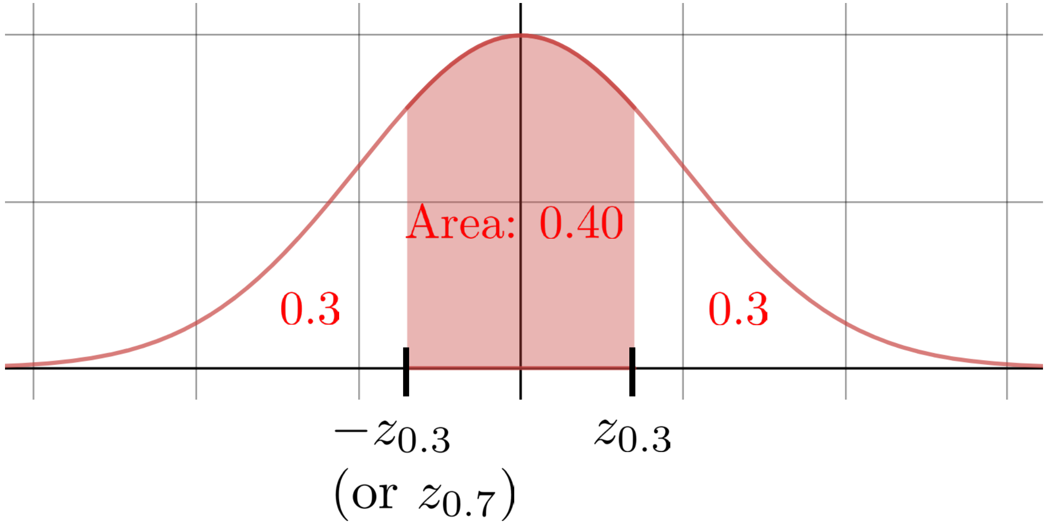
Changing the Confidence Level

Notation: Let z_α be the place on the horizontal axis of $Z \sim N(0,1)$ where the area to the right (under the density curve) is exactly α.



z_α is a "quantile" and must be found using tables or in R.

Your Turn! Put notation on the horizontal labels for this picture:




```
> qnorm(0.3, lower = F)
[1] 0.5244005
```

Claim: The CP $\boxed{\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)}$ has CL $1-\alpha$.

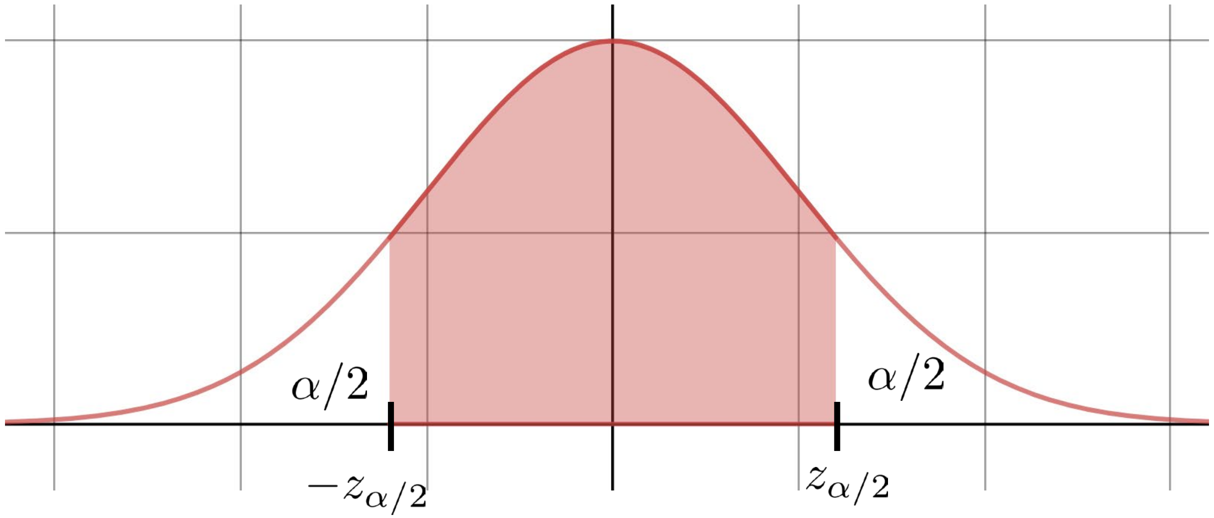
$$-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \overline{X} - \mu \implies -z_{\alpha/2} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$$

$$\mathrm{CL} = P\left(\underbrace{\overline{X} - z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}} < \mu}_{\overline{X}-\mu < z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \implies \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}} \overbrace{< \overline{X} + z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}}\right) = P\left(-z_{\alpha/2} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right)$$

$$= P\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right)$$

$$= 1 - \frac{\alpha}{2} - \frac{\alpha}{2}$$

$$= \boxed{1-\alpha}.$$



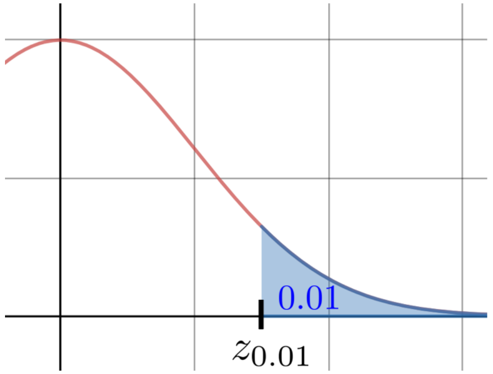
Your Turn: Sleep Cycle 😴

Find a 98% CI for our sleep data ($\overline{x} = 7.5$ where $X \sim N(\mu, 1)$) using R and explain *intuitively* why the answer should be wider than the 95% CI.

Set $98 = 100(1-\alpha) \implies \alpha = 0.02$. So, $z_{\alpha/2} = z_{0.01}$.

Our CI is $\left(7.5 - 2.326\frac{1}{\sqrt{4}}, 7.5 + 2.326\frac{1}{\sqrt{4}}\right)$

$= \boxed{(6.337 \text{ hours}, 8.663 \text{ hours})}$.



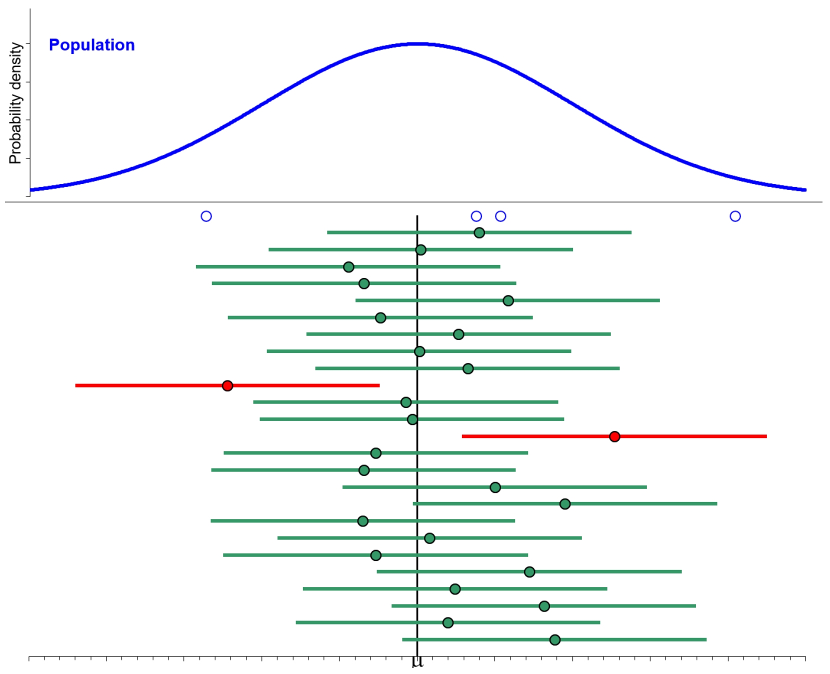

```
> qnorm(0.01, lower = F)
[1] 2.326348
```

Intuitively, to be **more** sure an interval captures μ, you would want a **wider** interval. Metaphor: To be more sure of catching a fish, throw a wider net.

Remember: We will **never** know if $\mu \in (6.337, 8.663)$ or not. Similarly, you never know if your safety is ensured on a particular car ride/flight.

Red Light, Green Light

Explore our CP visually here.



You must be **very** careful with your statements about a particular CI! Probabilities always refer to random processes.

Incorrect: $\mu \in$ our CI.

Incorrect: There is a 95% chance that $\mu \in$ our CI.

Correct: 95% of the random CIs contain μ.

Correct: **If** $\mu \in$ our CI (see last slide), then college students sleep, on average, between 6.337 and 8.883 hours per night.

Your Turn: CI Mania!

You will generate five random data sets of size n from $N(\mu, 3^2)$. For each, you'll build a 65% CI. What is the probability 4 or more of the CIs will capture μ?

Let p be the probability a random 65% CI captures μ. Based on the definition of a CI, $p = 0.65$.

Let X be the number (out of 5) of our CIs that capture μ.
We know: $X \sim Binom(n, p) = Binom(5, 0.65)$.

We want $P(X \geq 4) = P(X = 4) + P(X = 5)$

$$= \binom{5}{4}(0.65)^4(0.35)^1 + \binom{5}{5}(0.65)^5(0.35)^0 \approx \boxed{0.428}.$$