Standard Normal Distribution

Grinnell College

March 26, 2025

Starting Questions

1. Suppose you have a sample with a mean value of 0. What would happen to the mean if you added 10 to all of the observations. What would happen to the standard deviation?
2. Suppose you had a sample with a mean value of 0. What would happen to the mean if you multiplied all of the values by 10. What would happen to the standard deviation
3. Think of the 95% confidence interval for a normal distribution,

$$\overline{x} \pm 1.96 \times \frac{\sigma}{\sqrt{n}} = \left(\overline{x} - 1.96 \times \frac{\sigma}{\sqrt{n}},\ \overline{x} + 1.96 \times \frac{\sigma}{\sqrt{n}}\right)$$

   what quantiles of the distribution would each endpoint represent?

Starting Questions 2

1. Do quantiles exist for all data, regardless of their distribution? If so, how does the 0.05 quantile for a normal distribution compare to a 0.05 quantile of a skewed distribution?
2. Explain why for an 80% confidence interval we need to use the quantiles 0.1 and 0.9 to determine the critical values
3. Suppose that I have two datasets, each with 100 observations
   - Dataset A has a mean of $\mu_A = 50$ and $\sigma_A = 5$
   - Dataset B has a mean of $\mu_B = 25$ and $\sigma_B = 10$
     1. How many observations should fall between the 0.1 and 0.9 quantiles for A?
     2. How many observations should fall between the 0.1 and 0.9 quantiles for B?
     3. If I find the sample mean for each dataset, $\overline{x}_A$ and $\overline{x}_B$, which sample mean will have a larger 95% confidence interval?
     4. If I find the sample mean for each dataset, $\overline{x}_A$ and $\overline{x}_B$, which confidence interval will be centered further to the right?

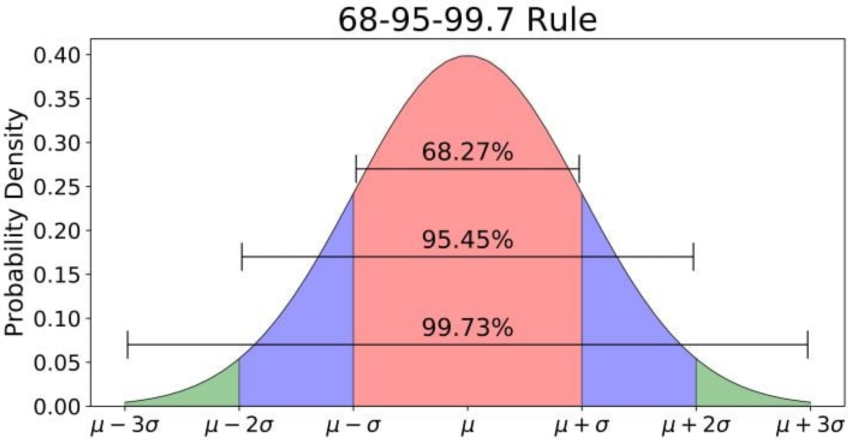
Critical values

Recall that the CLT gives us the distribution of $\overline{X}$,

$$\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

From this, we are interested in providing an interval of plausible values for μ, based on our observation of $\overline{X}$, typically based on percentiles

One of the great benefits of knowing our sampling distribution is that it allows us to relate our percentiles in terms of standard errors


68-95-99.7 Rule
Probability Density
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
68.27%
95.45%
99.73%
μ − 3σ
μ − 2σ
μ − σ
μ
μ + σ
μ + 2σ
μ + 3σ

Quantiles

Consider an expression written like:

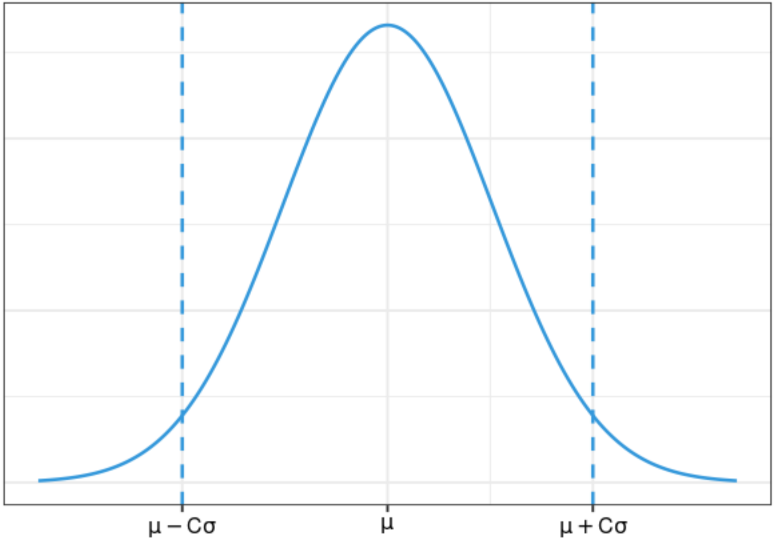
$$\overline{x} \pm C \times \frac{\hat{\sigma}}{\sqrt{n}}$$

This C term, called a **critical value** gives me my relationship between percentiles and standard deviation for normal.

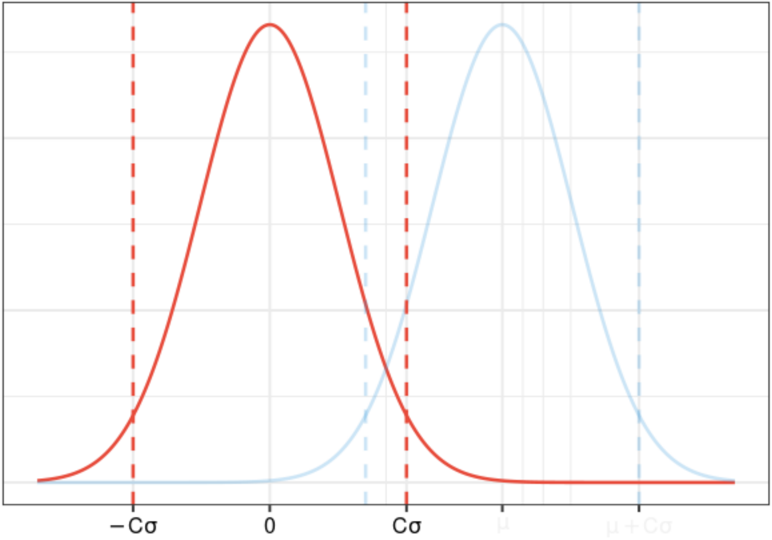
For example, based on the empirical rule on the previous slide, we know that setting $C = 2$ will cover 95.45% of our distribution

What would be of interest is a way to find values of C for any level of confidence

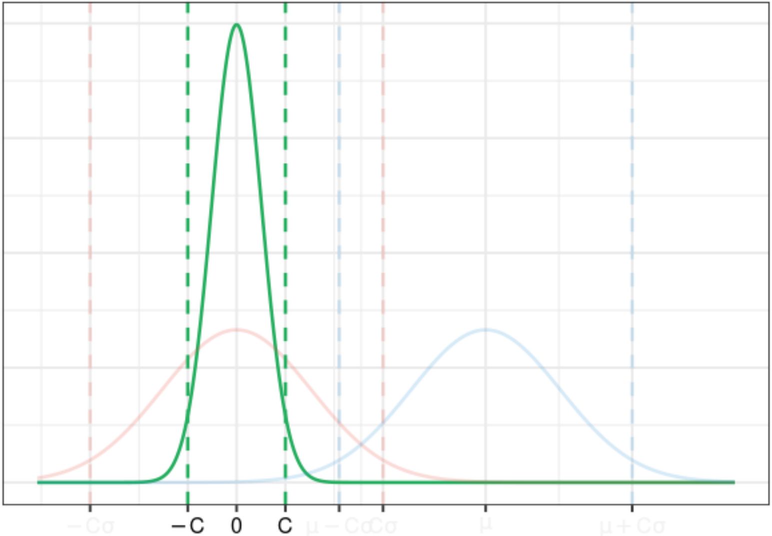
$$X \sim N(\mu, \sigma)$$



$$(X - \mu) \sim N(0, \sigma)$$



$$\frac{X-\mu}{\sigma} \sim N(0,1)$$



Standardization

Previously we saw that

$$Z = \frac{X - \mu}{\sigma}$$

would create a standardized variable with mean 0 and sd 1. For a normally distributed variable like $\overline{X}$, this results in the following:

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

We call this a **standard normal** distribution

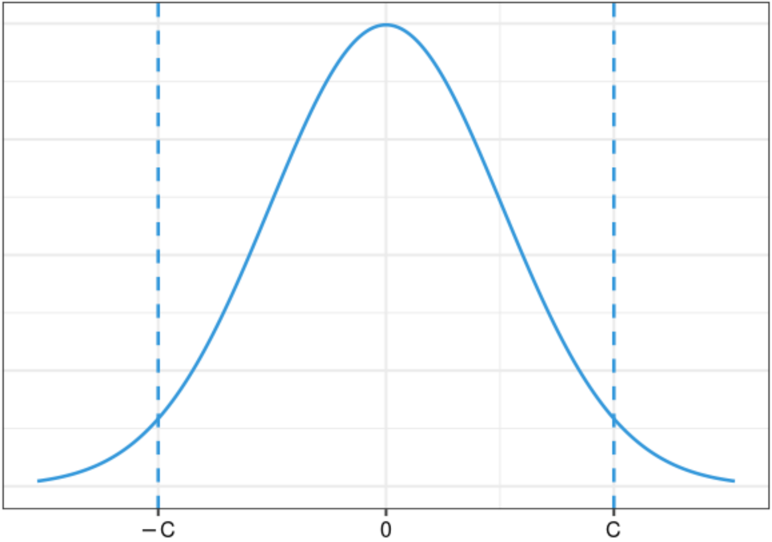
We can find the values for C by considering a standard normal distribution where $\mu = 0$ and $\sigma = 1$

$$\begin{aligned} \mu \pm C \times \sigma &= 0 \pm C \times 1 \\ &= \pm C \\ &= (-C, C) \end{aligned}$$

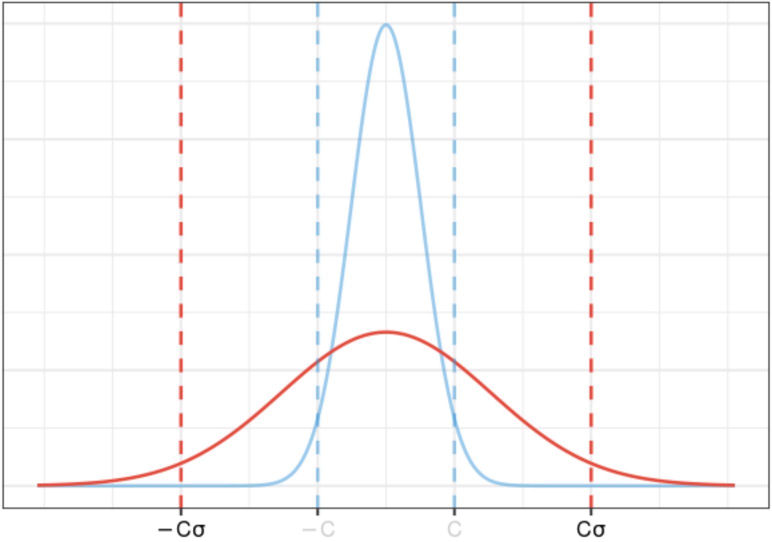
If we want an $m\%$ confidence interval, then, we must choose the values of C such that the interval $(-C, C)$ covers the middle $m\%$ of a standard normal distribution

For a 95% confidence interval, then, that means we need the 0.025 and 0.975 quantiles (i.e., this is the same 2.5th and 97.5th percentiles)

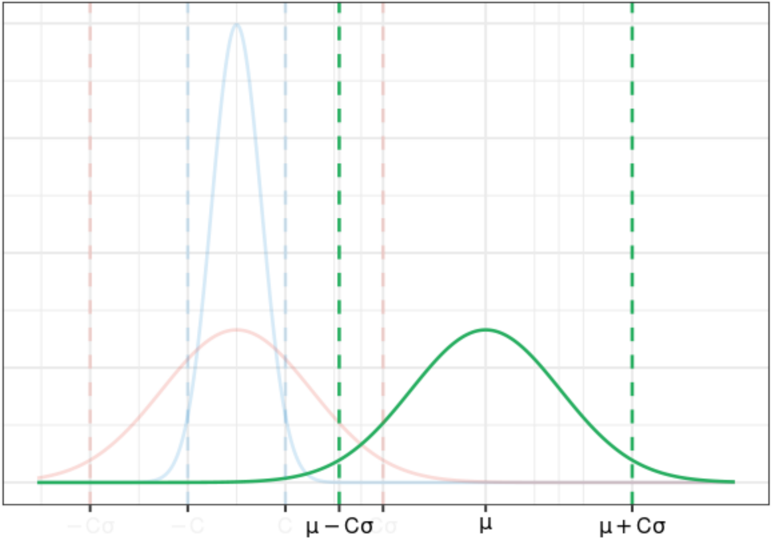
$$\frac{X - \mu}{\sigma} \sim N(0,1)$$



$$(X - \mu) \sim N(0, \sigma)$$



$$X \sim N(\mu, \sigma)$$



Quantiles

We can find the quantiles of a normal distribution with the R function `qnorm` (for **q**uantile of **norm**al) which takes as arguments the quantiles we want, as well as the mean and the standard deviation of the distribution:

```
> quants <- c(0.025, 0.975)

> qnorm(quants, mean = 0, sd = 1)
[1] -1.96  1.96
```

This means that for a *true* 95% confidence interval, we should be using

$$\overline{x} \pm 1.96 \times \frac{\sigma}{\sqrt{n}}$$

Example

To illustrate, consider our penguin dataset, where we consider the flipper length (mm) of male gentoo penguins. Our summary statistics give us the following:

$$\overline{x} = 199.94, \quad \hat{\sigma} = 5.9766, \quad n = 34$$

To find a 95% confidence interval, we could use our formula, $\overline{x} \pm C \times \frac{\hat{\sigma}}{\sqrt{n}}$ *or* we could use the `qnorm` function, passing in the mean and standard error from the CLT:

```
## Using qnorm function
> qnorm(c(0.025, 0.975), mean = 199.91, sd = 5.9766 / sqrt(34))
[1] 197.90 201.92

## Using our formula
> 199.91 + c(-1.96, 1.96)*(5.9766 / sqrt(34))
[1] 197.90 201.92
```

Note

It is worth noting – the values for our critical value for the normal distribution will be the same *regardless* of the sample mean and sample standard deviation

This is because they are based on the theoretical standard normal distribution

Review

Big day today:

- By assuming a distribution, we can use quantiles to determine our **critical values** for constructing confidence intervals
- We can standardize sampling distribution to derive the **standard normal distribution**
- The `qnorm` function will return for us the quantiles of a normal distribution
- Because *critical values* are based on the standard normal distribution, they will be the same regardless of our sample mean and sample standard deviation