Standard Normal

Grinnell College

March 4, 2026

Warmup

- If I have 1,000 observations, how many of them will fall between the 10th and 90th percentile?
- How does sampling distribution differ from distribution of a sample?
- What are distributional parameters of normal distribution? What does this mean?
- If I have two samples with:
  - Sample 1: $n_1 = 25$ and $\sigma_1 = 10$
  - Sample 2: $n_2 = 50$ and $\sigma_2 = 15$

  which sample will have the least variability in its estimate of $\overline{X}$?

Review

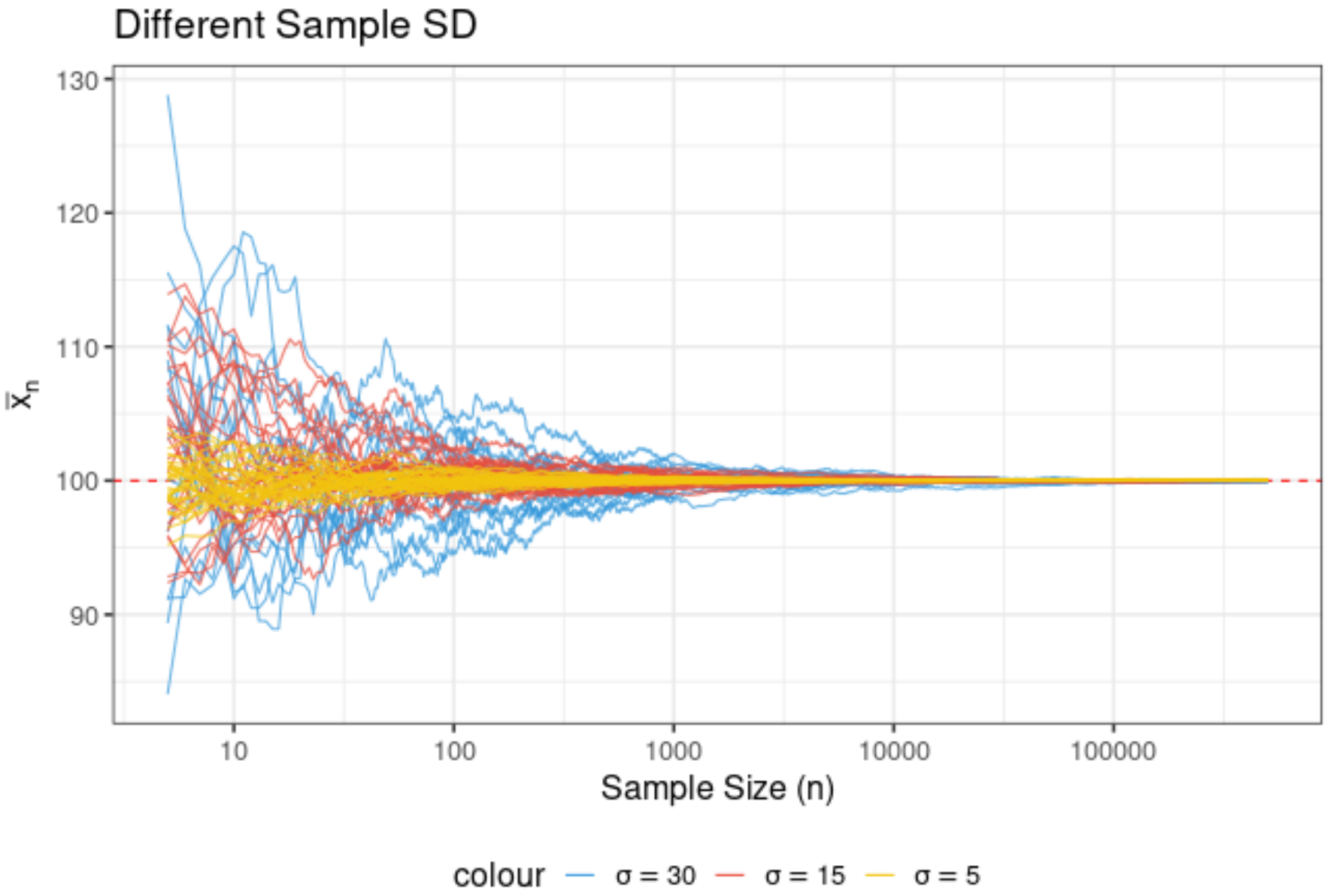
The **Law of Large Numbers** guarantees that, as the number of observations n in my sample increases, my estimate of the parameter will converge to the true value

A **sampling distribution** refers to the distribution of a sample statistic (i.e., $\overline{x}$) if we were to repeatedly sample from a population and recompute the statistic

- What values would they take?
- How frequently would they appear?

The **Central Limit Theorem** states that is my statistic is an average or a proportion, then the sampling distribution of my statistic will be approximately normal, with

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


Different Sample SD
130
120
110
100
90
$\bar{x}_n$
10
100
1000
10000
100000
Sample Size (n)
colour
σ = 30
σ = 15
σ = 5

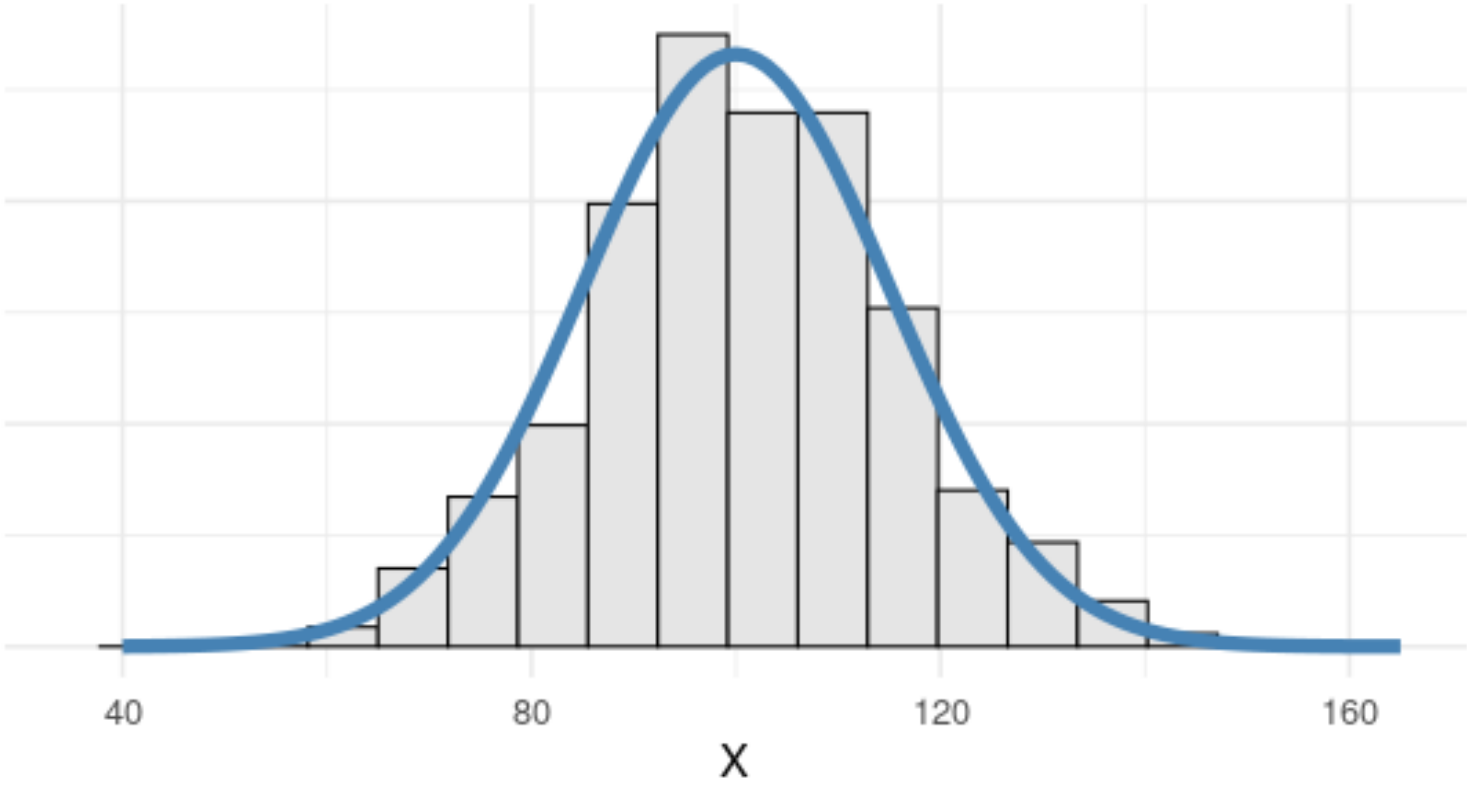
Question

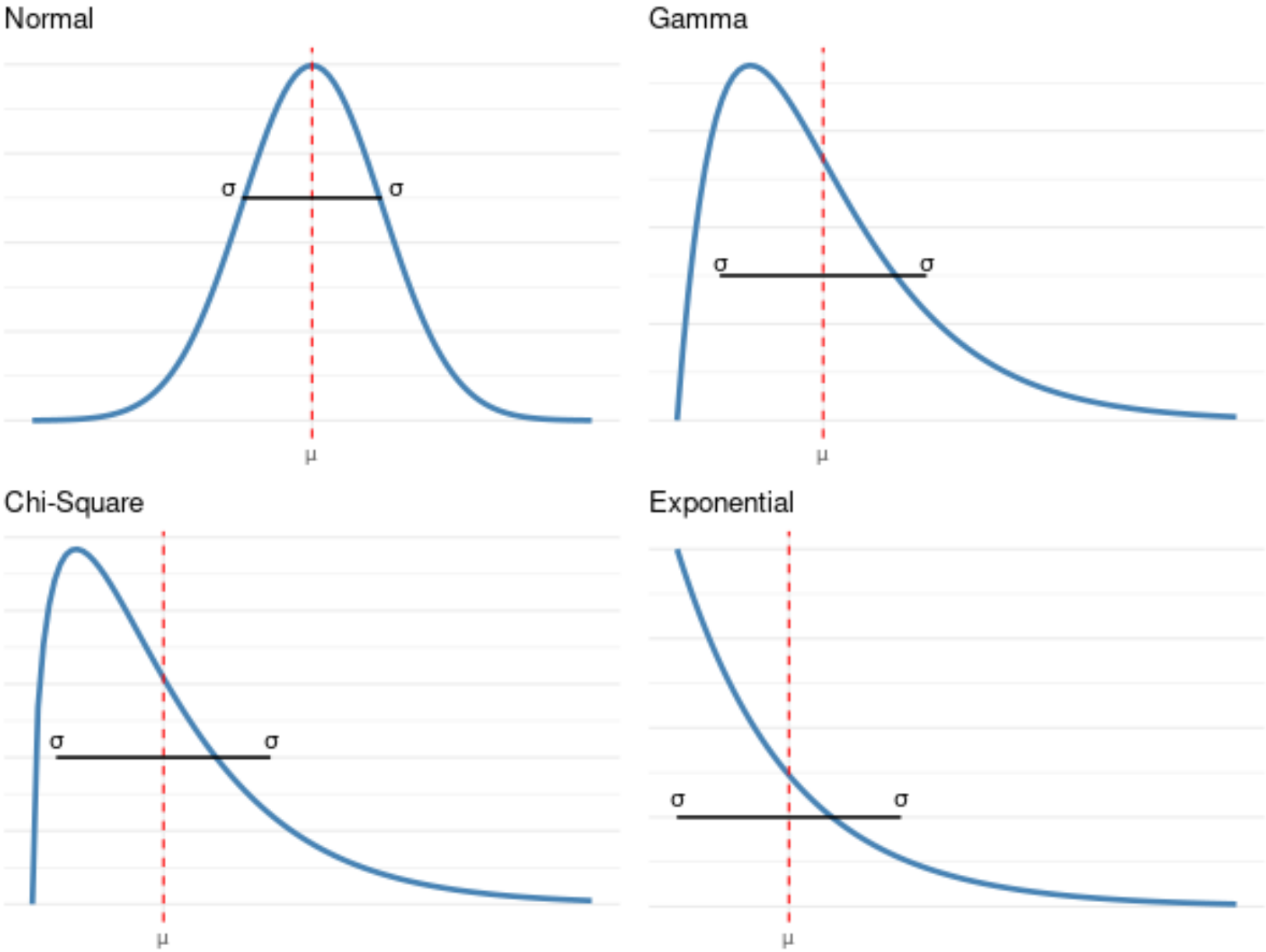
If I know that asymptotically the value of my statistic will get closer and closer to the true value of the parameter, why am I interested in the behavior of a sampling distribution for a fixed sample size?

Notes on Normal

The **normal distribution** describes a distribution that is

- Bell-shaped
- Symmetric about the mean
- Has two distributional parameters, the mean μ and standard deviation σ



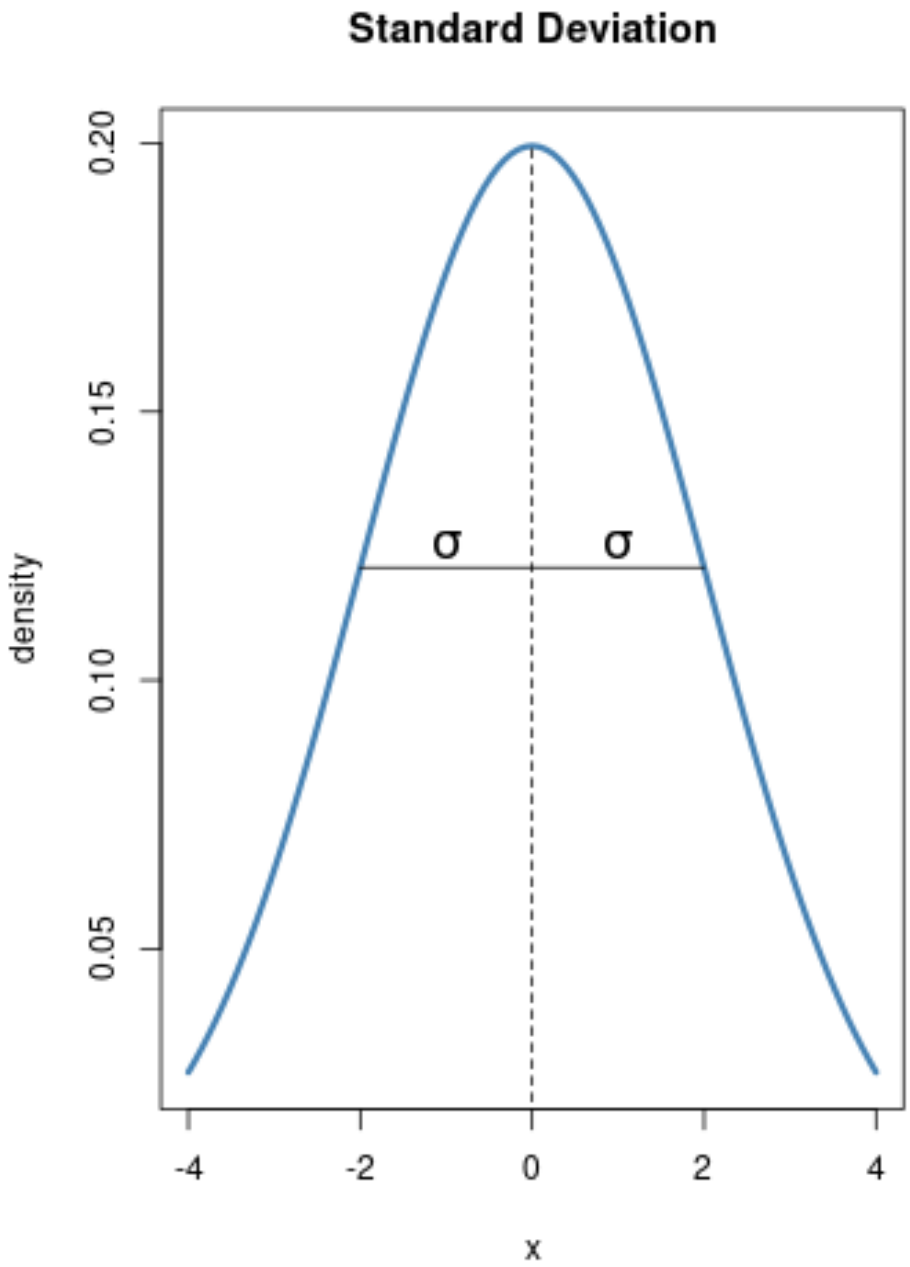

Normal
σ
σ
μ
Gamma
σ
σ
μ
Chi-Square
σ
σ
μ
Exponential
σ
σ
μ

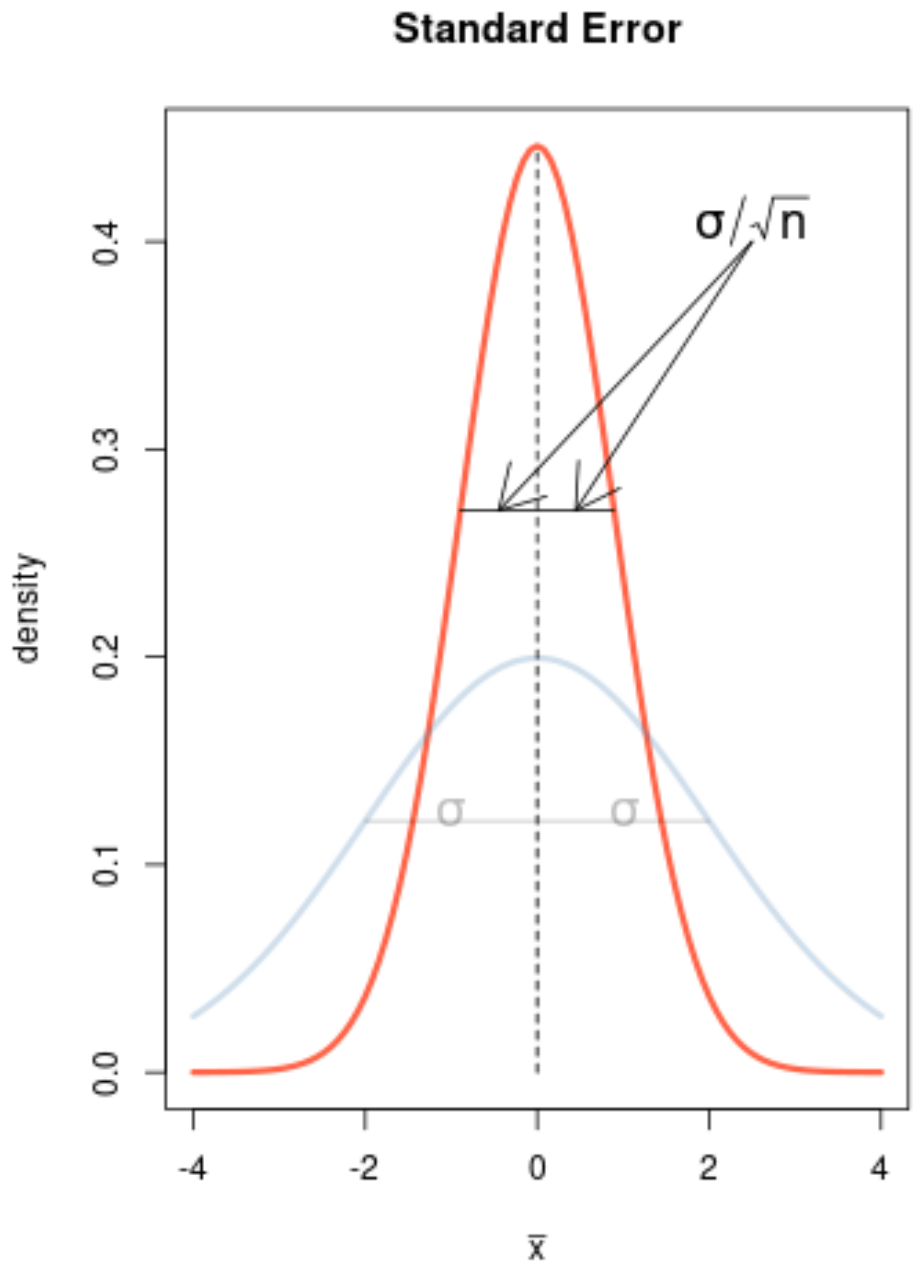
Some Terms to Know

Standard Deviation: A description of the variability in our *observations* describing average distances from the average or mean. It is often denoted σ

Standard Error: A description of variability in our *sampling distribution*. We will denote standard error as SE, with $SE = \sigma/\sqrt{n}$, where n is the number of observations in our sample.

Note that the standard error *is* the standard deviation of the sampling distribution


Standard Deviation
density
0.05
0.10
0.15
0.20
σ
σ
-4
-2
0
2
4
x


Standard Error
density
0.0
0.1
0.2
0.3
0.4
σ/√n
σ
σ
-4
-2
0
2
4
x̄

Standardization

As a consequence of the fact that

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

I know that if I were to collect many many samples of $\overline{X}$, the average value of these $\overline{X}$ would be near μ. Further, the average distance of each $\overline{X}$ from μ would be $\sigma/\sqrt{n}$

What would happen if I were to then standardize each of these values, taking

$$Z_i = \frac{\overline{X}_i - \mu}{\sigma/\sqrt{n}}?$$

Standard Normal Distribution

The **standard normal distribution** (typically represented with the variable Z) is normal distribution with a mean value of 0 and a standard deviation of 1

$$Z \sim N(0, 1)$$

Because standardizing a distribution only shifts and scales the distribution (without changing it's shape), it follows that

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

Example – Penguin

Suppose I have a population of Gentoo penguins with a true average bill length of $\mu = 47.56$ and a standard deviation of $\sigma = 3.1$

Suppose I collect a sample of size $n = 20$ and find my sample mean of $\overline{x} = 48.94$. Is this a value I should expect to get from a random sample? Or does it appear to be an outlier?

Example – Penguin

We can start by standardizing to see how this sample mean falls relative to the distribution of means we might get

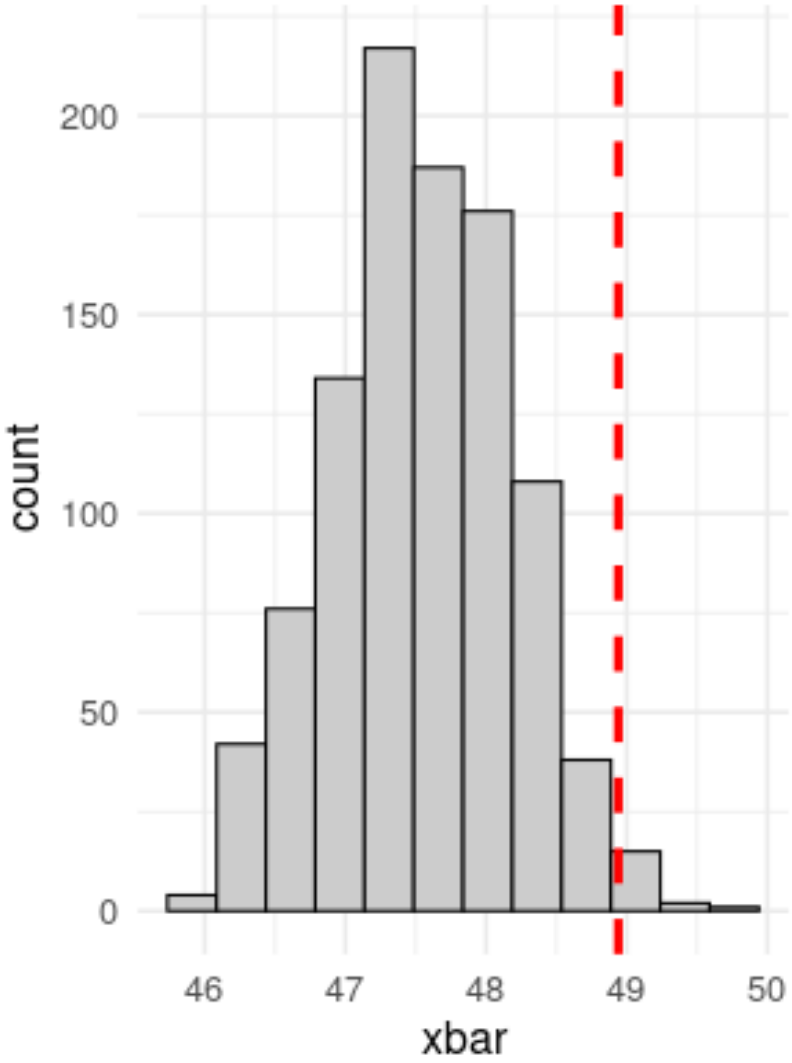
$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

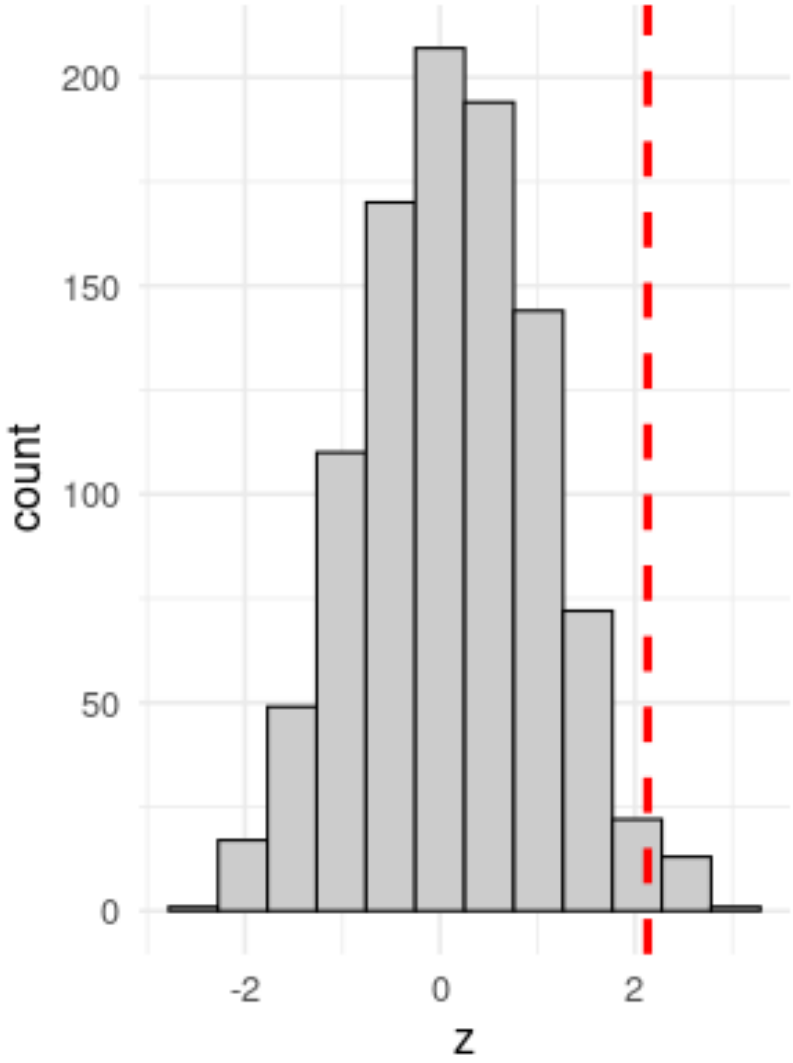
$$= \frac{48.94 - 47.46}{3.1/\sqrt{20}}$$

$$= 2.13$$

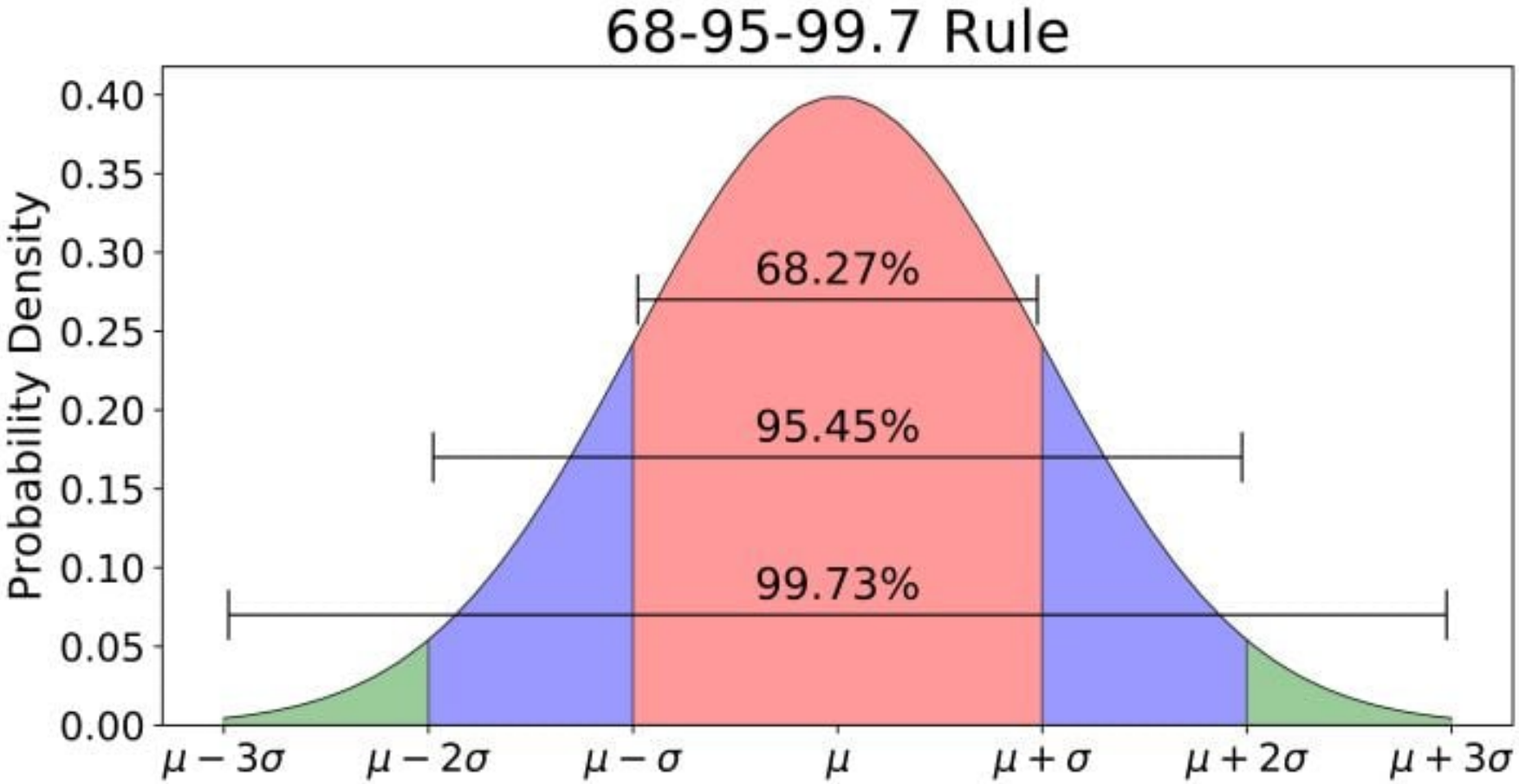
Indicating that the mean of this sample is about 2.13 standard deviations away from what we would expect

Example – Penguins





Empirical Rule



A Simulation of Sorts

We see from a simulation of sampling, from the 1000 standardized values we collected, 701 were within one standard deviation and 978 were within 2 standard deviations

```
> xbar <- lapply(1:1000, function(x) {
  sample(gentoo$bill_length_mm, size = 20) %>% mean()
})

> z <- (xbar - 47.46)/(3.1/sqrt(20)) # Create std. values

> sum(z < 1 & z > -1)
[1] 701
> sum(z < 2 & z > -2)
[1] 978
> length(z)
[1] 1000
```

Unlikely values

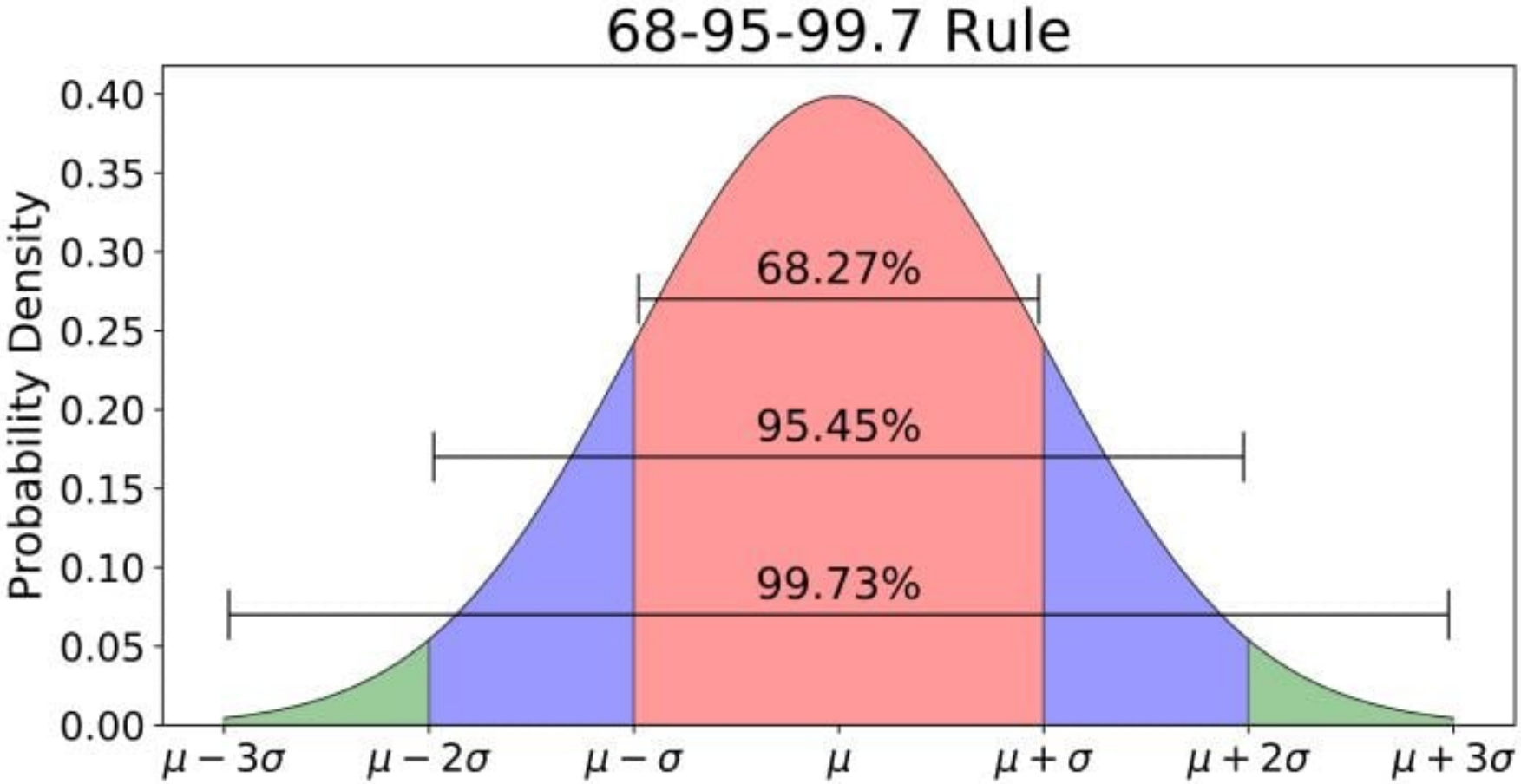
Benefits of a distribution

Empirical Distribution:
Suppose I have gone out and collected a sample:

- If I wanted to find the median of this dataset, what would I do?
- What if I wanted to find Q1 and Q3 of this dataset?

Named Distribution:
What if instead I wanted to find the median and Q1 and Q3 of a normal distribution with mean value μ and standard deviation σ?

Empirical Rule



Determining Range

If we can determine that the probability of our standardized values being between -1 and 1 to be

$$P(-1 < Z < 1) = 0.68$$

and

$$P(-2 < Z < 2) = 0.9545$$

Then theoretically to get the middle $M\%$ of values we should be able to find values C such that

$$P(-C < Z < C) = M\%$$

Why might this be useful?

Review

AHHHHHHHHH!!!!!!!!!!!!!!!!!!