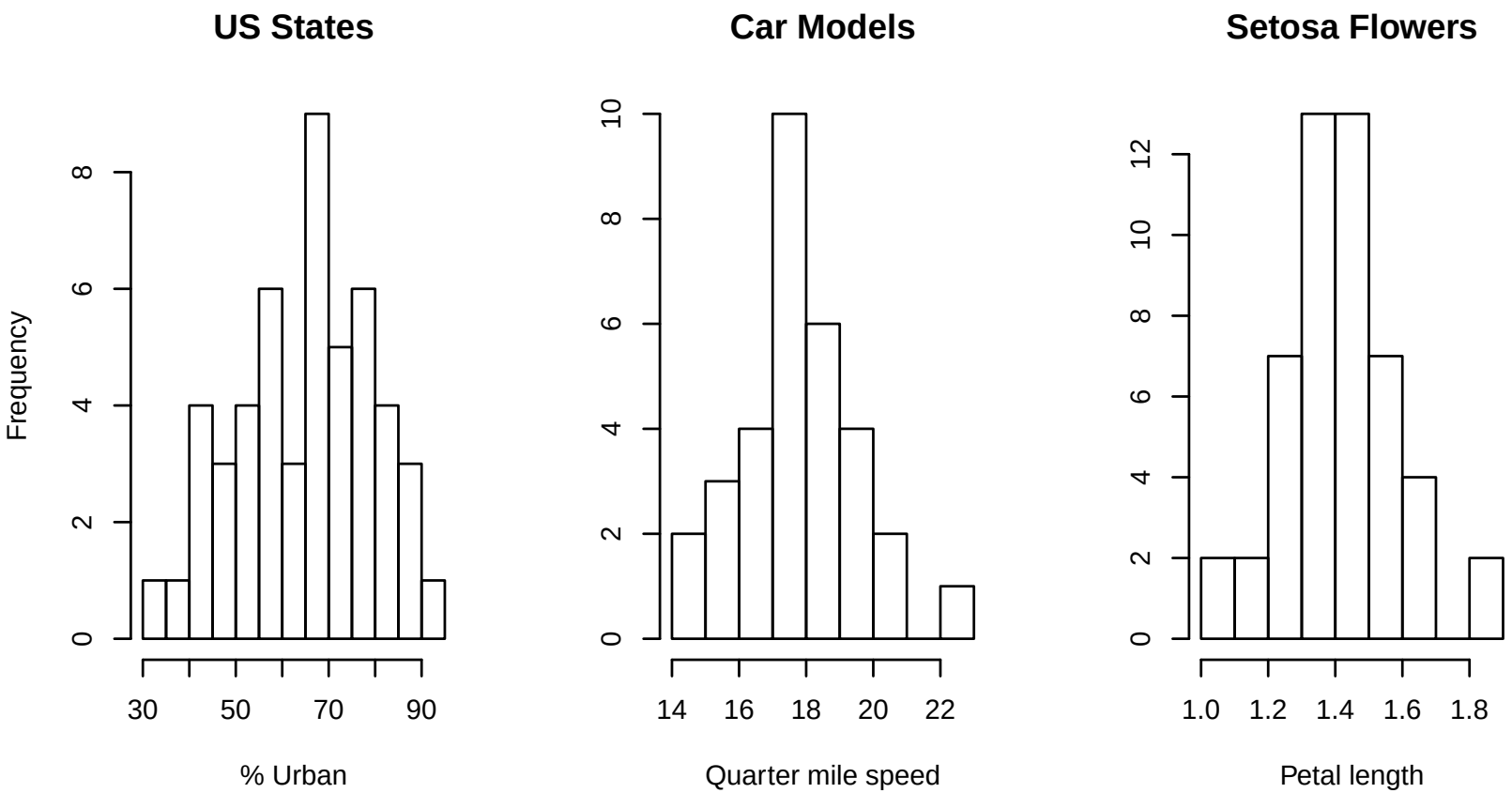
Normal Distributions and The Central Limit Theorem

Ryan Miller

Different but Similar Data?



In a practical sense, these data are very different. But, ignoring their units, do you see any similarities?

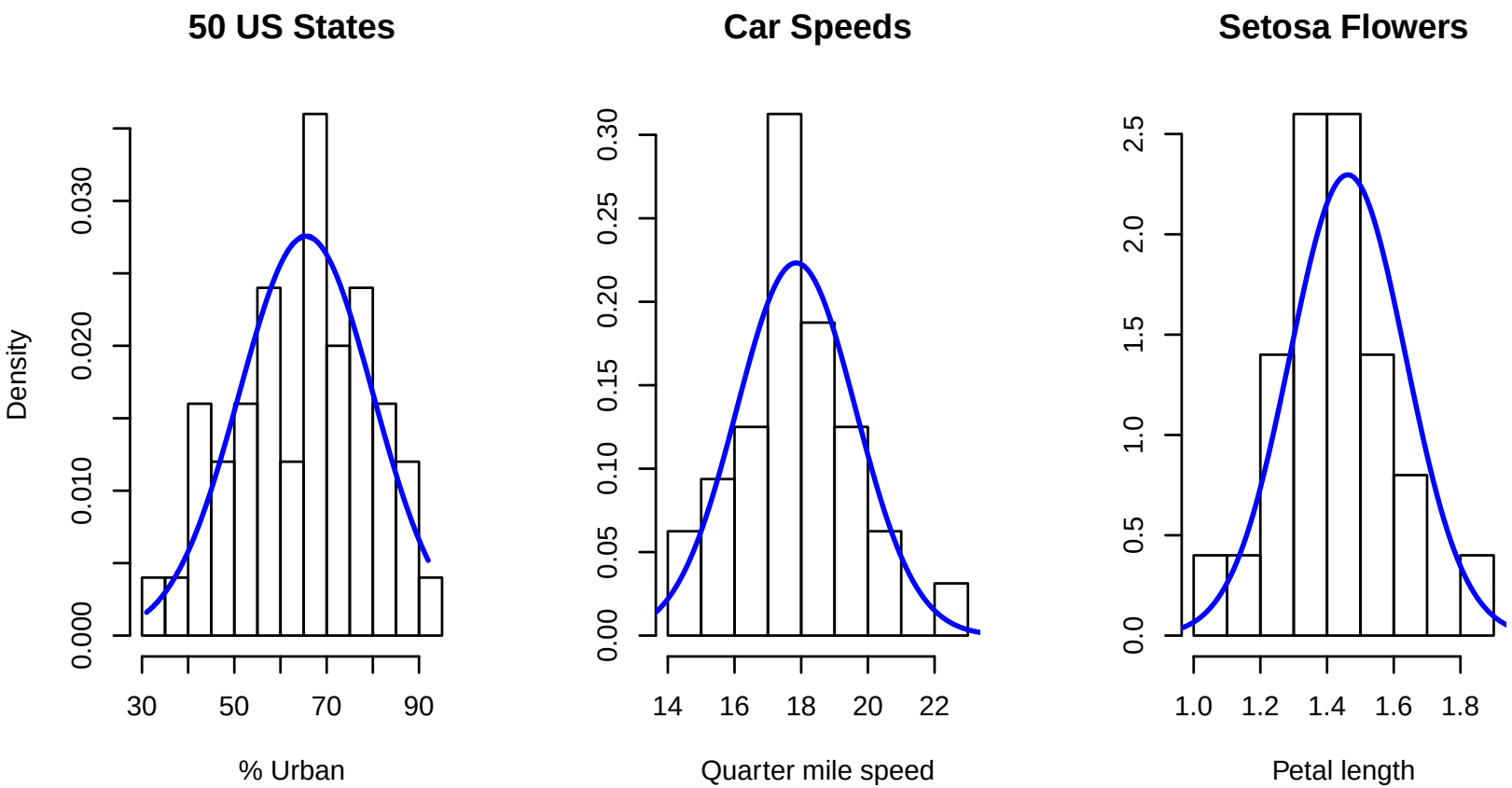
The Normal Distribution

- Mathematicians noticed long ago that a lot of different variables tended to have very similar looking distributions
- These distributions can be characterized by the curve:

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

- This curve defines the **Normal Distribution**
  - μ is the center (mean) of the distribution
  - σ is the standard deviation of the distribution
- You aren't expected to know the formula of the normal curve, but you should know that it depends upon μ and σ

The Normal Distribution



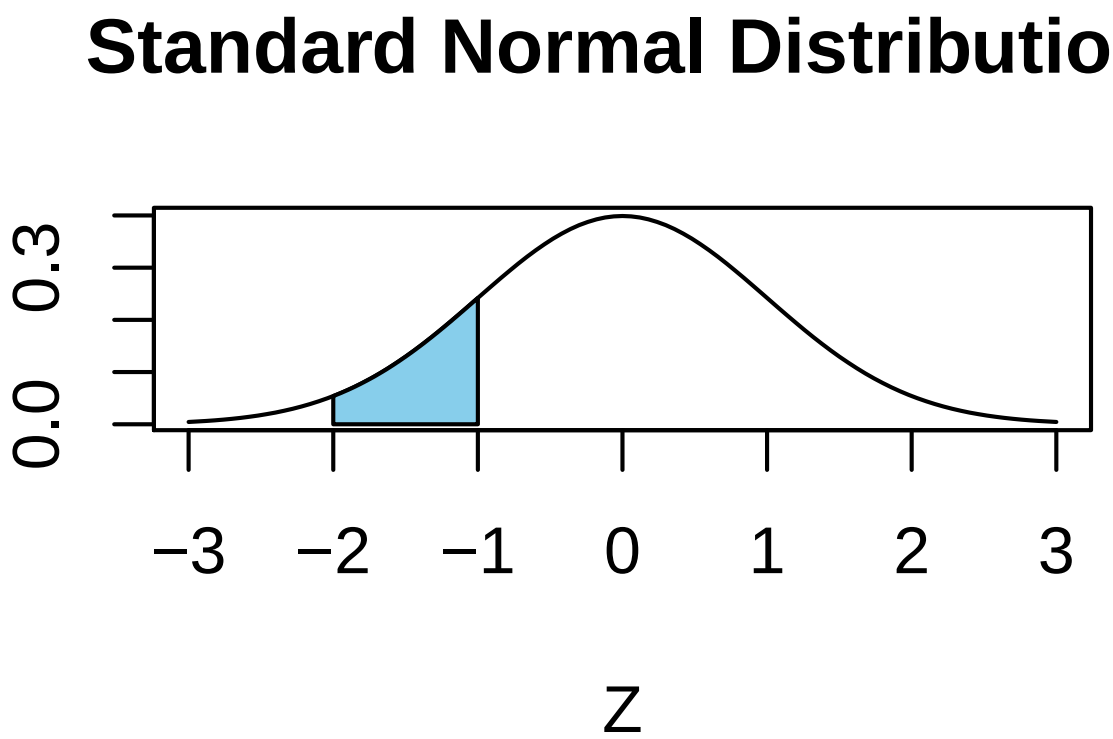
The normal curve is a reasonable approximation of each distribution

The Standard Normal Distribution

- The normal curve can accommodate a variety of different means (μ) and standard deviations (σ)
- It is sometimes convenient to standardize the data, giving us the **standard normal distribution**
  - The standard normal distribution is *parameterized* by $\mu = 0$ and $\sigma = 1$
  - The standard normal distribution displays z-scores instead of the data's original units
  - Any normal curve can be standardized by subtracting the mean and then dividing by the standard deviation (how we calculate z-scores)

Probabilities

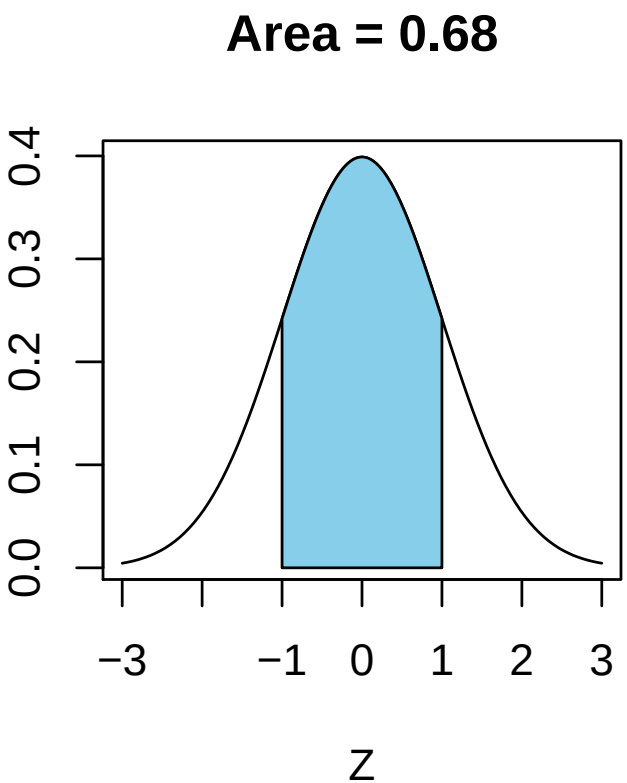
The probability of a given interval of Z values can be found using the area under the curve:

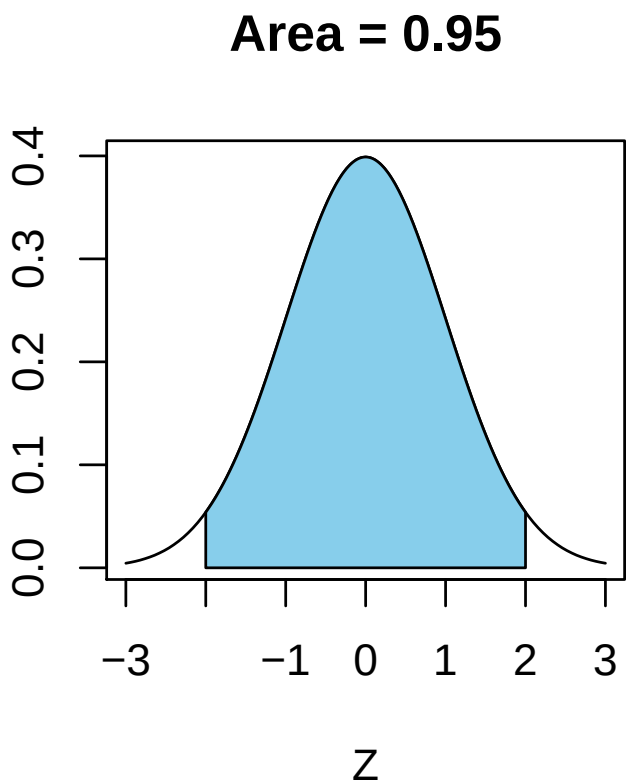


The interval from $Z = -2$ to $Z = -1$ has an area of 0.136, so $Pr(-2 < Z < -1) = 0.136$

Probabilities

This is the origin of the 68%/95% rule for the amount of values that are 1 and 2 standard deviations from the mean:





Probabilities

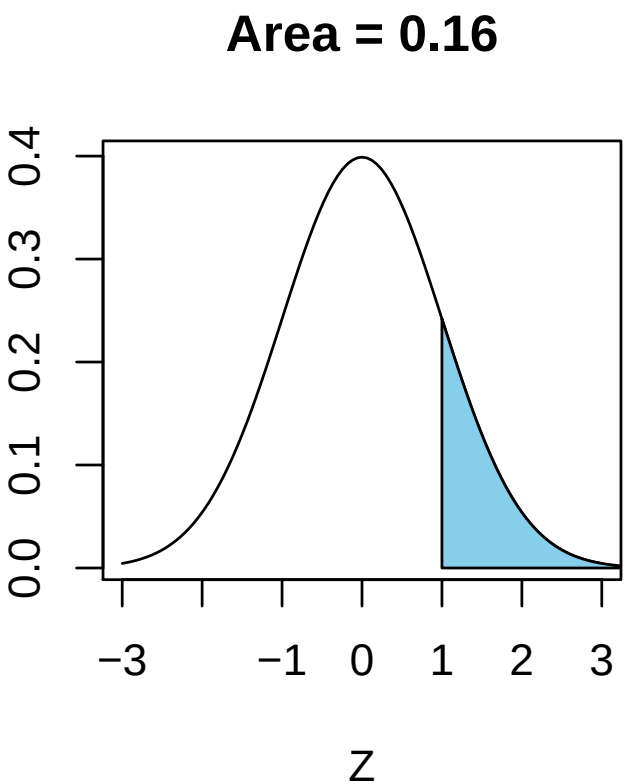
- Generally we calculate the area under a curve by integration
  - Unfortunately there isn't a closed form integral for the normal curve
- However, calculating the area under the normal curve is such a common occurrence in statistics that people have developed very efficient algorithms for doing so
- Historically, the output of these algorithms was aggregated into tables, which allowed for the calculation probabilities without a computer
- Obviously, the preferred option in modern times is to use a computer, which allows us to be the most precise and efficient

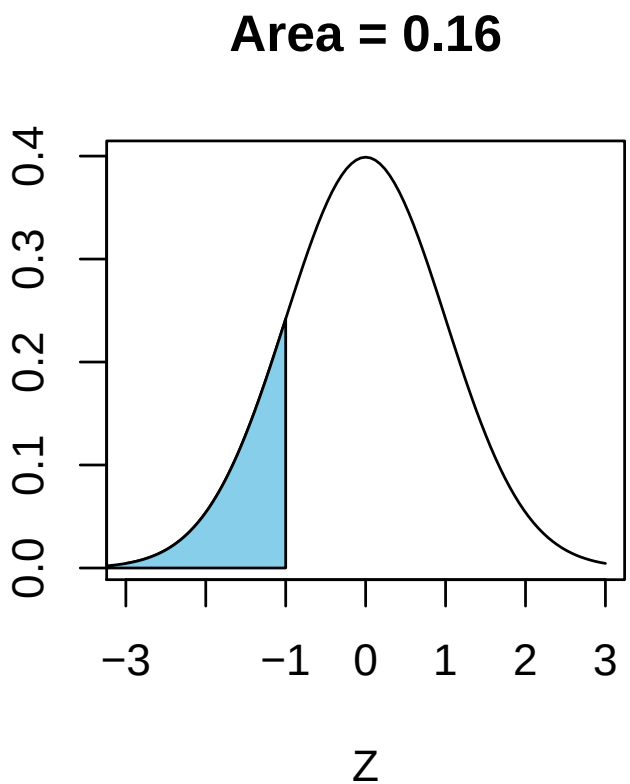
Probabilities

- Most of these algorithms (and tables) are geared towards calculating the area to the left of some threshold value
  - In Minitab, we can use **Calc** -> **Probability Distributions** -> **Normal** and then select "Cumulative probability" and "Input constant"
  - This calculates the area to the left of the constant that was input
- Because these algorithms provide only left tail areas, it important is to understand how to utilize the symmetry of normal curve and a few basic probability rules

Example #1

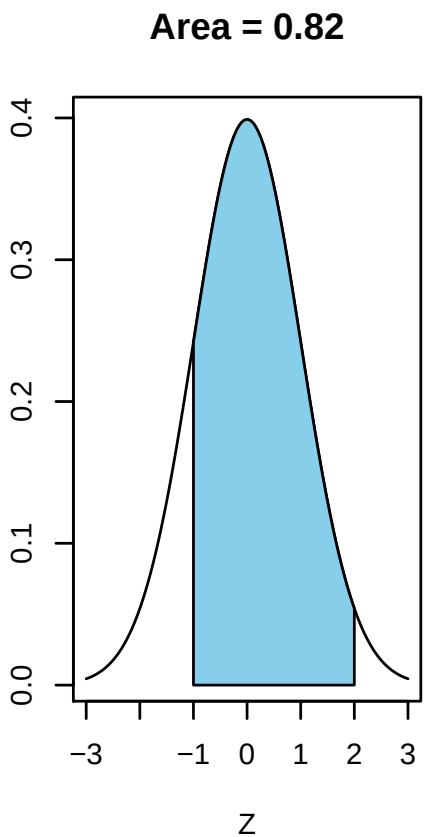
For the standard normal distribution: $Pr(Z \geq t) = Pr(Z \leq -t)$

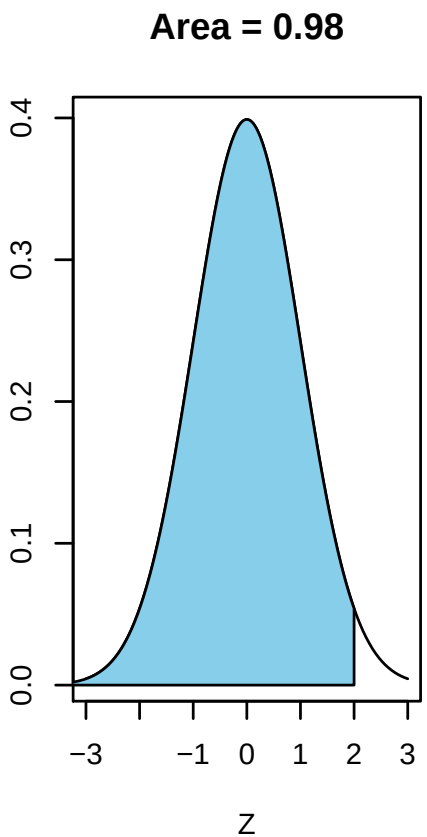


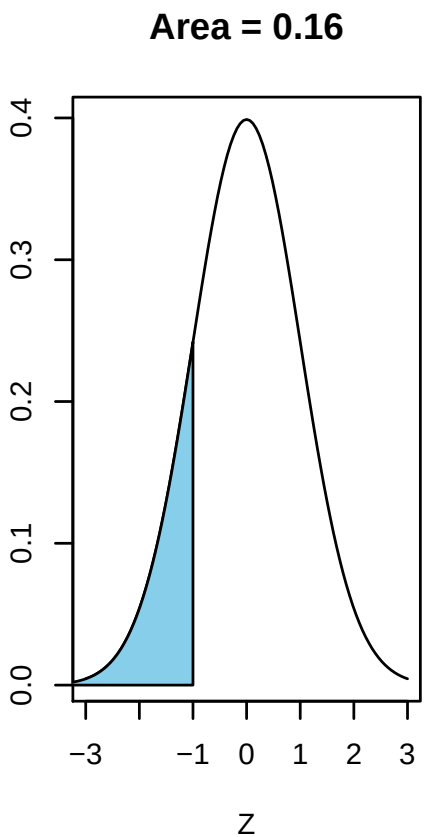


Example #2

For the standard normal distribution:
$Pr(a \leq Z \leq b) = Pr(Z \leq b) - Pr(Z \leq a)$

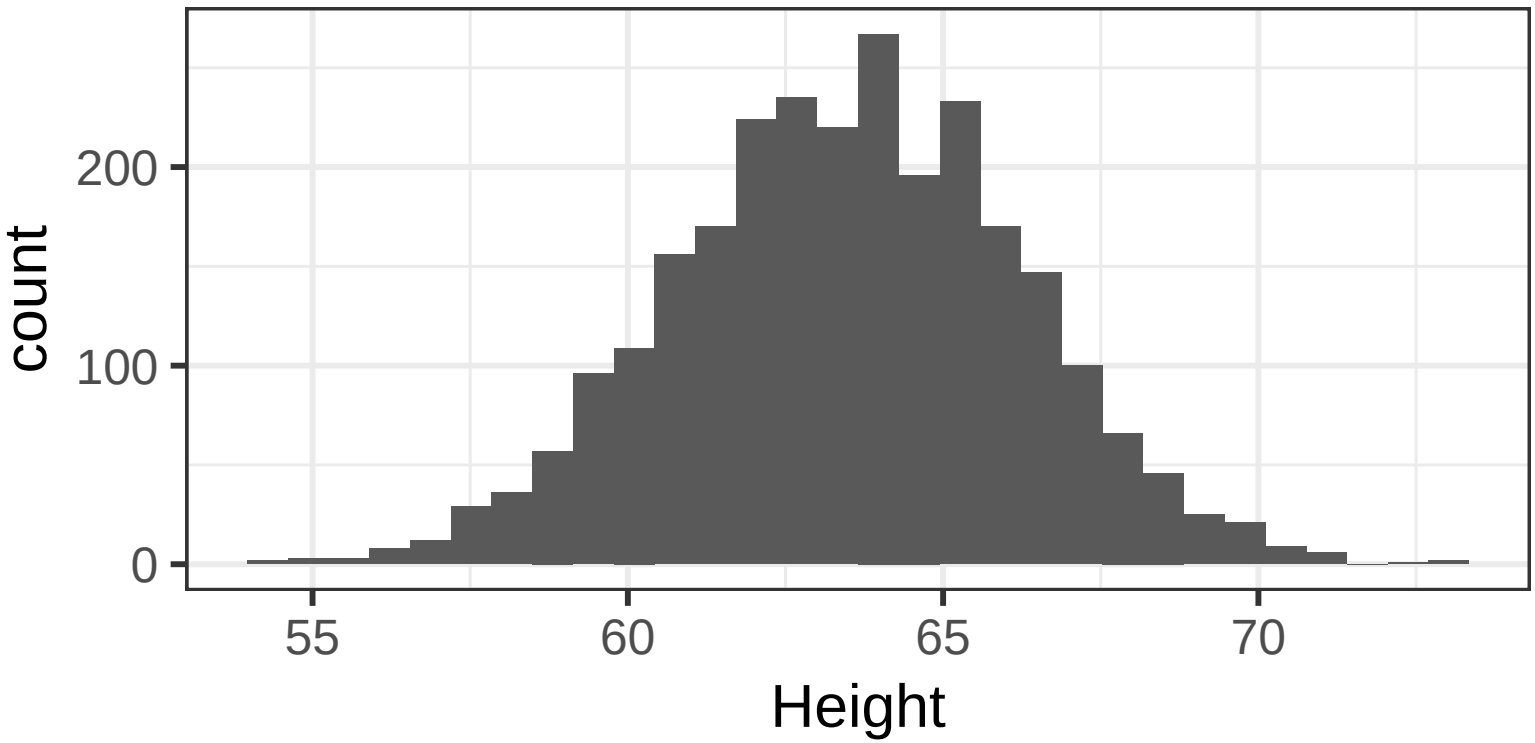






Practicing the Normal Approximation

The National Health and Nutrition Examination Survey (NHANES) collected the heights 2,649 adult women. The data have a mean of 63.5 inches and a standard deviation of 2.75 inches:



1. Estimate the percentage of women who are under 5 ft tall
2. Estimate the percentage of women between 5'3 and 5'6

Practicing Normal Approximations

- 5 ft (60 in) is 1.27 standard deviations below the mean
  - Using the standard normal distribution: $Pr(Z < -1.27) = 0.102$
  - In the actual sample, 282 of 2649 women (10.6%) were under 5 ft tall
- 5'3 (63 in) is 0.18 standard deviations below the mean, 5'6 (66 in) is 0.91 standard deviations above the mean
  - Using the standard normal distribution: $Pr(-.18 < Z < .91) = 0.819 - 0.429 = 0.390$
  - In the actual sample 1029 of 2649 women (38.8%) were between 5'3 and 5'6

The Central Limit Theorem

- You may have noticed that in many cases the sampling, bootstrap, and randomization distributions that are well-approximated by the normal distribution - This is not a coincidence
- The **Central Limit Theorem**, one of the most important results in all of statistics, establishes the normality of many common statistics given a sufficiently large sample size. These statistics include:
  - means
  - proportions
  - differences in means
  - differences in proportions
- We will get into the details for each of these statistics soon, but for now we will practice the general method of using the normal distribution to perform hypothesis tests and construct confidence intervals

Hypothesis Testing using the Normal Distribution

When the randomization distribution is symmetric and bell-shaped, it can be approximated by a normal distribution with a mean equal to the hypothesized null value and a standard deviation equal to the standard error of the randomization distribution:

$$N(\text{null value}, SE)$$

Using this approximation, the p-value can be calculated by finding the area beyond the observed statistic

Test Statistics

- Since we can convert any normal distribution to a standard normal distribution, the standard normal distribution has been widely used for hypothesis testing
- This testing procedure involves finding the *p*-value by using a standardized **test statistic**, often called the *z-statistic*:

$$z_{test} = \frac{\text{sample statistic} - \text{null value}}{SE}$$

- The *p*-value is the area of the standard normal distribution beyond this statistic
- A hypothesis test using a normal approximation is sometimes called a **z-test**

Practice

Since the 1800's, many experts have agreed that the average body temperature for healthy humans is 98.6 degrees F. However, recently there has been speculation that this may change over time. The StatKey dataset "Body Temperature" contains the body temperatures of a random sample of 50 adults taken in 1996.

1. Use StatKey to perform a two-sided randomization test assessing whether the average body temperature in 1996 is really 98.6
2. Use the SE and sample statistic to construct a z-statistic and carry out the hypothesis test using the standard normal distribution
3. Compare the p-values you calculated in 1 and 2

Confidence Intervals using the Normal Distribution

- Recall that we used the bootstrap distribution to construct 95% confidence intervals by finding the interval containing the middle 95% of bootstrap statistics
- We could do this using percentiles, but sometimes we could us the "2SE" approach:

$$\text{sample statistic} \pm 2 * SE$$

- For the appropriate statistics, given a sufficiently large sample size, the bootstrap distribution is well-approximated by the normal distribution:

$$N(\text{sample statistic}, SE)$$

- We can find $P\%$ confidence intervals using: z^*, the *critical value* that captures the middle $P\%$ of this distribution

Finding the endpoints

- So far we've used $z* = 2$ to construct 95% confidence intervals, this is actually slightly inaccurate. Common values of $z*$ and the corresponding confidence levels are given below:

Confidence Level	80%	90%	95%	99%
z^*	1.282	1.645	1.960	2.576

- Other critical values can be found in StatKey under "Theoretical Distributions" using "Two-Tail"

Practice

For the "Body Temperature" Data, use StatKey to:

1. Find 90% and 98% percentile bootstrap confidence intervals for the population mean
2. Use a normal approximation of the bootstrap distribution to find 90% and 98% confidence intervals for the population mean
3. Compare your results for 1 and 2

Summary

So far we've seen two very straightforward formulas that can be applied when the randomization or bootstrap distribution of a sample statistic is approximately normal.

Confidence Intervals:

$$\text{sample statistic} \pm z^* * SE$$

where z^* is chosen from the standard normal distribution based upon the desired confidence level

Hypothesis Tests:

$$z_{test} = \frac{\text{sample statistic} - \text{null value}}{SE}$$

where the p-value is then found by locating z_{test} on the standard normal distribution

Summary

- Each of these approximations assumes that we know SE, the standard error of the sample statistic
- We've already learned that we can find SE using bootstrapping or randomization
- Next we will learn how to estimate SE for various types of data without having to go to the trouble of simulating thousands of bootstrap or randomization samples

Conclusion

Right now you should...

1. Feel comfortable find areas using the standard normal distribution
2. Be able to conduct hypothesis tests using a normal approximation when you're given the standard error
3. Be able to construct confidence intervals using a normal approximation when you're given the standard error

These notes cover Sections 5.1 and 5.2 of the textbook, I encourage you to read through those sections and their examples