# Approximating with a Distribution

Javier E. Flores

March 6, 2019

# "Dad" Joke of the Day

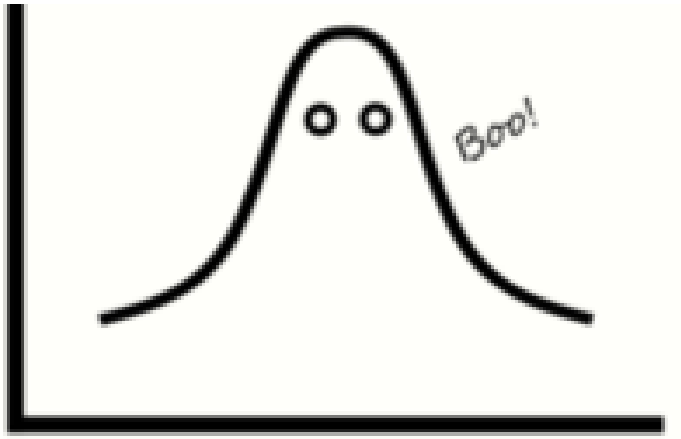
You've heard of sampling distributions,

We've talked about bootstrap distributions,

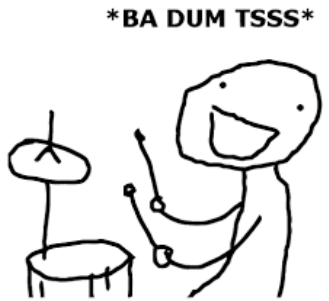
and we've even learned about randomization distributions...

# "Dad" Joke of the Day

But what do we call this distribution?

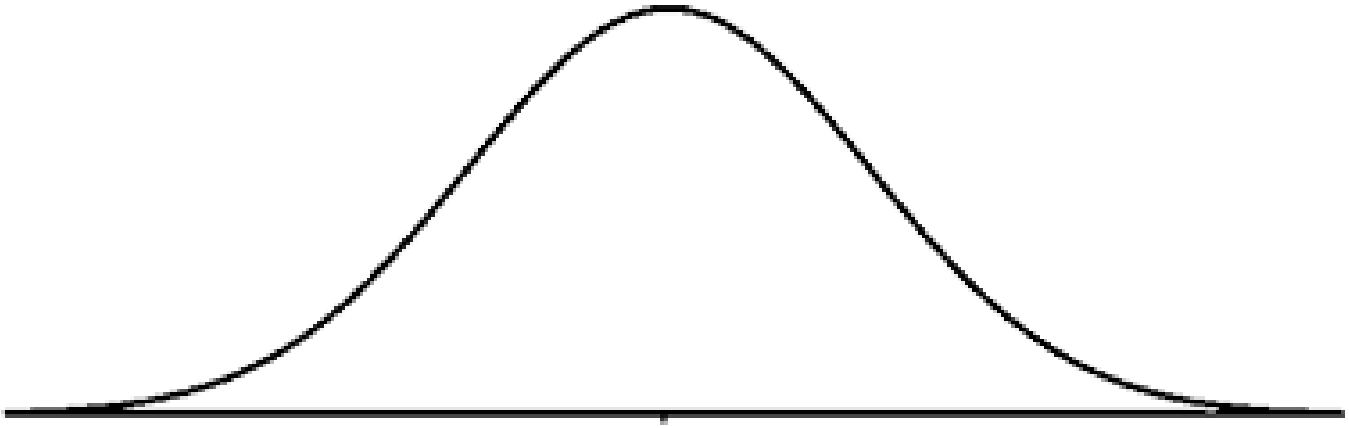



That's right, it's the (para)normal distribution!



# Normal Distribution

All jokes aside, during this lecture we will discuss the **normal distribution** and its importance to statistics.

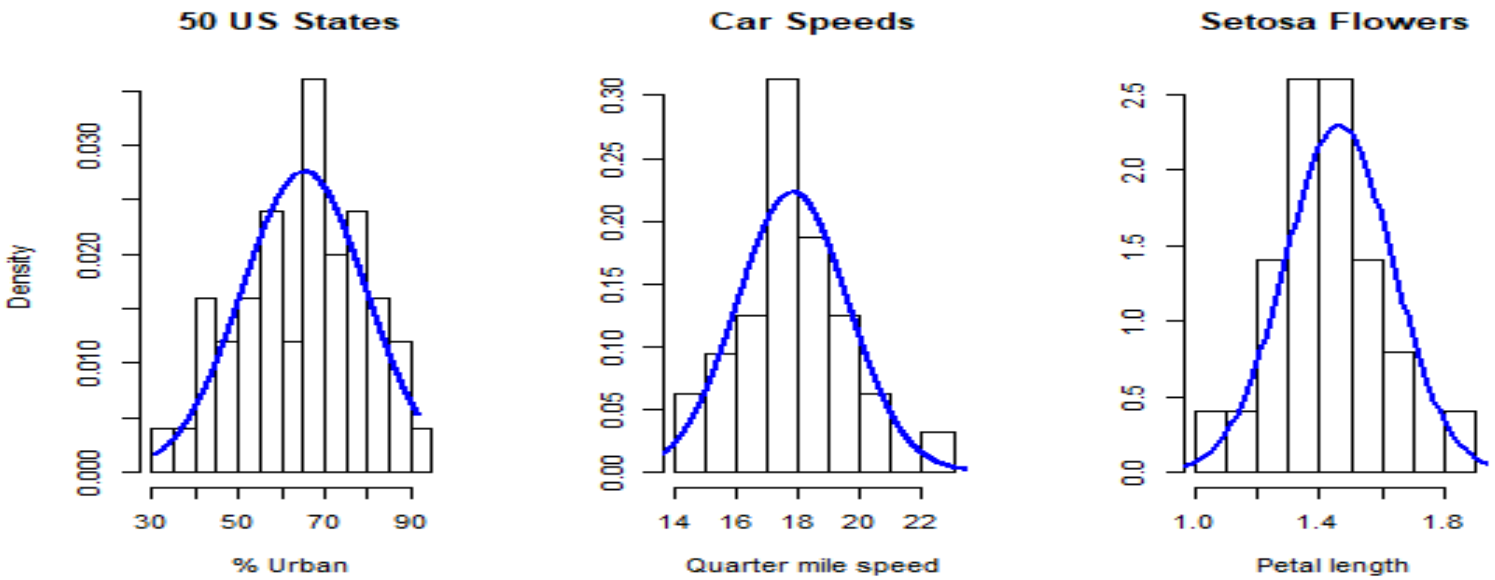


# Normal Distribution

One of the most remarkable facts about this distribution is its prevalence in the real world.

There are endless examples of variables whose distributions are approximately normal.

The variables below are just a few of these:



# Normal Distribution

While each of these examples may be approximated by a normal distribution, the approximating normal distribution isn't the same for each.

Each normal distribution is centered at a different value and has a different width.

This implies that each distribution is characterized by a specific mean (center, $\mu$) and standard deviation ("width", $\sigma$).

We can see this by looking at the (ugly) formula that defines the normal curve:

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

There is a clear dependence on both $\mu$ and $\sigma$.

(You aren't expected to know or memorize this formula!)

# Standard Normal Distribution

Despite the fact that each normal distribution may be characterized by a different mean ($\mu$) and standard deviation ($\sigma$), the data giving rise to the distribution can always be standardized.

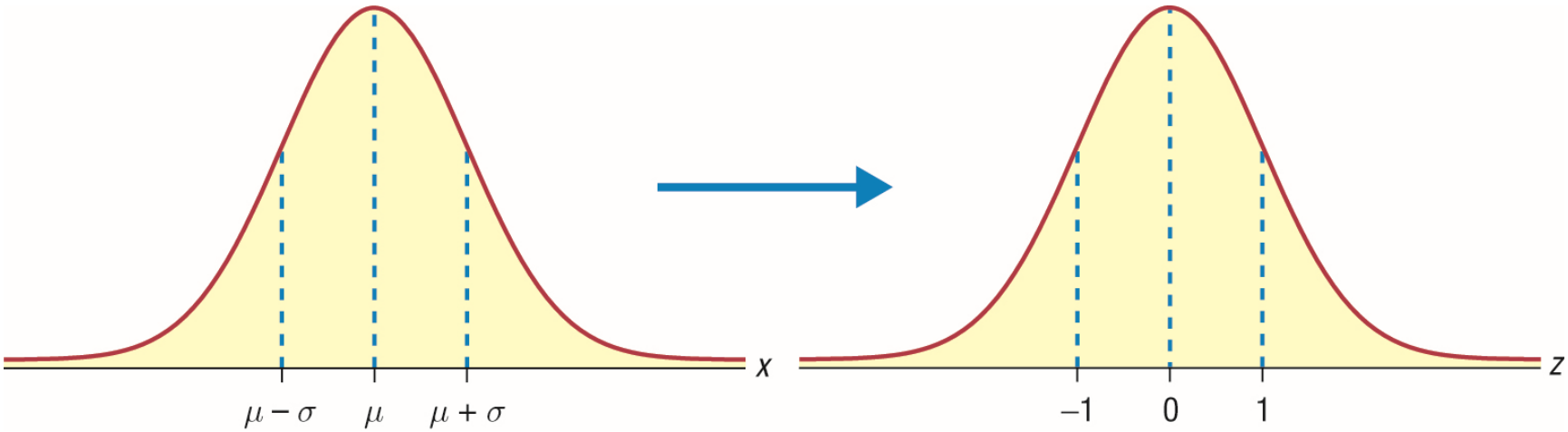
In doing so, we obtain the **standard normal distribution**.

The standard normal distribution is *parameterized* by $\mu = 0$ and $\sigma = 1$. (Does this seem familiar?)

Recall that z-scores are the way that we standardize variables. We know that z-scores have a mean of 0 and standard deviation of 1.

The standard normal distribution is just the distribution of the z-scores of a normally distributed variable!

# Standard Normal Distribution

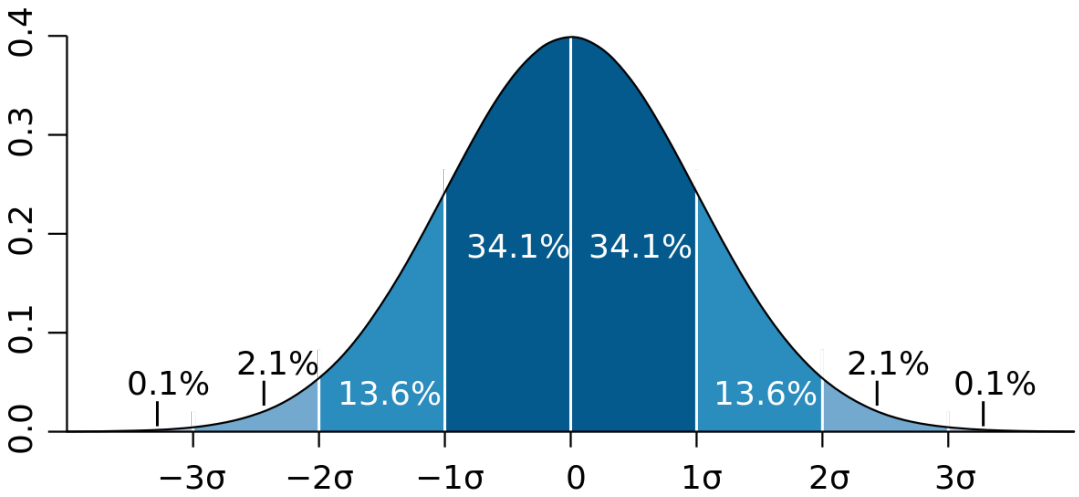



The units of the standard normal distribution (right figure) are in terms of the z-scores.

Oftentimes it is more convenient to work with the standard normal distribution than with the "non-standard" normal distribution.

# Probabilities

One example in which working with the standard normal distribution is more convenient is when computing probabilities.



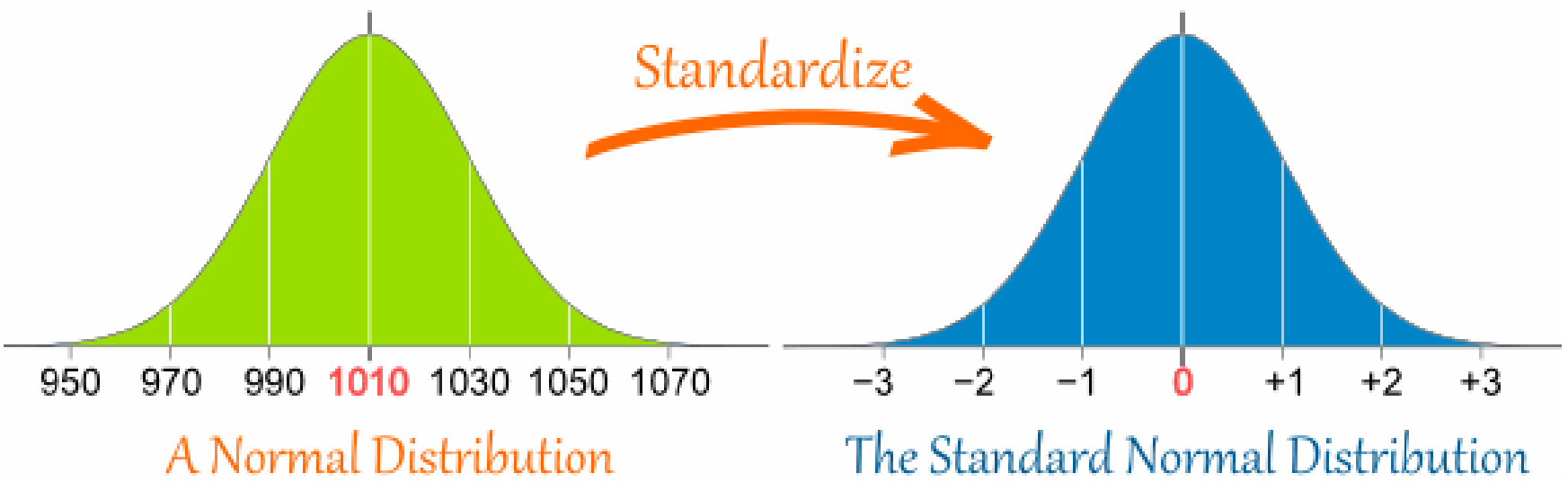

As indicated by the figure above, areas under portions of the standard normal curve are known. These areas correspond to certain probabilities.

For example, the area between 0 and 1 is 0.341 (i.e. 34.1%) so $Pr(0 < Z < 1) = 0.341$.

# Probabilities

Assuming we didn't have the standardized distribution below and to the right, we wouldn't be able to easily find $Pr(1010 < X < 1030)$ from the figure on the left.

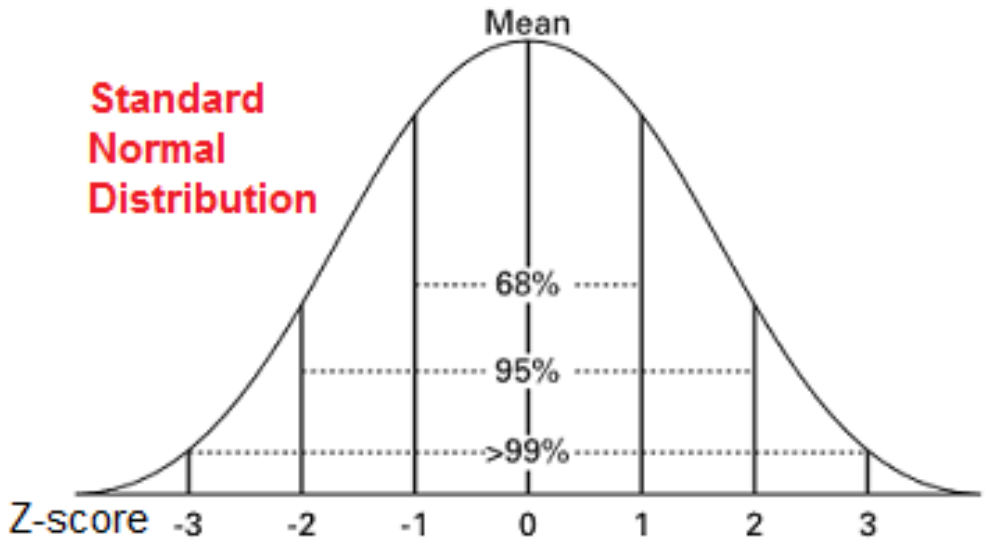
Only after standardizing would we see that
$Pr(1010 < X < 1030) = Pr(0 < Z < 1) = 0.341$.



# Probabilities

The following are three probability/z-score cutoffs you should commit to memory.

There is a 68% probability for values to fall between -1 and 1; a 95% probability between -2 and 2; and a 99% probability for values to fall between -3 and 3.



## Probabilities

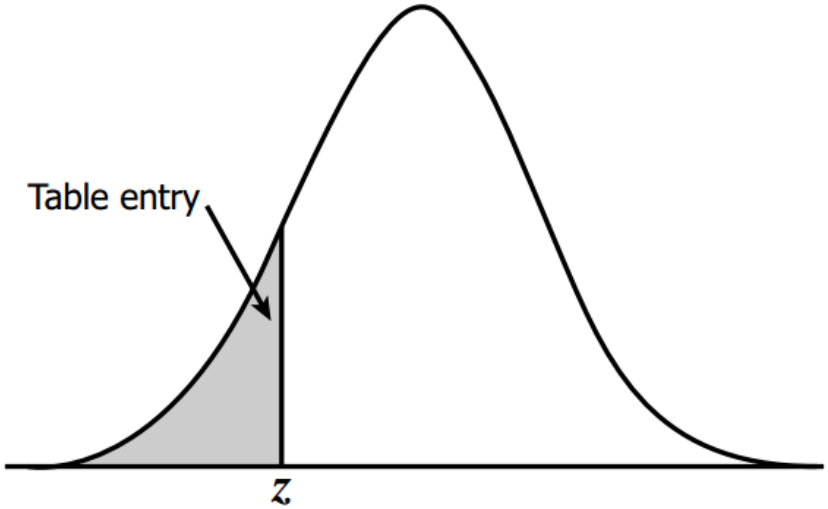
If we wanted to compute probabilities for some other interval, say $Pr(0.2 < Z < 1.8)$, we would need to use calculus (integration) in order to find the area under the curve.

As it would turn out, there is not a closed form solution to the integral of the normal curve and so statisticians have relied on numerical methods to calculate these probabilities.

The results of these methods have been aggregated into tables which allow the calculation of specific probabilities without a computer.

Given that that the year is 2019 and computers are abundant, we don't have to necessarily use these tables. We can just use tools like Minitab to compute probabilities of interest for us.

# Probabilities



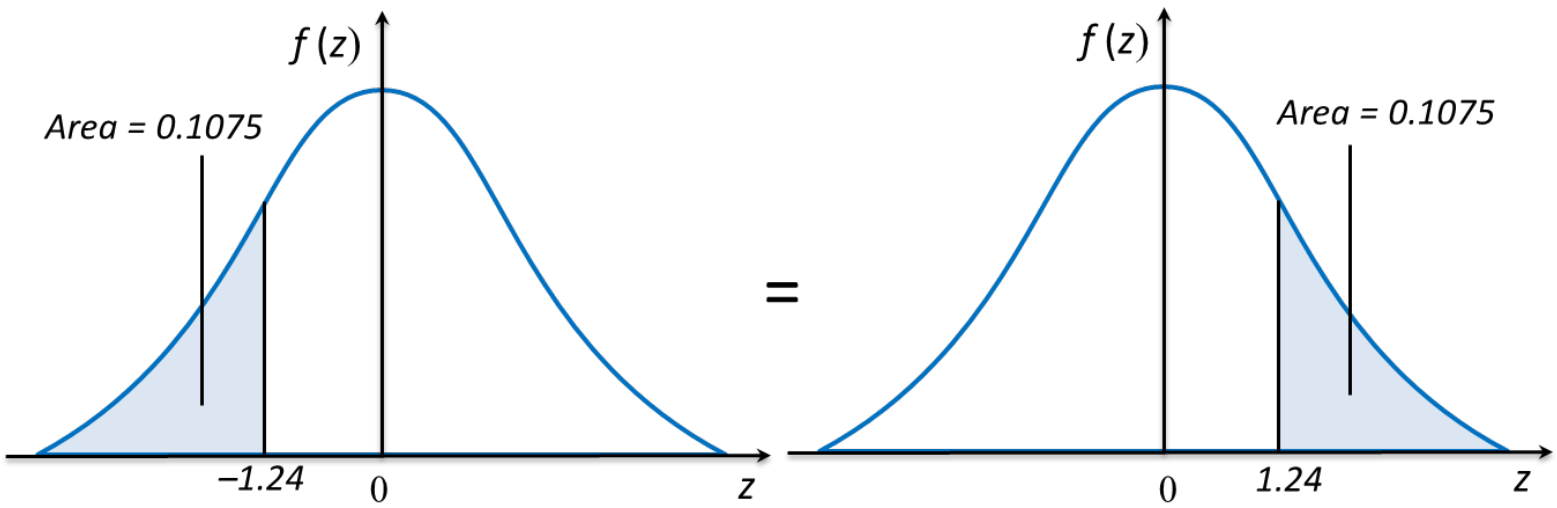

Historically, probability tables contain the area under the curve to the left of some value, $z$.

In Minitab, we can obtain this probability by selecting **Calc** -> **Probability Distributions** -> **Normal** and choosing **Cumulative Probability** and **Input Constant**.

Because this yields only left tail areas, it is important to understand how to leverage symmetry and some basic probability rules.

# Rule #1

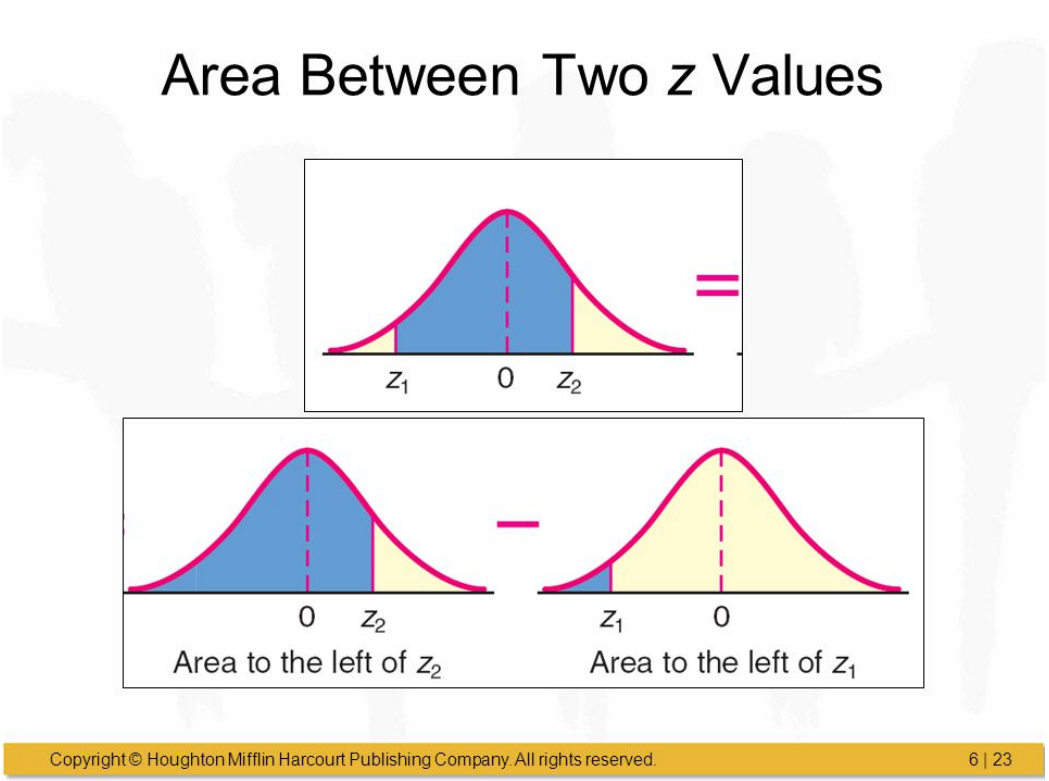
For the standard normal distribution,
$Pr(Z \geq z_1) = Pr(Z \leq -z_1)$.




From the figure above,
$Pr(Z \geq 1.24) = 0.1075 = Pr(Z \leq -1.24)$.

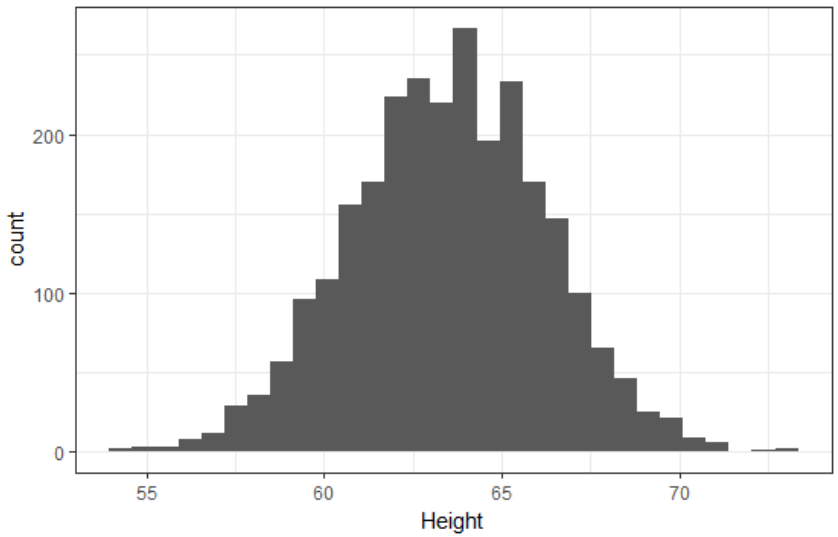
# Rule #2

For the standard normal distribution,
$Pr(z_1 \leq Z \leq z_2) = Pr(Z \leq z_2) - Pr(Z \leq z_1)$.



# Practice

The National Health and Nutrition Examination Survey (NHANES) collected the heights of 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 (60 in)
2) Estimate the percentage of women who are between 5'3 and 5'6
3) Estimate the percentage of women who are over 6 ft (72 in)

# Solution

We first compute the z-score for 5 ft to obtain -1.27. Using the standard normal distribution, we find that $Pr(Z < -1.27) = 0.102$.

In the actual sample, 282 out of 2649 women (10.6%) were under 5 ft.

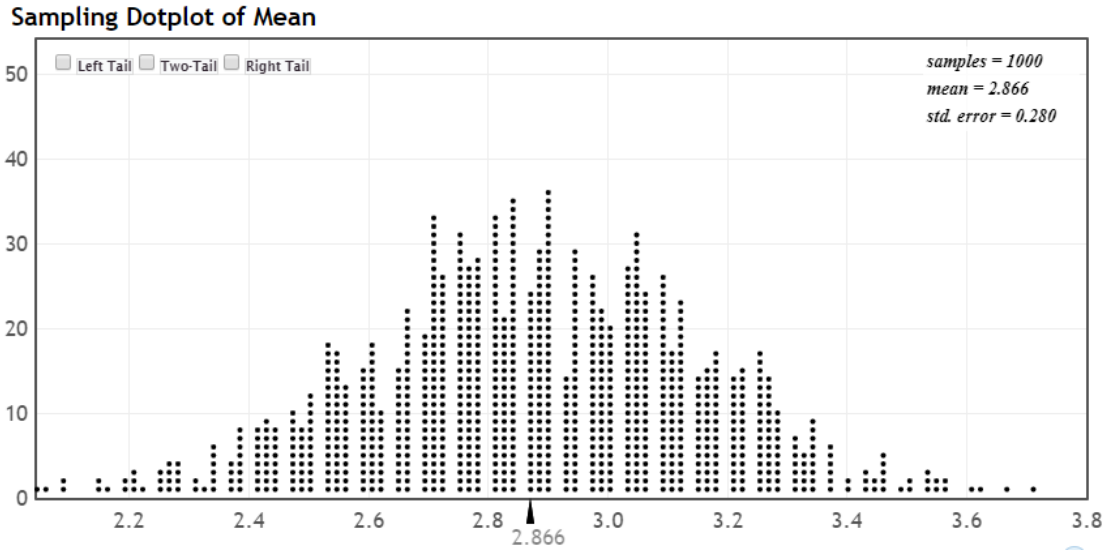
The z-score for each height is -0.18 and 0.91. Using the standard normal distribution, we find that $Pr(-0.18 < Z < 0.91) = Pr(Z < 0.91) - Pr(Z < -0.18) = 0.819 - 0.429 = 0.390$.

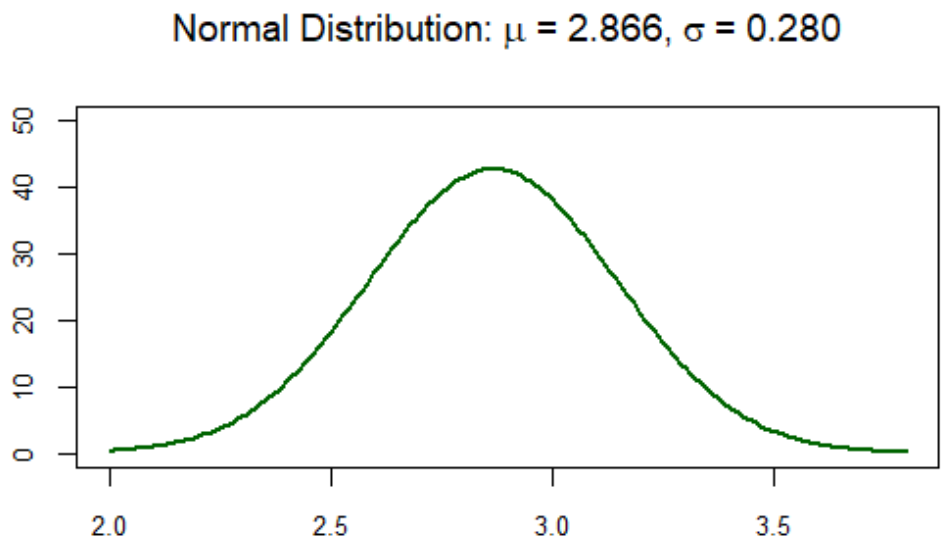
In the actual sample, 1029 out of 2649 women (38.8%) were between 5'3 and 5'6.

Finally, the z-score here is 3.09. Using the standard normal distribution, we find that $Pr(Z > 3.09) = Pr(Z < -3.09) = 0.001$.

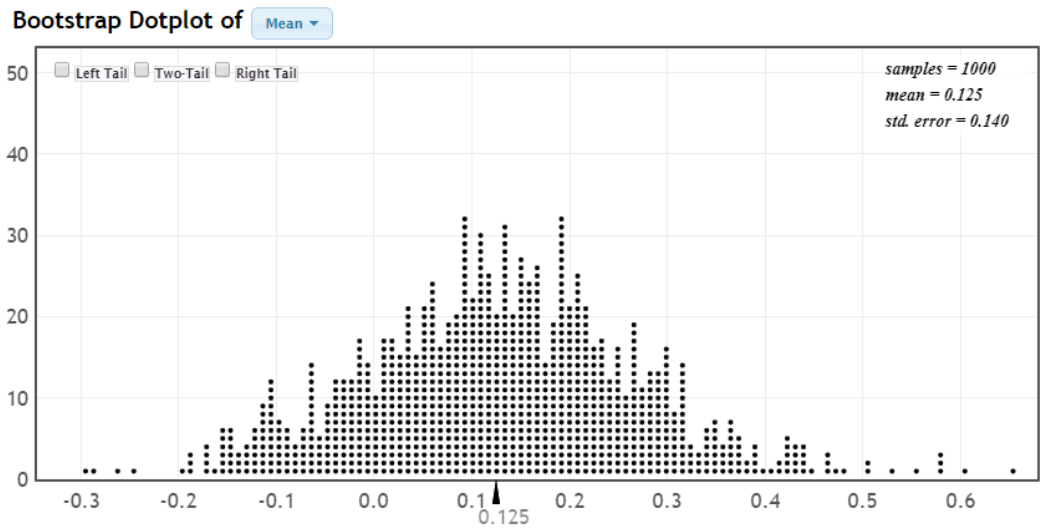
In the actual sample, 3 out of 2649 women (0.1%) were over 6 ft.

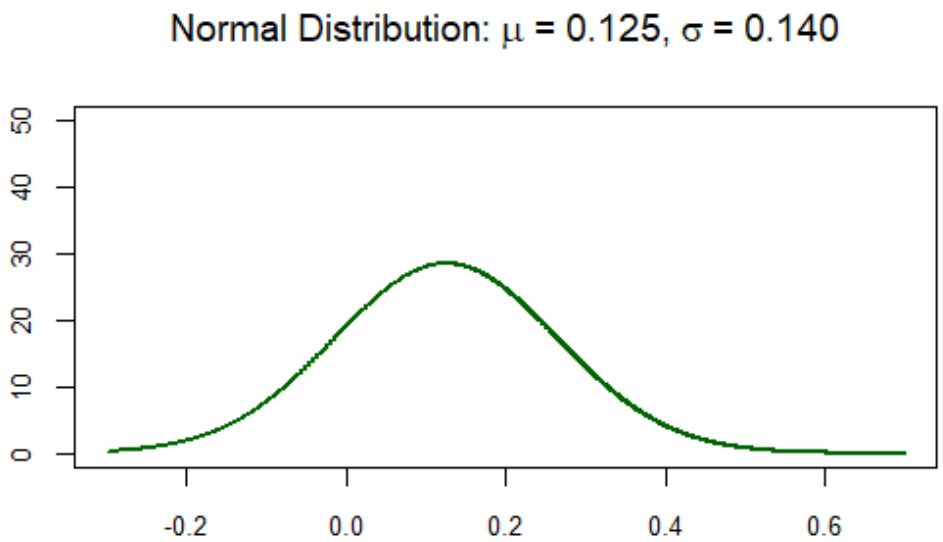
# Sampling Distribution



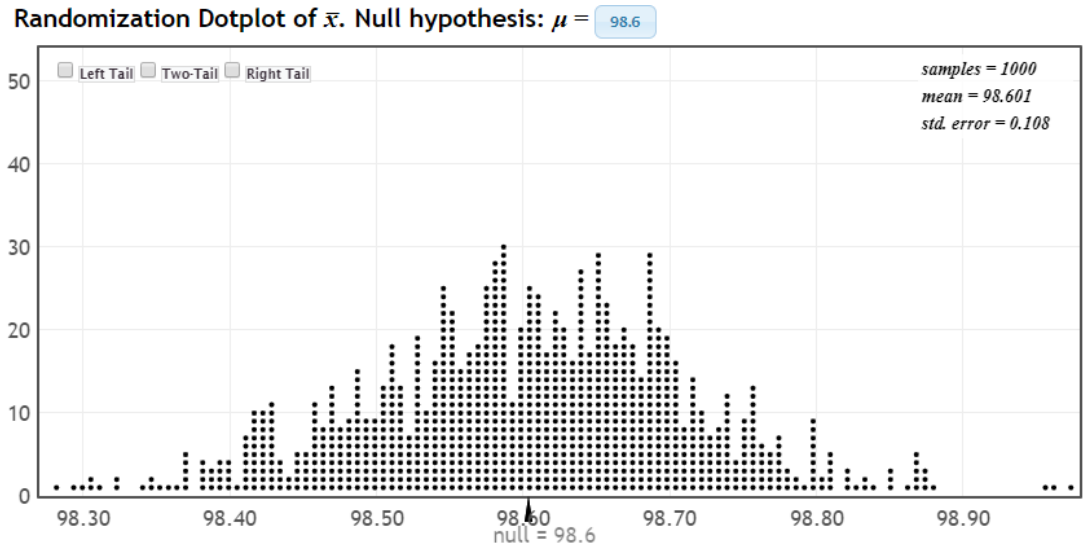



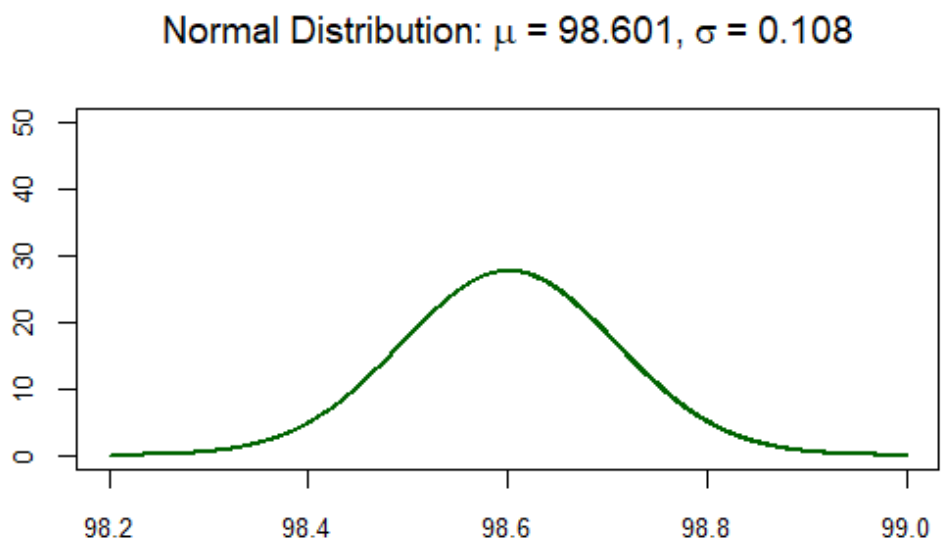
# Bootstrap Distribution



# Randomization Distribution



# Randomization Distribution

Clearly, each of the previous distributions - sampling, bootstrap, and randomization - can be approximated using some normal distribution.

This is not unique to the examples chosen, but rather a consequence of one of the most important results in all of statistics: the **Central Limit Theorem (CLT)**.

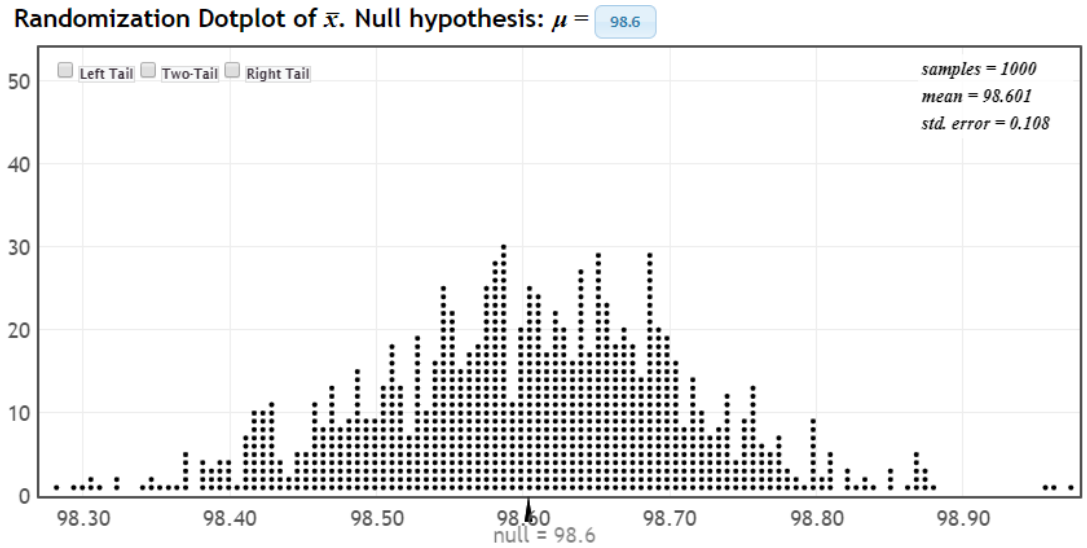
Given a "sufficiently large" sample size, the CLT establishes the normality of many common statistics. These include:

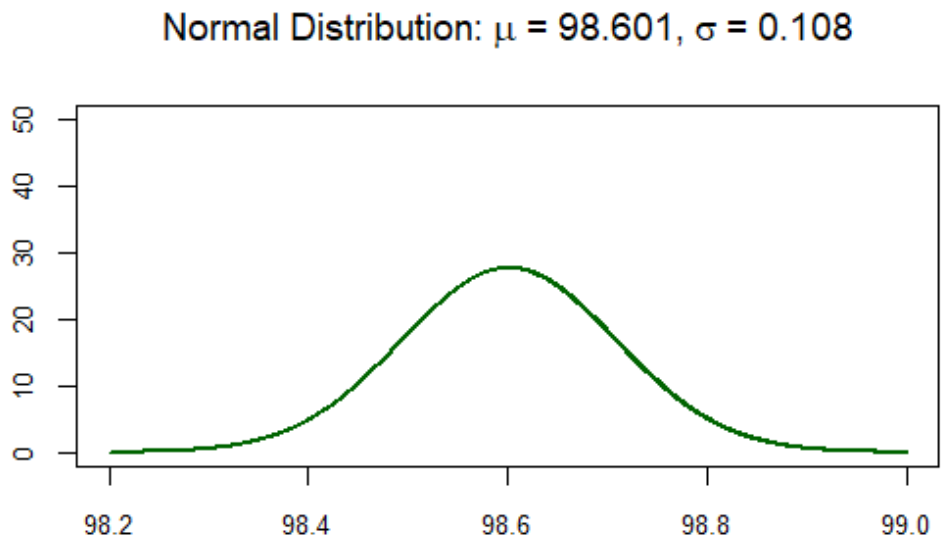
- means
- proportions
- differences in means
- differences in proportions

In the coming chapters we'll discuss the estimation and testing methods which rely on CLT-established normality for each of these statistics.

For now, we'll discuss the general approach of using the normal distribution to perform SHT and to construct confidence intervals.

# SHT using the Normal Distribution



# SHT using the Normal Distribution

The previous example shows us that, when the randomization distribution is symmetric and bell-shaped, it can be approximated by a normal distribution.

The approximating distribution should have a mean equal to the hypothesized null value (98.6 in the previous example) and a standard deviation equal to the standard error of the randomization distribution (0.108 in the previous example).

Using this approximation, we can apply the probability tools we learned for normal distributions in order to calculate p-values.

The p-values are calculated by finding the area beyond the observed sample statistic.

# SHT using the Normal Distribution

Since we learned that working with a standard normal distribution is often more convenient, we often work with the **z-statistic**, or standardized test statistic.

The z-statistic is simply the z-score of your sample statistic under the assumed null distribution:

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

Using the z-statistic as opposed to the sample statistic allows us to use the standard normal distribution to compute the p-value.

A hypothesis test using a normal approximation is sometimes called a **z-test**.

# Practice

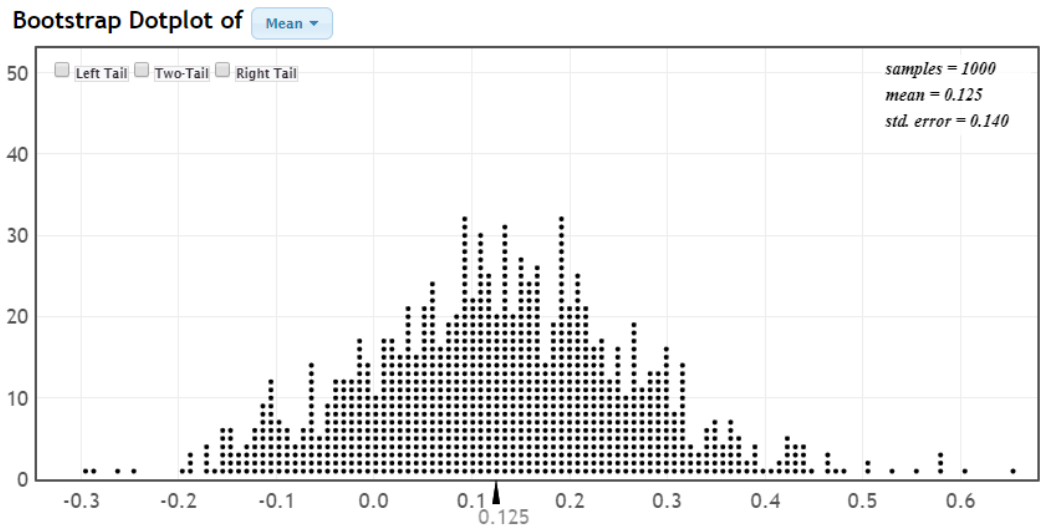
It is commonly accepted that the average healthy human has a body temperature of 98.6 degrees Fahrenheit. However, recently there has been speculation that this temperature may change over time. Certain lifestyle choices and even simply aging has been shown to affect normal body temperature. The StatKey dataset "Body Temperature" contains the body temperatures of a random sample of 50 adults taken in 1996.

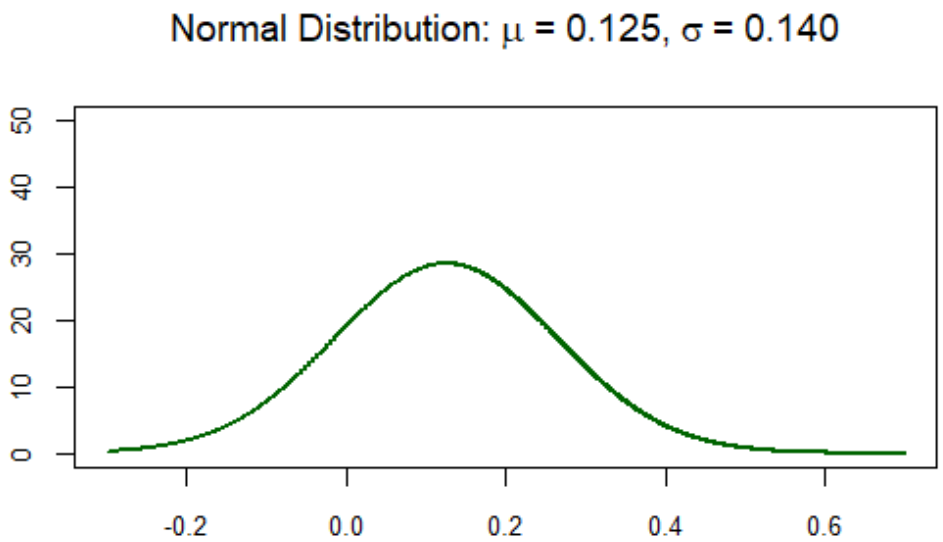
1) Perform a two-sided randomization test assessing whether the average body temperature in 1996 is 98.6.

2) Use the SE of the randomization distribution from 1) and your sample statistic to construct a z-statistic and perform a z-test.

3) Compare the p-values obtained in both approaches.

# CI's using the Normal Distribution



# CI's using the Normal Distribution

The previous example shows us that, when the bootstrap distribution is symmetric and bell-shaped, it can be approximated by a normal distribution.

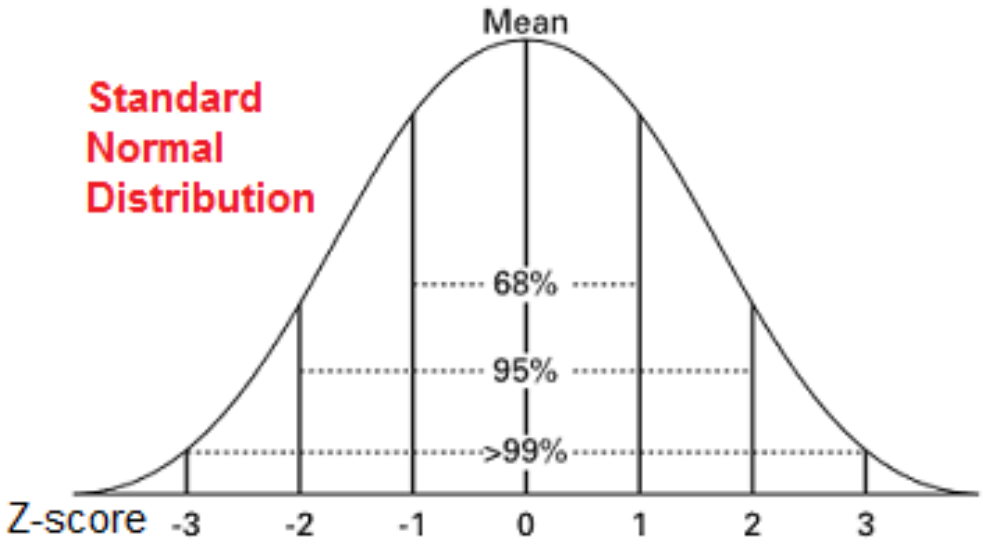
The approximating distribution should have a mean equal to the original sample statistic (0.125 in the previous example) and a standard deviation equal to the standard error of the bootstrap distribution (0.140 in the previous example).

Rather than perform hypothesis tests with this approximating distribution, we can construct confidence intervals.

# CI's using the Normal Distribution

The "SE approach" to constructing bootstrap confidence intervals is actually based on a normal approximation.

Recall that, for a standard normal distribution, we have that 95% of values fall between 2 standard deviations of the mean:



# CI's using the Normal Distribution

For *bell-shaped* and *symmetric* bootstrap distributions, which can be approximated by a normal distribution, you'll recall that the "SE approach" stated a 95% confidence interval is found by computing:

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

This is based on the idea that, like the approximating normal distribution, 95% of the values of our bootstrap distribution should fall between 2SE (i.e. 2 standard deviations) of the mean.

# CI's using the Normal Distribution

Using strictly the normal approximation of a bootstrap distribution, we can compute confidence intervals at various confidence levels (not just 95%).

In general, we can find the $P\%$ confidence interval by using $z_{crit}$, the **critical value** that captures the middle $P\%$ of the approximating normal distribution.

Commonly used values of $z_{crit}$ and their corresponding confidence levels are provided in the table below:

| **Confidence Level** | 80% | 90% | 95% | 99% |
|---|---|---|---|---|
| $z_{crit}$ | 1.282 | 1.645 | 1.960 | 2.576 |

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

## Practice

Using the "Body Temperature" Data in StatKey:

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 the the two sets of intervals.

# Summary

**Confidence Intervals**:

$$\text{sample statistic} \pm z_{crit} * SE$$

where $z_{crit}$ is chosen from the standard normal distribution based upon the desired confidence level.

**Hypothesis Testing**:

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

where the p-value is found by calculating area(s) defined by $z_{test}$ on the standard normal distribution.

# A Peek at What's Next...

In each of the methods learned today - confidence interval construction and hypothesis testing using normal distributions - we've assumed knowledge of SE.

In the past, we've learned how we can *estimate* the SE using bootstrapping or randomization.

In the coming lectures, we'll learn other methods of estimating the standard error for various types of data. These methods will not rely on simulating thousands of samples as is the case for randomization and bootstrap-based methods.

# Wrap-Up

Right now, you should...

- Know the basic probability rules and properties of the standard normal distribution
- Be comfortable applying these properties to find areas under the standard normal curve
- Conduct z-tests and construct confidence intervals using a normal approximation

These notes cover sections 5.1 and 5.2 of the textbook. Please read through the section and its examples along with any links provided in this lecture.