# INTRODUCTION TO DATA SCIENCE

**JC and EG**

Special Thanks to Dr. Arlo Clark-Foos of the University of Michigan for some of today's slides!

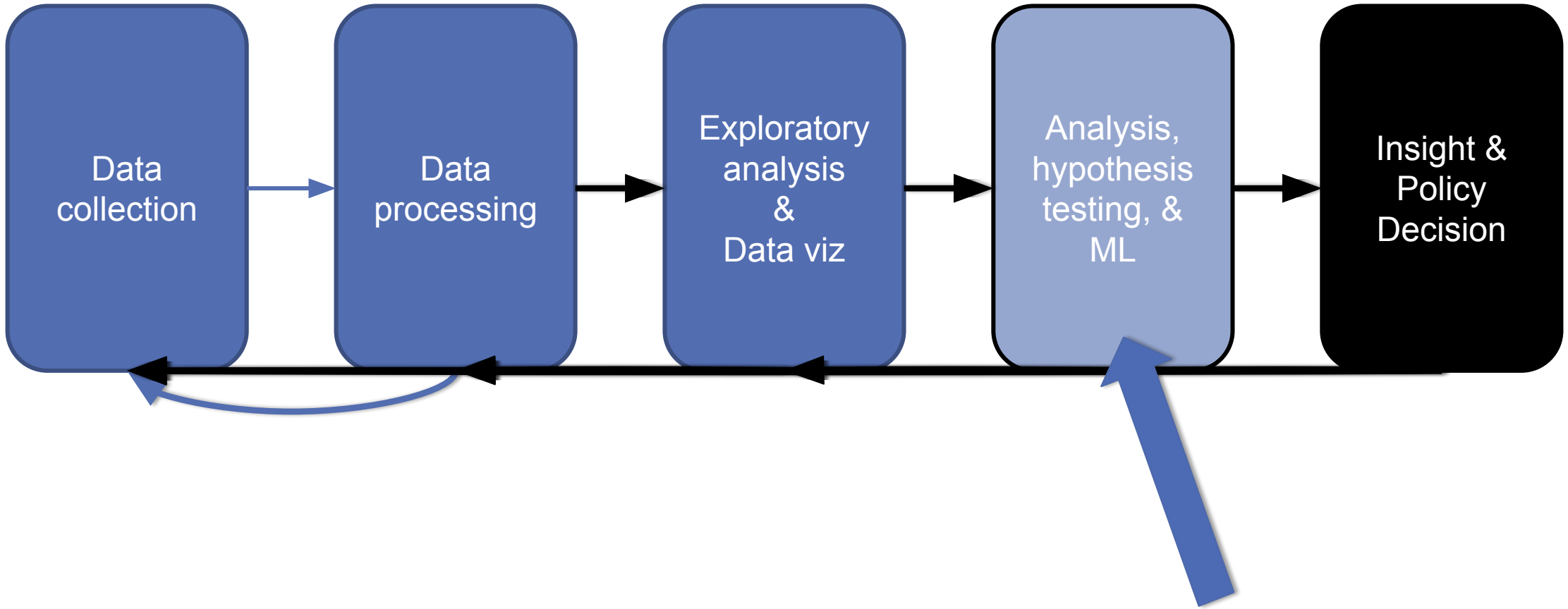
**Lecture #17 – 03/27/2022**

**CMSC320**
**Mondays & Wednesdays**
**3:30-4:45**

**https://cmsc320.github.io/**

COMPUTER SCIENCE
UNIVERSITY OF MARYLAND

# THE DATA LIFECYCLE



# TODAY'S LECTURE

**Hypothesis Testing**

- What is it?
- What are we looking to prove?
- **How can we use it to justify our hypothesis?**

**z-Score**

**t-Score**

# Hypothesis Testing

It's easy to maker claims and hypothesize, but how can we test the likelihood that the hypothesis is true?

$H_0$- Null Hypothesis that represents the default position, status quo

$H_1$- Alternative Hypothesis that we are comparing to

p-Value: Compute the probability that $H_0$ is true, that we will see a value at least as extreme as those we observed

# p-values

Instead of looking at our data based on some probability cutoff– can we compute the probability assuming that $H_0$ is true?

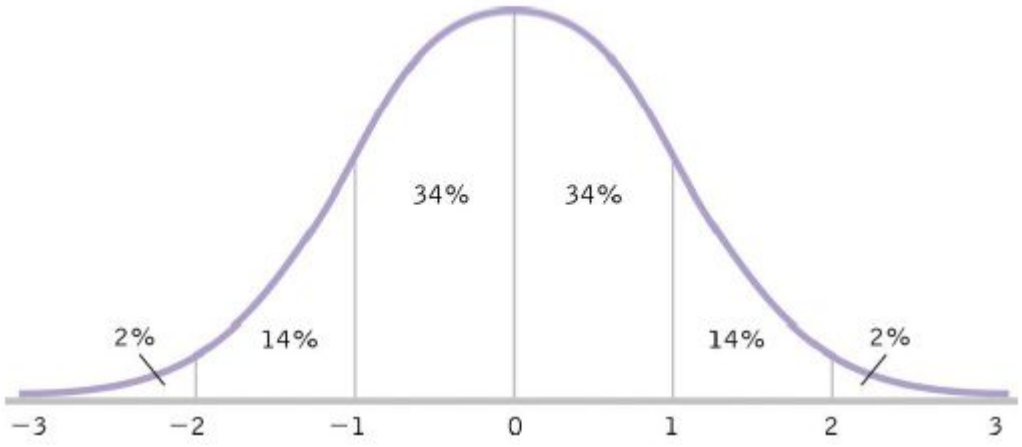
Calculated by adding together the probability that random chance generated the data, something else that is equals and something else that is rarer. For normally distributed values, use a density function.

Smaller the p-value the stronger the evidence to reject the null hypothesis.

Typically our p-value cutoff will be 5%

# Review: Standardization

**Allows us to easily see how one score (or sample) compares with all other scores (or a population).**



# CDC Example

**Jessica is 15 years old and 66.41 in. tall**

**For 15 year old girls, μ = 63.8, σ = 2.66**

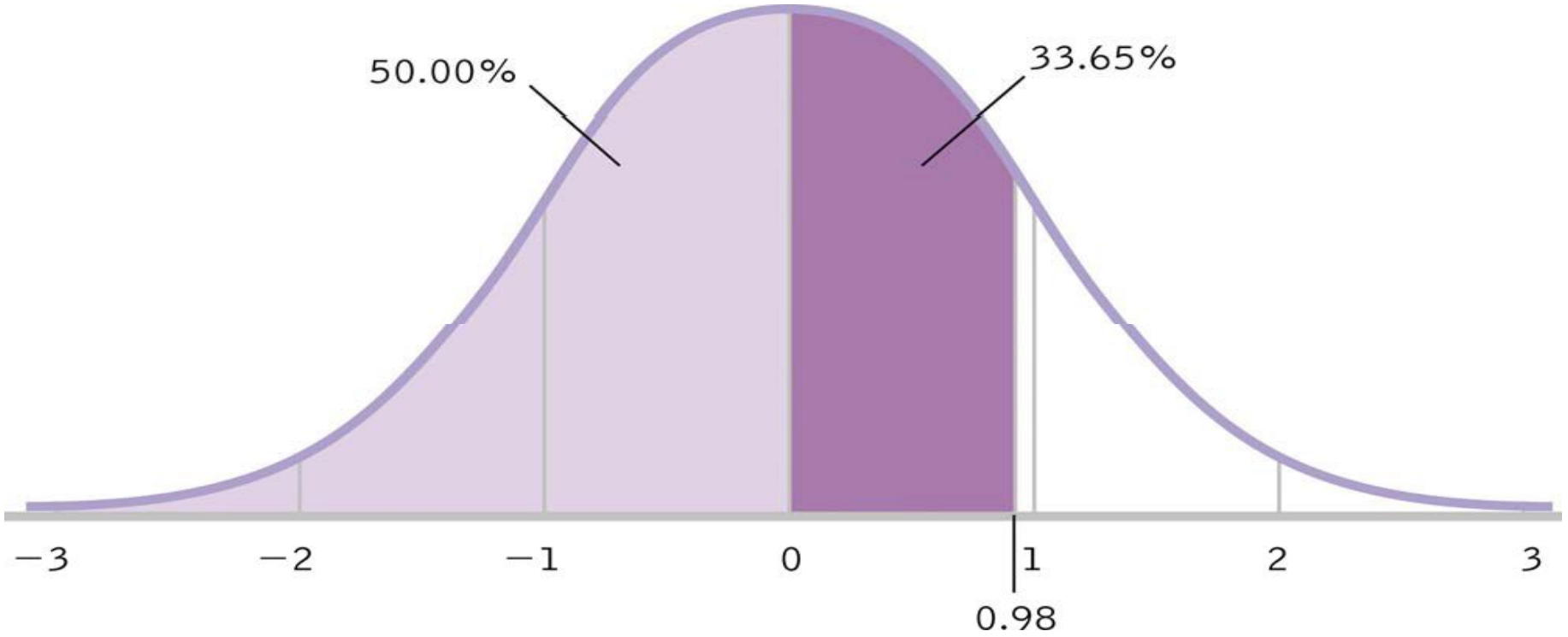
$$z = \frac{(X - \mu)}{\sigma} = \frac{(66.41 - 63.8)}{2.66} = 0.98$$

| z | % BETWEEN MEAN AND z |
|---|---|
| . | . |
| . | . |
| . | . |
| 0.97 | 33.40 |
| 0.98 | 33.65 |
| 0.99 | 33.89 |
| 1.00 | 34.13 |
| 1.01 | 34.38 |
| 1.02 | 34.61 |

# CDC Example: Jessica

**1. Percentile: How many 15 year old girls are shorter than Jessica?**

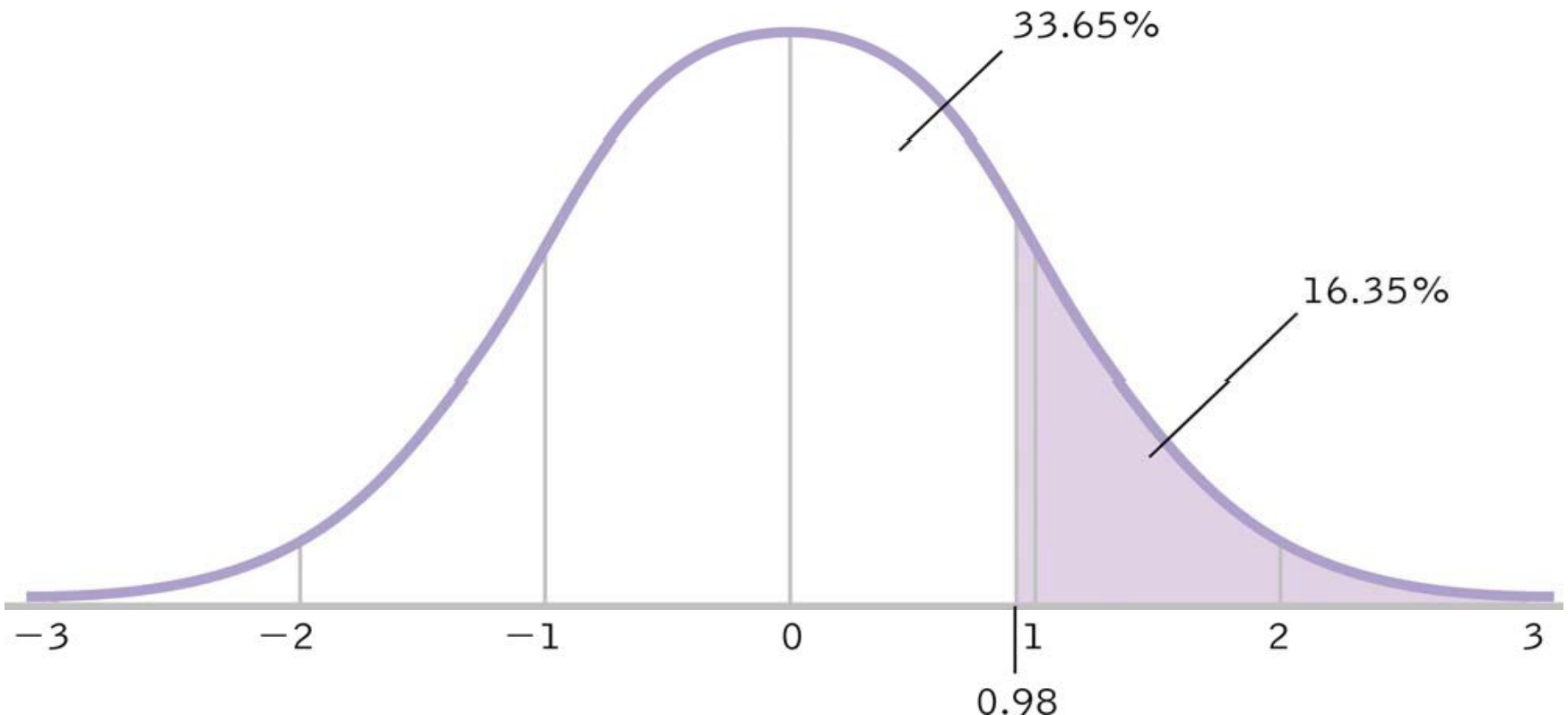
**50% + 33.65% = 83.65%**



# CDC Example: Jessica

**2. What percentage of 15 year old girls are taller than Jessica?**

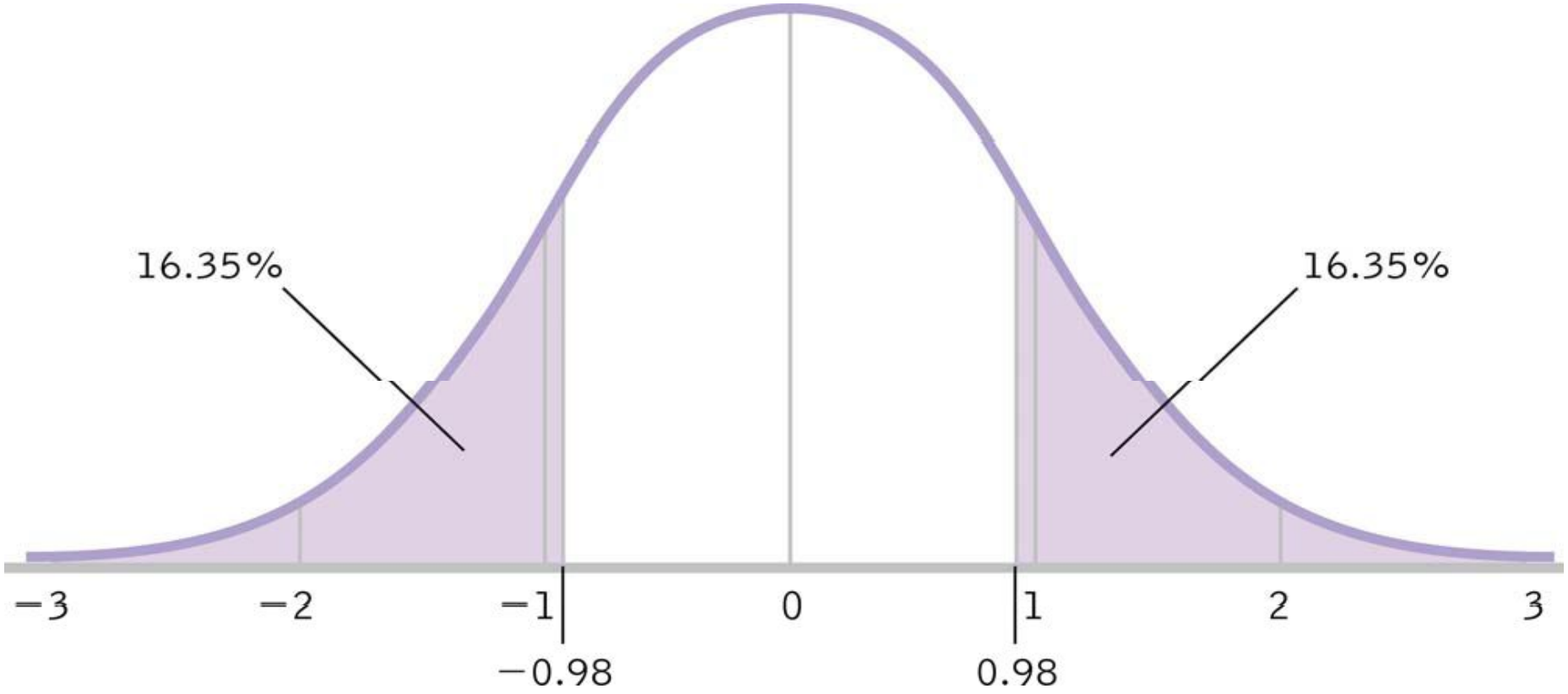
**50% - 33.65% OR 100% - 83.65% = 16.35%**



# CDC Example: Jessica

**3. What percentage of 15 year old girls are as far from the mean as Jessica (tall or short)?**

**16.35 % + 16.35% = 32.7%**



# CDC Example: Manuel

**Manuel is 15 years old and 61.2 in. tall**

**For 15 year old boys, μ = 67, σ = 3.19**

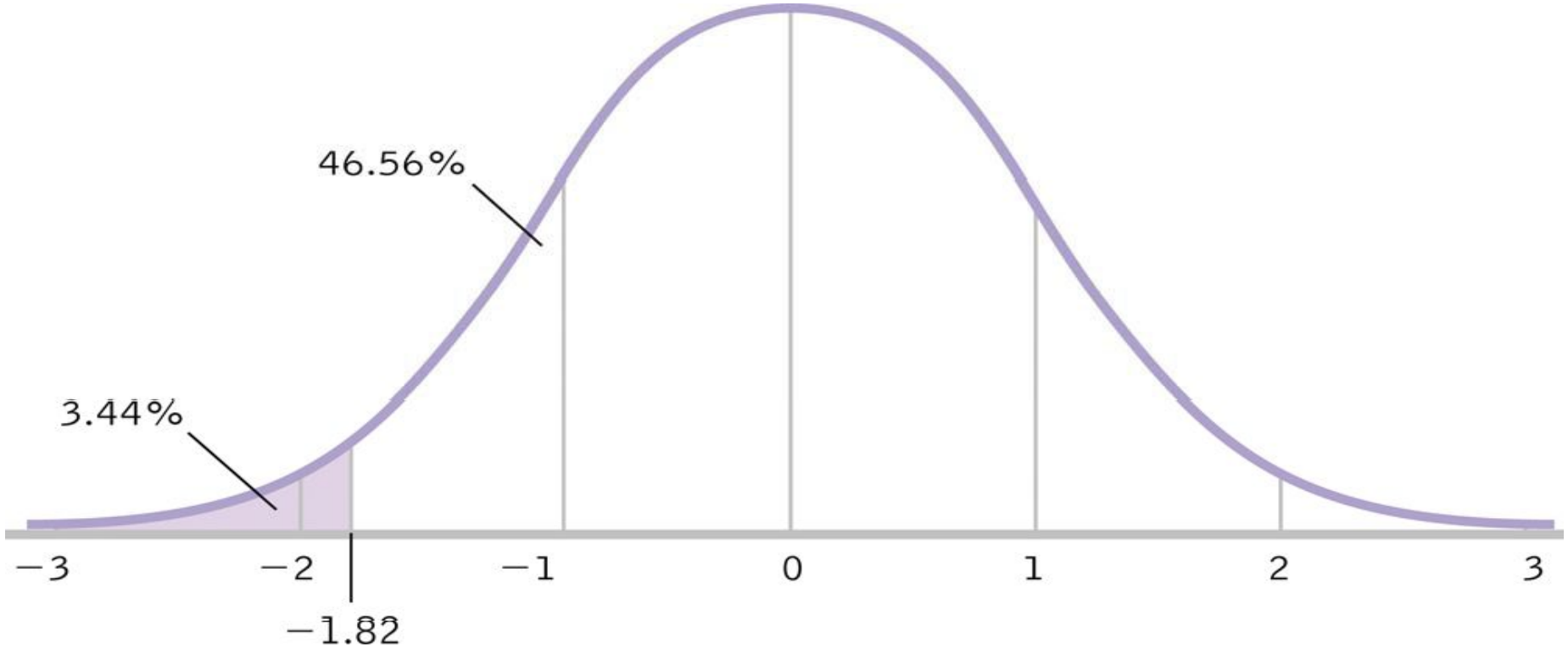
$$z = \frac{(X - \mu)}{\sigma} = \frac{(61.2 - 67)}{3.19} = -1.82$$

**Consult z table for 1.82 -> 46.56%**

# CDC Example: Manuel

**1. Percentile**

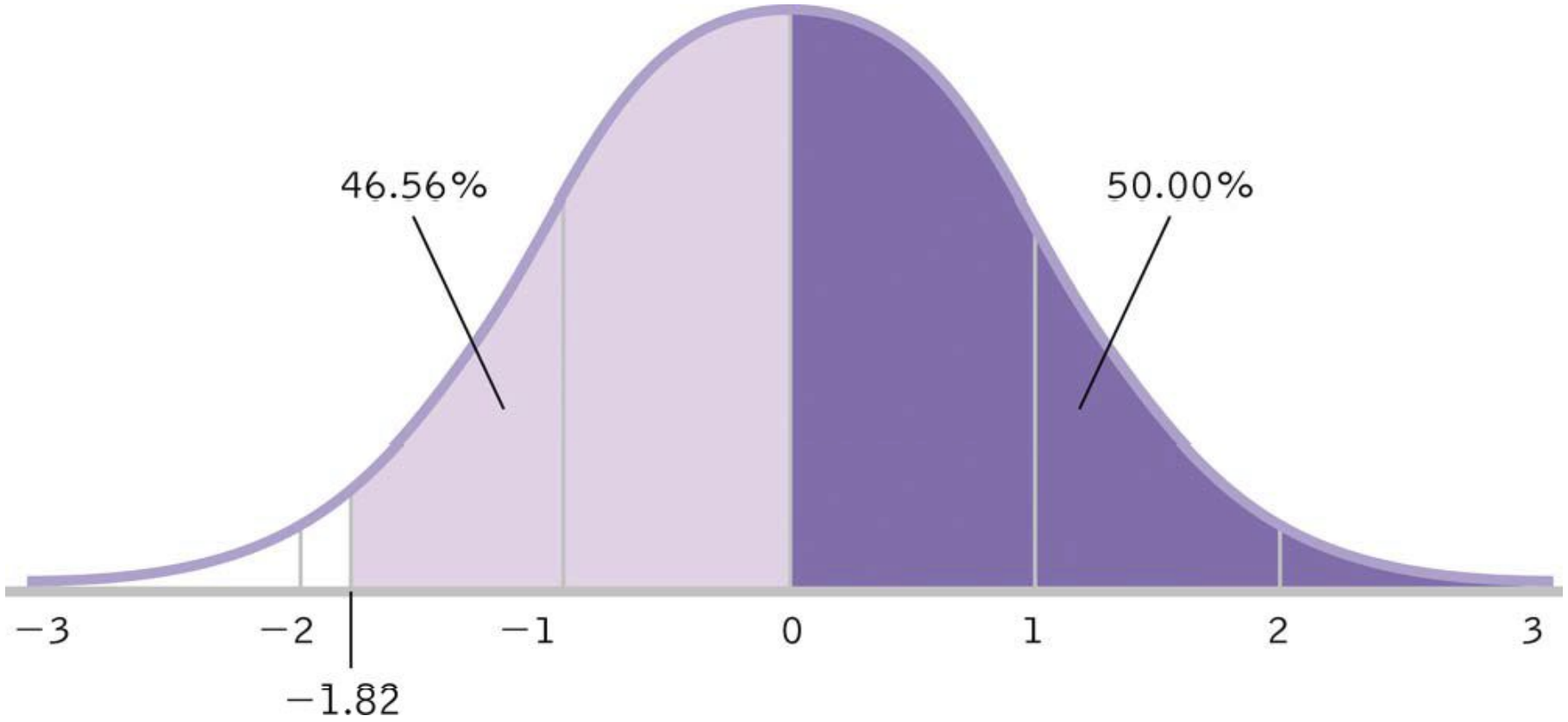
**Negative z, below mean: 50% - 46.56% = 3.44%**



# CDC Example: Manuel

**2. Percent Above Manuel**

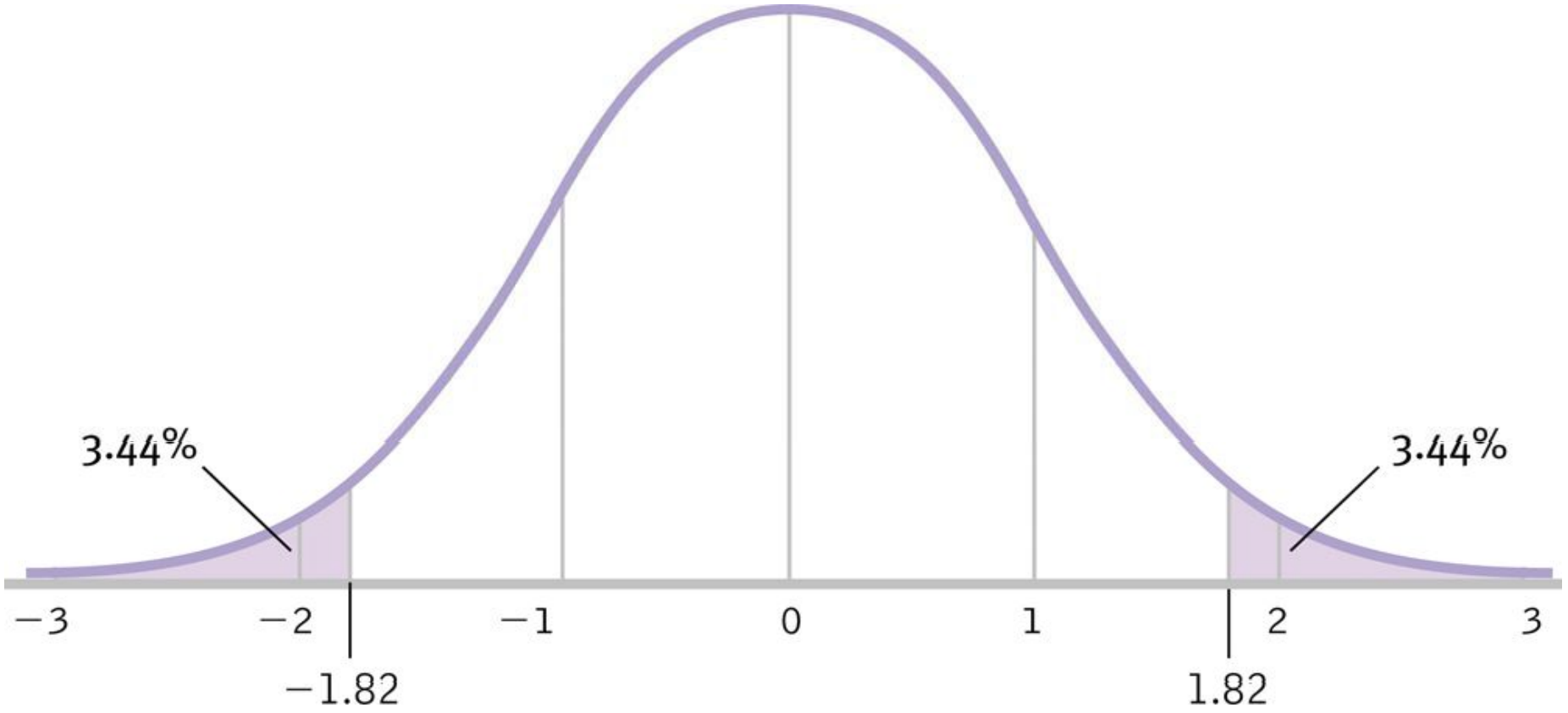
**100% - 3.44% = 96.56 %**



# CDC Example: Manuel

**3. Percent as extreme as Manuel**

**3.44% + 3.44% = 6.88%**



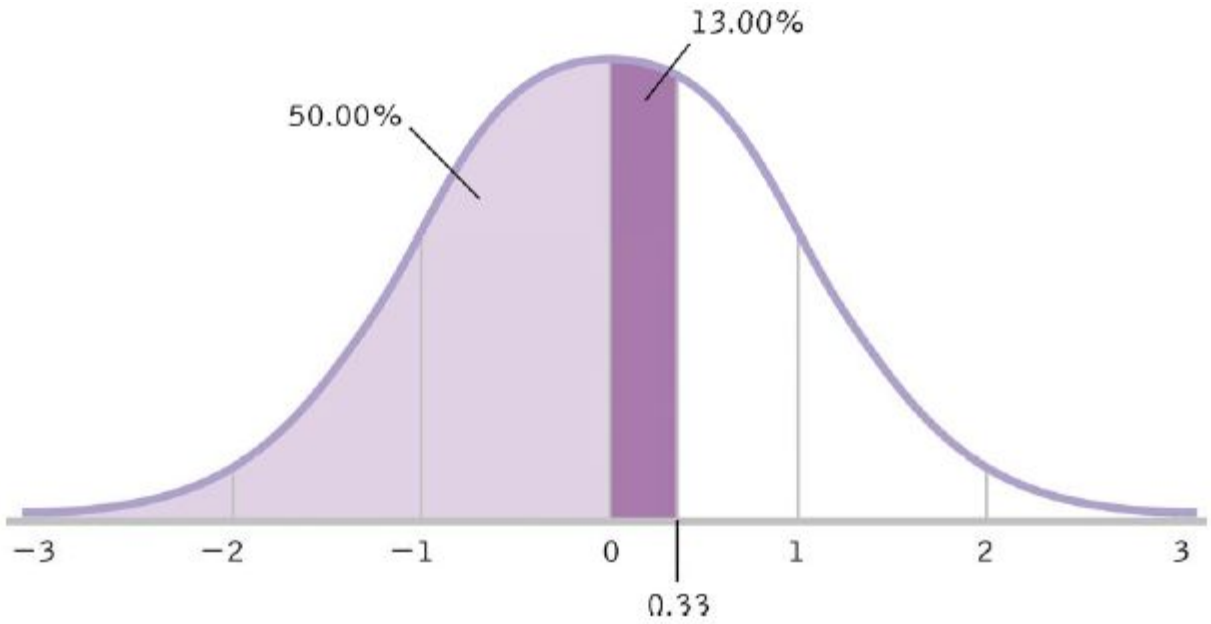
# Percentages to Z-Scores

**SAT Example: μ = 500, σ = 100**

**You find out you are at 63rd percentile**

**Consult z table for 13% -> z = .33**

**X = .33(100) + 500 = 533**



# Z-Table and Distribution of Means

**Remember that if we use distribution of means, we are using a sample and need to use standard error. How do UMD students measure up on the GRE?**

**μ = 554, σ = 99** **M = 568, N = 90**

**$\mu_M$ = μ = 554**

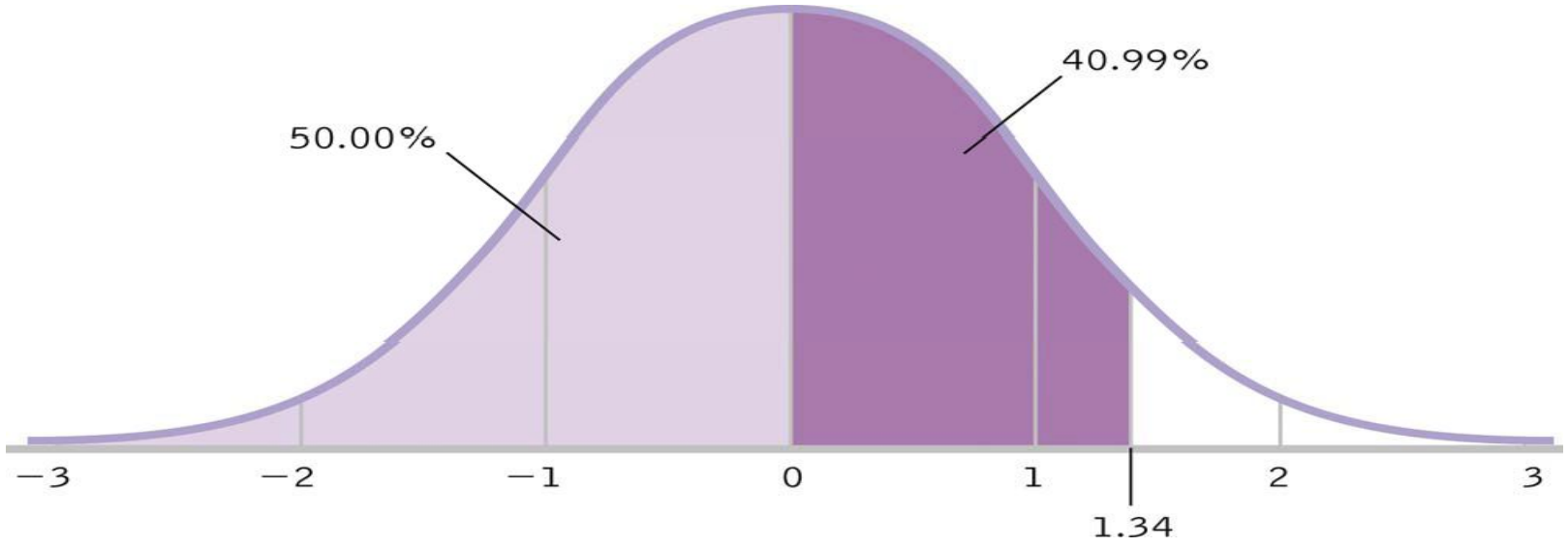
$$\sigma_M = \frac{\sigma}{\sqrt{N}} = \frac{99}{\sqrt{90}} = 10.436$$

# UMich Dearborn & GRE Scores

$$z = \frac{(M - \mu_M)}{\sigma_M} = \frac{(568 - 554)}{10.436} = 1.34$$

**Consult z table for z = 1.34 -> 40.99 %**

**50% + 40.99% = 90.99%**



# Assumptions of Hypothesis Testing

1. The DV is measured on an interval scale
2. Participants are randomly selected
3. The distribution of the population is approximately normal

Robust: These hyp. tests are those that produce fairly accurate results even when the data suggest that the population might not meet some of the assumptions.

- Parametric Tests
- Nonparametric Tests

# Assumptions of Hypothesis Testing

**TABLE 8-2. THE THREE ASSUMPTIONS FOR HYPOTHESIS TESTING**

We must be aware of the assumptions for the hypothesis test that we choose, and we must be cautious in choosing to proceed with a hypothesis test even though our data may not meet all of the assumptions. Note that in addition to these three assumptions, for many hypothesis tests, including the *z* test, the independent variable must be nominal.

| THE THREE ASSUMPTIONS | BREAKING THE ASSUMPTIONS |
|---|---|
| 1. Dependent variable is measured on an interval scale. | Usually OK if the data are not clearly nominal or ordinal. |
| 2. Participants are randomly selected. | OK if we are cautious about generalizing. |
| 3. Population distribution is approximately normal. | OK if the sample includes at least 30 scores. |

# Testing Hypotheses (6 Steps)

1. **Identify the population, comparison distribution, inferential test, and assumptions**
2. **State the null and research hypotheses**
3. **Determine characteristics of the comparison distribution**
    a. Whether this is the whole population or a control group, we need to find the mean and some measure of spread (variability).

# Testing Hypotheses (6 Steps)

4. **Determine critical values or cutoffs**
    a. How extreme must our data be to reject the null?
    b. Critical Values: Test statistic values beyond which we will reject the null hypothesis (cutoffs)
        i. p levels (α): Probabilities used to determine the critical value
5. **Calculate test statistic (e.g., z statistic)**
6. **Make a decision**
    a. Statistically Significant: Instructs us to reject the null hypothesis because the pattern in the data differs from what we would expect by chance alone.

# The z Test: An Example

**$\mu = 156.5$, $\sigma = 14.6$, $M = 156.11$, $N = 97$**

1. **Populations, distributions, and assumptions**
    a. Populations:
        i. All students at UMD who have taken the test (not just our sample)
        ii. All students nationwide who have taken the test
    b. Distribution: Sample -> distribution of means
    c. Test & Assumptions: z test
        i. Data are interval
        ii. We hope random selection (otherwise, less generalizable)
        iii. Sample size > 30, therefore distribution is normal

# The z Test: An Example

2. State the null and research hypotheses

$$H_0: \mu 1 \leq \mu 2$$
$$H_1: \mu 1 > \mu 2$$

or

$$H_0: \mu 1 = \mu 2$$
$$H_1: \mu 1 \neq \mu 2$$

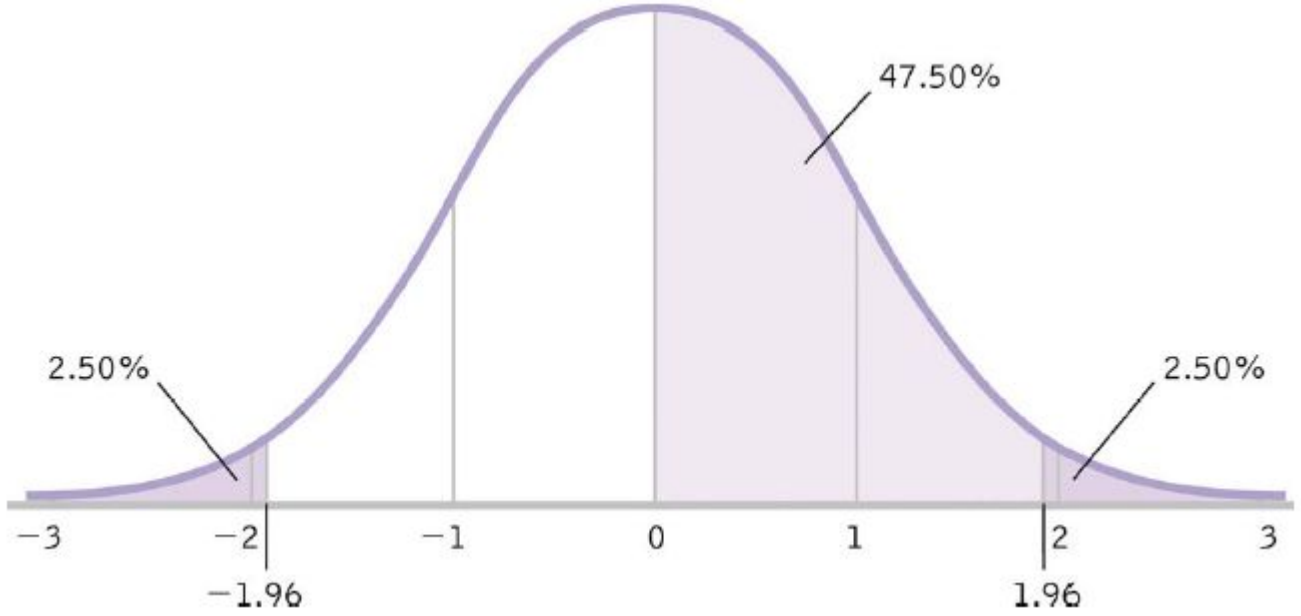
# The z Test: An Example

3. **Determine characteristics of comparison distribution.**
   a. Population: μ = 156.5, σ = 14.6
   b. Sample: M = 156.11, N = 97

$$\sigma_M = \frac{\sigma}{\sqrt{N}} = \frac{14.6}{\sqrt{97}} = 1.482$$

# The z Test: An Example

4. **Determine critical value (cutoffs)**
   a. In Data Science, we use p = .05
      i. p = .05 = 5% -> 2.5% in each tail
   b. 50% - 2.5% = 47.5%
   c. Consult z table for 47.5% -> z = 1.96

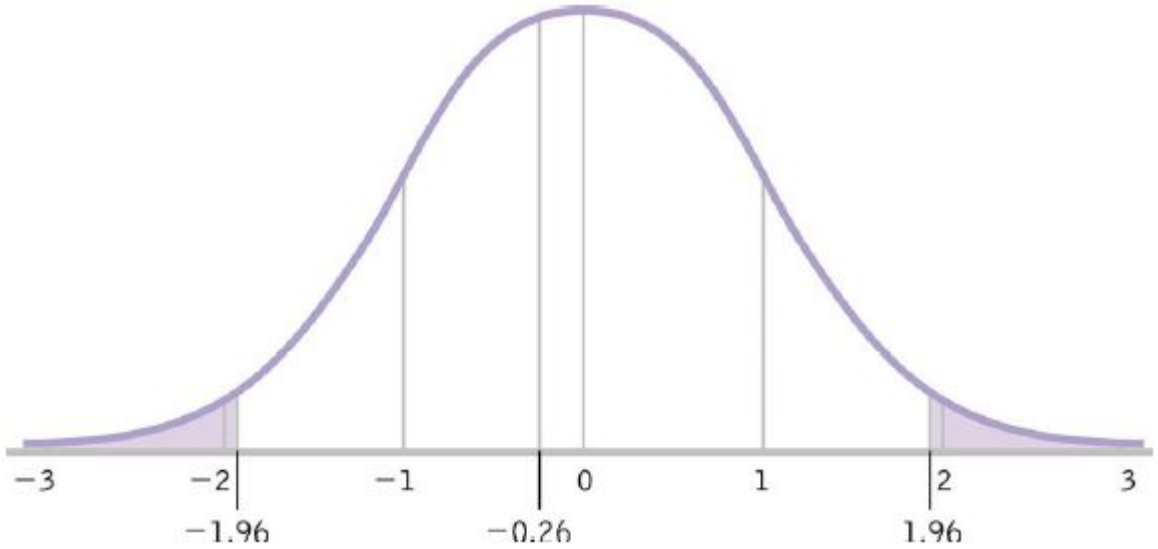


# The z Test: An Example

5. **Calculate test statistic**

$$z = \frac{(M - \mu_M)}{\sigma_M} = \frac{(156.11 - 156.5)}{1.482} = -0.26$$

6. **Make a Decision**



# Increasing Sample Size

- By increasing sample size, one can increase the value of the test statistic, thus increasing probability of finding a significant effect

- Example: Psychology GRE scores:

  Population: μ = 554, σ = 99

  Sample: M = 568, N = 90

$$\sigma_M = \frac{\sigma}{\sqrt{N}} = \frac{99}{\sqrt{90}} = 10.436$$

$$z = \frac{(M - \mu_M)}{\sigma_M} = \frac{(568 - 554)}{10.436} = 1.34$$

Population: μ = 554, σ = 99

Sample: M = 568, N = 200

$$\sigma_M = \frac{\sigma}{\sqrt{N}} = \frac{99}{\sqrt{200}} = 7.00$$

$$z = \frac{(M - \mu_M)}{\sigma_M} = \frac{(568 - 554)}{7.00} = 2.00$$

# Calculating t-Statistic

**Because our sample size is less than 30, this will now have a t distribution!**

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

**We can then use the t-Score to calculate the probability of getting a t value at least t extreme. Look it up on a t-table!**

# z-Score and t-Scores in Python

**Use SciPy!**

```
import scipy.stats as stats
values = [4,5,6,6,6,7,8,12,13,13,14,18]

zscores = stats.zscore(values)
```

```
import scipy.stats as stats


stats.t.ppf(q=1-.05/2,df=22)
```

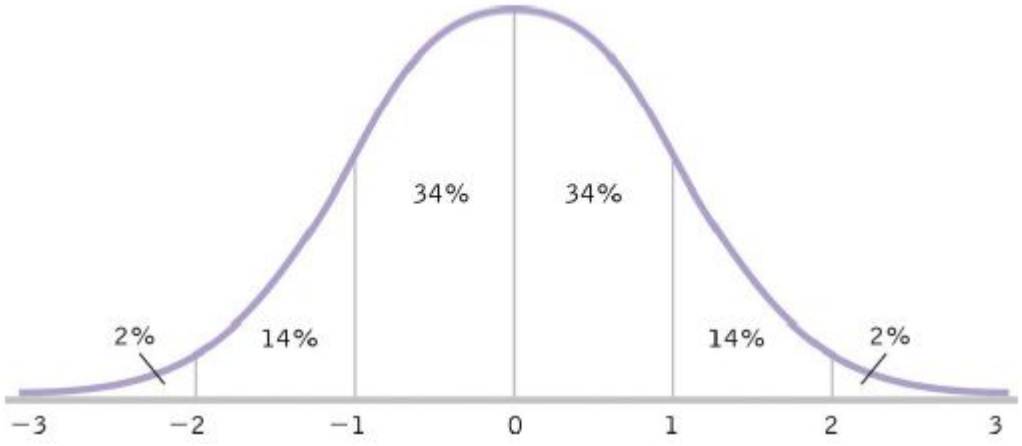
# Two Sample z and t-Test

1. Bivariate independent variable (A, B groups)
2. Continuous dependent variable
3. Each observation of the dependent variable is independent of the other observations of the dependent variable (its probability distribution isn't affected by their values).

**Exception: For the paired t-test, we only require that the pair-differences (Ai - Bi) be independent from each other (across i). [Note: "independent" and "dependent" are used in two different senses here. Just think of a "dependent variable" as one thing, and "observations that are dependent" as another thing.]

# Two Sample z and t-Test

4. **Dependent variable has a normal distribution, with the same variance, σ2, in each group (as though the distribution for group A were merely shifted over to become the distribution for group B, without changing shape):**



# Two Sample z and t-Test

**Used for determining if there is a statistical difference between two the means of two groups.**

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{(s^2(\frac{1}{n_1} + \frac{1}{n_2}))}}$$