Hypothesis Testing, Probability and Distributions

Pablo E. Gutierrez-Fonseca

Fall 2024

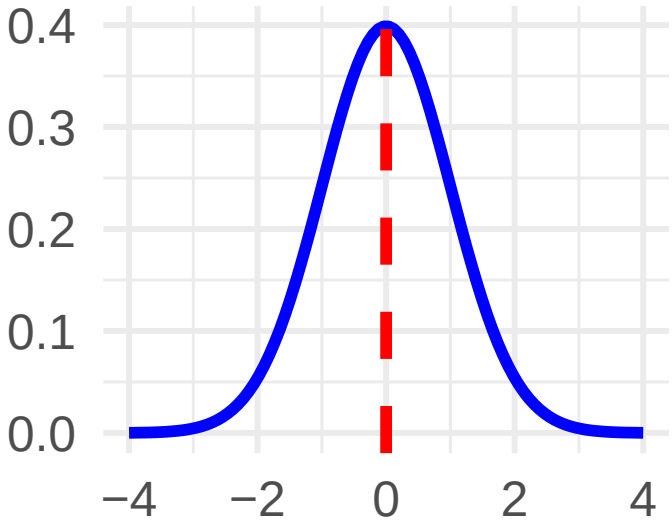
Normal distribution: Introduction

- Why did we focus on normality?
- The **normal distribution** is a key tool for determining the probability of a given value occurring in a population that follows this distribution.
- It allows us to make inferences about a population by calculating how likely it is for data to fall within certain ranges.
- Many statistical tests assume data follows a **normal distribution**, which helps in determining **significance** and making reliable conclusions.

University of Vermont

Normal distribution: Introduction

- The normal distribution has two parameters: the mean (μ) and the standard deviation (σ).
- If a **variable x** follows a normal distribution with a specific mean (μ) and standard deviation (σ), we write this as:

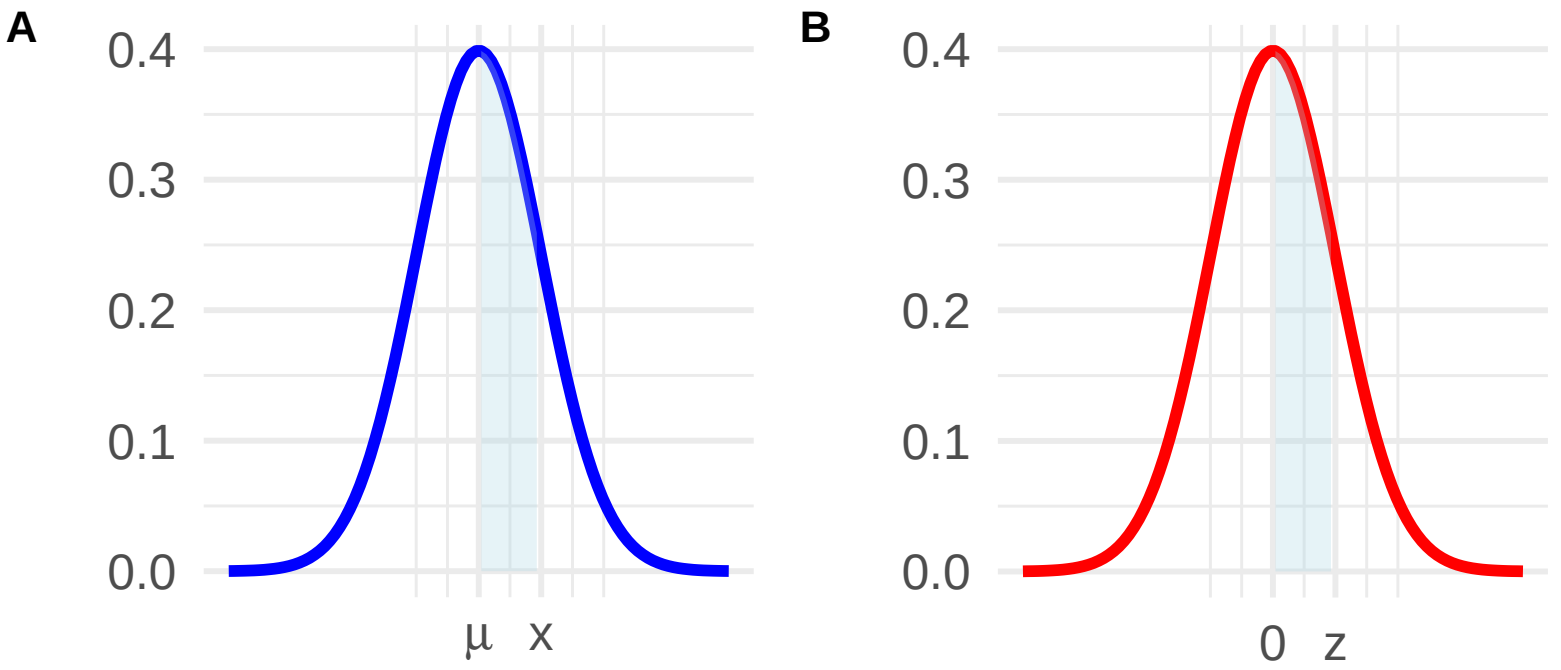


$$x \sim N(\mu, \sigma)$$

University of Vermont

Converting to Standard Normal Distribution

- To simplify comparisons between different normal distributions, we often convert values to the **standard normal distribution**.
- The standard normal distribution has a $\mu = 0$ and a $\sigma = 1$.
- Any normal distribution can be converted into the standard normal distribution by using the **z-score** formula: $z = \frac{x-\mu}{\sigma}$



University of Vermont

Hypothesis Testing, Probability and Distributions

- Hypothesis Testing
  - Null and alternative hypothesis
  - Test statistic
- Probability and Distributions
- z-score

University of Vermont

Hypothesis Testing: Null and alternative hypothesis

- A **null hypothesis** (denoted by H_0) is a statement of the status quo, one of **no difference or no effect**.
  - If the null hypothesis is not rejected, no changes will be made.

$$H_0\colon\ \mu = \mu_0$$

- An **alternative hypothesis** (denoted by H_1) is one in which **some difference or effect is expected**.
- One sided alternative hypothesis:

$$H_1\colon\ \mu < \mu_0 \text{ or } H_1\colon\ \mu > \mu_0$$

- Two sided alternative hypothesis:

$$H_1\colon\ \mu \neq \mu_0$$

University of Vermont

Hypothesis Testing: Test statistic

- All inferential tests use a formula that calculates a **test statistic**, quantifying the relationship or difference you are testing.

- Independent t-test:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

  - Dependent t-test

$$t = \frac{\bar{D}}{\frac{s_D}{\sqrt{n}}}$$

  - One sample z-test

$$z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

- F-test (ANOVA)

$$F = \frac{MST}{MSE}$$

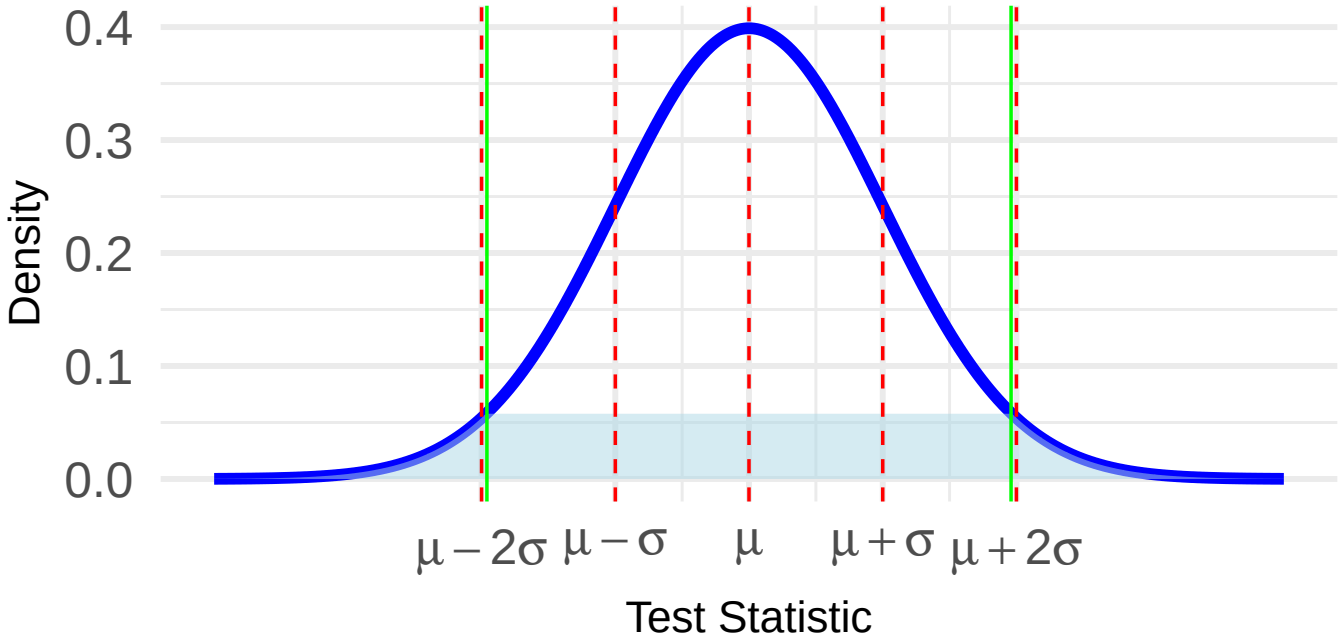
- Pearson correlation

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

University of Vermont

Hypothesis Testing: Test statistic

- The **normal distribution curve** allows us to calculate the **probability** of obtaining a **test statistic** as extreme as the one observed, assuming the **Null Hypothesis is true**.
- The **test statistic** is a value used in making a decision about the null hypothesis, and is found by converting the sample statistic to a score with the assumption that the null hypothesis is true



Hypothesis Testing: Test statistic

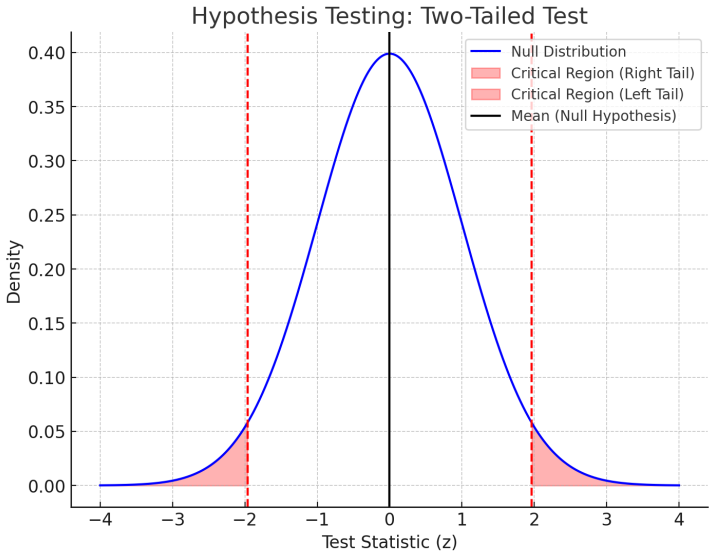
- If the **probability** (i.e., *the area under the curve*) of obtaining such an extreme test statistic is less than our chosen significance level (*usually 0.05*), we consider the result to be **statistically significant**.
- In such cases, we may **reject the Null Hypothesis**, suggesting there is evidence for a difference or effect.

University of Vermont

Hypothesis Testing: Test statistic

- Another way to think about this:

- If the **probability that our data fits the null distribution (i.e., the null hypothesis is true) is less than 5%**, we conclude that the data **does not fit the null**.
- This indicates a significant deviation from what we would expect by chance.
- We then **reject the null hypothesis**.
- The result is considered **statistically significant**.



University of Vermont

Hypothesis Testing, Probability and Distributions

- Hypothesis Testing
- Probability and Distributions
- z-score

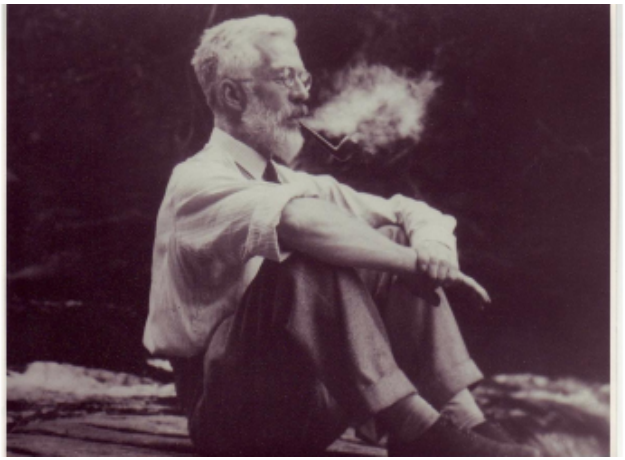
University of Vermont

Before: p-value, significance and alfa level

Why p-value of less than **0.05**?

"It is usual and convenient for experimenters to take 5% as a standard level of significance, in the sense that they are prepared to ignore all results which fail to reach this standard, and, by this means, to eliminate from further discussion the greater part of the fluctuations which chance causes have introduced into their experimental results."

- Ronald Aylmer Fisher (1890-1962)



University of Vermont

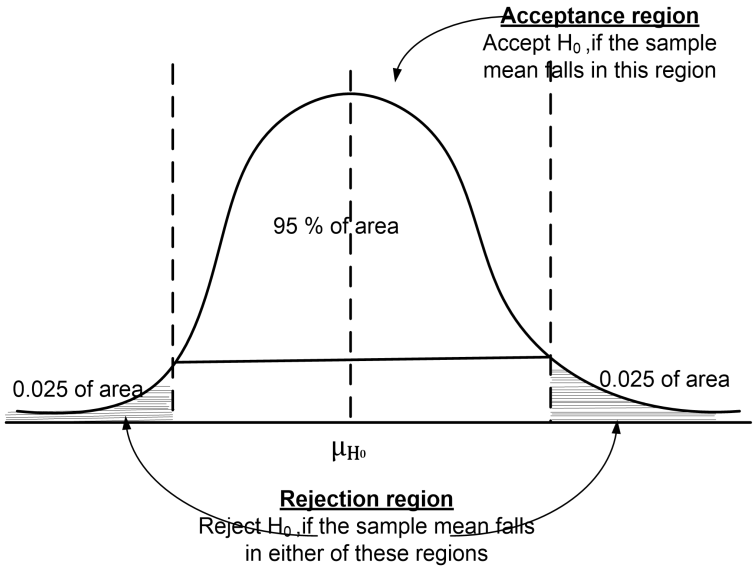
What is a P-Value?

- The **p-value** (or p-value or probability value) is the probability of getting a value of the **test statistic** that is **at least as extreme** as the one representing the sample data, assuming that the null hypothesis is true
- Interpretation:
  - A small p-value indicates that the observed result is unlikely under the null hypothesis, suggesting evidence against it.
  - A large p-value suggests that the observed result is consistent with the null hypothesis.

University of Vermont

Critical Region

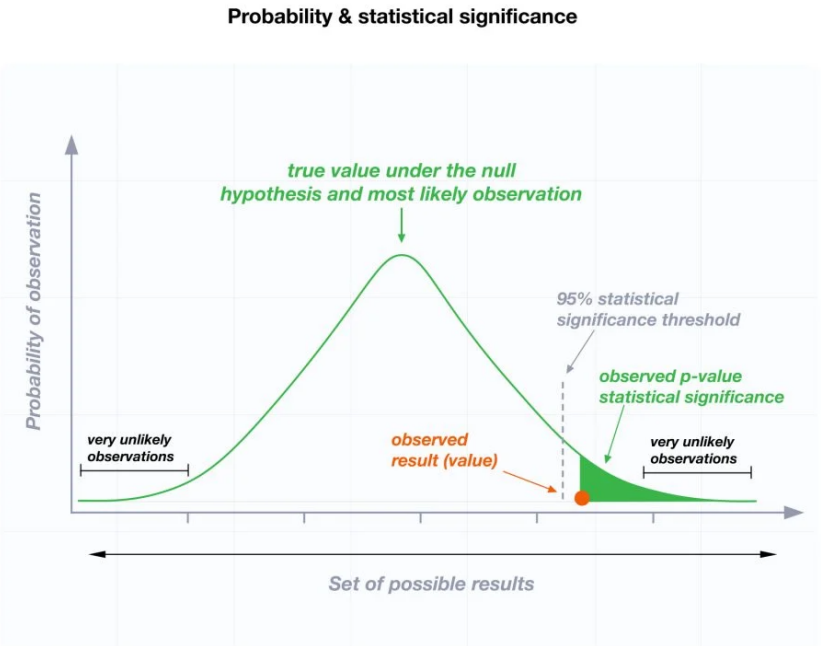
- The critical region (or rejection region) is the set of all values of the test statistic that cause us to reject the null hypothesis.



- Acceptance and rejection regions in case of a two-tailed test with 5% significance level.

What is Significance?

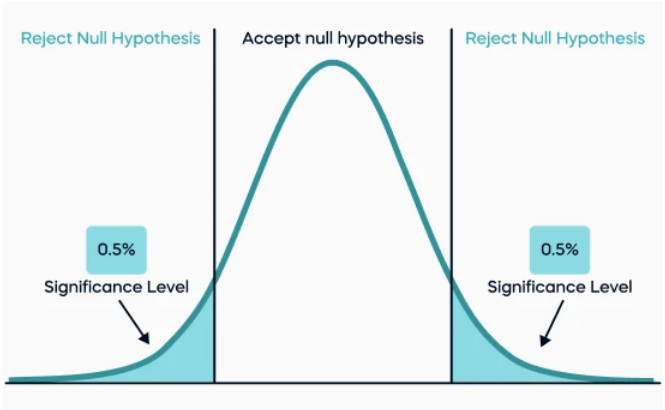
- A statistical result is **significant** if it is **unlikely to have occurred by chance**.
- Despite natural variability in the population, the probability of observing a value this extreme due to random variability is **low** (though not impossible).
- We use probabilities (**p-values**) and an **alpha threshold (commonly 0.05)** to determine whether a result is significant.



University of Vermont

What is Significance?

- Significance refers to the risk of rejecting the null hypothesis when it is actually true.
- It tells us the probability that our result happened by chance alone.
- A p-value of 0.05 (5%) means there's a 5% chance the result is due to random chance.
- A p-value of 0.01 (1%) means there's a 1% chance the result happened by chance.
  - A p-value of 0.01 indicates a low chance (1%) of the result occurring by chance, reflecting a more rigorous threshold for significance.



Focus - Types of Error

Focus - Types of Error

Scenario	Null Hypothesis is True	Null Hypothesis is False
Reject Null Hypothesis	**Type I Error**: Incorrectly rejecting the null hypothesis.	**Correct Decision**: Correctly rejecting the null hypothesis.
Fail to Reject Null Hypothesis	**Correct Decision**: Correctly not rejecting the null hypothesis.	**Type II Error**: Incorrectly failing to reject the null hypothesis.

University of Vermont

Focus - Types of Error

Scenario	Null Hypothesis is True	Null Hypothesis is False
Reject Null Hypothesis	**Type I Error**: Incorrectly rejecting the null hypothesis.	**Correct Decision**: Correctly rejecting the null hypothesis.
Fail to Reject Null Hypothesis	**Correct Decision**: Correctly not rejecting the null hypothesis.	**Type II Error**: Incorrectly failing to reject the null hypothesis.

- Failing to reject the null hypothesis means our data didn't show a significant effect. It doesn't prove the null hypothesis is true; it just means we didn't find strong evidence against it.

University of Vermont

- **Explanation**:
  - **Type I Error**: False positive. We conclude there is an effect or difference when there is none.
  - **Type II Error**: False negative. We fail to detect an effect or difference when one exists.
  - **Correct Decisions**: Accurately concluding the presence or absence of an effect or difference based on the truth of the null hypothesis.

University of Vermont

Calculating Significance

- To quantify a probability, you first need to calculate a **test statistic** and locate it on the normal probability curve.
- The normal curve acts as a statistical translator.
  - It helps you standardize your test statistic to a common scale.
  - This standardized value is then used to determine the probability of obtaining such a result in a standard normal population.

University of Vermont

Hypothesis Testing, Probability and Distributions

- Hypothesis Testing
- Probability and Distributions
- z-score

University of Vermont

z-score

- **z-scores** link measured or hypothesized values to probabilities.
- A z-score (or standard score) indicates how many standard deviations a value x is above or below the mean on the normal curve.
- It helps standardize values and connect them to probabilities.
- The z-score is calculated using the formula:

$$z = \frac{(x - \mu)}{\sigma}$$

- where:
  - x = the value of interest.
  - μ = mean of the population.
  - σ = standard deviation of the population.

University of Vermont

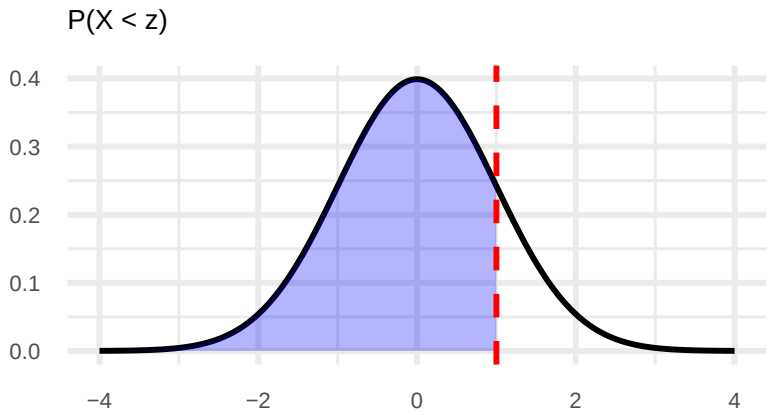
z-scores: Linking observations to probabilities

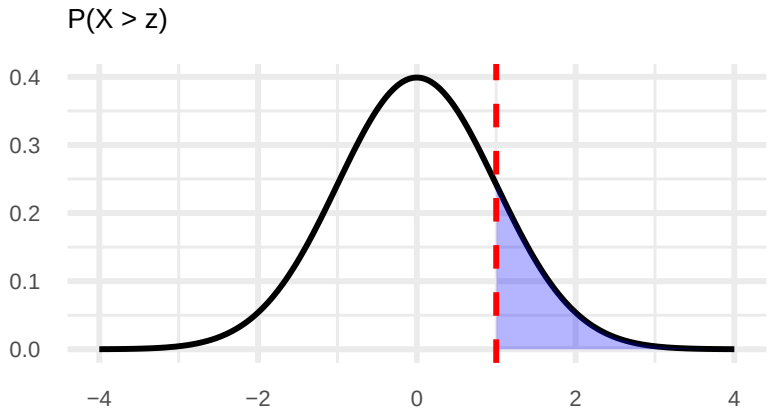
- **z-scores** link observations to probabilities.
- Using a z-score for probability:
  - z-scores are essentially the **x-axis** of the standard normal distribution.
  - They normalize any data set so that the mean is 0 and the standard deviation is 1.
  - The area under the curve tells you the probability of a certain Z-score occurring.
  - By using the Z-score, we can determine the probability associated with different values.

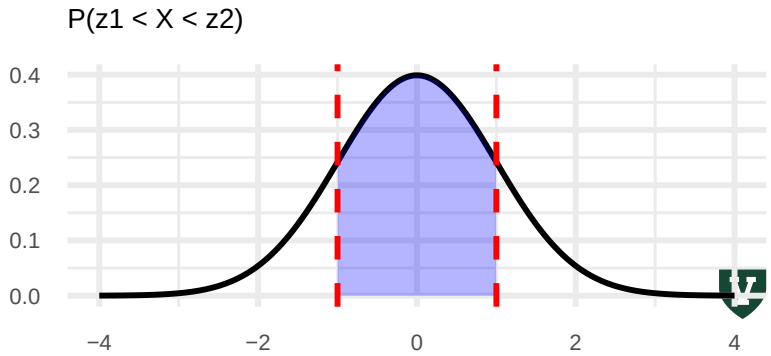
University of Vermont

Finding probabilities for z-scores

- **P(X < z)**: Denotes the probability of a value falling **less than** a given Z-score (z).
- **P(X > z)**: Denotes the probability of a value falling **above** a given Z-score (z).
- **P(z1 < X < z2)**: Denotes the probability of a value falling **between** two different Z-scores ($z1$ and $z2$).





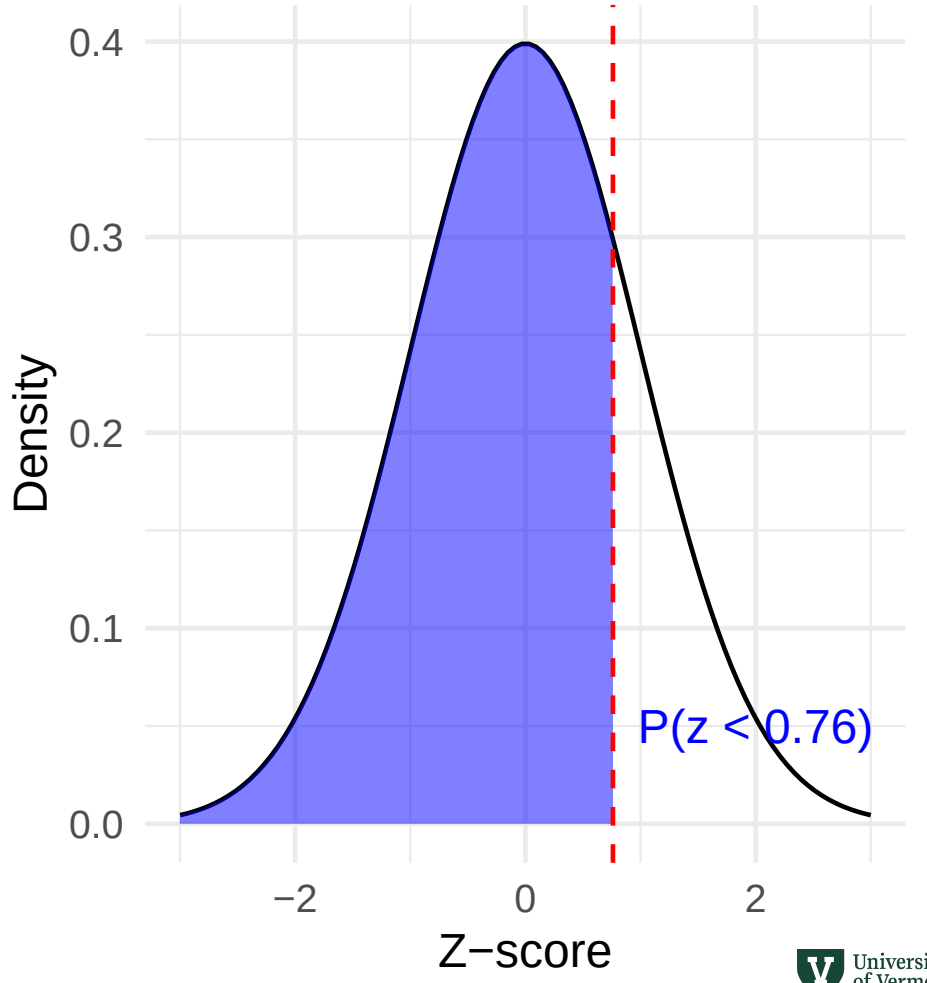


University of Vermont

Example

Example

- What percent of the area under the curve **falls below** a z-score of 0.76?
- P (z<0.76)

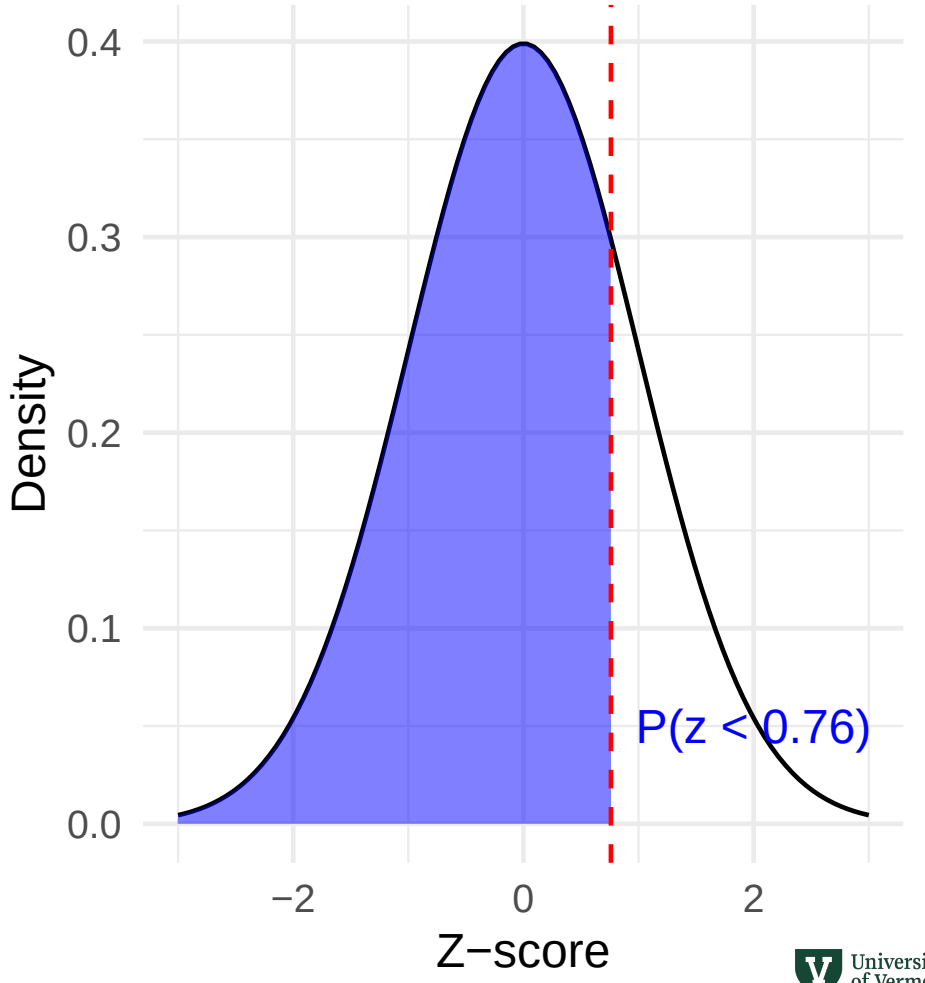


Example

- What percent of the area under the curve **falls below** a z-score of 0.76?
- P (z<0.76)

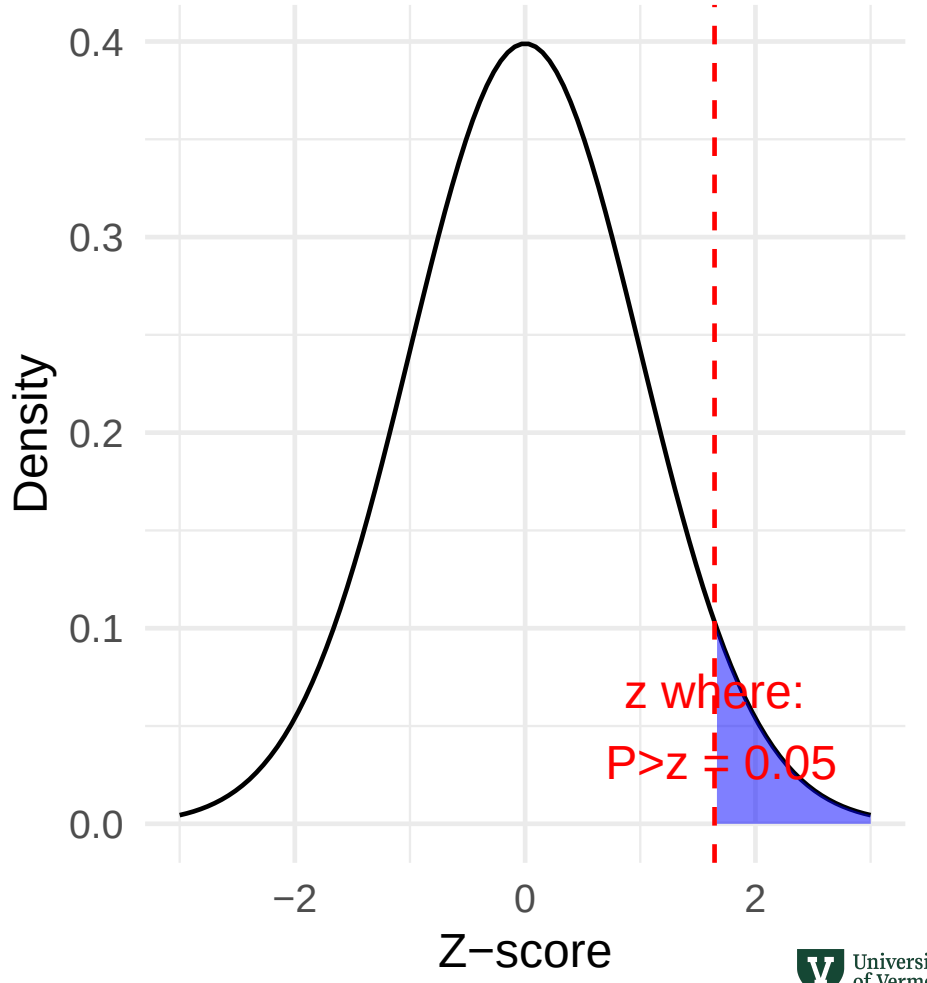
```
# Calculate the cumulative
# probability for z = 0.76
z_score <- 0.76
p_value <- pnorm(z_score)
# Convert to percentage
percent_area <- p_value * 100
percent_area
```

```
## [1] 77.63727
```



Example

- What is the z-score beyond which only 5% of all possible outcomes are higher?
- (P>z) = 0.05

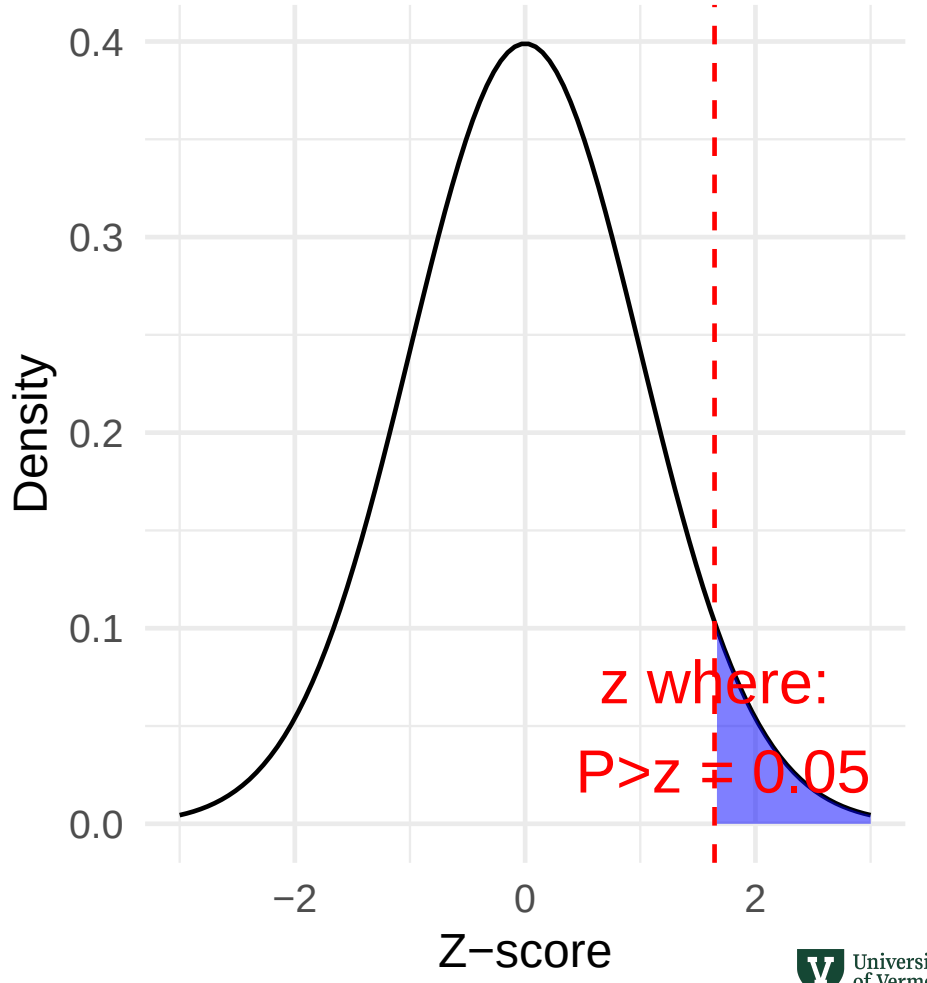


Example

- What is the z-score beyond which only 5% of all possible outcomes are higher?
- (P>z) = 0.05

```
# Calculate the z-score for
# the upper 5% (0.95 cumulative
# probability)
p_value <- 0.95
z_score <- qnorm(p_value)
# Print the result
z_score
```

```
## [1] 1.644854
```



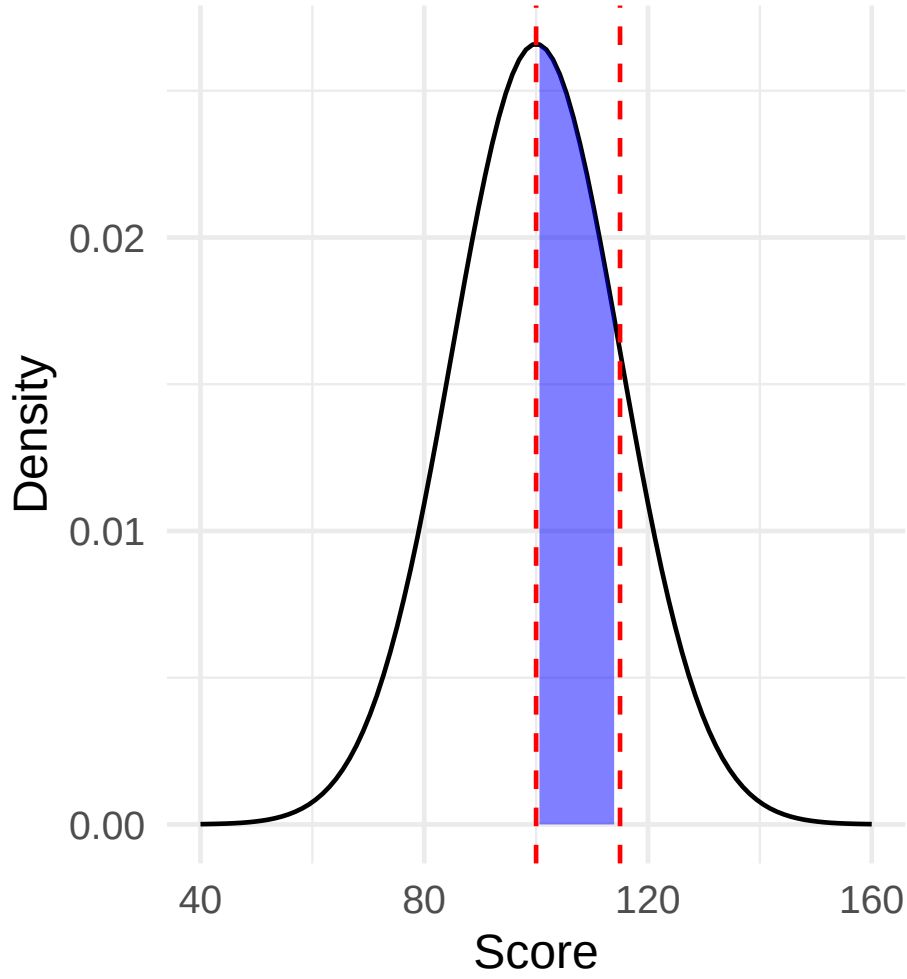
Example

- In a distribution with a mean of 100 and a standard deviation of 15, what is the probability that a score will fall between 100 and 115 ?

$$Z = \frac{(x - \mu)}{\sigma}$$

- Before you can find any probabilities you have to find z-scores
  - Z = (100 - 100) / 15 = 0
  - Z = (115 - 100) / 15 = 1

```
## [1] "Probability of falling
```



University of Vermont

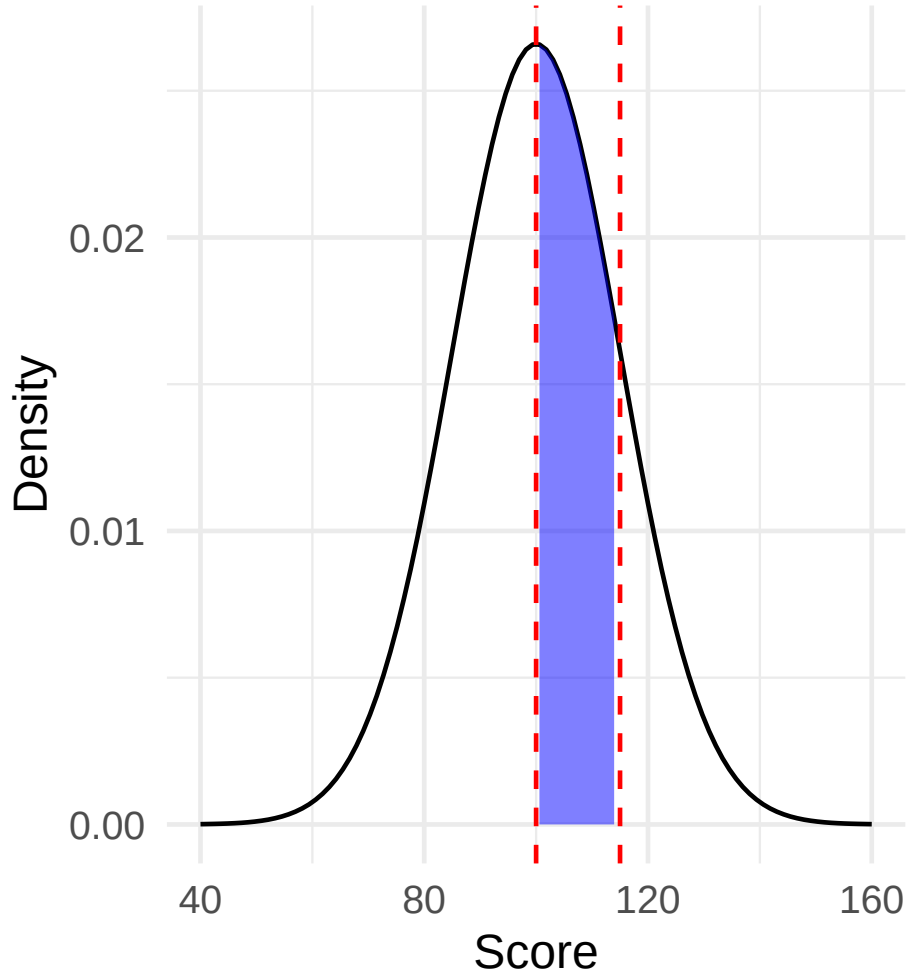
Example

- In a distribution with a mean of 100 and a standard deviation of 15, what is the probability that a score will fall between 100 and 115 ?

```
## [1] "Probability of falling
```

```
# Parameters
mean <- 100
sd <- 15
# Values
low_value <- 100
high_value <- 115

# Calculate the z-scores
z_low <- (low_value-mean)/sd
z_high <- (high_value-mean)/sd
```



Example

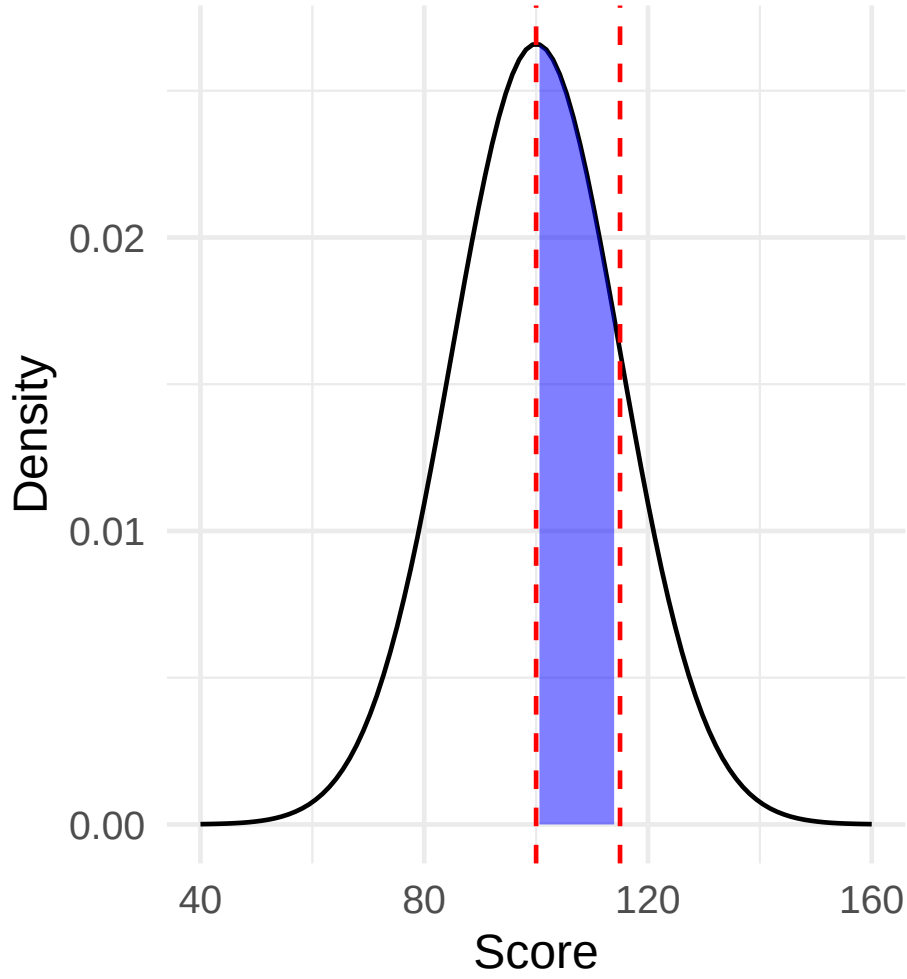
- In a distribution with a mean of 100 and a standard deviation of 15, what is the probability that a score will fall between 100 and 115 ?

```
# Find the probabilities
p_low <- pnorm(z_low)
p_high <- pnorm(z_high)

# Probability of falling
# between 100 and 115
probability <- p_high - p_low
probability
```

```
## [1] 0.3413447
```

```
## [1] "Probability of falling
```



University of Vermont

Example

Example: Probability of a Cat Living to a Certain Age

- The lifespan of domestic cats is normally distributed with a mean of 15.7 years and a standard deviation of 1.6 years.

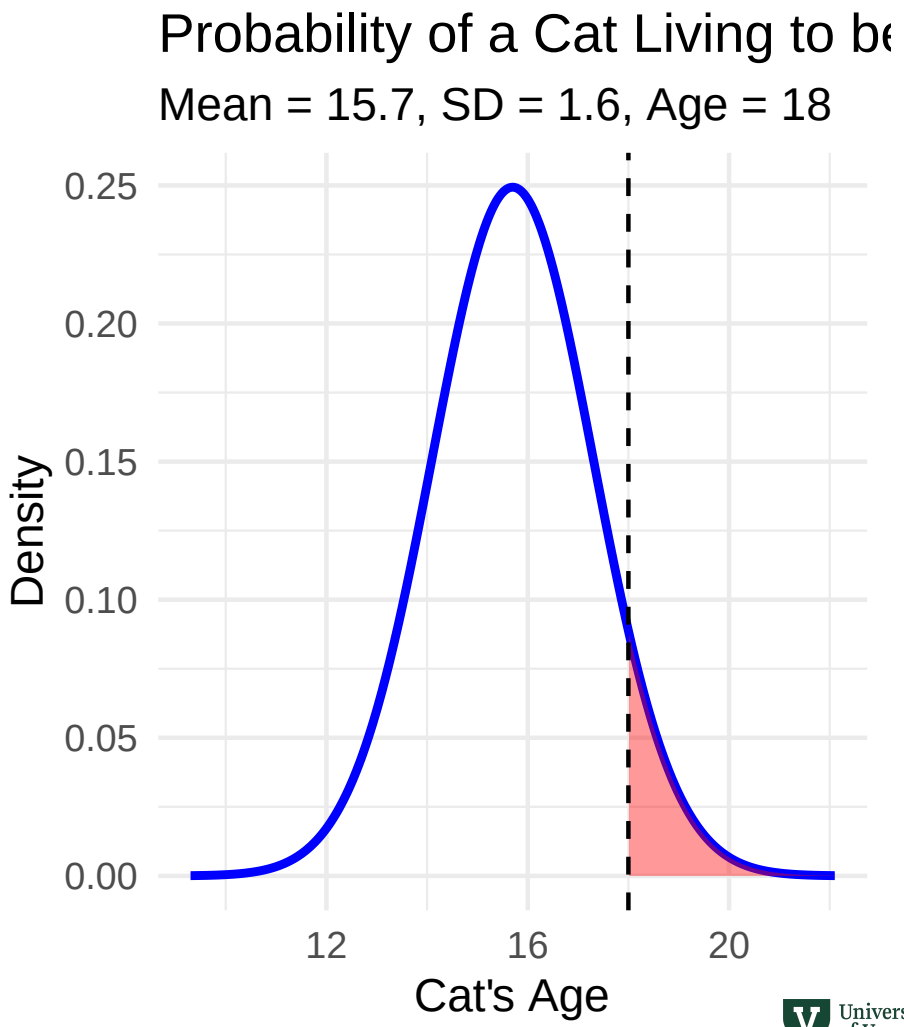
- **Question**: What is the probability that a cat will live to be as old as Allison's 18-year-old cat?

University of Vermont

Example: Probability of a Cat Living to a Certain Age

- The lifespan of domestic cats is normally distributed with a mean of 15.7 years and a standard deviation of 1.6 years.

- **Question**: What is the probability that a cat will live to be as old as Allison's 18-year-old cat?
- We're looking for the probability $P(X > 18)$, which represents the probability that a cat will live longer than 18 years.



Steps 1:

- **Calculate the Z-score**:

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

- where:
  - X is the value (18 years),
  - μ is the mean (15.7 years),
  - σ is the standard deviation (1.6 years).

$$Z = \frac{18 - 15.7}{1.6} = \frac{2.3}{1.6} \approx 1.4375$$

University of Vermont

Step 2.

- Find the Probability:

```
# Given values
mean_lifespan <- 15.7
sd_lifespan <- 1.6
age_allison_cat <- 18
# Calculate Z-score
z_score <- (age_allison_cat - mean_lifespan) / sd_lifespan
# Find the probability that a cat lives longer than 18 years
probability <- 1 - pnorm(z_score)
probability
```

```
## [1] 0.07528799
```

- Thus, the probability that a cat will live to be as old as or older than 18 years is approximately **0.0749** or **7.49%**.

University of Vermont

Step 2.

- Find the Probability:

```r
# Given values
mean_lifespan <- 15.7
sd_lifespan <- 1.6
age_allison_cat <- 18
# Calculate Z-score
z_score <- (age_allison_cat - mean_lifespan) / sd_lifespan
# Find the probability that a cat lives longer than 18 years
probability <- 1 - pnorm(z_score)
probability
```

```
## [1] 0.07528799
```

- This example walks you through calculating the probability in R using `pnorm()`, which calculates the cumulative probability under the normal distribution.

Another real life problems

Another real life problems

- **Context**: The EPA is assessing drinking water standards to protect public health.

- **Problem**: If the EPA sets the maximum allowable lead concentration in drinking water at 1ppm, and the lead concentrations in public buildings are normally distributed with a mean of 0.6 ppm and a standard deviation of 0.2ppm.
  - **What proportion of public buildings will exceed this threshold and require lead remediation?**

University of Vermont

Another real life problems

- **Context**: The EPA is assessing drinking water standards to protect public health.

- Analysis Required:
  - Calculate the z-score for the threshold of 1 ppm.
  - Determine the proportion of buildings exceeding this lead level using the normal distribution.

```
# Parameters
mean <- 0.6
sd <- 0.2
# Threshold 4 lead remediation
threshold <- 1
# Calculate the z-score for the th
z_score<-(threshold-mean)/sd
# Calculate the proportion of
#buildings above the threshold
threshold<-1-pnorm(z_score)
threshold
```

```
## [1] 0.02275013
```

University of Vermont