Statistical Inference

Overview

1. **Population vs. Sample**
2. **Inventing Null Worlds**
3. **The Central Limit theorem**
4. **Theoretical Distributions**

Population vs. Sample

Why do we do statistics ?

To make inferences about a population based on observing only a sample

Are action movies better higher than comedies?

Data → Calculation → Estimate → Truth

Category	Description	Notation
Data	IMDB ratings	D
Calculation	Average action rating − average comedy rating	$\bar{D} = \frac{\sum D_{\text{Action}}}{N} - \frac{\sum D_{\text{Comedy}}}{N}$
Estimate	$\bar{D}$ in a sample of movies	$\hat{\delta}$
Truth	Difference in rating for *all* movies	δ

Greek, Latin, and extra markings

Greek

- Letters like δ are the ***truth***
- Letters with extra markings like $\hat{\delta}$ are our ***estimate*** of the truth based on our sample

Latin

- Letters like D are ***actual data*** from our sample
- Letters with extra markings like $\bar{D}$ are ***calculations*** from our sample

Your turn #1: Calculating an estimate

Collect IMDB ratings for a bunch of films via the `ggplot2movies` package.

1. Install the package (use either console or the Rstudio interface. Do not use a Script)
2. Load the package in your script.
3. Load the movies data (type: `data("movies")`)
4. Make a new cleaned data frame by

- selecting only the `title`, `year`, `rating`, `Action` and `Comedy` columns
- filtering out films that classify as both Action and Comedy
- making a new variable `genre` (using `mutate()` and `case_when()`) which takes the values “Action” or “Comedy”
- removing the now obsolete `Action` and `Comedy` columns (use `select` and `–`)

5. Calculate the average ratings for the two genres

08:00

So, are action movies better than comedies?

```
# A tibble: 2 × 2
  genre  avg_rating
  <fct>       <dbl>
1 Action       5.24
2 Comedy       5.97
```

$$\delta \hat{=} \bar{D} = 5.24 - 5.97 = -0.73$$

Action movies seem to be slightly worse. But…

We don’t know if the estimate we found in this sample is actually true for the population of **all films**

Inventing Null Worlds

Simulated Null World

- Let's try to imagine a world with no differences between action and comedy movies
- We simulate data with ratings for 1'000'000 movies where there is no difference (the true δ is 0). Imagine that's the population, i.e. all movies ever made.

```r
set.seed(1234) # For reproducibility

imaginary_movies <- tibble(
  movie_id = 1:1000000,
  rating = sample(seq(1, 10, by = 0.1), size = 1000000, replace = TRUE),
  genre = sample(c("Comedy", "Action"), size = 1000000, replace = TRUE)
)
```

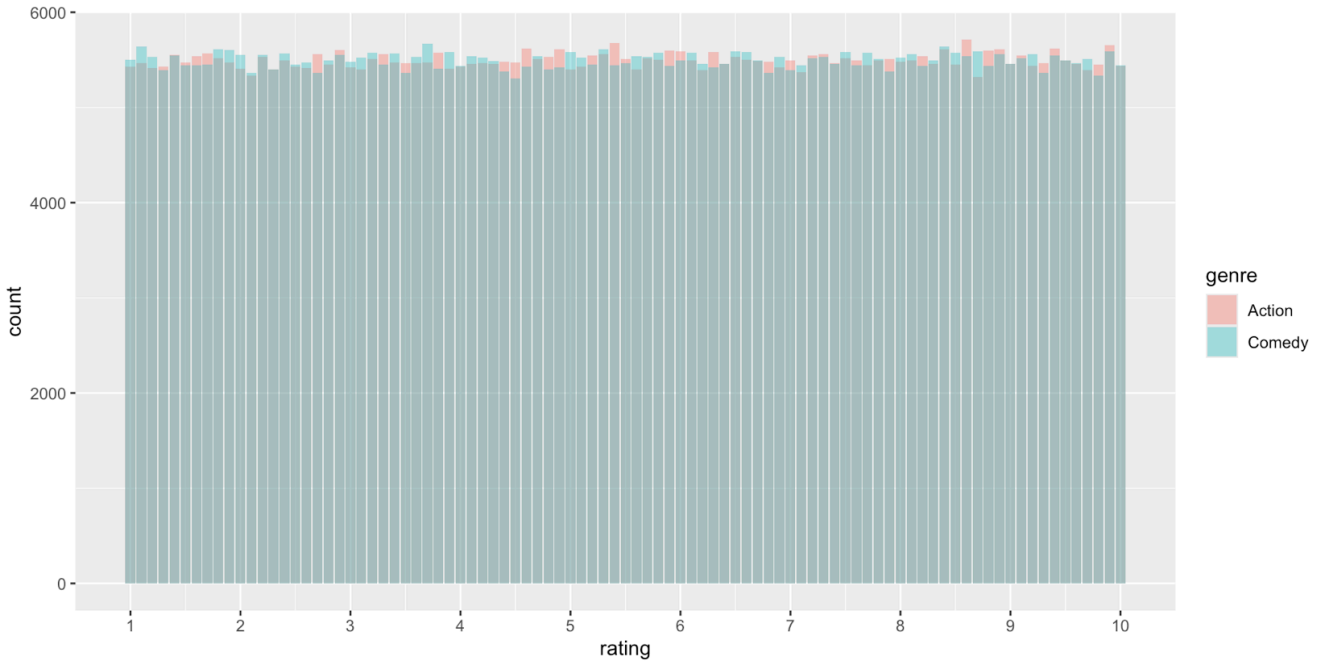
Simulated Null World

Our simulated action movies and comedies don't all have the same rating, but on average there's (almost) no difference

```
imaginary_movies |>
  group_by(genre) |>
  summarize(avg_rating = mea
```

```
# A tibble: 2 × 2
  genre  avg_rating
  <chr>       <dbl>
1 Action       5.51
2 Comedy       5.50
```

```
ggplot(imaginary_movies,
       aes(x = rating, fill = genre)) +
  geom_bar(alpha = 0.4, position = "identity") +
  scale_x_continuous(breaks = seq(1,10))
```



Sampling & Estimating in the Null world

In the actual IMDB data, we looked at a sample of about 20'000 films.

We can randomly pick a sample of that same size from our simulated population

```
# draw a sample of 20'000 films
imaginary_sample <- imaginary_movies |>
  sample_n(20000)
```

In this sample, we actually find a small difference

```
# compute rating difference in the sample
estimate <- imaginary_sample |>
  group_by(genre) |>
  summarize(avg_rating = mean(rating)) |>
  summarise(diff = avg_rating[genre == "Action"] - avg_rating[genre == "Comedy"]) %>%
  pull(diff)

estimate
```

```
[1] -0.03535811
```

Let's repeat this process of sampling and estimating 1000 times, and store the results.

```r
n_simulations <- 1000
differences <- c() # make an empty vector

for (i in 1:n_simulations) {
  # draw a sample of 20'000 films
  imaginary_sample <- imaginary_movies |>
    sample_n(20000)
  # compute rating difference in the sample
  estimate <- imaginary_sample |>
    group_by(genre) |>
    summarize(avg_rating = mean(rating)) |>
    summarise(diff = avg_rating[genre == "Action"] - avg_rating[genre == "Comedy"]) %>%
    pull(diff)

  differences[i] <- estimate
}
```

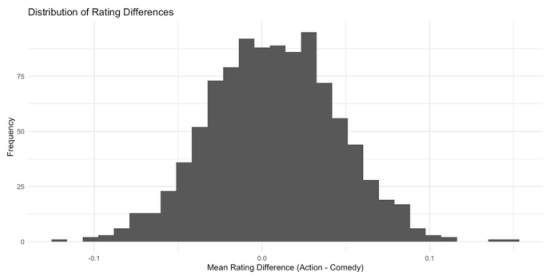
We can plot the results for an overview

```r
n_simulations <- 1000
differences <- c() # make an empty vect

for (i in 1:n_simulations) {
  # draw a sample of 20'000 films
  imaginary_sample <- imaginary_movies
    sample_n(20000)
  # compute rating difference in the sa
  estimate <- imaginary_sample |>
    group_by(genre) |>
    summarize(avg_rating = mean(rating)
    summarise(diff = avg_rating[genre =
    pull(diff)

  differences[i] <- estimate
}
```

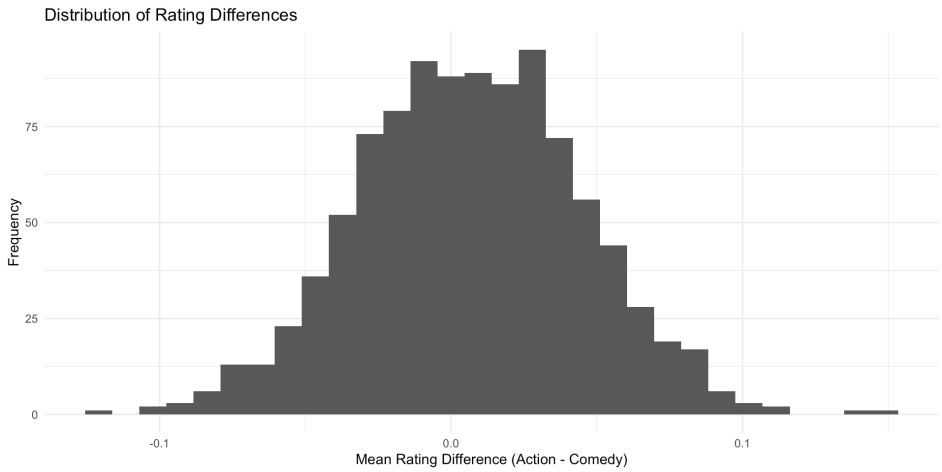
```r
ggplot(data.frame(differenc
  geom_histogram() +
  labs(title = "Distributic
       x = "Mean Rating Dif
       y = "Frequency") +
  theme_minimal()
```



Check $\hat{\delta}$ in the null world

Does the estimate we found in the IMDB data ($\hat{\delta}$ = -0.73) fit well into the world where the true difference δ is 0?

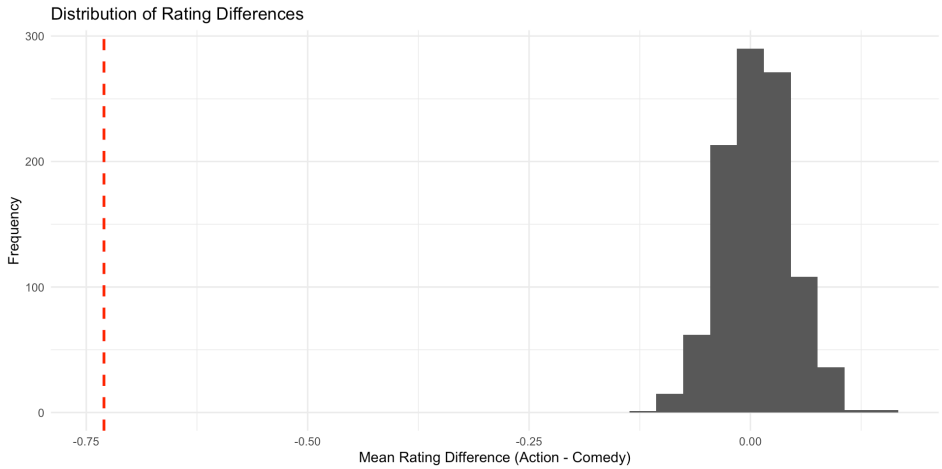
```
ggplot(data.frame(differences), aes(x = difference
  geom_histogram() +
  labs(title = "Distribution of Rating Differences
       x = "Mean Rating Difference (Action - Comed
       y = "Frequency") +
  theme_minimal()
```



Check $\hat{\delta}$ in the null world

Does the estimate we found in the IMDB data ($\hat{\delta}$ = -0.73) fit well into the world where the true difference δ is 0?

```
ggplot(data.frame(differences), aes(x = difference
  geom_histogram() +
  geom_vline(xintercept = -0.73, color = "red", si
  labs(title = "Distribution of Rating Differences
       x = "Mean Rating Difference (Action - Comed
       y = "Frequency") +
  theme_minimal()
```



That seems fairly rare for a null world!

So, again, are action movies better than comedies?

- We can now pretty confidently say that in a world where there is no difference, observing what we observed is super unlikely.
- Therefore, we're pretty confident that in fact there is a difference.
- (We still don't know what the true difference is, but at least we can say it's unlikely to be 0)
- 🎉 Congratulations, if you got that, you got the whole intuition behind hypothesis testing.

All this is good, but how (un)likely **exactly** is it to observe our $\hat{\delta}$ in the null world?

That is where the central limit theorem and theoretical distributions come into play…

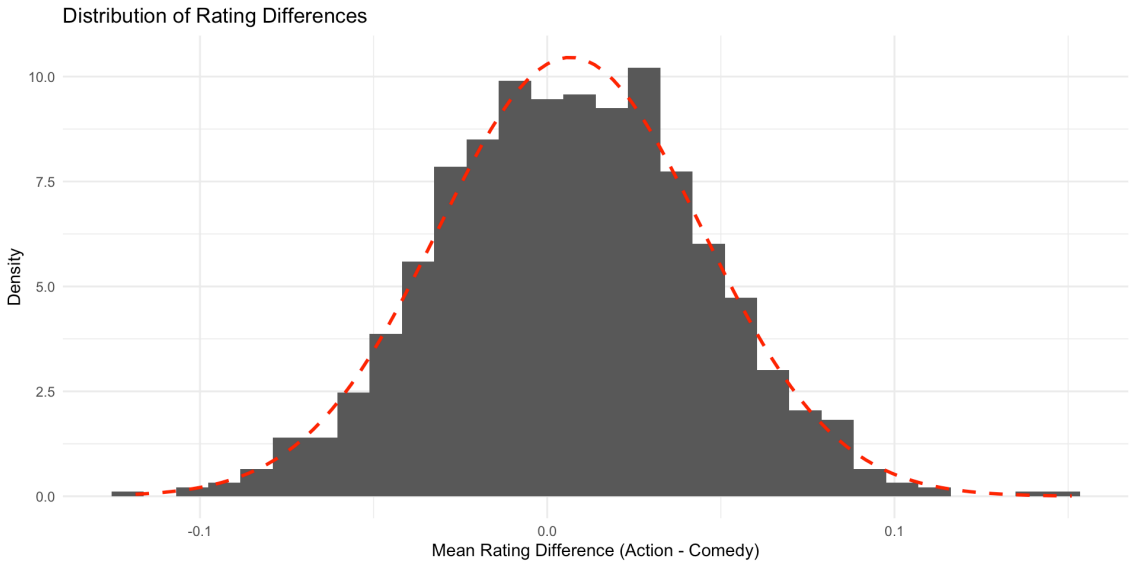
The central limit theorem

You have seen before the estimated mean differences of our imaginary samples (the $\delta\hat{s}$), somehow magically, form a curve that is…

- bell-shaped
- centered around the true value (δ), which in our case was 0.

This distribution of estimates is also called the **sampling distribution**.

The central limit theorem states that, with many observations, the sampling distribution approximates a theoretical distribution - the normal distribution.



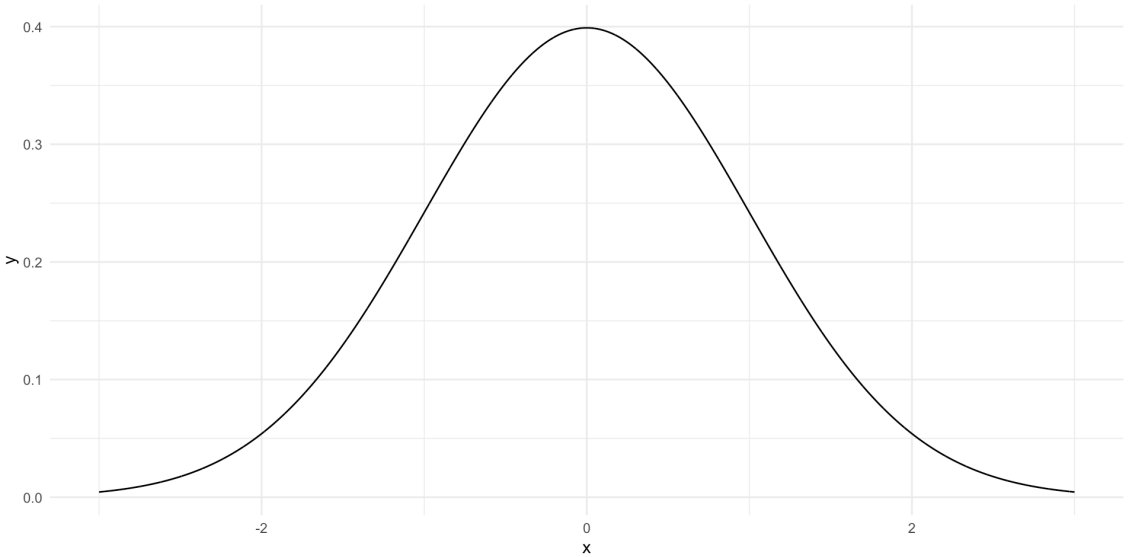
Theoretical distributions

Quick recap

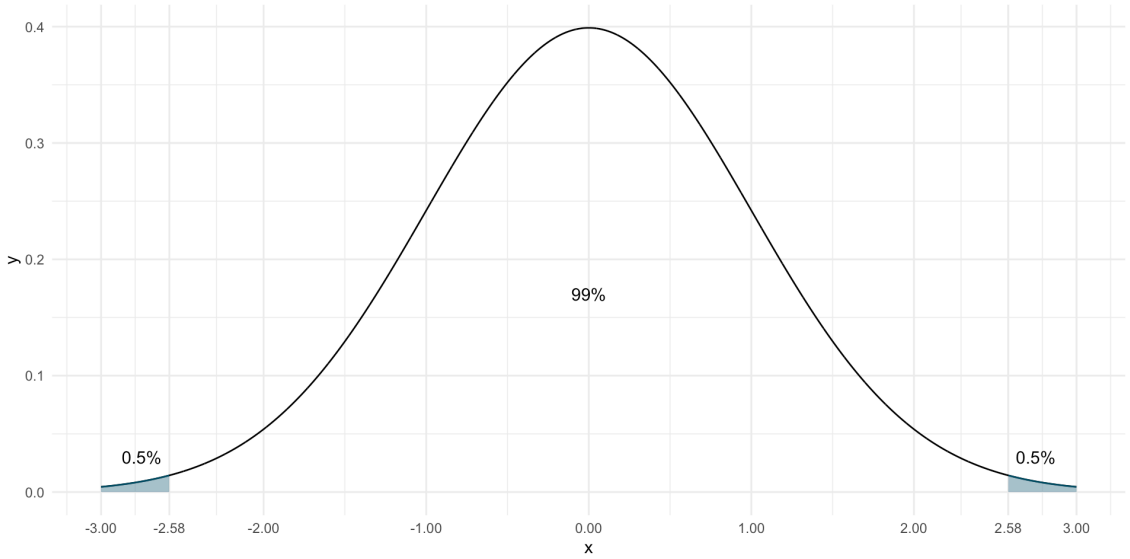
- Remember our problem: We were not sure how (un)likely exactly our observation was in the Null world
- Thanks to the central limit theorem, we know that sampling distributions approximate theoretical distributions.
- And for theoretical distributions, thanks to math, we know exactly how likely a certain value is 🎉

The most famous bell-shaped distribution is the (standard) normal distribution.

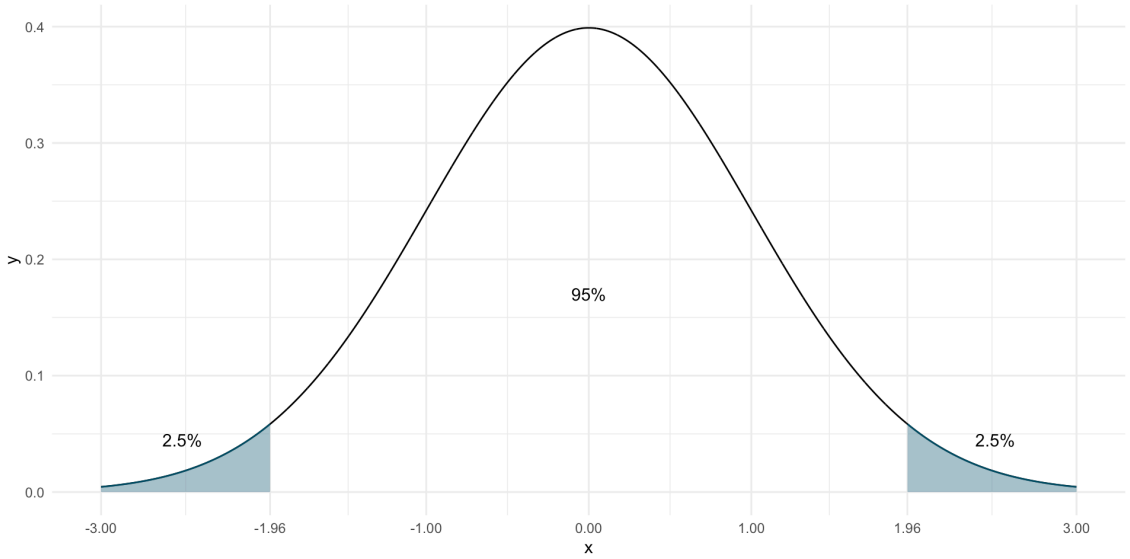
The standard normal distribution is centered around 0 and has a standard deviation of 1.



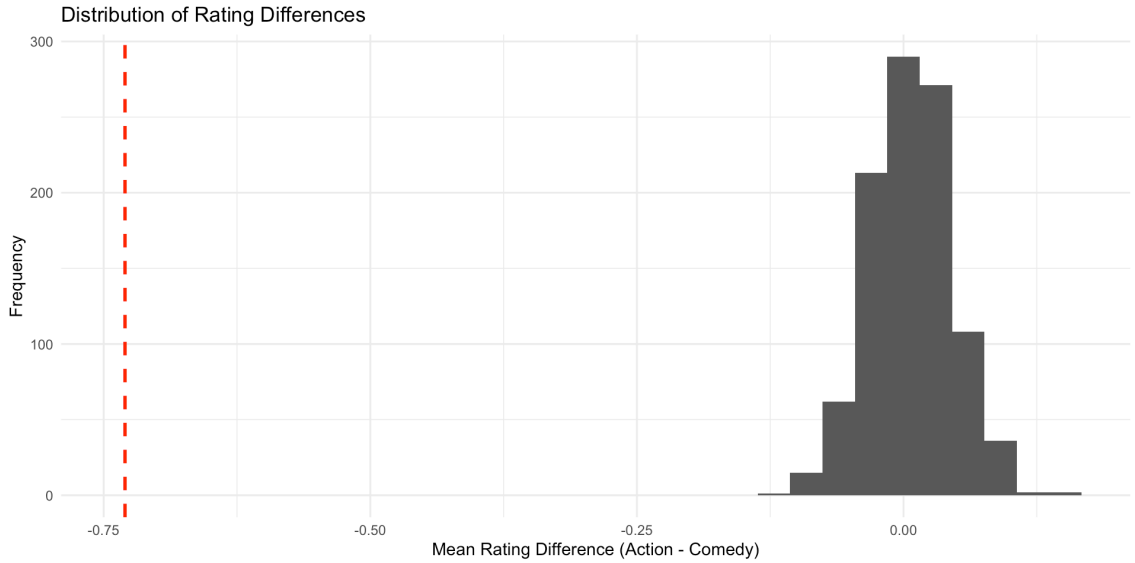
We know, e.g., that 99% of the distribution lie between $\pm$ 2.58



Or that 95% of the distribution lie between $\pm$ 1.96



Now, all we need to do is bring our sampling distribution on the scale of a standard normal distribution.



Now, all we need to do is bring our sampling distribution on the scale of a standard normal distribution.

We achieve this by

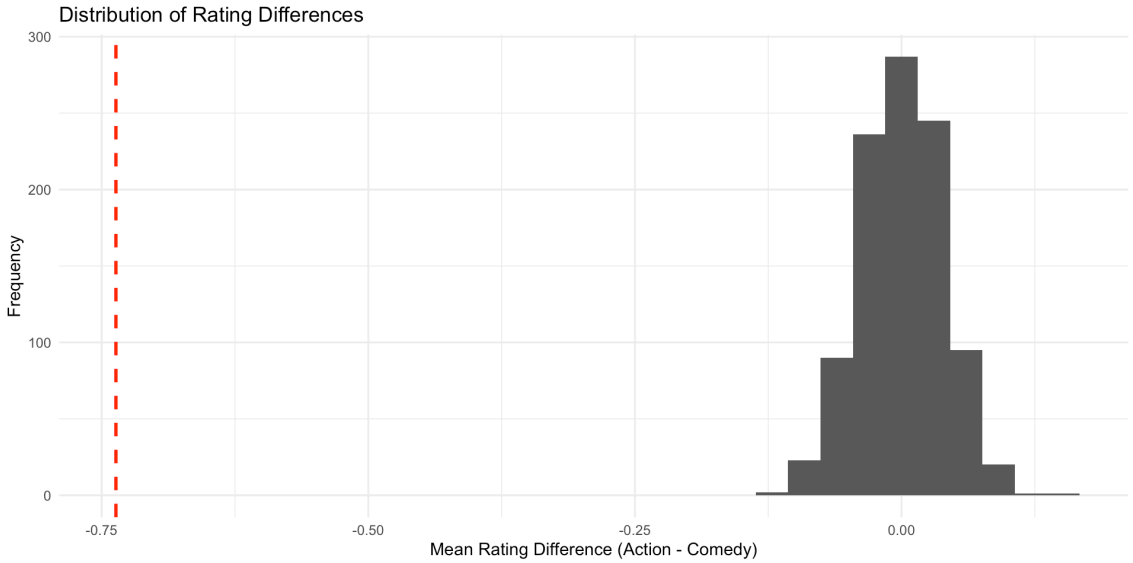
1. Subtracting the mean from all values (in our case, that is 0, so nothing happens)

```
differences_mean_centered <- difference
```

In our case, that is (almost) 0, so not much happens

```
mean(differences)
```

```
[1] 0.006633846
```



Now, all we need to do is bring our sampling distribution on the scale of a standard normal distribution.

We achieve this by

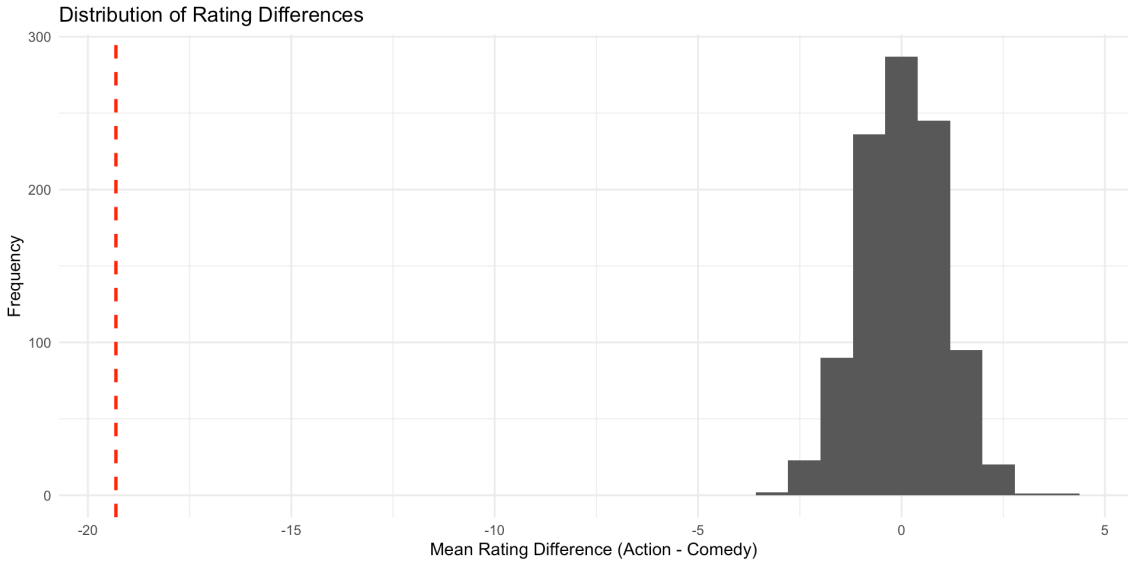
1. Subtracting the mean from all values
2. Dividing by the standard deviation

```
1  differences_scaled <- differences_mean_
```

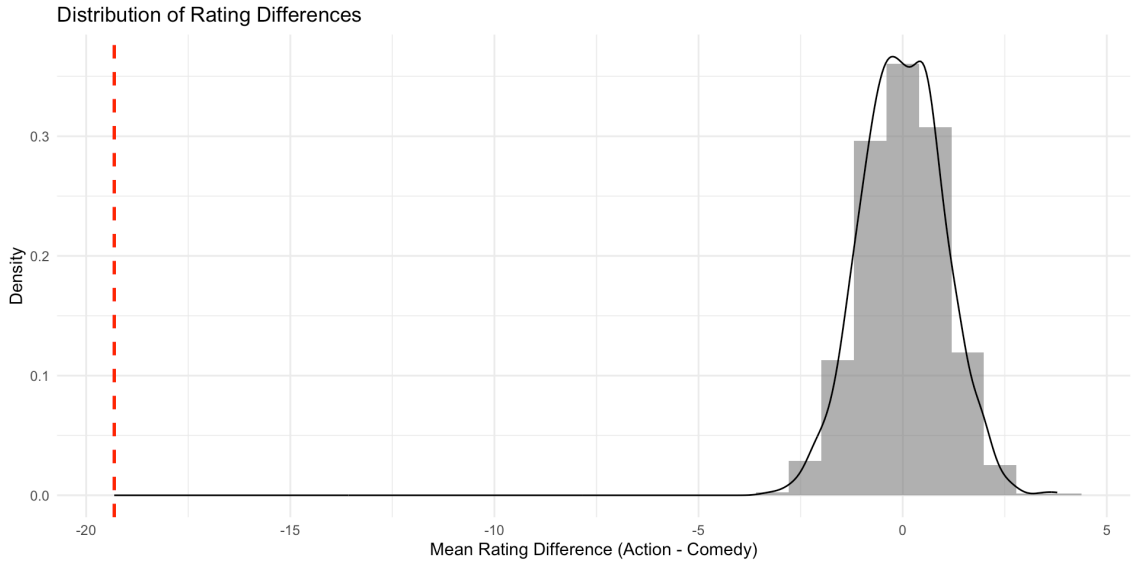
Since the sd is small than 1, our values become bigger

```
1  sd(differences_mean_centered)
```

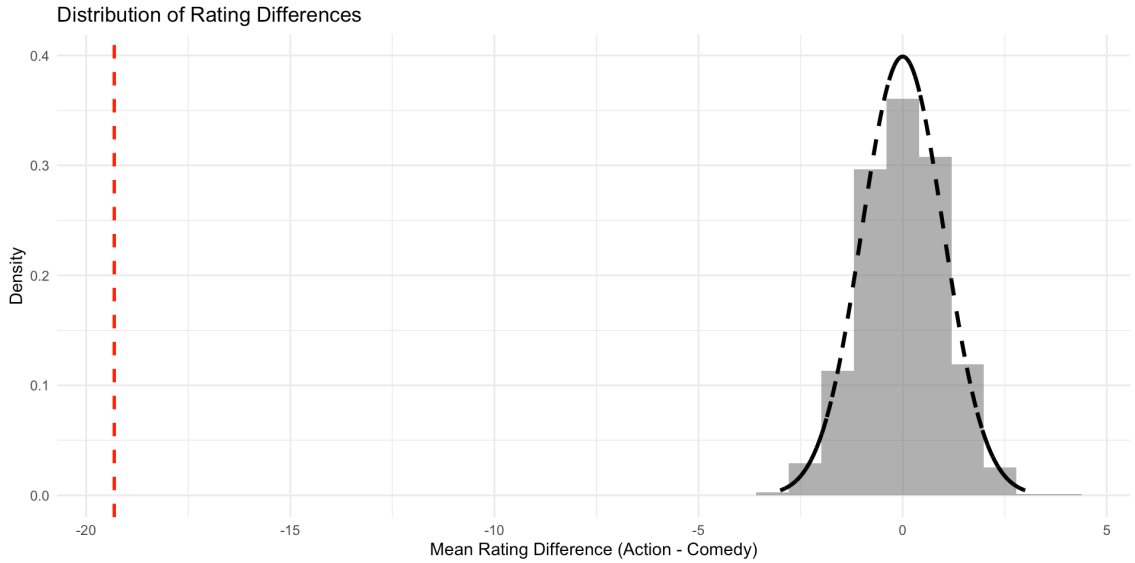
```
[1] 0.03814568
```



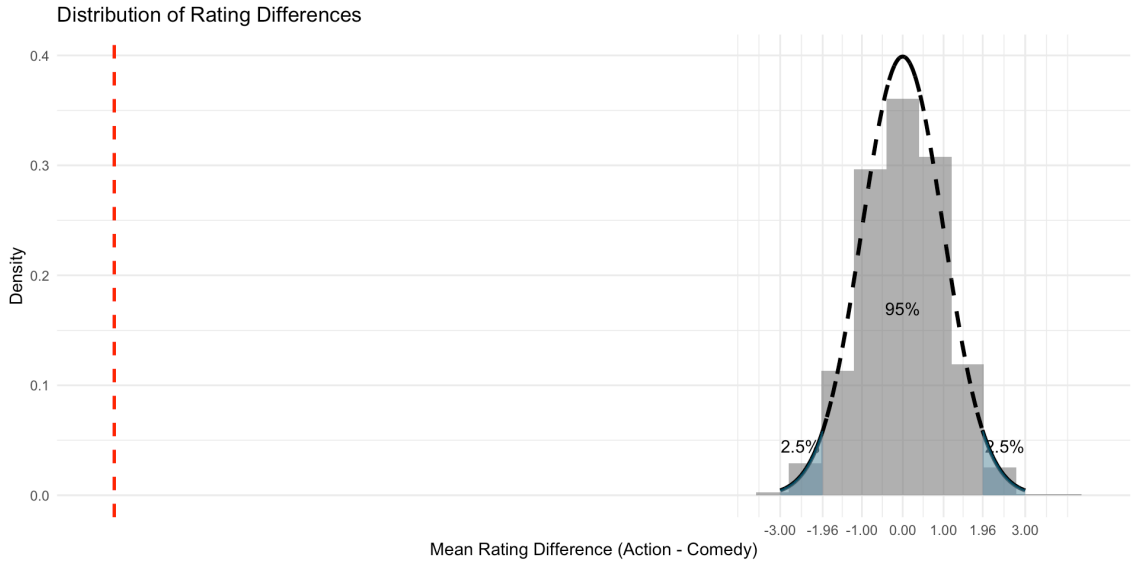
Instead of a histogram, we can use a density plot (which uses the same y-axis as the normal distribution, namely density)



Finally, we can lay over the standard normal distribution



Now we can say for sure that in our Null world, chances that we get an estimate as extreme as the one in our IMBD data is **less than 5%**



You can even calculate the exact probability of observing the estimate in a null world...

1. Bring your estimate on a scale of the standard normal distribution (that is also called a z-value)

```
estimate <- 0.73
sd_sampling_distribution <- sd(differences)

z_scaled_estimate = estimate / sd_sampling_distribution

z_scaled_estimate
```

```
[1] 19.13716
```

> **Note**
>
> You don't need to do a simulation of your sampling distribution all the time. In general, we obtain the standard deviation of the (imaginary) sampling distribution with math. This standard deviation is so important that it has its own name: the **Standard Error (SE)**

You can even calculate the exact probability of observing the estimate in a null world...

1. Bring your estimate on a scale of the standard normal distribution (that is also called a z-value)

```r
estimate <- 0.73
sd_sampling_distribution <- sd(differences)

z_scaled_estimate = estimate / sd_sampling_distribution

z_scaled_estimate
```

```
[1] 19.13716
```

2. Look up the corresponding probability (luckily, in R that's very easy)

```r
# the pnorm() function gives the cumulative probability from the standard normal distribution

# Two-tailed (i.e. a value "at least as extreme as", in both directions)
probability <- 2 * (1 - pnorm(z_scaled_estimate))

# in our case, the probability is reeaally low (practically 0)
probability
```

```
[1] 0
```

Note

In the real world, people actually use a slightly different version of the standard normal distribution, the t-distribution. The principle, however, is the same.

🎉 The probability that you have just calculated is also called p-value 🎉

It’s the probability of observing an estimate at least as extreme as the one in our sample, in a world where there is no true effect (the Null world).

Hypothesis testing in a nutshell

- **Step 1: Calculate an estimate based on your sample ($\hat{\delta}$).**
  This is the main measure you care about: the difference in means, the average, the median, the proportion, the difference in proportions, etc.
- **Step 2: Use simulation to invent a world where the true effect (δ) is null.**
  Simulate what the world would look like if there was no difference between two groups, or if there was no difference in proportions, or where the average value is a specific number.
- **Step 3: Look at $\hat{\delta}$ in the null world.**
  Put the sample statistic in the null world and see if it fits well.
- **Step 4: Calculate the probability that $\hat{\delta}$ could exist in the null world.**
  This is the p-value, or the probability that you'd see a $\hat{\delta}$ at least that high in a world where there's no difference.
- **Step 5: Decide if $\hat{\delta}$ is statistically significant.**
  Choose some evidentiary standard or threshold for deciding if there's sufficient proof for rejecting the null world. Standard thresholds (from least to most rigorous) are 0.1, 0.05, and 0.01.

An applied example

Are action movies better than comedies?

We can use a single command in R to test this hypothesis

```r
# Perform a t-test to compare ratings between Action and Comedy movies
t.test(rating ~ genre, data = movie_data)
```

```
    Welch Two Sample t-test

data:  rating by genre
t = -26.537, df = 5578.2, p-value < 2.2e-16
alternative hypothesis: true difference in means between group Action and group Comedy is not equal to 0
95 percent confidence interval:
 -0.7907698 -0.6819730
sample estimates:
mean in group Action mean in group Comedy
            5.237372             5.973744
```

We get a very small p-value

```r
# Using format() to use non-scientific notation
format(2.2204460493e-16, scientific = FALSE)
```

```
[1] "0.00000000000000022204460"
```

That’s it for today :)