# Testing Errors and Multiple Comparisons

Ryan Miller

Grinnell College
Statistics

# Introduction

Previously, we learned how to use *one-way ANOVA* to evaluate the *global hypothesis*:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$$

The ANOVA $F$-test measures the evidence the sample data provide against this hypothesis by comparing the *sum of squared residuals* for the null model (involving a single mean) and an alternative model (involve group-specific means)

# Introduction (cont.)

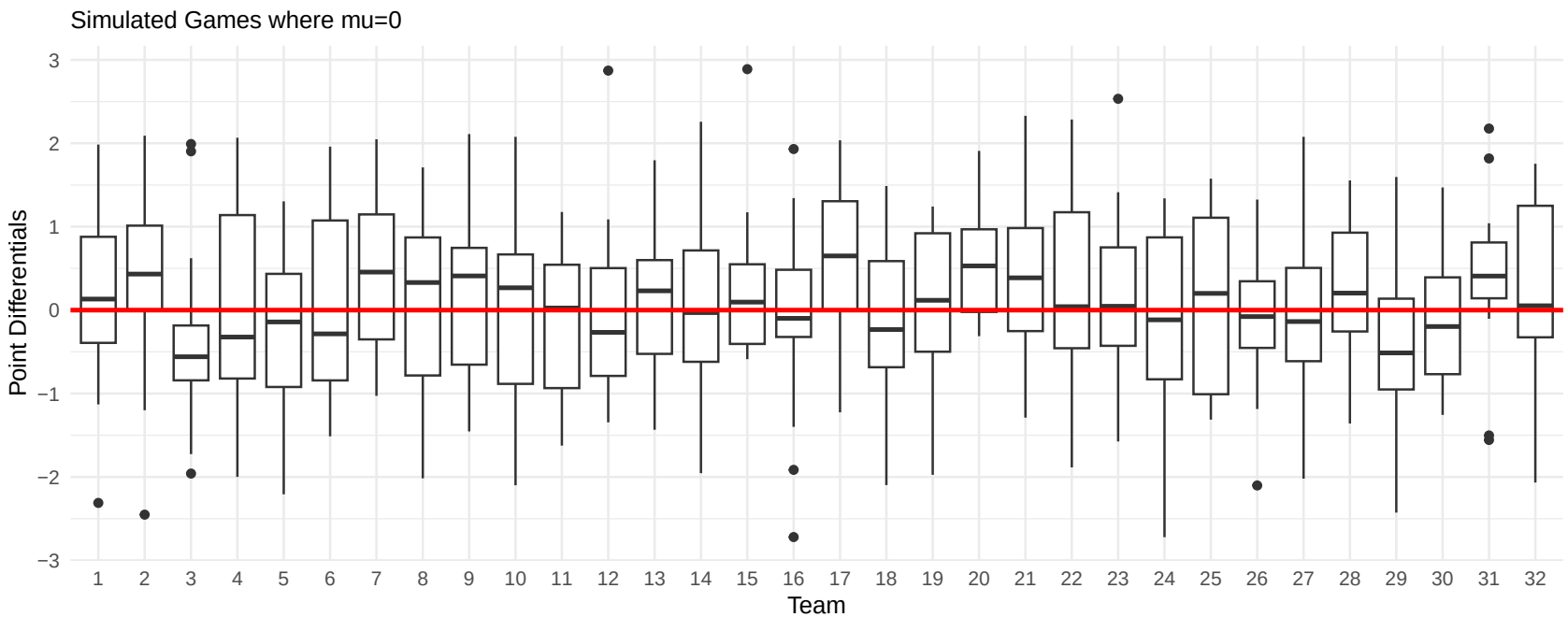
One-way ANOVA provides an answer to whether some groups have different means than other groups, but it *does not* tell us *which groups* are different. For this, we'd need to test numerous *pairwise hypotheses*:

1. $H_0 : \mu_1 = \mu_2$
2. $H_0 : \mu_1 = \mu_3$
3. $H_0 : \mu_2 = \mu_3$
4. ...

However, there are problems with performing a large number of hypothesis tests...

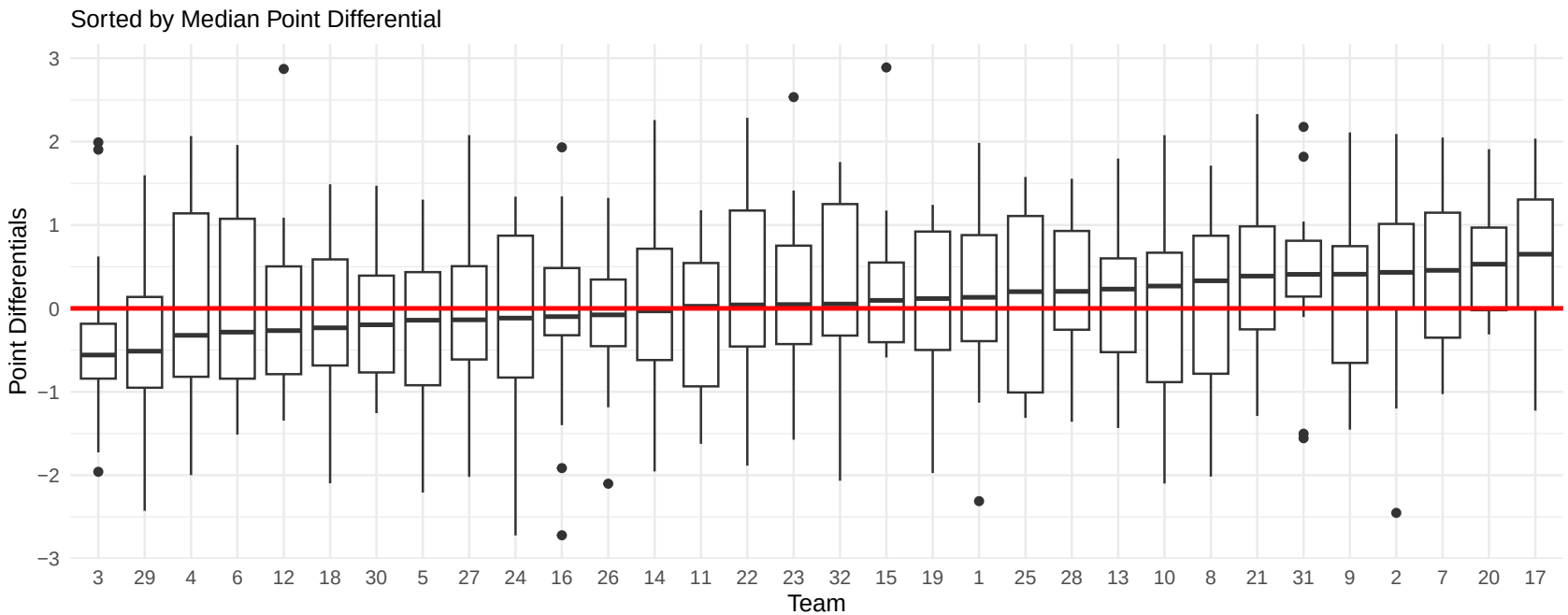
# NFL Example

- Suppose we generate $n = 17$ point differentials for 32 NFL teams such that the *true mean is exactly zero*
  - Considering $H_0 : \mu = 0$ for every team, we have 32 samples of size $n = 17$ where the null hypothesis is true



# NFL Example (cont.)

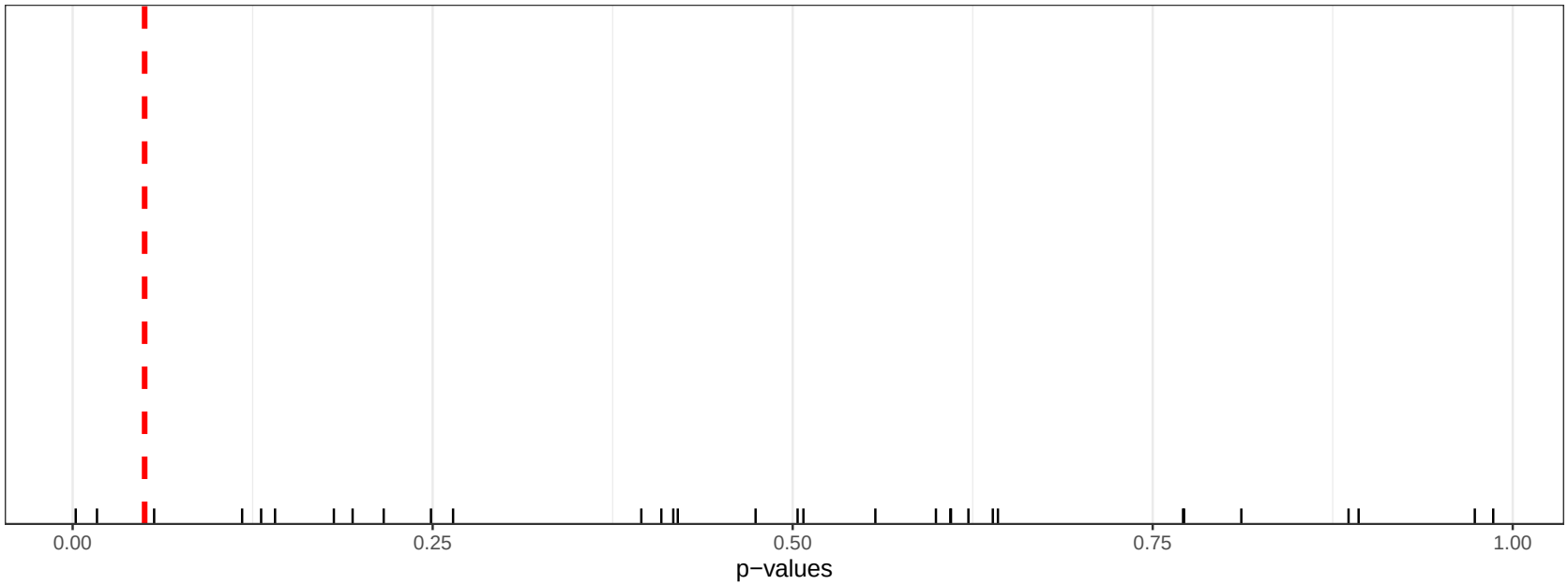
Below is the same simulated data, but sorted by median differential.




**Question**: If a T-test of $H_0 : \mu = 0$ is performed on each team, how many of these tests will have $p$-values less than 0.05?

# NFL Example (cont.)

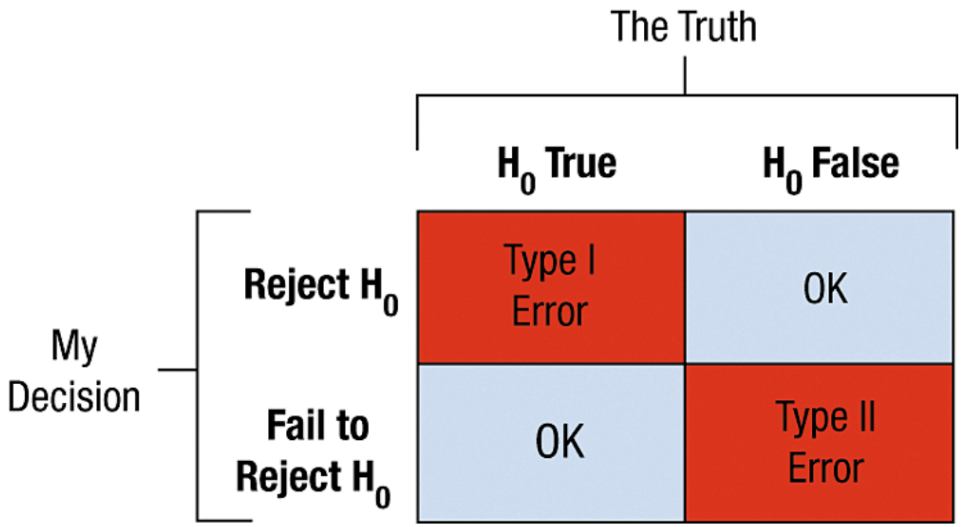
Two of the simulated teams, #20 (p=0.002) and #17 (p=0.017), seemingly provide strong evidence against $H_0 : \mu = 0$:




However, the distribution of $p$-values across all of these simulated teams is uniform...

# Testing Errors

In reality, any conclusion drawn from a hypothesis test may or may not be correct:




- A **type I error** occurs when the null hypothesis is *rejected*, but in reality it is *true*
- A **type II error** occurs when the null hypothesis *cannot be rejected*, but in reality it is *false*

# Testing Errors and Significant Thresholds

- A major reason why $p$-values tend to be compared to a *significance threshold* of $\alpha = 0.05$ is that this procedure will control the *rate of type I errors* to be no more than 5% in circumstances where $H_0$ is true
  - In our NFL simulation, we had 32 samples where $H_0$ was true, and we observed 2/32 produced $p$-values less than 0.05 (type I errors)
    - This isn't surprising, because we expect one type I error for every 20 hypothesis tests performed using $\alpha = 0.05$
- What could we (as statisticians) do to make fewer type I errors?
  - What consequence would this have on the prevelance of type II errors?

# Practice

Jury trials in the US use the premise "innocent until proven guilty". Relating this to hypothesis testing, we can view a trial as a test of $H_0$ : Person A is innocent vs. $H_a$ : Person A is guilty

1) In words, what would a Type I and Type II error each represent in this scenario?
2) Which error would be worse? How might you choose $\alpha$ to be mindful of the trade-off between Type I and Type II errors?

# Practice (solution)

1) A Type I error is convicting an innocent person. A Type II error is letting a guilty person go free.
2) A Type I error should be viewed as worse, so we might set a very strict decision threshold (ie: $\alpha = 0.01$ or even $\alpha = 0.001$). This is what courts actually do, as the standard of "beyond a reasonable doubt" is generally considered to be a very high bar.

# Family-wise Error Rates

- The scientific principle of replication helps prevent the pervasiveness of type I errors in controlled experiments
  - An experiment might have a 5% chance of producing a false positive result (type I error), but the chances of three replications of the experiment each independently producing a type I error is $0.05^3 = 0.000125$, or roughly 1 in 10,000
- Things are more problematic for studies that aren't replicable, particularly when they involve a *family* of related tests, such in our NFL example
  - In this example, we had 32 opportunities to make a type I error, so the chances of at least one false positive result were high

# Controlling the Family-wise Error Rate

We can calculate the probability of *at least one type I error* in 32 independent hypothesis tests where $H_0$ is true and $\alpha = 0.05$:

$$\begin{aligned} Pr(\text{At least one type I error}) &= 1 - Pr(\text{No type I errors}) \\ &= 1 - (1 - 0.05)^{32} = 80.63\% \end{aligned}$$

# Controlling the Family-wise Error Rate

We can calculate the probability of *at least one type I error* in 32 independent hypothesis tests where $H_0$ is true and $\alpha = 0.05$:

$$\begin{aligned} Pr(\text{At least one type I error}) &= 1 - Pr(\text{No type I errors}) \\ &= 1 - (1 - 0.05)^{32} = 80.63\% \end{aligned}$$

This suggests a simple correction to significance threshold: $\alpha^* = \alpha/h$, where $h$ is the number of hypothesis tests being performed:

$$\begin{aligned} Pr(\text{At least one type I error}) &= 1 - Pr(\text{No type I errors}) \\ &= 1 - (1 - 0.05/32)^{32} \approx 5\% \end{aligned}$$

# The Bonferroni Adjustment

Setting $\alpha^* = \alpha/h$ is known as the **Bonferroni Adjustment**. If we apply this to our NFL example, we should compare $p$-values to the adjusted threshold of $\alpha^* = 0.05/32 = 0.0016$ to ensure there is at most of 5% chance of *making any type I errors* across the entire family of tests:

| p-value | signif using 0.05? | signif using 0.0016? |
|---|---|---|
| 0.00213 | yes | no |
| 0.01691 | yes | no |
| 0.05659 | no | no |
| 0.11768 | no | no |
| 0.13089 | no | no |
| 0.14055 | no | no |

# Adjusted *p*-values

- Some applications, such as genetics, frequently use thousands of hypothesis tests
  - Because very small numbers are difficult, you'll commonly see **adjusted p-values** in these studies
  - Bonferroni adjusted *p*-values multiply the original *p*-value by $h$ (the number of tests) and are compared directly with the target family-wise Type I error rate (ie: $\alpha = 0.05$)

| p-value | adjusted p-value | significant? |
|---|---|---|
| 0.00213 | 0.06829 | no |
| 0.01691 | 0.54101 | no |
| 0.05659 | 1.00000 | no |
| 0.11768 | 1.00000 | no |

# Practice

A genetic association study tested for differences in gene expression between two types of leukemia. The study tested 7129 genes.

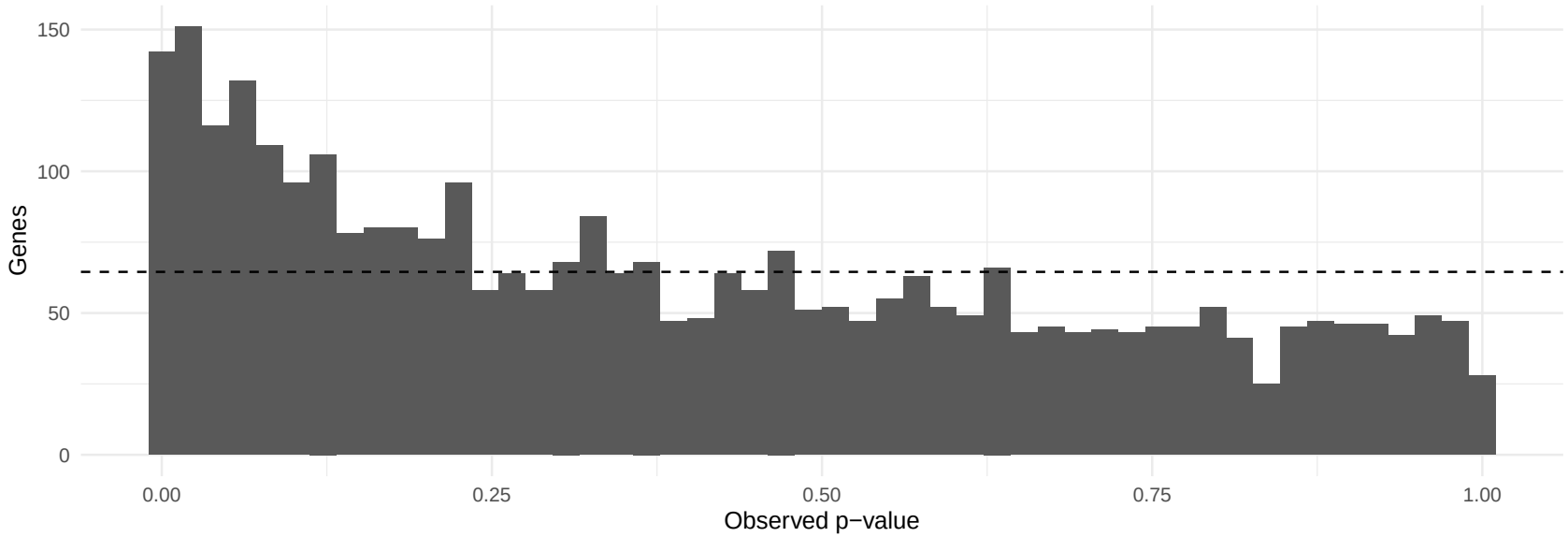
1) If all 7129 tests were done using $\alpha = 0.01$, and there are no genetic differences between these two types of leukemia, how many "statistically significant" genes would be expected?
2) Suppose 783 genes had $p$-values less than 0.01, do you believe there is an association between some genes and the type of leukemia?
3) Suppose you wanted to use the Bonferroni adjustment to ensure a Type I error rate no larger than 5%. What would your adjusted significance threshold be?
4) Suppose the "most significant" gene had a $p$-value of 0.000001, what is its *Bonferroni Adjusted p-value*?

# Practice (solution)

1) You'd expect $7129 * 0.01 = 71$ Type I errors
2) Yes, there were over 10 times (712) more significant results than expected
3) $\alpha^* = 0.05/7129 = 0.000007$
4) The adjusted $p$-value is $0.000001 * 7129$, or $p^* = 0.007$

# False Discovery Rates

A genomics study measured the expression levels of 17,322 genes to identify genes that are co-expressed with BRCA1, a gene that is well-known to be associated with breast cancer. For each gene a hypothesis test was performed, and the $p$-values of these tests are displayed using a histogram:



# False Discovery Rates

- Suppose we apply the Bonferroni adjustment to control the family-wise type I error rate at 10%
  - $\alpha^* = 0.1/3226 = 0.00003$
  - The study yields 2 statistically significant genes (with $p$-values less than 0.00003)
- Suppose we seek to control the false discovery rate at 10%
  - This isn't as easy to do "by hand", but the procedure in R identifies 24 genes
  - Among these 24 genes we'd expect 2 or 3 to be false positives

# Post-hoc Testing and ANOVA

- In our previous lab on ANOVA, we used the `TukeyHSD()` function to perform pairwise tests
  - This approach applies a method family-wise type I error control (similar to Bonferroni) but is slightly more powerful due to its exclusive focus on differences in means
  - In the `R` output below you should notice adjusted $p$-values are reported

```
##                 diff       lwr      upr     p adj
## ATL-ARI 17.1666667 -15.57986 49.91320 0.9806062
## BAL-ARI 16.9166667 -21.07713 54.91046 0.9984036
## BUF-ARI 13.1666667 -18.62122 44.95455 0.9995627
## CAR-ARI  0.9666667 -34.67467 36.60800 1.0000000
## CHI-ARI 16.4523810 -16.29415 49.19891 0.9892498
```

# Conclusion

Below are the main ideas you should understand:

1. Statistical tests do not provide definitive conclusions, your decision might be a type I or type II error
2. The chances of making at least one type I error increase dramatically as you perform more hypothesis tests within a single study, but this can be corrected for using family-wise error rate methods or false discovery rate methods.
3. False discovery rate methods are *less strict* as they allow a certain percentage of findings to be false discoveries, but this has the benefit of reducing the type II error rate
4. The pairwise $p$-values from post-hoc testing functions like `TukeyHSD()` are adjusted to control the family-wise error rate. We could use `p.adjust()` to control the false discovery rate instead.