Tim Book

# A/B Testing

GENERAL ASSEMBLY

# Recap of Experimental Design

Last time, given some research questions, we discussed how we could control for sources of variation and design a method of **collecting the right data to answer those questions, the right way**.

# Recap of Experimental Design... from a DS perspective

You might not expect to find yourself in a role where you're conducting surveys or administering scientific experiments. And thus you might think experimental design is outside the realm of interest, but it is not.

Knowing where your data came from is Step 0 to any project involving data.

# Recap of Experimental Design... from a DS perspective

After all... we are very accustomed to collecting data via **web scraping**. Have you considered all the sources of variation **before** scraping?

- Different social media sites might have differing sentiments towards a certain topic.
- Maybe if someone uses a certain hashtag, they have a different sentiment?
- Do verified accounts act differently that non-verified ones?

Answering these questions may be a task for **A/B testing**.

# What is A/B Testing?

Now that our data’s been collected, we’re going to discuss how to summarize our findings with **A/B testing**.

We’ve already learned a little about **hypothesis testing**, a category of statistical techniques that give us a mathematically rigorous way of deciding whether or not two or more categories differ with respect to some measurement.

# What is A/B Testing?

A/B testing is more of a business term that doesn’t have a single agreed-upon definition. Most sources largely equate the term A/B testing with hypothesis testing, so we will too.

**However, A/B testing is usually discussed in the context of conducting an experiment**. The “A” and “B” refer to a treatment and control group that we would test the difference between.

Recap of Hypothesis Testing

# Hooray!

# Definition of Hypothesis Testing

Hypothesis testing is the **scientific** act of using **statistics** to determine if your hypothesis is true within probabilistic reason.

You begin by specifying two hypothesis:

- **The null hypothesis:** The conventional wisdom you seek to disprove
- **The alternative hypothesis:** The exciting finding you wish to prove

# Oversimple Example

I wish to prove that a coin is unfair.

$$H_0 \ : \text{the coin is fair}$$

$$H_A \ : \text{the coin is unfair}$$

# Oversimple Example

I wish to prove that a coin is unfair.

$$H_0 : p = 0.5$$

$$H_A : p \neq 0.5$$

# Oversimple Example

I now begin flipping this coin and recording my results:

**HHHTH HHTHH HTHHH THHHH**

$$\hat{p} = 0.20$$

Is the coin fair?

# Oversimple Example

I now begin flipping this coin and recording my results:

**HHH**

$\hat{p} = 1.00$

Is the coin fair?

# Steps for Conducting an Experiment

1. *Define the objectives of the experiment*
2. *Identify all sources of variation*
3. *Choosing a rule for assigning units to treatments*
4. *Decide on the measurement to be made, experimental procedure, and* how you plan to analyze results

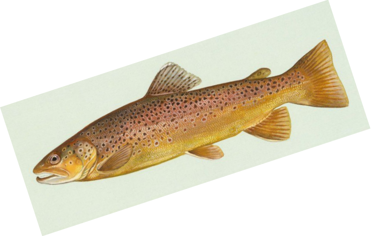
**EDA + A/B Testing**

# Example: *t*-test

Brown trout were placed into two separate tanks (10 fish each) and fed different concentrations of sulfamerazine with their food. After 35 days, their hemoglobin levels were recorded.

$$\bar{y}_1 = 7.20$$
$$\bar{y}_2 = 8.69$$

# Example: *t*-test



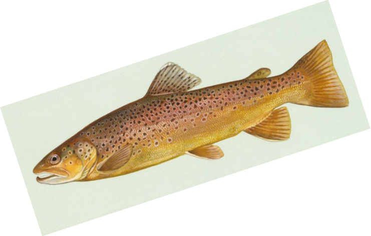
We'll conduct a **two-sample *t*-test**, which can determine whether or not two differing **population means** differ based on a sample.

$$H_0 : \mu_1 = \mu_2$$
$$H_A : \mu_1 \neq \mu_2$$

$$\bar{y}_1 = 7.20$$
$$\bar{y}_2 = 8.69$$

Based on these summary statistics, do the two hemoglobin levels differ?

# Example: *t*-test

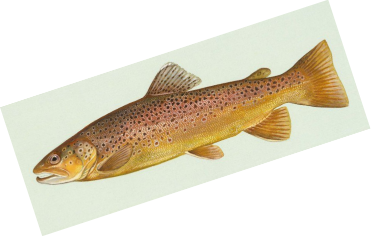


$$H_0 : \mu_1 = \mu_2$$
$$H_A : \mu_1 \neq \mu_2$$

$$t = -3.3$$
$$p\text{-val} = 0.004$$

***Yes!*** Remember, ***p-value's low, H₀'s gotta go!*** Because our *p*-value was low, we reject our null hypothesis and conclude that the sulfamerazine concentration causes differing hemoglobin levels.

# Example: *t*-test



$$H_0 : \mu_1 = \mu_2$$
$$H_A : \mu_1 \neq \mu_2$$

$$t = -3.3$$
$$p\text{-val} = 0.004$$

***Yes!*** Remember, ***p-value's low, $H_0$'s gotta go!*** Because our *p*-value was low, we reject our null hypothesis and conclude that the sulfamerazine concentration **causes** differing hemoglobin levels.

**Can I say this?!**

# Example continued: A New Perspective

Hemoglobin

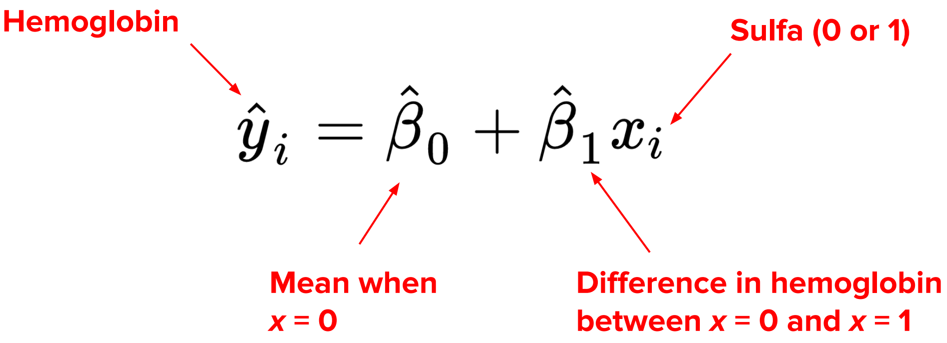
Sulfa (0 or 1)

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

???

???

# Example continued: A New Perspective



$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Hemoglobin

Sulfa (0 or 1)

Mean when
*x* = 0

Difference in hemoglobin
between *x* = 0 and *x* = 1

# Example continued: A New Perspective

Hemoglobin

Sulfa (0 or 1)

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Mean when
*x* = 0

Difference in hemoglobin
between *x* = 0 and *x* = 1

$$H_0 : \mu_1 = \mu_2$$
$$H_A : \mu_1 \neq \mu_2$$

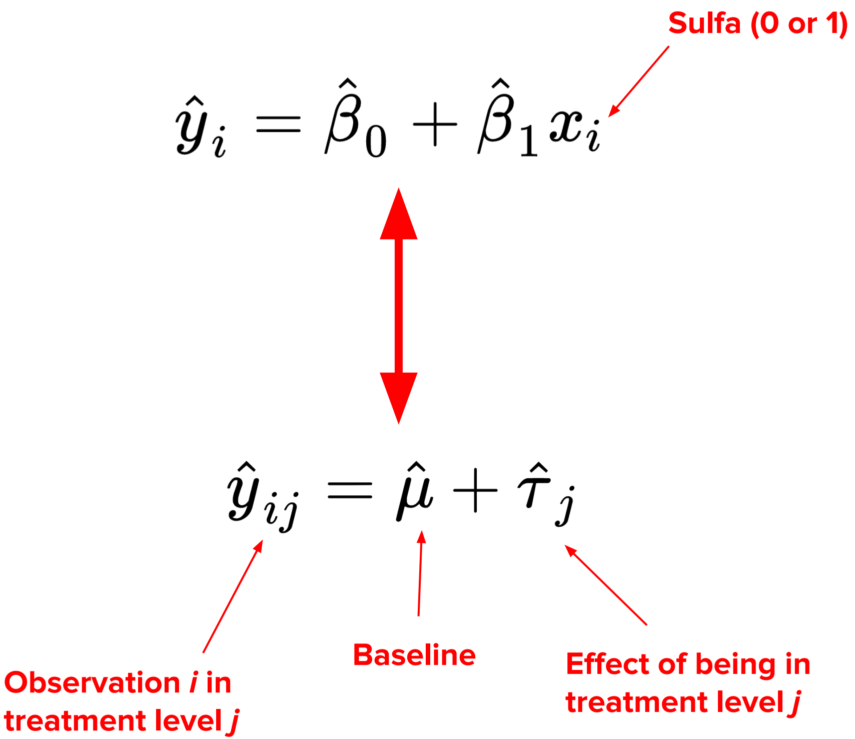
Logically equivalent!

$$\longleftrightarrow$$

*"Does sulfa concentration affect hemoglobin levels?"*

$$H_0 : \beta_1 = 0$$
$$H_A : \beta_1 \neq 0$$

# Example continued: A New Perspective



Sulfa (0 or 1)

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$\hat{y}_{ij} = \hat{\mu} + \hat{\tau}_j$$

Observation *i* in treatment level *j*

Baseline

Effect of being in treatment level *j*

The mother of all hypothesis tests...

# ANOVA Tests

# ANOVA

**ANOVA (Analysis of Variance)** is sometimes referred to as a hypothesis test, but it's actually an entire category of hypothesis testing.

An ANOVA test tests whether or not two or more means are equal.

The following are the hypotheses of a **One-Way ANOVA**:

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

$$H_A : \mu_i \neq \mu_j \text{ for some } i \neq j$$

# ANOVA

The following are the hypotheses of a **One-Way ANOVA** of three means:

$$H_0 : \mu_1 = \mu_2 = \mu_3$$
$$H_A : \mu_i \neq \mu_j \text{ for all } i \neq j$$

Why is this not the same as saying

$\mu_1 \neq \mu_2 \neq \mu_3$ ?

# Example: Soap

In this experiment, experimenters wish to determine the rate at which three different soaps dissolve in water. Soaps are cut into cubes, weighed, and put into separate portions of a muffin tin and then filled with water. After a specified amount of time, the cubes are removed and weighed.

**Soap types:**

```
1 = Regular
2 = Deodorant
3 = Moisturizing
```

# Example: Soap

jupyter To the notebook!

# Results: ANOVA Test

We conducted this test for the mean weight loss of 3 different soap types:

$$H_0 : \mu_1 = \mu_2 = \mu_3$$
$$H_A : \mu_i \neq \mu_j \text{ for all } i \neq j$$

And got these results:

$$F = 104.45, p\text{-val} = 5.9 \times 10^{-7}$$

# Results: ANOVA Test

$$F = 104.45, p\text{-val} \ = 5.9 \times 10^{-7}$$

Since our p-value was low, we reject our null hypothesis and conclude that the three soaps exhibit different weight losses when in contact with water.

Which differ, and how? **More on this soon.**

# Results: ANOVA Table

The result of an ANOVA test is actually an **ANOVA table**. Being able to read an ANOVA table is a critical statistical skill. The **one-way ANOVA table** looks like this:

| | *d.f.* | Sum Sq | Mean Sq | $F$ | $p$-val |
|---|---|---|---|---|---|
| Treatment | $k$ - 1 | $\sum(\hat{y}_i - \bar{y})^2$ | $\frac{SSTR}{k-1}$ | $\frac{MSTR}{MSE}$ | $P(F_{k-1,n-k} > F)$ |
| Error | $n$ - $k$ | $\sum(y_i - \hat{y}_i)^2$ | $\frac{SSE}{n-k}$ | | |
| Total | $n$ - 1 | $\sum(y_i - \bar{y})^2$ | | | |

# Results: ANOVA Table

The result of an ANOVA test is actually an **ANOVA table**. Being able to read an ANOVA table is a critical statistical skill. The **one-way ANOVA table** looks like this:

| | $d.f.$ | Sum Sq | Mean Sq | $F$ | $p$-val |
|---|---|---|---|---|---|
| Treatment | $k$ - 1 | $\sum(\hat{y}_i - \bar{y})^2$ | $\frac{SSTR}{k-1}$ | $\frac{MSTR}{MSE}$ | $P(F_{k-1,n-k} > F)$ |
| Error | $n$ - $k$ | $\sum(y_i - \hat{y}_i)^2$ | $\frac{SSE}{n-k}$ | | |
| ~~Total~~ | ~~$n$ - 1~~ | ~~$\sum(y_i - \bar{y})^2$~~ | | | |

**Often omitted in software**
**(it's the sum of all above rows anyway)**

# Results: ANOVA Table

The result of an ANOVA test is actually an **ANOVA table**. Being able to read an ANOVA table is a critical statistical skill. The **one-way ANOVA table** looks like this:

| | *d.f.* | Sum Sq | Mean Sq | $F$ | $p$-val |
|---|---|---|---|---|---|
| Treatment | $k$ - 1 | $\sum(\hat{y}_i - \bar{y})^2$ | $\frac{SSTR}{k-1}$ | $\frac{MSTR}{MSE}$ | $P(F_{k-1,n-k} > F)$ |
| Error | $n$ - $k$ | $\sum(y_i - \hat{y}_i)^2$ | $\frac{SSE}{n-k}$ | | |

**Error between our model and truth**

**Difference between our model and null model**
**(ie, high if our model is doing a good job)**

**Ratio should be high if our treatment was effective!**

# ANOVA’s Dilemma

After looking at graphs, statistics, and doing an ANOVA test, you may now be tempted to say that one mean is higher than another mean - but technically you can’t! You didn’t test this!

You may then be tempted to perform **individual two-sample *t*-tests**, but you need to be careful! Why?

# Think about it...

Suppose that, on any given day, there is a 5% chance of rain. Which is most likely?

**A**

It will not rain Monday.

**B**

It will not rain Monday, Tuesday, Wednesday, Thursday, ***and*** Friday.

# ANOVA’s Dilemma

You’re doing the same thing with every hypothesis test you do. You’re free to do these tests, but you need to account for these **multiple comparisons** with some **multiple comparisons correction**.

# Multiple Comparisons

The simplest (and therefore the most conservative) is the **Bonferroni multiple comparisons correction**. If you’re carrying out *m* hypothesis tests, simply replace:

$$\alpha \mapsto \frac{\alpha}{m}$$

*Carlo Emilio Bonferroni*

Real talk

# More Advanced ANOVAs (if time)

# Example: Complete Two-Way ANOVA

An experiment was done to measure the effect of different factors on individuals' reaction times. Respondents were given a **cue**, warning them the stimulus was coming it was either auditory or visual. After the cue, the stimulus would arrive in either 5, 10, or 15 seconds. Their reaction to the stimulus was recorded.

***Response*** = Reaction time (seconds)
***Treatment A*** = Cue (1: auditory, 2: visual)
***Treatment B*** = Wait time (1: 5sec, 2: 10sec, 3: 15sec)

# Example: Complete Two-Way ANOVA

This corresponds to the following linear model equation:

$$\hat{y}_{ijk} = \hat{\mu} + \hat{\alpha}_j + \hat{\beta}_k + (\widehat{\alpha\beta})_{jk}$$

# Example: Complete Two-Way ANOVA

This corresponds to the following linear model equation:

$$\hat{y}_{ijk} = \hat{\mu} + \hat{\alpha}_j + \hat{\beta}_k + (\widehat{\alpha\beta})_{jk}$$

**Baseline**

**Effect of cue type *j***

**Effect of wait time *k***

**Interaction. Do different cues affect the response at different wait times? This term is what makes this model “complete” - and it is very much the norm when building these models!**

# Example: Complete Two-Way ANOVA

To the notebook!

# Example: Complete Two-Way ANOVA Table

| | *d.f.* | Sum Sq | Mean Sq | *F* | *p*-val |
|---:|:-:|:-:|:-:|:-:|:-:|
| Cue | 1 | 0.023544 | 0.023544 | 81.375 | 0.000001 |
| Wait Time | 2 | 0.001158 | 0.000579 | 2.0013 | 0.177799 |
| Cue x Wait | 4 | 0.000846 | 0.000423 | 1.4626 | 0.270135 |
| Error | 12 | 0.003472 | 0.000289 | | |

# Example: Complete Two-Way ANOVA Table

| | *d.f.* | Sum Sq | Mean Sq | *F* | *p*-val |
|---|---|---|---|---|---|
| **Cue** | **1** | **0.023544** | **0.023544** | **81.375** | **0.000001** |
| Wait Time | 2 | 0.001158 | 0.000579 | 2.0013 | 0.177799 |
| Cue x Wait | 4 | 0.000846 | 0.000423 | 1.4626 | 0.270135 |
| Error | 12 | 0.003472 | 0.000289 | | |

Since our *p*-value was low, we can reject the null and conclude that the cue type affects reaction time.

# Example: Complete Two-Way ANOVA Table

| | *d.f.* | Sum Sq | Mean Sq | *F* | *p*-val |
|---|---|---|---|---|---|
| Cue | 1 | 0.023544 | 0.023544 | 81.375 | 0.000001 |
| **Wait Time** | **2** | **0.001158** | **0.000579** | **2.0013** | **0.177799** |
| Cue x Wait | 4 | 0.000846 | 0.000423 | 1.4626 | 0.270135 |
| Error | 12 | 0.003472 | 0.000289 | | |

Since our *p*-value was above 0.05, we fail to reject the null and cannot conclude that the wait time affects reaction time.

# Conclusions

- A/B testing is the rule of law when stating statistical “facts”
- Pretty much all A/B tests can also be written as linear models, and thus any *F*-tests of the same hypotheses will be equivalent
- ANOVA tests lie at the heart of most statistical tests, and hence many real-life experiments boil down to some sort of *F*-test.
- Consequently, there are many, many types and variants of the ANOVA.
- **Being able to read basic ANOVA tables might be the #1 takeaway from this whole lesson!**

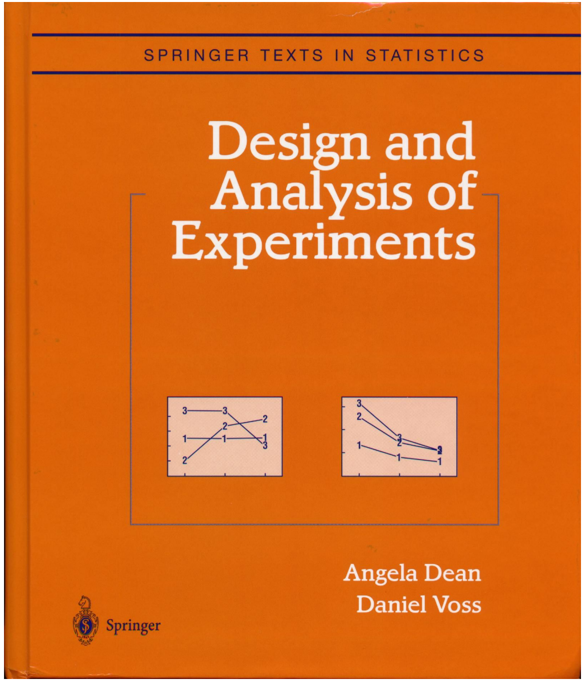
# Credit where credit is due

## Angela Dean

Emeritus Professor, OSU

## Dan Voss

Professor, Wright State



SPRINGER TEXTS IN STATISTICS

Design and Analysis of Experiments

Angela Dean
Daniel Voss

Springer