Multimedia ANOVA Testing for Website Conversions

Carsten Lange

1 Introduction

Imagine you are in charge of optimizing a company's website. The goal is simple: increase conversions — defined here as users clicking on the ordering form. The marketing team proposes a solution: **Enhance the website with either audio or video content to better engage visitors.**
But will these multimedia upgrades actually drive more clicks, or will the effort fall flat?

To answer this question, the company decides to run a controlled experiment. Every website visitor is randomly redirected to one of three versions of the site:

- the original (**Control**),
- one enhanced with **Audio**, or
- one enhanced with **Video**.

Afterward, the team tracks hourly conversions for each of the three scenarios (*Control*, *Audio*, *Video*). The result generated a randomized dataset, ideal for testing whether the *Audio*, *Video* treatments have a measurable impact on user behavior.

In this article, we will walk through a simulated version of this marketing experiment using artificial data adapted from the well-known Palmer Penguins dataset. The setup mirrors how real companies might structure and evaluate a digital content strategy.

2 Experimental Design

To assess the impact of multimedia content on user behavior, the company conducted a randomized test involving three website variants:

- **Control**: the original site with no enhancements

- **Audio**: the same site augmented with audio content
- **Video**: the site upgraded with embedded video content

Every visitor arriving at the website was randomly assigned to one of these three groups. The primary metric of **conversion** was defined as whether users clicked on the ordering form. Each record reflects one hour of user data for the three scenarios leading to a total of 342 observations.

3 Data Generation and Structure

To simulate this experiment, a proxy dataset was created using the well-known **Palmer Penguins** dataset. In this synthetic version:

- The variable **flipper length** was reinterpreted to represent the **number of conversions** within a one-hour period.
- The **species** of the penguins was repurposed as the **treatment group** (**Control**, **Audio**, or **Video**).

Although artificial, this approach allows us to explore realistic patterns in conversion behavior under different digital content strategies.

The code below shows how the Palmer Penguin dataset was used to generate the dataset and which R libraries were used:

```
library(knitr)
#library(kableExtra)
library(tidyverse)
library(palmerpenguins)

DataWeb=penguins |>
       select(Conversions=flipper_length_mm, Treatment=species) |>
       drop_na()
levels(DataWeb$Treatment)=c("Control", "Audio", "Video")
```

Each observation shows for different web design scenarios how many conversions resulted within 1 hour. The first 6 records are shown in Table 1:

```
set.seed(123)
knitr::kable(sample_n(DataWeb,6))
```

Table 1: First Six Observations from the Dataset

Conversions	Treatment
215	Video
198	Control
216	Video
201	Audio
189	Control
195	Audio

4 Group and Overall Means

The code below generates group means for the *Control*, *Audio*, and *Video* groups, together with the overall mean (*GrandMean*) across all groups and outputs the results in Table 2:

```
GrandMean  = mean(DataWeb$Conversions)

# Conditional Means by Treatment
MeansByTreatment <- DataWeb %>%
  group_by(Treatment) %>%
  summarise(Mean = mean(Conversions), N=n())|>
  bind_rows(tibble(
    Treatment = "Overall",
    Mean = GrandMean,
    N = 342
  ))
N=342
# View results
knitr::kable(MeansByTreatment)
# |>
#    kable_styling(full_width = FALSE, position = "center")%>%    scroll_box(width = "400px"
```

Table 2: Overall Mean and Means of the Treatments

Treatment	Mean	N
Control	189.9536	151
Audio	195.8235	68
Video	217.1870	123
Overall	200.9152	342

At a glance, both multimedia-enhanced designs resulted in a higher average number of conversions per hour compared to the control group.

Notably, the *Video* variant had the highest mean, with an average of 217.19 conversions — nearly 27 more than the Control.

The Audio variant also outperformed the Control, though with a smaller gap.

The Control group had the lowest mean but the largest sample size.

The Grand Mean of 200.92 provides a useful benchmark for understanding overall performance across all treatments, especially when considering statistical comparisons or variance analysis.

Figure 1 provides additional evidence for this impression:

```
ggplot(DataWeb, aes(y=0, x = Conversions, color = Treatment)) +
  geom_jitter(width = 0.1) +
  geom_vline(xintercept = MeansByTreatment[[1,2]], color = "red",linewidth=0.7) +
  geom_vline(xintercept = MeansByTreatment[[2,2]], color = "green",linewidth=0.7) +
  geom_vline(xintercept = MeansByTreatment[[3,2]], color = "blue",linewidth=0.7) +
  geom_vline(xintercept = mean(DataWeb$Conversions), color = "black",linewidth=1)+
  labs(y = "") +              # Set y-axis title
  scale_y_continuous(breaks = NULL)+
  annotate("text", x = MeansByTreatment[[1,2]], y = 0.37,
    label = round(MeansByTreatment[[1,2]],2), color = "red", angle = 90,
    vjust = -0.5, size = 3)+
  annotate("text", x = MeansByTreatment[[2,2]], y = 0.37,
    label = round(MeansByTreatment[[2,2]],2), color = "green", angle = 90,
    vjust = -0.5, size = 3) +
  annotate("text", x = MeansByTreatment[[3,2]], y = 0.37,
    label = round(MeansByTreatment[[3,2]],2), color = "blue", angle = 90,
    vjust = -0.5, size = 3) +
  annotate("text", x = mean(DataWeb$Conversions), y = 0.37,
    label = round(mean(DataWeb$Conversions),2), color = "black", angle = 90, vjust = -0.5, si
```

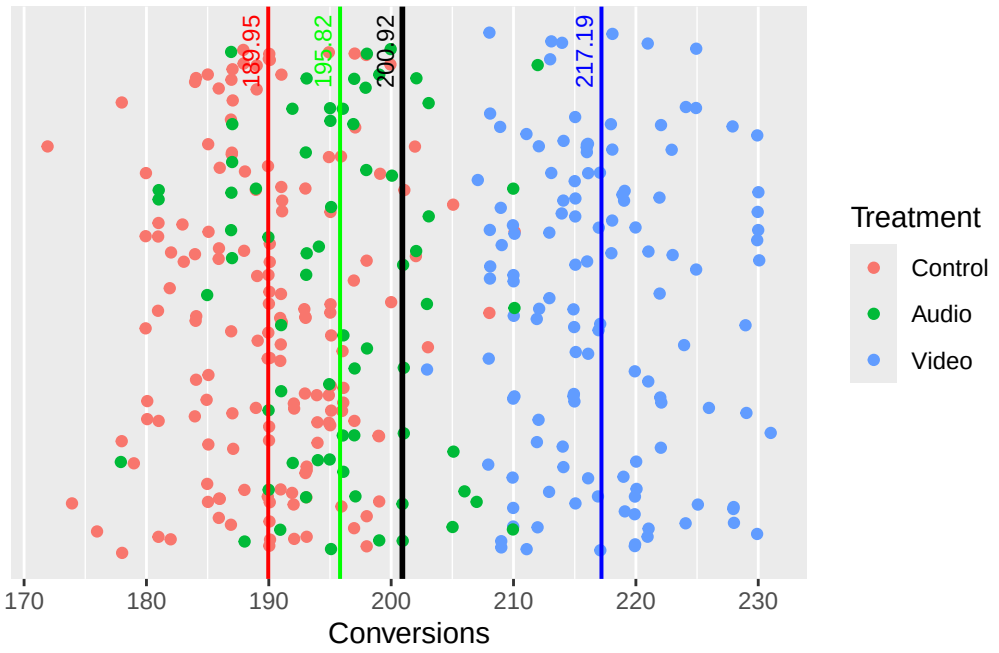

Figure 1: Treatment Means and Grand Means with Data Distribution

Can we scientifically determine whether these observed differences are statistically significant?

5 Why Not Use Pairwise A/B Tests?

A tempting idea in this setting is to run multiple pairwise **A/B Tests** to compare each design against the *Control*:

- **Audio** vs. **Control**
- **Video** vs. **Control**

While intuitive, this approach introduces big statistical concern: increased risk of false positives (*Type I Error*). That is, by running multiple comparisons independently, we raise the chance of incorrectly concluding that a difference exists when it does not.

Even if each individual test maintains a 5% error rate ($\alpha = 0.05$), the combined probability of making at least one false discovery across two tests rises to:

$$1 - (1 - 0.05)^2 = 0.0975 \approx 10\%$$

With more variations, the problem compounds quickly. For example, if we had four treatments in addition to the *Control*, the chance of a false positive somewhere among those comparisons would rise to nearly 20%:

$$1 - (1 - 0.05)^2 = 0.1855 \approx 20\%$$

Clearly, this inflation of error makes multiple pairwise
A/B Tests **statistically unreliable**.

A Better Alternative:
Testing all treatments at once instead of running several separate tests. This is where *Analysis of Variance* or short *ANOVA* comes in:

- *ANOVA* allows to test one or more groups of variables simultaneously
- *ANOVA* avoids the pitfalls of multiple testing

However, if the ANOVA result is significant, we know only that at least one, but not which one(s) of the variables are significant. Afterward, more testing is needed to find out which variable is significant.

6 Regression and ANOVA

One way to approach *ANOVA* is through regression analysis. The idea is simple: if any of the treatment types (*Audio* or *Video*) has an impact on conversions, then including these variables in a regression model should improve its explanatory power.

This is exactly what we explore below. We fit a linear model with **Treatment** as a categorical predictor of **Conversions**:

$$Conv_i = \beta_1 Audio_i + \beta_2 Video_i + \beta_0$$

Here, $Audio_i$ and $Video_i$ are dummy variables for group membership. The **Control group** is omitted and serves as the **reference category**.

The code below runs the analysis in *R*, stores the model in the object `ModelLM`, and displays the summary output in Table 3 with `summary()`.

```
ModelLM = lm(Conversions ~ Treatment, data = DataWeb)

TableLM=summary(ModelLM) |>
broom::tidy()
knitr::kable(TableLM)
```

Table 3: Output from the Linear Regression

term	estimate	std.error	statistic	p.value
(Intercept)	189.953642	0.5404799	351.453643	0
TreatmentAudio	5.869887	0.9699452	6.051772	0
TreatmentVideo	27.233349	0.8066819	33.759714	0

At first glance, the *t-values* for the treatment coefficients may suggest statistical significance. However, due to the **multiple testing problem**, we cannot rely on these individual *t-tests* for a *multi-variable* comparison. Doing so would increase the risk of an inflated *Type I error* — falsely concluding that a difference exists when it does not.

6.1 From Regression to ANOVA

So, if multiple *t-test* are not an option:

What can we do instead?

Rather than testing each treatment effect separately, we assess their combined influence on the model. In other words:

Does adding the Treatment variables simultaneously improve the model vs. no treatment effect at all?

This is precisely what *ANOVA* measures in the context of regression — it compares the **explained variance of the full model (with Treatment)** to a **restricted model (intercept-only in our case)** using an **F-test.**

The **restricted model** model can be written like this:

$$Conv_i = \beta_0 \qquad \text{with:} \quad \beta_0 = \frac{\sum_{i=1}^{N} Conv_i}{N}$$

In other words, the *restricted model* predicts every observation using only the **Grand Mean** (mean of all *Conversions*).

To quantify the improvement of the *Full* model over the *Restricted* model based on the Total Sum of Squared Errors (SSE), we calculate the *F-value* as follows:

$$F = \left(\frac{SSE_{restr} - SSE_{full}}{df_{restr} - df_{full}} \right) \div \left(\frac{SSE_{full}}{df_{full}} \right) \tag{1}$$
$$\iff$$

$$F = \overbrace{\frac{SSE_{restr} - SSE_{full}}{SSE_{full}}}^{\substack{\text{Proportional Improvement} \\ \text{(restr. to full model)}}} \cdot \overbrace{\frac{df_{full}}{df_{restr} - df_{full}}}^{C_{onst}} \tag{2}$$

Source: https://online.stat.psu.edu/stat501/lesson/6/6.2

The larger the $F-$value, the more improves the full model over the restricted model — the more important are the treatments (*Audio*, *Video*) for conversions.

6.2 Break Down the F-Value Formula

The Right Multiplier (is constant):

The second term in the formula — the right-hand multiplier — is constant once the experimental design is fixed. It depends only on the degrees of freedom of the models:

- **Number of observations:** $N = 342$
- **Degrees of freedom for the restricted model (intercept only):** $df_{\text{restr}} = 342 - 1 = 341$
- **Degrees of freedom for the full model (intercept + 2 treatments):** $df_{\text{full}} = 342 - 3 = 339$

We can calculate C^{onst} as follows:

$$C^{\text{onst}} = \frac{df_{\text{full}}}{df_{\text{restr}} - df_{\text{full}}} = \frac{342 - 3}{(342 - 1) - (342 - 3)} = \frac{339}{2}$$

The Left Multiplier: Model Improvement

The left term in the formula captures how much better the full model is in reducing the squared error:

$$\frac{SSE_{restr} - SSE_{full}}{SSE_{full}} \tag{3}$$

With the right multiplier of the F-value equation being constant, it is only the left multiplier that determines if F is large or not.

$$F = \overbrace{\frac{SSE_{restr} - SSE_{full}}{SSE_{full}}}^{\text{Proportional Improvement (restr. to full model)}} \cdot \overbrace{\frac{df_{full}}{df_{restr} - df_{full}}}^{C_{onst}}$$

6.3 Calculating the SSEs in R

To compute the SSE values for both models and derive the F-statistic, we use the `anova()` function for a linear model:

```
ModelAnova=anova(ModelLM)
```

The code below outputs the related *ANOVA* table and adds a row for the *restricted model* (Intercept).

```
ModelAnovaRestr=anova(lm(Conversions~1, data=DataWeb))

FancyOutput=rbind(broom::tidy(ModelAnovaRestr),
                  broom::tidy(ModelAnova)) |>
                  select(-p.value)
FancyOutput[1,1]="Intercept (ModelRestr)"
FancyOutput[3,1]="Residuals (ModelFull)"
colnames(FancyOutput)=c("Term","df", "SSE","MeanSq","F-Value")
knitr::kable(FancyOutput)
```

Table 4: Ammended anova() Output

Term	df	SSE	MeanSq	F-Value
Intercept (ModelRestr)	341	67426.54	197.7318	NA
Treatment	2	52473.28	26236.6419	594.8016
Residuals (ModelFull)	339	14953.26	44.1099	NA

In the ANOVA output generated by `anova(ModelLM)`, the **F-value** is already reported. However, we will recalculate it below to illustrate the procedure *under the hood* of the `anova()` function.

Step 1:
The **Sum of Squared Errors** for the **restricted model**, which includes only the intercept (i.e., the grand mean of conversions), is found in the **first row** of the ANOVA Table 4 , under the SSE column ($SSE_{\text{restr}} = 67{,}426$).

Step 2:
The **Sum of Squared Errors** for the **full model**, which includes the predictors *Audio* and *Video*, is listed in the **Residuals** row (third row) of Table 4 ($SSE_{\text{full}} = 14{,}953$).

> **Note:** The first row of the Table 4 (labeled *Intercept*) is included primarily for convenience and interpretability. However, all values required to compute the F-statistic using Equation 2 can be derived entirely from the **second and third rows** which is the output produced by `anova()`.
>
> Specifically:
>
> - $SSE_{restr} - SE_{full} = 67{,}426 - 14{,}953 = 52{,}473.28$ and
> - $df_{restr} - df_{full} = 341 - 339 = 2$
>
> are both reported in row 2 of Table 4.

Step 3:
We now plug the values from Step 1 and Step 2 together with the related degrees of freedom into the F-statistic formula from Equation 2:

$$F = \underbrace{\frac{SSE_{\text{restr}} - SSE_{\text{full}}}{SSE_{\text{full}}}}_{\substack{\text{Proportional Improvement} \\ \text{(restr. to full model)}}} \cdot \underbrace{\frac{df_{\text{full}}}{df_{\text{restr}} - df_{\text{full}}}}_{\text{Constant}}$$

Plugging in the values we get:

$$F = \underbrace{\frac{67{,}426 - 14{,}953}{14{,}953}}_{\approx 3.51} \cdot \underbrace{\frac{339}{341 - 339}}_{=169.5} = 3.51 \cdot 169.5 \approx \boxed{594.80}$$

The sum of squared errors (SSE) in the restricted model is approximately **3.5 times larger** than in the full model. Thus reflecting a big improvement through incorporating the treatment effects (*Audio* and *Video*).

Multiplying this improvement factor by the constant (169.5) yields the **F-value of 594.8**, as reported in the ANOVA output. This same value also appears in the `summary()` output for the linear regression model, confirming consistency across both methods of evaluation.

7 F-value: When is Large Large Enough for Significance (needs editing work by CL!)

In the previous section we derived an F-value of $F = 594.8$ and an F-value of that the **sum of squared errors** in the restricted model was approximately **3.5 times larger** than in the model that included the treatment groups: *Audio*, *Video*, and (implicitly) *Control*.

This suggests that the chance of all three group means being equal — and thus, their inclusion in the *full model* having no impact — is **extremely small**.

Still, in order to make a **scientific claim**, we need to test this assumption formally by stating and rejecting an (admittedly ridiculous) null hypothesis:

Hypothesis 0:

$$Mean(Conv_{Control}) = Mean(Conv_{Audio}) = Mean(Conv_{Video})$$

If **Hypothesis 0** were true, then including the treatment variables in the model would **not** reduce the error. The full model would perform just as poorly as the restricted one — meaning the F-statistic should be around **1**. Thus, we can simplify our **Hypothesis 0** to:

Hypothesis 0:

$$F = 1$$

Only for the curious reader: You might wonder — should the F-value be 0 under *Hypothesis 0*? After all, if there is no improvement, the numerator of the F-value (i.e., $SSE_{\text{restr}} - SSE_{\text{full}}$) in Equation 2 would be zero.

That intuition is partly correct: if *Hypothesis 0* is true, the **expected value** of this difference is indeed zero:

$$E(SSE_{\text{restr}} - SSE_{\text{full}}) = 0$$

However, the **variance** of $(SSE_{\text{restr}} - SSE_{\text{full}})$ is **not zero**, because repeated sampling would yield slightly different SSEs for each model. In fact, $SSE_{\text{restr}} - SSE_{\text{full}}$ if *Hypothesis 0* is true:

$$Var(SSE_{restr} - SSE_{Full}) = Var(SSE_{Full})$$

Since we find these two variances in the divident and divisor of Equation 1, $F = 1$ — if the *Hypothesis 0* is true.

Although we obtained an F-value of $F = 594.8$, we cannot immediately rule out the possibility that this result was due to **random chance** — that is, a fluke in our sample — and that the true *F-value* might still be 1.

However, we can assess how (un)likely it would be to obtain such a high *F-value* if *Hypothesis 0* were true.

To do this, we calculate a **p-value**, which represents the probability of observing an *F-value* as extreme as the one we found — **assuming Hypothesis 0 is true**.

To compute the *p-value* from the *F-distribution*, we need to know the degrees of freedom for the two variances we compare according to Equation 2. The first one is $df_{restr} - df_{full}$ (341-339=2) and the second one is df_{full} (339).

Since we trying to reject the hypothesis with a sample *F-value* that is far to the right of $F = 1$ we are performing a *right-tail* test. Below is the related code and the result:

```r
PValue=pf(594.8, df1 = 2, df2 = 339, lower.tail = FALSE)
```

```r
cat("The probability to get such an high F-Value, \nif the Hyppothesis 0 were true,
is (in scientific notation):\n", PValue)
```

```
The probability to get such an high F-Value,
if the Hyppothesis 0 were true,
is (in scientific notation):
 1.352198e-111
```

```r
cat("The probability to get such an high F-Value, \nif the Hyppothesis 0 were true,
is (written out):\n", format(PValue, scientific = FALSE))
```

```
The probability to get such an high F-Value,
if the Hyppothesis 0 were true,
is (written out):
 0.000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
```

With such a low probability for a *Type-1* error we can safely state:

One of our treatments, did significantly influence the conversions

To identify which treatment *Audio* or *Video* (or possibly both) influenced conversions, we need further testing, which exceeds the scope of this article.

Happy analytics!