# NFL Passing Yards Study

Kevin Mouck, Zachary Kemp, Nate Smith

2023-05-03

## Executive Summary

To understand if the division of National Football League (NFL) teams affects a team's passing yards per game, we took data from the 2022 NFL season and created a study. Using the division showed that there was no statistical significance between a team's division and their passing yards per game. Our additional question using post hoc contrasts also revealed that there was no statistically significant difference in the conference that the team played in and the passing yards per game. The limitations of our study and future research revolving around this topic are discussed.

## Introduction and Background

The NFL consists of 32 teams that are split into 2 conferences (AFC and NFC) and 8 divisions, with 4 divisions in each conference. The divisions include the AFC East, AFC North, AFC South, AFC West, NFC East, NFC North, NFC South, and NFC West. However, this was not always the case. When the NFL merged with the American Football League in 1970, where the NFL name was kept, there were only 4 total divisions. As expansion teams were added, division realignment kept changing, until the addition of the 32nd team to the NFL in 2002. This is when the current division realignment occurred, giving us 4 divisions in each conference (National Football League, 2023).

The old divisions and addition of teams in new geographical areas of the United States caused rivalries to form between teams that were carried over into the current divisions. With the same teams remaining in the same divisions over this time period, we wanted to see if they kept their own style of playing football or if each division changed together with the trends of the league.

The primary research questions we are interested is: Do NFL divisions make a difference in a team's passing yards per game?

If we find significant results to this question, we would also like to explore what divisions perform the best and worst in terms of passing yards per game, as well as if the conference (AFC or NFC) affects the passing yards per game.

## Study Design

For our study design, we wrote each of the 32 team names on a slip of paper and divided them up by the 8 divisions. We then proceeded to put all 4 teams from a division in a hat and selected two slips of paper that gave us the two teams from the division. We replicated this process 7 more times, giving us the 16 teams needed, 2 per division.

This is a quasi-experiment study where we have a non-random assignment of divisions with a random sampling of 16 teams from the 32 total teams. As this is a quasi-experiment instead of an experiment,

we cannot make any causal claims and can only have a generalizable correlational statement between NFL divisions and passing yards per game of NFL teams.

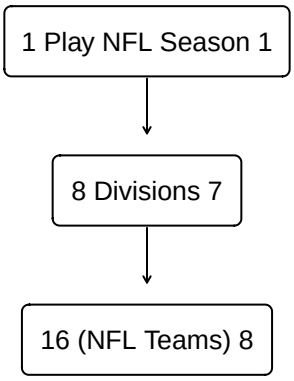


Figure 1: Hasse Diagram for Passing Yards Study

Figure 1 shows a Hasse diagram for the study where a one-way ANOVA model is applicable. We have one fixed effect categorical factor (divisions), our response is the passing yards per game of the NFL teams, and we have a sample size of 16 NFL teams. This leaves us sufficient degrees of freedom to estimate the main effect and the residuals, allowing us to utilize a one-way ANOVA model. This is not factorial, ANCOVA, or a randomized complete block design because the only factor we are assessing is the division that the team plays in and how that affects the passing yards per game of the team. Therefore, one-way ANOVA is the best method because it allows us to only use the one factor to see how the response is impacted.

Our null hypothesis is that there is no statistically significant impact of NFL division on a team's passing yards per game. Our alternative hypothesis is that the NFL division has a statistically significant affect on a team's passing yards per game. These hypotheses can be represented as:

$$H_0 : y_{ij} = \mu_{..} + \epsilon_{ij}$$
$$H_1 : y_{ij} = \mu_{..} + \alpha_i + \epsilon_{ij}$$

where $y_{ij}$ represents a team's passing yards per game, $\mu_{..}$ the baseline performance of NFL teams passing yards per game, $\alpha_i$ the additional performance due to being in each factor level, and $\epsilon_{ij}$ the residuals for each team $j$ on factor group $i$.

Our overall Type I risk will be set at 10%. We will control the Simultaneous Confidence Interval (SCI) error rate for multiple comparisons using Tukey's HSD. We chose to control SCI using Tukey's HSD because we want to look at all possible pairwise comparisons of the divisions. Our unusualness threshold will also be set at 5%.

# Data Exploration

Table 1: Summary Statistics for Passing Yards per Game

| | n | Min | Q1 | Median | Q3 | Max | MAD | SAM | SASD | Sample Skew | Sample Ex. Kurtosis |
|---|---|---|---|---|---|---|---|---|---|---|---|
| AFC East | 2 | 208.0 | 222.350 | 236.70 | 251.050 | 265.4 | 42.551 | 236.70 | 40.588 | 0 | -2.75 |
| AFC North | 2 | 178.8 | 184.250 | 189.70 | 195.150 | 200.6 | 16.160 | 189.70 | 15.415 | 0 | -2.75 |
| AFC South | 2 | 171.4 | 186.775 | 202.15 | 217.525 | 232.9 | 45.590 | 202.15 | 43.487 | 0 | -2.75 |
| AFC West | 2 | 269.6 | 276.650 | 283.70 | 290.750 | 297.8 | 20.905 | 283.70 | 19.940 | 0 | -2.75 |
| NFC East | 2 | 219.8 | 225.225 | 230.65 | 236.075 | 241.5 | 16.086 | 230.65 | 15.344 | 0 | -2.75 |
| NFC North | 2 | 130.5 | 151.275 | 172.05 | 192.825 | 213.6 | 61.602 | 172.05 | 58.761 | 0 | -2.75 |
| NFC South | 2 | 176.2 | 199.600 | 223.00 | 246.400 | 269.8 | 69.386 | 223.00 | 66.185 | 0 | -2.75 |
| NFC West | 2 | 213.3 | 216.675 | 220.05 | 223.425 | 226.8 | 10.008 | 220.05 | 9.546 | 0 | -2.75 |

Table 1 shows the passing yards per game separated by division. This is a balanced design with two teams represented in each division, meaning that we can use any type of sums of squares. To explore the amount of variation within the divisions, we can use the Sample Arithmetic Standard Deviation (SASD). This statistic provides a value that is a measure of the distance between pairs of observations within each division. The values for the SASD vary widely between the divisions, with the NFC West having the smallest and the NFC West having the largest. There are no apparent groupings of divisions when looking at the SASD. The Sample Arithmetic Mean (SAM) can be interpreted for the AFC East as the collection of teams in this division amassing 237 times as many yards per games as teams in the division. This interpretation is adequate for the rest of the divisions as well, with their own value of the SAM. Looking at the SAM, there appears to be differences in the passing yards per game in each division.

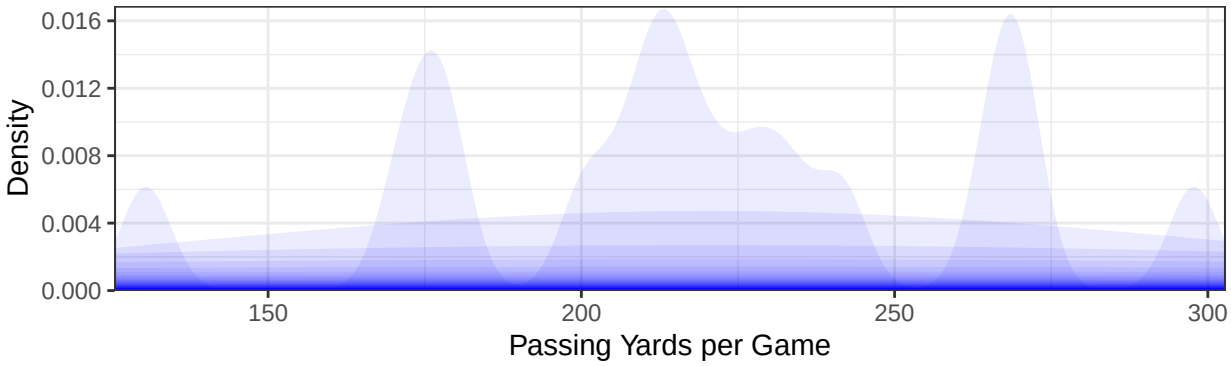


Figure 2: Shadowgram for Passing Yards Study

Figure 2 shows a shadowgram of the data. If division has a strong impact on the passing yards per game, there would be eight distinct modal clumps in the shadowgram. However, there seems to be only one clump in the shadowgram, which suggests that if there are effects of division on passing yards per game, they might be weak.

Figure 3 shows box plots for each division. From Figure 1 we can see that the sample arithmetic median for the NFC North is lower in value than the rest of the divisions while the AFC West had the highest median passing yards per game. The NFC North and NFC South had the highest spread while the NFC West had the lowest spread. There are no outliers identified in these box plots.

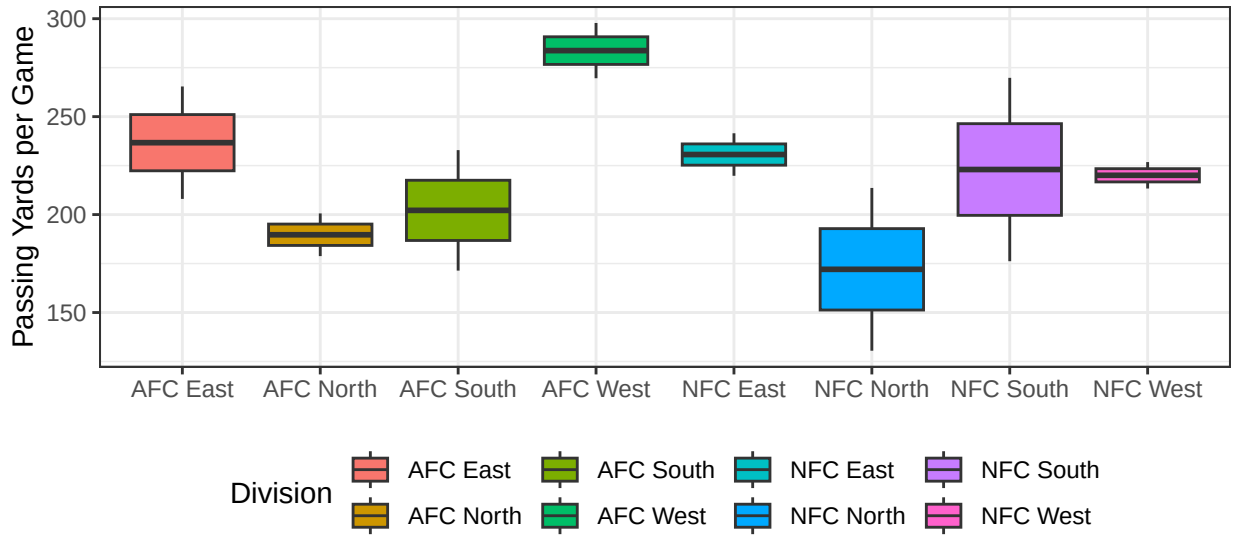


Figure 3: Side-by-side Box Plots for Passing Yards per Game

# Results

We must test for three assumptions in order to use the one-way ANOVA F test. These assumptions include the residuals following a Normal distribution, homoscedasticity, and the independence of observations.

## Assumptions

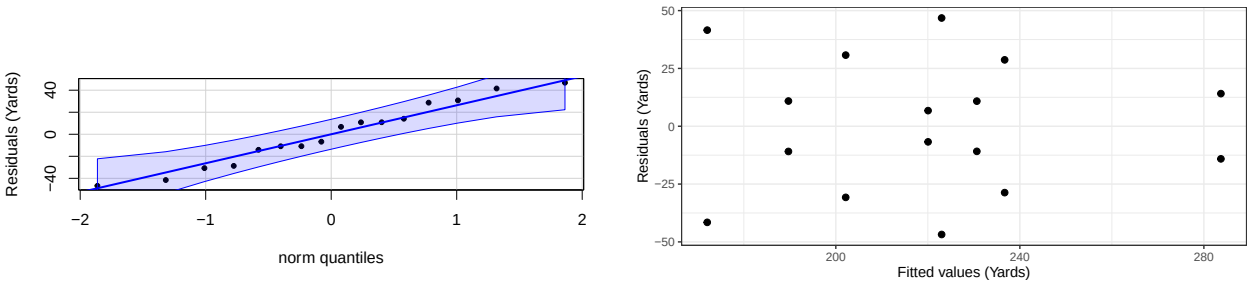


Figure 4: Assessing Assumptions with QQ Plot (Left) and Strip Chart (Right)

The left Figure 3 shows a QQ plot of the residuals with a 90% confidence envelope. All of the residuals fall into the confidence envelope. Since this is a 90% confidence envelope, we want no more than 10% of the residuals on the outside. With all of the residuals inside of the confidence envelope, we can say that our residuals satisfy the assumption of coming from a Normal distribution.

The right of figure 3 is a strip chart which allows us to assess the assumption of homoscedasticity. The fifth group has the most vertical height which appears to be more than double the height of the second and last group. These large differences in vertical space should draw caution to our assumption of homoscedasticity. However, there is no apparent pattern in the groups so that major attribute of the strip chart is not violated. Because we have a balanced design that leads to more robustness and there is no clear pattern, we can assume homoscedasticity.

The last assumption is the independence of observations. Because we do not know the measurement order, we cannot use an index plot to look at this assumption and rather we must look into the study design. The chosen teams were selected randomly from their division, so we can assume that there is an independence of observations.

## ANOVA One-Way Omnibus Results

Table 2: Modern ANOVA Table for Passing Yards Study

| Source | SS | df | MS | F | p-value | Eta Sq. | Omega Sq. | Epsilon Sq. |
|---|---|---|---|---|---|---|---|---|
| Division | 15988.84 | 7 | 2284.12 | 1.4816 | 0.2958 | 0.5645 | 0.174 | 0.1835 |
| Residuals | 12333.60 | 8 | 1541.70 | | | | | |

Table 2 shows that a team's division accounts for 1.48 times as much variation as the residuals. Under the null model that the division does not affect their passing yards per game, we would observe this F value or one more extreme around 30% of the time; our p-value is greater than our unusualness threshold of (0.296 $>$ 0.05) meaning that we fail to reject the null hypothesis and act as if the division of a NFL team does not affect their passing yards per game. This is supported by our values of omega and epsilon squared where around 17% of the variation in passing yards per game could be explained by a team's division. The large value of eta squared may be an overestimated effect size in our model due to our small sample size in each group.

Table 3: Point Estimates from the Passing Yards Study

| | Estimate |
|---|---|
| Grand Mean | 219.75 |
| AFC East | 16.95 |
| AFC North | -30.05 |
| AFC South | -17.60 |
| AFC West | 63.95 |
| NFC East | 10.90 |
| NFC North | -47.70 |
| NFC South | 3.25 |
| NFC West | 0.30 |

Table 3 shows that ignoring the divisions, the teams' passing yards per game is 219.75 times as large as the total number of teams. The teams in the AFC West had an additional performance of 63.95 yards per game per team while the teams in the NFC North had a deficit of 47.7 yards per game per team. The NFC West was right around the baseline with only an additional 0.3 yards per game per team.

## Post Hoc Contrasts

We will not conduct any post hoc pairwise comparisons given that we failed to reject the null hypothesis. However, we can still examine our question involving if either the AFC or NFC had a difference in the passing yards per game. To do this we well form a set of linear contrasts using Tukey's HSD to control the SCI.

Table 4: NFL Conferences Contrasts

| Contrast | Difference | SE | DF | t Statistic | p-value | Cohen's d | Prob. of Superiority |
|---|---|---|---|---|---|---|---|
| AFC vs. NFC | 66.5 | 78.529 | 8 | 0.847 | 0.422 | 0.599 | 0.664 |

Table 4 shows the results of the contrast between the AFC and the NFC, with a Tukey HSD adjusted p-value. We will reject the null hypothesis for this contrast with a p-value of 0.422. The effect size is small for this group and so is the probability of superiority, which tells us that picking a random team from each the AFC and NFC will result in the AFC team having a larger value for passing yards per game 66% of the time.

# Discussion

We found that the NFL divisions do not appear to affect the passing yards per game of NFL teams. It was also found that the conference does not have a significant impact on the passing yards per game of the teams. If we want to figure out what significantly affects the yards per game of NFL teams, we will need to look at more than just the division the team plays in.

This insignificant result does make sense when looking at the composition of NFL divisions, which are formed from the geography of the teams as well as old rivalries. These divisions were not created based on certain play styles. Thus we can see how that each division has evolved with the rest of the league in terms of the NFL. As the amount of passes thrown in game and the number of passing yards at the end of the season has increased, no division has stuck to the old ways of running the ball the majority of the game, and no division has significantly surpassed all the others in their passing attack capabilities.

An aspect of the teams that could have affected the passing yards per game differently for each team include the number of quarterbacks that played for each team during the season. None of our 16 teams in the study had the same quarterback play every game. Playing multiple quarterbacks during the season will usually have an affect on the passing yards per game of the whole season, as even one bad game from a backup quarterback could alter the results. When multiple quarterbacks play during the season, it is usually due to injury to the best quarterback on the team, or from poor performances from the starting quarterback. Both of these situations will cause a change in the passing yards per game as worse or better quarterbacks come in and throw for different amounts of yards during the game.

Aspects such as multiple quarterbacks, as well as offensive scheme, strength of schedule, and rankings of wide receivers on each team could all have limited our study. These attributes of teams may have a more significant impact on the passing yards per game of a NFL team compared to the division they play in, and are worth looking into to see how passing yards per game can be tracked from these aspects.

# References

Hatfield, N. J. (2023). Unit 2: Study Design. In STAT 461: Analysis of Variance (ANOVA) Course Materials. Pennsylvania State University.

Hatfield, N. J. (2023). Unit 3: Oneway ANOVA. In STAT 461: Analysis of Variance (ANOVA) Course Materials. Pennsylvania State University.

Wikimedia Foundation. (2023, May 1). National Football League. Wikipedia. https://en.wikipedia.org/wiki/National_Football_League

## Code Appendix

```
knitr::opts_chunk$set(echo = TRUE)
packages <- c("tidyverse", "knitr", "kableExtra", "sass",
"parameters", "emmeans", "DescTools", "multcompView", 'hasseDiagram')
lapply(packages, library, character.only = TRUE)
options(knitr.kable.NA = "")
options(contrasts = c("contr.sum", "contr.poly"))
source("https://raw.github.com/neilhatfield/STAT461/master/rScripts/ANOVATools.R")
source("https://raw.github.com/neilhatfield/STAT461/master/rScripts/shadowgram.R")
modelLabels <- c("1 Play NFL Season 1", "8 Divisions 7", "16 (NFL Teams) 8")
modelMatrix <- matrix(
  data = c(FALSE, FALSE, FALSE, TRUE, FALSE, FALSE, TRUE, TRUE, FALSE),
  nrow = 3,
  ncol = 3,
  byrow = FALSE
)
hasseDiagram::hasse(
 data = modelMatrix,
 labels = modelLabels
)
footballData <- read.csv('PassingYards.csv', header = TRUE, sep = ',')

footballData$DIV <- as.factor(footballData$DIV)

colnames(footballData)[colnames(footballData) == "DIV"] ="Division"

footballData$Division <- factor(
  x = footballData$Division,
  levels = c('AFC East','AFC North','AFC South','AFC West',
             'NFC East', 'NFC North', 'NFC South', 'NFC West')
)
footballModel <- aov(
  formula = Y.G ~ Division,
  data = footballData
)
scoreStats <- psych::describeBy(
  x = footballData$Y.G,
  group = footballData$Division,
  na.rm = TRUE,
  skew = TRUE,
  ranges = TRUE,
  quant = c(0.25, 0.75),
  IQR = FALSE,
  mat = TRUE,
  digits = 4
)

scoreStats %>%
  tibble::remove_rownames() %>%
  tibble::column_to_rownames(
    var = "group1"
  ) %>%
```

```
  dplyr::select(
    n, min, Q0.25, median, Q0.75, max, mad, mean, sd, skew, kurtosis
  ) %>%
  knitr::kable(
    caption = "Summary Statistics for Passing Yards per Game",
    digits = 3,
    format.args = list(big.mark = ","),
    align = rep('c', 11),
    col.names = c("n", "Min", "Q1", "Median", "Q3", "Max", "MAD", "SAM", "SASD",
                  "Sample Skew", "Sample Ex. Kurtosis"),
    booktabs = TRUE
  )  %>%
  kableExtra::kable_styling(
    font_size = 12,
    latex_options = c("scale_down", "HOLD_position")
  )

shadowgram(
  dataVec = footballData$Y.G,
  label = "Passing Yards per Game",
  layers = 50,
  color = "blue",
  aStep = 4
)
ggplot2::ggplot(
data = footballData,
mapping = aes(x = Division, y = Y.G, fill = Division)
) +
ggplot2::geom_boxplot(
na.rm = TRUE
) +
ggplot2::theme_bw() +
xlab(NULL) +
ylab("Passing Yards per Game") +
theme(
legend.position = "bottom"
)
car::qqPlot(
x = footballModel$residuals,
distribution = "norm",
envelope = 0.90,
id = FALSE,
pch = 20,
ylab = "Residuals (Yards)"
)

ggplot(
data = data.frame(
residuals = footballModel$residuals,
fitted = footballModel$fitted.values
),
mapping = aes(x = fitted, y = residuals)
) +
```

```
geom_point(size = 2) +
theme_bw() +
xlab("Fitted values (Yards)") +
ylab("Residuals (Yards)")
parameters::model_parameters(
  model = footballModel,
  effectsize_type = c("eta", "omega", "epsilon")
) %>%
dplyr::mutate(
  p = ifelse(
    test = is.na(p),
    yes = NA,
    no = pvalRound(p, digits = 4)
)
) %>%
knitr::kable(
  digits = 4,
  col.names = c(
    "Source", "SS", "df", "MS", "F", "p-value",
    "Eta Sq.", "Omega Sq.", "Epsilon Sq."),
  caption = "Modern ANOVA Table for Passing Yards Study",
  booktabs = TRUE,
  align = c("l", rep("c", 8))
) %>%
kableExtra::kable_styling(
  font_size = 10,
  latex_options = c("HOLD_position")
)
pointEst <- dummy.coef(footballModel)

pointEst <- unlist(pointEst)
names(pointEst) <- c("Grand Mean", "AFC East", "AFC North", 'AFC South', 'AFC West',
                     'NFC East', 'NFC North', 'NFC South', 'NFC West')
data.frame("Estimate" = pointEst) %>%
knitr::kable(
digits = 2,
caption = "Point Estimates from the Passing Yards Study",
booktabs = TRUE,
align = "c"
) %>%
kableExtra::kable_styling(
font_size = 12,
latex_options = c("HOLD_position")
)
c1 <- c(1,1,1,1,-1,-1,-1,-1)
contrasts(footballData$Division) <- cbind(c1)
footballContrasts <- aov(
    formula = Y.G ~ Division,
    data = footballData
)
contrastOut <- summary(
object = footballContrasts,
split = list(
```

```
Division = list(
"AFC vs. NFC" = 1
)
)
)
footballMeans <- emmeans::emmeans(
object = footballModel,
specs = ~ Division
)


FBContrasts1 <- emmeans::contrast(
object = footballMeans,
method = list(
"AFC vs. NFC" = c1
),
adjust = "tukey"
)

as.data.frame(FBContrasts1) %>%
dplyr::mutate(
cohen = effectsize::t_to_d(t = t.ratio, df_error = df)$d,
ps = probSup(cohen)
) %>%
kable(
digits = 3,
caption = "NFL Conferences Contrasts",
col.names = c("Contrast", "Difference", "SE", "DF", "t Statistic",
paste0("p-value"),
"Cohen's d", "Prob. of Superiority"),
align = "lccccccc",
booktabs = TRUE,
escape = FALSE
) %>%
kableExtra::kable_styling(
bootstrap_options = c("striped", "condensed"),
font_size = 12,
latex_options = c("HOLD_position", "scale_down")
)
```