Analysis of an Agricultural Dataset from an Argentinian Corn Field

Yield monitor data for the Las Rosas farm with variable nitrogen.

Jack Nguyen, Nikita Kiselov

2025-12-18

1 Introduction

1.1 Dataset Overview

The dataset that we chose documents an experimental study investigating corn yield response to nitrogen fertilizer treatments while accounting for field characteristics. It was collected from the Las Rosas farm estate in Rio Cuarto, Cordoba, Argentina, by Anselin, Bongiovanni, and Lowenberg-DeBoer (2004). The data was collected using a yield monitor from strip trials during the harvests in the years 1999 and 2001 to determine the potential of spatial econometric analysis in estimating the site-specific crop response when it comes to applying nitrogen treatment. There are 3443 observations from corn plants and 8 variables, including the year, latitude, longitude, yield in quintals/ha, amount of nitrogen in kg/ha, topographic factor, brightness value, and nitrogen factor. Of these variables, the topographic and nitrogen factors are qualitative, while the rest are quantitative.

1.2 Summary Statistics

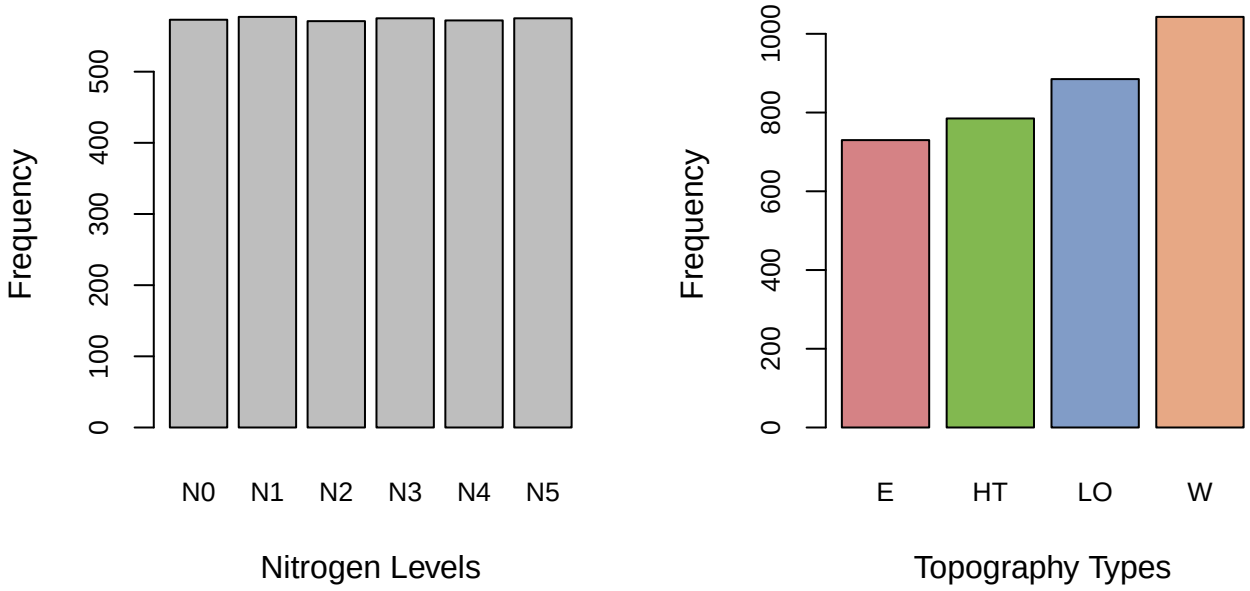
Based on the summary statistics in Table 1, corn yield responses to nitrogen fertilization show a clear positive trend. Mean yields increase progressively from 65.0 quintals/ha with no nitrogen (N0) to 72.8 quintals/ha at the highest application level (N5), representing an 11.8% improvement. Variability also shifts systematically, with the standard deviation decreasing from 20.9 to 20.1 and the interquartile range narrowing from 36.0 to 29.8 across treatments, suggesting more consistent yield responses at higher nitrogen levels. These patterns indicate both increased productivity and reduced yield variability with nitrogen application across the experimental field. We omit mode because the yield variable is continuous and repetition of values is more likely a coincidence than an important trend.

Table 1: Summary Statistics of Yield (quintals/ha) by Nitrogen Fertilizer Level

Level	Size	Mean	Median	SD	IQR	Range
N0	573	65.0	62.0	20.9	36.0	12.66 - 108.84
N1	577	68.6	65.2	19.2	29.8	27.44 - 110.54
N2	571	69.7	67.1	19.3	28.1	31.79 - 112.85
N3	575	70.3	66.8	19.2	27.7	19.41 - 110.12
N4	572	72.6	69.2	19.1	27.5	32.05 - 117.9
N5	575	72.8	70.3	20.1	29.8	31.79 - 117.19

Also, for each nitrogen level (according to Figure 1), the crops were given approximately the same amount of nitrogen treatment. However, this is not the case in terms of topography levels, as they each have different areas. West Slope (W) was most common, while East Slope (E) was least common (Figure 1).

Figure 1: Distribution of nitrogen and topography treatment levels



1.3 Farm topography and yield

The farm has 4 distinct topographic areas due to hilly terrain (Figure 2). These seem to significantly affect the mean corn yield, as seen in Figure 3, with Hilltop (HT) having the poorest harvest and Low East having the best harvest. The data from the year 1999 has less variation than 2001 and is therefore worse for making statistically significant conclusions. Therefore, we use data from 2001 for the rest of the report.

Figure 2: Farm topography in 2001

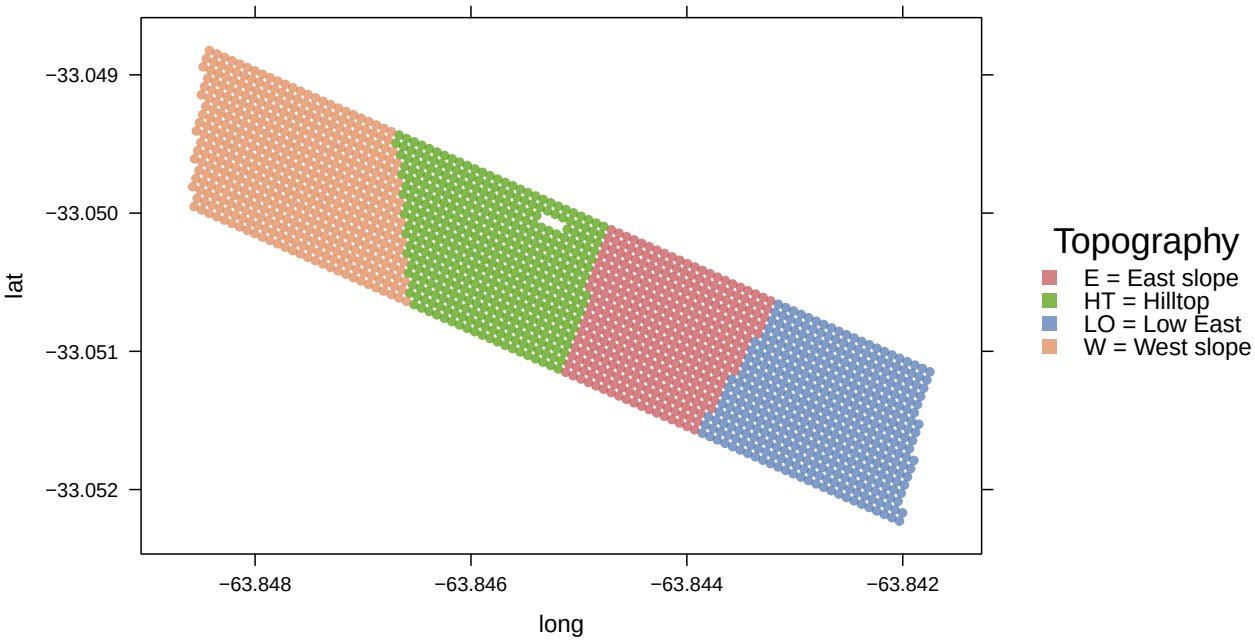
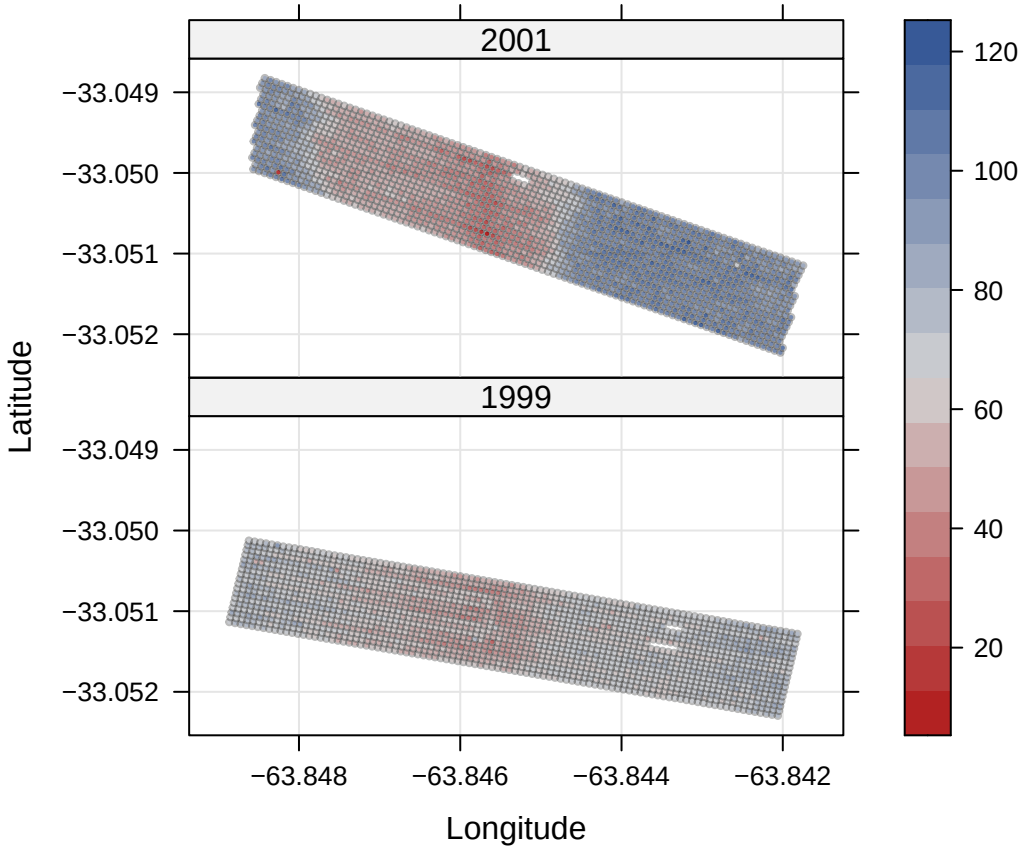
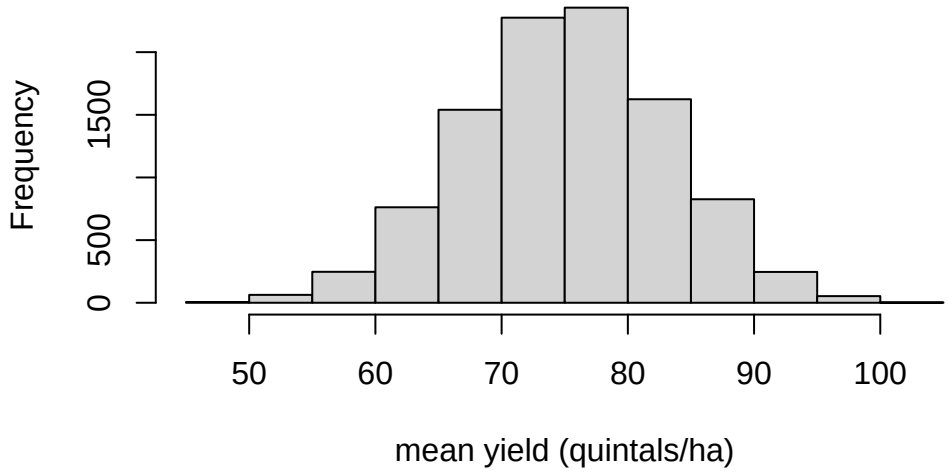


Figure 3: Corn yield distribution by location



1.4 Sampling distribution

Figure 4: Sampling distribution of mean corn yield in 2001 of 10000 samples of size 10



As seen in Figure 4, the sampling distribution is symmetric around 75 and approximately normal. Although the sample size ($n = 10$) is relatively small for CLT to apply ($n < 30$), the resulting sampling distribution appears approximately normal, suggesting that the underlying yield distribution is not strongly skewed.

2 One Parameter Test

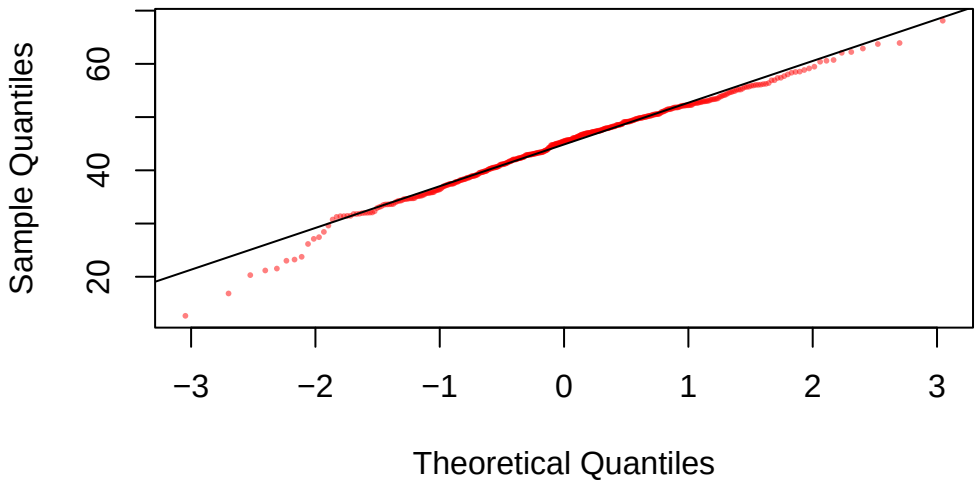
The mean corn yield in 2001 across all topographies was 75.2 quintals/ha. However, as seen in Figure 3, the corn yield differed between various parts of the farm. To test whether some topographies affect corn growth, we compare the true mean corn yield on Hilltop (HT) topography (μ) to the mean yield of the entire farm in 2001. We chose Hilltop since its yield distribution is closest to normal among all 4 topographies. The overall farm mean of 75.2 quintals/ha is used as a benchmark value for comparison. We use a one-sample two-tailed t-test at $\alpha = 0.01$, with the following hypotheses:

$$H_0 : \mu = 75.2 \quad \text{(same as farm mean)}$$

$$H_a : \mu \neq 75.2 \quad \text{(different from farm mean)}$$

There are two assumptions to assess: normality and independence. Figure 5 shows that the yield distribution is very close to normal apart from a few outliers on the lower end that are below the line. This implies that there are more very low yields than normality would predict. This is expected, however, since some plants could have failing yield due to parasites or other environmental factors.

Figure 5: Q-Q plot for corn yield on hilltop topography in 2001



Independence can not be fully claimed since the data was collected from a single field. Therefore, the result can not be generalized to corn yield on Hilltop topography in general. However, if the scope of this analysis is limited to the Las Rosas farm and the nearby area, the sample contains enough independent measurements ($n = 431$) and is representative, so the t-test can proceed.

Table 2: t-test results for hilltop true mean yield

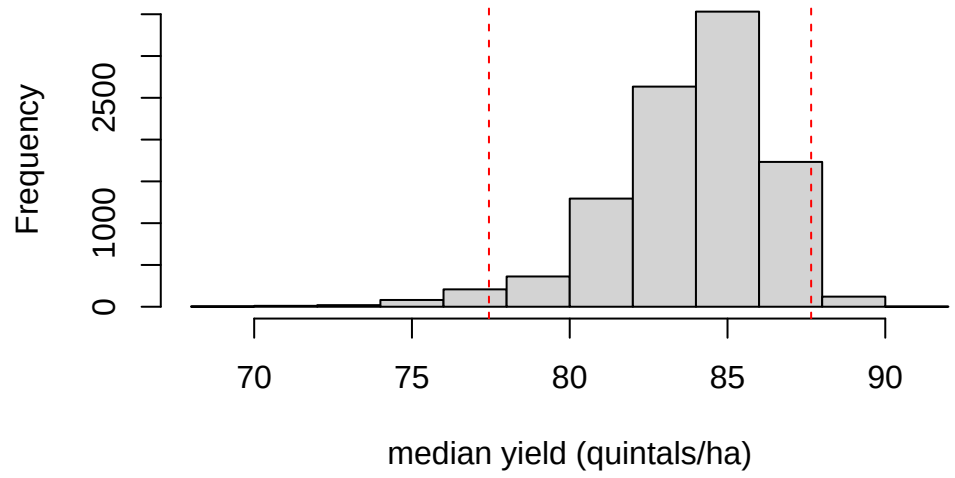
p-value	df	99% CI
<1e-100	430	43.68 - 45.69

Based on the data from Las Rosas farm in 2001, we are 99% confident that the true mean yield of corn on Hilltop topography of the farmland is between 43.68 and 45.69. Since this interval does not contain 75.2, there is statistically significant evidence against the null hypothesis at 1% significance level. We conclude that the true mean corn yield on Hilltop topography is different from the mean yield of all topographies on the Las Rosas farm.

3 Bootstrap Approach

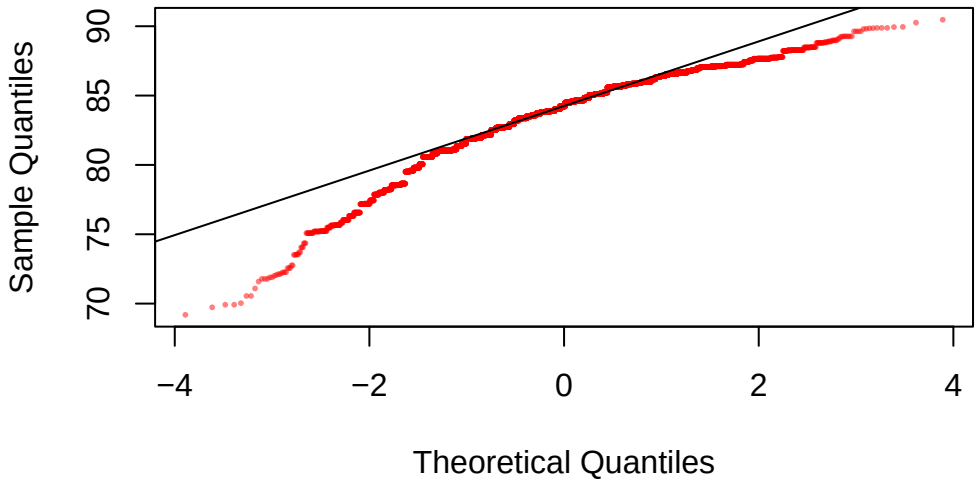
Due to topographic variations, the sampling distribution for the corn yield median is not normal as seen in Figure 6. This invalidates the use of a direct t-confidence interval for the median. Instead, we use a bootstrap technique to create a confidence interval for the true median 2001 corn yield.

Figure 6: Bootstrapped distribution for median 2001 corn yield of 10000 samples with replacement. The 95% confidence percentile confidence interval is shown in red.



The Q-Q plot in Figure 7 shows a concave curve, implying that the bootstrap distribution is left-skewed (Kross (2016)). Therefore, a percentile-based confidence interval is appropriate and the bootstrapped t-confidence interval is not.

Figure 7: Q-Q plot for the bootstrapped medians



The constructed 95% percentile-based confidence interval is shown in Figure 6. We are 95% confident that the true median corn yield on the Las Rosas farm in 2001 was between 77.44 and 87.65 quintals/ha.

4 Analysis of Variance

Anselin, Bongiovanni, and Lowenberg-DeBoer (2004) studied 6 treatment levels for the nitrogen factor in corn yield. We perform an ANOVA test on the corn yield measurements from 2001 to test if the means of the treatment levels are different. The hypotheses are:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_6 \quad \text{(same means for each nf level)}$$

$$H_a : \mu_i \neq \mu_j \quad \text{(at least two means are different)}$$

Table 3: Results for ANOVA test on difference of corn yield based on nitrogen treatment

Source of Variation	df	Sum of Squares	Mean Square	f	p-value	Rejection Region
Treatments	5	7544.2	1508.8	2.304	0.042	f > 2.219
Error	1699	1112697.0	654.9			
Total	1704	1120241.2				

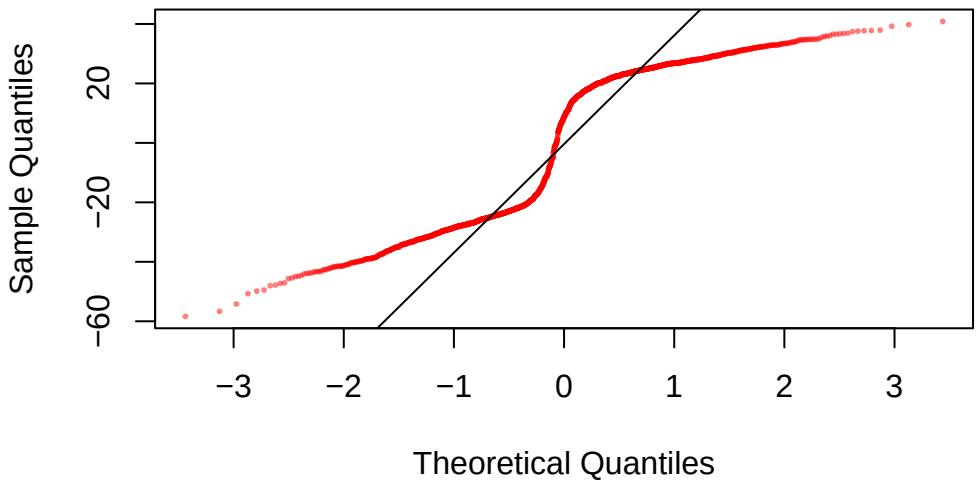
The p-value for the test in Table 3 is p=0.042 with rejection region f $\geq$ 2.219. The two assumptions for ANOVA are normality and equal variances. We test for equal variances using Levene's test with the null hypothesis of equal variance and alternative hypothesis that at least two variances differ.

Table 4: Results for Levene test on variance of corn yield based on nitrogen treatment

df	F value	p-value
5	1.4017	0.221

Since the p-value for the Levene's test is p>0.05 (Table 4), there is not enough evidence to reject the null hypothesis, so we can assume equal variances. On the other hand, Q-Q plot of the residuals in Figure 8 obtained from ANOVA in Table 3 shows an S-curve, which implies a heavy-tailed distribution (Kross (2016)). Therefore, normality assumption is not satisfied and ANOVA is not applicable.

Figure 8



To test for the equality of means despite lack of normality, we turn to the Kruskal–Wallis test (Soetewey (2022)). This test doesn't assume normality or equal variance and checks for a difference in the underlying

distributions. Given that the factor levels have similar variances from Levene's test, any remaining difference in distribution is primarily due to a shift (e.g., median). Therefore, we can substitute the Kruskal–Wallis test for ANOVA and still make inferences about differences of medians. The switch from mean to median is necessitated by the Kruskal–Wallis test (Soetewey (2022)).

Table 5: Results for the Kruskal–Wallis test on effect of nitrogen treatment

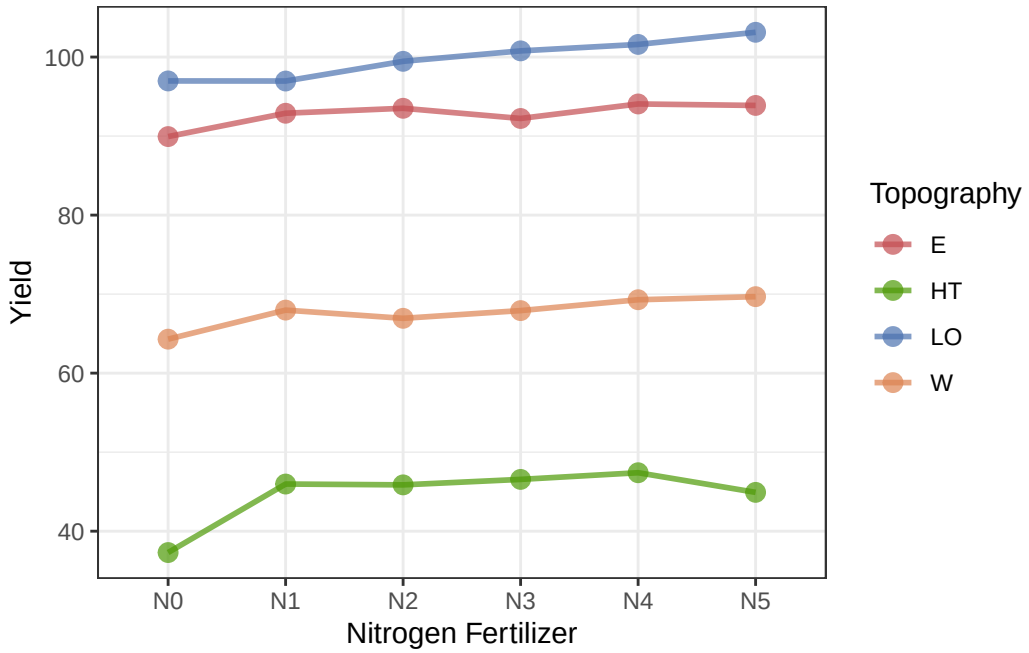
df	chi-squared	p-value
5	31.4789	7.53e-06

The results for the Kruskal–Wallis test in Table 5 indicate that there is statistically significant evidence at 5% confidence level that there is a difference in median corn yield between nitrogen treatment levels. This agrees with the ANOVA result in Table 3 even though it was not applicable here. The next step with ANOVA/Kruskal–Wallis test would be to use multiple comparisons to determine which pairs of levels have differences in mean/median.

5 Multiple Comparisons

The Las Rosas farm from Anselin, Bongiovanni, and Lowenberg-DeBoer (2004) contained 4 different topographies that affected corn yield in addition to the 6 nitrogen fertilizer treatment levels used by the researchers. To study the combined effects of these two factors on the corn yield in 2001, we first construct an interaction plot in Figure 9.

Figure 9: Interaction between nitrogen and topography factors on corn yield



The yield generally increases with higher nitrogen treatment as seen in Figure 9. However, each line has a distinct treatment at which this trend breaks (HT at N5, W at N1, E at N3, and LO at N0). There is clearly an interaction between the two factors, so we must use a non-additive fixed effects model in the following form:

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$$

Where γ_{ij} are the interaction parameters, ϵ_{ijk} are the residuals, and α_i/β_j are factor level effects. Next, we perform two-factor ANOVA based on this model. This test assumes normality and equal variances, which we verify next. The 3 null hypotheses tested are:

$$H_{0AB} : \gamma_{ij} = 0 \, \forall i, j \quad \text{(there is no interaction)}$$

$$H_{0A} : \alpha_i = 0 \, \forall i \quad \text{(there is no difference in effect of nitrogen)}$$

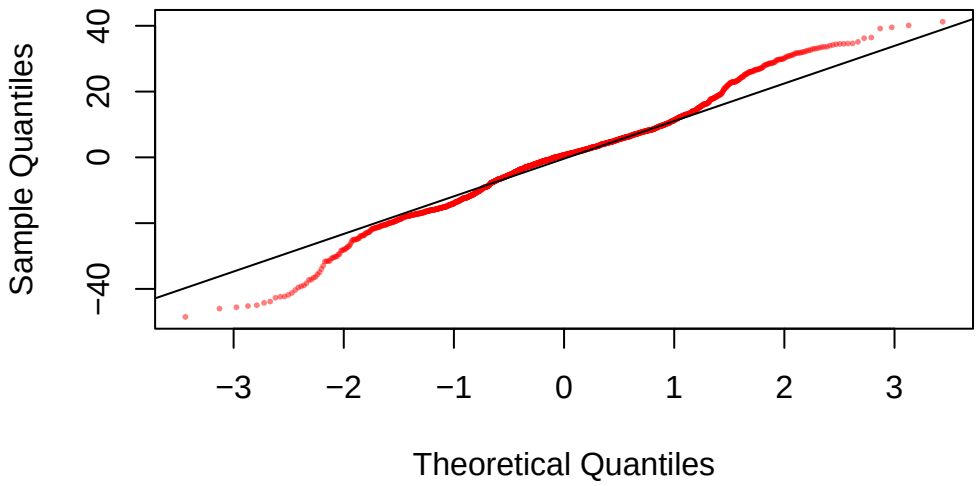
$$H_{0B} : \beta_j = 0 \, \forall j \quad \text{(there is no difference in effect of topography)}$$

Table 6: Results for Multifactor ANOVA on combined effect of nitrogen and topography on corn yield

Source of Variation	df	Sum of Squares	Mean Square	f	p-value	Rejection Region
Nitrogen	5	7544.2	1508.8	8.113	1.41e-07	f > 2.219
Topography	3	797539.4	265846.5	1429.46	<1e-100	f > 2.61
Interaction	15	2530.6	168.7	0.907	0.556	f > 1.672
Error	1681	312627.1	186			
Total	1704	1120241.2				

The Q-Q plot in Figure 10 for the residuals shows no strong deviation from normality. However, there are still bends at endpoints, indicating that the residual distribution has fat tails (Kross (2016)). Since ANOVA is reasonably robust against slight violations of normality, the normality assumption is valid.

Figure 10: Q-Q plot for the residuals of the two-factor ANOVA test on nitrogen/topography



On the other hand, Figure 11 does not satisfy the equal variances condition. Going from left to right, there is a significant difference in the spread of residuals. Therefore, the two-factor ANOVA test can not be justified (Soetewey (2023)). We turn to Scheirer–Ray–Hare test (Mangiafico (n.d.)), which is an extension of the Kruskal–Wallis test that was used in Section 4. It is also a non-parametric test that checks if the samples came from the same underlying distribution without assuming normality.

Figure 11: Variations of ANOVA residuals for each fitted value (tests for homogeneity)

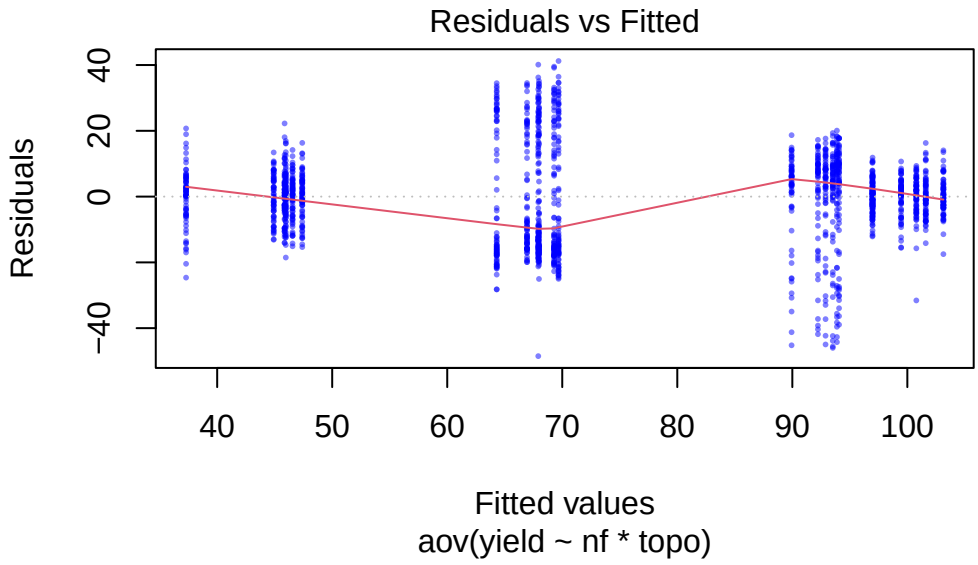


Table 7: Results for Scheirer–Ray–Hare test on combined effect of nitrogen and topography on corn yield

Source of Variation	df	Sum of Squares	H	p-value
Nitrogen	5	7135008	29.436	1.9e-05
Topography	3	281854775	1162.796	<1e-100
Interaction	15	1728372	7.13	0.954
Error	1681	121826047		
Total	1704	412544202		

Based on both results of the two-factor ANOVA (Table 6) and the Scheirer–Ray–Hare (Table 7), there is statistically significant evidence at 5% significance level ($p < 0.05$) against hypotheses H_{0A} and H_{0B}, but not H_{0AB}. This means that both nitrogen and topography have a significant effect on mean/median corn yield; however, their interaction is not a significant factor in mean/median corn yield. Note that conclusions from the Scheirer–Ray–Hare test are about medians (Mangiafico (n.d.)) in contrast to means from the two-factor ANOVA.

Instead of the Tukey procedure used with ANOVA, we use the Pairwise Permutation method to determine which pairs of treatment levels have significantly different medians. This method is non-parametric and complements the earlier use of the Scheirer–Ray–Hare test. The results of the multiple comparisons are shown in Table 8.

Table 8: Pairs of treatment levels with statistically significant mean differences at p=0.05

Comparison	Stat	p-value	p-adjust
E - HT = 0	25.05	<1e-100	<1e-100
E - LO = 0	-8.171	3.05e-16	1.83e-15
E - W = 0	16.44	<1e-100	<1e-100
HT - LO = 0	-28.31	<1e-100	<1e-100
HT - W = 0	-18.19	6.64e-74	3.98e-73
LO - W = 0	22.11	<1e-100	<1e-100

No significant difference was found between nitrogen levels in Table 8. This does not agree with the result from Section 4 and Table 7 because the Pairwise Permutation method uses the Bonferroni adjustment for p-value, which is more conservative due to making no assumptions about the underlying distribution. In conclusion, based on 2001 data from the Las Rosas farm, there was a difference between median corn yields for all pairs of topographies; however, there was no significant difference between medians of nitrogen treatment levels.

6 References

Anselin, Luc, Rodolfo Bongiovanni, and Jess Lowenberg-DeBoer. 2004. "A Spatial Econometric Approach to the Economics of Site-Specific Nitrogen Management in Corn Production." *American Journal of Agricultural Economics* 86 (3): 675–87. https://doi.org/https://doi.org/10.1111/j.0002-9092.2004.00610.x.

Kross, Sean. 2016. "A q-q Plot Dissection Kit." February 29, 2016. https://seankross.com/2016/02/29/A-Q-Q-Plot-Dissection-Kit.html.

Mangiafico, Salvatore S. n.d. "Scheirer–Ray–Hare Test." Accessed December 15, 2025. https://rcompanion.org/handbook/F_14.html.

Soetewey, Antoine. 2022. "Kruskal-Wallis Test, or the Nonparametric Version of the ANOVA." March 24, 2022. https://statsandr.com/blog/kruskal-wallis-test-nonparametric-version-anova/.

———. 2023. "Two-Way ANOVA in r." June 19, 2023. https://statsandr.com/blog/two-way-anova-in-r/.

7 Code Appendix

The code below shows all the R code used to generate the tables and figures in this document. The code is displayed here for transparency and reproducibility but is hidden in the main body of the report.

```
library(tidyverse)
library(knitr)
library(kableExtra)
library(agridat)
data(lasrosas.corn)
dat <- lasrosas.corn

set.seed(101)

pformat <- function(x) {
  ifelse(x<1e-100, "<1e-100",
  ifelse(x < 0.01,
         format(x, scientific = TRUE, digits = 3),
         round(x, 3)))
}

topo_cols <- adjustcolor(c("#C44E52","#4E9A06","#4C72B0","#DD8452"), alpha.f = 0.7)

dat %>% group_by(nf) %>%
  summarise(
    N = n(),
    Mean = mean(yield),
    Median = median(yield),
    SD = sd(yield),
    IQR = IQR(yield),
    Range = paste(format(min(yield), nsmall = 1), "-", format(max(yield), nsmall = 1)),
    .groups = "drop"
  ) %>% kable(
      digits = 1,
      col.names = c("Level", "Size", "Mean", "Median", "SD", "IQR", "Range"))
par(mfrow = c(1, 2))

barplot(table(dat$nf), main = "",
        xlab = "Nitrogen Levels", ylab = "Frequency",
        cex.names = 0.8,cex.axis = 0.8,cex.main = 1.2)

barplot(table(dat$topo), main = "",
        xlab = "Topography Types", ylab = "Frequency",
        cex.names = 0.8,col = topo_cols,cex.axis = 0.8,cex.main = 1.2)
library(lattice)
library(latticeExtra)

dat2001 <- subset(dat, year == 2001)

legend_labels <- c("E = East slope","HT = Hilltop",  "LO = Low East", "W = West slope")
xyplot(lat ~ long, data = dat2001, groups = topo, pch = 16, col = topo_cols,
  key = list(
    space = "right", title = "Topography",
    points = list(col = topo_cols, cex = 1.2, pch = 15),
```

```
    text = list(legend_labels)
  )
)
redblue <- colorRampPalette(c("firebrick", "lightgray", "#375997"))
levelplot(yield ~ long*lat|factor(year), data=dat,
          xlab="Longitude", ylab="Latitude",
          scales=list(alternating=FALSE),
          prepanel = prepanel.default.xyplot,
          panel = panel.levelplot.points,
          type = c("p", "g"), aspect = "iso", col.regions=redblue, cex = 0.4)
means <- replicate(10000,mean(sample(dat2001$yield, size = 10, replace = TRUE)))
hist(means, main = "", xlab = "mean yield (quintals/ha)")
dat2001hills <- subset(dat2001, topo == "HT")
qqnorm(dat2001hills$yield,main = "", pch = 16, cex = 0.4, col = rgb(1, 0, 0, 0.5));
qqline(dat2001hills$yield)
ttest_result <- t.test(dat2001hills$yield, mu = 75.2, conf.level = 0.99)
results_table <- tibble(
  "p-value" =  pformat(ttest_result$p.value),
  df = ttest_result$parameter,
  "99% CI" = paste(round(ttest_result$conf.int[1], 2), "-",
            round(ttest_result$conf.int[2], 2))
)

kable(results_table,row.names = FALSE,align = c("l", "l", "l"))
bootstrap_medians <- replicate(10000,median(sample(dat2001$yield, size =
                                  length(dat2001$yield), replace = TRUE)))
percentile_ci <- quantile(bootstrap_medians, probs = c(0.025, 0.975))
hist(bootstrap_medians, main = "", xlab = "median yield (quintals/ha)")
abline(v = percentile_ci[1], col = "red", lty = "dashed")
abline(v = percentile_ci[2], col = "red", lty = "dashed")
qqnorm(bootstrap_medians,main = "", pch = 16, cex = 0.4, col = rgb(1, 0, 0, 0.5));
qqline(bootstrap_medians)
library(car)
anova_result <- aov(yield ~ nf, data = dat2001)
anova_summary <- summary(anova_result)[[1]]
anova_table <- tibble(
  "Source of Variation" = c("Treatments", "Error", "Total"),
  df = c(anova_summary$Df[1], anova_summary$Df[2], sum(anova_summary$Df)),
  "Sum of Squares" = c(round(anova_summary[1, "Sum Sq"], 1),
                       round(anova_summary[2, "Sum Sq"], 1),
                       round(sum(anova_summary[, "Sum Sq"]), 1)),
  "Mean Square" = c(round(anova_summary[1, "Mean Sq"], 1),
                    round(anova_summary[2, "Mean Sq"], 1), ""),
  f = c(round(anova_summary[1, "F value"], 3), "", ""),
  "p-value" = c(pformat(anova_summary[1, "Pr(>F)"]),"",""),
  "Rejection Region" = c(paste("f > ", round(qf(0.95, anova_summary$Df[1],
                             anova_summary$Df[2]), 3)),"","")
)

kable(anova_table)
levene_result <- leveneTest(yield ~ nf, data = dat2001)

levene_table <- tibble(
```

```
  df = levene_result$Df[1],
  "F value" = round(levene_result$"F value"[1],4),
  "p-value" = pformat(levene_result$"Pr(>F)"[1])
)

kable(levene_table)
qqnorm(anova_result$residuals,main = "", pch = 16, cex = 0.4, col = rgb(1, 0, 0, 0.5));
qqline(anova_result$residuals)
kruskal_result <- kruskal.test(yield ~ nf, data = dat2001)

kruskal_table <- tibble(
  df = kruskal_result$parameter,
  "chi-squared" = round(kruskal_result$statistic,4),
  "p-value" = pformat(kruskal_result$p.value[1])
)

kable(kruskal_table)
library(ggplot2)

ggplot(dat2001, aes(x = nf, y = yield, color = topo, group = topo)) +
  stat_summary(fun = mean, geom = "line", size = 1) +
  stat_summary(fun = mean, geom = "point", size = 3) +
  scale_color_manual(values = topo_cols) +
  labs(x = "Nitrogen Fertilizer", y = "Yield", color = "Topography") +
  theme_bw()
multinova_result <- aov(yield ~ nf * topo, data = dat2001)

multinova_summary <- summary(multinova_result)[[1]]

multinova_table <- tibble(
  "Source of Variation" = c("Nitrogen", "Topography", "Interaction", "Error", "Total"),
  df = c(multinova_summary$Df[1],multinova_summary$Df[2],
         multinova_summary$Df[3],multinova_summary$Df[4],
         sum(multinova_summary$Df)),
  "Sum of Squares" = c(round(multinova_summary[1, "Sum Sq"], 1),
                       round(multinova_summary[2, "Sum Sq"], 1),
                       round(multinova_summary[3, "Sum Sq"], 1),
                       round(multinova_summary[4, "Sum Sq"], 1),
                       round(sum(multinova_summary[, "Sum Sq"]), 1)),
  "Mean Square" = c(round(multinova_summary[1, "Mean Sq"], 1),
                    round(multinova_summary[2, "Mean Sq"], 1),
                    round(multinova_summary[3, "Mean Sq"], 1),
                    round(multinova_summary[4, "Mean Sq"], 1),""),
  f = c(round(multinova_summary[1, "F value"], 3),
        round(multinova_summary[2, "F value"], 3),
        round(multinova_summary[3, "F value"], 3),"",""),
  "p-value" = c(pformat(multinova_summary[1, "Pr(>F)"]),
                pformat(multinova_summary[2, "Pr(>F)"]),
                pformat(multinova_summary[3, "Pr(>F)"]),"",""),
  "Rejection Region" = c(
    paste("f > ", round(qf(0.95, multinova_summary$Df[1], multinova_summary$Df[4]), 3)),
    paste("f > ", round(qf(0.95, multinova_summary$Df[2], multinova_summary$Df[4]), 3)),
    paste("f > ", round(qf(0.95, multinova_summary$Df[3], multinova_summary$Df[4]), 3)),
```

```
    "", "")
)

kable(multinova_table)
qqnorm(multinova_result$residuals,main = "", pch = 16, cex = 0.4,
       col = rgb(1, 0, 0, 0.5));
qqline(multinova_result$residuals)
plot(multinova_result, which = 1,main = "", pch = 16, cex = 0.4,
     col = rgb(0, 0, 1, 0.5), id.n=0)
library(rcompanion)
sink("/dev/null")
hare_result <- scheirerRayHare(yield ~ nf*topo, data = dat2001)
sink()
hare_table <- tibble(
  "Source of Variation" = c("Nitrogen", "Topography", "Interaction", "Error", "Total"),
  df = c(hare_result$Df[1],hare_result$Df[2],hare_result$Df[3],
         hare_result$Df[4],sum(hare_result$Df)),
  "Sum of Squares" = c(round(hare_result[1, "Sum Sq"], 1),
                       round(hare_result[2, "Sum Sq"], 1),
                       round(hare_result[3, "Sum Sq"], 1),
                       round(hare_result[4, "Sum Sq"], 1),
                       round(sum(hare_result[, "Sum Sq"]), 1)),
  H = c(round(hare_result$H[1], 3),round(hare_result$H[2], 3),
        round(hare_result$H[3], 3),"",""),
  "p-value" = c(pformat(hare_result[1, "p.value"]),
                pformat(hare_result[2, "p.value"]),
                pformat(hare_result[3, "p.value"]),"","")
)

kable(hare_table)
perma_result <- pairwisePermutationTest(yield ~ topo,data = dat2001,method = "bonf")
sig_comparisons <- perma_result[perma_result$p.adjust < 0.05, ]

sig_table <- tibble(
        "Comparison" = sig_comparisons$Comparison,
        "Stat" = sig_comparisons$Stat,
        "p-value" = pformat(as.numeric(sig_comparisons$p.value)),
        "p-adjust" = pformat(sig_comparisons$p.adjust)
)

kable(sig_table)
```