

# Chapter 6

## A Short Discussion of Observational Studies

*“Thou shall adjust for what thou can not control.”*

In most scientific studies, the groups being compared do not consist of identical experimental units that have been randomly assigned to receive a treatment. Instead, the groups might be extremely heterogeneous on factors that might be related to a specific response on which you wish to compare the groups. Inferences about the nature of differences among groups in such **observational studies** can be flawed if this heterogeneity is ignored in the statistical analysis.

The following problem emphasizes the care that is needed when analyzing **observational studies**, and highlights the distinction between the **means** and **lsmeans** output for a two-way table. The **data are artificial**, but the conclusions are consistent with an interesting analysis conducted by researchers at Sandia National Laboratories.

A representative sample of 550 high school seniors was selected in 1970. A similar sample of 550 was selected in 1990. The final SAT scores (on a 1600 point scale) were obtained for each student<sup>1</sup>.

---

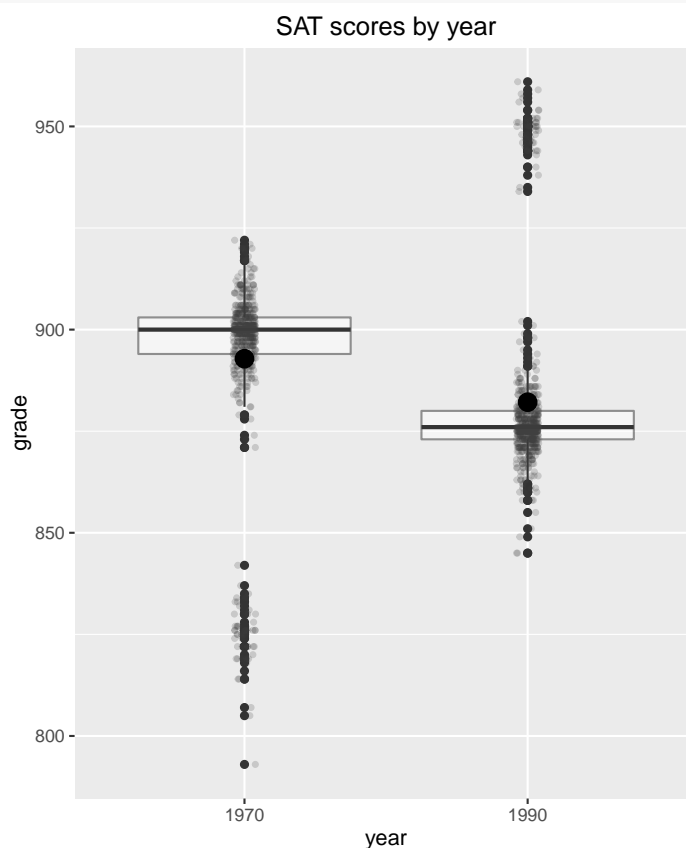
<sup>1</sup>The fake-data example in this chapter is similar to a real-world SAT example illustrated in this paper: “Minority Contributions to the SAT Score Turnaround: An Example of Simpson’s Paradox” by Howard Wainer, *Journal of Educational Statistics*, Vol. 11, No. 4 (Winter, 1986), pp. 239–244 <http://www.jstor.org/stable/1164696>.

The boxplots for the two samples show heavy-tailed distributions with similar spreads. Given the large sample sizes, the  $F$ -test comparing populations is approximately valid even though the population distributions are non-normal.

```
#### Example: SAT
sat <- read.table("http://statacumen.com/teach/ADA2/ADA2_notes_Ch06_sat.dat", header = TRUE)
sat$year <- factor(sat$year)
sat$eth <- factor(sat$eth )

# calculate means by year (also calculated below to illustrate lsmeans())
library(plyr)
sat.mean.y <- ddply(sat, .(year), summarise, m = mean(grade))

# Interaction plots, ggplot
p <- ggplot(sat, aes(x = year, y = grade))
p <- p + geom_boxplot(alpha = 0.5)
p <- p + geom_point(position = position_jitter(w = 0.1, h = 0), colour="gray25", size=1, alpha = 0.2)
p <- p + geom_point(data = sat.mean.y, aes(y = m), size = 4)
#p <- p + geom_line(data = sat.mean.y, aes(y = m), size = 1.5)
p <- p + labs(title = "SAT scores by year")
print(p)
```



A simple analysis might compare the average SAT scores for the two years, to see whether students are scoring higher, lower, or about the same, over time. The one-way **lsmeans** and **means** breakdowns of the SAT scores are identical; the average SAT scores for 1970 and 1990 are 892.8 and 882.2, respectively. The one-way ANOVA, combined with the observed averages, indicates that

the typical SAT score has decreased significantly (10.7 points) over the 20 year period.

```
lm.g.y <- lm(grade ~ year, data = sat
            , contrasts = list(year = contr.sum))
library(car)
# type III SS
Anova(lm.g.y, type=3)

## Anova Table (Type III tests)
##
## Response: grade
##          Sum Sq   Df   F value    Pr(>F)
## (Intercept) 866418325     1 1.7076e+06 < 2.2e-16 ***
## year          31410     1 6.1904e+01 8.591e-15 ***
## Residuals    557117 1098
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

library(plyr)
# balanced with respect to year, so means and lsmeans match
sat.mean.y <- ddply(sat, .(year), summarise, m = mean(grade))
sat.mean.y

##   year      m
## 1 1970 892.8418
## 2 1990 882.1545

library(lsmeans)
lsmeans(lm.g.y, list(pairwise ~ year), adjust = "bonferroni")

## $`lsmeans of year`
##   year  lsmean      SE   df lower.CL upper.CL
##  1970 892.8418 0.9604853 1098 890.9572 894.7264
##  1990 882.1545 0.9604853 1098 880.2700 884.0391
##
## Confidence level used: 0.95
##
## $`pairwise differences of contrast`
##   contrast      estimate      SE   df t.ratio p.value
## 1970 - 1990 10.68727 1.358331 1098   7.868 <.0001
```

Should we be alarmed? Should we be concerned that students entering college have fewer skills than students 20 years ago? Should we be pumping billions of dollars into the bloated bureaucracies of our public school systems with the hope that a few of these dollars might be put to good use in programs to enhance performance? This is the consensus among some people in the know, all of whom wax eloquently about the impending inability of the U.S. to compete in the new global economy.

The SAT study is not a well-designed experiment, where a scientist has controlled all the factors that might affect the response (the SAT score) other than the treatment (the year). Even without these controls, there is no randomization of treatments to students selected from a target population.

The SAT study is an **observational study** of two distinct populations. The observed differences in SAT scores may indeed be due to a decrease in performance. The differences might also be due to factors that make the two populations incomparable for assessing changes in performance over time.

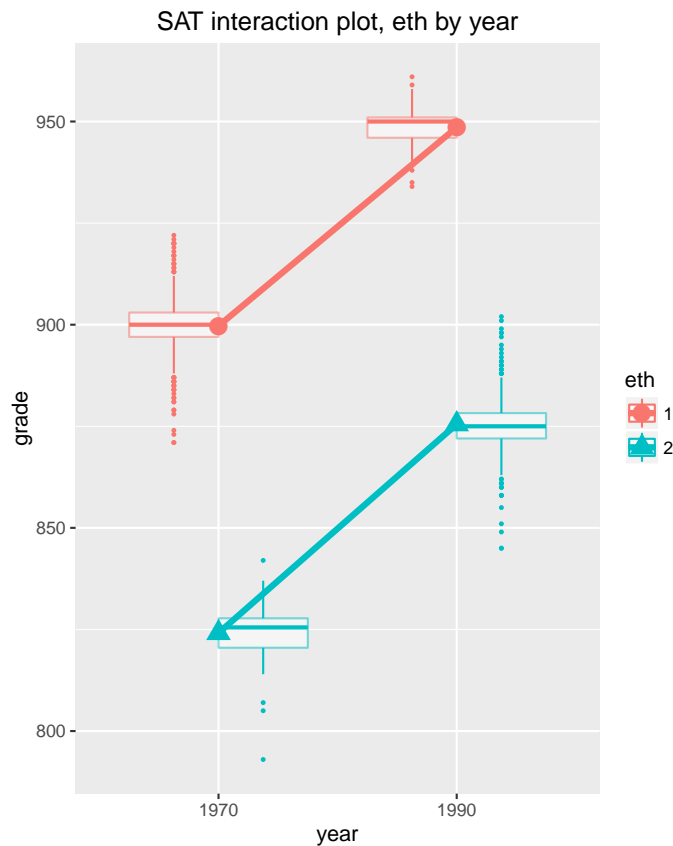
My hypothetical populations have students from two ethnic groups (1 and 2). If you construct box-plots of the SAT scores for the four combinations of ethnicity and year, you see that the typical SAT score **within** each ethnic group has **increased** over time, whereas the typical SAT score **ignoring** ethnicity decreased over time. Is this a paradox, and what are appropriate conclusions in the analysis?

```
sat.mean.ye <- ddply(sat, .(year,eth), summarise, m = mean(grade))
sat.mean.ye

##   year eth      m
## 1 1970  1 899.712
## 2 1970  2 824.140
## 3 1990  1 948.560
## 4 1990  2 875.514

# Interaction plots, ggplot
library(ggplot2)
p <- ggplot(sat, aes(x = year, y = grade, colour = eth, shape = eth))
p <- p + geom_boxplot(alpha = 0.5, outlier.size=0.5)
p <- p + geom_point(data = sat.mean.ye, aes(y = m), size = 4)
p <- p + geom_line(data = sat.mean.ye, aes(y = m, group = eth), size = 1.5)
p <- p + labs(title = "SAT interaction plot, eth by year")
print(p)

#p <- ggplot(sat, aes(x = eth, y = grade, colour = year, shape = year))
#p <- p + geom_boxplot(alpha = 0.5, outlier.size=0.5)
#p <- p + geom_point(data = sat.mean.ye, aes(y = m), size = 4)
#p <- p + geom_line(data = sat.mean.ye, aes(y = m, group = year), size = 1.5)
#p <- p + labs(title = "SAT interaction plot, year by eth")
#print(p)
```



I fit a two-factor model with year and ethnicity effects plus an interaction. The two-factor model gives a method to compare the SAT scores over time, after **adjusting** for the effect of ethnicity on performance. The  $F$ -test for comparing years adjusts for ethnicity because it is based on comparing the average SAT scores across years after averaging the cell means over ethnicities, thereby eliminating from the comparison of years any effects due to changes in the ethnic composition of the populations. The two-way analysis is preferable to the unadjusted one-way analysis which **ignores** ethnicity.

```
lm.g.y.e.ye <- lm(grade ~ year * eth, data = sat
                  , contrasts = list(year = contr.sum, eth = contr.sum))
## CRITICAL!!! Unbalanced design warning.
## The contrast statement above must be included identifying
## each main effect with "contr.sum" in order for the correct
## Type III SS to be computed.
## See http://goanna.cs.rmit.edu.au/~fscholer/anova.php
library(car)
# type III SS
Anova(lm.g.y.e.ye, type=3)
## Anova Table (Type III tests)
##
## Response: grade
```

```
##           Sum Sq   Df   F value  Pr(>F)
## (Intercept) 286085884   1 5.7022e+06 < 2e-16 ***
## year        228283     1 4.5501e+03 < 2e-16 ***
## eth         501984     1 1.0005e+04 < 2e-16 ***
## year:eth     145       1 2.8904e+00 0.08939 .
## Residuals   54988 1096
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The year and ethnicity main effects are significant in the two factor model, but the interaction is not. The marginal **lsmeans** indicate that the average SAT score increased significantly over time when averaged over ethnicities. This is consistent with the cell mean SAT scores increasing over time **within** each ethnic group. Given the lack of a significant interaction, the expected increase in SAT scores from 1970 to 1990 **within each** ethnic group is the difference in marginal averages:  $912.0 - 861.9 = 50.1$ .

```
library(plyr)
# unbalanced, don't match (lsmeans is correct)
sat.mean.y <- ddply(sat, .(year), summarise, m = mean(grade))
sat.mean.y

##   year      m
## 1 1970 892.8418
## 2 1990 882.1545

library(lsmeans)
lsmeans(lm.g.y.e.ye, list(pairwise ~ year), adjust = "bonferroni")
## NOTE: Results may be misleading due to involvement in interactions
## $`lsmeans of year`
##   year lsmean      SE   df lower.CL upper.CL
## 1970 861.926 0.5253021 1096 860.8953 862.9567
## 1990 912.037 0.5253021 1096 911.0063 913.0677
##
## Results are averaged over the levels of: eth
## Confidence level used: 0.95
##
## $`pairwise differences of contrast`
##   contrast      estimate      SE   df t.ratio p.value
## 1970 - 1990  -50.111 0.7428893 1096  -67.454 <.0001
##
## Results are averaged over the levels of: eth
# unbalanced, don't match (lsmeans is correct)
sat.mean.e <- ddply(sat, .(eth), summarise, m = mean(grade))
sat.mean.e

##   eth      m
```

```

## 1 1 904.1527
## 2 2 870.8436

lsmeans(lm.g.y.e.ye, list(pairwise ~ eth), adjust = "bonferroni")
## NOTE: Results may be misleading due to involvement in interactions
## $`lsmeans of eth`
## eth lsmean SE df lower.CL upper.CL
## 1 924.136 0.5253021 1096 923.1053 925.1667
## 2 849.827 0.5253021 1096 848.7963 850.8577
##
## Results are averaged over the levels of: year
## Confidence level used: 0.95
##
## $`pairwise differences of contrast`
## contrast estimate SE df t.ratio p.value
## 1 - 2 74.309 0.7428893 1096 100.027 <.0001
##
## Results are averaged over the levels of: year
# unbalanced, but highest-order interaction cell means will match
sat.mean.ye <- ddply(sat, .(year,eth), summarise, m = mean(grade))
sat.mean.ye
## year eth m
## 1 1970 1 899.712
## 2 1970 2 824.140
## 3 1990 1 948.560
## 4 1990 2 875.514

lsmeans(lm.g.y.e.ye, list(pairwise ~ year | eth), adjust = "bonferroni")
## $`lsmeans of year | eth`
## eth = 1:
## year lsmean SE df lower.CL upper.CL
## 1970 899.712 0.3167691 1096 899.0905 900.3335
## 1990 948.560 1.0017118 1096 946.5945 950.5255
##
## eth = 2:
## year lsmean SE df lower.CL upper.CL
## 1970 824.140 1.0017118 1096 822.1745 826.1055
## 1990 875.514 0.3167691 1096 874.8925 876.1355
##
## Confidence level used: 0.95
##
## $`pairwise differences of contrast, eth | eth`
## eth = 1:
## contrast estimate SE df t.ratio p.value
## 1970 - 1990 -48.848 1.050604 1096 -46.495 <.0001
##
## eth = 2:
## contrast estimate SE df t.ratio p.value
## 1970 - 1990 -51.374 1.050604 1096 -48.899 <.0001

```

```

lsmeans(lm.g.y.e.ye, list(pairwise ~ eth | year), adjust = "bonferroni")
## $`lsmeans of eth | year`
## year = 1970:
##   eth  lsmean      SE    df lower.CL upper.CL
##   1    899.712 0.3167691 1096 899.0905 900.3335
##   2    824.140 1.0017118 1096 822.1745 826.1055
##
## year = 1990:
##   eth  lsmean      SE    df lower.CL upper.CL
##   1    948.560 1.0017118 1096 946.5945 950.5255
##   2    875.514 0.3167691 1096 874.8925 876.1355
##
## Confidence level used: 0.95
##
## $`pairwise differences of contrast, year | year`
## year = 1970:
##   contrast estimate      SE    df t.ratio p.value
##   1 - 2         75.572 1.050604 1096  71.932  <.0001
##
## year = 1990:
##   contrast estimate      SE    df t.ratio p.value
##   1 - 2         73.046 1.050604 1096  69.528  <.0001

```

As noted in the insulin analysis, the marginal **lsmeans** and **means** are different for unbalanced two-factor analyses. The marginal **means** ignore the levels of the other factors when averaging responses. The marginal **lsmeans** are averages of cell means over the levels of the other factor. Thus, for example, the 1970 **mean** SAT score of 892.8 is the average of the 550 scores selected that year. The 1970 **lsmeans** SAT score of 861.9 is midway between the average 1970 SAT scores for the two ethnic groups:  $861.9 = (899.7 + 824.1)/2$ . Hopefully, this discussion also clarifies why the year marginal **means** are identical in the one and two-factor analyses, but the year **lsmeans** are not.

The 1970 and 1990 marginal **means** estimate the typical SAT score ignoring all factors that may influence performance. These marginal averages are not relevant for **understanding** any trends in performance over time because they do not account for changes in the composition of the population that may be related to performance.

The average SAT scores (ignoring ethnicity) decreased from 1970 to 1990 because the ethnic composition of the student population changed. Ten out of



every eleven students sampled in 1970 were from the first ethnic group. Only one out of eleven students sampled in 1990 was from this group. Students in the second ethnic group are underachievers, but they are becoming a larger portion of the population over time. The decrease in average (**means**) performance inferred from comparing 1970 to 1990 is **confounded** with the increased representation of the underachievers over time. Once ethnicity was taken into consideration, the typical SAT scores were shown to have increased, rather than decreased.

In summary, the one-way analysis ignoring ethnicity is valid, and allows you to conclude that the typical SAT score has decreased over time, but it does not provide any insight into the nature of the changes that have occurred. A two-factor analysis backed up with a comparison of the marginal **lsmeans** is needed to compare performances over time, adjusting for the changes in ethnic composition.

The Sandia study reached the same conclusion. The Sandia team showed that the widely reported decreases in SAT scores over time are due to changes in the ethnic distribution of the student population over time, with individuals in historically underachieving ethnic groups becoming a larger portion of the student population over time.

A more complete analysis of the SAT study would adjust the SAT scores to account for other potential confounding factors, such as sex, and differences due to the number of times the exam was taken. These confounding effects are taken into consideration by including them as effects in the model.

The interpretation of the results from an observational study with several effects of interest, and several confounding variables, is greatly simplified by eliminating the insignificant effects from the model. For example, the year by ethnicity interaction in the SAT study might be omitted from the model to simplify interpretation. The year effects would then be estimated after fitting a two-way additive model with year and ethnicity effects only. The same approach is sometimes used with designed experiments, say the insulin study that we analyzed earlier.

**An important caveat** The ideas that we discussed on the design and analysis of experiments and observational studies are universal. They apply regardless of whether you are analyzing categorical data, counts, or measurements.