solution-4.R

wickhamc

Fri Oct 30 11:44:49 2015

(**Grading Note** You were awarded points on your observations on plots, and correct statements about robustness, not your closeness to these solutions.)

# Problem 1

1. This example has within group dependence since the measurments are taken across years and we expect some serial dependence between consecutive years. This may also result in some between group dependence, since year 10 and year 11, might be expected to be similar, but are in different groups. There might also be some between group dependence since the roads in the 2nd group are the same roads in the 1st group. Roads are also interconnected, bad traffic at a given road can cause more accidents at that road but also cause more accidents at roads that are connected to it. Note that this is only a problem if we treat the observations as number of accidents on each road. If instead our observations are total yearly number of accidents over all roads this dependence isn't a concern.
2. This example has between group dependence. The similar genetic makeup of the twins leads to dependence in their scores.
3. This example has within group dependence. People from the same household will tend to have similar respiratory capacity. You might also imagine this as a spatial dependence. Members of the same household are in close proximity and are expected to be exposed to the same air quality conditions which may result in more similar respiratory health, and at a higher level, houses in close proximity are expected to be exposed to the same air quality conditions which may result in more similar respiratory health across households.

# Problem 2

a

library(reshape2)
library(ggplot2)
source(url("http://stat511.cwick.co.nz/code/stat\_qqline.r"))

## Loading required package: proto

cdc <- read.csv(url("http://stat511.cwick.co.nz/homeworks/cdc.csv"))
cdc$wt\_diff <- with(cdc, weight - wtdesire)

qplot(wt\_diff, data = cdc) + facet\_wrap(~ exerany, ncol = 1, scale = "free\_y")

## stat\_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat\_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.

## Warning: closing unused connection 5
## (http://stat511.cwick.co.nz/code/stat\_qqline.r)



qplot(sample = wt\_diff, data = cdc) +
 facet\_wrap(~ exerany, ncol = 1) +
 stat\_qqline()



sd(subset(cdc, exerany == 0)$wt\_diff)/sd(subset(cdc, exerany == 1)$wt\_diff)

## [1] 1.526975

Looking at the histograms the spread in desired weight loss, is roughly similar between the two groups. (If you calculated the actual ratio of sample sds, you get about 1.5, this a big difference, but it actually mostly due to a few outliers in the not exercising group). There is no evidence of a gross violation of the equal population standard deviations assumption.

The normal probability plots and histograms suggest that these sample data are not coming from a Normal population. There is a large peak at zero desired weight loss, and longer tails than we might expect from a Normal distribution. However, we have large sample sizes ($n\_{0}$ = 279, $n\_{1}$ = 721), so expect robustness to this violation.

The two sample t-test should be valid.

(Charlotte's conclusion: I'd do the t-test, but I'd spend some time checking out those outliers.)

b

library(Sleuth3)
head(ex0125)

## Group Zinc
## 1 A 1.31
## 2 A 1.45
## 3 A 1.12
## 4 A 1.16
## 5 A 1.30
## 6 A 1.50

qplot(Zinc,data=ex0125) + facet\_wrap(~Group,ncol=1)

## stat\_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat\_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.



qplot(sample=Zinc,data=ex0125) + facet\_wrap(~Group,ncol=1) + stat\_qqline()



The histograms show evidence that the two groups have different spreads, although the magnitude of this difference might be just attributible to the sampling variation. The sample sizes are similar so we have some robustness to this assumption in this case. This is a randomized experiment, so this also provides some evidence the additive treatment model may not be appropriate.

The normal probability plots show some evidence of non-Normality, but again with a smallish sample it is hard to know if this is sampling variation. The sample sizes of 20 and 19 may be large enough to rely on robustness to this assumption.

(Charlotte's conclusion: the possibility of unequal standard deviations is of biggest concern. A randomization test could be done here but using the more general null and alternative. It might be more appropriate to think of something other than an additive treatment effect model.)

c

ex0318 <- data.frame(expenditure = c(20.1, 22.9, 18.8, 20, 20.9, 22.7, 21.4,
 20, 38.5, 25.8, 22, 23, 37.6, 30, 24.5), group = rep(c("Nontrauma", "Trauma"),
 c(8, 7)))
head(ex0318)

## expenditure group
## 1 20.1 Nontrauma
## 2 22.9 Nontrauma
## 3 18.8 Nontrauma
## 4 20.0 Nontrauma
## 5 20.9 Nontrauma
## 6 22.7 Nontrauma

qplot(expenditure,data=ex0318) + facet\_wrap(~group,ncol=1)

## stat\_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat\_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.



qplot(sample=expenditure,data=ex0318) + facet\_wrap(~group,ncol=1, scale = "free") + stat\_qqline()



The histograms show the two groups have very different spreads, an indication that the equal population standard deviation assumption is violated.
We may have some robustness to this violation with the roughly equal sample sizes. The assumption of normality seems reasonable, but we must be a bit cautious with such small sample sizes.

(Charlotte's conclusion: Again the unequal standard deviations are the most serious concern especially when we can't verify the Normality assumption. I'd probably do Welch's t-test, and reason that at least the populations aren't grossly non-Normal, and the samples size might be big enough).