Stat 242 Quiz – Topics Drawn from Sections 5.5 and Chapter 3

What's Your Name? ________________________________

In each of the following data problems there is a potential violation of the assumption of independence. Explain what the potential problem is in a sentence or two.

1. Researchers interested in learning the effects of speed limits on traffic accidents recorded the number of accidents per year for each of 10 consecutive years on roads in a state with speed limits of 90 km/h. They also recorded the number of accidents for the next 7 years on the same roads after the speed limit had been increased to 110 km/h. They conducted statistical analyses to compare the mean number of accidents per year when the speed limit was 90 km/h and the mean number of accidents per year when the speed limit was 110 km/h.

The same roads were measured multiple times at both speed limits, and those measurements are probably not independent of each other. For example, some roads may be more dangerous and may consistently have more accidents. Multiple measurements for the same road cannot be treated as independent. Additionally, the number of accidents may be more similar in consecutive years than in years that are far apart. Measurements in consecutive years may not be independent.

2. Researchers interested in investigating the effect of indoor pollution on respiratory health randomly selected houses in a particular city. Each house was monitored for nitrogen dioxide concentration and categorized as being either high or low on the nitrogen dioxide scale. Each member of the household was measured for respiratory health in terms of breathing capacity. They conducted statistical analyses to compare the mean breathing capacity score for people in houses with high nitrogen dioxide levels and the mean breathing capacity score for people in houses with low nitrogen dioxide levels.

The individuals within a given household may be related, and may therefore have similar respiratory health for genetic reasons; their measurements cannot be treated as independent.

(this page left intentionally blank)

Stat 242 Quiz – Topics Drawn from Sections 5.5 and Chapter 3

What's Your Name? ______________________________

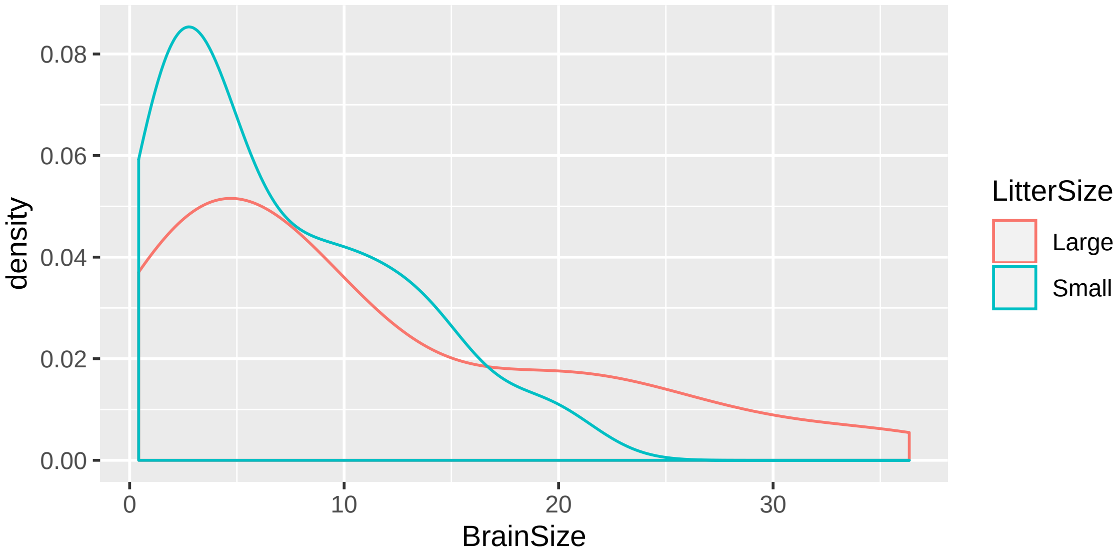
We have data for the relative brain weights (brain weight divided by body weight) for 51 species of mammal whose average litter size is less than 2, and for 45 species of mammal whose average litter size is greater than or equal to 2. (Across both groups, we have data for 96 species.)

The following R code displays the first few rows of the data set, plots of the data, and the standard deviations of the relative brain weights for the observations within each group.

```
head(brain_sizes)
```

```
##   BrainSize LitterSize
## 1      0.42      Small
## 2      0.86      Small
## 3      0.88      Small
## 4      1.11      Small
## 5      1.34      Small
## 6      1.38      Small
```

```
ggplot(data = brain_sizes, mapping = aes(x = BrainSize, color = LitterSize)) +
  geom_density()
```



```
brain_sizes %>%
  group_by(LitterSize) %>%
  summarize(
    sd_brain_size = sd(BrainSize)
  )
```

```
## # A tibble: 2 x 2
##   LitterSize sd_brain_size
##   <fct>              <dbl>
## 1 Large               9.84
## 2 Small               5.46
```

1. Discuss any conditions for the anova model that are not satisfied.

The standard deviations are quite different for the two groups. Although the ratio of standard deviations is less than 2, we'd really prefer them to be more similar.

Relative brain sizes in each group are not normally distributed; in the density plot, both distributions are skewed right.

I often have doubts about independence. In this case, I wonder if perhaps some similar species might be included in our data set. For example, if there were two species of mouse included in the data, I would expect their residuals to be similar.

2. Suggest a strategy that could be used to help address the issues you identified in problem 1.

We could try a transformation of brain size.