Project 1

Reproductive Research using RMarkdown

Huiyue Li, Ying Gao, & Mian Wei<- Huiyue Li

Introduction

This exercise will use the **esoph** dataset, which is part of base R. The *esoph* dataset contains data from a case-control study of esophageal cancer in Ille-et-Vilaine, France.

The *esoph* dataset has 88 rows and 5 variables, including:
1. Age group
2. Alcohol consumption
3. Tobacco consumption
4. Number of cases
5. Number of controls

Data Details

- This dataset is aggregated, which means each row represents one distinct group that includes the number of cases and controls in each group.
- There are six age groups:
  - 25-34
  - 35-44
  - 45-54
  - 55-64
  - 65-74
  - 75+
- There are four alcohol consumption groups:
  - 0-39 *g/day*
  - 40-79 *g/day*
  - 80-119 *g/day*
  - 120+ *g/day*
- There are four tobacco consumption groups:
  - 0-9 *g/day*
  - 10-19 *g/day*
  - 20-29 *g/day*
  - 30+ *g/day*

Data Processing

We want to know the percent of cases by exposure group. The code below calculates the total number of participants and the percentages for cases.

```
library(dplyr)

#create new variables
esoph$total = esoph$ncases + esoph$ncontrols
esoph$cases_p = esoph$ncases/esoph$total

#calculate mean percent of cases by age group
cases_by_age = esoph %>% group_by(agegp) %>% summarize(mean=100*mean(ca
ses_p))
```

Results

Table

Create a table showing the mean percentage of cases by age group.

Table 1: Percent Cases by Age

Age Group	Percent Cases
25-34	3.33
35-44	6.55
45-54	23.57
55-64	30.33
65-74	31.25
75+	32.62

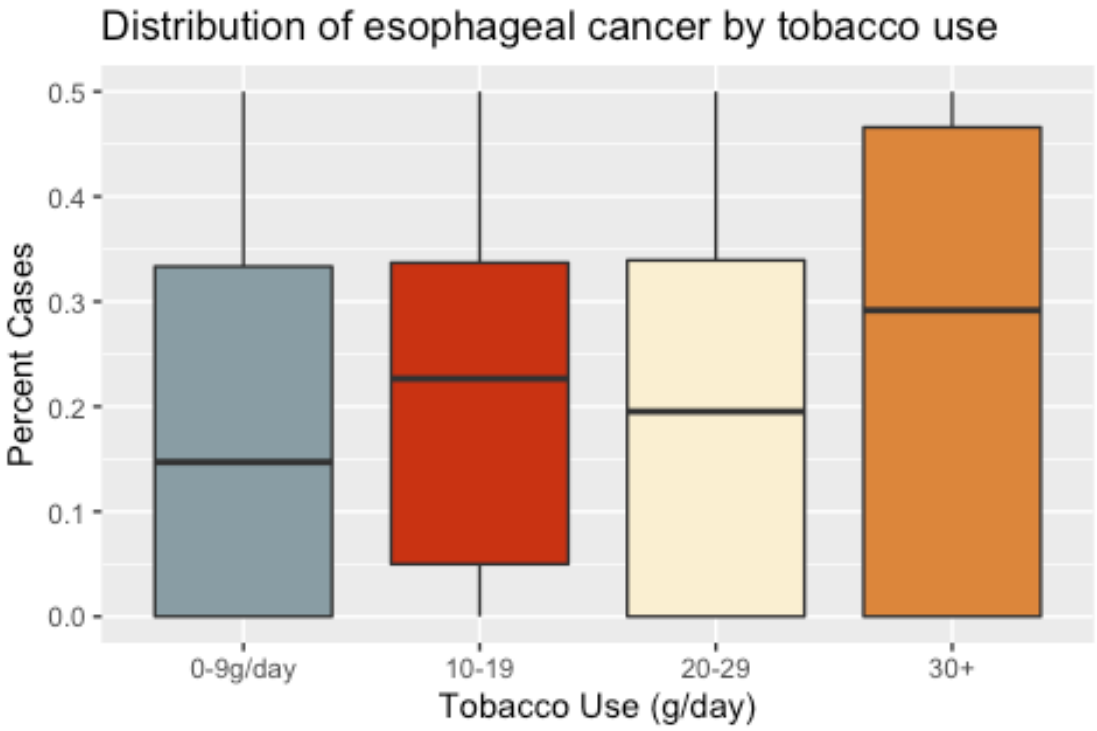
Figure

The figure below shows the relationship between tobacco usage and esophageal cancer.

```
#code to create side-by-side boxplots

library(ggplot2)
library(wesanderson)

colors = wes_palette("Royal1", 4, type = c("discrete"))
ggplot(data=esoph, aes(x=tobgp, y=cases_p)) + geom_boxplot(fill=colors)
 + ggtitle("Distribution of esophageal cancer by tobacco use") + xlab("
Tobacco Use (g/day)") + ylab("Percent Cases")
```


Distribution of esophageal cancer by tobacco use
Percent Cases
0.5
0.4
0.3
0.2
0.1
0.0
0-9g/day
10-19
20-29
30+
Tobacco Use (g/day)