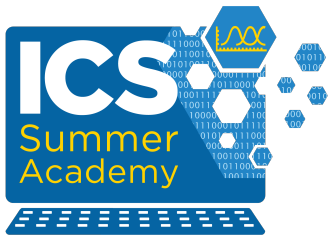
AIzheimer's Data Analysis

Babak Shahbaba and Sam Behseta
UC Irvine and California State University Fullerton

July and August, 2022



Bivariate Data Summarization and Inference

- In this lesson, we review the effective ways of summarizing and analyzing data associated with two variables.

Bivariate Data Summarization and Inference

- In this lesson, we review the effective ways of summarizing and analyzing data associated with two variables.
- There are multiple scenarios: both variables are categorical, one is categorical while the other is continuous, and both are numerical (and continous).

Bivariate Data Summarization and Inference

- In this lesson, we review the effective ways of summarizing and analyzing data associated with two variables.
- There are multiple scenarios: both variables are categorical, one is categorical while the other is continuous, and both are numerical (and continous).
- In principle, we can have a third kind of variable, called discrete, which is somewhat broader than categorical. We can have features that refer to number of incidence or counts of a random phenomenon. Think of *size of a family*, as a feature. This variable is not categorical per se, but the good news is historically, the methods for categorical variables mostly co-incides with the theory and methods for expressing and understanding discrete variables. We'll get to discuss that concept soon!

Summarizing and Visualizing Two Categorical Variables

- Historically, the main them here revolve around what is known as a *contingency table*. This is a table that reflects the frequency (or relative frequency) of the joint categories of each variable.

Summarizing and Visualizing Two Categorical Variables

- Historically, the main them here revolve around what is known as a *contingency table*. This is a table that reflects the frequency (or relative frequency) of the joint categories of each variable.
- We can summarize the information through marginal and total sums or proportions of the table.

Summarizing *diagnosis* and *female* Jointly

- Let's recall the information we have from these variables:

```
diag=as.factor(diagnosis)
summary(diag)

female=as.factor(female)
summary(female)
```

Summarizing *diagnosis* and *female* Jointly

- Let's recall the information we have from these variables:

```
diag=as.factor(diagnosis)
summary(diag)

female=as.factor(female)
summary(female)
```

- We would like to be able to cross-correlate them into a table. The command is not surprisingly called *table*!

```
>t<- table(diagnosis,female)
>t
         female
diagnosis    0    1
        0  529 1005
        1  327  286
        2  295  258
```

Summarizing *diagnosis* and *female* Jointly

- This is a so-called 3 by 2 contingency table. Before fully immersing ourselves to it, let us first extract more information from this table!

```
tab_sum <- addmargins(t, FUN = sum)
tab_sum
```

Summarizing *diagnosis* and *female* Jointly

- This is a so-called 3 by 2 contingency table. Before fully immersing ourselves to it, let us first extract more information from this table!

```
tab_sum <- addmargins(t, FUN = sum)
tab_sum
```

- Tell me what I did in the code below, and why that is important in the context of our analysis?

```
t.rate=t/sum(t)
t.rate
tab_rate <- addmargins(t.rate, FUN = sum)
tab_rate
```

Visualizing *diagnosis* and *female* Jointly

- Let's proceed with a familiar way of representing this data. Make sure to analyze the code carefully!

```
counts<-c(529,1005,327,286,295,258)
diagnosis<-c("normal","mild","severe")
female<-c("no","yes")
new.data<-data.frame(counts,diagnosis,female)
```

```
> new.data
  counts diagnosis female
1    529    normal     no
2   1005      mild    yes
3    327    severe     no
4    286    normal    yes
5    295      mild     no
6    258    severe    yes
```

Visualizing *diagnosis* and *female* Jointly

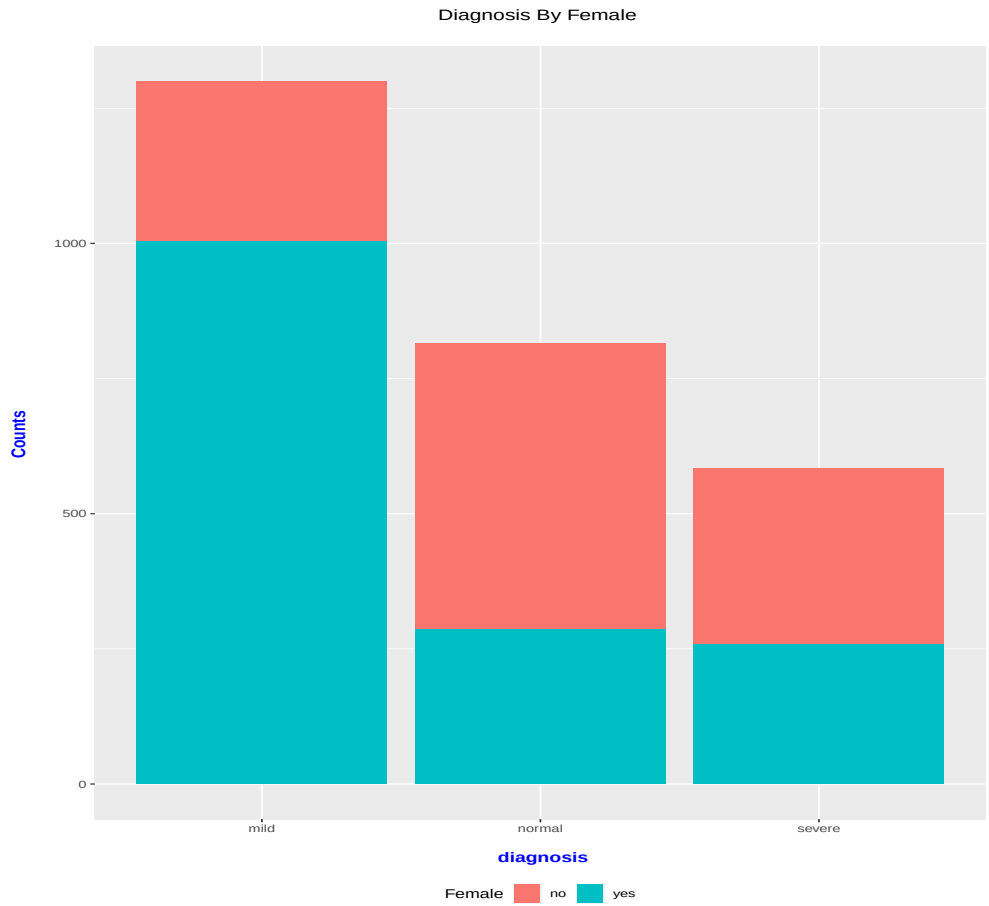
- *ggplot2* allows us to make the visualization elaborate!

```
ggplot(data = new.data, aes(x = diagnosis, y = counts, fill = female)) +
  geom_bar(stat = "identity") +
  labs(x = "\n diagnosis", y = "Counts \n",
       title = "Diagnosis By Female \n",
       fill = "Female") +


theme(plot.title = element_text(hjust = 0.5),
axis.title.x = element_text(face="bold", colour="blue", size = 12),
axis.title.y = element_text(face="bold", colour="blue", size = 12),
legend.position = "bottom")
```

Visualizing *diagnosis* and *female* Jointly

- And eventually, below is what would emerge from the code:



Testing the Independence of *diagnosis* and *female*

- The R command is *chisq.test*. Let's apply it on this data, and also verify its values with simple calculations. Remember that

$$\chi^2 = \sum_{i=1}^{nrow} \sum_{j=1}^{ncol} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Testing the Independence of *diagnosis* and *female*

- The R command is *chisq.test*. Let's apply it on this data, and also verify its values with simple calculations. Remember that

$$\chi^2 = \sum_{i=1}^{nrow} \sum_{j=1}^{ncol} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- Now let's run the χ^2 test for these two variables:

```
c<-chisq.test(t)
c
c$observed
c$expected
```

Testing the Independence of *diagnosis* and *female*

- The R command is *chisq.test*. Let's apply it on this data, and also verify its values with simple calculations. Remember that

$$\chi^2 = \sum_{i=1}^{nrow} \sum_{j=1}^{ncol} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- Now let's run the χ^2 test for these two variables:

```
c<-chisq.test(t)
c
c$observed
c$expected
```

- Let's further check the output of the χ^2 test:

```
my.chisq=sum((c$observed-c$expected)^2/c$expected)
my.chisq
1-pchisq(my.chisq,2)
```

Class Activity 1 – Group Work

- Collapse the variable *educ* into three groups: 0-12, 12-17, 17-25 years of education. Called the newly created variable *educ.cat*!
- Perform a full analysis of cross-correlation between *female* and *educ.cat*.

Hint: Below are two ways to accomplish the disrectization task. I like to refer to first approach as the *puzzle-solving* approach! The second one utilizes the important package *deplyr*:

- Method 1:

```
educ.cat=numeric(2700)
educ.cat[educ>=0 & educ <=12]<-1
educ.cat[educ>12 & educ <=17]<-2
educ.cat[educ>17 & educ <=25]<-3
educ.cat<-as.factor(educ.cat)
head(educ.cat)
```

- Method 2:

```
educ1.cat <- educ %>%
  mutate(category=cut(educ, breaks=c(0,12,17,25), labels=c("1","2","3")))
educ1.cat
```

Summarizing and Visualizing a Categorical Versus a Continuous Variable

- This is somewhat straightforward, as the task would revolve around summarizing, and thereby contrasting, the continuous vairable across the levels of the categorical one.

Summarizing and Visualizing a Categorical Versus a Continuous Variable

- This is somewhat straightforward, as the task would revolve around summarizing, and thereby contrasting, the continuous vairable across the levels of the categorical one.
- We can summarize the data, say with its mean or the median, for each level of the categorical variable, taking into account the variation in each level.

Summarizing and Visualizing a Categorical Versus a Continuous Variable

- This is somewhat straightforward, as the task would revolve around summarizing, and thereby contrasting, the continuous vairable across the levels of the categorical one.
- We can summarize the data, say with its mean or the median, for each level of the categorical variable, taking into account the variation in each level.
- The easiest way to see this is to use the familiar variables *diagnosis* and *vol* (please see the previous lecture for the full description of *vol*.)

Summarizing *vol* By *diagnosis*

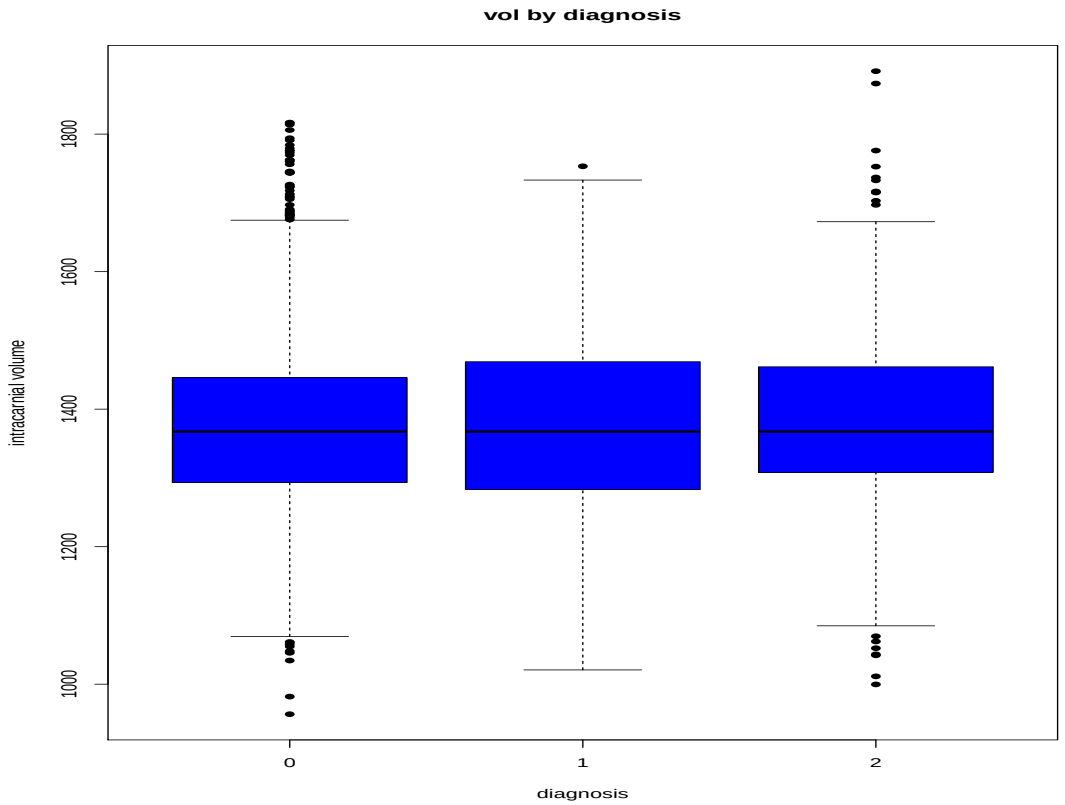
We can begin by summarizing the brain volume associated with each category of diagnosis:

```
summary(vol[diag==0])
summary(vol[diag==1])
summary(vol[diag==2])
```

Visualizing *vol* By *diagnosis*

- Let us quickly complement that with a *side-by-side boxplot*:

```
boxplot(vol~diag1,col="blue",pch=16,xlab="diagnosis",
ylab="intracarnial volume")
```



Two-Sample Hypothesis Test with R

- Below, we will test the hypothesis that whether the mean of the variable *naccicv* or *vol* is the same among the *female* and *others* groups:

```
vol.f<-vol[female=="yes"]
vol.o<-vol[female=="no"]

summary(vol.f)
summary(vol.o)

> t.test(vol.f,vol.o)

Welch Two Sample t-test

data:  vol.f and vol.o
t = 0.47365, df = 2697.3, p-value = 0.6358
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -7.717054 12.632580
sample estimates:
mean of x mean of y
 1378.147  1375.689
```

Relationship Among Two Continuous Variables

- Let's consider the correlation between the right hippocampus volume of the subjects versus their right hippocampus volume. There should be a significantly high correlation among the two.

```
plot(lhippo,rhippo,pch=16,xlab="left hippocampus volume",
ylab="right hippocampus volume")

cor(lhippo,rhippo)
cor.test(lhippo,rhippo)
```