Depicting the Distribution of Two Discrete Variables

Using Side-by-Side Pie Charts and Mosaic Charts

Christopher Kuetsinya

MATH 6490 Statistical Graphics

October 22, 2024

Introduction

- In most cases, the data we seek to analyze contains two or more variables
- We can utilize statistical graphs to find patterns and variations
- In a setting of two discrete variables, what graphical displays can we use?

Side-by-Side Pie Charts & Mosaic Charts

Objective:

- To illustrate the bivariate distribution of the two discrete variables
- Visualizing Conditional Distributions

Side-by-side pie charts and mosaic charts are typically used to compare the proportional distribution of categories across different groups.

Side-by-Side Pie Charts

It uses two or more pie charts to compare how two categorical variables are distributed across different groups or categories.

- Each pie chart represents the distribution of one variable
- Place them next to each other to compare the proportions of categories between the two variables visually
- Different sizes of pie charts can be used to depict variations in proportions
- Comparison is based upon area as the measure of size and not radius.

Example: Eye Color vs. Hair Color Distribution

	Hair Color			
Eye Color	**Black**	**Brunette**	**Red**	**Blond**
Brown	68	119	26	7
Hazel	15	54	14	10
Green	5	29	14	16
Blue	20	84	17	94

Side-by-Side Pie Charts using pie function

```
haireye<-matrix(data=c(94,16,10,7,17,14,14,26,84,29,54,119,20,5,15,68),
                nrow=4,ncol=4,byrow=TRUE,
                dimnames=list(c("Blond","Red","Brunette","Black"),
                              c("Blue","Green","Hazel","Brown")))
haireye<-t(haireye)

par(fin=c(4.45,4.45),pin=c(4.45,4.45),mfrow=c(2,2),mai=c(0.1,0.3,0.2,0.3))

pie(haireye[,4],clockwise=TRUE,col=gray.colors(4)[4:1],main="Black")
pie(haireye[,3],clockwise=TRUE,col=gray.colors(4)[4:1],main="Brunette")
pie(haireye[,2],clockwise=TRUE,col=gray.colors(4)[4:1],main="Red")
pie(haireye[,1],clockwise=TRUE,col=gray.colors(4)[4:1],main="Blond")
```

Side-by-Side Pie Charts using pie function

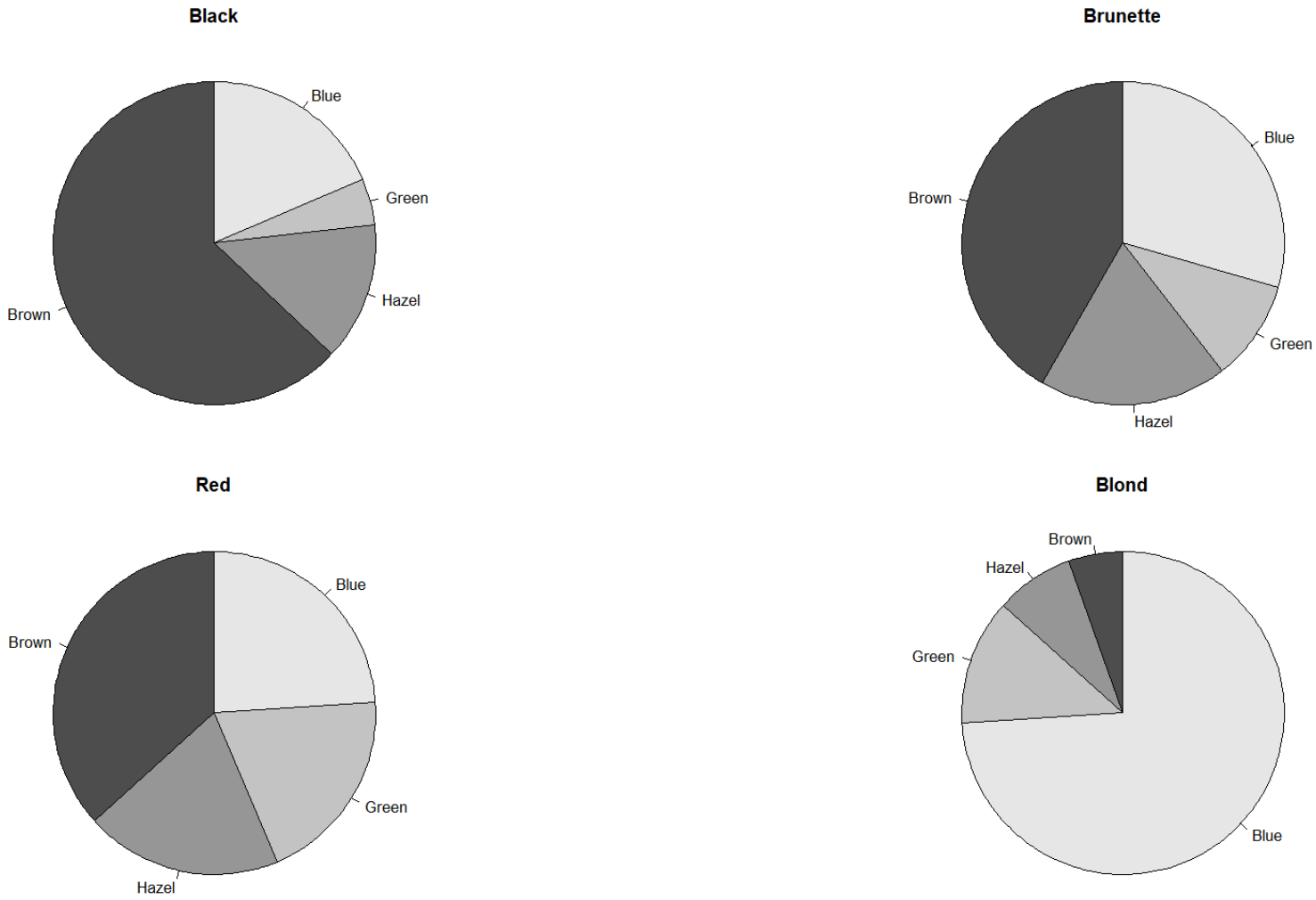

Figure 1 Side-by-side pie chart of conditional eye-color relative frequency grouped within hair color

Side-by-Side Pie Charts using pie function

- Side-by-side pie charts cannot be used to depict the bivariate distribution of discrete variables if the pies are the same size.
- Adjust the size of the pie to depict the relative count for each eye color
- Comparison is based upon area as the measure of size and not radius

```
hesum<-apply(haireye,1,sum)
hesum<-0.8*sqrt(hesum/hesum[4])
par(fin=c(4.45,4.45),pin=c(4.45,4.45),mfrow=c(2,2),mai=c(0.2,0.4,0.2,0.2))

pie(haireye[4,],clockwise=TRUE,col=gray.colors(4)[4:1],main="Brown",radius=hesum[4])
pie(haireye[3,],clockwise=TRUE,col=gray.colors(4)[4:1],main="Hazel",radius=hesum[3])
pie(haireye[2,],clockwise=TRUE,col=gray.colors(4)[4:1],main="Green",radius=hesum[2])
pie(haireye[1,],clockwise=TRUE,col=gray.colors(4)[4:1],main="Blue",radius=hesum[1])
```

Side-by-Side Pie Charts using pie function

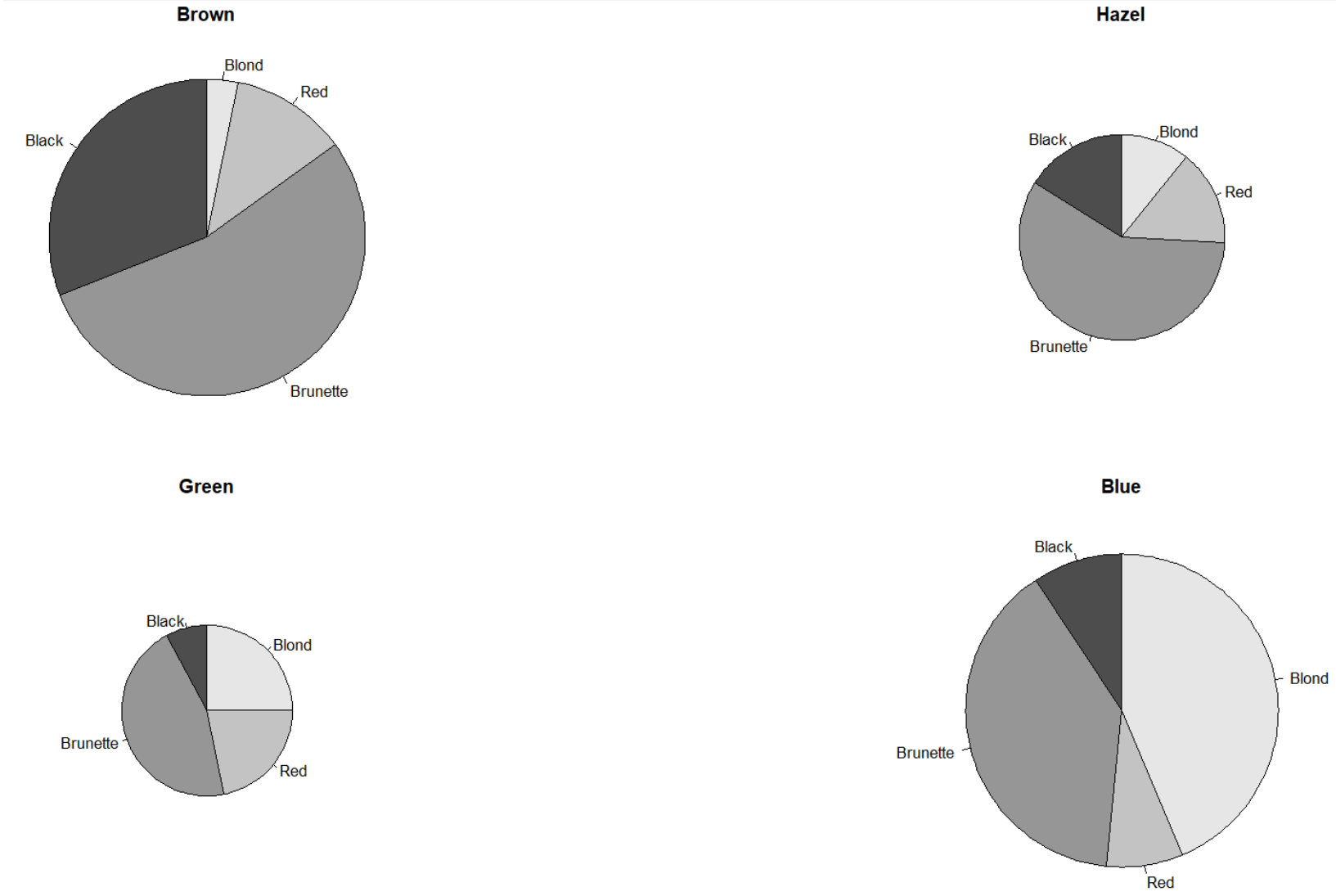

Figure 2 Side-by-side pie chart of conditional eye-color relative frequency grouped within hair color

Limitations of Side-by-Side Pie Charts

- Difficult to compare sizes across different pies
- Making radius proportional to the marginal counts could lead to exaggeration of the differences among the marginal counts
- When variables have many groups the pie chart becomes messy

"A table is nearly always better than a dumb pie chart" – Tufte

Mosaic Charts

It is a graphical display of the distribution of two discrete variables in which each count is represented by a rectangle of area proportional to the count.

- Area of each tile is proportional to frequency.
- Focuses attention on **bivariate relationships**.

Mosaic Charts using mosaicplot function

```
haireye<-matrix(data=c(20,5,15,68,84,29,54,119,17,14,14,26,94,16,10,7),
                nrow=4,ncol=4,byrow=TRUE,
                dimnames=list(c("Black","Brunette","Red","Blond"),
                              c("Blue","Green","Hazel","Brown")))
haireye<-t(haireye)
mosaicplot(haireye,main=" ",las=1,cex=0.75)
```

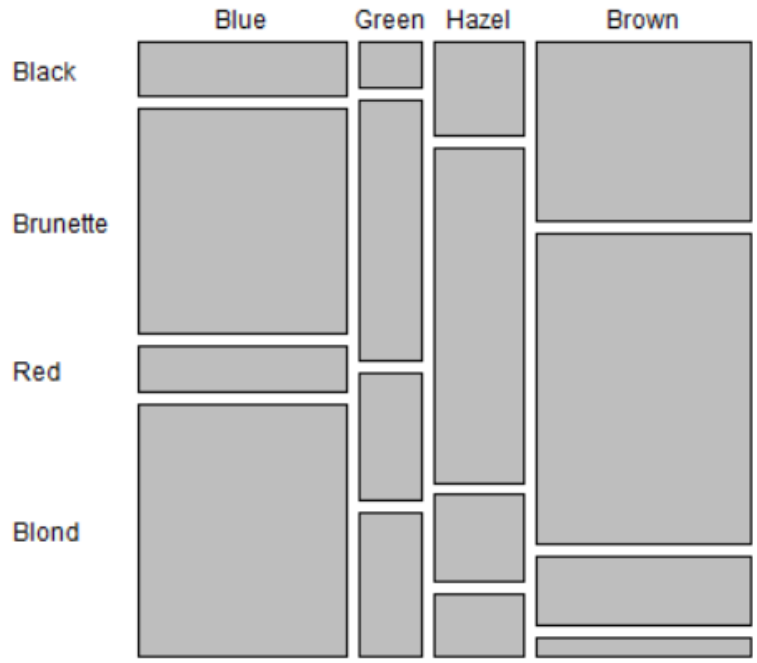

Figure 3 Mosaic chart of eye and hair color

Mosaic Charts using mosaicplot function

- Gray shading can be added to make the mosaic plot more eye-catching
- Obtained by setting color=TRUE

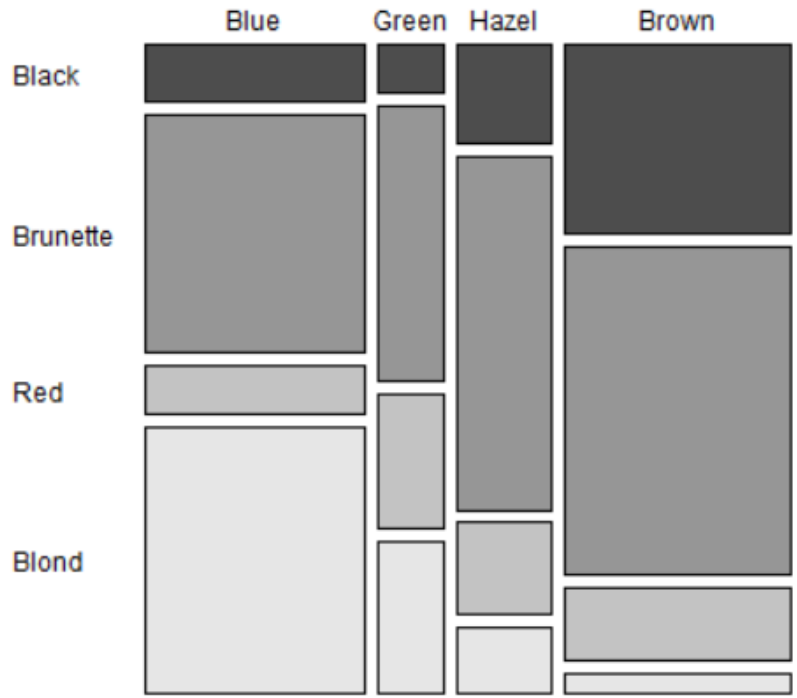

Figure 3 Mosaic chart of eye and hair color with grayscales for hair color

Mosaic Charts using mosaicplot function

- Shading or adding colors in a mosaic plot can also help make conditional comparisons
- We can also specify color=c("black","brown","red","gold")

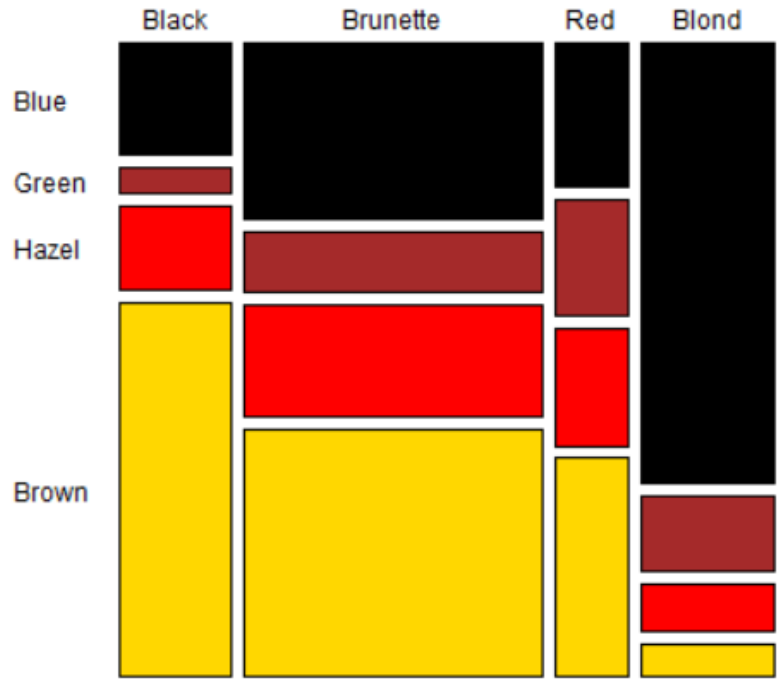

Figure 3 Mosaic chart of eye and hair color with colors for hair color

Interpretation of Bivariate graphs

The objective is to identify variations in the distribution proportions and make conditional comparisons.

So from our figures, we can see that

- The proportion of blonds decreases with darkness of eye color
- An increase in the proportion of black-haired individuals is associated with darkness of eye color

Critique of Mosaic Charts

- Requires comparison of **areas**, which is less intuitive than comparing lengths
- Effective for eye-catching presentations but may not always provide clarity
- Cleveland and McGill suggest that **grouped dot charts** are still the gold standard.

Conclusion: When to Use These Charts?

Recommendations:

- Avoid side-by-side pie charts—tables or grouped dot charts are more effective
- Use mosaic charts only when focusing on relationships between two variables
- **Grouped dot charts** remain the preferred visualization for clarity and accuracy

Remember: Choose charts that provide the most **clarity and ease of interpretation**!