Today: What Is Data?
Graphics Principles

Sam Ventura
36-315

Department of Statistics
Carnegie Mellon University

January 23, 2017

structured (vs. unstructured)

What Is Data?

Data: Information organized in some fixed / standard / easy-to-understand way (humans or computers)

Ex) temperature measurements, housing prices, heights / weights / BMI / heart rate / cholesterol / etc

Image classification — MNIST → images, video

Tweets, stock prices

rows
↑
etc

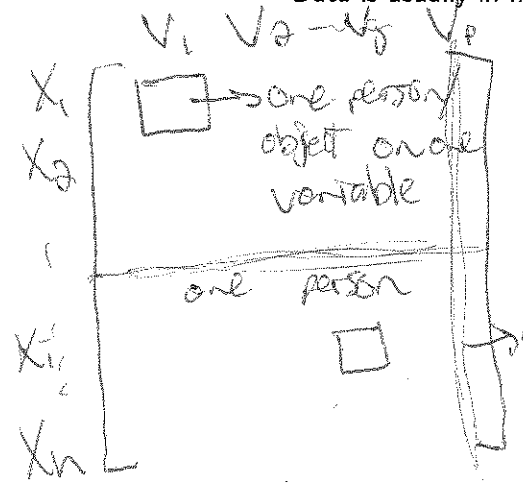
How Do We Describe Data?

Two measurements used to describe datasets:

n: # of observations (people, objects, survey responses, days in the month)

p or d: # of variables / covariates / features / questions → columns

Data is usually in matrix form:



* rows = observations

↳ single row contains all information on all variables for one person/obj

* columns = variables

↳ one column has all answers to a single ? for all rows

Types of Data (usually)

Categorical: Qualitative, describes categories / classes / qualities of observation

CS: boolean, string / text, integer, factor

sub groups {
ordinal / ordered: e.g. "strongly agree" "agree" "neutral" ...
nominal / unordered: e.g. race, gender, nationality

Continuous: Quantitative, (usually) real-valued, numerical data

CS: floats, integer, doubles, long.

Graphics and Their Goals (from Tufte)

Graphics: visually display measured quantities by combining points, lines, coordinate system, numbers, symbols, words, shading, color

Goals: show data!

- induce viewer to think about substance, not graphical methodology
- avoid **distorting** the data
- present numbers in small space
- make large, complicated datasets more coherent
- encourage comparison of different pieces of data
- reveal data at several levels of detail
- **describe, explore, tabulate, identify relationships**
- be closely integrated with statistical/verbal descriptions of dataset

Graphs that do not meet these goals are not successful

Graphs leading viewers to make misleading conclusions should be avoided

optimal: LF = 1

increased effect: LF > 1

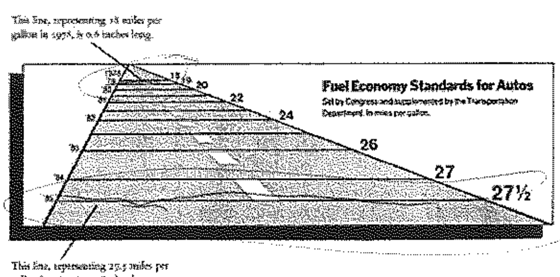
decreased effect: LF < 1

Distortion = "graph doesn't match the data"

Visual representation of data is inconsistent with numerical representation

$$\text{Lie factor} = \frac{\text{size of "effect" in graph}}{\text{size of "effect" in data}}$$

"effect" = change in amount of some feature/variable



Tufte suggests optimizing the Lie Factor:

in data

Actual % increase: $\frac{|27.5 - 18|}{18} \times 100\% \approx 52.8\%$

Graphical % increase: $\frac{|5.3\text{ in} - 0.6\text{ in}|}{0.6\text{ in}} \times 100\% \approx 783\%$

$LF = \frac{783\%}{52.8\%} \approx 14.83$

"Decorating" and Data-Ink

Graphics should not draw the viewer's attention away from the data.
Extras get in the way.

Note: Decoration does not refer to appropriate graph labeling.
Labels should always be clear, detailed, and thorough.
Label key parts of the data. Add text explanations if necessary.

Data Ink should primarily present information about the data:
the non-erasable, non-redundant core of a graphic

Tufte suggests using the *data-ink ratio*:

$$DI = \frac{\text{ink used to represent data}}{\text{total ink in graphic}} \times 100\%$$

% of ink devoted to non-redundant information in graph

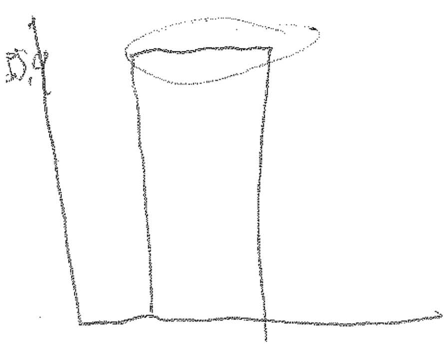
Ideally, maximize DI ratio (max of 1)

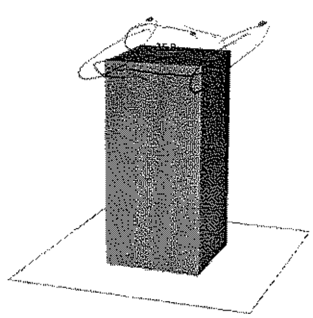
"Decorating" / Data-Ink

Two ways to increase the proportion of data-ink:

Remove non-data-ink:

Remove redundant data-ink:





1) height of front face
2) ------ side face (right)
3) ------ back line
4) ------ left line
5) 35.9
6) area of front face
7) height of location of 35.9

Today: Grammar of Graphics
1-D Categorical
Friday: ggplot2, 1-D Categorical

January 25, 2017

ggplot2: Based on "The Grammar of Graphics" (Wilkinson, 2005)

Each plot can be broken down into core components.
Wilkinson defines core components in book.

Hadley Wickham puts this into practice in R via ggplot2.

ggplot(cars93) + geom_point(aes(x = fuel.tank.capacity, y = MPG.City, color = Type))

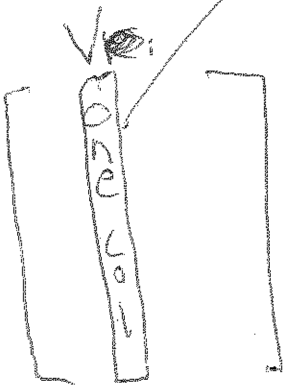
R Package ggplot2 – Hadley Wickham

Core components of a plot:

1. **data**: in ggplot2, data must be stored as an R data frame
2. **coordinate system**: describes 2-D space that data is projected onto e.g., Cartesian coordinates, polar coordinates, map projections, ...
3. **geometries**: describe type of geometric objects that represent data e.g., points, lines, polygons, ...
4. **aesthetics**: describe visual characteristics that represent data ~~e.g., for~~ example, position, size, color, shape, transparency, fill
5. **scales**: for each aesthetic, describe how it is is converted into values that are displayed on the actual graph e.g., log scales, color scales, size scales, date scales, ...
6. **stats** : describe statistical transformations that help summarize data e.g., counts, means, medians, regression lines, ...
7. **facets**: describe how data is split into subsets and displayed as multiple small graphs (particularly important for categorical data!)

map pieces of the data to pieces of the graph

# of obs in data



1-D Categorical Data

vector of length n

Recall: Data can be **categorical** or **continuous**

Categorical data can be **ordered** or **unordered / nominal**

1-D Categorical Data Structure:

for any obs. in this vector, that obs can take one of the following categories $\{C_1, C_2, \ldots, C_K\}$

How could we summarize this data? → "counts" / frequencies / proportions / percentages in each category

What information would you report?

- range of percentages
- smallest / least frequent category?
- largest / most frequent category?
- "outliers", e.g. category w/ only one obs

1-D Categorical Data

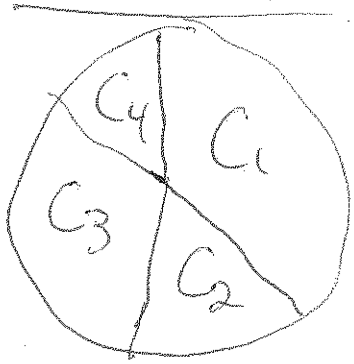
To show the differences among the categories, need to use *area plots*:

Differences in area correspond to differences in category frequency (count, prop. perc.)

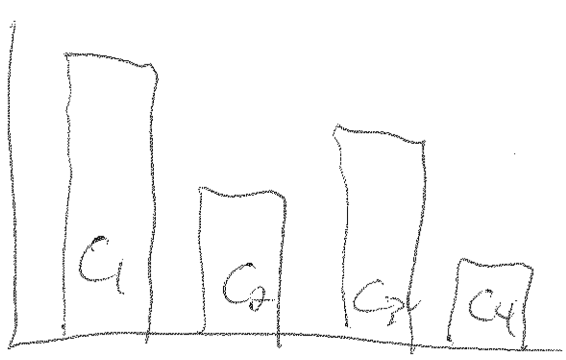
Each area of graph corresponds to a category

Examples of area plots?

pie chart



bar chart

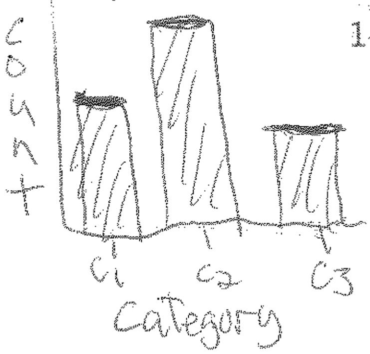


spine charts

rose diagrams

Decreasing? →

unordered → Never say increasing or dec



1-D Categorical Data – Bar Charts

Bar Charts: rectangular bar is created for each unique categorical value. The area and height of the bar is proportional to % of observations with the categorical value. Bars usually have equal width.

Width of each bar $\propto$ 1 (nothing)

Area = width × height

Area and height are $\propto$ the same thing

Area, height $\propto$ "count" in each category

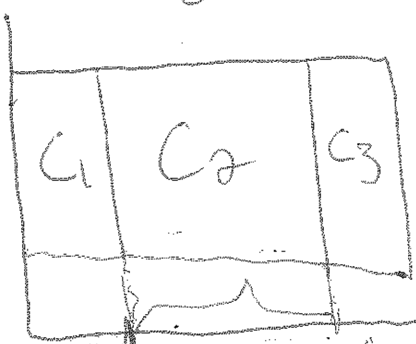
5/6

Y-axis can also be proportion, %

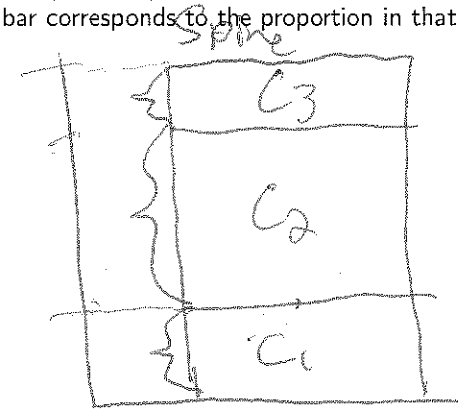
1-D Categorical Data – Spine Charts

Spine Charts: rectangular bar is created for each unique categorical value. The height (or width) of all bars is equal, and the width (or height) of the bar corresponds to the proportion in that category.

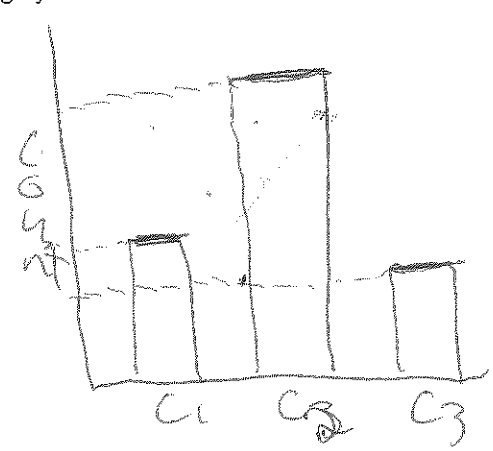
Spine



"count" / "total"



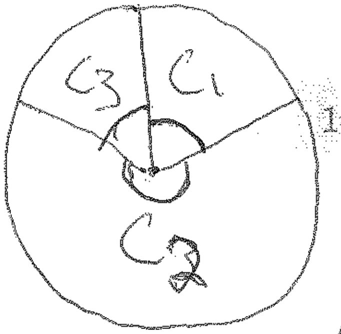
Bar



Spine: very hard to visually determine category counts

Bar: very easy

Issues:
- difficult to compare areas/angles
- Bad w/ lots of categories
- No total # of obs; ~~hard~~ hard to determine exact counts/proportions



1-D Categorical Data – Pie Charts

Pie Charts: circle divided up into sections ("pie slices") such that the area of each section is proportional to the number of observations with each unique categorical value.

Area of a circle: $A = \pi r^2$

Area of a slice: sector $A_{sector} = \pi r^2 \cdot \frac{\theta}{360}$

θ = angle in degrees (0 to 360)

$r \propto 1$ (nothing)

$A_{sector} \propto$ "counts" in each category
$\theta \propto$ "counts" in each category
} or proportions

1-D Categorical Data – Rose Diagrams

Rose Diagrams: circle sections are created for each category. All sections have the same width/arc/angle. The radius of each section is proportional to the category frequency. Sections are called "petals". Developed by Florence Nightingale (example posted to Blackboard).

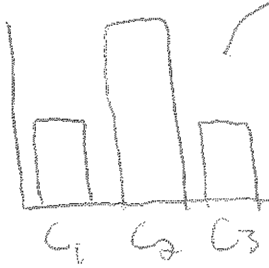
$\theta \propto 1$

$r \propto$ "counts" OR $r \propto \sqrt{count}$

$A_{sector} \propto (\text{"counts"})^2$ $\quad$ $A \propto count$

Issue = distortion

issue = requires extra data manipulation



$\rightarrow P(X=c_1)=1/4,\ P(X=c_2)=1/2,\ P(X=c_3)=1/4$

$\rightarrow$ or pie, rose, spine

What Does a Bar Chart Show? "Distribution" of a categorical variable

Marginal Distribution: $\rightarrow$ "truth" or "true distribution"

Categorical variable X has K categories — interested in $P(X=c_K)$ for each c_K

Empirical Distribution: $\rightarrow$ "what did we observe" "observed distribution"

"our best estimate of the true (marginal) distribution of the variable that we plotted given the information / data that we have"

$\hat{p}(X=c_1)=0.26$

$\hat{p}(X=c_2)=0.48$

$\hat{p}(X=c_3)=0.26$

$\boxed{\text{we want to visualize uncertainty}}$!

Bar Charts: Counts vs. Proportions

Counts/Frequencies of each category: $\rightarrow$ Advantage: we get an idea of sample size

$\rightarrow$ no sample size

Proportions: But, we do get interesting statistical info

estimates of $P(X=c_j)=\hat{p}_j=\dfrac{\#\text{ of obs. in category } j}{\text{total }\#\text{ of obs.}}$

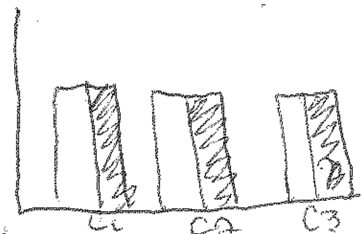
standard error on $\hat{p}_j$: $se(\hat{p}_j)=\sqrt{\dfrac{\hat{p}_j(1-\hat{p}_j)}{N}}$ $\rightarrow$ sample size

Confidence interval around $\hat{p}_j$: $\hat{p}_j \pm z_{\alpha/2}\cdot\sqrt{\dfrac{\hat{p}_j(1-\hat{p}_j)}{N}}$

$\alpha=0.95 \rightarrow 95\%$ CI

$\alpha=0.99 \rightarrow 99\%$ CI

Chi-Square Test for Independence



Chi-squared test: Statistical test used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories (of a categorical variable).

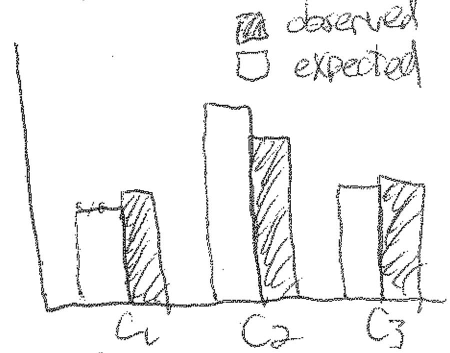
2-D Categorical: Used to test differences in the conditional distributions (more on this Wednesday)

1-D Categorical: assume we have K categories

$\boxed{\text{setup \#1}}$

$H_0: p_1 = p_2 = p_3 = \cdots = p_K$ → "all proportions are equal"

H_a: at least one of the p_j's is different



$\boxed{\text{setup \#2}}$

$H_0: p_1 = p_1^*, p_2 = p_2^*, \ldots, p_K = p_K^*$

H_a: at least one is not the same as expected

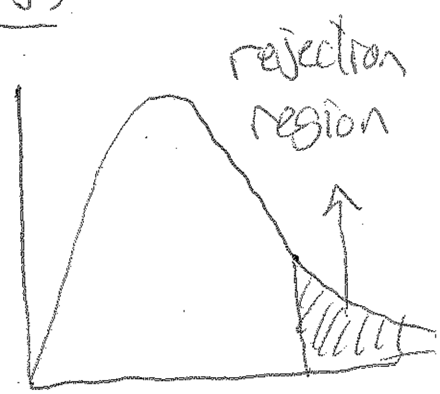
p_j^* = expected proportion in Category j

Computing and Interpretting the Chi-Square Test

Test Statistic: $\chi^2 = \sum_{j=1} \frac{(O_j - E_j)^2}{E_j}$

O_j = observed # in ~~each~~ category j

E_j = expected # — — — — — —



Interpretation:

Large χ^2 statistic means what we observed is very different from what we expected, so we will reject the null hypothesis (H_0) and conclude that there is evidence that the null is not true at some level α

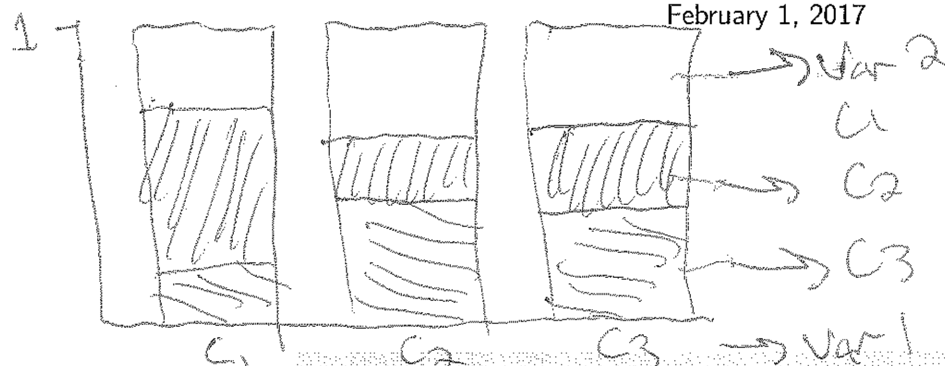
Today: Marginal and Conditional Distributions
2-D Caterogical Data
Contingency Tables and Mosaic Plots
Friday: 2-D Categorical

Sam Ventura
36-315

Department of Statistics
Carnegie Mellon University

February 1, 2017

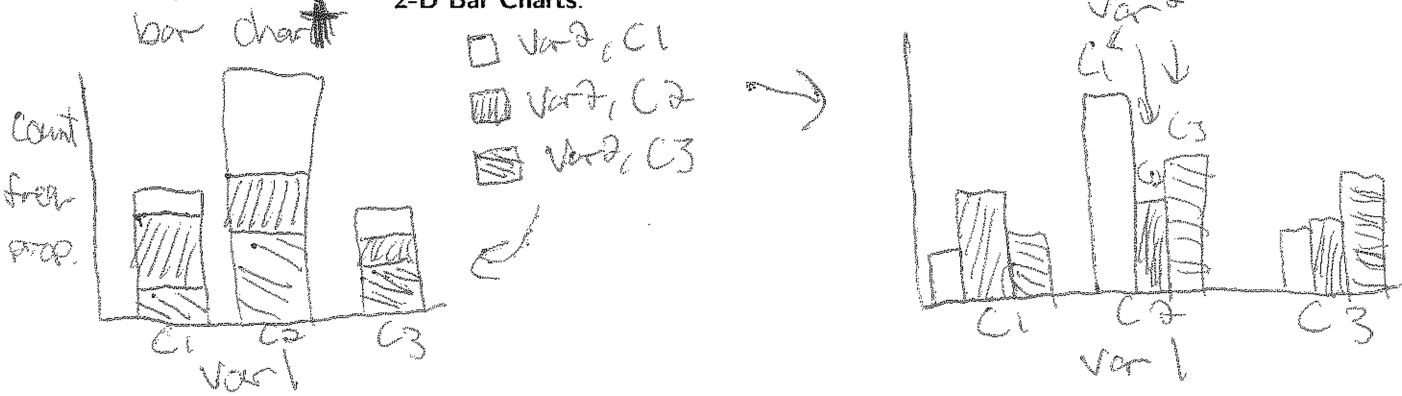
Proportional bar chart



2-D Categorical: Stacked and Side-by-Side Bar Charts

Stacked bar chart

2-D Bar Charts:



What does a 2-D bar chart show?:

Stacked	Side-by-side	proportional
- Marginal of V1: easy	Marg. V1: OK	Marg: of V1
- conditional distributions of V2 given V1:	cond. V2|V1	cond. of V2|V1

Contingency Tables → 2 variables

Contingency Tables Give Us:

Empirical {
- Marginal distribution (of each variable individually)
- conditional dist'ns (of $V_1 | V_2$ or $V_2 | V_1$)
- Joint dist'n (of both V_1 and V_2 simultaneously)

Two Categorical Variables:

Var 1: K_1 categories ; Var 2: K_2 categories

What Are We Interested In?

n = total # of obs
$n_{i\cdot}$ = total # in row i
$n_{\cdot j}$ = total # in column j

we know
- Observed counts in each cell
  ↳ O_{ij} = # obs in cat i of V_1 and cat j of V_2

3/7

we want to know: Expected counts in each cell → E_{ij}

Contingency Tables and Marginal/Conditional Distributions

Variable 2

Variable 1

	C_1^*	C_2^*	...	$C_{K_2}^*$	
C_1	O_{11}	O_{12}	...	O_{1K_2}	$n_{1\cdot}$
C_2	O_{21}	O_{22}	...	O_{2K_2}	$n_{2\cdot}$
⋮	⋮	⋮		⋮	⋮
C_{K_1}	O_{K_11}	O_{K_12}	...	$O_{K_1K_2}$	$n_{K_1\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	...	$n_{\cdot K_2}$	n

→ Marginal dist'n of Var 1

→ Conditional dist'n of V2 given $V_1 = C_2$

→ cond. dist'n of V1 given $V_2 = C_2^*$

↳ Marginal dist'n of Var 2

4/7

Today: 2-D Caterogical Data
Independence and Mosaic Plots

Sam Ventura
36-315

Department of Statistics
Carnegie Mellon University

February 6, 2017

for two events, A and B

Recall: Independence Rules from Probability

$P(A) = P(A|B)$ → marginal dist'n of Var 1 = conditional dist'n of Var 1 given Var 2

$P(B) = P(B|A)$ → " " Var 2 = " " Var 2 given Var 1

$P(A \cap B) = P(A)P(B)$ → joint dist'n of Var 1, Var 2 = (marginal of Var 1) × (marginal of Var 2)

Can input contingency tables into chi-square tests for independence

E.g. `chisq.test(table(var1, var2))`

→ contingency table in R

More on this in Lab 04

→ Test for independence of two categorical variables

i = category i of Var 1
j = category j of Var 2

E_{ij} assumes Var 1 $\perp\!\!\!\perp$ Var 2

independent

Pearson Residuals

Pearson Residuals: Scaled difference between observed/expected

$$r_{ij} = \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}}} = \frac{\text{Observed} - \text{Expected}}{\sqrt{\text{Expected}}}$$

$r_{ij} \overset{\approx}{=} 0$: we observed roughly as many obs. in categories i and j as we would expect if Var 1 $\perp\!\!\!\perp$ Var 2

$r_{ij} > 0$: "too many" observed
$r_{ij} < 0$: "too few" observed
} assuming Var 1 $\perp\!\!\!\perp$ Var 2

3/4

Result: r_{ij}'s are asymptotically Normally distributed.

$r_{ij} \sim N(\mu = 0, \sigma^2 = 1)$

$|r_{ij}| > 2 \Rightarrow$ significant at the $\alpha = 0.05$ level

$|r_{ij}| > 4 \Rightarrow$ significant at the $\alpha = 0.01$ level

Mosaic Plots

Mosaic Plots: Area plot for two categorical variables

Each cell in a contingency table gets a box

↳ Area of each box $\propto$ % of obs. in the corresponding cell of the table

↳ width of each box $\propto$ marginal distn of Var 1

↳ heights of each box $\propto$ conditional distn of Var 2 | Var 1

Can color the boxes by their differences from what was expected

Friday: Mosaic Plots in R