Today: What Is Data?
Graphics Principles
Friday: Introduction to R and Reproducibility
Monday: No class (Labor Day)

Sam Ventura
36-315

Department of Statistics
Carnegie Mellon University

August 31, 2016

What Is Data?

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

ex): Tweets, sports statistics, Temperature

Censuses / Surveys → collect info on population, demographics etc

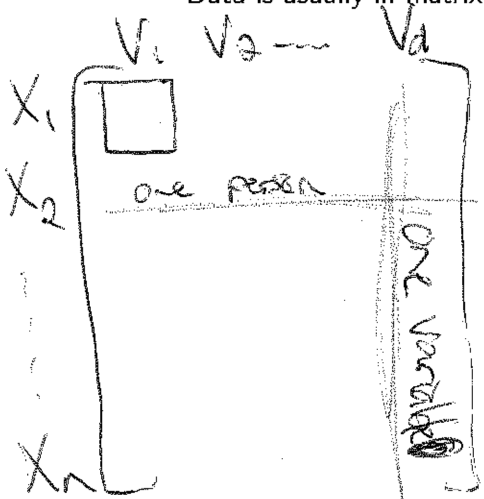
How Do We Describe Data?

Two measurements used to describe datasets:

n = # observations (people, subjects, objects, etc) → rows

p or d = # of variables / covariates / questions / etc → columns

Data is usually in matrix form:



* rows = observations

↳ single row has all answers / values / measurements / etc for single person / obs. / etc.

* columns = variables

↳ single column has all answers to a single ? for all people

Types of Data

Categorical: usually Qualitative, describes categories / classes / qualities of obs.

CS: strings, integers, factors, etc

Subcategories {
Ordinal: eg. strongly disagree, disagree, neutral, agree, SA
nominal: "unordered" eg. race, gender, names / text, etc

Continuous: Quantitative, real-valued, numerical data

CS: double, long, sometimes integer, float.

Notation: $X = \{X_1, X_2, \ldots, X_d\}$

$X_i \in \mathbb{R}$, $X \in \mathbb{R}^d$

"Decorating" / 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*:

$$DI = \frac{\text{data ink}}{\text{total ink on graphic}}$$

% of ink devoted to non-redundant / useful information.

Ideally → Maximize DI (max = 1)
won't quite get to 1, because of axes, grid lines etc

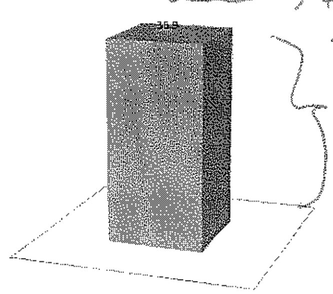
"Decorating" / Data-Ink

Two ways to increase the proportion of data-ink:

Remove non-data-ink:
↳ Ink that does not depict statistical info

In class wednesday 1/20, hands on map graphic

Remove redundant data-ink: → Ink that is unnecessarily redundant/repetitive.

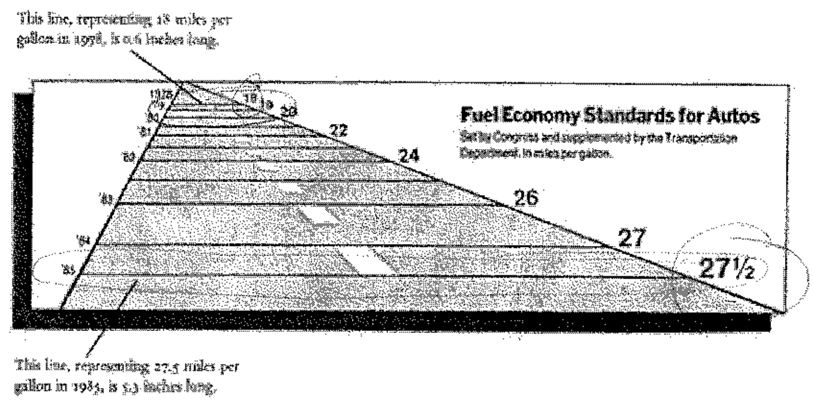


Indications of height:
1) height of front-left line on bar
2) height of front-right line on bar
3) ---- back-right ----
4) position of number
5) value of number
6) position of top line (front)
7) position of top line (back)
8) height of shading
9) etc....

Distortion

Visual representation of data is inconsistent with numerical representation

In other words: **The graph doesn't match the data**



Optimal: $LF \approx 1$

$LF > 1 \rightarrow$ enhance the effect

$LF < 1 \rightarrow$ decrease the effect

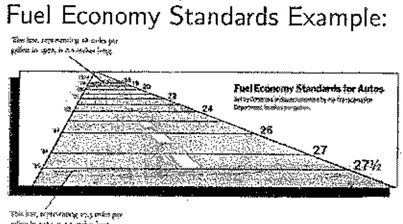
Lie Factor

Tufte suggests optimizing the Lie Factor:

$$LF = \frac{\text{size of "effect" in graphic}}{\text{- - - - - - in data}}$$

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

Fuel Economy Standards Example:



→ Actual % increase (in data)

$$\frac{|27.5 - 18|}{18} \approx 0.528$$

graphical increase (in graph)

$$\frac{|5.3\text{ in} - 0.6\text{ in}|}{0.6\text{ in}} = 7.83$$

$$LF = \frac{7.83}{0.528} = 14.83$$

father of graphics

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, or decorate
- 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

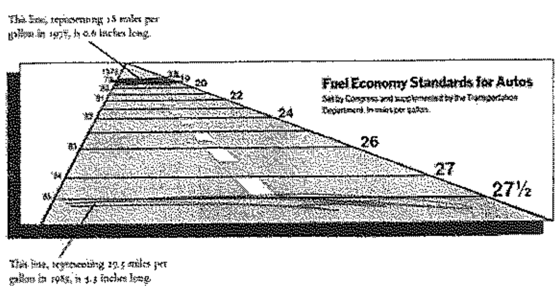
Distortion = "graph doesn't match the data"

Visual representation of data is inconsistent with numerical representation

optimal: $LF \approx 1$

~~enhanced~~ increased effect : $LF > 1$

decreased effect : $LF < 1$



Lie factor = $\dfrac{\text{size of "effect" in graphic}}{\text{size of effect in data}}$

effect = change in amount of some feature or variable

Tufte suggests optimizing the Lie Factor:

Actual % incr (in data) $= \dfrac{|27.5 - 18|}{18} \times 100 = 0.528 \times 100\%$

graphical % incr $= \dfrac{|5.3 \text{ in} - 0.6 \text{ in}|}{0.6 \text{ in}} = 7.83 \times 100\%$

$LF = \dfrac{7.83}{0.528} = 14.83$ → huge

"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

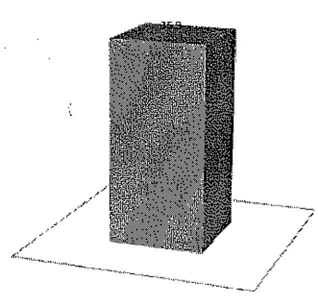
Tufte suggests using the *data-ink ratio*:

"Decorating" / Data-Ink

Two ways to increase the proportion of data-ink:

Remove non-data-ink:

Remove redundant data-ink:



Lab Exam On Wed, 10/19

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

September 7, 2016

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.

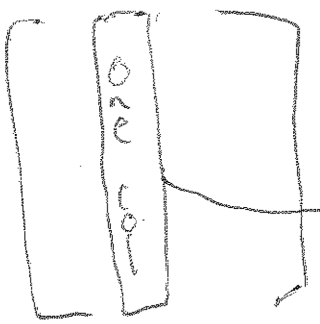
R Package ggplot2 – Hadley Wickham

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!)

Note: everyones needs to study definition of stats

Data



1-D Categorical Data

fixed # of categories K, $C_1, C_2, \ldots C_K$ in original dataset

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

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

1-D Categorical Data Structure: → vector of length n = # of rows = # of obs

for any obs in vector, $\in \{C_1, C_2, \ldots, C_K\}$

How could we summarize this data? → "counts": percentage in each category, proportion — — —, frequency — — —

What information would you report?

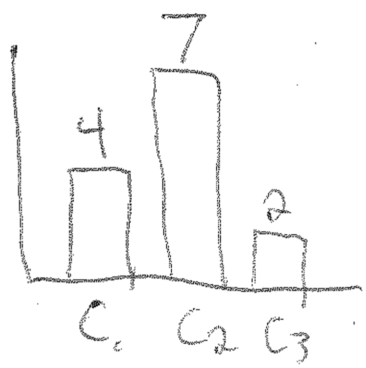
- counts / percentages / etc
- K = # of unique categories
- most / least frequent category
- "outliers" ≈ category w/ only one obs.
- ordered or unordered?

Monday: - "frequentist probabilities"
- deviation from what we expected to observe
- standard errors on category proportion

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 (each area corresponds to a category)



Bar chart

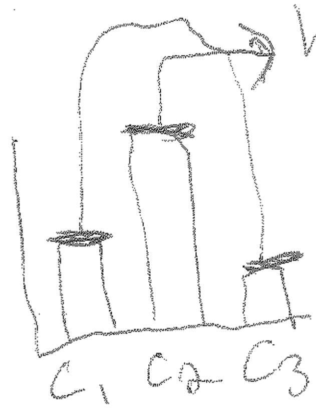
Examples of area plots?

pie charts

spine charts

rose diagrams

height of the bar $\propto$ count in each category



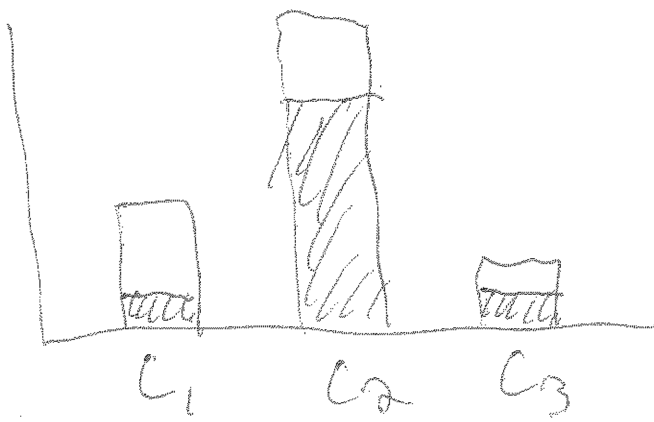
1-D Categorical Data – Bar Charts

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 the bar $\propto$ 1 (nothing)

Area = width × height

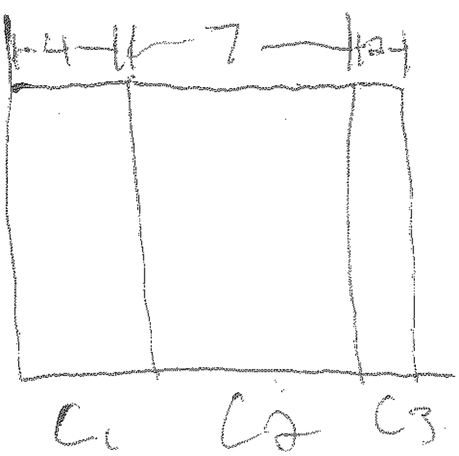
Area $\propto$ "count" each category



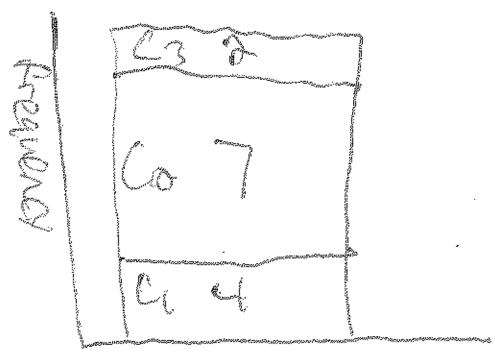
1-D Categorical Data – Spine Charts

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.

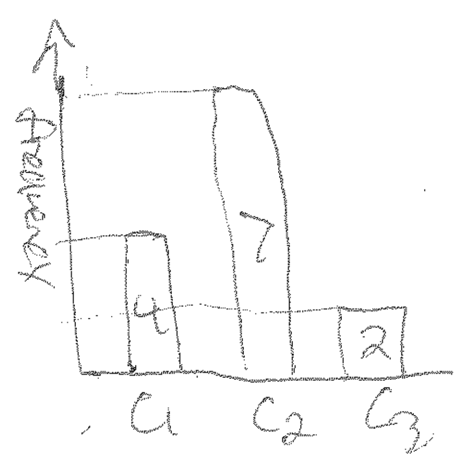
Spine



Spine



Bar



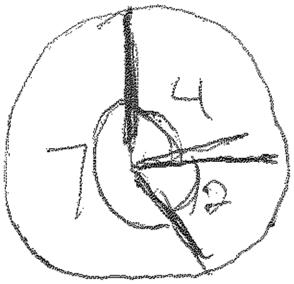
Spine: very hard to visually determine category counts

Bar: very easy

↳ compare to y-axis

measured in degrees

1-D Categorical Data – Pie Charts → Polar coordinates



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.

$\theta \propto$ "count" in each category

$r \propto 1$ (nothing)

$$A = r^2 \pi \qquad A_{ck} = \pi r^2 \cdot \frac{\theta}{360}$$

$A_{ck} \propto$ "count" in that category

good for displaying relative proportions

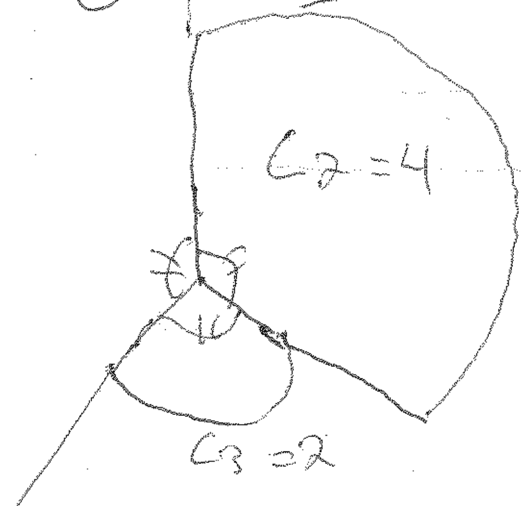
bad for most of everything else

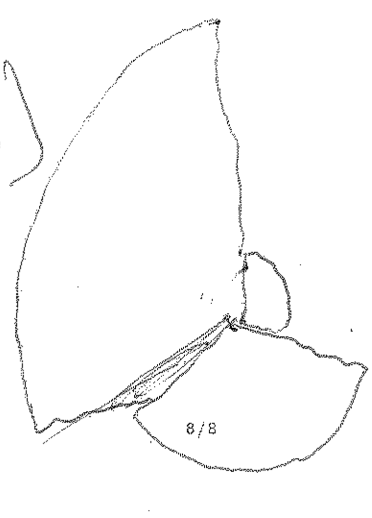
1-D Categorical Data – Rose Diagrams

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).

$r \propto$ "counts"

$\theta \propto 1$ (nothing)





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

September 7, 2016

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.

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!)

Data
variables

1-D Categorical Data

→ K categories $C_1, C_2, \ldots, C_K$

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

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

1-D Categorical Data Structure: → Vector of length n = # of rows / obs. in orig dataset

~~If $X \in$ ... then~~ $X \in \{C_1, C_2, \ldots, C_K\}$

each obs. in vector

How could we summarize this data? → percentages, proportions, frequencies of each category → "counts"
What information would you report?

- # of unique categories
- what are the unique categories?
- ordered or unordered?
- counts in each category

- most/least frequent categories
- deviation from what you expected to observe
- "frequentist probabilities" + standard errors

Model

1-D Categorical Data

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

In graph of 1-D categorical variable, we want to see differences in the area of ~~the~~ part of the graph corresponding to each category

Examples of area plots?

Bar ~~graph~~ chart

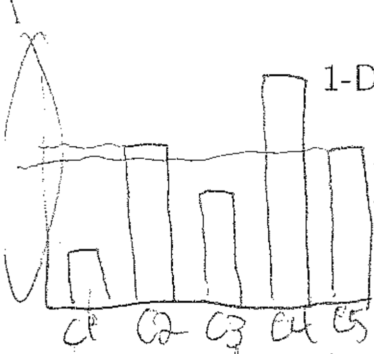
pie chart

Spine chart

rose diagram

→ freq. or prop.

area of rectangle = width × height



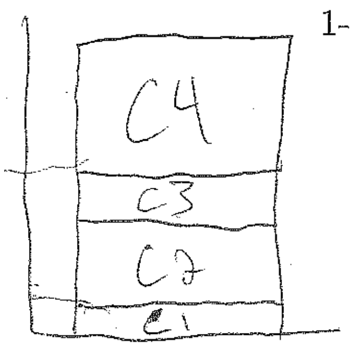
1-D Categorical Data – Bar Charts

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 bars ∝ nothing

heights ∝ frequency / proportion of observations that fall into that particular category

area ∝ same as height



1-D Categorical Data – Spine Charts

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

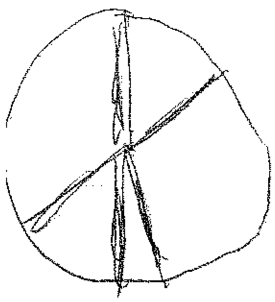
harder to compare stacked heights as opposed to bar chart

width ∝ equal

heights ∝ frequency / proportion

area ∝

But: STACKED → hard to compare.



1-D Categorical Data – Pie Charts Polar Coordinates

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. (frequency / proportion)

Θ = Theta ~~∝~~

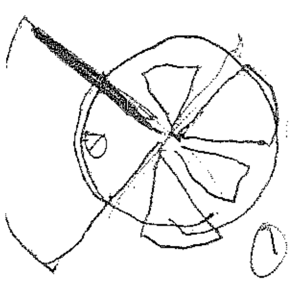
A = area ∝ frequency / proportion

r = radius ∝ nothing

↳ all are the same

Polar Coordinates

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 is proportional to the square root of the category frequency. Sections are called "petals". Developed by Florence Nightingale (example will be posted to Blackboard).

1 r = radius ∝ ~~...~~ freq / prop

2+3 ⇒ 3 $A \propto (\text{frequency / proportion})^2$

4

3 $A = \pi r^2 \Rightarrow A \propto r^2$

$\Rightarrow r \propto \sqrt{\text{freq/proportion}}$

0 θ = "theta" ∝ nothing

→ or, pie, spine, rose

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

we are interested in: **Marginal Distribution**: → "true distribution"

specifies the probability of observing each particular category

- X has K categories: $P(X=C_1)=1/4$, $P(X=C_2)=1/2$, $P(X=C_3)=1/4$

Empirical Distribution: → "observed distribution"

"our best estimate of the true marginal distribution of the variable that we plotted, given the data that we observed."

$p(X=C_1)=.26$
$p(X=C_2)=.52$
$p(X=C_3)=.22$

We want to visually display uncertainty

Bar Charts: Counts vs. Proportions

Counts/Frequencies of each category: get an idea of sample size

↱ No sample size,

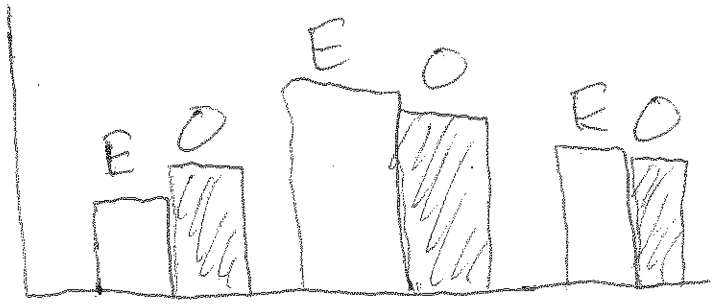
Proportions: But, we do get some useful statistical information

$$\text{estimate of } P(X=C_i) = \hat{p}_i = \frac{\#\text{ of observations in category } C_i}{\text{total }\#\text{ of obs }(N)}$$

standard error of $\hat{p}_i$: $se(\hat{p}_i) = S_{\hat{p}_i} = \sqrt{\frac{\hat{p}_i(1-\hat{p}_i)}{N}}$

CI around $\hat{p}_i$: $\hat{p}_i \pm Z_{0.99} \cdot \sqrt{\frac{\hat{p}_i(1-\hat{p}_i)}{N}}$ → 99% CI

↳ confidence interval



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).

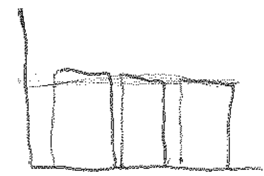
2-D Categorical: Used to test differences in the conditional distributions (more on this Wednesday)

1-D Categorical: Assume we have K categories

setup #1

$H_0: p_1 = p_2 = p_3 = \cdots = p_K$

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



setup #0

$H_0: p_1 = p_1^*$ and $p_2 = p_2^*$ and ... and $p_K = p_K^*$

H_a: at least one is not the same as expected.

p_K^* = expected proportion in category K

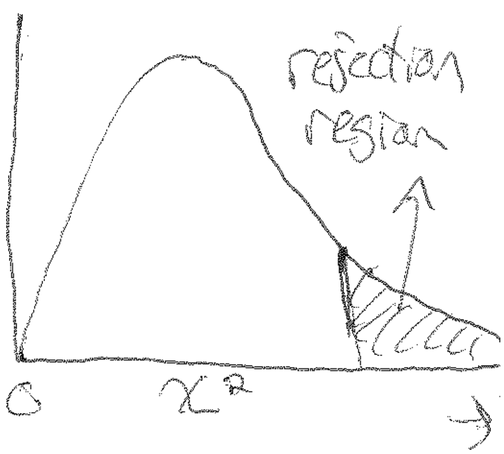
Computing and Interpretting the Chi-Square Test

Test Statistic: $\chi^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i}$

O_i = # observed in category i

E_i = # expected in category i

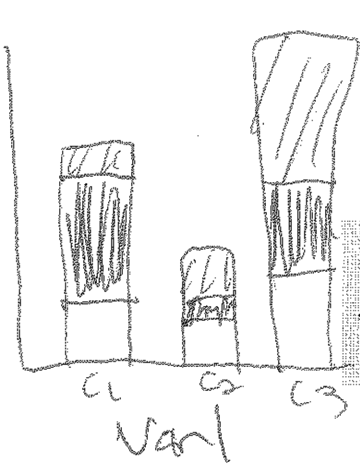
In R, chisq.test



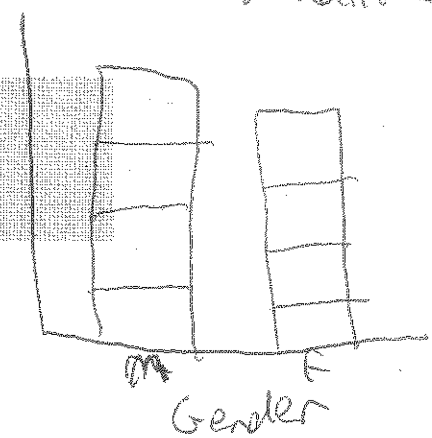
Interpretation:

Large χ^2 statistic → small p-value means we will reject the null hypothesis and conclude that there is evidence that the null is not true at our level α → significance threshold

aes(x = gender, fill = class_year)
↳ marginal ↳ conditional



Today: 2-D Caterogical Data
Independence and Mosaic Plots
1-D Continuous Data



Sam Ventura
36-315

Department of Statistics
Carnegie Mellon University

September 19, 2016

= cond. distn of Var 2 given Var1 = cat 1

2 categorical variables: Var1 + Var2
↓ ↓
K_1 K_2 categories

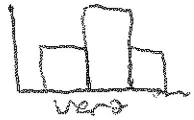
O_{ij} = # of obs. with category i in Var1 & cat j in Var2

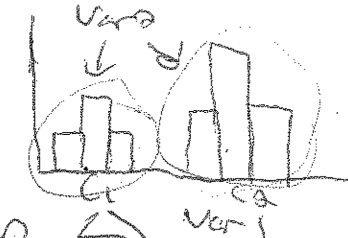
Contingency Tables and Marginal/Conditional Distributions

		Var2				marginal distn of Var1
		Cat 1	cat 2	— —	cat K_2	
Var1	Cat 1	O_{11}	O_{12}	— —	O_{1K_2}	$n_{1\cdot}$
	cat 2	O_{21}	O_{22}	— —	O_{2K_2}	$n_{2\cdot}$
	⋮	⋮	⋮	⋱	⋮	⋮
	cat K_1	O_{K_11}	O_{K_12}	— —	$O_{K_1K_2}$	$n_{K_1\cdot}$
		$n_{\cdot 1}$	$n_{\cdot 2}$	— …	$n_{\cdot K_2}$	

Marginal distn of Var2

Suppose we have two ~~cats~~ events A, B





$A \perp\!\!\!\perp B \iff$

Recall: Independence Rules from Probability

$P(A) = P(A|B) \rightarrow$ marginal distn of Var1 $=$ conditional distn of Var1 | Var2

$P(B) = P(B|A) \rightarrow$ — — — Var2 $=$ — — Var2 | Var1

$P(A \cap B) = P(A) \cdot P(B) \rightarrow$ Joint distn of Var1 and Var2 $=$ (marginal of Var1) $\times$ (marginal of Var2)

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

Pearson Residuals → assumes Var1 $\perp\!\!\!\perp$ Var2 → independent

Pearson Residuals: Scaled difference between observed/expected

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

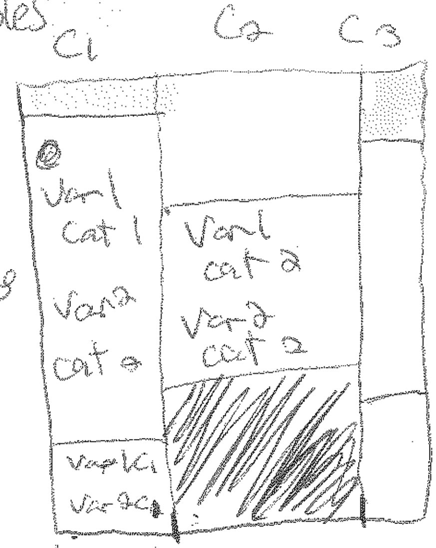
$r_{ij} > 0$: "too many" observed
$r_{ij} < 0$: "too few" observed } assuming Var1 and Var2 are independent

r_{ij}'s are Asymptotically Normally distributed!

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

$|r_{ij}| > 4 \Rightarrow$ — — — — — — $\alpha = 0.0001$ level color Mosaic plot according to the Pearson residuals in each cell

Mosaic Plots → visualizes contingency tables



Mosaic Plots: Area plot for two categorical variables

each cell in the contingency table gets a box in the plot

area of each box $\propto$ % of obs. in the corresponding cell of the cont. table

width of box$_i$ $\propto$ % of obs. in Var1 cat i
↳ marginal distn of Var 1

height of box j $\propto$ % of obs. in Var2 cat j given Var1 is in cat i → conditional distribution

Can color the boxes by their differences from what was expected

Friday: Mosaic Plots in R

1-D Continuous Data

Ex: time, age, temperature, distance, height, weight, rate, probabilities

Structure: $X = \{X_1, X_2, \dots, X_n\}$, $X_i \in \mathbb{R}$ ↱ $-\infty$ to ∞
$n \times 1$ vector (column of our data), each obs is a real #

Summary:
mean / average, median (or other measures of center)

spread: standard deviation, variance, IQR

range: min, max, quantile, percentiles

In R:

range(), min(), max()

mean(), sd(), median()

summary(), var(), quantile()

modality → unimodal, bimodal, multimodal

1-D Continuous Distributions

symmetry

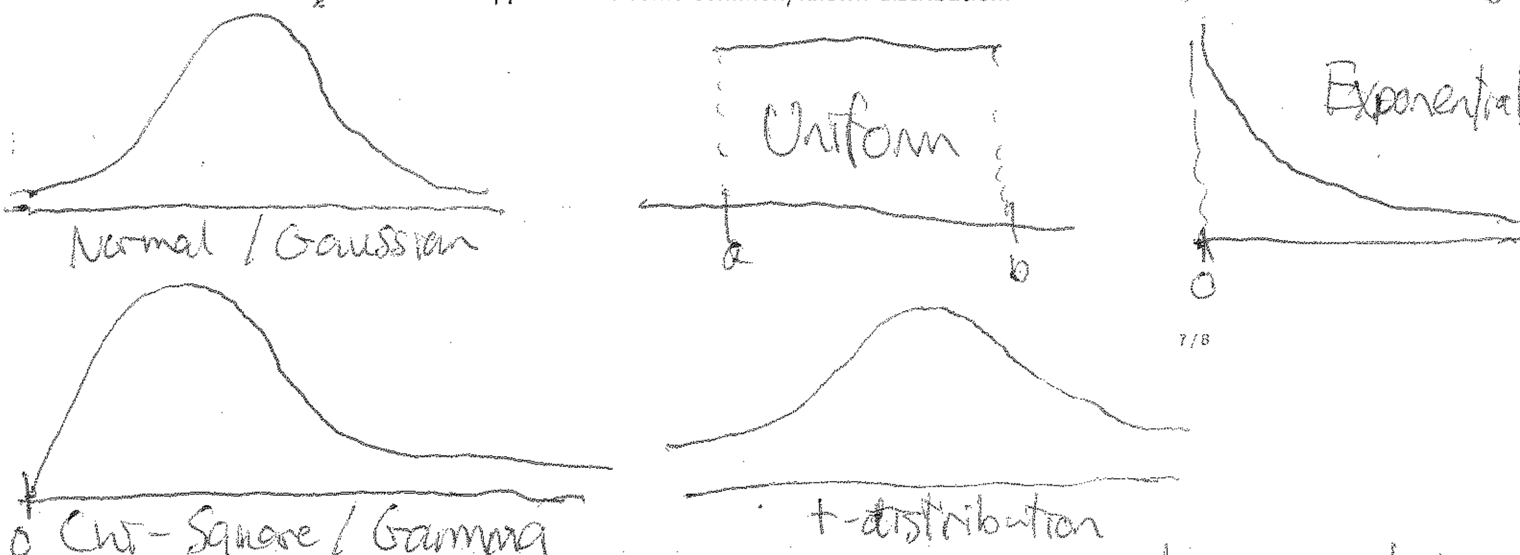
How do we describe continuous distributions?

outliers

Shape (skew)
Center (mean, median)
Spread (sd, var)

ranges of values that are most common → "high density" or least common ↳ "low density"

Do the data appear to fit some common/known distribution:



Visualizing 1-D Continuous Distributions

→ divide range into bins

How do we visualize continuous distributions?

Histogram vs Bar Chart
For continuous data
→ count obs in each bin

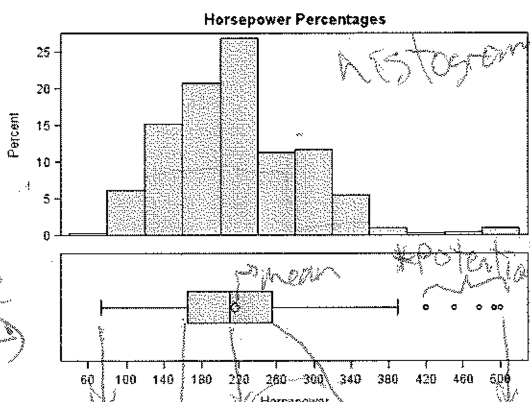
Today: 1-D Continuous

Sam Ventura
36-315
Today: Boxplots, Histograms, and Density Estimates
Conditional Distributions for Continuous Variables
Tartan Data Science Cup – Episode II → Sunday, 10/9

Department of Statistics
Carnegie Mellon University

September 26, 2016

Box Plots vs. Histograms: Advantages and Disadvantages



Boxplots

advantages:

- can easily see skew →
- get a sense of the center and the spread of the distribution
- get some specific values

Disadvantages:

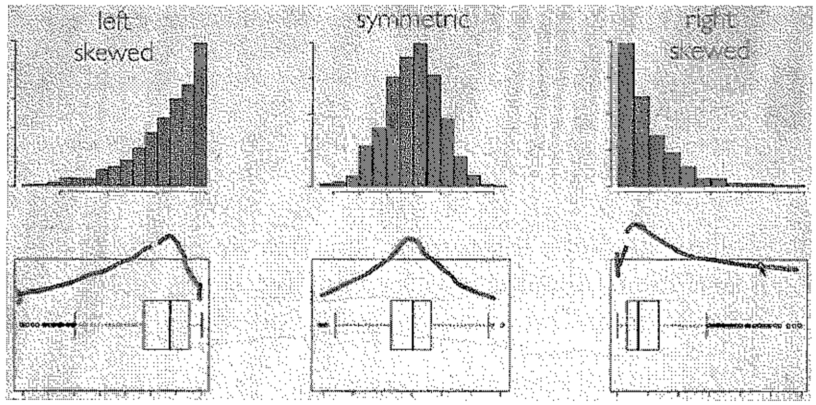
- no shape
- modality is ~~difficult~~ impossible to discern
- no idea about sample size
- hard to distinguish between very different distributions

Histograms

Adv: Area actually represents data
- can better see skew
- sample size can be estimated
- frequencies w/in each bin
- can usually see modality

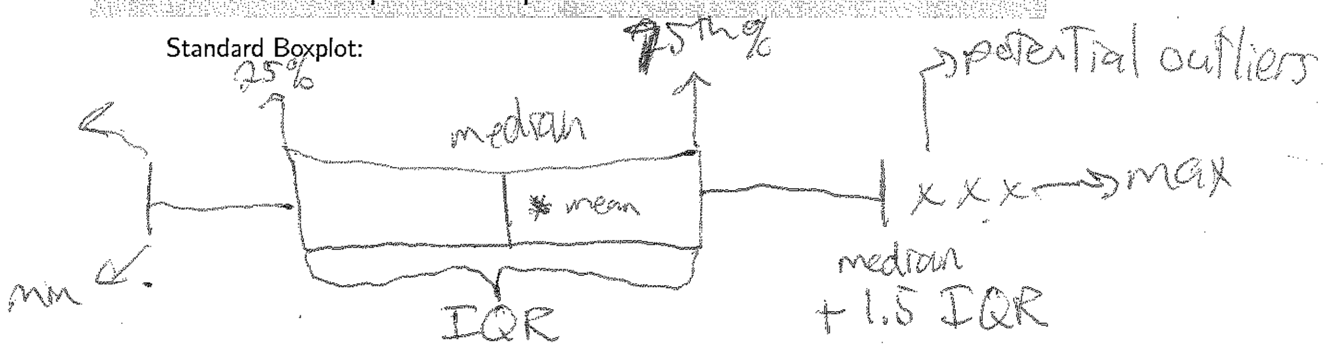
Disadvantages: we have to choose # of bins or the bin width

Determining Skew Visually



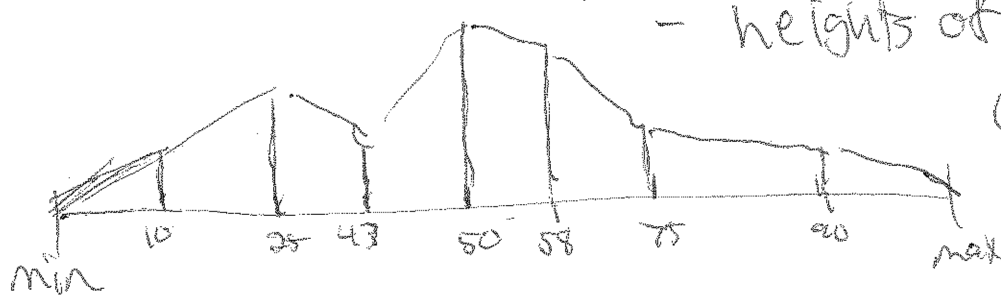
How Can We Improve Boxplots?

Standard Boxplot:



Improved Boxplot?

same thing, but w/ more %-iles
- heights of lines ∝ "density" of observations in that region



"smoothed out function of the data's (variables) distribution"

Continuous Densities

Theoretical: X is "random variable", $f_X(x) \rightarrow$ "probability density function"

If $X \sim N(\mu, \sigma^2)$

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{\sigma^2}}$$

If $X \sim Exp(\lambda)$

$$f_X(x) = \frac{1}{\lambda} e^{-x/\lambda}$$

What if we don't know the underlying distribuion? → "true distribution" → theoretical

How can we estimate the empirical distribution? → based on the data

parametric statistics: making assumptions about the underlying distribution of the data, and then finding optimal parameters

$\wedge$ = estimate

$\hat{f}$ = estimated density function

1-D Kernel Density Estimation → Non-parametric → NO assumptions

→ 100% based on the data

$$\boxed{\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)}$$

- n = # of observations
- h = "bandwidth" → "bin size"
  this is a parameter that you choose
  it dictates the "smoothness" vs. "jaggedness" "rigidity" of the resulting density estimate

small h: bumpy, rigid density estimate

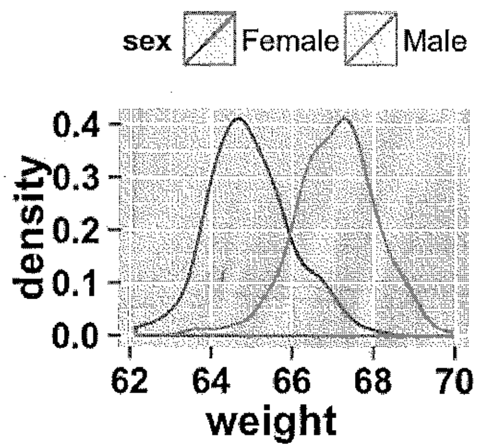
Large h: smooth density

- $K()$ = "Kernel function" → we get to choose this as well (parameter)
  → different functions will give us different features in the density estimate

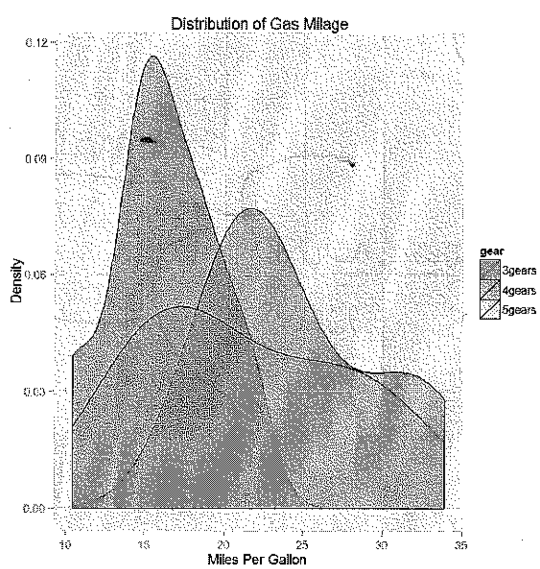
- x: the point at which we are estimating the density
  should be in the range of the observed data

→ contributes to the shape of the density

Density Estimates and Conditional Distributions



Density Estimates and Conditional Distributions



Today: 1-D Continuous

Sam Ventura

36-315

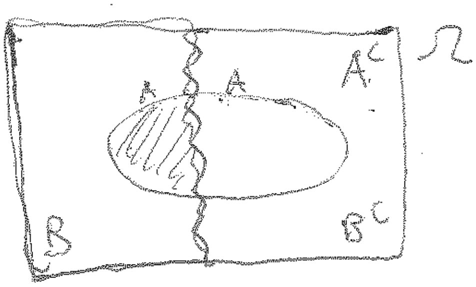
Today: Density Estimates, Kernels, Violin Plots
Rugs, Conditional Distributions, KS Tests

Department of Statistics
Carnegie Mellon University

September 28, 2016

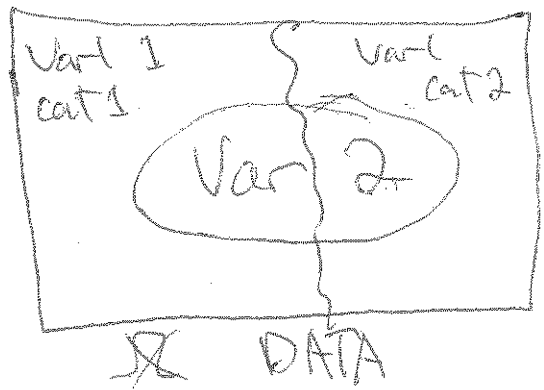
Conditional Distributions

In probability:



$$P(A) = \frac{\text{Area of } A}{\text{Area of } \Omega}$$

$$P(A|B) = \frac{\text{Area of } A \cap B}{\text{Area of } B}$$



Var 1: categorical

Var 2: continuous

marginal dist'n of Var 2

cond. dist'n of Var 2 | Var 1 = cat 2

↳ find dist'n of Var 2 within the subset of the data corresp. to Var 1 = cat 2

Histograms

Adv: easily understandable / similar to what we've learned

1-D Kernel Density Estimation

disadv: sensitive to changes in bin locations, bin widths, # of bins

- Not really ~~continuous~~ smooth

adv: - count scale OR proportion/%

- gives us estimate the sample size

disadv * - No sample size *

Density Estimates

disadv: sensitive to changes in bandwidth

adv: - put some mass in between observed points

- we can set specific estimates of the density at any point
- we're working w/ continuous data, and DEs gives a continuous curve

Kernels:

Gaussian / Normal
• smooth, default

triangular
• most of mass is centered ~~on the point~~ on each observation

uniform / boxcar / rectangular
• DEs will look like step functions

Epanechnikov
• smooth-ish

Good when you have fixed endpoints
↳ proportions / probabilities / %s
↳ time, distance, mass
↳ Beta, Exponential, Gamma