Lecture 3: Histograms

February 11, 2019

Overview

Course Administration

Kate

Good, Bad and Ugly

What is a Histogram?

ggplot and Histograms

Course Administration

1. Collect policy brief proposals
2. Make sure you're checking Piazza
3. Anything else?

Next Week's Good Bad and Ugly

Monday by 9 am. Earlier is ok.

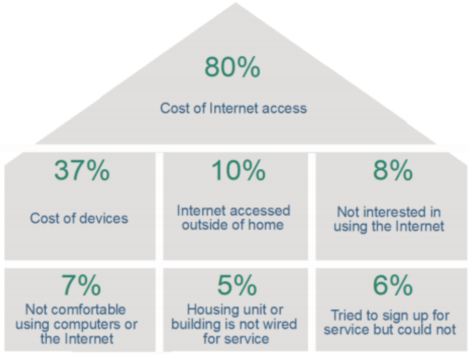
- PH
- JB

This Week's Good Bad and Ugly

- KJ
- KL
- AM

Kim's Example

Figure 3. Reasons For Lack of Internet Access in the Home, Among Households Not Connected



80%

Cost of Internet access

37%

Cost of devices

10%

Internet accessed outside of home

8%

Not interested in using the Internet

7%

Not comfortable using computers or the Internet

5%

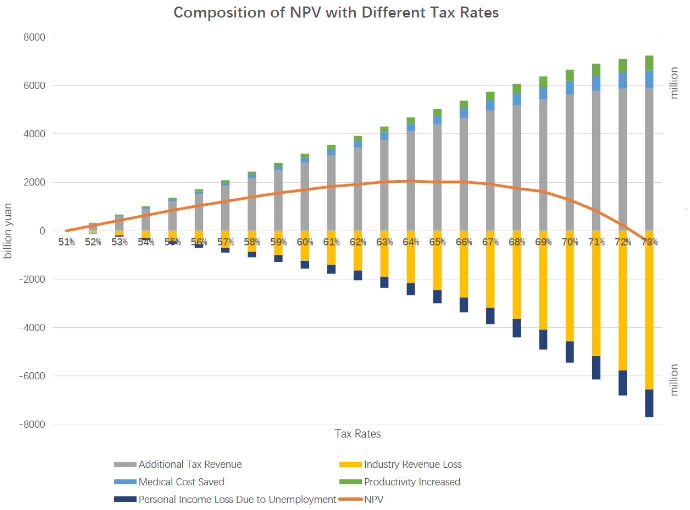
Housing unit or building is not wired for service

6%

Tried to sign up for service but could not

Source: ConnectHome Baseline Internet Access Survey; conducted November 2015–June 2016
Note: Respondents could check more than one category.

Kun Li’s Example



Alvan’s Example

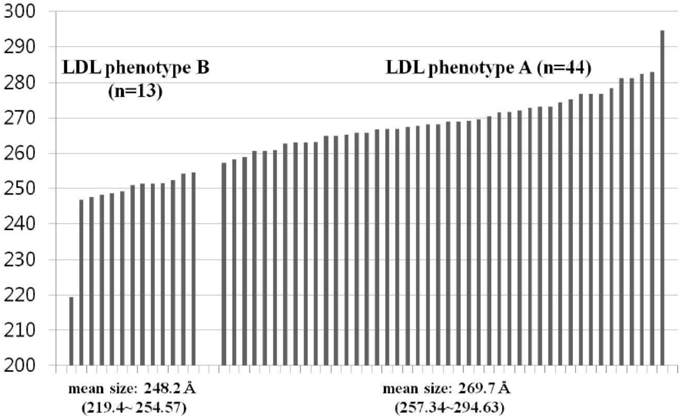


Fig. 1. Distribution of low-density lipoprotein (LDL) particle size in all study subjects (LDL phenotypes A and B). *LDL phenotype A group* (mean size: 269.7 Å, n = 44), subjects with buoyant-mode profiles [peak LDL particle diameter ≥ 264 Å] including intermediate LDL subclass pattern [256 Å ≤ peak LDL particle diameter ≤ 263 Å]; *LDL phenotype B group* (mean size: 248.2 Å, n = 13), subjects with dense-mode profiles [peak LDL particle diameter ≤ 255 Å]

Histograms

Histogram Shows the Distribution of **One** Variable

- Take a variable
- Make bins by value
- Count the number of observations in each bin
- Plot bars with that number

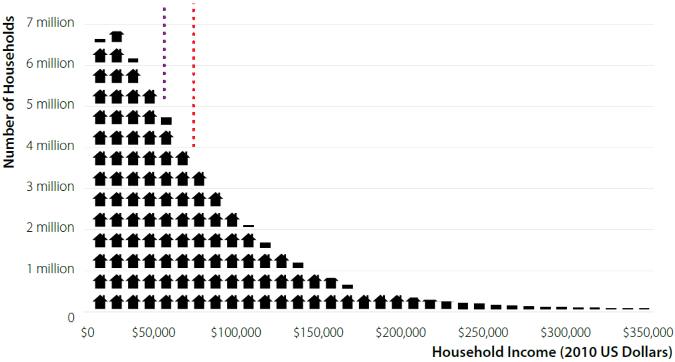
Histogram Shows the Distribution of **One** Variable

- Take a variable
- Make bins by value
- Count the number of observations in each bin
- Plot bars with that number
- Unlike bars charts, histogram bars touch, to indicate continuity
  - Which of Few's principles does this illustrate?

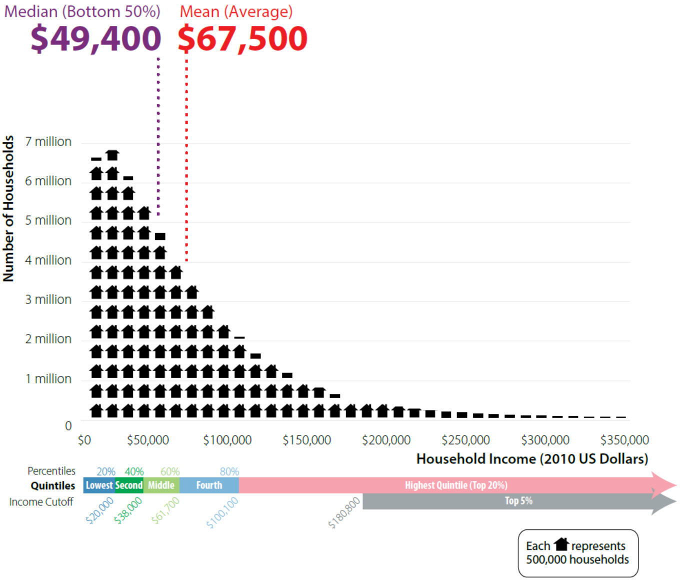
Two Examples

- Income distribution
- As a guide on a map

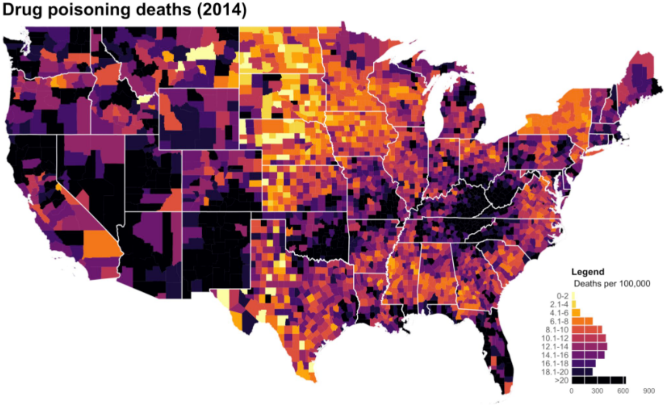
Mulbrandon’s Income Histogram



Mulbrandon’s Income Histogram



As a Map Legend



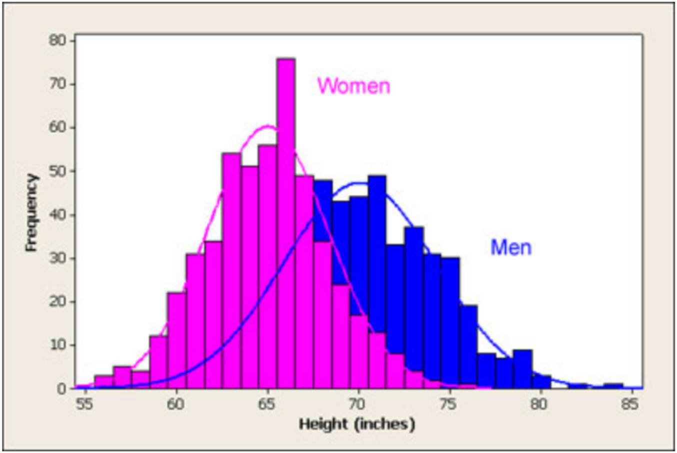
Source: https://blogs.cdc.gov/nchs-data-visualization/drug-poisoning-mortality/

From https://mathewkiang.com/2017/01/16/using-histogram-legend-choropleths/

Density Curves: Smoothed Histograms

- Imagine many very thin bars
- This yields a curve
- Sometimes it is more helpful to draw the curve

Height: Note the Curves



From http://www.usablestats.com/lessons/normal

R

Today

Z. Issues from last time: `ifelse()` and factors

A. What is `ggplot`?

B. The parts of `ggplot`

C. Histograms via `ggplot`

D. Density curves via `ggplot`

E. Additions to make your graph more clear

Z.1 ifelse()

```
dataf$var1 <- ifelse(CONDITION,
                     VALUE IF TRUE,
                     VALUE IF FALSE)
```

Z.1 ifelse()

```
dataf$var1 <- ifelse(CONDITION,
                     VALUE IF TRUE,
                     VALUE IF FALSE)
```

- you can separately evaluate all these parts
- check the condition

```
dataf$check <- (CONDITION)
```

Z.1 ifelse()

```
dataf$var1 <- ifelse(CONDITION,
                     VALUE IF TRUE,
                     VALUE IF FALSE)
```

- you can separately evaluate all these parts
- check the condition

```
dataf$check <- (CONDITION)
```

* check the evaluation dataf$check2 <- [VALUE IF FALSE]

Z.1 `ifelse()`

```
dataf$var1 <- ifelse(CONDITION,
                     VALUE IF TRUE,
                     VALUE IF FALSE)
```

- you can separately evaluate all these parts
- check the condition

```
dataf$check <- (CONDITION)
```

* check the evaluation dataf$check2 <- [VALUE IF FALSE]

* factors do strange things!

- `is.factor()`
- `as.factor()`

A. Graphing in R

- `ggplot` is the premier package for graphing in R
- There is a simple version of `ggplot` called `qplot`: we ignore it
- Developed in 2005 by Hadley Wickham
- In 2017, Wickham says “Ten years after ggplot2’s release, Wickham wonders how much longer his program will dominate chart making in R. ‘It really feels to me now like ggplot2 is ripe for disruption,’ said Wickham. ‘I’m surprised some young gun hasn’t come along, and thought, ’Wow this is crap,’ and done better. But so far, it hasn’t really happened.”’

Article link is here.

B. The Key Parts of a ggplot command

```
ggplot() +
  geom_something(data = , aes(x = xvar, [y = yvar])) +
  labs(title = "title here", x = "x label") +
  [things about scales] +
  theme([things you modify here])
```

C. Histograms in ggplot

```
ggplot() +
  geom_hist(data = DATA, aes(X = VARIABLE))
```

- R auto-chooses the number of bins
- And you can adjust

D. Density Curves in ggplot

```r
ggplot() +
  geom_density(data = DATA, aes(X = VARIABLE))
```

- R auto-chooses the number of bins
- And you can adjust

Adding in curves by groups

- suppose your data has height by gender
- each row indicates gender
- best if this variable is a factor

```r
ggplot() +
  geom_density(data = DATA,
               aes(X = VARIABLE, color = gender))
```

E. Additions for Clarity

Labels

```
ggplot() +
  ... +
  labs(title = "title here", x = "x title", y = "ytitle")
```

E. Additions for Clarity

Labels

```
ggplot() +
  ... +
  labs(title = "title here", x = "x title", y = "ytitle")
```

Subsets: Here where `DATA$VAR1 == 1`

```
ggplot() +
  geom_hist(data = DATA[which(DATA$VAR1 == 1),],
            aes(X = VARIABLE))
```