Exploratory graphs

Roger D. Peng, Associate Professor of Biostatistics

May 18, 2016

Why do we use graphs in data analysis?

- To understand data properties
- To find patterns in data
- To suggest modeling strategies
- To “debug” analyses
- To communicate results

Exploratory graphs

- To understand data properties
- To find patterns in data
- To suggest modeling strategies
- To “debug” analyses
- To communicate results

Characteristics of exploratory graphs

- They are made quickly
- A large number are made
- The goal is for personal understanding
- Axes/legends are generally cleaned up
- Color/size are primarily used for information

Background - perceptual tasks

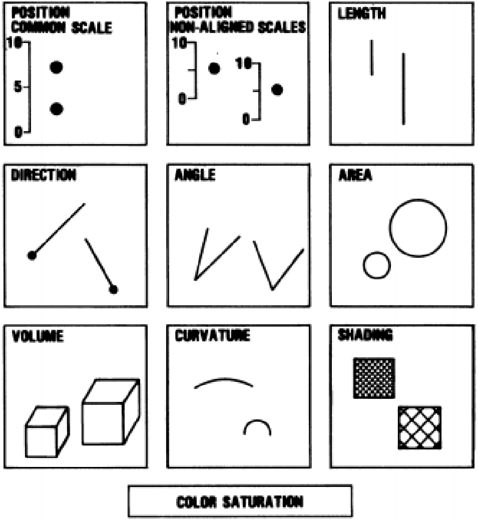
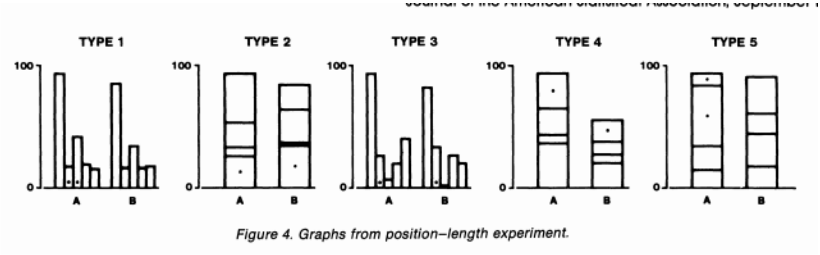


Figure 1. Elementary perceptual tasks.

Graphical perception: Theory, Experimentation, and Applications to the Development of Graphical Models

Position versus length

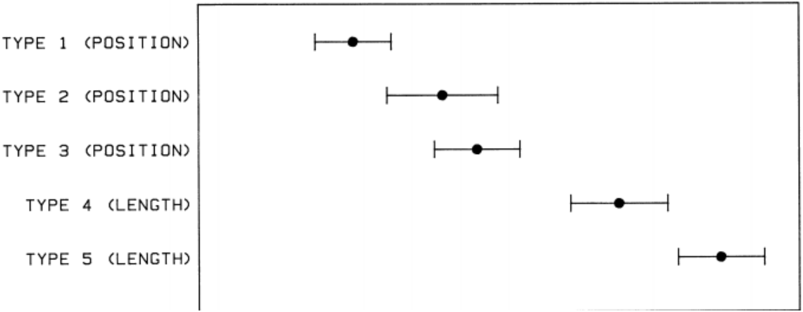


Journal of the American Statistical Association, September 1

Figure 4. Graphs from position–length experiment.

Graphical perception: Theory, Experimentation, and Applications to the Development of Graphical Models

Position versus length - results



Graphical perception: Theory, Experimentation, and Applications to the Development of Graphical Models

Position versus angle

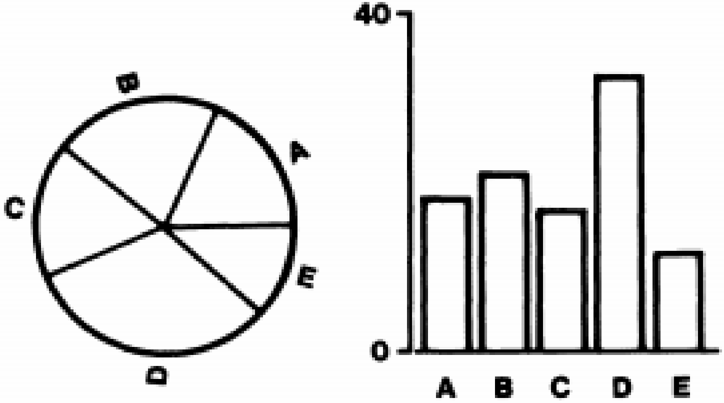
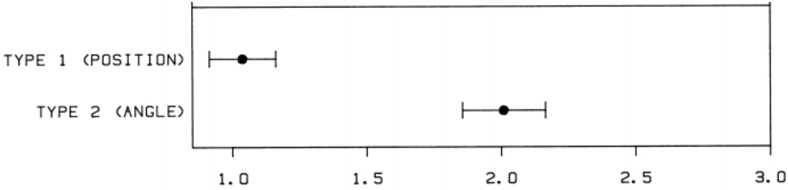


Figure 3. Graphs from position–angle experiment.

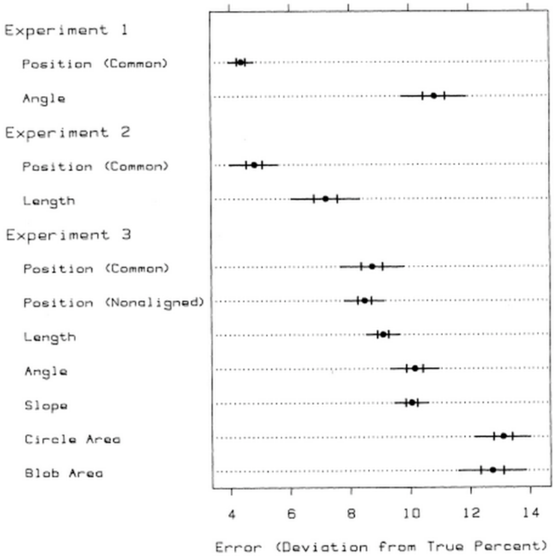
Graphical perception: Theory, Experimentation, and Applications to the Development of Graphical Models

Position versus angle - results



Graphical perception: Theory, Experimentation, and Applications to the Development of Graphical Models

More experimental results

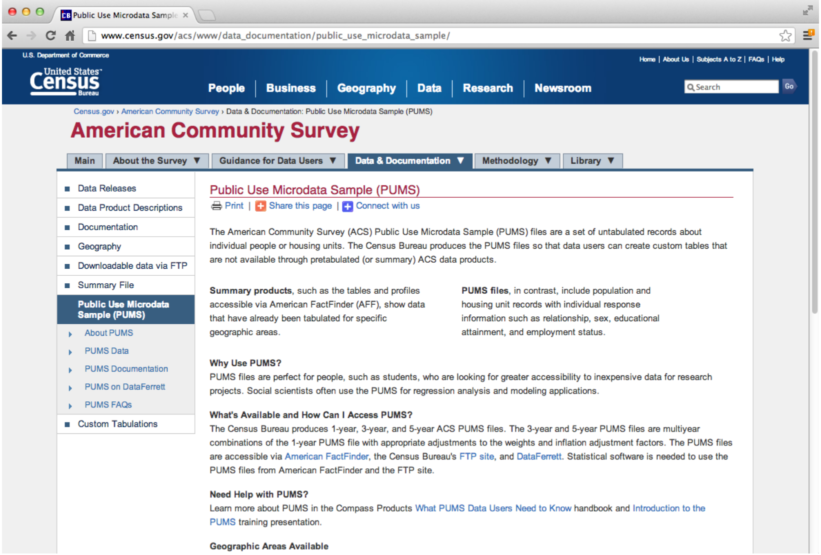


Graphical Perception and Graphical Methods for Analyzing Scientific Data

Summary

- Use common scales when possible
- When possible use position comparisons
- Angle comparisons are frequently hard to interpret (no piecharts!)
- No 3-D barcharts. Please.

Housing data



Public Use Microdata Sample
www.census.gov/acs/www/data_documentation/public_use_microdata_sample/
U.S. Department of Commerce
United States Census Bureau
Home | About Us | Subjects A to Z | FAQs | Help
People | Business | Geography | Data | Research | Newsroom
Search
Go
Census.gov › American Community Survey › Data & Documentation: Public Use Microdata Sample (PUMS)

American Community Survey

Main | About the Survey ▼ | Guidance for Data Users ▼ | Data & Documentation ▼ | Methodology ▼ | Library ▼
Data Releases
Data Product Descriptions
Documentation
Geography
Downloadable data via FTP
Summary File
Public Use Microdata Sample (PUMS)
About PUMS
PUMS Data
PUMS Documentation
PUMS on DataFerrett
PUMS FAQs
Custom Tabulations

Public Use Microdata Sample (PUMS)

Print | Share this page | Connect with us

The American Community Survey (ACS) Public Use Microdata Sample (PUMS) files are a set of untabulated records about individual people or housing units. The Census Bureau produces the PUMS files so that data users can create custom tables that are not available through pretabulated (or summary) ACS data products.

Summary products, such as the tables and profiles accessible via American FactFinder (AFF), show data that have already been tabulated for specific geographic areas.

PUMS files, in contrast, include population and housing unit records with individual response information such as relationship, sex, educational attainment, and employment status.

Why Use PUMS?
PUMS files are perfect for people, such as students, who are looking for greater accessibility to inexpensive data for research projects. Social scientists often use the PUMS for regression analysis and modeling applications.

What's Available and How Can I Access PUMS?
The Census Bureau produces 1-year, 3-year, and 5-year ACS PUMS files. The 3-year and 5-year PUMS files are multiyear combinations of the 1-year PUMS file with appropriate adjustments to the weights and inflation adjustment factors. The PUMS files are accessible via American FactFinder, the Census Bureau's FTP site, and DataFerrett. Statistical software is needed to use the PUMS files from American FactFinder and the FTP site.

Need Help with PUMS?
Learn more about PUMS in the Compass Products What PUMS Data Users Need to Know handbook and Introduction to the PUMS training presentation.

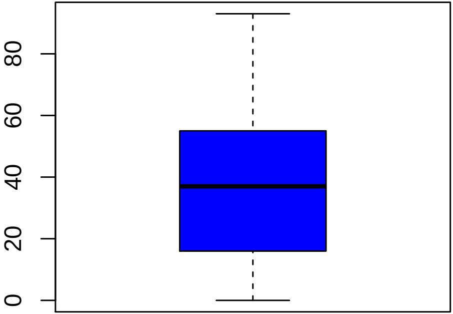
Geographic Areas Available

```r
pData <- read.csv("./data/ss06pid.csv")
```

Boxplots

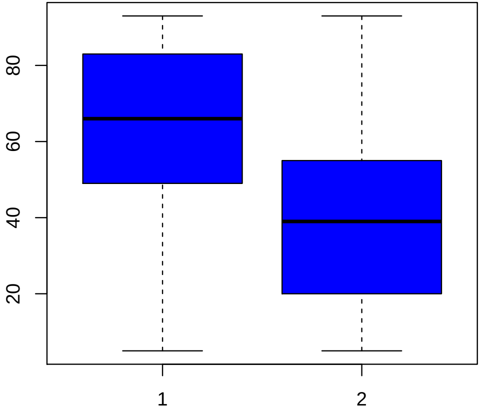
Important parameters: *col*, *varwidth*, *names*, *horizontal*

```
boxplot(pData$AGEP, col="blue")
```



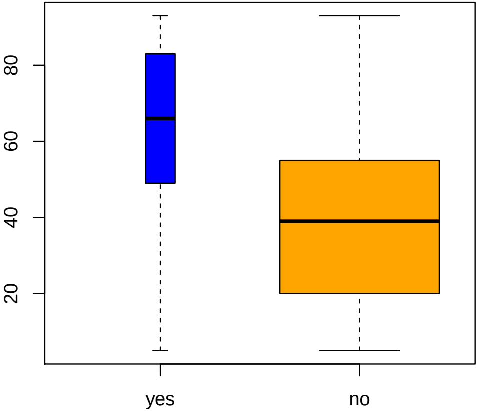
Boxplots

```
pData <- transform(pData, DDRS = factor(DDRS))
boxplot(AGEP ~ DDRS, data = pData, col = "blue")
```



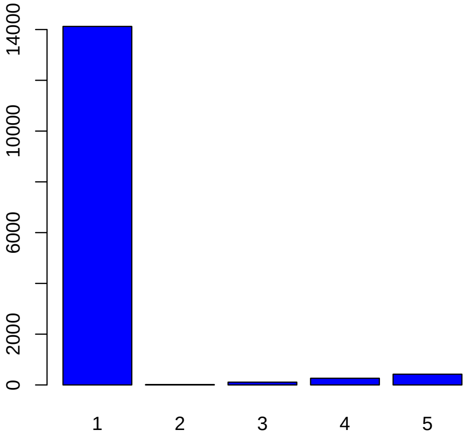
Boxplots

```
boxplot(AGEP ~ DDRS, data = pData, col = c("blue", "orange"
        names = c("yes", "no"), varwidth = TRUE)
```



Barplots

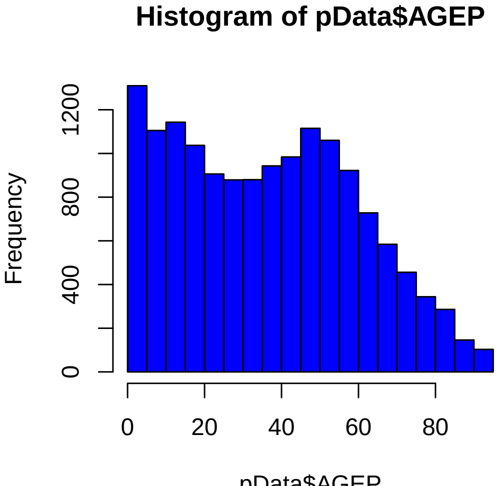
```
barplot(table(pData$CIT), col = "blue")
```



Histograms

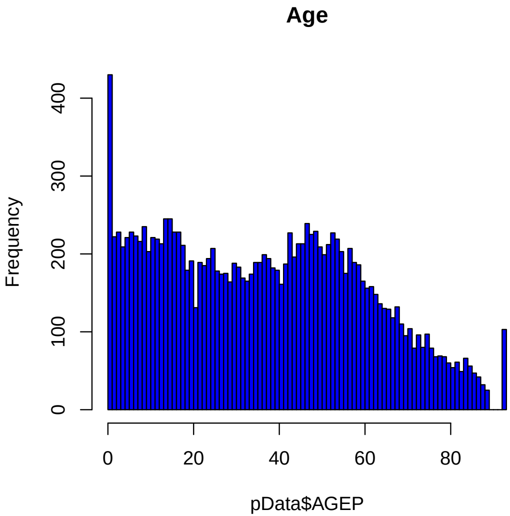
Important parameters: *breaks*,*freq*,*col*,*xlab*,*ylab*, *xlim*, ylim_ ,*main*

```r
hist(pData$AGEP, col = "blue")
```



Histograms

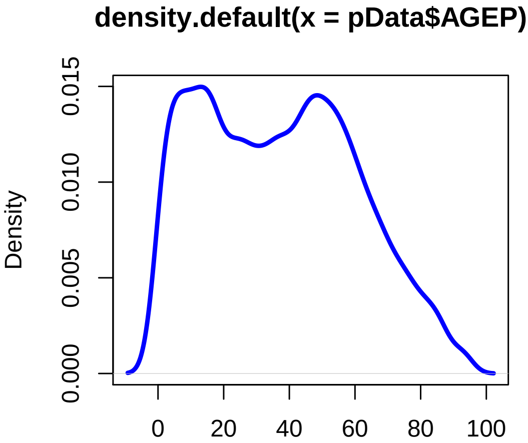
```
hist(pDataAGEP, col = "blue", breaks = 100, main = "Age")
```



Density plots

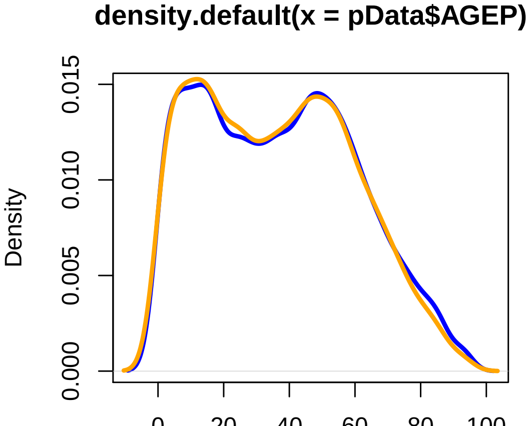
Important parameters (to plot): *col*,*lwd*,*xlab*,*ylab*,*xlim*,*ylim*

```
dens <- density(pData$AGEP)
plot(dens, lwd = 3, col = "blue")
```



Density plots - multiple distributions

```
dens <- density(pData$AGEP)
densMales <- density(pData$AGEP[which(pData$SEX == 1)])
plot(dens, lwd = 3, col = "blue")
lines(densMales, lwd = 3, col = "orange")
```


density.default(x = pData$AGEP)
Density