Exploratory Data Analysis

Maneesh Agrawala

CS 448B: Visualization
Spring 2016

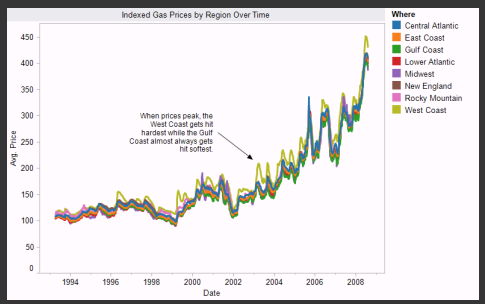
Assignment 2: Exploratory Data Analysis

Use Tableau to formulate & answer questions

First steps
- Step 1: Pick a domain
- Step 2: Pose questions
- Step 3: Find data
- Iterate

Create visualizations
- Interact with data
- Question will evolve
- Tableau

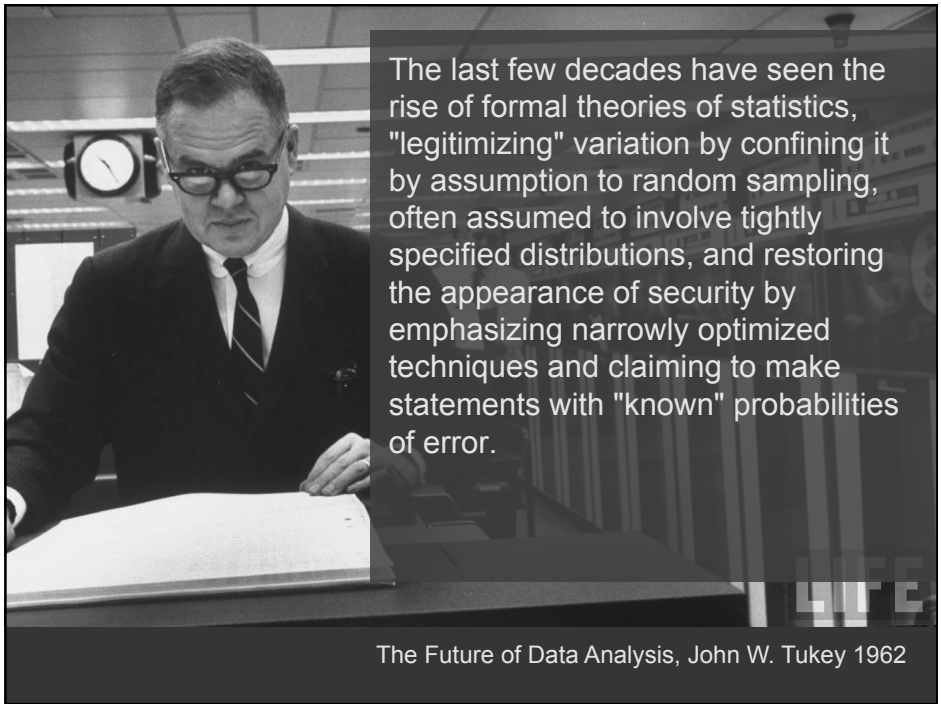
Make wiki notebook
- Keep record of all steps you took to answer the questions

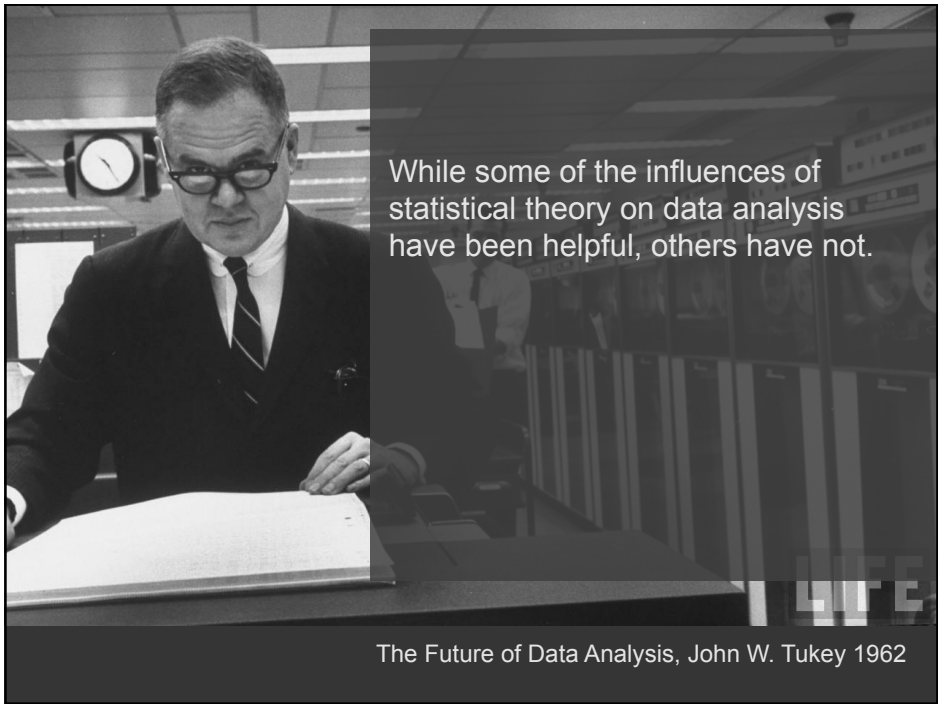


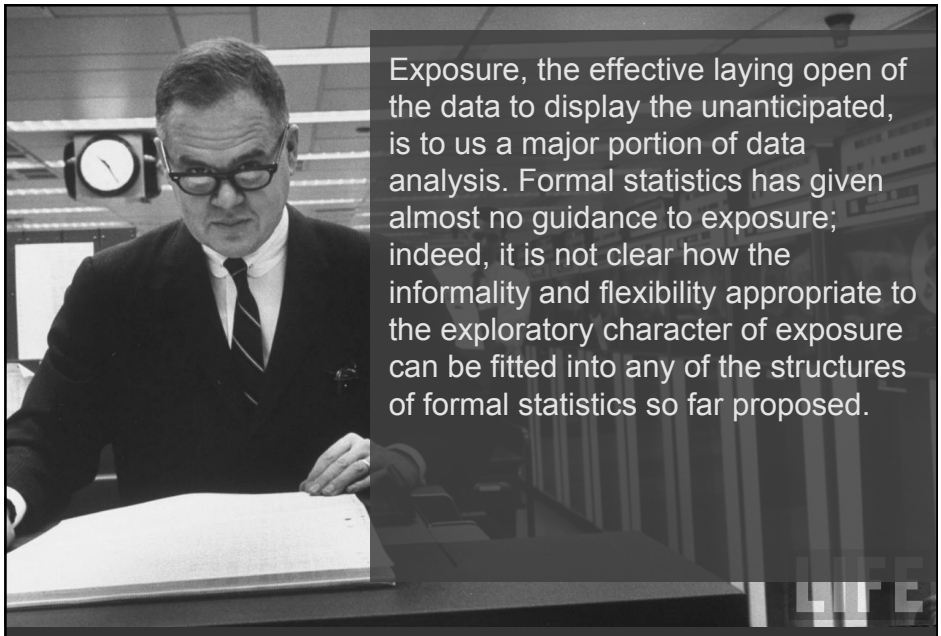
Due before class on Apr 18, 2016

Exploratory Data Analysis

The Future of Data Analysis, John W. Tukey 1962


The last few decades have seen the rise of formal theories of statistics, "legitimizing" variation by confining it by assumption to random sampling, often assumed to involve tightly specified distributions, and restoring the appearance of security by emphasizing narrowly optimized techniques and claiming to make statements with "known" probabilities of error.
The Future of Data Analysis, John W. Tukey 1962


While some of the influences of statistical theory on data analysis have been helpful, others have not.
The Future of Data Analysis, John W. Tukey 1962



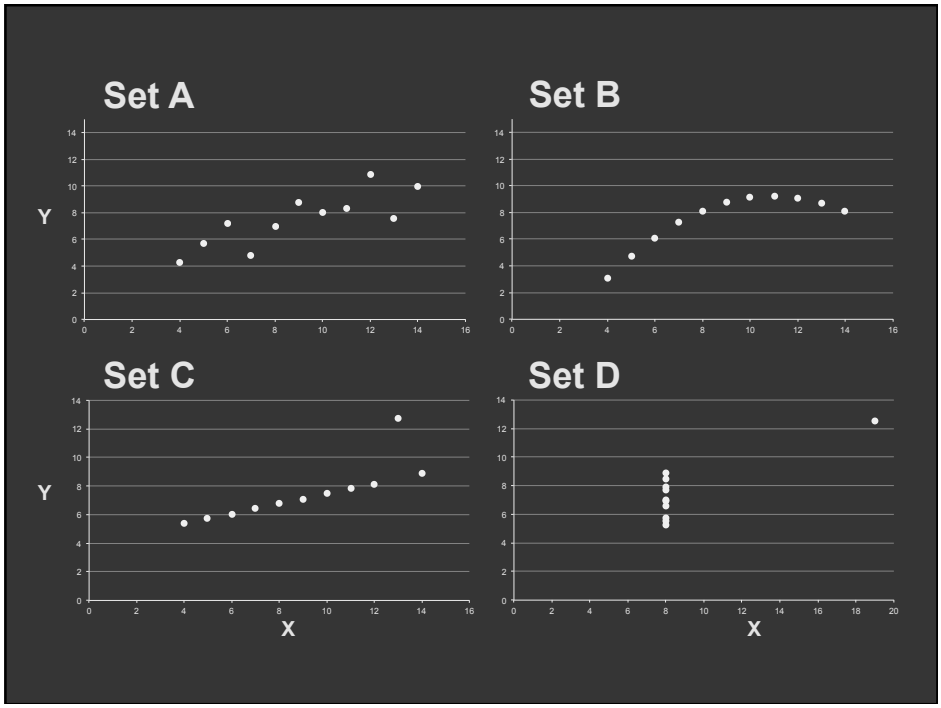
The Future of Data Analysis, John W. Tukey 1962

Set A		Set B		Set C		Set D	
X	Y	X	Y	X	Y	X	Y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.11	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Summary Statistics Linear Regression

$u_X = 9.0$ $\sigma_X = 3.317$ $Y = 3 + 0.5\ X$
$u_Y = 7.5$ $\sigma_Y = 2.03$ $R^2 = 0.67$

[Anscombe 73]



Topics

Exploratory Data Analysis

- Data Diagnostics
- Graphical Methods
- Data Transformation

Confirmatory Data Analysis

- Statistical Hypothesis Testing
- Graphical Inference

Data Diagnostics

Bureau of Justice Statistics – Data Online
http://bjs.ojp.usdoj.gov/

Reported crime in Alabama

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	4525375	4029.3	987	2732.4	309.9
2005	4548327	3900	955.8	2656	289
2006	4599030	3937	968.9	2645.1	322.9
2007	4627851	3974.9	980.2	2687	307.7
2008	4661900	4081.9	1080.7	2712.6	288.6

Reported crime in Alaska

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	657755	3370.9	573.6	2456.7	340.6
2005	663253	3615	622.8	2601	391
2006	670053	3582	615.2	2588.5	378.3
2007	683478	3373.9	538.9	2480	355.1
2008	686293	2928.3	470.9	2219.9	237.5

Reported crime in Arizona

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	5739879	5073.3	991	3118.7	963.5
2005	5953007	4827	946.2	2958	922
2006	6166318	4741.6	953	2874.1	914.4
2007	6338755	4502.6	935.4	2780.5	786.7
2008	6500180	4087.3	894.2	2605.3	587.8

Reported crime in Arkansas

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	2750000	4033.1	1096.4	2699.7	237
2005	2775708	4068	1085.1	2720	262
2006	2810872	4021.6	1154.4	2596.7	270.4
2007	2834797	3945.5	1124.4	2574.6	246.5
2008	2855390	3843.7	1182.7	2433.4	227.6

Reported crime in California

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	35842038	3423.9	686.1	2033.1	704.8
2005	36154147	3321	692.9	1915	712
2006	36457549	3175.2	676.9	1831.5	666.8
2007	36553215	3032.6	648.4	1784.1	600.2
2008	36756666	2940.3	646.8	1769.8	523.8

Reported crime in Colorado

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	4601821	3918.5	717.3	2679.5	521.6

Data “Wrangling”

One often needs to manipulate data prior to analysis. Tasks include reformatting, cleaning, quality assessment, and integration

Some approaches:

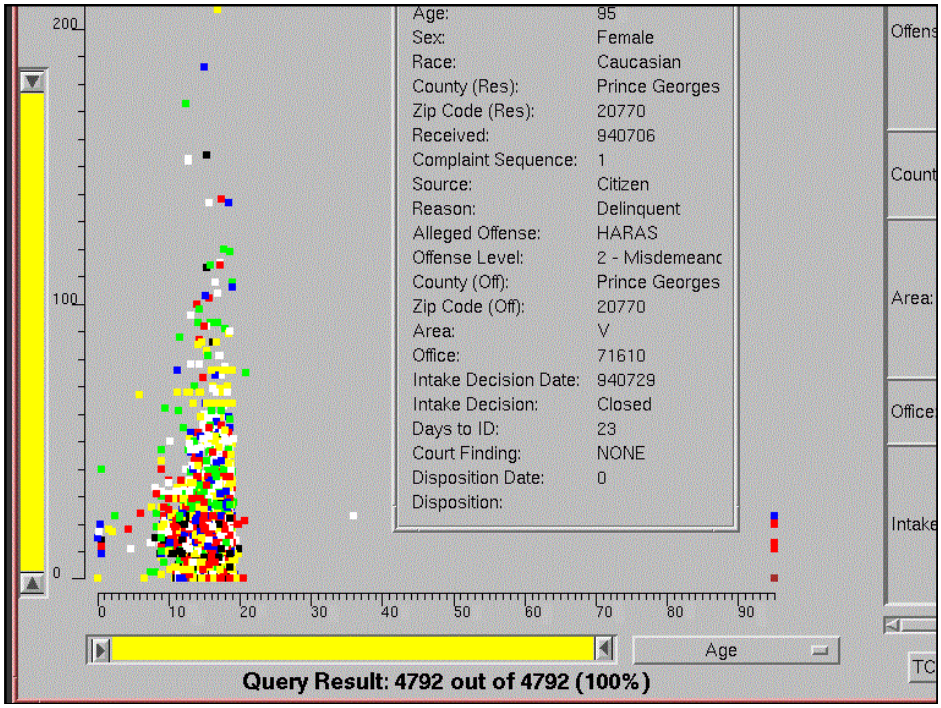
- Writing custom scripts
- Manual manipulation in spreadsheets
- Data Wrangler: http://vis.stanford.edu/wrangler
- Google Refine: http://code.google.com/p/google-refine

How to gauge the quality of a visualization?

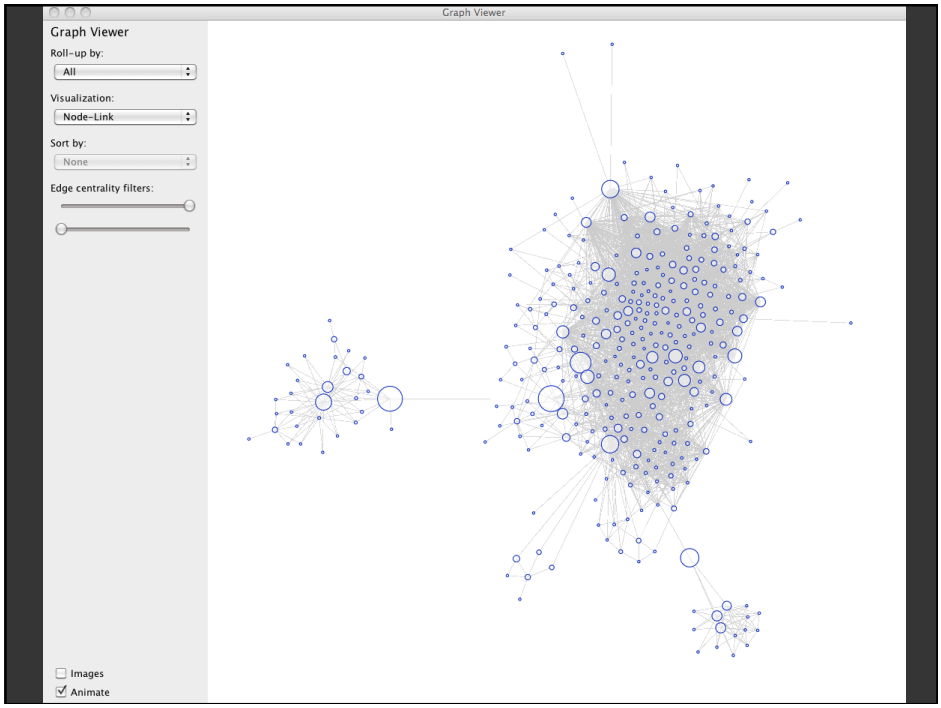
“The first sign that a visualization is good is that it shows you a problem in your data…

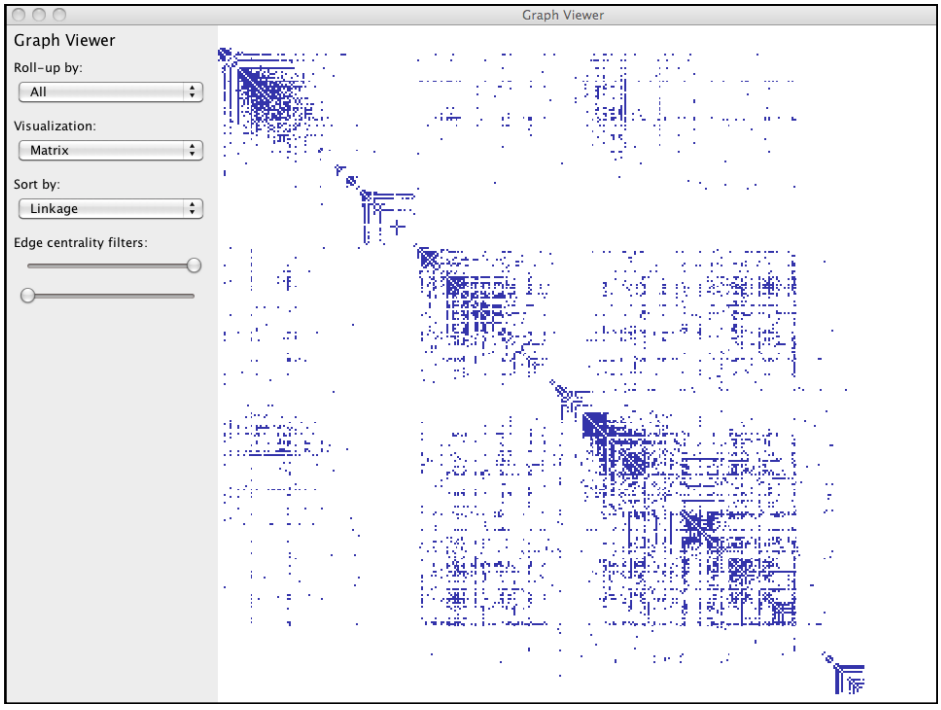
…every successful visualization that I've been involved with has had this stage where you realize, "Oh my God, this data is not what I thought it would be!" So already, you've discovered something.”

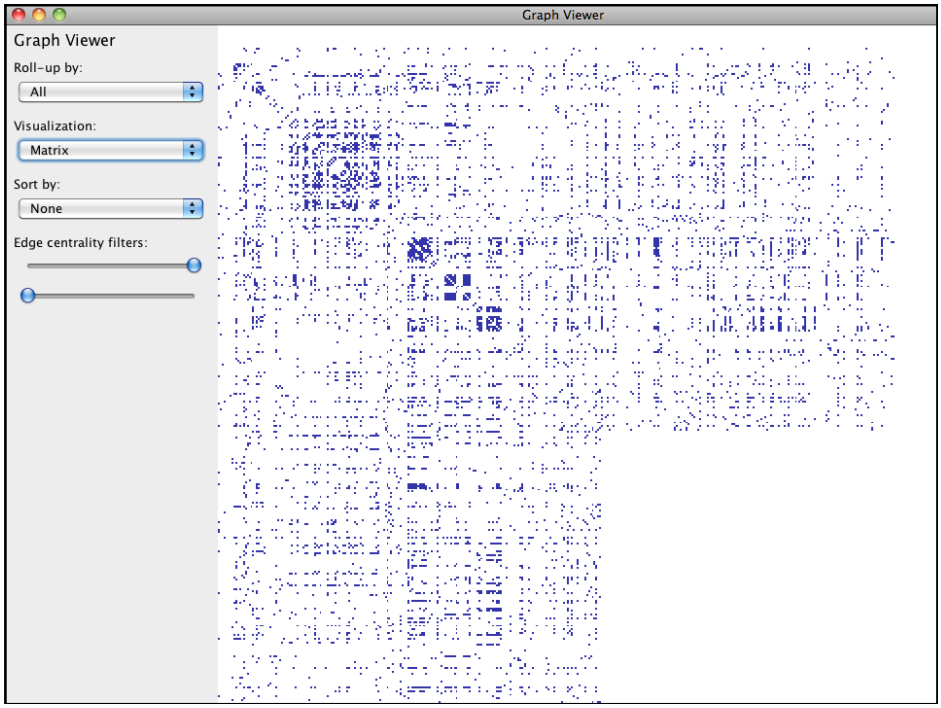
- Martin Wattenberg


200
100
0
0
10
20
30
40
50
60
70
80
90
Age:
Sex:
Race:
County (Res):
Zip Code (Res):
Received:
Complaint Sequence:
Source:
Reason:
Alleged Offense:
Offense Level:
County (Off):
Zip Code (Off):
Area:
Office:
Intake Decision Date:
Intake Decision:
Days to ID:
Court Finding:
Disposition Date:
Disposition:
95
Female
Caucasian
Prince Georges
20770
940706
1
Citizen
Delinquent
HARAS
2 - Misdemeanc
Prince Georges
20770
V
71610
940729
Closed
23
NONE
0
Offens
Count
Area:
Office
Intake
Age
TC
Query Result: 4792 out of 4792 (100%)


facebook
Profile edit
Friends
Networks
Inbox
home account privacy logout
Search
Applications edit
Photos
Groups
Events
Marketplace
The New York Times News Quiz
more
Send Bill a Gift
Send Bill a Message
Poke Bill!
Friends
See All
Melinda Gates
Steve Ballmer
Mark Zuckerberg
William Randolph
Bono
Bill Gates
is glad he finally joined facebook and hopes you will too!!! :)
Updated 6 minutes ago
Mini-Feed
Displaying 15 stories
See All
Update: Bill has posted a note:
Friends, I have finally caved and joined facebook, America's fastest-growing social-networking web site! At first I didn't join because you needed a college alumni address, and I never quite got one... Then when the place started opening up to high schools and corporations, with everyone and his grandmother joining, I wanted in. But by then I was mad I didn't have any shares in this $15 billion baby. ... So just now I decided to plunk down $240 million to buy 1.6% of the company from cool kid CEO Mark Zuckerberg. Sure I saw the potential for ad revenue right away -- but this is wild. I've never had my own Facebook page before!
Don't have a lot of friends yet but I've been running into people... Seeing their status updates... Wow, it's a great place to check up on my employees and my kids! They keep saying what their weekend plans are in their status bars. And of course I love how you can add all these little software applications to the page. Or write your very own!
I'm still getting the hang of the whole thing... It seems there are a variety of forums where I can vent to others and display details of my life? O.K.: It wasn't easy being so much smarter than everyone else, pretending to be a grown-up over the telephone so I could get grown-up jobs programming these new things called computers when I was still a child.
At college I led the anti-social group. Never led a social group, or had a social network... Ha, ha, now I've done even better than that: I've bought a piece of the national friend system! Take that cliquesters. Anyone who ever ignored me in the dining hall... Got friends? I own 1.6 % of your friends.
But don't worry about them. Send me a message! Write on my wall!
Bill and Mark Zuckerberg are now friends
Bill and Warren Buffett have joined the group Save the World Now through Creative Capitalism (3 Members)


Graph Viewer
Graph Viewer
Roll-up by:
All
Visualization:
Node-Link
Sort by:
None
Edge centrality filters:
Images
Animate


Graph Viewer
Graph Viewer
Roll-up by:
All
Visualization:
Matrix
Sort by:
Linkage
Edge centrality filters:



Visualize Friends by School?

Berkeley	|||||||||||||||||||||||||||||||
Cornell	||||
Harvard	|||||||||
Harvard University	|||||||
Stanford	||||||||||||||||||
Stanford University	|||||||||
UC Berkeley	||||||||||||||||||||
UC Davis	|||||||||
University of California at Berkeley	||||||||||||||
University of California, Berkeley	||||||||||||||||||
University of California, Davis	|||

Data Quality & Usability Hurdles

Missing Data	no measurements, redacted, …?
Erroneous Values	misspelling, outliers, …?
Type Conversion	e.g., zip code to lat-lon
Entity Resolution	diff. values for the same thing?
Data Integration	effort/errors when combining data

LESSON: **Anticipate problems with your data.**
Many research problems around these issues!

Exploratory Analysis: Effectiveness of Antibiotics

What questions might we ask?

Table 1: Burtin's data.

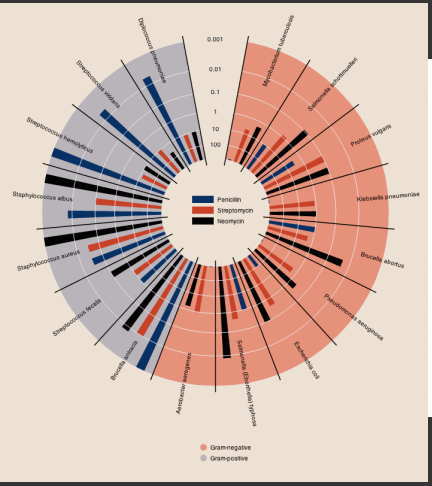
Bacteria	Antibiotic			Gram Staining
	Penicillin	Streptomycin	Neomycin	
Aerobacter *aerogenes*	870	1	1.6	negative
Brucella *abortus*	1	2	0.02	negative
Brucella *anthracis*	0.001	0.01	0.007	positive
Diplococcus *pneumoniae*	0.005	11	10	positive
Escherichia *coli*	100	0.4	0.1	negative
Klebsiella *pneumoniae*	850	1.2	1	negative
Mycobacterium *tuberculosis*	800	5	2	negative
Proteus *vulgaris*	3	0.1	0.1	negative
Pseudomonas *aeruginosa*	850	2	0.4	negative
Salmonella (Eberthella) *typhosa*	1	0.4	0.008	negative
Salmonella *schottmuelleri*	10	0.8	0.09	negative
Staphylococcus *albus*	0.007	0.1	0.001	positive
Staphylococcus *aureus*	0.03	0.03	0.001	positive
Streptococcus *fecalis*	1	1	0.1	positive
Streptococcus *hemolyticus*	0.001	14	10	positive
Streptococcus *viridans*	0.005	10	40	positive

The Data Set

Genus of Bacteria	String
Species of Bacteria	String
Antibiotic Applied	String
Gram-Staining?	Pos / Neg
Min. Inhibitory Concent. (g)	Number

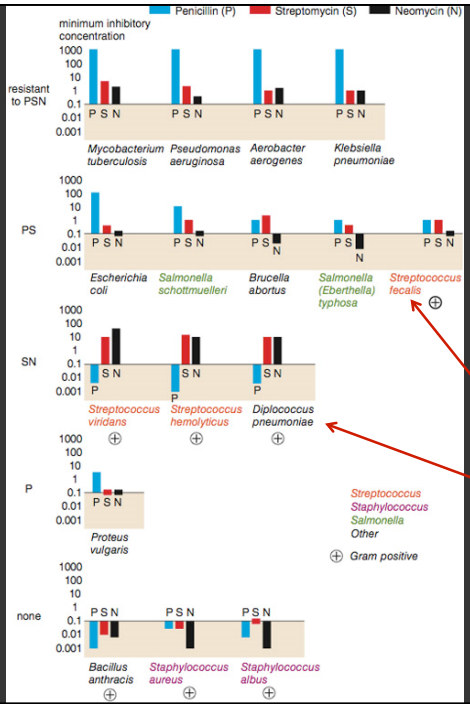
Collected prior to 1951

Will Burtin, 1951



Bacteria	Penicillin	Antibiotic Streptomycin	Neomycin	Gram stain
Aerobacter aerogenes	870	1	1.6	–
Brucella abortus	1	2	0.02	–
Bacillus anthracis	0.001	0.01	0.007	+
Diplococcus pneumoniae	0.005	11	10	+
Escherichia coli	100	0.4	0.1	–
Klebsiella pneumoniae	850	1.2	1	–
Mycobacterium tuberculosis	800	5	2	–
Proteus vulgaris	3	0.1	0.1	–
Pseudomonas aeruginosa	850	2	0.4	–
Salmonella (Eberthella) typhosa	1	0.4	0.008	–
Salmonella schottmuelleri	10	0.8	0.09	–
Staphylococcus albus	0.007	0.1	0.001	+
Staphylococcus aureus	0.03	0.03	0.001	+
Streptococcus fecalis	1	1	0.1	+
Streptococcus hemolyticus	0.001	14	10	+
Streptococcus viridans	0.005	10	40	+

How do the drugs compare?



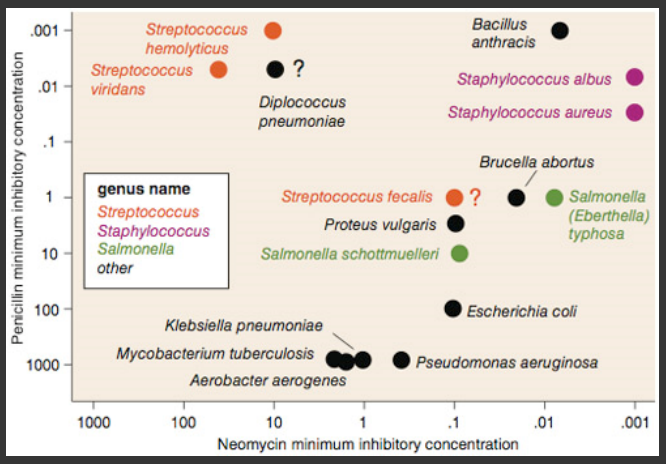
How do the bacteria group with respect to antibiotic resistance?

Not a streptococcus! (realized ~30 yrs later)

Really a streptococcus! (realized ~20 yrs later)

Wainer & Lysen

American Scientist, 2009



How do the bacteria group w.r.t. resistance?
Do different drugs correlate?

Wainer & Lysen

American Scientist, 2009

Lessons

Exploratory Process

1. **Construct graphics to address questions**
2. **Inspect "answer" and assess new questions**
3. **Repeat!**

Transform the data appropriately (e.g., invert, log)

"Show data variation, not design variation"

-Tufte

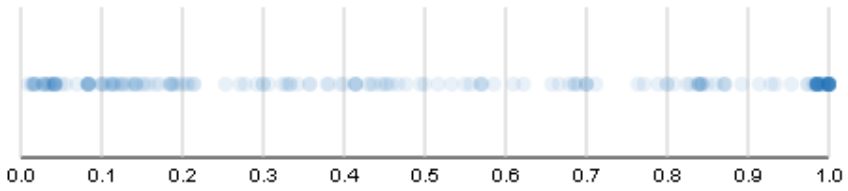
Exploratory Analysis: Participation on Amazon's Mechanical Turk

The Data Set (~200 rows)

Turker ID	**String**
Avg. Completion Percentage	**Number [0,1]**

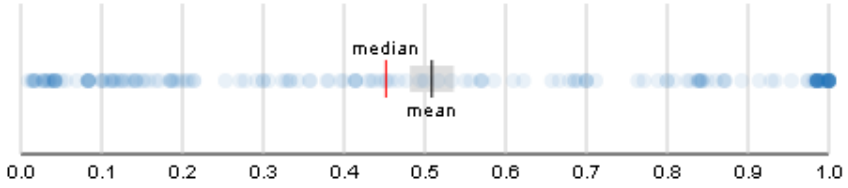
Collected in 2009 by Heer & Bostock.

What questions might we ask of the data?
What charts might provide insight?



Turker Completion Percentage

Dot Plot (with transparency to indicate overlap)

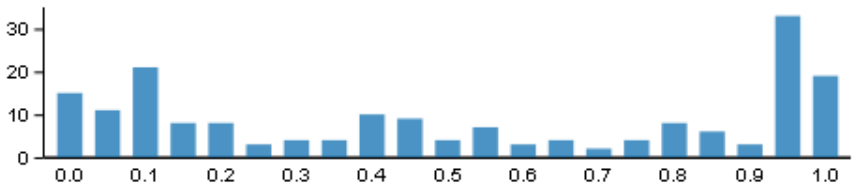


Turker Completion Percentage

Dot Plot w/ Reference Lines

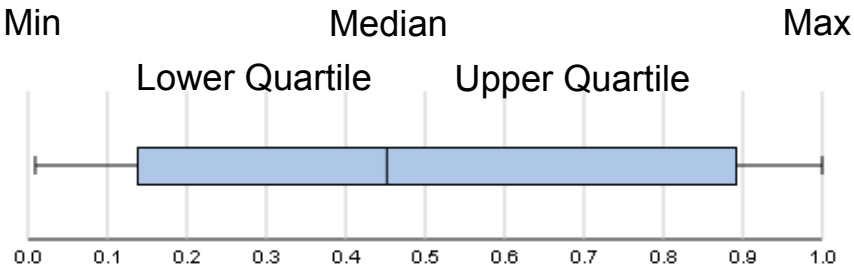
0	1 1 1 2 2 2 2 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 5 6 7 8 8 8 8 8 8 9
1	0 0 0 0 1 1 1 1 2 2 3 3 3 3 4 4 4 4 5 5 6 7 7 8 9 9 9 9 9
2	0 0 1 1 1 5 7 8 9
3	0 0 1 2 3 3 3 4 6 6 8 8
4	0 0 1 1 1 1 3 3 4 5 5 5 6 7 8 9
5	0 2 3 5 6 7 7 7 9
6	1 2 6 7 8 9 9 9
7	0 0 0 1 6 7 9
8	0 0 1 2 3 4 4 4 4 4 4 4 5 6 7 7 7 9
9	1 3 3 5 7 8 8 8 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9
10	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

Stem-and-Leaf Plot

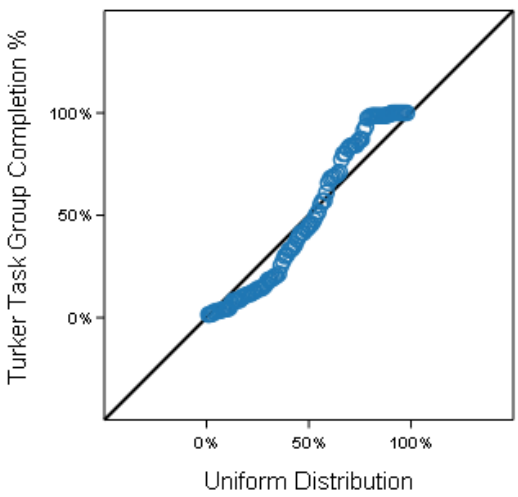


Turker Completion Percentage

Histogram (binned counts)



Box (and Whiskers) Plot

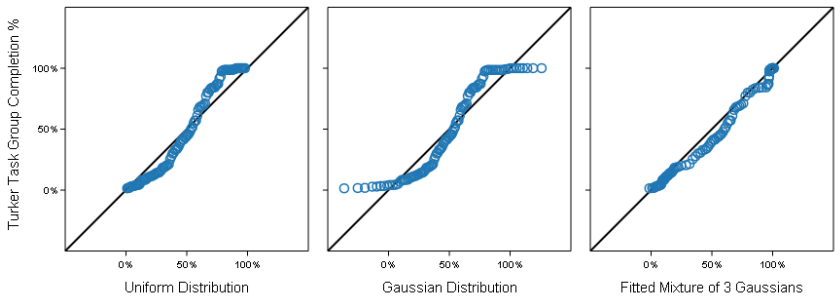


Used to compare two distributions; in this case, one actual and one theoretical.

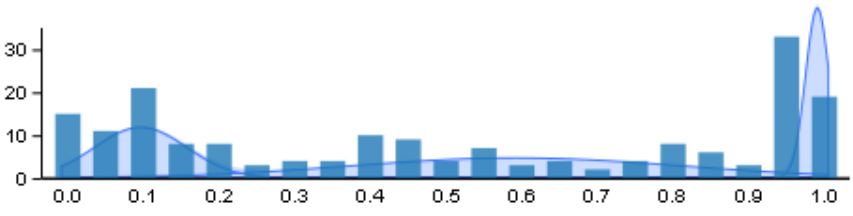
Plots the quantiles (here, the percentile values) against each other.

Similar distributions lie along the diagonal. If linearly related, values will lie along a line, but with potentially varying slope and intercept.

Quantile-Quantile Plot



Quantile-Quantile Plots



Turker Completion Percentage

Histogram + Fitted Mixture of 3 Gaussians

Lessons

Even for "simple" data, a variety of graphics might provide insight. Again, tailor the choice of graphic to the questions being asked, but be open to surprises.

Graphics can be used to understand and help assess the quality of statistical models.

Premature commitment to a model and lack of verification can lead an analysis astray.

Confirmatory Data Analysis

Some Uses of Formal Statistics

What is the probability that the pattern I'm seeing might have arisen by chance?

With what parameters does the data best fit a given function? What is the goodness of fit?

How well do one (or more) data variables predict another?

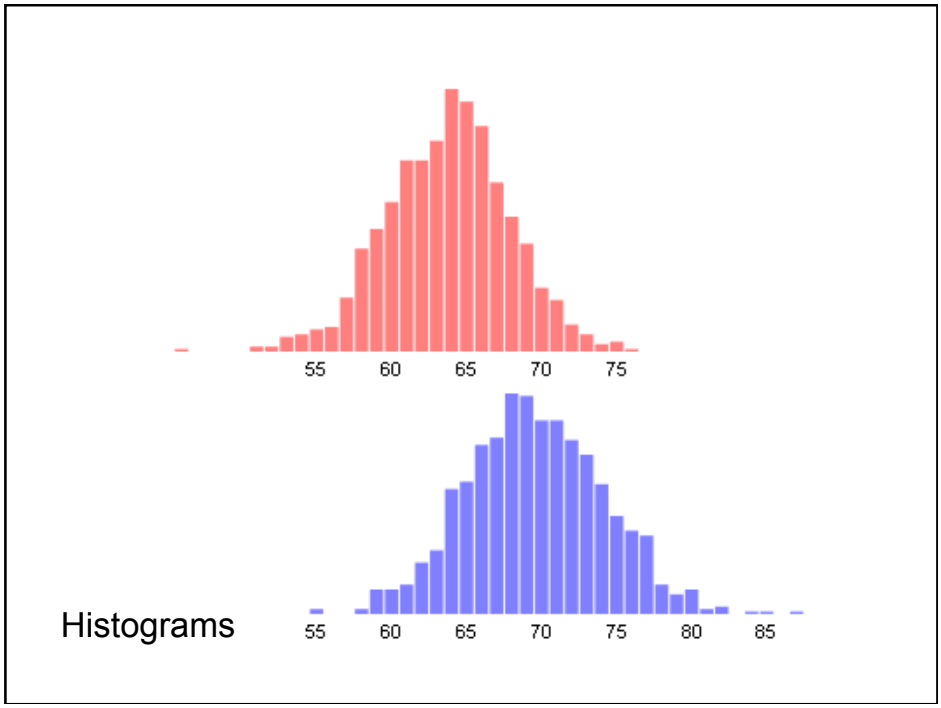
…and many others.

Example: Heights by Gender

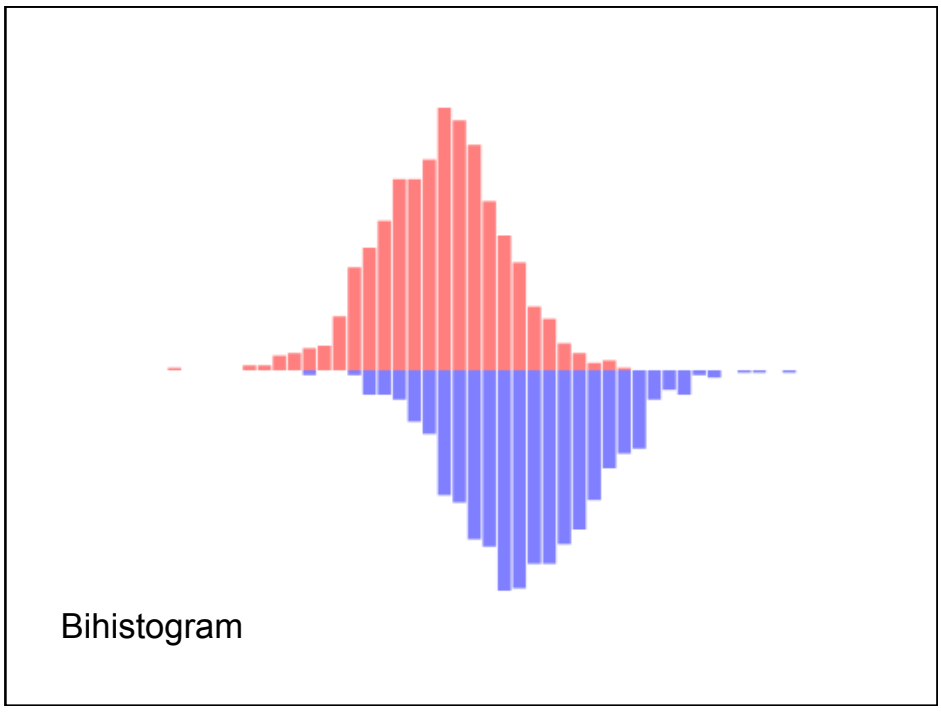
Gender	Male / Female	
Height (in)		Number
$\mu_m = 69.4$	$\sigma_m = 4.69$	$N_m = 1000$
$\mu_f = 63.8$	$\sigma_f = 4.18$	$N_f = 1000$

Is this difference in heights significant?

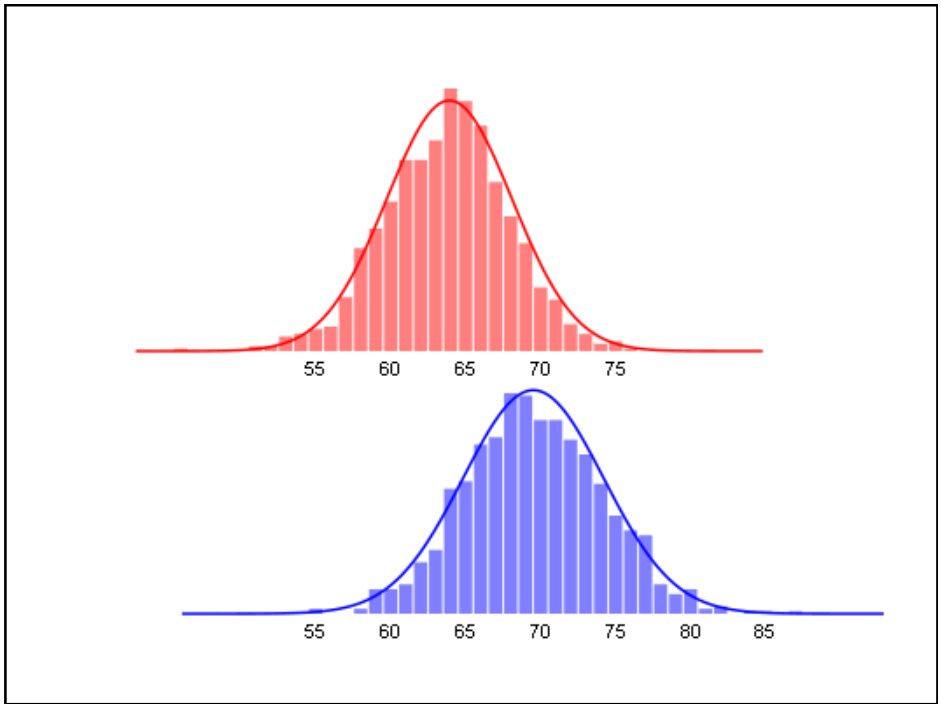
In other words: assuming no true difference, what is the prob. that our data is due to chance?

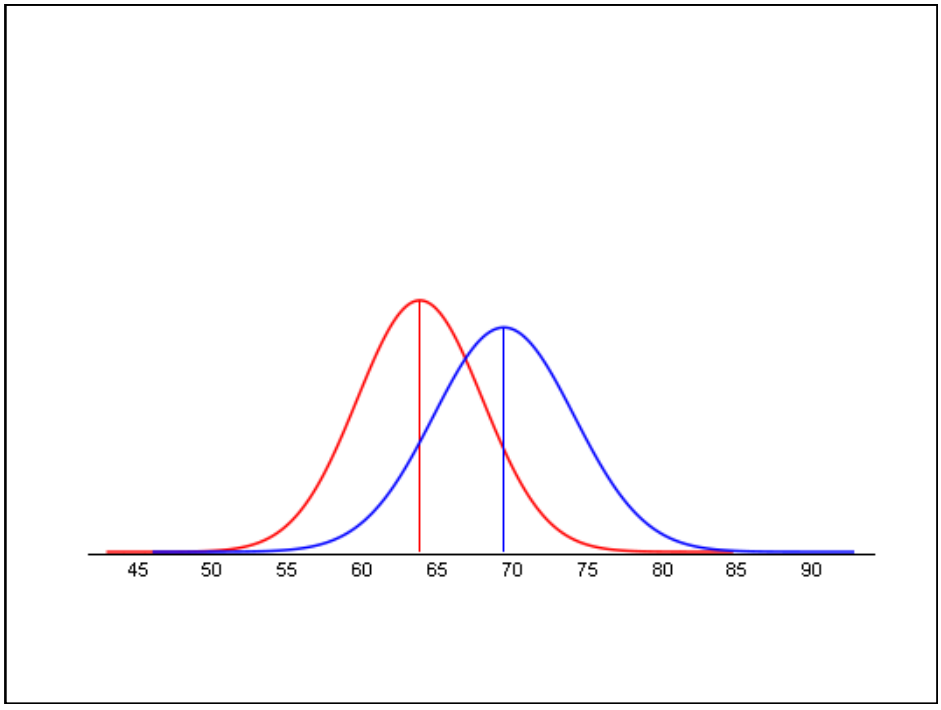


Histograms



Bihistogram


55
60
65
70
75
55
60
65
70
75
80
85


45
50
55
60
65
70
75
80
85
90

Formulating a Hypothesis

Null Hypothesis (H_0): $\mu_m = \mu_f$ (population)

Alternate Hypothesis (H_a): $\mu_m \neq \mu_f$ (population)

A statistical hypothesis test assesses the likelihood of the null hypothesis.

What is the probability of sampling the observed data assuming population means are equal?

This is called the *p* value

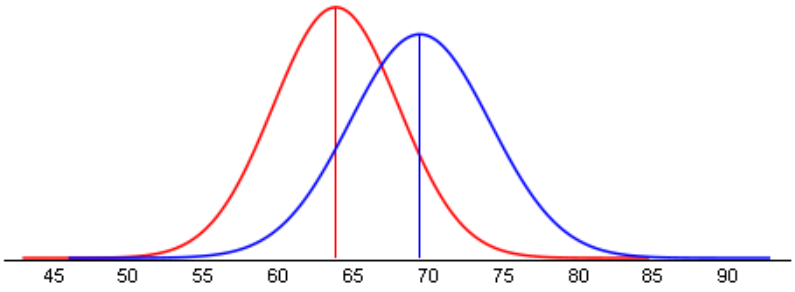
Testing Procedure

Compute a test statistic. This is a number that in essence summarizes the difference.

Compute test statistic

$$Z = \frac{\mu_m - \mu_f}{\sqrt{\sigma^2_m / N_m + \sigma^2_f / N_f}}$$

$\mu_m - \mu_f = 5.6$



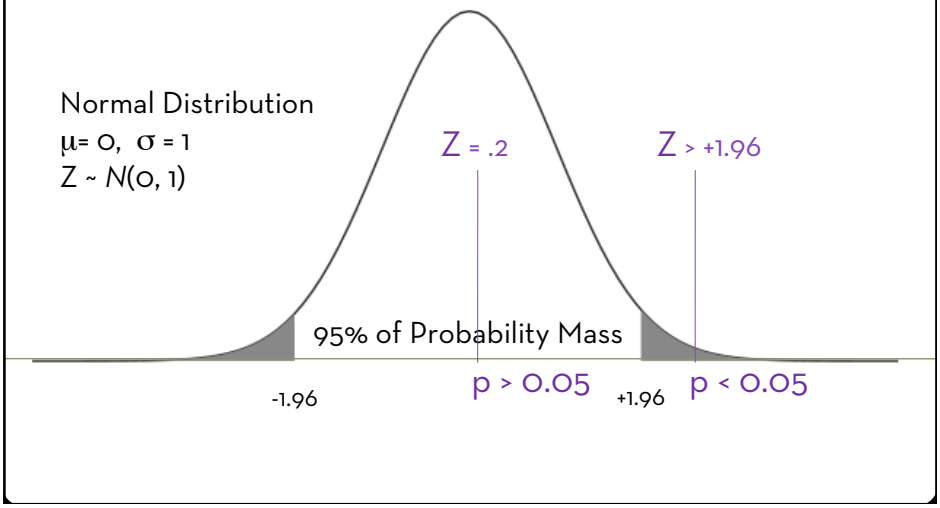
Testing Procedure

Compute a test statistic. This is a number that in essence summarizes the difference.

The possible values of this statistic come from a known probability distribution.

According to this distribution, look up the probability of seeing a value meeting or exceeding the test statistic. This is the *p* value.

Lookup probability of test statistic



Statistical Significance

- The threshold at which we consider it safe (or reasonable?) to *reject the null hypothesis*.
- If $p < 0.05$, we typically say that the observed effect or difference is statistically significant.
- This means that there is a less than 5% chance that the observed data is due to chance.
- Note that the choice of 0.05 is a somewhat arbitrary threshold (chosen by R. A. Fisher)

Common Statistical Methods

Question	Data Type	Parametric	Non-Parametric

Assumes a particular distribution for the data -- usually normal, a.k.a. Gaussian.

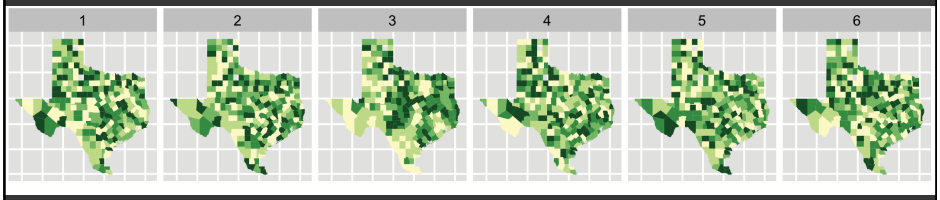
Does not assume a distribution. Typically works on rank orders.

Common Statistical Methods

Question	Data Type	Parametric	Non-Parametric
Do data distributions have different "centers"? (aka "location" tests)	2 uni. dists > 2 uni. dists > 2 multi. dists	t-Test ANOVA MANOVA	Mann-Whitney U Kruskal-Wallis Median Test
Are observed counts significantly different?	Counts in categories		χ^2 (chi-squared)
Are two vars related?	2 variables	Pearson coeff.	Rank correl.
Do 1 (or more) variables predict another?	Continuous Binary	Linear regression Logistic regression	

Graphical Inference

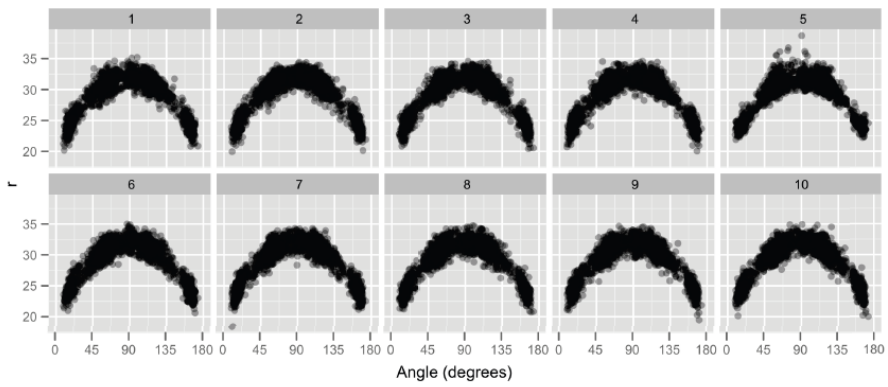
Buja Cook, Hoffman, Wickham et al.



Choropleth maps of cancer deaths in Texas.

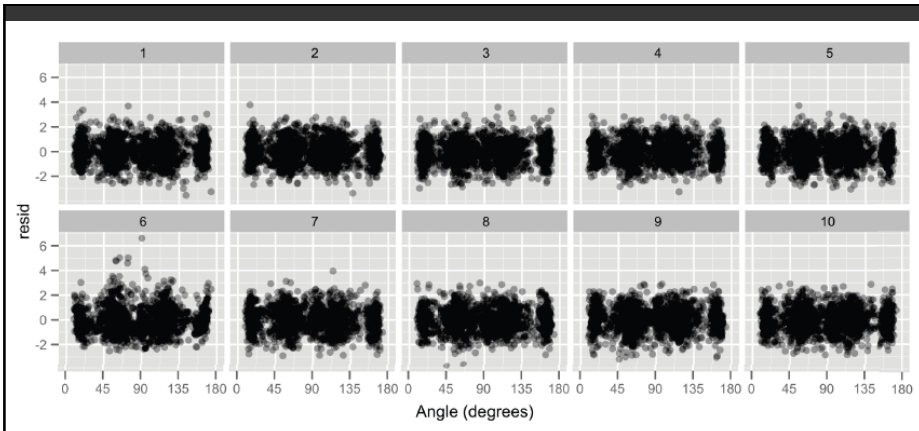
One plot shows a real data sets. The others are simulated under the null hypothesis of spatial independence.

Can you spot the real data? If so, you have some evidence of spatial dependence in the data.



Distance vs. angle for 3 point shots by the LA Lakers.

One plot is the real data. The others are generated according to a null hypothesis of quadratic relationship.



Residual distance vs. angle for 3 point shots.

One plot is the real data. The others are generated using an assumption of normally distributed residuals.

Summary

Exploratory analysis may combine graphical methods, data transformations, and statistics

Use questions to uncover more questions

Formal methods may be used to confirm, sometimes on held-out or new data

Visualization can further aid assessment of fitted statistical models