Multidimensional Visualization

Maneesh Agrawala

CS 294-10: Visualization
Fall 2013

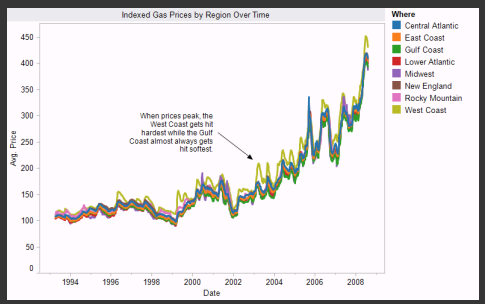
Assignment 2: Exploratory Data Analysis

Use existing software 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 Sep 30, 2013

Last Time:
Exploratory Data Analysis

Topics

Exploratory Data Analysis

- Data Diagnostics
- Graphical Methods
- Data Transformation

Confirmatory Data Analysis

- Statistical Hypothesis Testing
- Graphical Inference

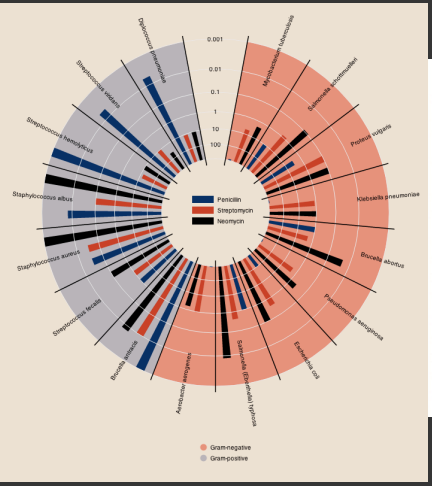
Exploratory Analysis: Effectiveness of Antibiotics

What questions might we ask?

Table 1: Burtin's data.

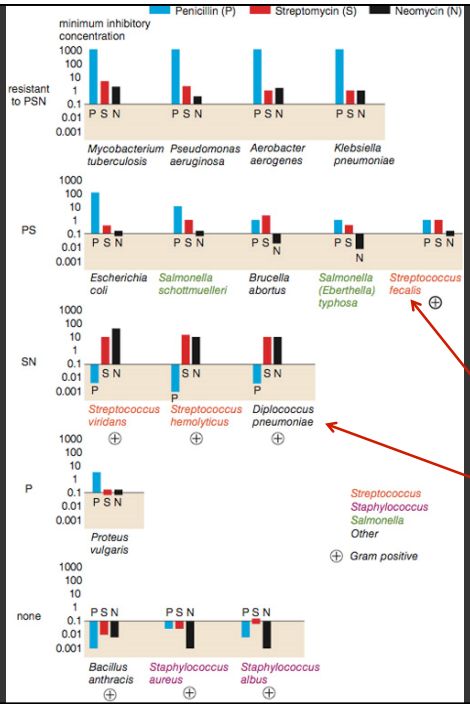
Bacteria	Antibiotic			Gram Staining
	Penicillin	Streptomycin	Neomycin	
Aerobacter *aerogenes*	870	1	1.6	negative
Brucella *abortus*	1	2	0.02	negative
Brucella *anthracis*	0.001	0.01	0.007	positive
Diplococcus *pneumoniae*	0.005	11	10	positive
Escherichia *coli*	100	0.4	0.1	negative
Klebsiella *pneumoniae*	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

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

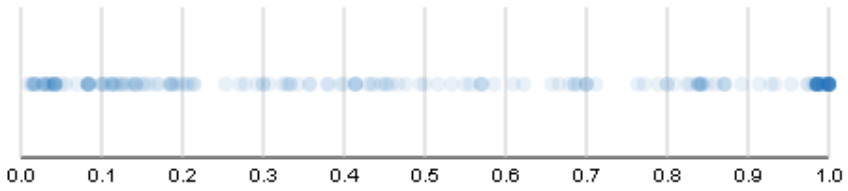
Exploratory Analysis: Participation on Amazon's Mechanical Turk

The Data Set (~200 rows)

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

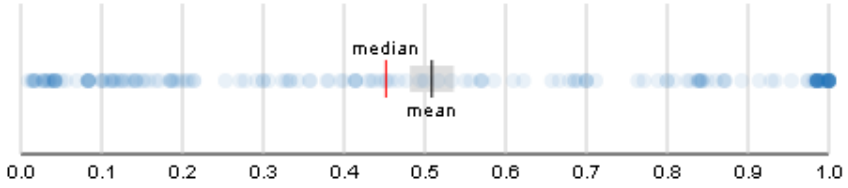
Collected in 2009 by Heer & Bostock.

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



Turker Completion Percentage

Dot Plot (with transparency to indicate overlap)

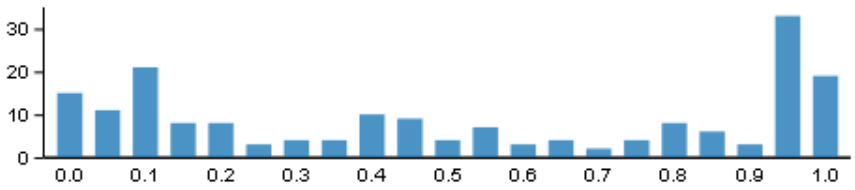


Turker Completion Percentage

Dot Plot w/ Reference Lines

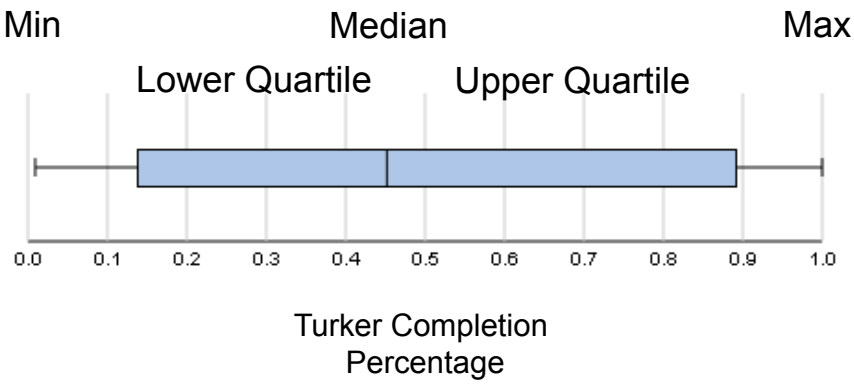
0	1 1 1 2 2 2 2 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 5 6 7 8 8 8 8 8 8 9
1	0 0 0 0 1 1 1 1 2 2 3 3 3 3 4 4 4 4 5 5 6 7 7 8 9 9 9 9 9
2	0 0 1 1 1 5 7 8 9
3	0 0 1 2 3 3 3 4 6 6 8 8
4	0 0 1 1 1 1 3 3 4 5 5 5 6 7 8 9
5	0 2 3 5 6 7 7 7 9
6	1 2 6 7 8 9 9 9
7	0 0 0 1 6 7 9
8	0 0 1 2 3 4 4 4 4 4 4 4 5 6 7 7 7 9
9	1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

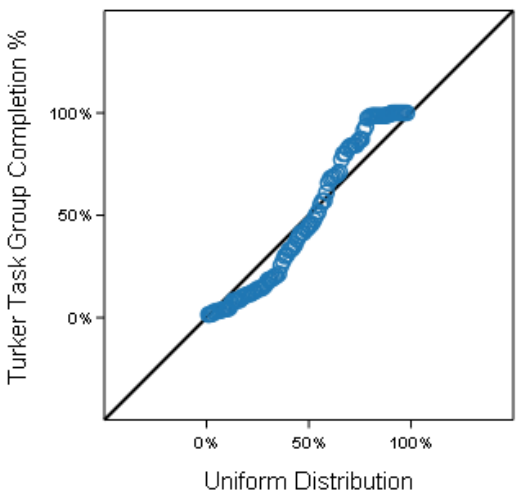


Turker Completion Percentage

Histogram (binned counts)



Box (and Whiskers) Plot

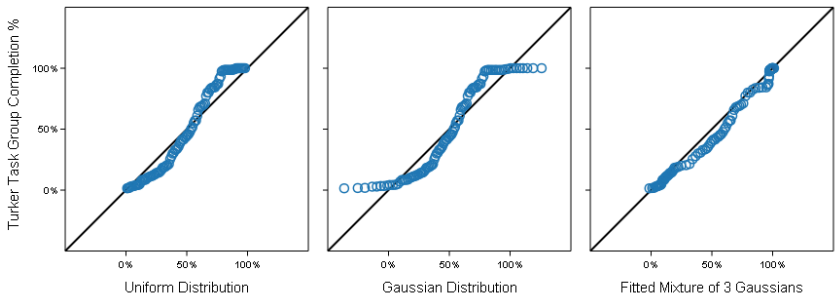


Used to compare two distributions; in this case, one actual and one theoretical.

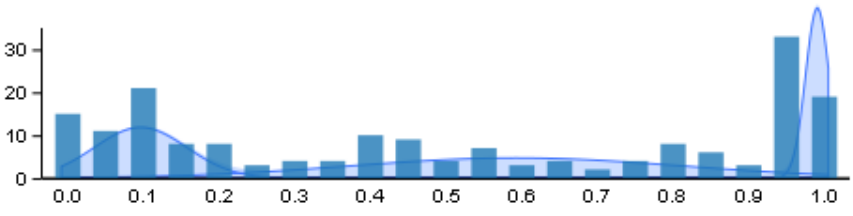
Plots the quantiles (here, the percentile values) against each other.

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?

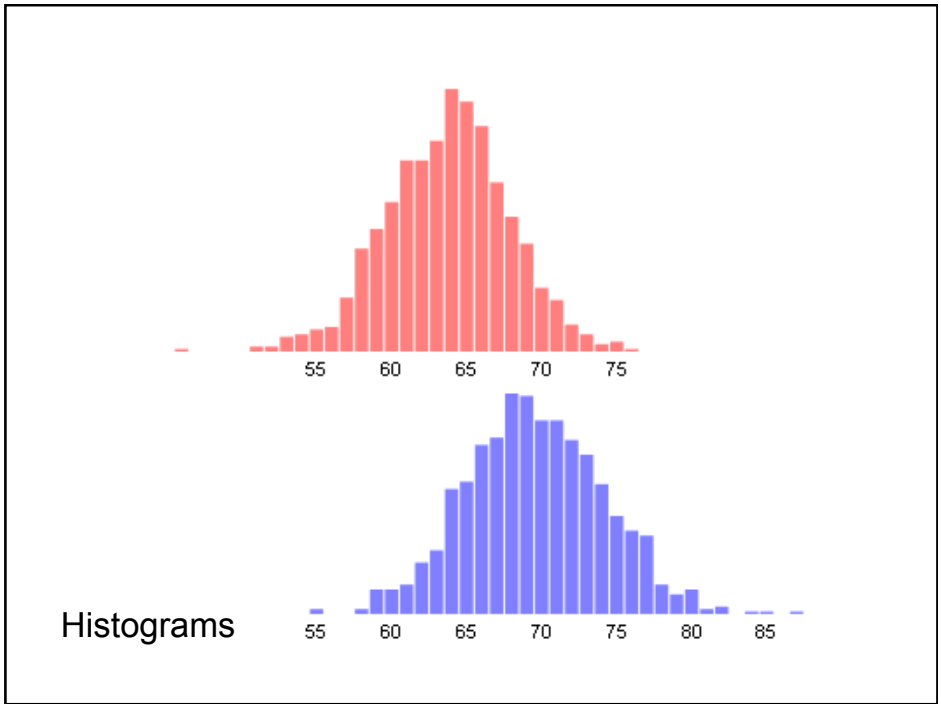
…and many others.

Example: Heights by Gender

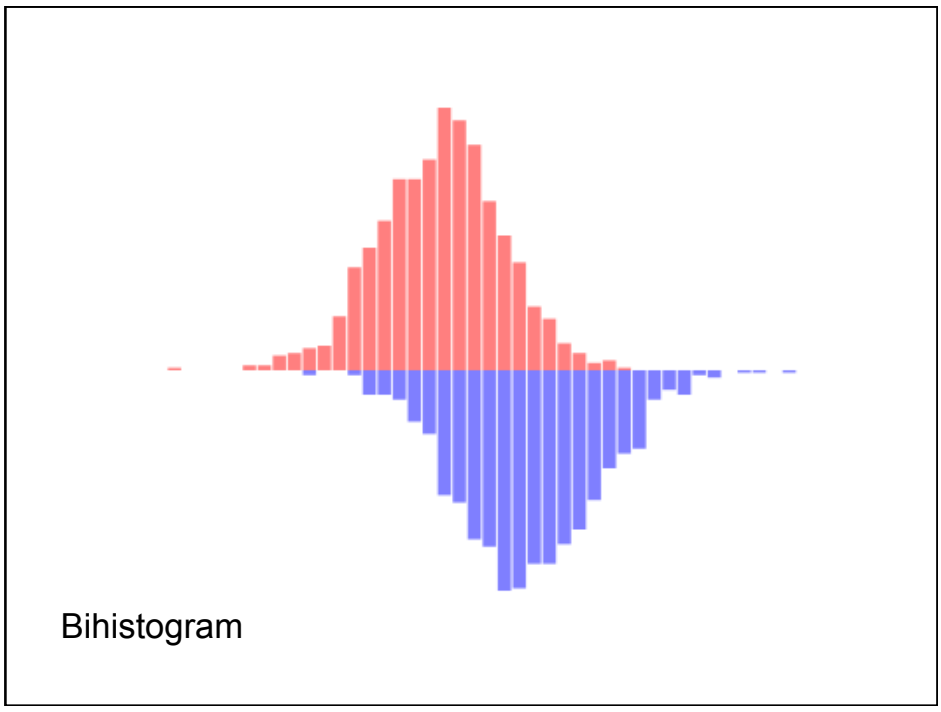
Gender	Male / Female	
Height (in)		Number
$\mu_m = 69.4$	$\sigma_m = 4.69$	$N_m = 1000$
$\mu_f = 63.8$	$\sigma_f = 4.18$	$N_f = 1000$

Is this difference in heights significant?

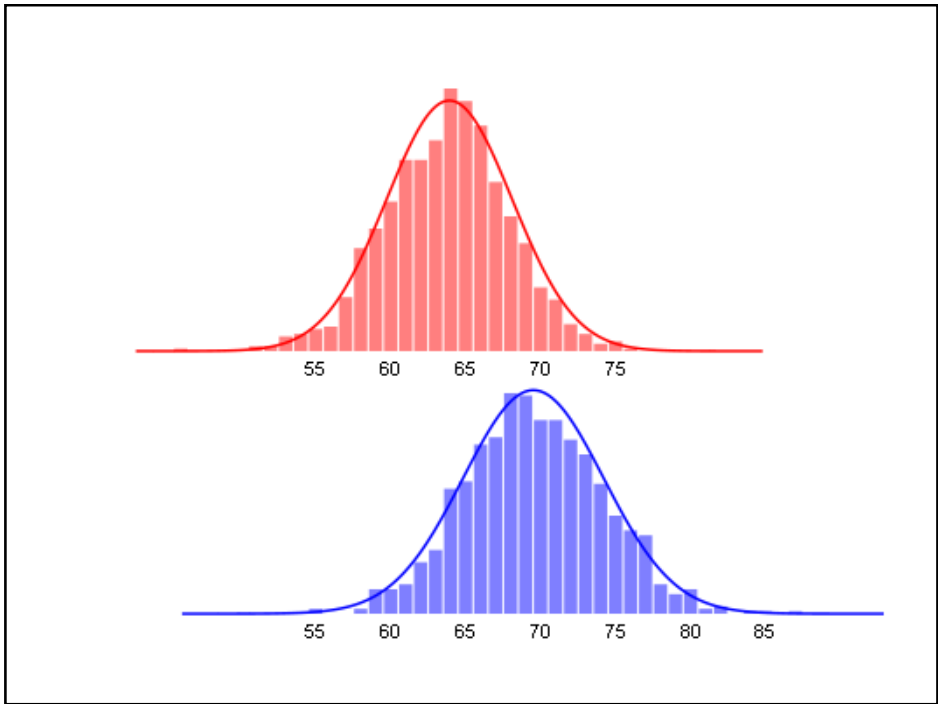
In other words: assuming no true difference, what is the prob. that our data is due to chance?

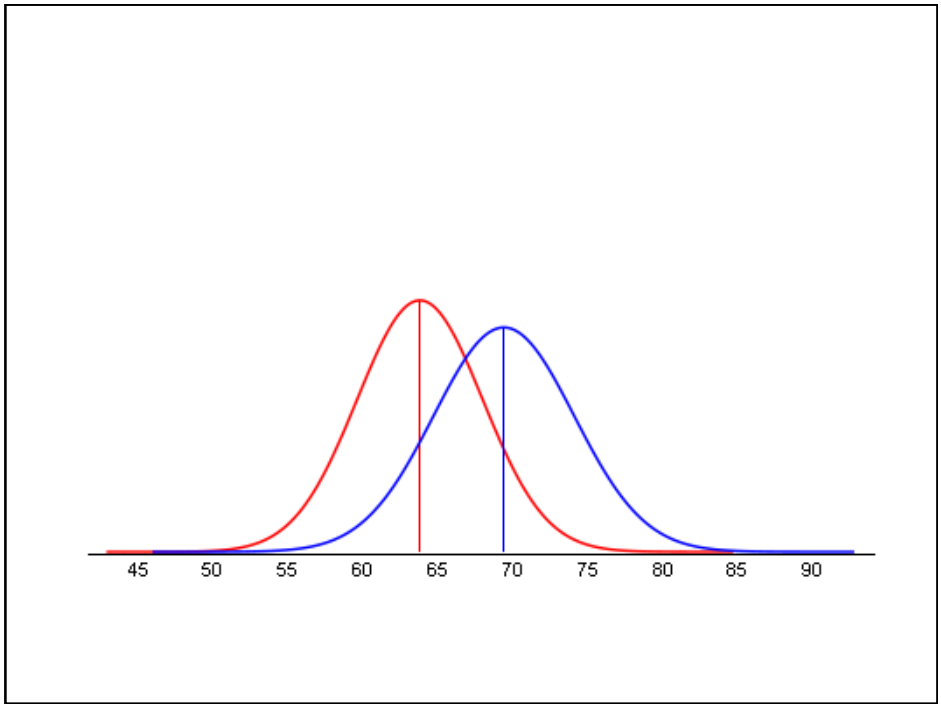


Histograms



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.

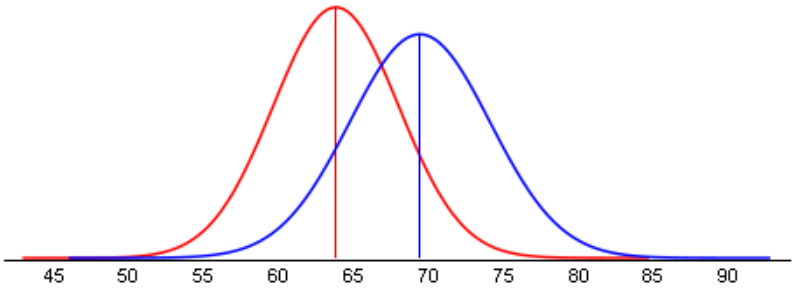
Testing Procedure

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

Compute test statistic

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

$\mu_m - \mu_f = 5.6$



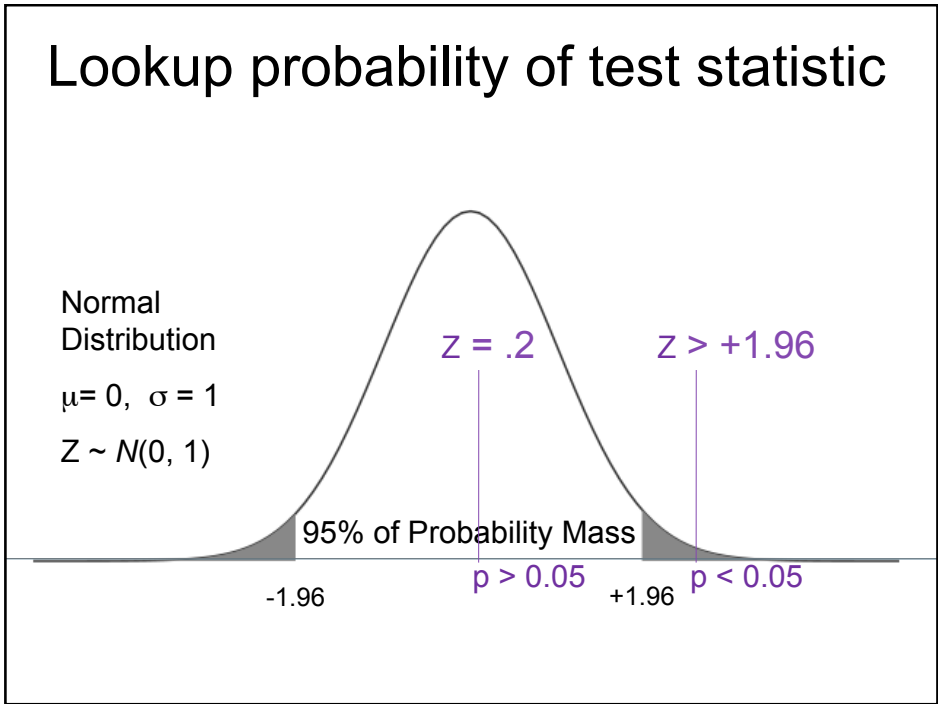
Testing Procedure

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.

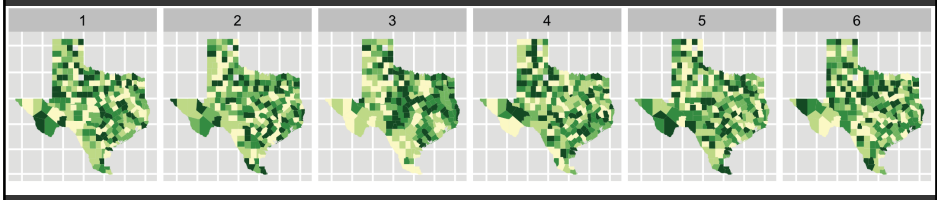
Does not assume a distribution. Typically works on rank orders.

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

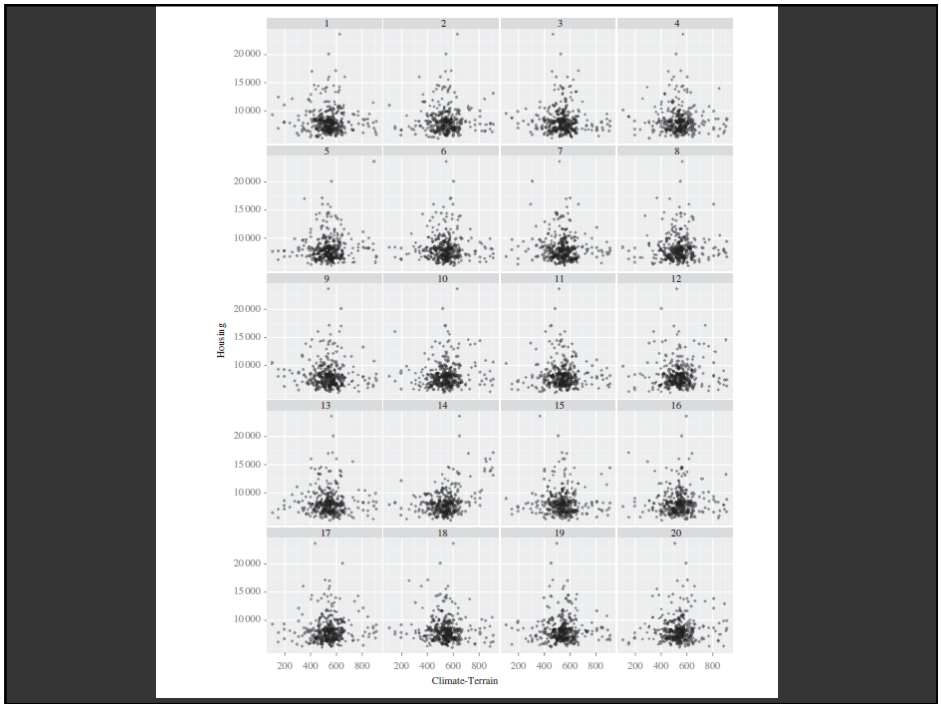
Buja Cook, Hoffman, Wickham et al.

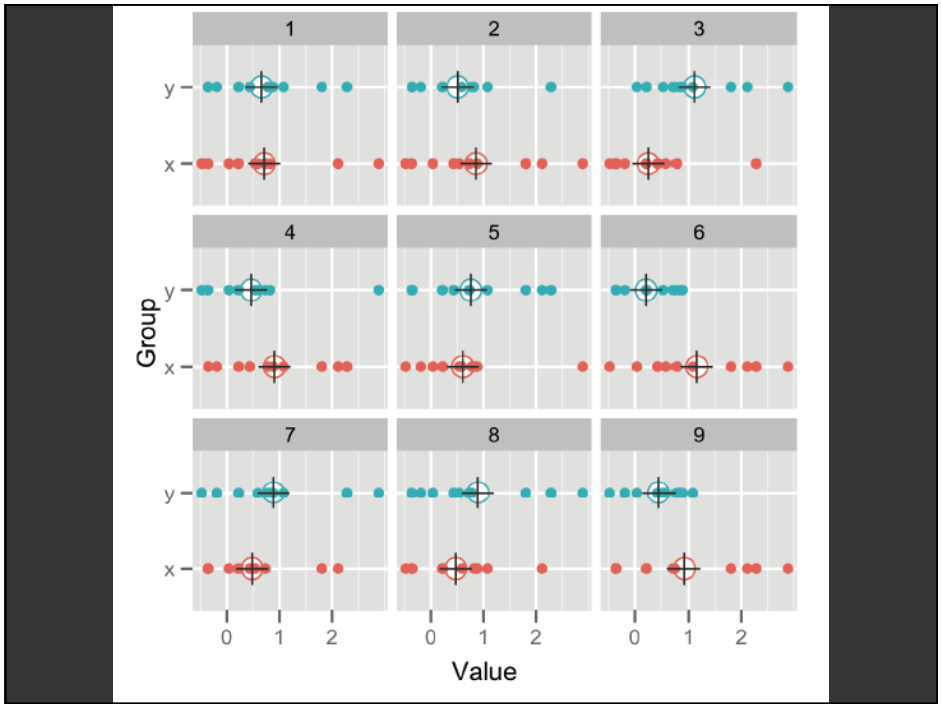


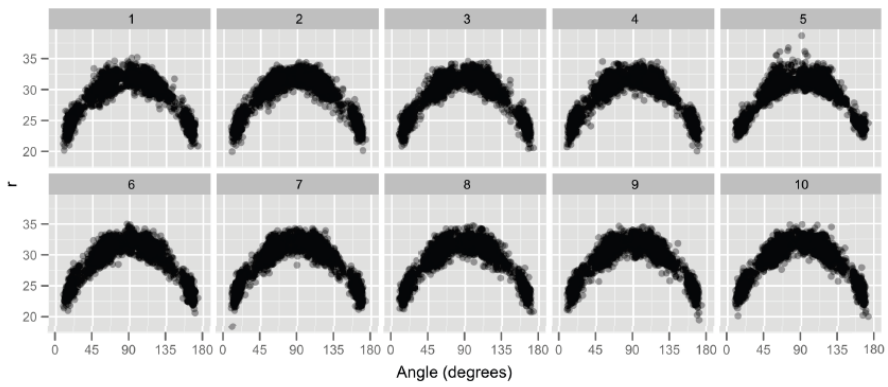
Choropleth maps of cancer deaths in Texas.

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.

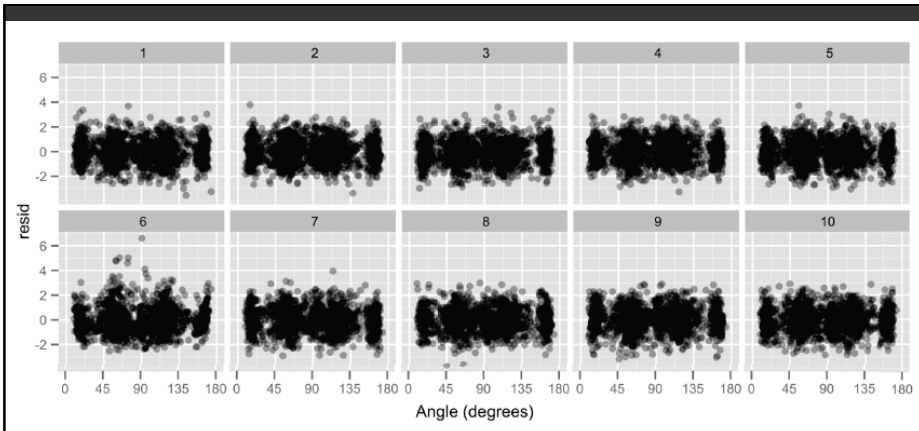

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
20000
15000
10000
Housing
200
400
600
800
Climate-Terrain


1
2
3
4
5
6
7
8
9
y
x
Group
0
1
2
Value



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