Lecture 1: Laying the Foundations + Terminology

Chapters 1.1-1.2

Goals for Today

- Go over the syllabus
- Show some fun examples
- Discuss how to evaluate the efficacy of a treatment
- Describe the different kinds of variables we'll consider

What is statistics?

(Direct from text) The general scientific process of investigation can be summed up as follows:

What is statistics?

(Direct from text) The general scientific process of investigation can be summed up as follows:

1. Identify the scientific question or problem

What is statistics?

(Direct from text) The general scientific process of investigation can be summed up as follows:

1. Identify the scientific question or problem
2. Collect relevant data on the topic

What is statistics?

(Direct from text) The general scientific process of investigation can be summed up as follows:

1. Identify the scientific question or problem
2. Collect relevant data on the topic
3. Analyze the data

What is statistics?

(Direct from text) The general scientific process of investigation can be summed up as follows:

1. Identify the scientific question or problem
2. Collect relevant data on the topic
3. Analyze the data
4. Form a conclusion

What is statistics?

(Direct from text) The general scientific process of investigation can be summed up as follows:

1. Identify the scientific question or problem
2. Collect relevant data on the topic
3. Analyze the data
4. Form a conclusion and communicate it

What is statistics?

(Direct from text) The general scientific process of investigation can be summed up as follows:

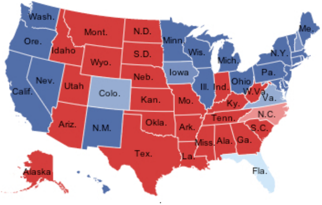
1. Identify the scientific question or problem
2. Collect relevant data on the topic
3. Analyze the data
4. Form a conclusion and communicate it

Statistics concerns itself with points 2 through 4.

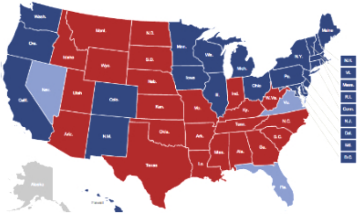
Your Majors

Biology	11	Economics	5
History	4	Environmental Studies	3
Mathematics	3	Psychology	3
Biochem and Molecular Biology	2	Chemistry	2
International Policy Studies	2	Linguistics	2
Undecided	2	Anthropology	1
Economics/Mathematics	1	Environmental Studies-Hist	1
Environmental Studies-Pol Sci	1	Physics	1
Sociology	1		

Example: 2012 Election - Nate Silver's Predictions vs Actual Results



Nate Silver's Map



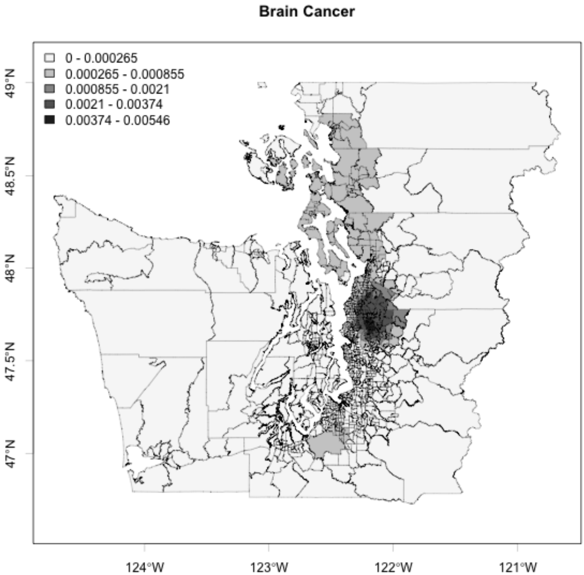
The Actual Map

Example: Brain & Breast Cancer in Western Washington

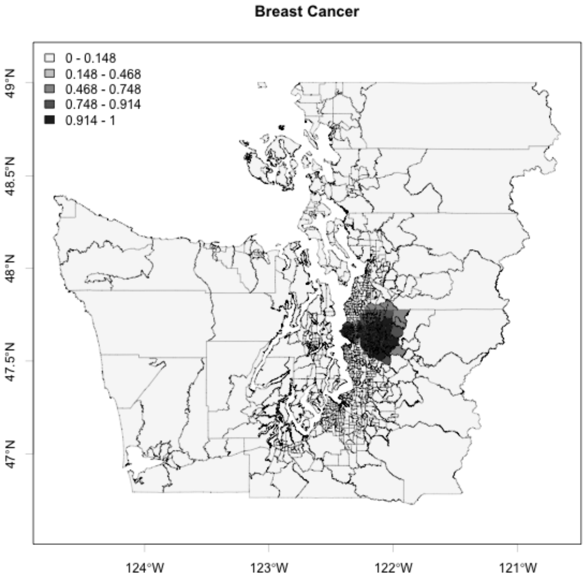
My PhD dissertation involved detecting cancer "clusters": areas of residual spatial variation of disease risk.

We modeled the (Bayesian) probability of cluster membership for each of the $n = 887$ census tracts in Western Washington in 2000, using cancer data from 1995–2005, controlling for age, race, and gender.

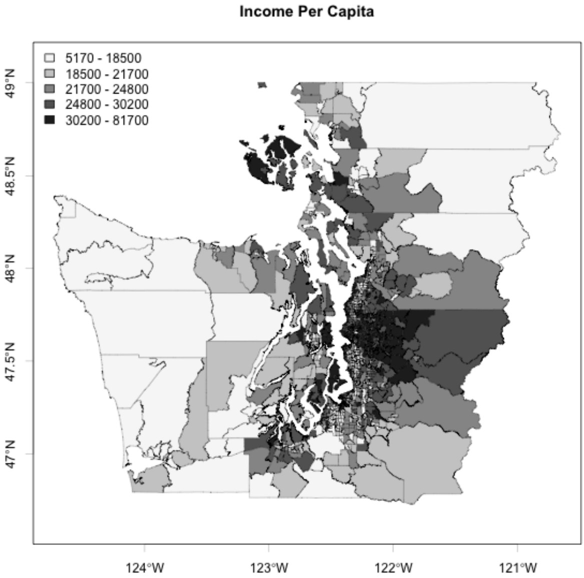
Brain Cancer Controlling for Age, Race, & Gender



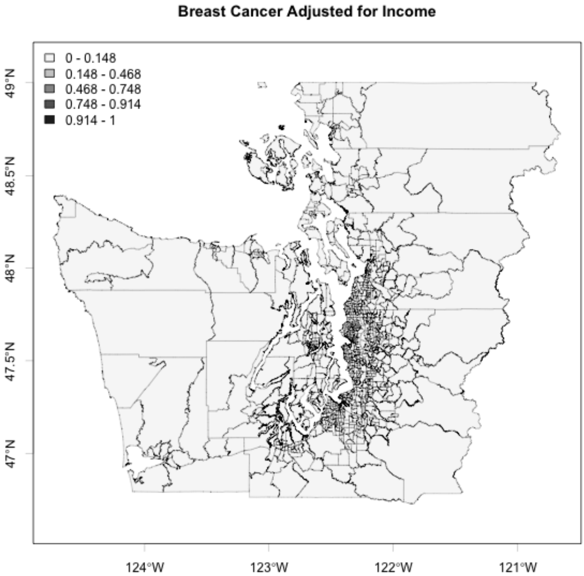
Breast Cancer Controlling for Age, Race, & Gender



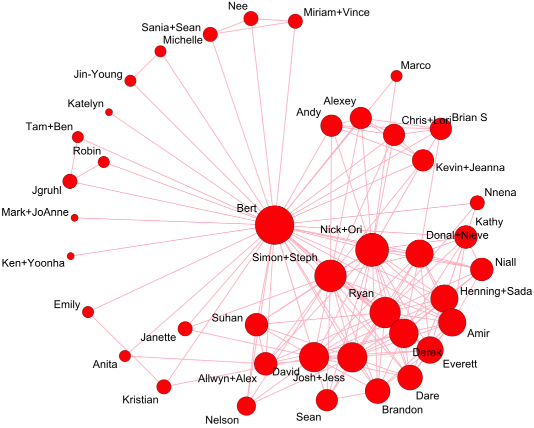
Income per Capita Quintiles



Breast Cancer Adjusted for Income as Well



Example: Social Network Display of a Recent Party I Had



Say we want answer the following questions:

Say we want answer the following questions:

- Does a new kind of cognitive therapy alter levels of depression in patients?

Say we want answer the following questions:

- Does a new kind of cognitive therapy alter levels of depression in patients?
- Or you question the effectiveness of antioxidants in preventing cancer.

Say we want answer the following questions:

- Does a new kind of cognitive therapy alter levels of depression in patients?
- Or you question the effectiveness of antioxidants in preventing cancer.
- Will reassuring potential new users to a gambling website that we won't spam them increase the sign-up rate?

Evaluating the efficacy of a ‘treatment’

In all the above cases, you are questioning the efficacy of a treatment/intervention. One way to evaluate the efficacy is via an experiment where you define

Evaluating the efficacy of a ‘treatment’

In all the above cases, you are questioning the efficacy of a treatment/intervention. One way to evaluate the efficacy is via an experiment where you define

- A control group: the “business as usual” baseline group

Evaluating the efficacy of a ‘treatment’

In all the above cases, you are questioning the efficacy of a treatment/intervention. One way to evaluate the efficacy is via an experiment where you define

- A control group: the “business as usual” baseline group
- A treatment group: the group that receives/is subject to the treatment/intervention

Evaluating the efficacy of a ‘treatment’

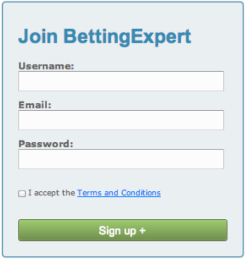
In all the above cases, you are questioning the efficacy of a treatment/intervention. One way to evaluate the efficacy is via an experiment where you define

- A control group: the “business as usual” baseline group
- A treatment group: the group that receives/is subject to the treatment/intervention

and make comparisons.

Website Experiments

Control:



Join BettingExpert

Username:

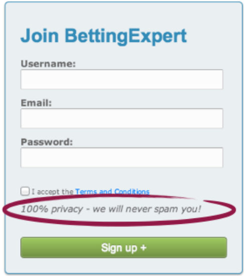
Email:

Password:

□ I accept the Terms and Conditions

Sign up +

Treatment:



Join BettingExpert

Username:

Email:

Password:

□ I accept the Terms and Conditions

100% privacy - we will never spam you!

Sign up +

Example of a treatment vs control

Two other examples in the media of late

- Facebook's tinkering with user's emotions (link)
- OkCupid's admission that they experiment on human beings (link)

Variables

A variable is a description of any characteristic whose value may change from one unit in the population to the next:

Data

At its simplest, data are presented in a data table or matrix where (almost always) each

Data

At its simplest, data are presented in a data table or matrix where (almost always) each

- row corresponds to cases or units of observation/analysis

Data

At its simplest, data are presented in a data table or matrix where (almost always) each

- row corresponds to cases or units of observation/analysis
- column represents the variables corresponding to a particular observation

Data

At its simplest, data are presented in a data table or matrix where (almost always) each

- row corresponds to cases or units of observation/analysis
- column represents the variables corresponding to a particular observation

It is almost always the case that

Data

At its simplest, data are presented in a data table or matrix where (almost always) each

- row corresponds to cases or units of observation/analysis
- column represents the variables corresponding to a particular observation

It is almost always the case that

- n is the number of observations

Data

At its simplest, data are presented in a data table or matrix where (almost always) each

- row corresponds to cases or units of observation/analysis
- column represents the variables corresponding to a particular observation

It is almost always the case that

- n is the number of observations
- p is the number of variables

Data Summaries

Consider the variable "federal spending per capita" in each of the 3,143 counties in the US. One can hardly digest this:

```
   [1]   6.068095   6.139862   8.752158   7.122016   5.130910   9.973062   9.311835  15.439218
   [9]   8.613707   7.104621   6.324061  10.640378   9.781442   8.982702   6.840035  20.330684
  [17]   9.687698  11.080738   7.839761   9.461856   9.650295   7.760627  25.774791  13.948106
   ...
[3121]   7.520731  10.246400   3.106800  17.679572   4.824044   7.247212   8.484211   8.794626
[3129]   9.829593   8.100945  17.090715   4.855849   6.621378  22.587359  10.813260  11.422522
[3137]   9.580265   4.368986   5.062138   6.236968   4.549105   8.713817   6.694784
```

Data Summaries

We can't interpret all the data at once; we need to boil it down via summary statistics, single numbers summarizing a large amount of data.

Data Summaries

We can't interpret all the data at once; we need to boil it down via summary statistics, single numbers summarizing a large amount of data.

Using the `summary()` command in `R`:

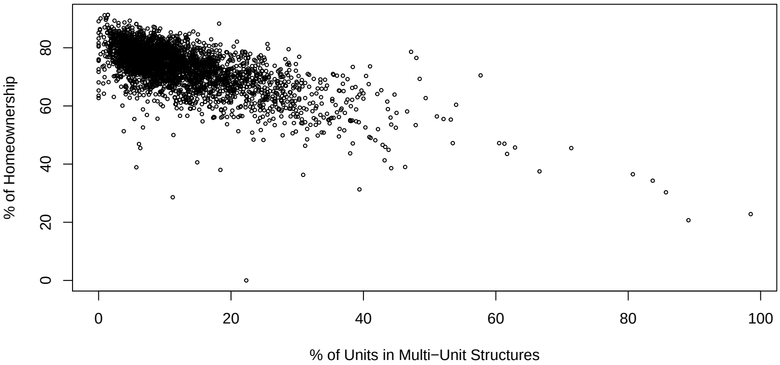
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  0.000   6.964   8.669   9.991  10.860 204.600       4
```

Relationships between variables

We can best display the relationship between two variables using a scatterplot AKA bivariate plot:

Relationships between variables

We can best display the relationship between two variables using a scatterplot AKA bivariate plot:



Relationships between variables

Almost always we are interested in the relationship between two or more variables.

Relationships between variables

Almost always we are interested in the relationship between two or more variables.

A pair of variables are either related in some way (associated) or not (independent). No pair of variables are both associated and independent.

Relationships between variables

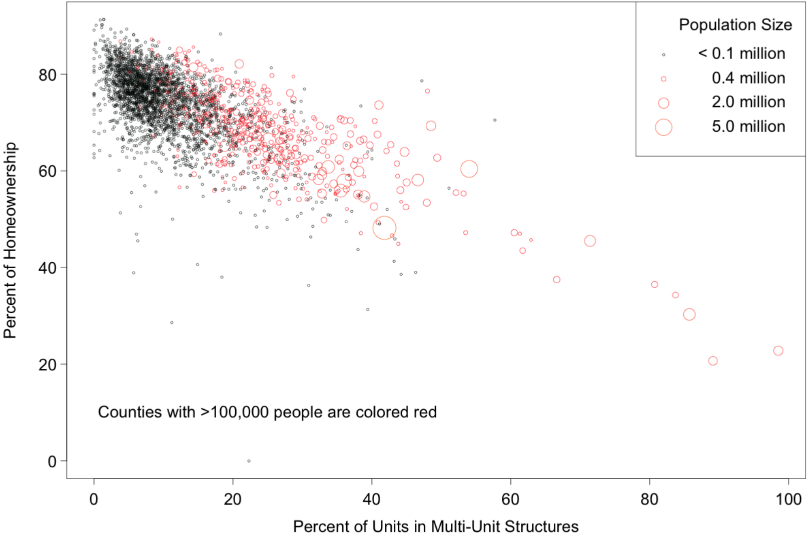
Almost always we are interested in the relationship between two or more variables.

A pair of variables are either related in some way (associated) or not (independent). No pair of variables are both associated and independent.

We can have either a negative association (as the value of one variable increases, the other decreases) or a positive association.

Relationships between variables

We can consider a third variable in the previous plot.


Population Size
< 0.1 million
0.4 million
2.0 million
5.0 million
Counties with >100,000 people are colored red
Percent of Homeownership
Percent of Units in Multi-Unit Structures