To Do List

Do Problems 1-20 in Chapter 1

Assignment 1 is due Friday September 23 at NOON.

Tutorial Test 1 is on Wednesday September 28 – see detailed instructions posted on Learn

Today's Class: Finish Graphical Summaries

1) Histograms
2) Empirical Cumulative Distribution Function (e.c.d.f.)
3) Boxplots
4) Run Charts
5) Scatterplots (Bivariate Data)
6) Bar Charts, Pie Charts

Clicker Question 1

Which one of the following statements is true for you?

A: I am a CS student and my hometown is in Canada.

B: I am not a CS student and my hometown is in Canada.

C: I am a CS student and my hometown is not in Canada.

D: I am not a CS student and my hometown I not in Canada.

CS/Hometown Data

What type of data are we collecting in this example?

Bivariate Categorical Data

PROGRAM/ HOMETOWN	Canadian Hometown	Non-Canadian Home town	Total
Computer Science	35	43	78
Non-Computer Science	18	69	87
Total	53	112	165

Does there appear to be a relationship between hometown and program?

Can't use a scatterplot!

Bivariate Categorical Data

Proportion of CS students with Canadian hometown = 35/78 = 0.448.

Proportion of Non-CS students with Canadian hometown = 18/87 = 0.206.

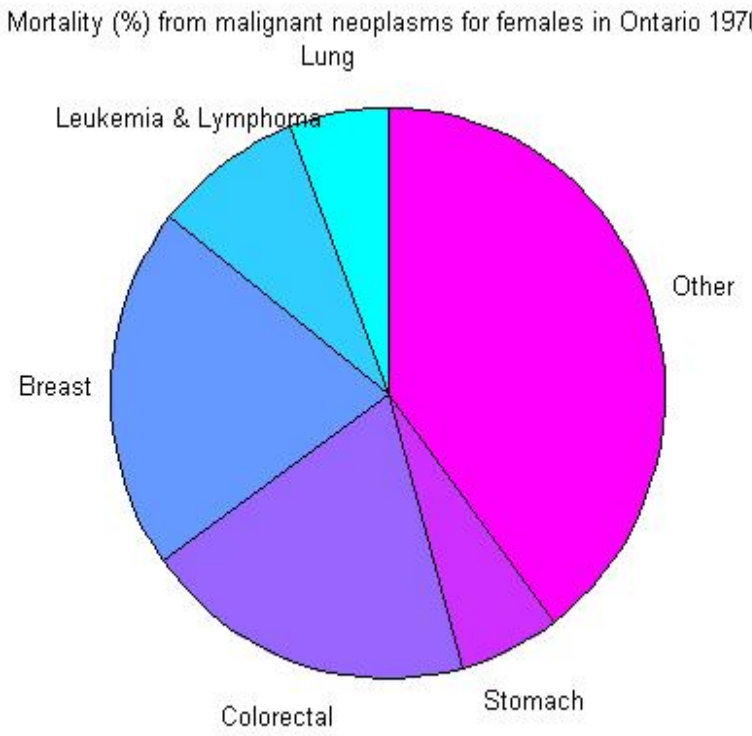
The relative risk of a Canadian hometown among CS students as compared to non-CS students = (35/78)/ (18/87) = 2.17.

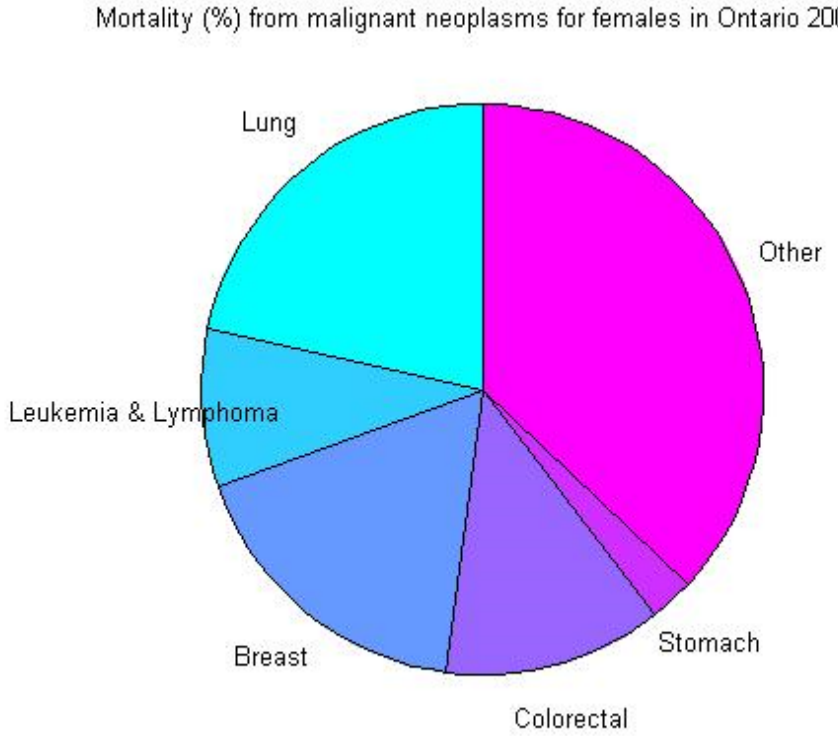
Example: Mortality (%) from malignant neoplasms for females in Ontario, 1970 and 2000.

Females	1970	2000
Lung	5.8	21.6
Leukemia & Lymphoma	8.6	9.0
Breast	20.5	17.4
Colorectal	19.5	12.6
Stomach	5.8	2.5
Other	39.8	37.0

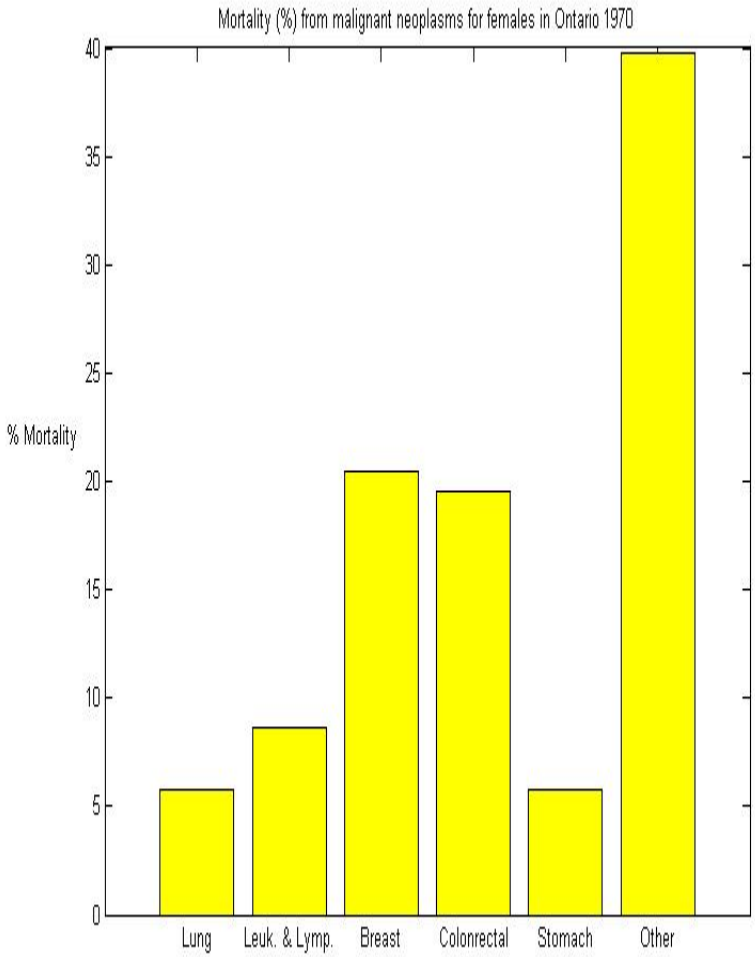
How can we best display these data graphically?

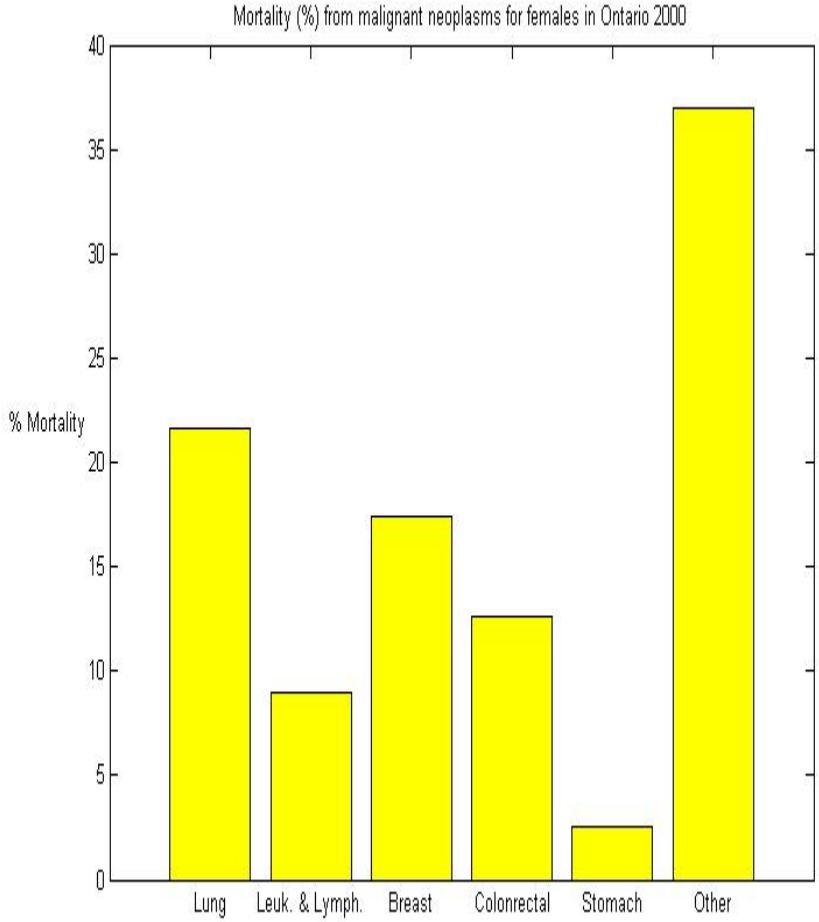
Pie Charts

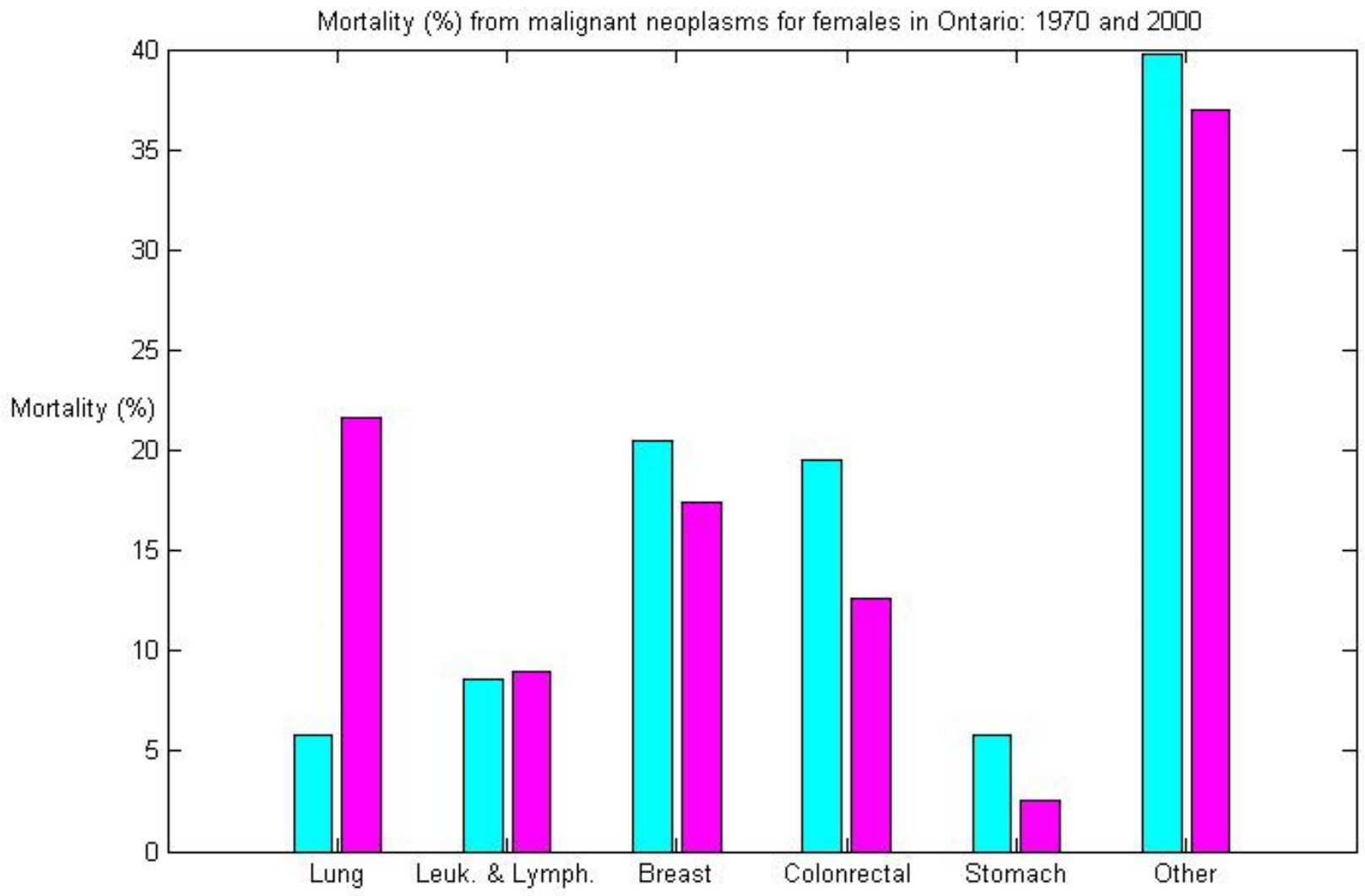

Mortality (%) from malignant neoplasms for females in Ontario 1970
Lung
Leukemia & Lymphoma
Other
Breast
Colorectal
Stomach


Mortality (%) from malignant neoplasms for females in Ontario 2000
Lung
Other
Leukemia & Lymphoma
Stomach
Breast
Colorectal

Bar Charts


Mortality (%) from malignant neoplasms for females in Ontario 1970
% Mortality
0
5
10
15
20
25
30
35
40
Lung
Leuk. & Lymp.
Breast
Colonrectal
Stomach
Other


Mortality (%) from malignant neoplasms for females in Ontario 2000
% Mortality
0
5
10
15
20
25
30
35
40
Lung
Leuk. & Lymph.
Breast
Colonrectal
Stomach
Other

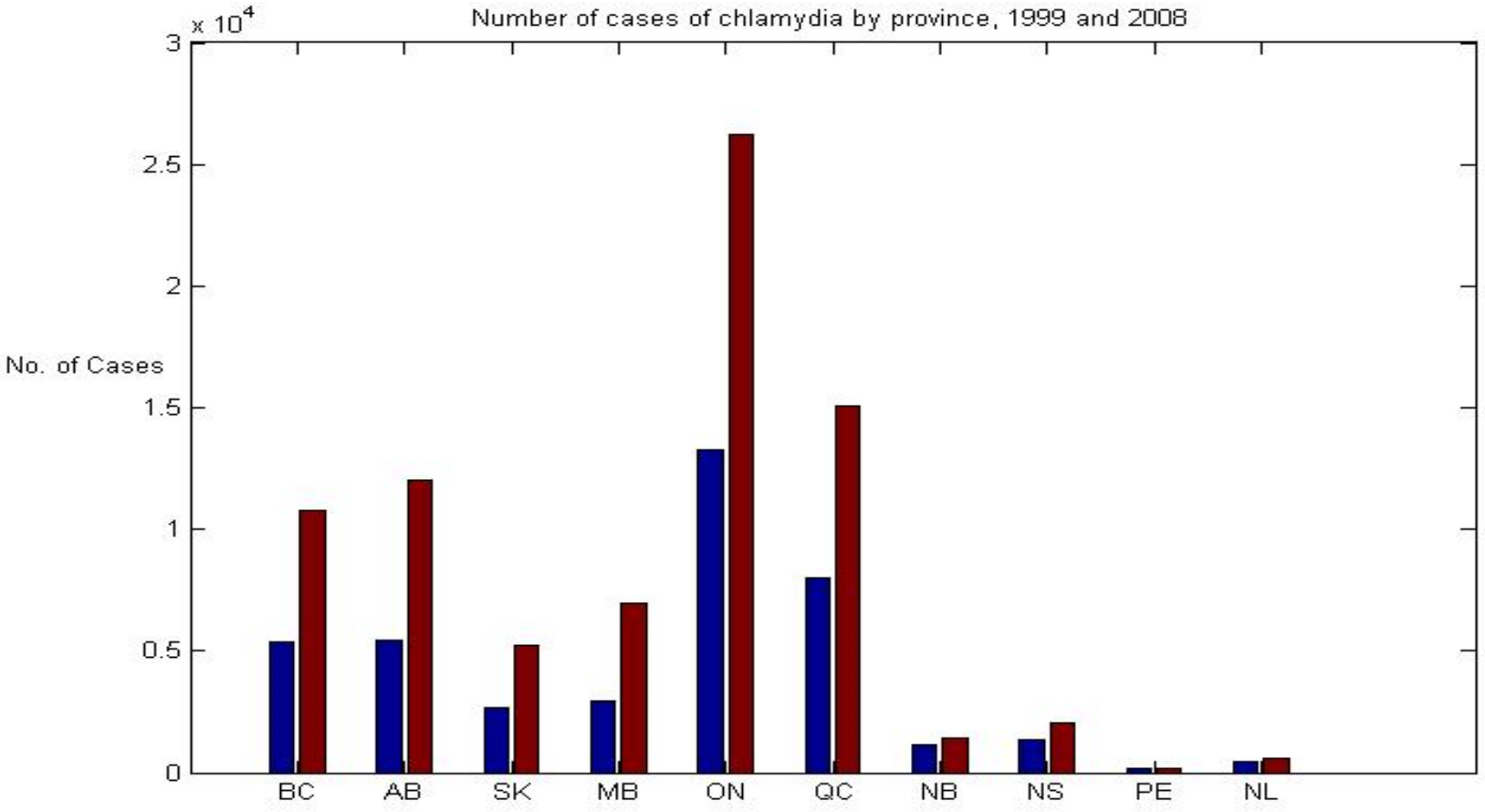

Mortality (%) from malignant neoplasms for females in Ontario: 1970 and 2000
Mortality (%)
0
5
10
15
20
25
30
35
40
Lung
Leuk. & Lymph.
Breast
Colonrectal
Stomach
Other

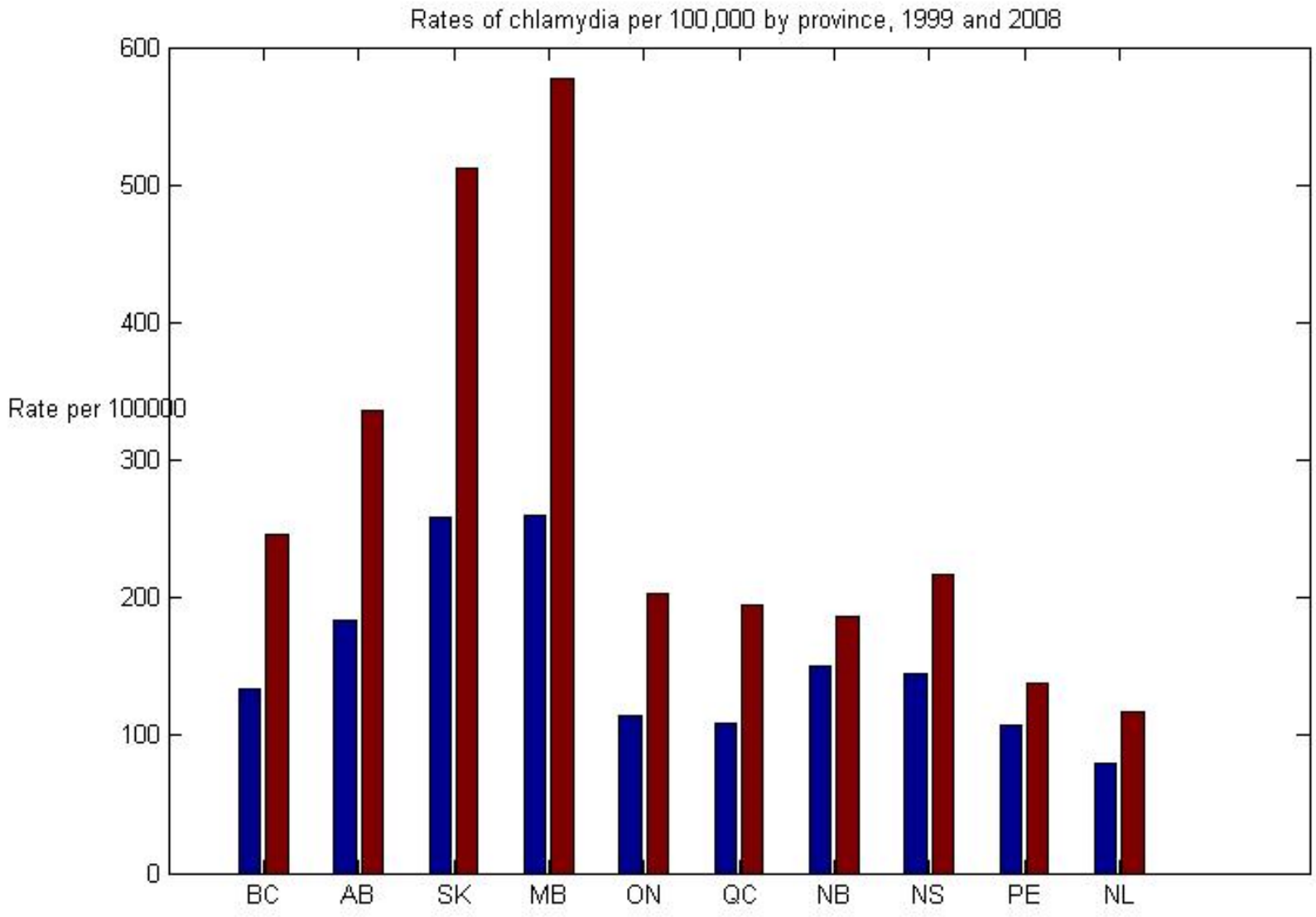
Example: Cases and Rates of Chlamydia

Table 1: Reported Cases and Rates[1] of Chlamydia by Province/Territory, 1999 and 2008, Canada

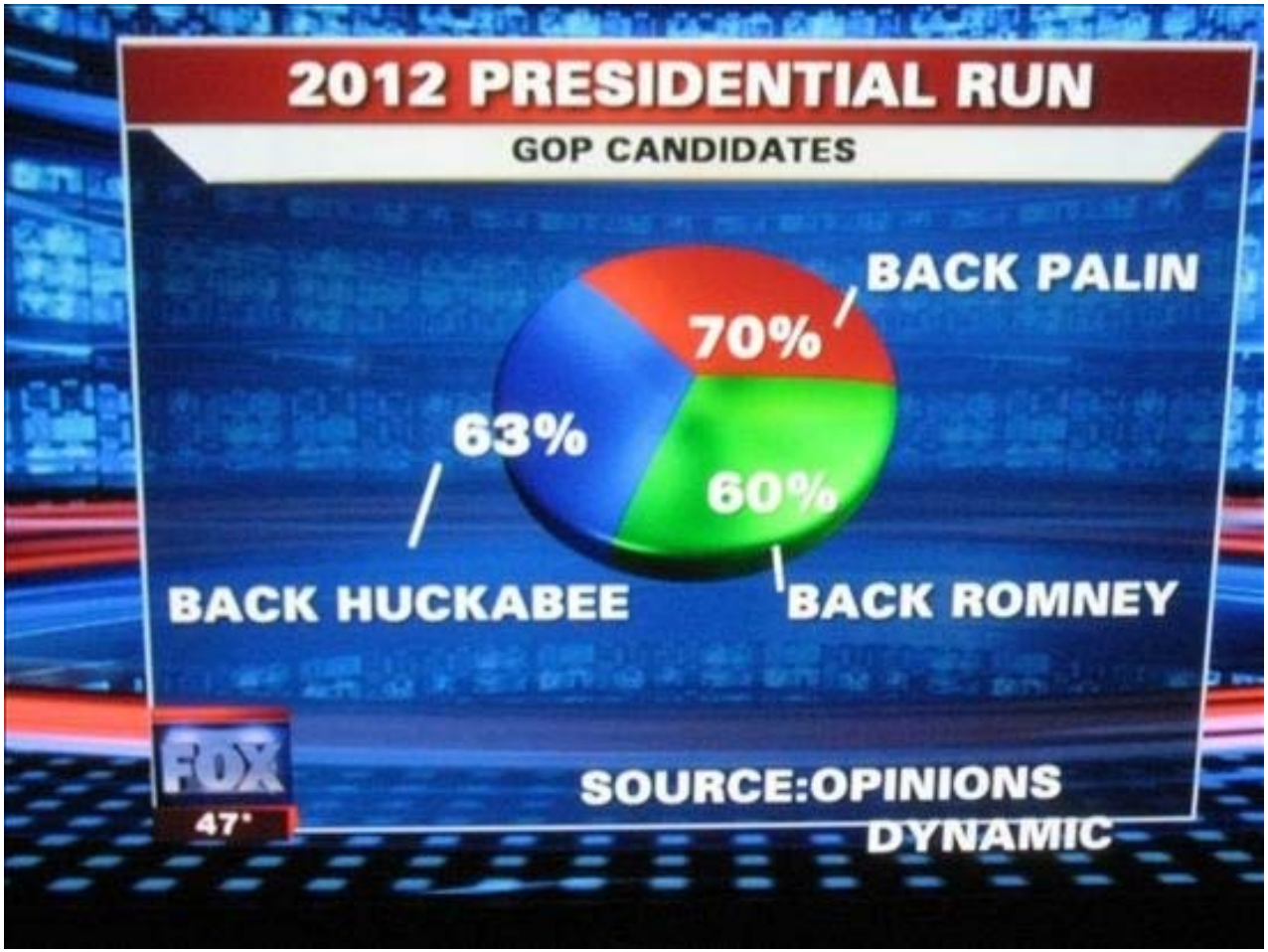
Jurisdiction	Number of Cases		Rates per 100,000[3]		Rate Change (%)
	1999	**2008**	**1999**	**2008**	**1999–2008**
Canada	*42,141*	*82,919*	*138.2*	*248.9*	80.2
BC	5,402	10,766	134.1	245.7	83.2
AB	5,416	12,047	**183.0**	**336.0**	83.6
SK	2,656	5,203	**259.0**	**512.1**	97.7
MB	2,967	6,965	**259.7**	**576.6**	122.0
ON	13,256	26,245	115.0	203.0	76.5
QC	7,968	15,043	108.4	194.1	79.0
NB	1,136	1,389	**150.6**	185.9	23.4
NS	1,364	2,033	**145.1**	216.7	49.3
PE	148	193	107.6	138.0	28.3
NL	433	596	80.1	117.3	46.5
YT	176	232	**567.0**	**700.0**	23.4
NT	1,219	870	**1,796.4**	**2,010.0**	N/A
NU[2]	N/A	1,337	N/A	**4,251.5**	N/A

Number of Cases by Province

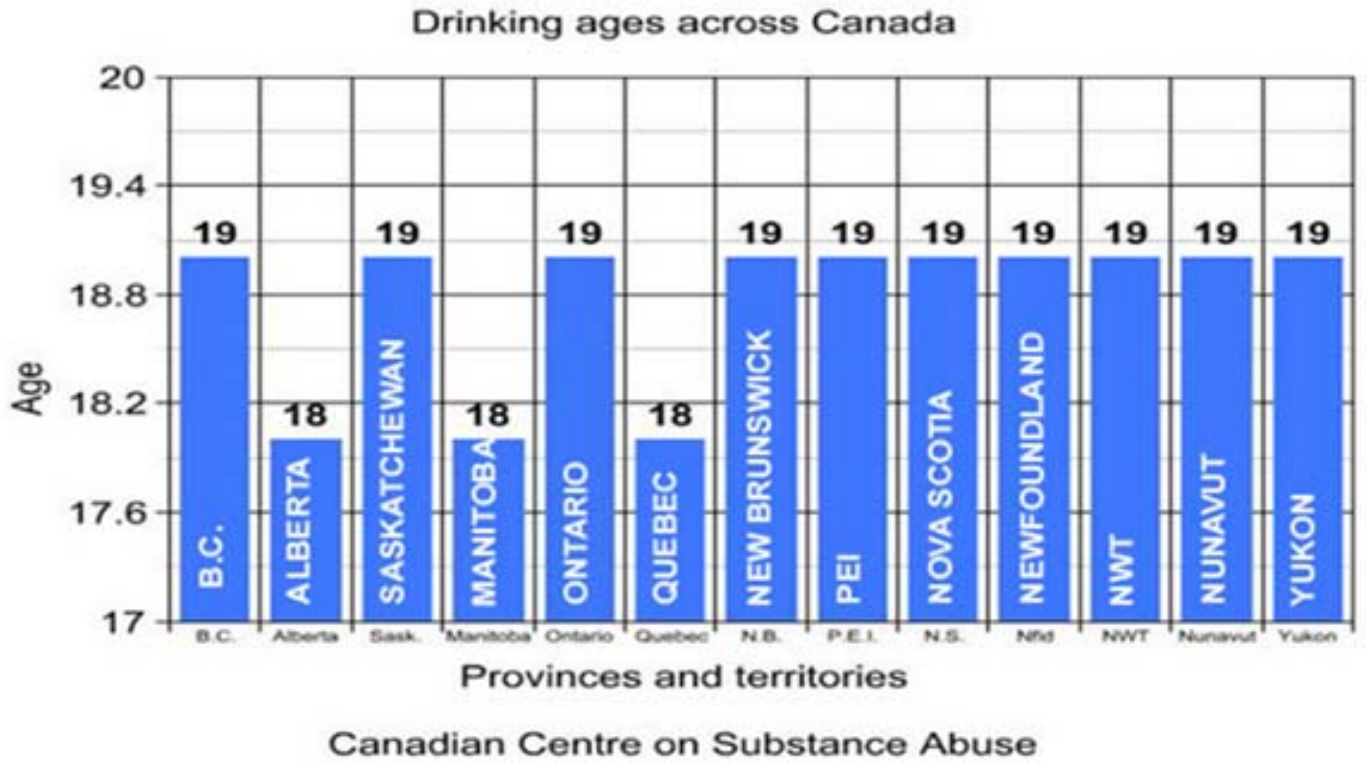



Rates of chlamydia per 100,000 by province, 1999 and 2008
Rate per 100000
600
500
400
300
200
100
0
BC
AB
SK
MB
ON
QC
NB
NS
PE
NL

Not all graphical representations are good



One of my favourites:



You have to be 19 in Saskatchewan to have a drink, while in Alberta and Manitoba, the drinking age 18. (CBC)

Remainder of Today's Class and Friday's Class:

1) Descriptive Statistics and Statistical Inference (Inductive versus Deductive Reasoning)
2) Types of Statistical Problems:
   (i) Estimation Problems
   (ii) Hypothesis Testing Problems
   (ii) Prediction Problems
3) Statistical Models – Why and How to Choose
4) Families of Statistical Models
5) Unknown Parameters in a Statistical Model
6) Estimates of Unknown Parameters in Statistical Models

Data Analysis and Statistical Inference

We collect data to increase our knowledge or as a basis for making decisions.

Proper analysis of the data is crucial.

Two broad aspects of the analysis and interpretation of data are:

1) Descriptive Statistics
2) Statistical Inference

Descriptive Statistics

Descriptive statistics is the portrayal of the data, or parts of it, in numerical and graphical ways to show features of interest.

The numerical and graphical summaries we have examined are all examples of descriptive statistics.

A more complex example is "knowledge discovery" and "data mining" (KDD) which refers to exploratory data analysis of databases. The emphasis is on descriptive statistics for very large data bases. The goal is to find interesting patterns and relationships.

Statistical Inference

When the data obtained in the study of a population or process are used to draw general conclusions about the population or process itself we call this process statistical inference.

Inductive versus Deductive Reasoning

When we reason from the specific (the observed data on a sample of units) to the general (the target population or process) we call this inductive reasoning. Therefore statistical inference is a form of inductive reasoning.

Note: in mathematics when we use general results (axioms) to prove theorems this is called deductive reasoning.

Methods of Statistical Inference

The methods of statistical inference that we will look at in STAT 231 will be used to examine 3 main types of problems:

1) **Estimation problems**

2) **Hypothesis testing problems**

3) **Prediction problems**

Estimation Problems

In an **estimation problem** we are interested in estimating one or more attributes of a process or population.

Examples:

1) Estimate the proportion of Ontario residents aged 14 - 20 who smoke.
2) Estimate the distribution of survival times for certain types of AIDS patients.
3) "Fit" or select a probability distribution for a process.

Hypothesis Testing Problems

In a **hypothesis testing problem** we use the data to assess the truth of some question or hypothesis.

Examples:

1) Is it true that in the 14-20 age group a higher proportion of females than males smoke?

2) Will the use of a new treatment with unpleasant side effects increase the average survival time of AIDS patients by at least 20 percent?

Prediction Problems

In a **prediction problem** we use the data to predict a future value of a variate for a unit to be selected from the population or process.

Examples:

1) Given the past performance of a stock and other data, predict the value of the stock at some point in the future.

2) Based on the results of a clinical trial predict how much an individual's blood pressure would drop for a given dosage of a drug.

Statistical Analyses

See other examples in Section 1.5 of the Course Notes.

We will be discussing methods of estimation, tests of hypotheses, and methods for prediction in much more detail in future lectures.

What we have done so far:

We have looked at numerical and graphical summaries of a dataset both univariate $\{y_1, y_2, \ldots, y_n\}$ and bivariate $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$.

We will use these summaries to help us chose a statistical model for the data.

A statistical model is a mathematical model that incorporates probability.

Why are Statistical Models Useful?

Statistical models can be used to describe many different processes such as:

- **the daily closing value of the Canadian dollar**
- **the number of hits on a government website describing how to apply for unemployment insurance**
- **catastrophic events such as oil spills, nuclear meltdowns, epidemics, diseases etc.**

Why are Statistical Models Useful?

Statistical models can also be used to describe:

- **the selection of units from a population or process (Labour Force Survey**
- **the measurement of variates**
- **the variability in the variate values for a population or process**

Why are Statistical Models Useful?

- Drawing conclusions from data involves some degree of uncertainty. Statistical models can be used to quantify this uncertainty.
- Often, questions of interest can be formulated in terms of parameters of the statistical model.

  e.g. Poisson(θ), θ = mean number of events

Why use a Statistical Model?

- Procedures for making decisions based on observed data can often be formulated in terms of statistical models.
- Models allow us to characterize processes and to simulate them via computer experiments.