Chapter 1: Stats Starts Here

What is statistics? A way of reasoning, along with a collection of tools and methods designed to help us understand the world. P2

Statistics is about variation p3

Things vary

- people are different
- can't see everything or measure it all
- what we *can* measure might be inaccurate.

How do we make sense of an imperfect picture of an imperfect world?

Chapter 2: Data

Individual Cases (Individuals) are the objects described by a set of data. Cases can be people, animals, things. p9

•**A variable** is any characteristic of an individual case (or an individual.) A variable can take different values on different cases.

• Example: A college's student data base.

–Individuals: students of the college.

– variables: date of birth, gender (female or male), choice of major, GPA etc.

•When you plan a statistical study or explore data from someone else's work, ask your self Why, Who, What and How (and if possible When and Where) p8

–Why? What purpose do the data have? Do we hope to answer some specific questions? Do we want to draw conclusions about individuals other than the ones actually have data for?

–Who? What cases (or individuals) do the data describe? How many individuals appear in the data?

–What? How many **variable**s do the data contain? **Exact definitions** of these variables? In what **units of measurement** is each variable recorded? Weights for example, might be recorded in pounds, in thousands of pounds or in kg.

•Some variables like gender, college major simply place individuals into categories. Others like height, gpa take numerical values with which we can do arithmetic.

- Categorical and Quantitative variables

–A **categorical variable** (or **qualitative variables**) places an individual into one of several groups or categories. These categories are sometimes called the **levels**.

The set of categories for a categorical variable is called a **nominal** scale.

Example: For the categorical variable, the mode of transportation to work we might use the nominal scale {Bus, subway, car, bicycle, walk}

–A **quantitative variable** takes numerical values for which arithmetic operations are defined.

A set of numerical values for a quantitative variable is sometimes called an **interval** scale.

For such variables we can answer questions like how much larger.

Example: Annual income of Canadians

Ordinal scale: This scale falls between nominal and interval scales. They have a natural ordering of values but undefined interval distances between the values.

Example: Social class – upper, middle or lower

Because their scale consists of a set of categories, the variables with an ordinal scale are often treated as qualitative and analyzed using methods for qualitative variables

However, in some respects, ordinal scale closely resembles interval scales.

Thus in some situations, quantitative treatment of ordinal data has benefits in some statistical methods for data analysis.

Example

Airlines monitored for safety and customer service. For each flight, carriers must report:

- flight number
- type of aircraft
- number of passengers
- how late the flight was (0=on time)
- any mechanical problems

Who, what, why, where, when, how?

- who: the individual cases (flights)
- what: the variables (as above):
  - categorical: flight number, type of aircraft, mechanical problems
  - quantitative: number of passengers, lateness (minutes).
- why: why these variables (help to assess safety and customer service)
- when: don't know
- where: worldwide
- how: from pilot's log

Identifier variables:

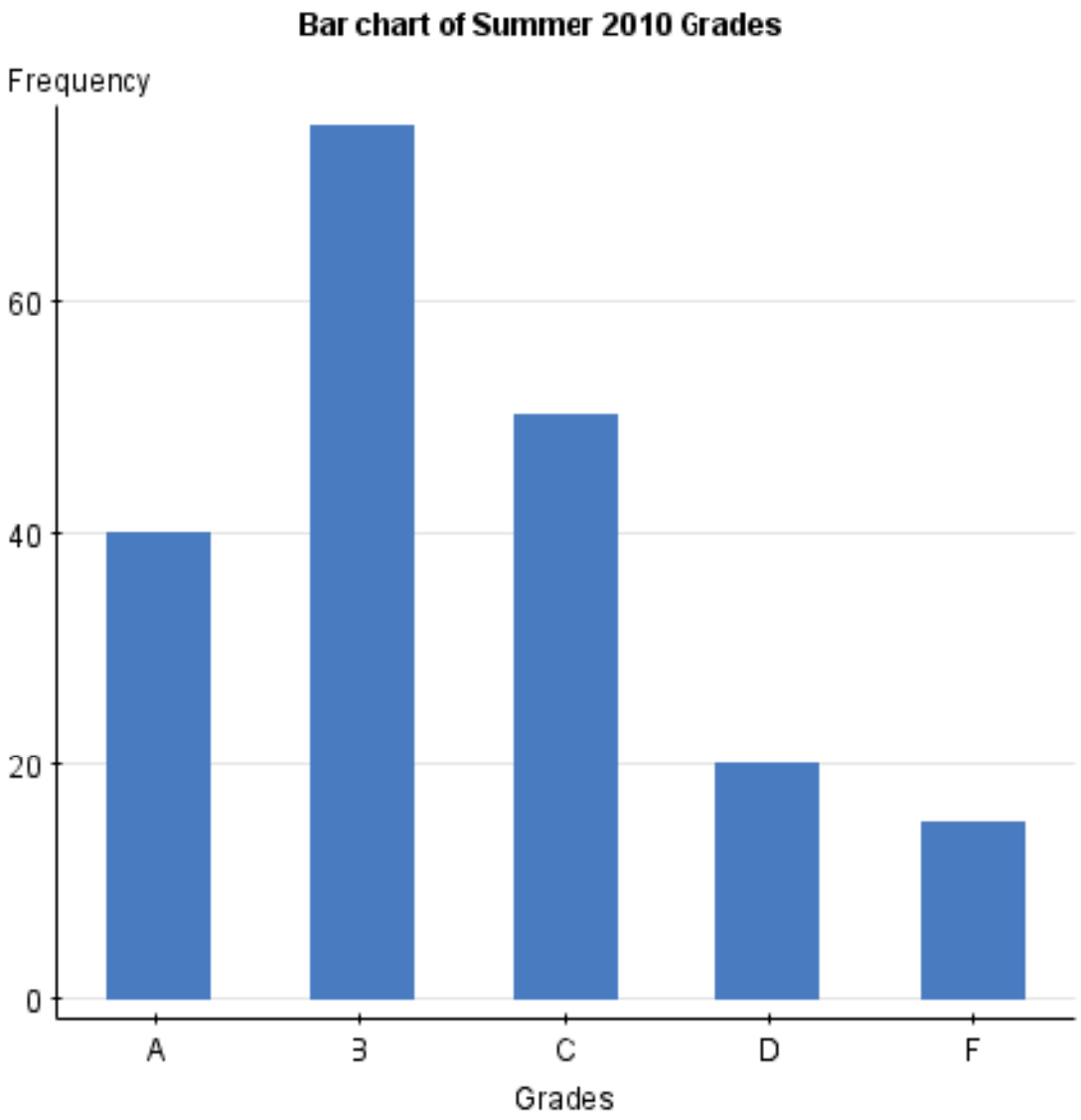
- identify individual cases
  - flight number

Chapter 3: Displaying and describing categorical data

Distribution of STAB22 grades

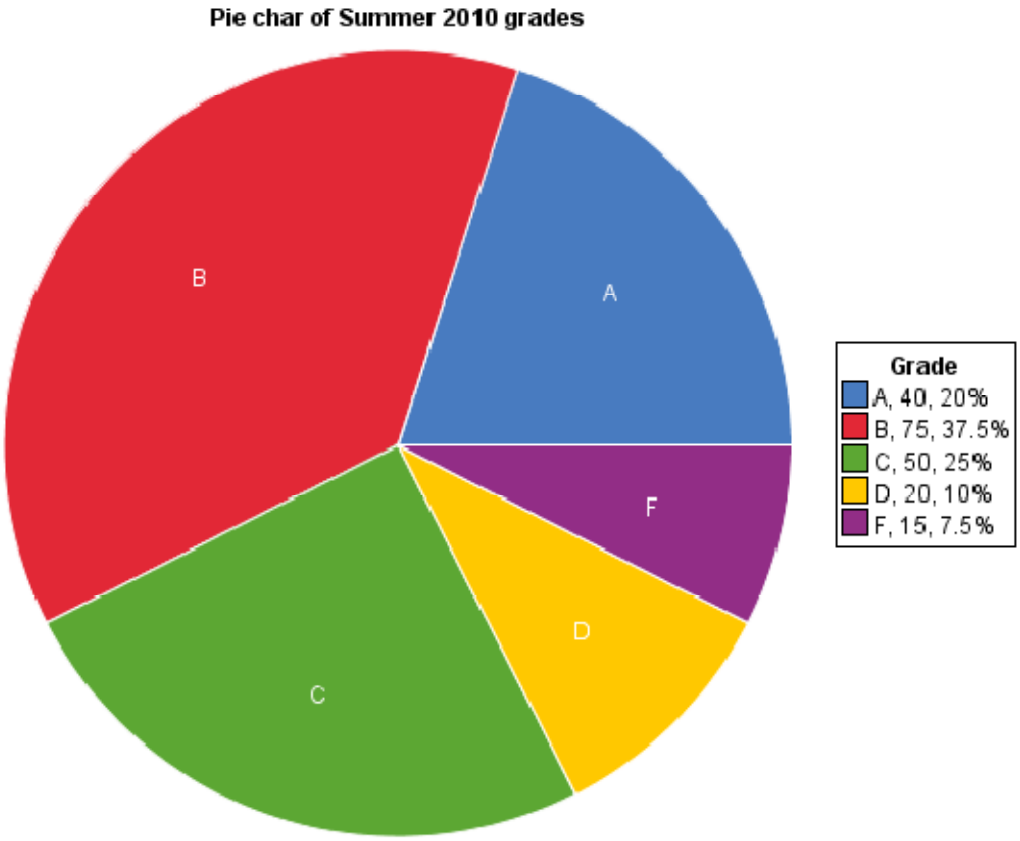
Note: These are hypothetical values.

Grade	Summer 2010	Summer 2011
A	40	45
B	75	70
C	50	58
D	20	19
F	15	8
Total	200	200



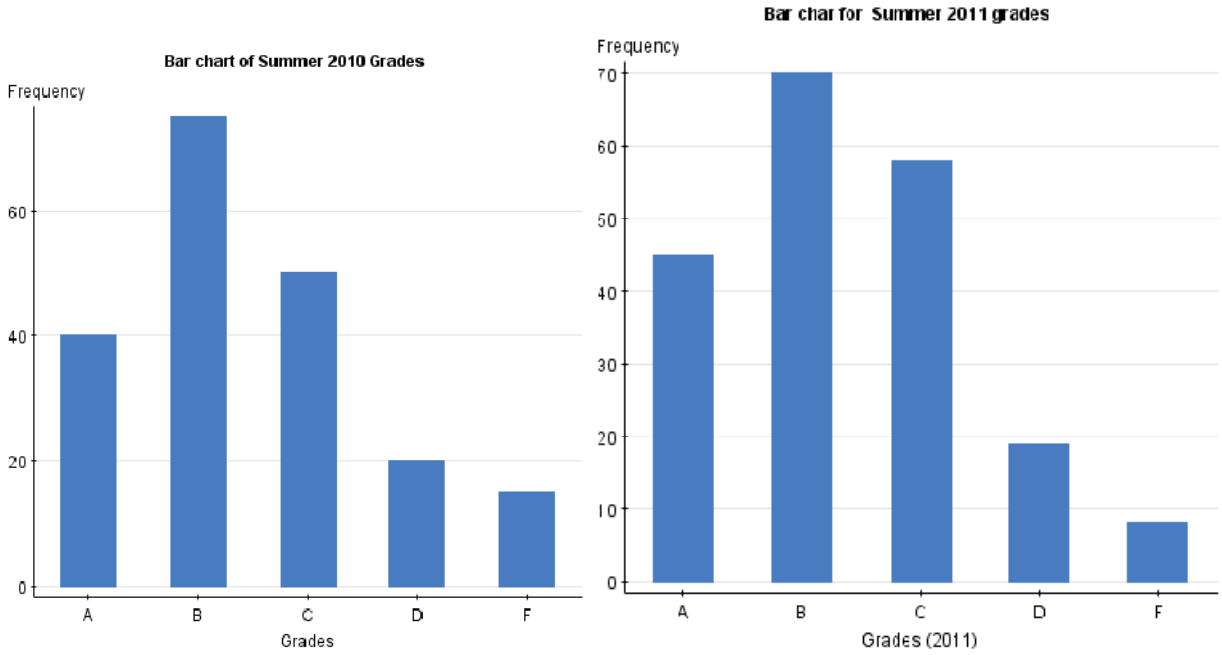
– *StatCrunch: graphics, bar plot, with summary*

Most students end up with a B



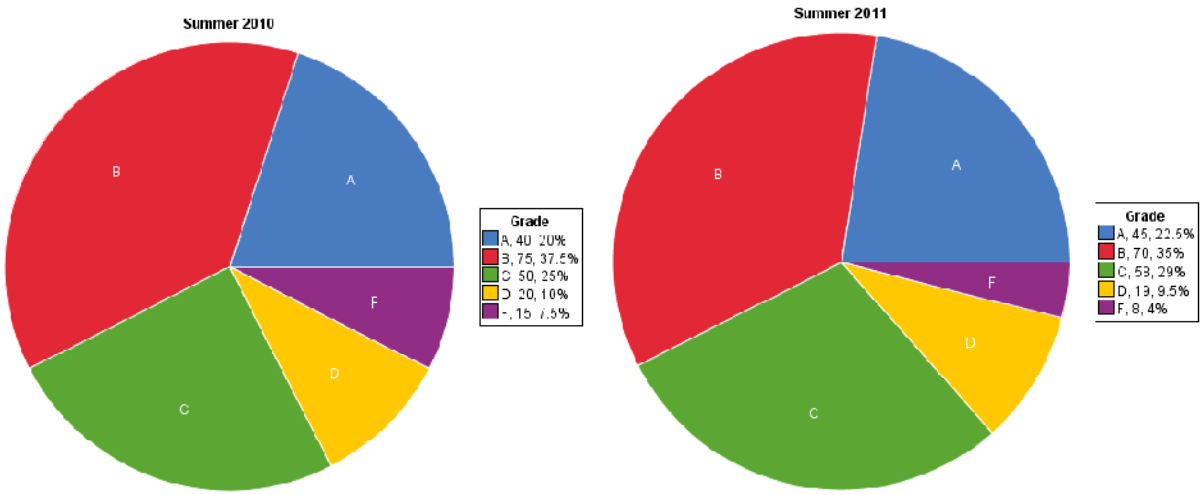
– *StatCrunch: graphics, pie chart, with summary.*

Bar charts

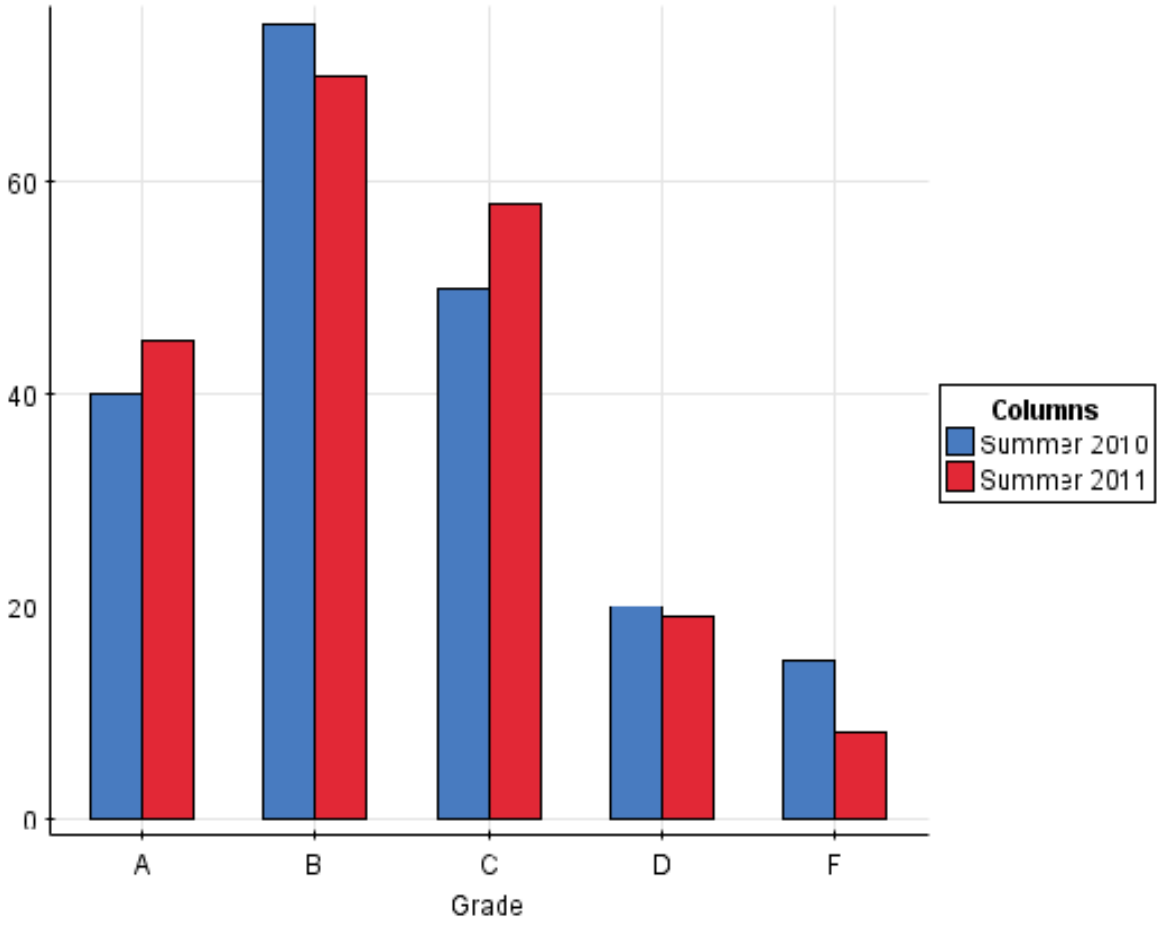


Hard to see much difference.

Pie charts



Again, not much difference apparent. Better: make bar chart for each year, but put bars side by side:



– *StatCrunch: Graphics, Chart, Columns.*

Contingency tables: two (or more) categorical variables p24

- University records applications to professional schools:

	Accepted	Rejected	Total
Males	490	210	700
Females	280	220	500
Total	770	430	1200

- 280 of the applicants were females who were accepted.
- How many of the applicants were males who were rejected?
  - *210*
- How many females applied altogether?
  - *500*

Joint distribution (percentage of total) p25

	Accepted	Rejected	Total
Males	490	210	700
Females	280	220	500
Total	770	430	1200

- More males than females applied (and more people accepted than not), so difficult to compare numbers.
- Compute *percentages.* (divide everything by 1200).
- *joint distribution.*
-
- StatCrunch: Stat, Tables, Contingency, With Summary.

Options

Contingency table results:

Rows: gender

Columns: None

Cell format
Count (Total percent)

	accepted	rejected	Total
males	490 (40.83%)	210 (17.5%)	700 (58.33%)
females	280 (23.33%)	220 (18.33%)	500 (41.67%)
Total	770 (64.17%)	430 (35.83%)	1200 (100.00%)

Marginal distributions p25

(a) Find the marginal distribution of gender.

(a) Find the marginal distribution for admission decision (i.e. whether accepted or rejected)

Conditional distribution p26

Options

Contingency table results:

Rows: gender

Columns: None

Cell format
Count (Row percent)

	accepted	rejected	Total
males	490 (70%)	210 (30%)	700 (100.00%)
females	280 (56%)	220 (44%)	500 (100.00%)
Total	770 (64.17%)	430 (35.83%)	1200 (100.00%)

- Joint distribution is “out of everything”.
- Doesn't answer question “are more males than females accepted”?
- For that: *out of males*, what % accepted – *row percents.*
- See males and females both add up to 100%.
- 70% of male applicants accepted, but only 56% of female applicants.
- Discrimination?

Column percents

- 63% of people accepted were males.
- 51% of people rejected were females.

Options

Contingency table results:

Rows: gender

Columns: None

Cell format
Count (Column percent)

	accepted	rejected	Total
males	490 (63.64%)	210 (48.84%)	700 (58.33%)
females	280 (36.36%)	220 (51.16%)	500 (41.67%)
Total	770 (100.00%)	430 (100.00%)	1200 (100.00%)

- Deciding between row and column percents

Three categorical variables and Simpson's paradox p33

Professional schools example: also recorded acceptance and rejection separately for law school and business school:

Law	accepted	rejected	total
males	10	90	100
females	100	200	300
total	110	290	400

Business	accepted	rejected	total
males	480	120	600
females	180	20	200
total	660	140	800

What would be appropriate percents to find here, and what do we conclude?

Professional schools

Options

Contingency table results:

Rows: law-gender

Columns: None

Cell format
Count (Row percent)

	accepted	rejected	Total
males	10 (10%)	90 (90%)	100 (100.00%)
females	100 (33.33%)	200 (66.67%)	300 (100.00%)
Total	110 (27.5%)	290 (72.5%)	400 (100.00%)

Options

Contingency table results:

Rows: business-gender

Columns: None

Cell format
Count (Row percent)

	accepted	rejected	Total
males	480 (80%)	120 (20%)	600 (100.00%)
females	180 (90%)	20 (10%)	200 (100.00%)
Total	660 (82.5%)	140 (17.5%)	800 (100.00%)

	Male % accepted	Female % accepted
Overall	70	56
Law school	10	33
Business school	80	90

– how is that possible???