Biostatistics and Informatics

BTM6000

Jochen G. Raimann

Session 1: Data

Learning objectives Session 1

- Introduction to the course
- Introduction and Comparison of R and Excel
- Outline of Syllabus and expectations
- Identifications of data types
- Creation and Interpretation of tables and graphs

COURSE LEARNING OUTCOMES

By the end of this course, students will be able to:

1. Explain the role of probabilistic models, inferential statistics, and descriptive statistics in science.

2. Design observational studies and experiments with appropriate scope of inference.

3. Collaborate with and coordinate the work of experimental scientists.

Required pre-installed software

- **MS Office** (or alternatively open-source alternatives i.e. LibreOffice) https://www.libreoffice.org
- **R base**
- **R Studio**

Required Textbook:

- Diez, D. M. (2015). *OpenIntro Statistics: Third Edition*. CreateSpace Independent Publishing Platform. A digital version is available free at *http://www.openintro.org/stat/textbook.php*.

Suggested additional reading

- Geoffrey R. Norman, Geoffrey R., and Sreiner, David L. (2014).*Biostatistics: The Bare Essentials 4th Edition*. People's Medical Publishing House - USA, Ltd.
- Irizarry, Rafael A. and Love, Michael I. (2016). *Data Analysis for the Life Sciences with R*. Chapman and Hall/CRC.
- Triolo, Marc M. and Triolo, Mario F. (2017). *Biostatistics for the Biological and Health Sciences, 2nd edition*. Pearson.

Grading

Evaluation criteria entire course

Assessment	Weight
Practice Labs	10%
Graded Labs (n=3)	30%
Quizzes (n=3)	30%
Final Project	30%
	100%

Evaluation criteria Final Project

Assessment	Weight
First Draft	10%
Analysis plan	30%
Final Report	40%
Presentation	20%
	100%

Final Project

There are 2 choices for the final project – every student is expected to choose one and to base the developed analysis plan on.

Option 1 (cross-sectional data):

Based on the provided NHANES dataset (actual study data) students are asked to develop a hypothesis and use simple analytic procedures to test this hypothesis. The hypothesis shall include the analytic testing of a physiological relationship the student is familiar with. The National Library of Medicine Pubmed.gov (https://www.ncbi.nlm.nih.gov/pubmed/) can be used to identify hypotheses that have been tested in the NHANES data and used for inspiration for developing own hypotheses. A glossary of the included parameters is provided on the following page.

Parameters in the dataset: age, gender, race, Hispanic ethnicity, albumin, creatinine, C-reactive protein, systolic and diastolic blood pressure, weight, height, waist circumference, Triceps skinfold

Final Project

There are 2 choices for the final project – every student is expected to choose one and to base the developed analysis plan on.

Option 2 (longitudinal, experimental data):

The provided dataset contains simulated (no real patients') data from a virtual randomized controlled trial. The diverse variables describe a before (i.e. baseline) and after period (i.e. follow up), an "intervention" parameter describes intervention versus control group. Students are asked to "invent" the trial with an established (patho-)physiological association, develop a study hypothesis for the conducted randomized controlled trial and to compare a) before and after period and b) control versus intervention arm using the methods taught in the course.

The students are asked to approach either project with creativity and an understanding where to find information (i.e. internet), how to approach and manage a dataset and how to formulate aims, null-hypotheses and how to approach data analytically using R. An analysis plan with a short background should be developed (sample on the following pages) and the final class will serve as the forum to present hypothesis and results in a brief presentation (<10 minutes) with a Q&A with the entire class.

Parameters in the dataset: age, gender, race, Hispanic ethnicity, albumin, creatinine, C-reactive protein, systolic and diastolic blood pressure, weight, height, waist circumference

Your instructor

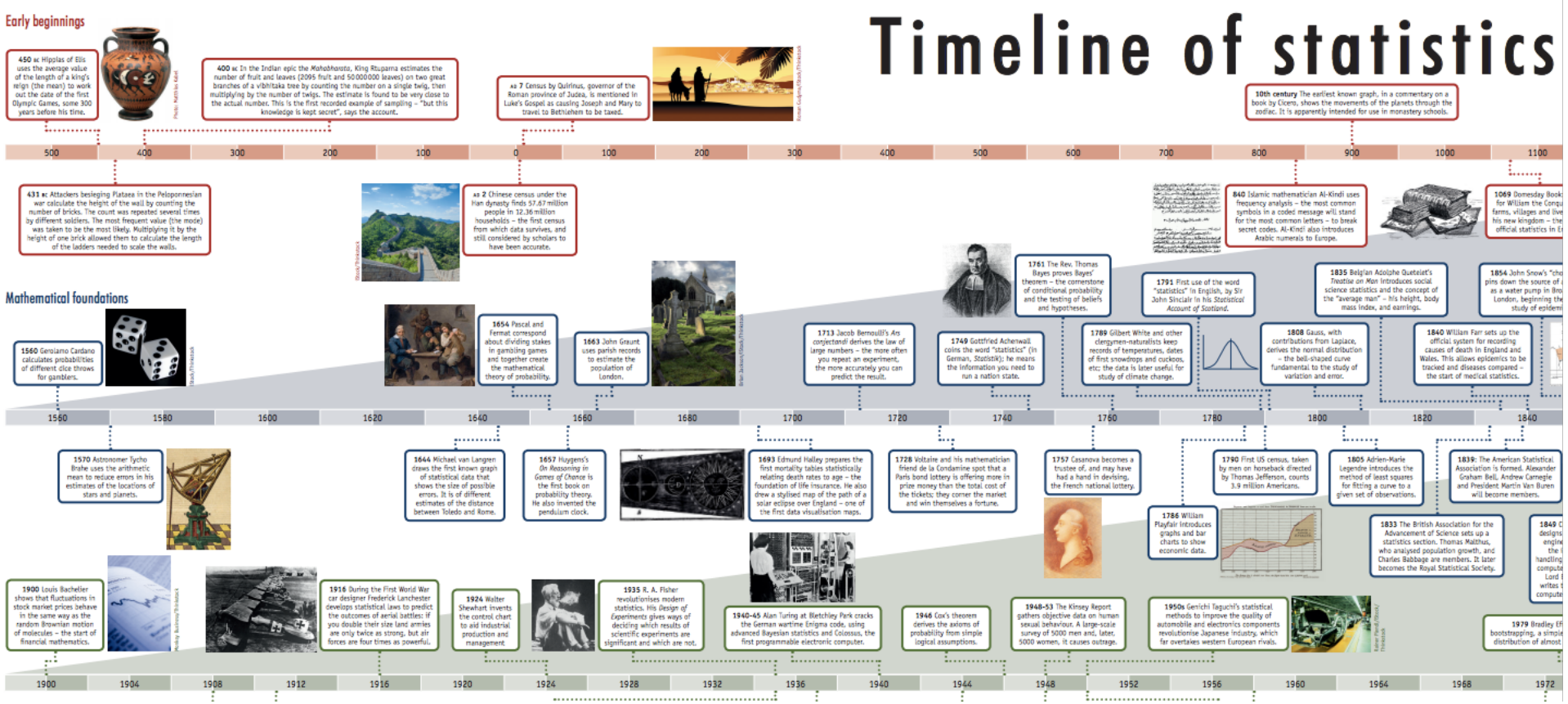
- Jochen G. Raimann, MD, PhD, MPH

- Research Interests

What is statistics?

- Major components of statistics are:
  - **Design**: planning how to obtain data
  - **Description**: Summarizing the data obtained
  - **Inference**: Making decisions and predictions that account for uncertainty and incomplete information

Timeline of statistics



450 BC Hippias of Elis uses the average value of the length of a king's reign (the mean) to work out the date of the first Olympic Games, some 300 years before his time.

431 BC Attackers besieging Plataea in the Peloponnesian war calculate the height of the wall by counting the number of bricks. The count was repeated several times by different soldiers. The most frequent value (the mode) was taken to be the most likely. Multiplying it by the height of one brick allowed them to calculate the length of the ladders needed to scale the walls.

http://www.statslife.org.uk/images/pdf/timeline-of-statistics.pdf

To begin with, some definitions

Population: the entire group of interest

- All people
- American adults
- Residents of New York

To begin with, some definitions

'London has been voted one of the world's favourite cities in the largest ever global survey of its kind by Ipsos MORI, covering 18,000 people in 24 countries.'

Sample: a subset of the group of interest

Note: a survey of the entire group is a ***census***

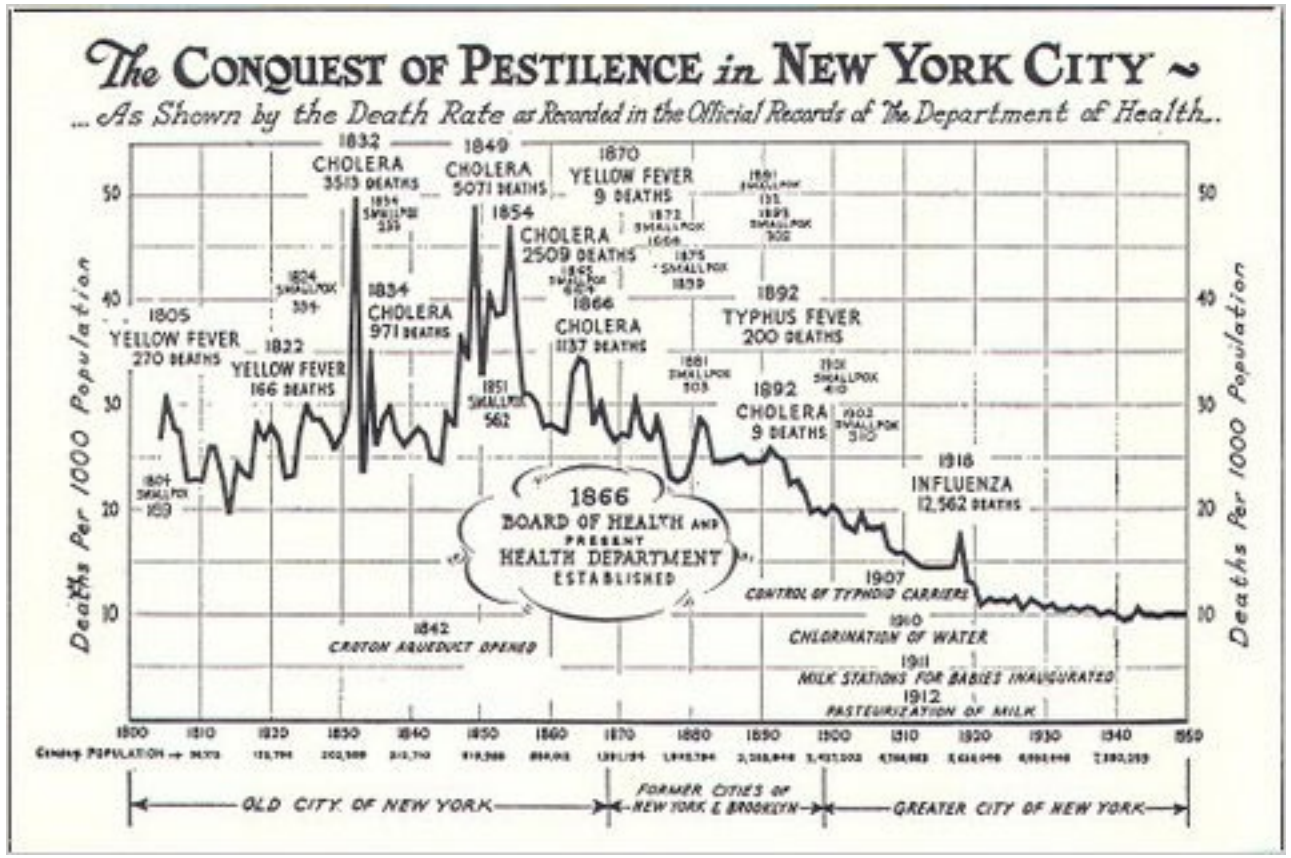
A single observation is also called a **record**

World's favourite city	Best city to live	Europe's top city	Best city to do business	Best city to visit
1. New York	1. Zurich	1. London	1. New York	1. Paris
2. London	2. Sydney	2. Zurich	2. Abu Dhabi	2. New York
3. Paris	3. London	3. Paris	3. Hong Kong	3. Rome
4. Abu Dhabi	4. Paris	4. Berlin	4. Tokyo	4. London
5. Sydney	5. New York	5. Rome	5. London	5. Sydney

Ipsos MORI Top Cities: Global Survey of the World's Favourite Cities from Ipsos MORI. Sept 4, 2013.

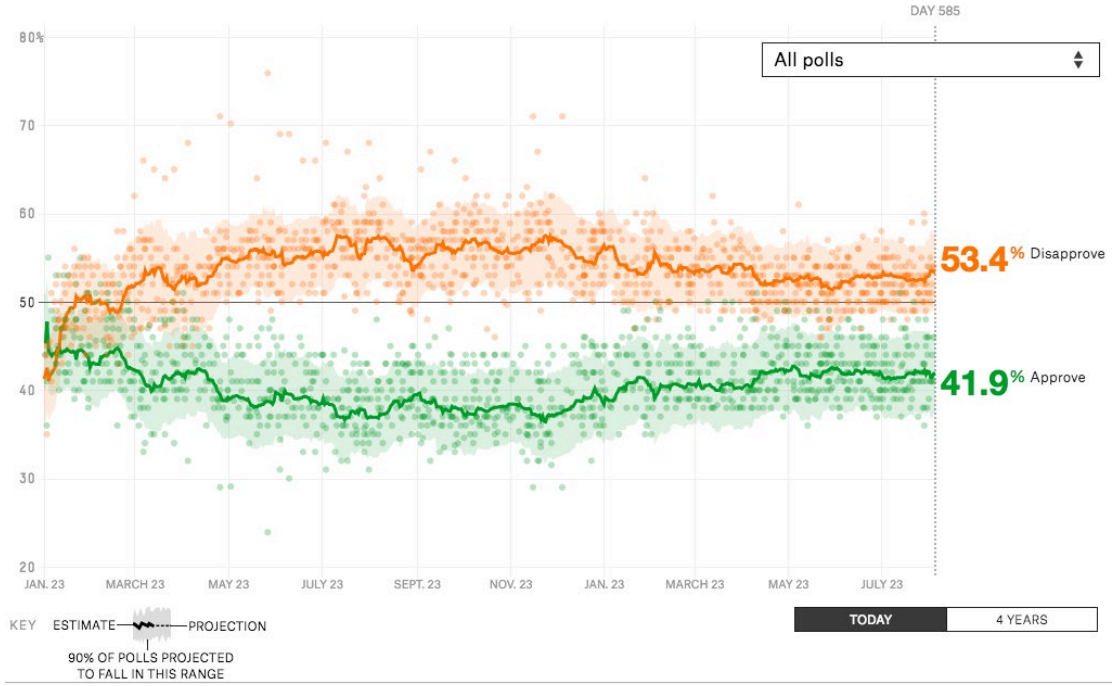
Parts of Statistics

- Descriptive Statistics



Parts of Statistics

- Inferential Statistics
  - *Inference about the population based on a sample*



Trump approval rating by multiple polls

How great is the uncertainty?

What are the sources of uncertainty?

Types of Data

- **Nominal data**
  - Binary / dichotomous, and categorical
- Ordinal data
  - Ordered data, rank data
- Discrete data
  - Integer data, count data
- Continuous data

Types of Data - nominal

- Binary / dichotomous
  - 0 – control, 1 – treatment
  - 0 – alive, 1 – dead
- Categorical
  - Blood types: 0=A, 1=B, 3=AB, 4=O

These are only *nominally* numbers

Types of Data

- Nominal data
  - Binary / dichotomous, and categorical
- **Ordinal data**
  - Ordered data, rank data
- Discrete data
  - Integer data, count data
- Continuous data

Types of Data - ordinal

"Performance scale":

- 0: fully active
- 1: can do light work
- 2: no work. Ambulatory > 50% of the time
- 3: ambulatory <50% of the time
- 4: disabled

Order important, *magnitude* unimportant

Rank data is an important type of ordinal data

Types of Data - ordinal

Number of deaths for leading causes of death in 2010 (CDC.gov):

1. Heart disease: 597,689
2. Cancer: 574,743
3. Chronic lower respiratory diseases: 138,080
4. Stroke (cerebrovascular diseases): 129,476
5. Accidents (unintentional injuries): 120,859

Magnitude of the difference doesn't matter

Only the order

Types of Data

- Nominal data
  - Binary / dichotomous, and categorical
- Ordinal data
  - Ordered data, rank data
- **Discrete data**
  - Integer data, count data
- Continuous data

Types of Data - discrete

Number of deaths for leading causes of death in 2010 (CDC.gov):

1. Heart disease: 597,689
2. Cancer: 574,743
3. Chronic lower respiratory diseases: 138,080
4. Stroke (cerebrovascular diseases): 129,476
5. Accidents (unintentional injuries): 120,859

Types of Data - discrete

Leading Causes of Death, New York City, 2002-2011

Source: Vital Statistics Data as of March, 2013
New York State Department of Health - Bureau of Biometrics and Health Statistics

Year and # of Deaths	#1 Cause of Death and # of Deaths Age-adjusted Death Rate	#2 Cause of Death and # of Deaths Age-adjusted Death Rate	#3 Cause of Death and # of Deaths Age-adjusted Death Rate	#4 Cause of Death and # of Deaths Age-adjusted Death Rate	#5 Cause of Death and # of Deaths Age-adjusted Death Rate
2011 **Total: 51,344**	Heart Disease 16,794 196 per 100,000	Cancer 12,476 [illegible]	Pneumonia and Influenza 2,490 29 per 100,000	Chronic Lower Respiratory Diseases (CLRD) 1,785 21 per 100,000	Stroke 1,778 21 per 100,000
2010 **Total: 50,852**	Heart Disease 17,669 211 per 100,000	Cancer 12,408 150 per 100,000	Pneumonia and Influenza 2,420 29 per 100,000	Chronic Lower Respiratory Diseases (CLRD) 1,723 21 per 100,000	Diabetes 1,673 20 per 100,000
2009 **Total: 51,446**	Heart Disease 19,715 229 per 100,000	12,277 144 per 100,000	Pneumonia and Influenza 2,188 26 per 100,000	Diabetes 1,653 19 per 100,000	Chronic Lower Respiratory Diseases (CLRD) 1,532 18 per 100,000
2008 **Total: 52,998**	Heart Disease 21,063 235 per 100,000	Cancer 12,238 142 per 100,000	Pneumonia and Influenza 2,284 25 per 100,000	Chronic Lower Respiratory Diseases (CLRD) 1,627 19 per 100,000	Diabetes 1,607 18 per 100,000
2007 **Total: 52,864**	Heart Disease 21,173 243 per 100,000	Cancer 12,450 148 per 100,000	Pneumonia and Influenza 2,353 27 per 100,000	Unintentional Injury 1,658 19 per 100,000	Stroke 1,601 19 per 100,000
2006 **Total: 54,225**	Heart Disease 21,681 255 per 100,000	Cancer 12,289 148 per 100,000	Pneumonia and Influenza 2,549 30 per 100,000	Stroke 1,713 20 per 100,000	Diabetes 1,678 20 per 100,000
2005 **Total: 55,827**	Heart Disease 22,514 263 per 100,000	Cancer 12,487 152 per 100,000	Pneumonia and Influenza 2,887 34 per 100,000	Diabetes 1,781 22 per 100,000	Stroke 1,666 20 per 100,000
2004 **Total: 56,043**	Heart Disease 22,429 271 per 100,000	Cancer 12,649 157 per 100,000	Pneumonia and Influenza 2,951 35 per 100,000	Stroke 1,820 22 per 100,000	Diabetes 1,715 21 per 100,000
2003 **Total: 57,869**	Heart Disease 23,722 293 per 100,000	Cancer 12,895 162 per 100,000	Pneumonia and Influenza 2,677 33 per 100,000	Diabetes 1,870 24 per 100,000	Stroke 1,867 23 per 100,000
2002 **Total: 58,327**	Heart Disease 24,281 307 per 100,000	Cancer 12,867 165 per 100,000	Pneumonia and Influenza 2,513 32 per 100,000	Stroke 1,877 24 per 100,000	Chronic Lower Respiratory Diseases (CLRD) 1,737 22 per 100,000

Standardized: per 100,000

Fluctuations in #4/5 positions

Discreteness becomes unimportant with many possible values

Magnitude matters, arithmetic applies

Types of Data

- Nominal data
  - Binary / dichotomous, and categorical
- Ordinal data
  - Ordered data, rank data
- Discrete data
  - Integer data, count data
- **Continuous data**

Types of Data - continuous

- Time
- Weight
- Length
- Mass

May be bounded or unbounded

All values are possible within bounds

Magnitude matters, arithmetic applies

Summarization of Data

- Very, very, very important
- Do it for every dataset you analyze seriously
- Multiple options for every data type *and combination of data types*

Numeric summary and graphical summary

Summarization of Data

- Nominal data
  - Binary / dichotomous, and categorical
- Ordinal data
  - Ordered data, rank data
- Discrete data
  - Integer data, count data
- Continuous data

Pie gone wrong!

- What's wrong with this?

THIS WILL BE THE ONLY PIE CHART YOU WILL SEE IN THIS COURSE!

Frequency Tables

- Also called contingency tables
- Provides the total number of each variable

Frequency Tables

Single nominal or ordinal variable

- Clinical trial:

Treatment	Control
50	50

- Severity:

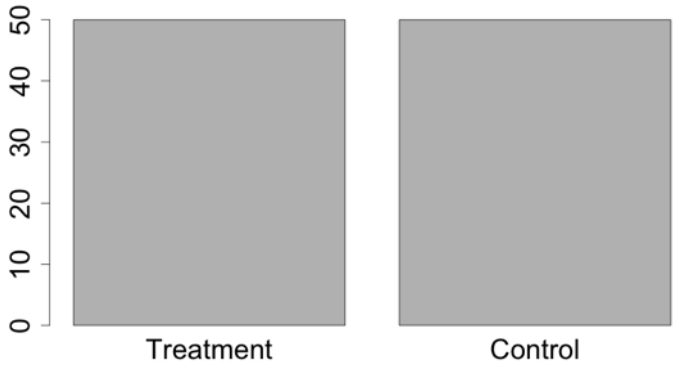
Mild	Moderate	Severe
20	30	50

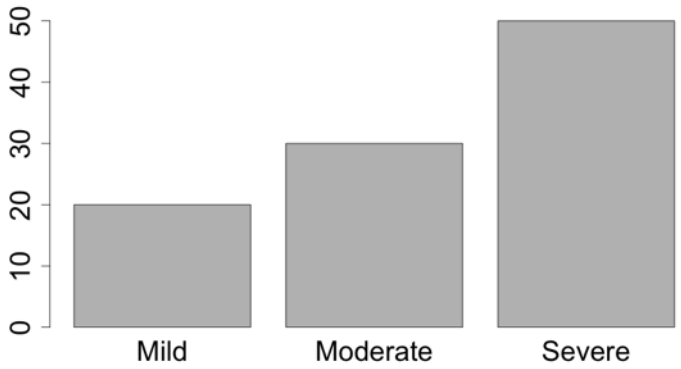
Frequency Tables

Two nominal or ordinal variables:

	Treatment	Control		
Mild	15	5	20	0.2
Moderate	15	15	30	0.3
Severe	20	30	50	0.5
Marginal totals →	50	50		
Marginal probabilities →	0.5	0.5		

Graphical Displays: Bar Chart

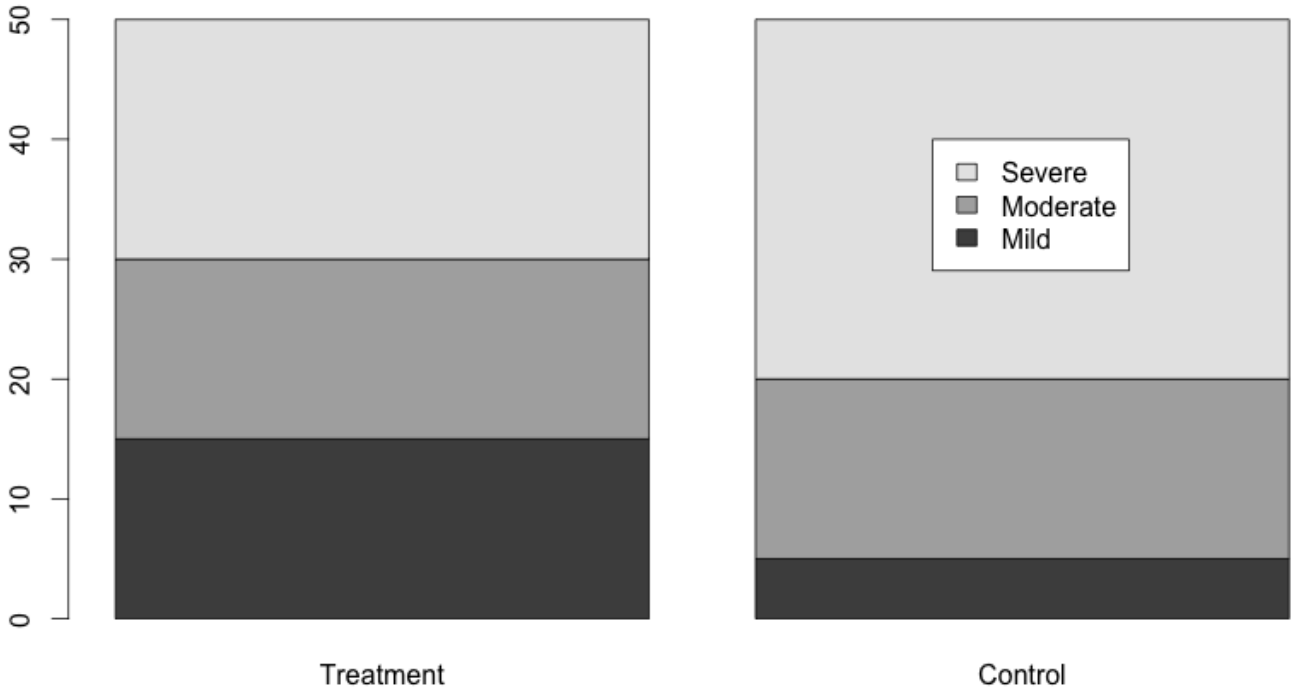




Single nominal or ordinal variable

Graphical Displays: Bar Chart

Two nominal or ordinal variables



Summarization of Data

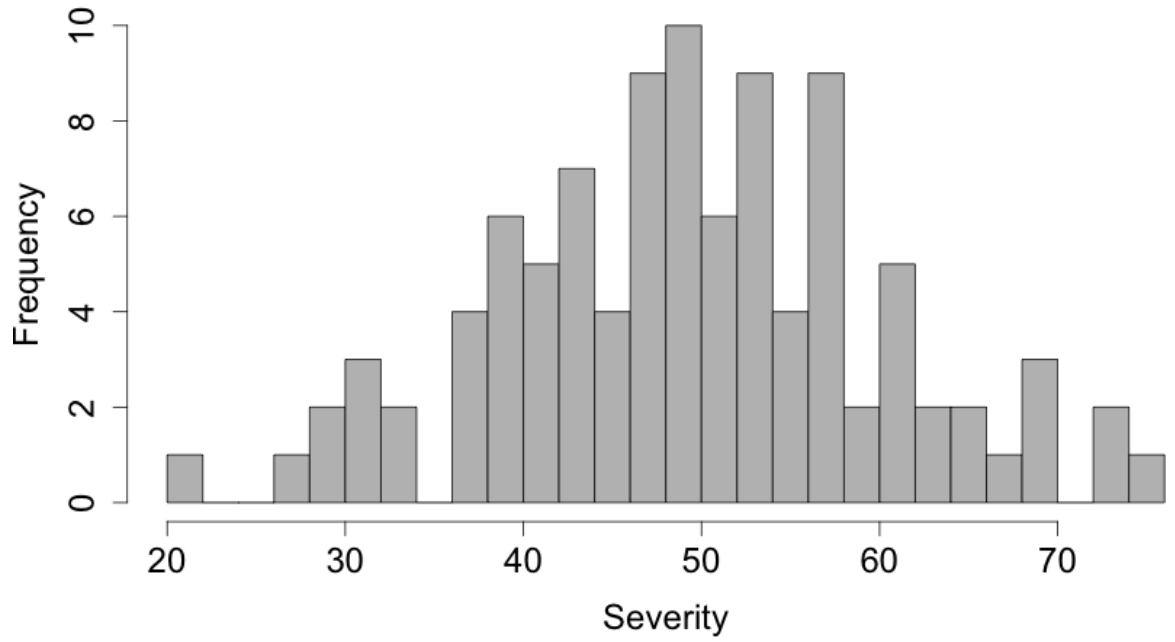
- Nominal data
  - Binary / dichotomous, and categorical
- Ordinal data
  - Ordered data, rank data
- Discrete data
  - Integer data, count data
- Continuous data

Frequency Tables

- Discrete and continuous data
  - select appropriate bins
  - Discretize (or further discretize)
  - Then do the same as for nominal and ordinal data
  - smaller bins = bigger table.

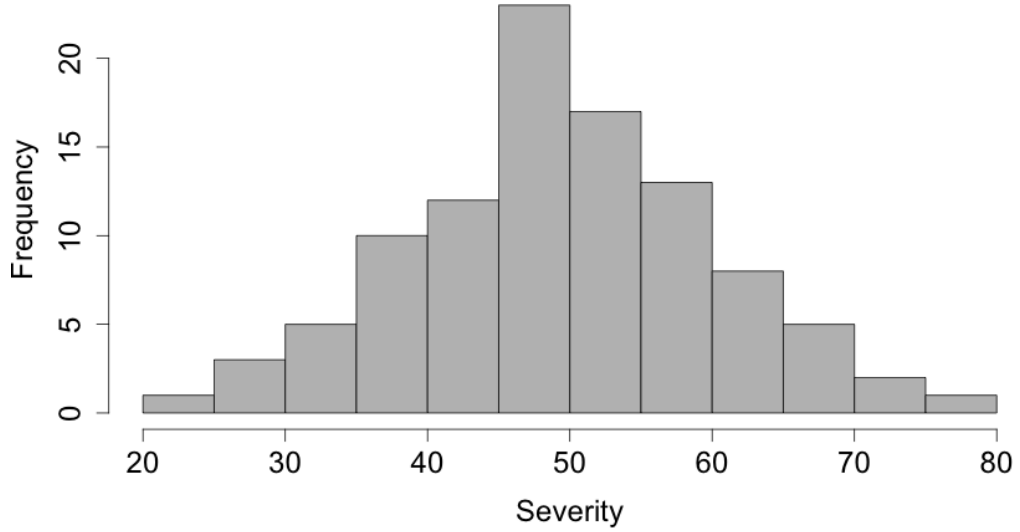
		Treatment	Control
0-30	Mild	15	5
30-50	Moderate	15	15
50-100	Severe	20	30

Histogram:
discretization of continuous data



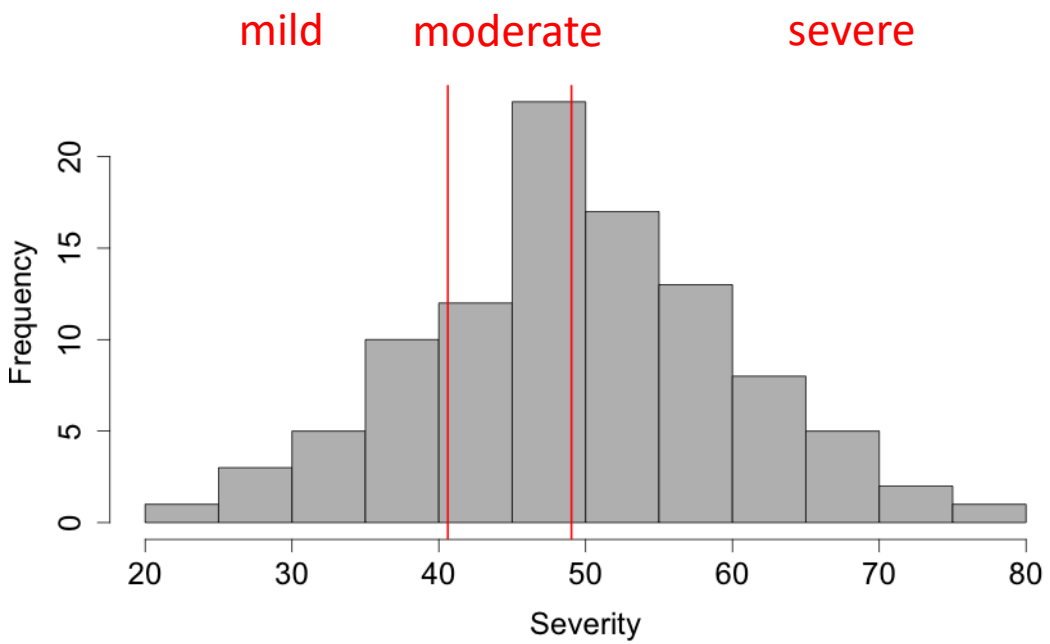
Can show more bins than a table, but it's the same

Histogram:
discretization of continuous data



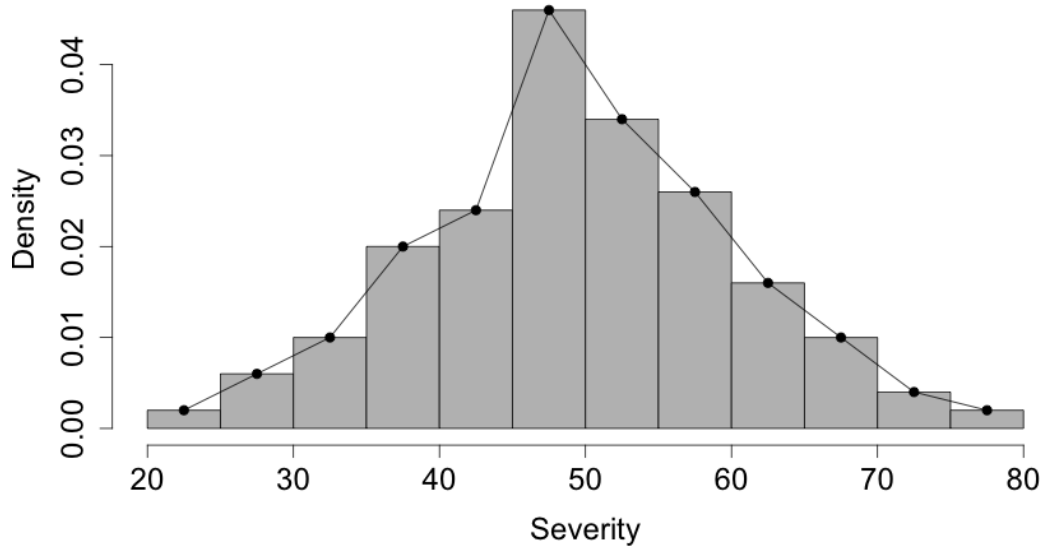
Number of bins is a subjective choice

Histogram:
discretization of continuous data



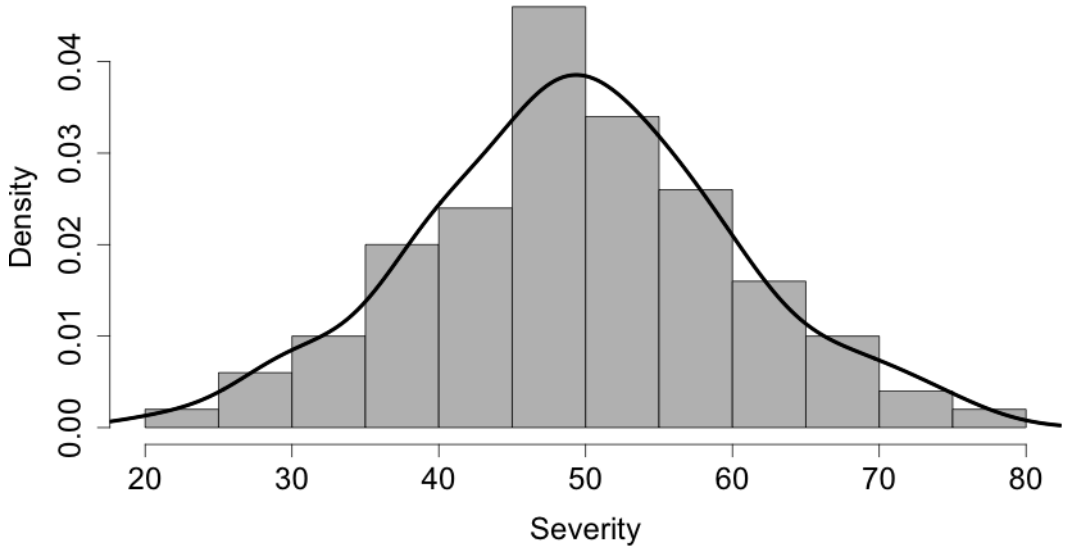
In this case, selection of mild/moderate/severe cutoffs is arbitrary

Histogram:
discretization of continuous data



“Frequency polygons”

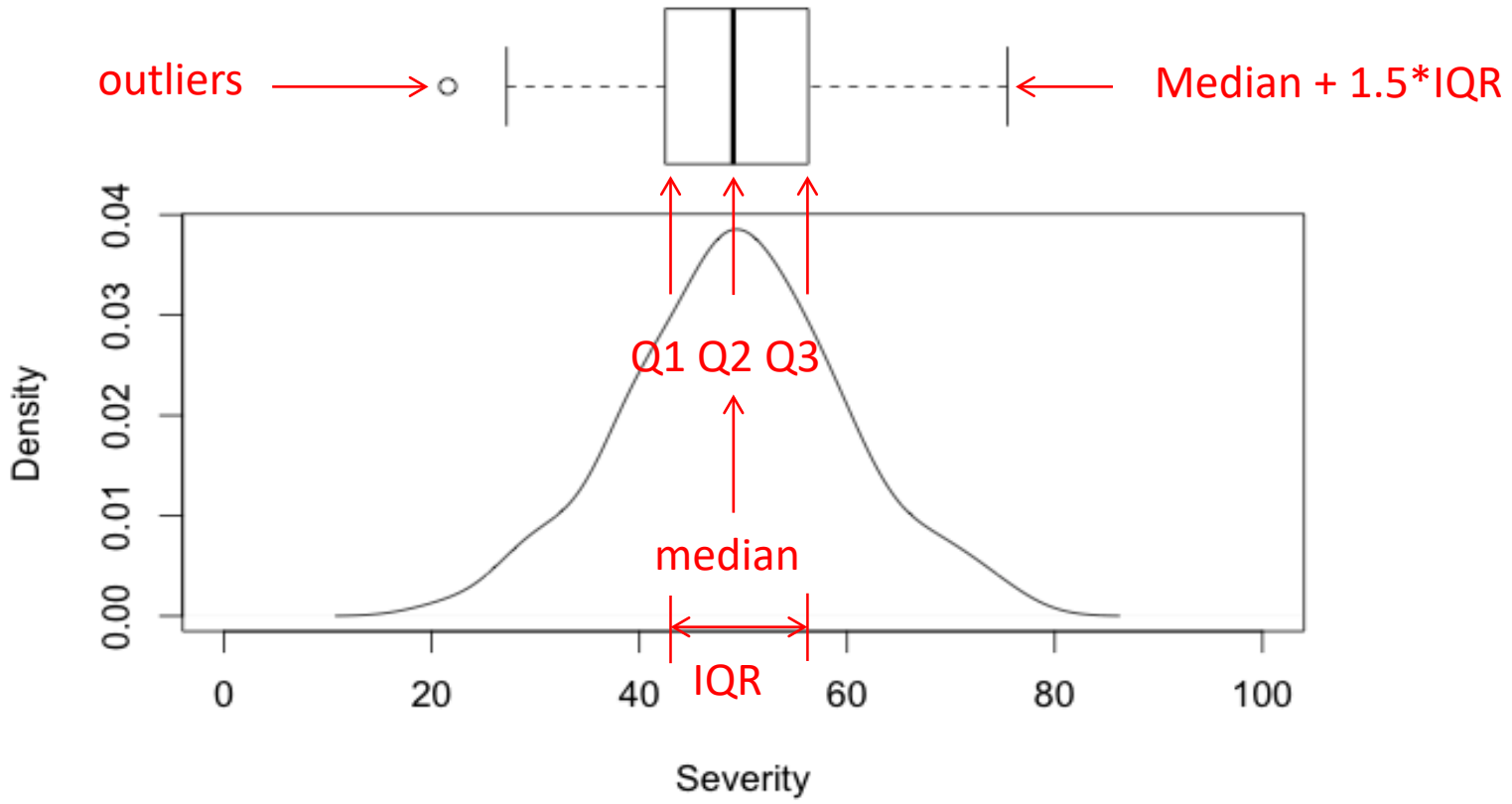
Histogram:
discretization of continuous data



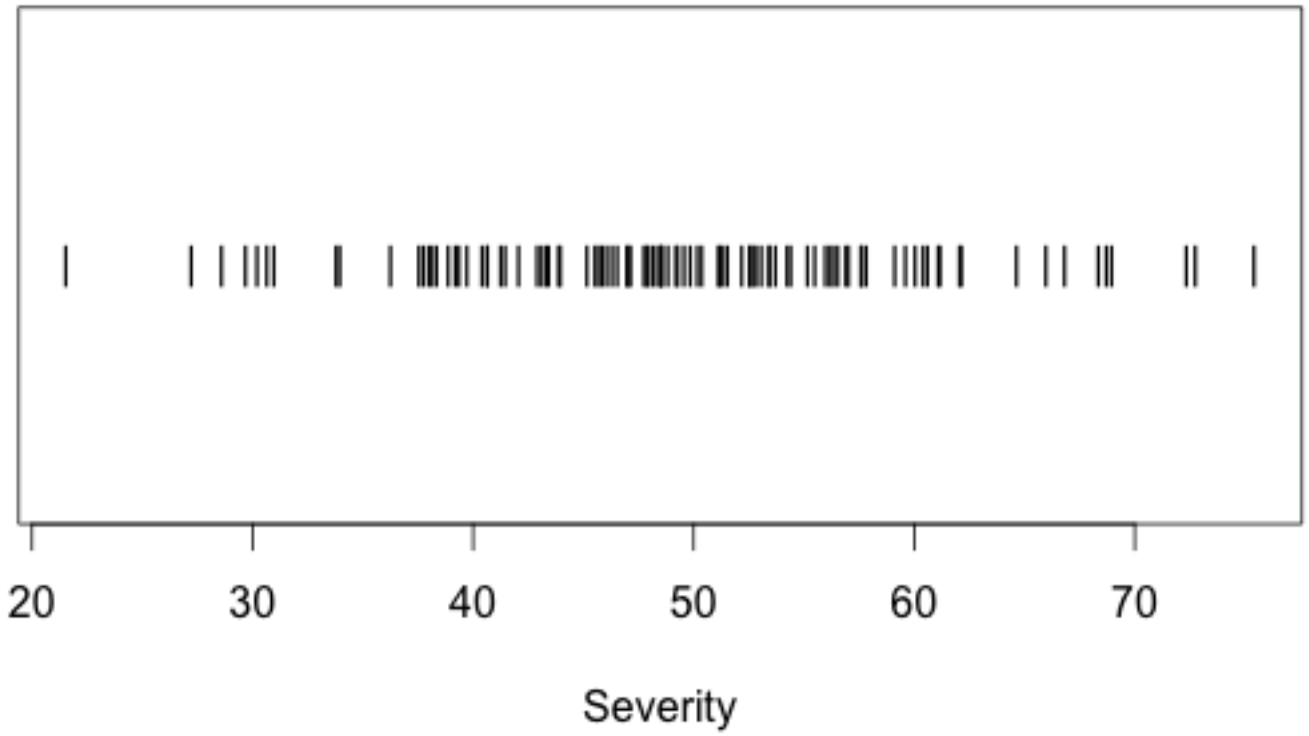
Can estimate a smooth probability density

Box plot

Percentile = % of data below a given value
Quartile = 25, 50, 75th percentile
Median = 50th percentile
IQR = Interquartile Range = 25th – 75th percentile

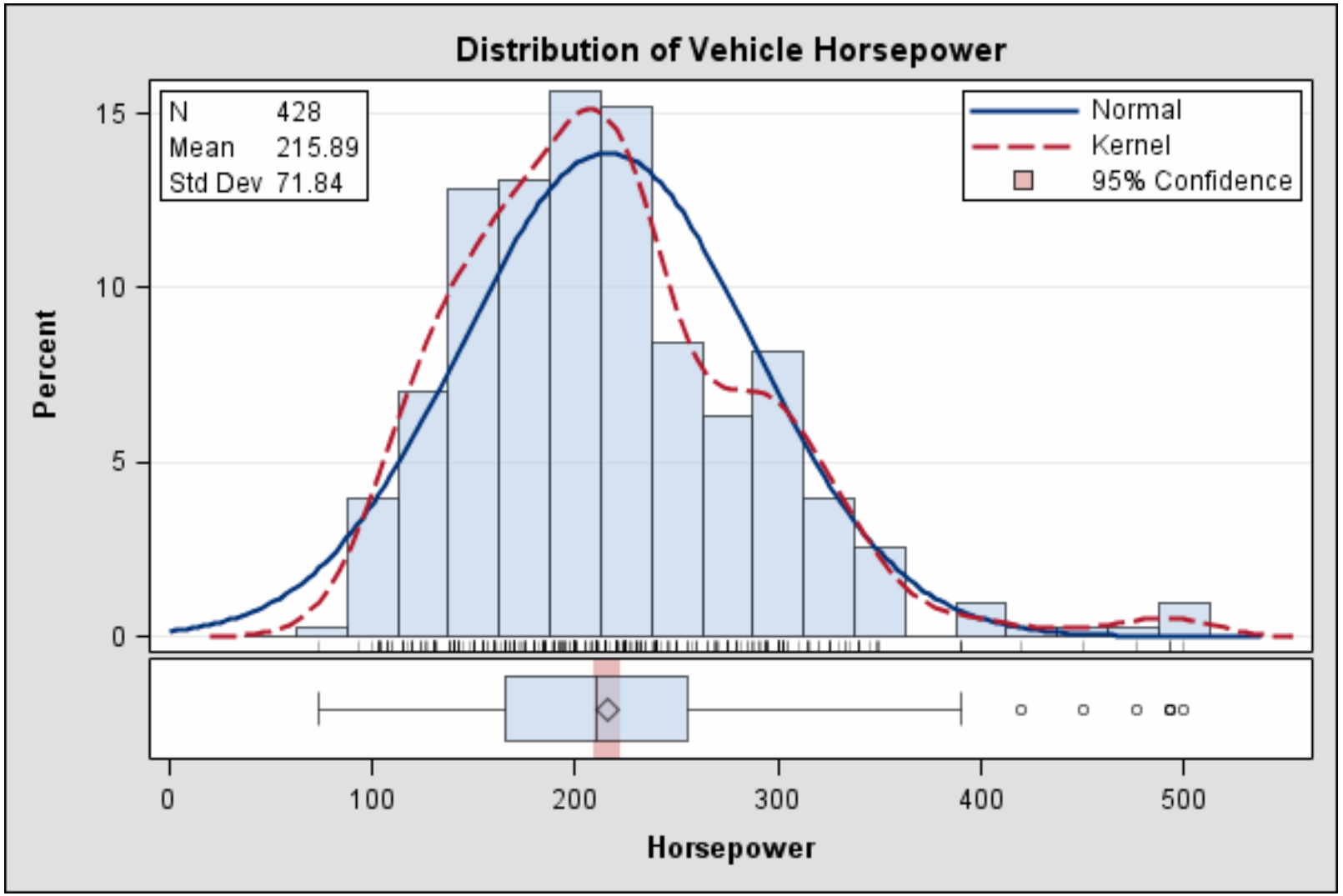


Exact plotting of data



One-dimensional scatterplot, or stripchart

Mash-up combinations…



http://support.sas.com/kb/35/addl/fusion_35171_1_gtlhandout_histogram.gif

Summarization of Data

- Nominal and ordinal data
  - Frequency tables and plots
- Discrete and continuous data
  - With further discretization:
    - Frequency tables and plots
    - Histogram
    - Boxplot
    - Frequency polygon
  - Without further discretization
    - Scatterplot, strip-chart