MS&E 125: Intro to Applied Statistics

Exploratory Data Analysis

Professor Udell

Management Science and Engineering
Stanford

March 23, 2023

Announcements

Questions from forum

Our programming language policy

- we'll do demos and provide homework starter code in python
- you're welcome to use any language you like (that your TAs can read) for homework or project
- TAs will only support python

Topics to review

We will cover (most of) these in section, too:

Why look at the data?

explore

- detect errors in data
- check assumptions
- select appropriate models
- understand relationships among the features
- understand relationships between features and labels

Why look at the data?

explore

- detect errors in data
- check assumptions
- select appropriate models
- understand relationships among the features
- understand relationships between features and labels

communicate

- convince others of your findings

How to look at the data?

- inspect raw data
- summary statistics
- visualize

Python and Jupyter

- Python is a programming language:
  it parses human-readable code to machine-readable code,
  executes it, returns the answer
- Jupyter is a protocol for interacting with a programming language.
- Jupyter stores inputs and outputs as `.ipynb` files.
- Jupyter notebooks display inputs and outputs in a browser
- Google Colab is an interface to a webserver running Python

Python and Jupyter

- Python is a programming language:
  it parses human-readable code to machine-readable code,
  executes it, returns the answer
- Jupyter is a protocol for interacting with a programming language.
- Jupyter stores inputs and outputs as `.ipynb` files.
- Jupyter notebooks display inputs and outputs in a browser
- Google Colab is an interface to a webserver running Python

how to access?

- install VSCode with Python extension
- install Python with Anaconda distribution
- use Google Colab

Python 3.9–11 are all fine

Summary statistics

univariate

- mean, median, mode
- max, min, range
- variance
- ...

explore via Python + Jupyter notebook

`https://colab.research.google.com/github/stanford-mse-125/demos/blob/main/sales.ipynb`

Summary statistics

univariate

- mean, median, mode
- max, min, range
- variance
- ...

explore via Python + Jupyter notebook

`https://colab.research.google.com/github/stanford-mse-125/demos/blob/main/sales.ipynb`

multi- (but usualy just bi-)variate

- correlation, covariance
- ...

The perils of summary statistics

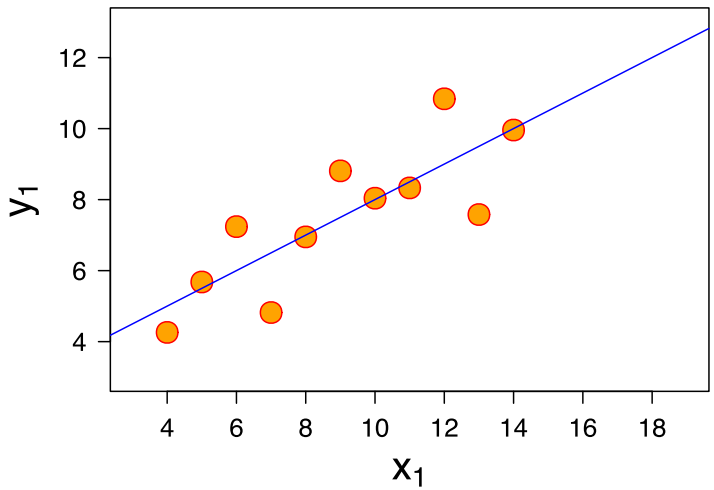
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

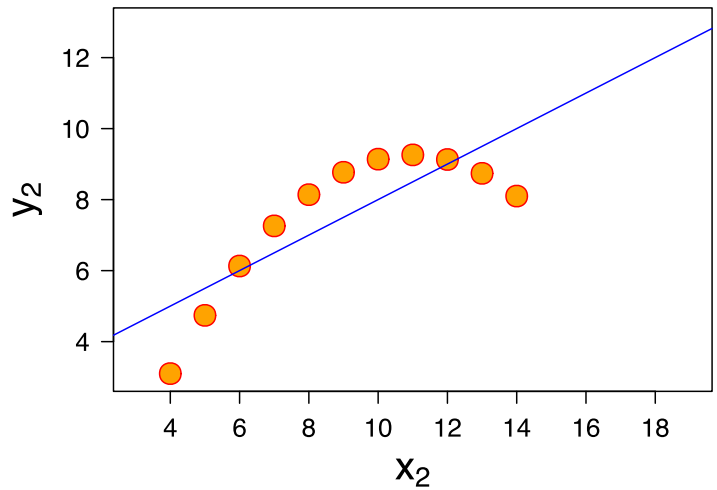
The perils of summary statistics

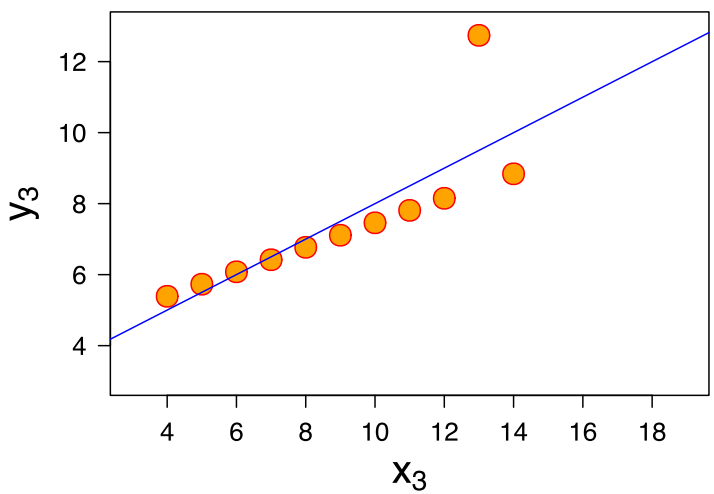
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

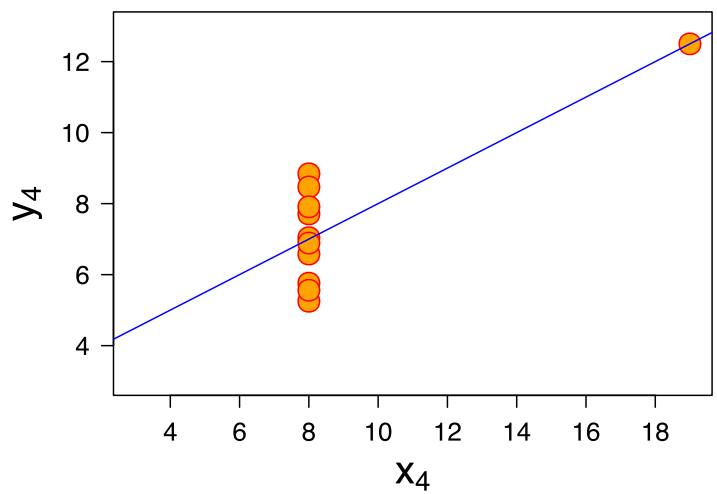
same mean, variance, correlation, line of best fit. . .

The perils of summary statistics









The perils of summary statistics: modern update

https://www.autodeskresearch.com/publications/samestats

Choosing a plot type

- Beware of pie charts; bar charts make comparisons easier.
- Beware of line plots; if your data is not continuous, try scatter plot instead.
- Beware of scatter plots; if you've got a lot of data, visualize the density instead
  - histogram, heat map, or contour plot
  - or at least make points transparent
- Visualize uncertainty or spread.
  - error bars, box plots, violin plots
- Consider the scale of your axes. Log scale or not?
  - log scale axis, not log scale data

Plotting parameters

- plot type
- scale and order ▷ not just alphabetic!
- color ▷ colorblindness is common!
- annotations and labels ▷ tell your story!
- orientation ▷ avoid head tilting
- size and aspect ratio ▷ ~~screenshot~~save as pdf

Principles of visual communication

- make comparisons easy
- maximize data-to-ink ratio
- label everything (axes, legends, etc.)
- final plot should tell a story

Beware of bad data

Label: Number of Days Physical Health Not Good
Section Name: Healthy Days — Health Related Quality of Life
Core Section Number: 2
Question Number: 1
Column: 91-92
Type of Variable: Num
SAS Variable Name: PHYSHLTH
Question Prologue:
Question: Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good?

Value	Value Label	Frequency	Percentage	Weighted Percentage
1 - 30	Number of days	159,327	36.43	35.59
88	None	269,145	61.53	62.53
77	Don't know/Not sure	7,602	1.74	1.58
99	Refused	1,336	0.31	0.30
BLANK	Not asked or Missing	26	.	.

Take away

- look at your data
- decide what you want to communicate

Questions?