

Statistics for Data Science

Lecture 1

Dennis Fok (Econometric Institute)

September – October, 2024

Who am I

Dennis Fok

- PhD in Econometrics
- Professor of *Econometrics and Data Science*
- Director of Econometric Institute, Erasmus School of Economics
- Research interests:
 - Modeling individual behavior
 - Marketing models
 - Panel data models
 - Simulation-based estimation methods
 - High-dimensional data
- Publications:
 - Marketing (*Marketing Science*, *Journal of Marketing Research*, *International Journal of Research in Marketing*)
 - Econometrics (*Journal of Econometrics*, *Journal of Applied Econometrics*)

The Erasmus University logo, featuring a stylized, handwritten-style script of the word "Erasmus" in black.

Course setup

Background of this course

Statistics:

- Most scary course –or– exiting and fun?
- Basis for many courses to follow!

Goals:

- (Re-)introduce statistics
- Apply everything in Python (or R)
- Not just know *how* to do things: also understand *why*!
- Critical thinking!

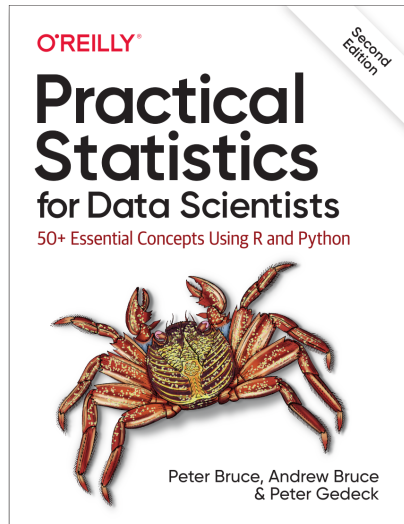
Setup of course

People involved

- Me!
- + (technical) assistant

Material

- Handouts of slides (most important)
- Book: Bruce, Bruce, and Gedeck, “Practical Statistics for Data Scientists” (we do chapters 1 – 5)
- Additional exercises



Study advice

Steps to take:

- ① Preparation: read book
- ② Lectures
 - Theory
 - In-class practice
- ③ After lecture:
 - Reread book (try out code examples in book)
 - Practice
 - ▶ Weekly assignment
 - ▶ Apply methods in own work environment!
- ④ “Final” assignment is to be submitted in parts
- ⑤ Questions and discussions
 - During class
 - Through *Discussions* on Canvas



Outline of course

- 1 Basics of statistics and inference
- 2 Distributions, descriptive statistics and hypothesis testing
- 3 Testing for differences
- 4 Linear regression model
- 5 Diagnostics for multiple regression + model selection
- 6 Generalized linear models (logistical regression)
- 7 Bayesian statistics

Statistics

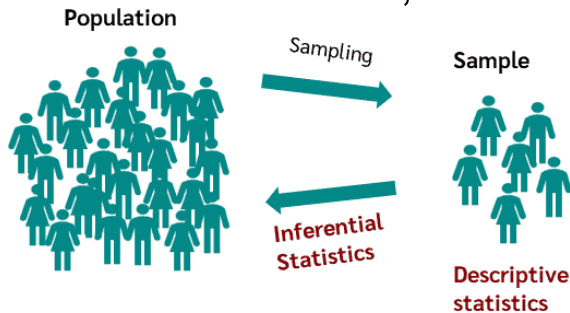
Goal of statistics

Goals:

- 1 **Summarize** properties of data
- 2 Make statements on (differences across) datasets
→ Statistical **hypothesis testing**
- 3 **Estimate** properties of (assumed) data generating process

} Descriptive statistics

} Inferential statistics



Erasmus

Usually statements on the **population** are the target!

Two important things to keep in mind:

- ① How good are these statements?
- ② Estimation uncertainty will always be present
- ③ **All** methods have associated assumptions (also ML/AI methods)!
 - Properties of methods derived under these assumptions
 - What if assumptions are not correct?

Some key concepts in statistics

- 1 Data and variables
- 2 Samples and population
- 3 Variation and uncertainty
- 4 Models

Key concept 1: Data

Key starting point is always: **data** or **dataset** → collection of observed **variables** or **features**


Classification of data/variables

- Role in the analysis
 - Dependent/Response/Outcome variable
 - Independent/Explanatory variable
- Measurement type
 - Numeric
 - ▶ Continuous (eg. temperature)
 - ▶ Discrete (eg. a count)
 - Categorical (aka: *factor with levels*)
 - ▶ Binary (eg. yes/no)
 - ▶ Nominal (no ordering: eg. color)
 - ▶ Ordinal (with ordering: eg. disagree/neutral/agree)

Notes:

- No clear dependent variable → *Exploratory statistics*
- Measurement type (dependent) variable → Determines *type of statistical analysis*
- In R/Python: data type may determine “actions” by functions

Terminology

- Data frame:  **DataFrame** objects
 - Data like a spreadsheet
 - Structured, rectangular data
- Other data shapes: possible, but more advanced material

Key concept 2: samples and population

- Source of data
 - Experimental
 - Observational
- Independent observations?
 - Repeated observations?
 - Hierarchical clustering? (eg. Children within a Class within a School)
- Random sample from population?
 - What is the population of interest?
 - What effects to control for?

Beware of *selective sampling / sampling bias*

Question

How (not) to get random sample for political survey?



Examples of non-random samples

- Survey on “random” people at the local market
- Response to a (e)mail/or online survey (response rate matters)
- ...

→ Compare “population” to “sample” to spot (potentially big) problems

Key concept 3: Variation and uncertainty

- Variation across samples is always expected
 - measurement error
 - different respondents
 - random variation
 - ...
- When is variation larger than expected?
- Comparing (assumed) truth (=unobserved) versus measurement/estimation (=observed)
 - Expectations vs. (sample) means
- Statistical concept: significance
 - A difference (*assumed* truth – observed) is significant:
 - Size of found difference is *unlikely* under the assumed truth
 - Not significant:
 - The found difference is not larger than what can be expected by chance alone
 - Not significant does not mean no true difference!
(and other way around?)

The Erasmus University logo, featuring a stylized, handwritten-style signature of the word "Erasmus" in black.

Key concept 4: Models

Statistical/Econometric models:

- Set of **assumptions** made about the *data generating process*
- Allow for description and prediction (and sometimes prescription)
- Important for all statistical procedures (even for “just testing”)
- “All models are wrong, but some are useful” (Box & Draper, 1987)
- Know which assumptions are crucial!
- Model choice and testing of assumptions are important
- How to fix things?

Models in Python (sneak preview)

Many models are available in Python packages (eg in statsmodels), examples:

- Linear model `m = smf.ols(..)`
- Generalized linear model `m = smf.glm(..)`
- Linear mixed effects `m = smf.mixedlm(..)`
- etc.

Most models allow for a large range of functions/results

- `r = m.fit()`: fit the model to data and get result named r
- `r.summary()`: print summary
- `r.predict()`: give fitted values
- `r.params`: give coefficients
- `sm.stats.anova_lm(r)`: analysis of variance of fitted model
- etc

(after `import statsmodels.api as sm` and
`import statsmodels.formula.api as smf`)

Organization of (statistical) analysis in Python

Steps to take:

- Import data (using pandas DataFrame: `import pandas as pd`)
 - Create a data frame directly: `data = pd.DataFrame(..)`
 - Load from file, eg. `data = pd.read_csv("file.csv")`
- Select and transform data (if necessary)
- Explore data (spot & fix errors)
 - `data.plot(title="Title text", ..)`
 - `data.describe()`
- Perform statistical calculations
- Present results

Organize all of this in a script, such that the results can be **replicated!**

→ See programming course for more info

The Erasmus University logo, featuring a stylized, handwritten-style signature of the word "Erasmus" in black.

In-class assignment

Assignment

- See the file `Day-1-AssignmentPython.pdf` on canvas.
- Do exercise 1.1

Descriptive statistics

Explore data

Use **descriptive statistics** to understand your data

Graphical:

- various plots: `dataname.plot.scatter(...)`,
for example `dataname.plot.scatter('xvar', 'yvar')`
- histograms: `dataname.seriesname.plot.hist()`
- density: `dataname.seriesname.plot.density()`
- boxplots: `dataname.boxplot('varname')`

where `dataname` refers to a dataframe and `seriesname` to a variable within the data)

Things to look for

- Degree of variation in variables
- Shape of distributions
- Signs of relations between variables (eg. correlation)
- Strange observations: Outliers!

The Erasmus University logo, featuring the word "Erasmus" in a stylized, handwritten script.

Summary statistics

- Graphical summaries of data are useful
- Numerical summaries → basis for further analysis

Consider n observations on a variable: X_1, X_2, \dots, X_n

Measures of **location/central tendency**

- Mode (`🐞 dataname.seriesname.mode()`): most frequently observed value
- Mean (`🐞 .mean()` or `🐞 np.mean(...)` from `🐞 import numpy as np`)
(note `dataname.seriesname` should come before the `.` or instead of the `..`)

$$\frac{X_1 + X_2 + \dots + X_n}{n} = \frac{\sum_{i=1}^n X_i}{n} = \bar{X}$$

- Median=50%quantile (`🐞 .median()`): 50% of observations is smaller
(median is much less sensitive to outliers than mean)

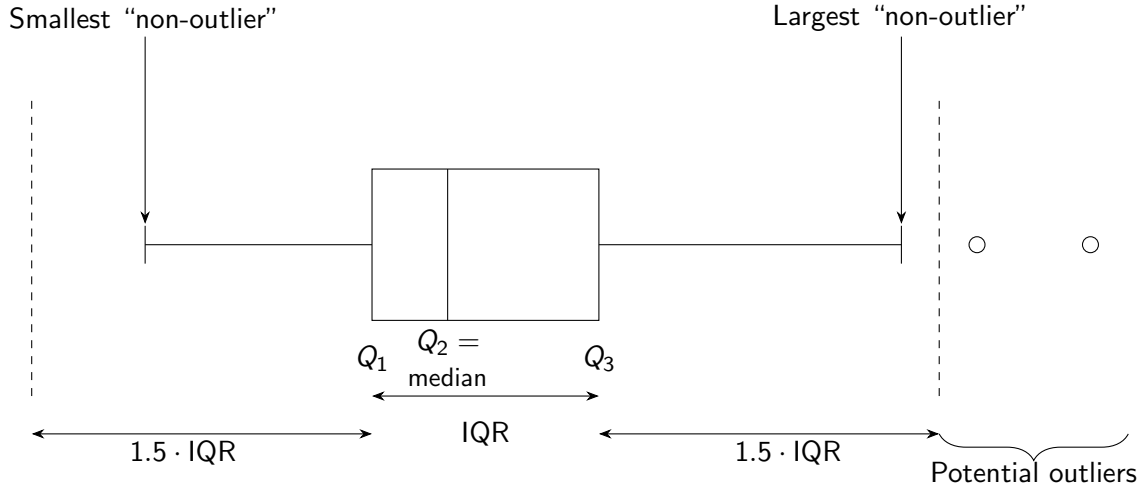


Measures of variation

Possible measures of variation

- Range (max-min): use `.min()` and `.max()`
- Inter-quantile range
 - 75% quantile – 25% quantile *or*
 - 3rd quartile – 1st quartile
 - `.quantile(0.75) - .quantile(0.25)`
 - Also useful to detect outliers
 - Common definition of outlier: obs. more than $1.5 \times \text{IQR}$ below 1st or above 3rd quartile

Putting some things together: Boxplots



Other measures of variation


- Mean deviation from mean? → will always be zero
 - Mean absolute deviation from mean?
 - Very useful (robust to outliers) *but*
 - Absolute values are mathematically difficult
- Use mean *squared* deviation from mean

Mean squared deviation & Degrees of freedom


The mean squared deviation is a crucial tool in statistics!

Given a sample X_1, \dots, X_n . Define **sum of squares** $= \sum_i (X_i - \bar{X})^2$

Important detail: how to define “mean”?

- Naive definition: sum over all i (all observations) and divide by n
- However: we used the data to calculate \bar{X} !
- Here we **know** that $\sum_i (X_i - \bar{X}) = 0$
→ We “lose” the information of one observation, **degrees of freedom** becomes $n - 1$
- **Estimated variance** of X ( `.var()`)

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

- In general: degrees of freedom = no. obs – no. estimated parameters
- **Standard deviation** $= \sqrt{\text{Variance}}$ ( `.std()`)



In-class assignment

Assignment

- See the file `Day-1-AssignmentPython.pdf` on canvas.
- Do exercise 1.2

Estimation uncertainty

Sample mean $\frac{1}{n} \sum_i X_i$ and sample variance $\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ are both **estimates!**

Therefore:

- Different sample \rightarrow different findings
- There is estimation uncertainty

Estimates of what?

\rightarrow Corresponding *population* concepts (remember the concept *inferential statistics*)

- Expected value: $E[X]$
- (Population) variance: $\text{Var}[X] = E[(X - E[X])^2]$

Higher order moments

Until now:

- Central tendency (eg. mean)
- Measures of variation (eg. variance)

Moments of a random variable X

- First moment: $E[X] = \mu$
- Second (central) moment: $E[(X - \mu)^2] = \text{Var}[X] = \sigma^2$
- Third (standardized) moment: $E[\frac{(X - \mu)^3}{\sigma^3}] = \text{skewness}$
- Fourth (standardized) moment: $E[\frac{(X - \mu)^4}{\sigma^4}] = \text{kurtosis}$

→ Can **estimate** all of these using data

 **.skew()** or  **.kurtosis()**

(do check exact definition of what is calculated!)



Moments for normal distribution

If $X \sim N(\mu, \sigma^2)$

- mean = $E[X] = \mu$
→ location
- variance = $E[(X - \mu)^2] = \sigma^2$
→ spread/variation
- skewness = $E[\frac{(X - \mu)^3}{\sigma^3}] = 0$
→ skewed or symmetric?
- kurtosis = $E[\frac{(X - \mu)^4}{\sigma^4}] = 3$
→ “peakedness”

Notes

- Often we look at excess kurtosis = kurtosis - 3
- Can test moments against values for normal distribution (more in later lectures)

Overview of descriptive statistics

Getting a quick overview

- `dataframe.describe()` from pandas package
- Various packages will give you options for descriptive statistics
- If you do not have a package yet:
 - Install it first (see programming course). This is needed only once.
 - Next load it with `import packagename` (in each session where you use it)
 - Abbreviate the package name (for later use): use eg `import pandas as pd` instead

Also possible: bivariate (or multivariate) descriptives

- scatter plot
- conditional boxplot
- correlation
- contingency table/cross table
- ...

The Erasmus University logo, featuring the word "Erasmus" in a stylized, handwritten script.

Assignment

Before next time

Assignment for next week

- Read
 - Chapter 1 (this week's material)
 - Chapter 2 (next week)
- Try some examples in the book yourself (see [here](#) for data and code)
- Finish today's assignments (1.1 - 1.4)
- Continue to practice using own data (or the housing data)
 - Create simple plots
 - Calculate summary statistics
 - Inspect distributions of some variables
(also consider transformations of variables)
 - Visualize relations between variables
- Optional: Exercise 2 (Volkswagen prices)