Erasmus

School of

Economics

Erasmus University Rotterdam



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:



Slide 2 of 30

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*)

Course setup

# Background of this course

Statistics:

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



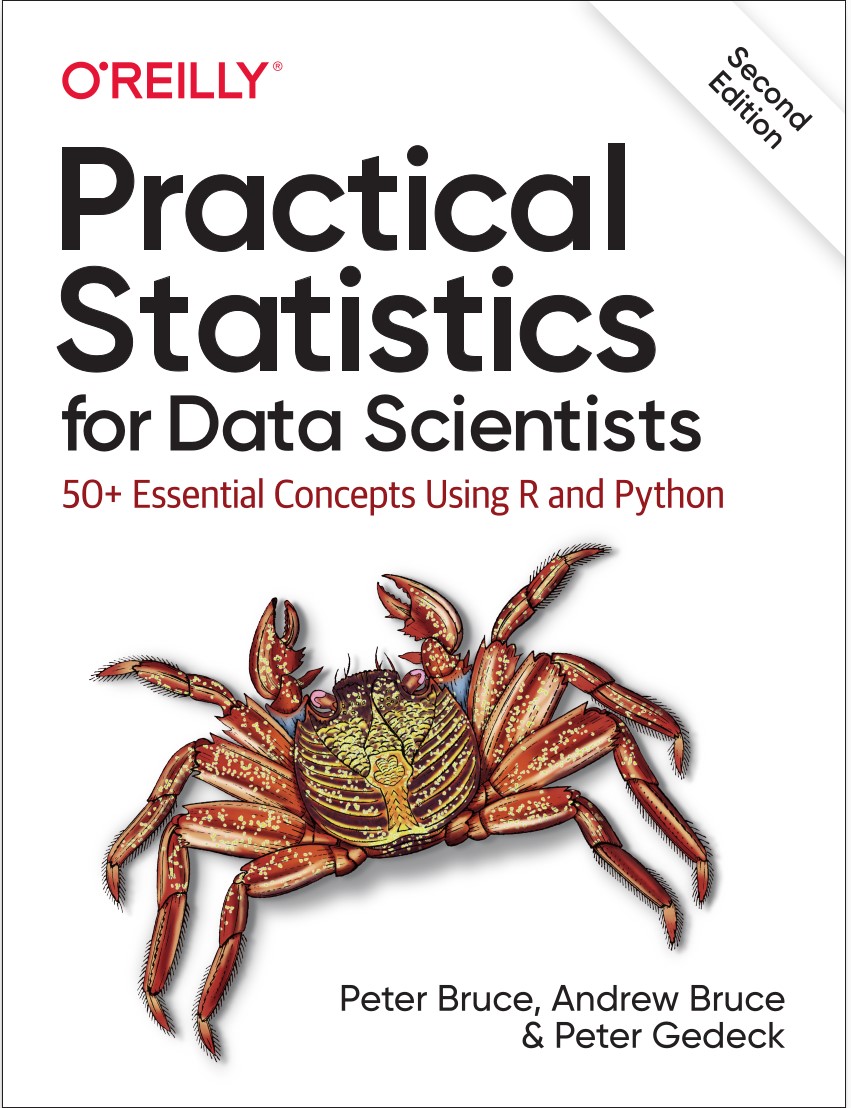
Slide 3 of 30

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:



Slide 5 of 30

Reread book (try out code examples in book)

Practice

▶ Weekly assignment

▶ Apply methods in own work environment!

1. “Final” assignment is to be submitted in parts
2. 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



Slide 6 of 30

1. Testing for differences
2. Linear regression model
3. Diagnostics for multiple regression + model selection
4. Generalized linear models (logistical regression)
5. Bayesian statistics

Statistics

Goal of statistics



Goals:

1

Summarize

properties of data

2

Make statements on (differences across) datasets

→

Statistical

hypothesis testing

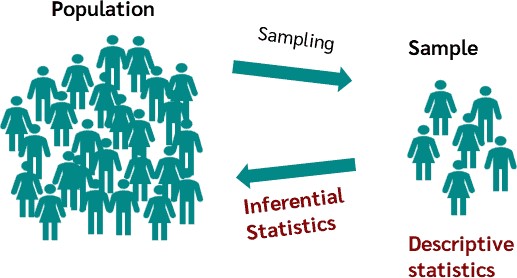
3

Estimate

properties of (assumed) data generating

Inferential statistics

Descriptive statistics



process

# Inference/Inferential statistics

Usually statements on the population are the target!



Slide 8 of 30

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

# Data sets

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



Beware of

*selective sampling*

/

*sampling bias*

Question



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

* 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?

# 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?)

# 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

In-class assignment

# Assignment



Slide 18 of 30

* 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!

# Summary statistics

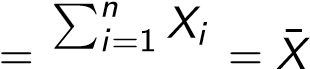
* Graphical summaries of data are useful
* Numerical summaries → basis for further analysis

Consider *n* observations on a variable: *X*1*,X*2*,...,Xn*

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 ..)

*X*1 + *X*2 + *...* + *Xn* 

*n n*

* 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×IQR below 1st or above 3rd quartile

Putting some things together: Boxplots



*Q*

1

*Q*

2

=

median

*Q*

3

IQR

1

*.*

5

·

IQR

1

*.*

5

·

IQR

Potential outliers

Largest “non-outlier”

Smallest “non-outlier”

# 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*,...,Xn*. Define sum of squares = P*i*(*Xi* − *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 *X*¯!
* Here we know that P*i*(*Xi* − *X*¯) = 0

→ We “loose” the information of one observation, degrees of freedom becomes *n* − 1 • Estimated variance of *X* (.var())

*n* 1

*s*2 = X(*Xi* − *X*¯)2 *n* − 1

*i*=1

* In general: degrees of freedom = no. obs − no. estimated parameters

√

* Standard deviation = Variance ( .std())



In-class assignment

# Assignment

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

# Estimation uncertainty

*n*

Sample mean 1 X*Xi* and sample variance 1 X(*Xi* − *X*¯)2 are both estimates! *n n* − 1 *i i*=1

Therefore:

* Different sample → different findings
* There is estimation uncertainty

Estimates of what?

→ Corresponding *population* concepts (remember the concept *inferential statistics*)

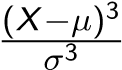
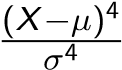
* Expected value: E[*X*]
* (Population) variance: Var[*X*] = E[(*X* − E[*X*])2]

# Higher order moments

Until now:

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

Moments of a random variable *X*

* First moment: E[*X*] = *µ*
* Second (central) moment: E[(*X* − *µ*)2] = Var[*X*] = *σ*2
* Third (standardized) moment: E[ ] = skewness
* Fourth (standardized) moment: E[ ] = kurtosis

→ Can estimate all of these using data

.skew() or .kurtosis()

(do check exact definition of what is calculated!)

# Moments for normal distribution

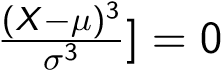
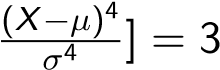
If *X* ∼ *N*(*µ,σ*2)

* mean=E[*X*] = *µ*

→ location

* variance=E[(*X* − *µ*)2] = *σ*2

→ spread/variation

* skewness=E[ → skewed or symmetric?
* kurtosis= E[

→ “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

* dataframename.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
* ...

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](https://github.com/gedeck/practical-statistics-for-data-scientists) for data and code)



Slide 30 of 30

* 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)