Introduction to Probability and Statistics

Jerome Dumortier

25 August 2025

Topics Covered

Introductions

- Name, degree sought, full-time/part-time, work experience
- Research and teaching activities

Syllabus

- Office hours, readings, grading, etc.

Topics

- Statistics in the real-world
- Course overview
- Types of data and levels of measurement
- R/RStudio
- Artificial intelligence

Syllabus in a Nutshell I

Office hours

- By appointment via email
- Come to office hours early in the semester
- Step learning curve regarding R/RStudio

Readings

- Lecture notes available at https://jrfdumortier.github.io/dataanalysis/
- No required textbook but potential reference books available for free through the library
  - 3.1 R Resources and Help
- Not related to R but to probability and statistics more generally: A Modern Introduction to Probability and Statistics

Syllabus in a Nutshell II

Assignments and grading

- 7 assignments to be submitted each as one PDF
- Read Assignment Formatting Guidelines
- Grading on linear scale without curve to avoid grading relative to classmates

Online course evaluations

- You receive 2 percentage points on your final grade if more than 90% of students fill out the online course evaluation at the end of the semester.

No restriction on the use of artificial intelligence (more on that later)

Risk and Uncertainty in Everyday Life

Grades

- Uncertainty surrounding class grade during a semester
- Association of probabilities with each grade

Fire station calls

- Number and location of calls
- Number of fire trucks and other vehicles required

Two outcomes does not mean a 50% chance for each to happen

- Success of a free throw by Stephen Curry
- Flight delay due to fog

Statistics in the News

Election outcomes

- 2002 French Presidential Election
  - Two-stage election
  - Final round: Jacques Chirac (82.2%) and Jean-Marie Le Pen (17.8%)
- 2016 U.S. Presidential Election
  - FiveThirtyEight forecast of Donald Trump winning: 28.6%
  - Cognitive biases versus data as an explanation
- 2024 U.S. Presidential Election
  - 7 swing states that all voted for Donald Trump: Chance of 0.78% of that outcome in the case of independence of events

Path of hurricane Sandy

Financial Economics

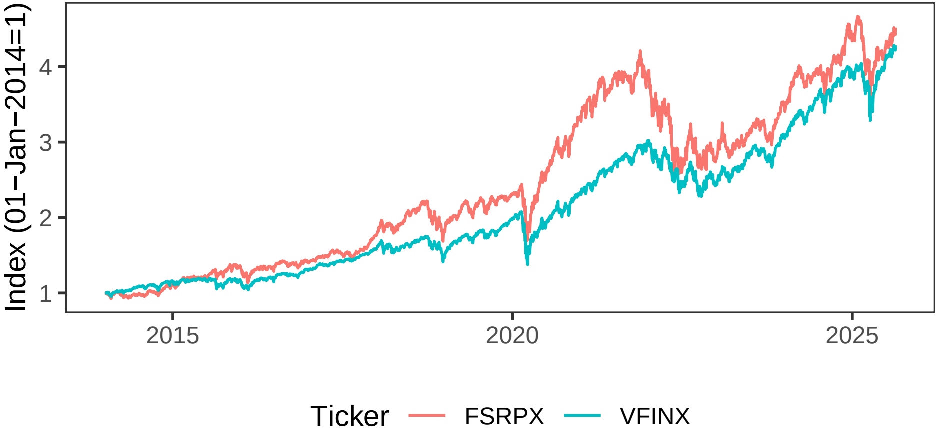
Evolution of the stock market

- Importance of correlation among stocks and mutual funds

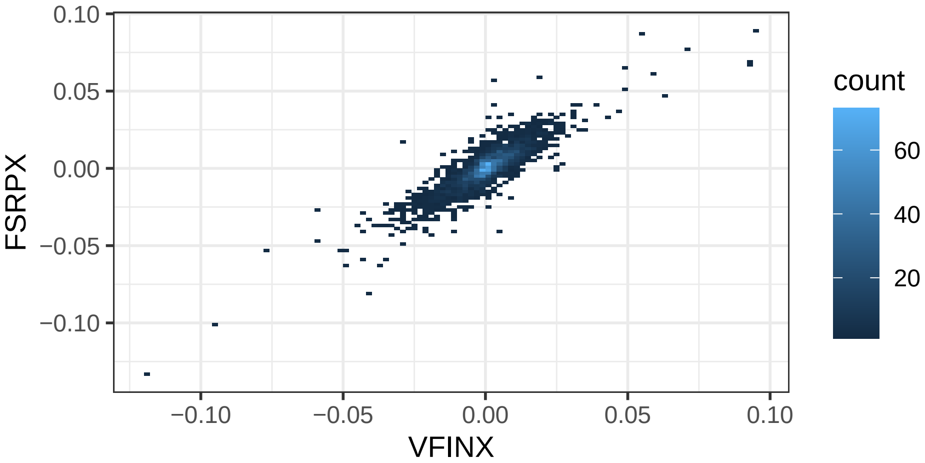
Next slides

- FSRPX: Fidelity Select Retailing Portfolio
- VFINX: Vanguard 500 Index Fund Investor Shares

FSRPX and VFINX: Evolution



FSRPX and VFINX: Returns Scatter Plot



Scottish Ministers' Widows' Fund

Preceding work

- Edmond Halley's (same as comet) life tables for the city of Breslau (today Wrocław) in 1693
- Detailed work on birth and death by age

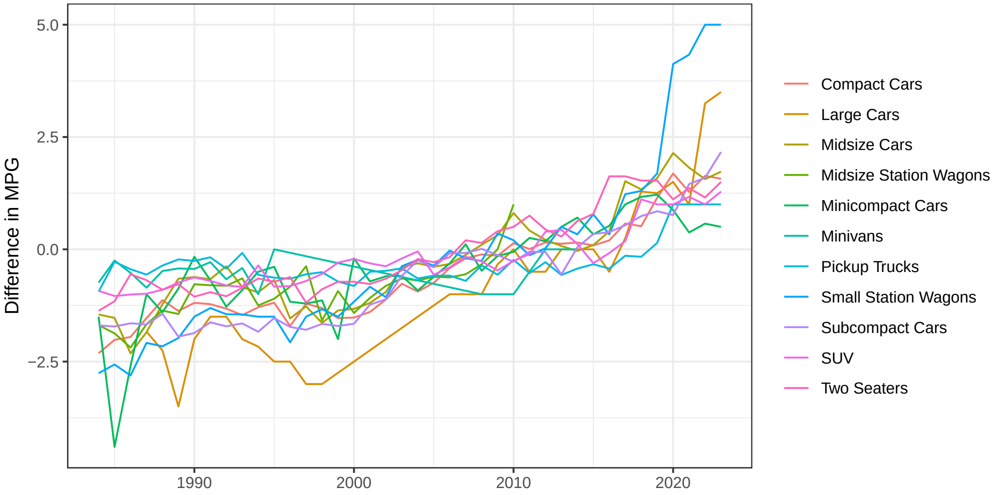
Insurance fund calculations in 1744 by Alexander Webster and Robert Wallace

- Payments to widows and heirs after death of ministers
- Required information: Number of clergymen, deaths per year, life expectancy of surviving family, time of remarriage, etc.
- Calculation of annual payments into the fund

Fund balance (in pound sterling) in 1765

- Estimated: 58,348
- Actual: 58,347

Automatic vs. Manual Transmission



Data, Probability, Statistics, and Regression

Data and data visualization

- Descriptive statistics and graphical presentation of data

Probability

- Providing basis for modeling populations, experiments, and any other random phenomena

Statistics

- Learning about the population based on a sample

Regression analysis

- Mathematical relationship among variables

Difference between probability and statistics: Bucket example

Data and Descriptive Statistics Overview

Data

- Raw observations collected from surveys, experiments, or administrative records.
- Examples in public policy: Annual household income, homicide rates, citizens' opinions on a policy

Descriptive statistics

- Summarizes and simplifies data to reveal patterns
- Examples: Average donations per year to a nonprofit
- Common descriptive statistics are related to central tendency (e.g., mean, median) and dispersion, i.e., spread of the data measured by range, variance, or standard deviation
- Purpose: Understand the typical values, detect variability and outliers, and provide a foundation for further analysis

Data Visualization Overview

Graphical representation of data to make patterns intuitive and interpretable

- Increasing importance for communicating with the public
- Examples previously used: Pattern of financial returns and fuel economy

Common techniques used

- Histograms, bar charts, boxplots
- Scatterplots, line charts

Benefits:

- Identify trends, relationships, and anomalies
- Communicate results clearly to policymakers and stakeholders

Probability Overview

Study of uncertainty and randomness

- Given a model, what are the chances of an event happening

Public policy examples

- Chance (probability) of a voter turnout of over 60%
- Probability of extreme weather events

Topic covered

- Probability theory: Many examples of flipping coins and rolling dice
- Probability distributions: How to model random outcomes

Foundation for statistics

Difference between Population and Sample

Population

- A population is the collection of all possible individuals, entities, objects, or measurements of interest for a particular investigation. A sample is any portion or subset of the population. A *parameter* characterizes the population and is usually unknown (forever).

Sample

- A statistic is any measurable characteristic of a sample. Statistical analysis utilizes statistics from representative samples to infer the parameters of an entire population.

Using a sample rather than the population

- Cost considerations
- Possible destruction of observation units (e.g., mileage of tires)
- Unfeasible to study all units of observations

Statistics Overview

Inference

- Given data, what can we infer about the population or phenomenon?

Examples:

- Estimating average income in a city
- Calculating unemployment rates from survey data
- Relies on probability to quantify uncertainty in inferences

Topics covered

- Sampling
- Confidence intervals
- Hypothesis testing

Regression Overview

Regression analysis

- Statistical method to model the relationship between a dependent variable (Y) and one or more independent variables (X)
- Correlation is not causation!

Types of regression models

- Simple linear regression (i.e., one independent variable), e.g., effect of education on income
- Multiple regression (i.e., multiple independent variables), e.g., effect of education, experience, and age on income
- Other regression models: Logistic regression (binary outcomes), Poisson regression (count data), etc.

Example: Price of a used car as a function of mileage

Variables

Qualitative variables

- Non-numeric, e.g., gender, political affiliation, state of residence
- Can be transformed into numerical value, i.e., “dummy variables’ ’ in regression analysis

Quantitative variables

- Numeric, e.g, age, income, GPA, number of kids

Quantitative variables can be either

- Discrete: Take two close values and there is no value in between, e.g., number of people in a class
- Continuous: Take two close values and there is always (!) a value in between, e.g., weight of a people

Levels of Variable Measurements

Nominal

- Categories, e.g., eye color, gender, religious affiliation, mode of transportation to O'Neill IU Indianapolis
- No natural ordering

Ordinal

- Categories, e.g., level of happiness, Homeland Security Advisory System
- Natural ordering, i.e., data can be ordered

Interval

- Intervals between levels are equally spaces and differences between variables have a meaning
- Examples: Income, GPA, etc.
- Most commonly used in this class.

Introduction to R and RStudio

R

- A programming language and environment for statistical computing and graphics
- Widely used in public policy, economics, and data science
- Open-source and free

RStudio - An integrated development environment (IDE) for R - Provides a user-friendly interface with: Script editor, console, environment/history panels as well as plots, files, and packages panels

Reasons to use R and RStudio

- Powerful statistical and graphical capabilities
- Active community with thousands of packages for specialized analyses
- Reproducible research with R Markdown and Shiny

Advantages and Disadvantages

Advantages

- Free and open-source
- Supports advanced statistics and machine learning
- Excellent for data visualization and reporting
- Reproducible workflows with R Markdown
- Active and helpful community
- Maybe most importantly: Easy to use with AI

Disadvantages

- Steeper learning curve than spreadsheet software
- Some packages may have inconsistent syntax
- Memory-intensive with very large datasets
- Limited GUI support compared to commercial software (e.g., SPSS, Stata)