# Statistics

## Week 2: Statistics Models

Etienne Wijler

Econometrics and Data Science
Econometrics and Operations Research
Bachelor Program

VRIJE UNIVERSITEIT AMSTERDAM
SCHOOL OF BUSINESS AND ECONOMICS

# The goal of statistics

Goal: Statistics is about answering questions or making decisions in the face of uncertainty.

Examples: statistics are used across all scientific and business disciplines to answer questions such as:

- What is the probability that a destructive tornado hits the US next year?
- Is a new medical procedure better than the older one?
- How sure are we about the predictions of a political election?

# The statistics approach

1. Formulate the research question.
2. Collect relevant data $(x_1, \ldots, x_n)$.
3. Formulate a statistical model (week 2).
4. Estimate the parameters of the statistical model (week 3-6).
5. Conduct inference and quantify uncertainty (week 7-12).

# The statistics approach

1. Formulate the research question.
2. Collect relevant data $(x_1, \dots, x_n)$.
3. Formulate a statistical model (week 2).
4. Estimate the parameters of the statistical model (week 3-6).
5. Conduct inference and quantify uncertainty (week 7-12).

# Dealing with uncertainty

Problem: fully describing the dynamics of the process that we are studying is almost always too complicated.

- **Tornado**: wind, humidity, air currents, climate change, ...
- **Medicine**: lifestyle, genetics, stress, external environment, ...
- **Elections**: social environments, stigmas, personal circumstances, ...

Solution: assume that the observed data is a realization of a random vector from some unknown distribution $f$!

Note: we surely won't be able to capture all the uncertainty, but by approximating reality we can still do way better than simply guessing!

# Data Generating Process

Goal: estimate the unknown distribution $f$ from which the data is drawn.

**Definition (Data Generating Process)**

Let our data $\boldsymbol{x} = (x_1, \ldots, x_n)$ be a realization from the random vector $\boldsymbol{X} = (X_1, \ldots, X_N)$ with distribution $f := f(\boldsymbol{x} \mid \theta_0)$. Then, $f_{\boldsymbol{X}}$ is referred to as the Data Generating Process (DGP) .

# Statistical Model

Idea: Formulate a set of candidate distributions that (hopefully) contains the DGP.

## Definition (Statistical Model)

A statistical model for $(X_1, \ldots, X_n)$ is a collection of probability distribution functions $\mathcal{M} = \{f(x \mid \theta) \mid \theta \in \Theta\}$, where $\Theta$ is a set and $\theta$ is an indexing parameter.

Simplification: While $\Theta$ can be any set, we often use external knowledge to restrict $\Theta$ and simplify the statistical analysis.

## Definition (Parametric models)

A statistical model is called parametric if there exists a $k \in \mathbb{N}$ such that $\Theta \subseteq \mathbb{R}^k$.

# Model specification

### Definition (Correct specification)

Let $\mathcal{M} = \{f(\boldsymbol{x} \mid \boldsymbol{\theta}) \mid \boldsymbol{\theta} \in \Theta\}$ be a statistical model for $X_1, \ldots, X_n$ with DGP $f(\boldsymbol{x} \mid \boldsymbol{\theta}_0)$. We say that $\mathcal{M}$ is correctly specified if $f(\boldsymbol{x} \mid \theta_0) \in \mathcal{M}$ or $\boldsymbol{\theta}_0 \in \Theta$.

Note: A correctly specified model contains the DGP!

Question: If our model is correctly specified, we may be able to "find the DGP" inside of it. But how?

Answer: Since the DGP is indexed by $\theta_0$, we should try to estimate $\theta_0$!

# Scope of this course

- Statistical models are always approximations of reality.
- The more we simplify reality,
  - the easier our life at steps 4 and 5 becomes,
  - the less realistic and generalizable our conclusions are.
- In this course, we (almost) exclusively limit ourselves to:
  - parametric models,
  - independent and identically distributed data,
  - correctly specified models.

# Scope of this course

- Statistical models are always approximations of reality.
- The more we simplify reality,
  - the easier our life at steps 4 and 5 becomes,
  - the less realistic and generalizable our conclusions are.
- In this course, we (almost) exclusively limit ourselves to:
  - parametric models,
  - independent and identically distributed data,
  - correctly specified models.

# Models for iid data

## Definition

If $X_1, \ldots, X_n$ are iid with unknown pdf $g(x)$, then we call $X_1, \ldots, X_n$ a random sample from the population $g(x)$.

- If the underlying data generating process (DGP) is iid, then the pdf splits

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} g(x_i)$$

- Then, we can specify a statistical model based on univariate distribution functions:

$$\mathcal{N} = \{g(x \mid \theta) \mid \theta \in \Theta\} \text{ instead of } \mathcal{M} = \left\{ \prod_{i=1}^{n} g(x_i \mid \theta) \mid \theta \in \Theta \right\}.$$

# Examples of statistical models: coin wager

*I have a coin and offer you a bet for thousand euro that the next coin flip will be heads.*

- Research question (broad): Should you take the bet?
- Research question (specific): Is the probability of flipping heads less than 50%?
- Data collection: You may flip the coin 100 times before deciding.
- Statistical Model: $\{\text{Bernoulli}(p) \mid p \in [0, 1]\}$.
- Parameter estimation: *coming up next week*
- Inference: Evaluate if $p_0 = \mathbb{P}(X_1 = 1) < 0.5$. Yes? Take the bet!

# Examples of statistical models: milk sales

*You own a store and want to optimize your inventory of milk based on demand and storage costs.*

- Research question (broad): How much milk should you buy every morning?
- Research question (specific): What is the minimum amount of milk I should buy such that no customer finds an empty store with 99% certainty?
- Data collection: Record daily number of customers for 3 months.
- Statistical Model: $\{\text{Binomial}(k, p) \mid k \in \mathbb{N}, p \in [0, 1]\}$.
- Parameter estimation: *coming up next week*
- Inference: Determine $m$ such that $\mathbb{P}(X_1 > m) \leq 0.01$.

# Examples of statistical models: celestial distance

*A physicist wants to find the distance between celestial bodies, but is only able to take inexact measurements.*

- Research question: What is the distance between the two celestial bodies?
- Data collection: Measure the distance $n$ times.
- Statistical Model: $\{\text{Normal}(\mu, \sigma) \mid \mu \geq 0, \sigma^2 > 0\}$.
- Parameter estimation: an intuitive estimator would be $\frac{1}{n}\sum_{i=1}^{n} X_i$.
- Inference: Can we construct a $L(\boldsymbol{X})$ and $U(\boldsymbol{X})$, such that $\mathbb{P}_{\mu_0}\left(L(\boldsymbol{X}) \leq \mu_0 \leq U(\boldsymbol{X})\right) = 0.95$?

# Scope of this course

- Statistical models are always approximations of reality.
- The more we simplify reality,
  - the easier our life at steps 4 and 5 becomes,
  - the less realistic and generalizable our conclusions are.
- In this course, we (almost) exclusively limit ourselves to:
  - parametric models,
  - independent and identically distributed data,
  - correctly specified models.

# Model validation

Important: throughout this course, we assume correct specification of our models.

However, in practice there is uncertainty about the model choice.

Validation: we briefly review two popular methods to validate the chosen model

1. Histograms: good as a visual first impression
2. QQ-plots: general visual tool based on quantiles

# The histogram function

- Let $\boldsymbol{x}$ denote the data, drawn from population $g$.
- Let $a_0 < a_1 < \ldots < a_m$ be an even partition of the range of the $x_i$, i.e. $a_j - a_{j-1} = c$ for $1 \leq j \leq m$.
- For any $y \in \mathbb{R}$, the histogram function $h_n$ is defined as

$$
\begin{aligned}
h_n(y) &= \sum_{j=1}^{m} \sum_{i=1}^{n} \mathbb{1}_{\{a_{j-1} < y \leq a_j\}} \mathbb{1}_{\{a_{j-1} < x_i \leq a_j\}} \\
&= \sum_{j=1}^{m} \mathbb{1}_{\{a_{j-1} < y \leq a_j\}} \left( \sum_{i=1}^{n} \mathbb{1}_{\{a_{j-1} < x_i \leq a_j\}} \right).
\end{aligned}
$$

# Histograms as density approximators

Idea: use histograms to approximate density.
Problem: a histogram does not integrate to 1, but to $c \cdot n$. (why?)
Solution: Rescale the histogram function:

$$\tilde{h}_n(y) = \frac{1}{cn} \sum_{j=1}^{m} \sum_{i=1}^{n} \mathbb{1}_{\{a_{j-1} < y \leq a_j\}} \mathbb{1}_{\{a_{j-1} < x_i \leq a_j\}}$$

Motivation: If $n$ and $m$ are large, then the histogram can give a good approximation of the density $g$. To motivate this, take a $y \in (a_{j-1}, a_j]$. Then,

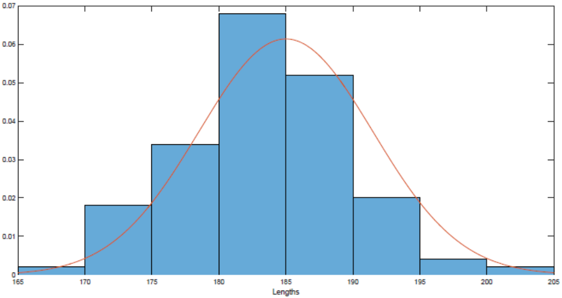
$$\tilde{h}_n(y) = \frac{1}{cn} \sum_{i=1}^{n} \mathbb{1}_{\{a_{j-1} < x_i \leq a_j\}} \overset{\text{(i)}}{\approx} \frac{1}{c} P(a_{j-1} < X_1 \leq a_j) = \frac{1}{c} \int_{a_{j-1}}^{a_j} g(x)dx \overset{\text{(ii)}}{\approx} g(y),$$

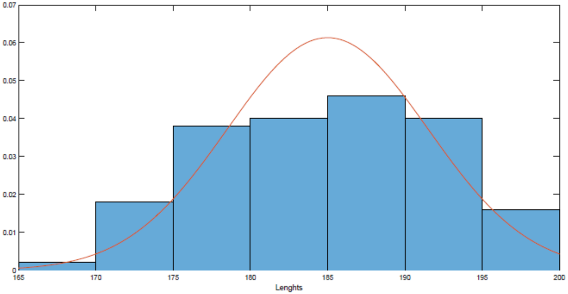
where (i) follows from LLN and (ii) holds if $g$ does not vary too much on $(a_{j-1}, a_j]$.

# Histogram examples

Disadvantage: Histograms tend to require a lot of data points to provide good approximations and are sensitive to the bin width.

Example: Below are two histograms based on $n = 100$ draws of the Normal$(185, 36)$.



# QQ-plots

You suspect that the random sample $X_1, \ldots, X_n$ has population pdf $h$ and CDF $H$. Let $g$ and $G$ denote the true pdf and CDF.

Goal: Check whether $h = g$ and $H = G$.

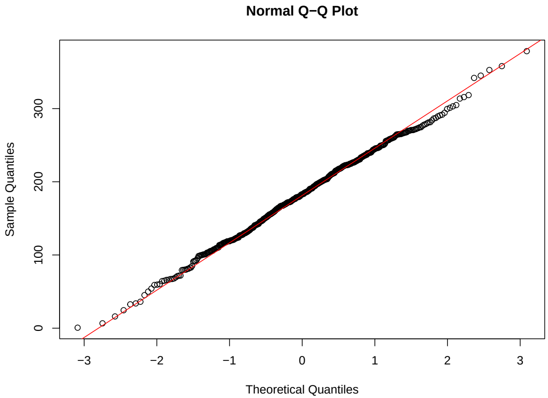
Idea: Compare the quantiles predicted by $H$ to the empirical quantiles of the observed data.

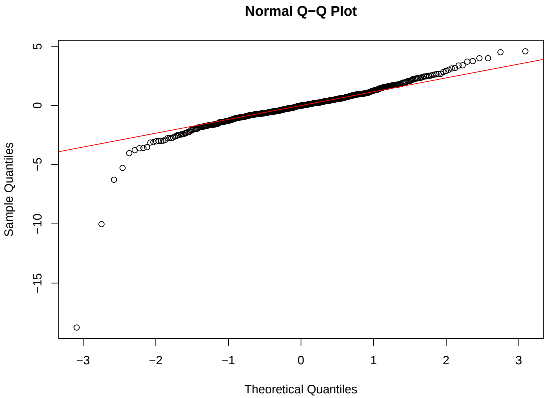
Approach: Order the observed data $x_{(1)} \leq \ldots \leq x_{(n)}$ and plot the points

$$\left(x_{(k)}, H^{-1}\left(\frac{k}{n+1}\right)\right).$$

Interpretation: If $G = H$, the points should lie on a straight line.

# QQ-plot: examples




Normal QQ-plot with $X \sim N(185, 63)$.

Normal QQ-plot with $X \sim t(3)$.

# QQ-plot: motivation

- Let $Y$ be a random variable with distribution $g$. Then by symmetry we have that

$$
\begin{aligned}
P(Y \leq X_{(1)}) &= P(X_{(1)} < Y \leq X_{(2)}) = \cdots \\
&= P(X_{(n-1)} < Y \leq X_{(n)}) = P(Y > X_{(n)}) = \frac{1}{n+1}.
\end{aligned}
$$

- It follows that the order statistics can be used as an approximation for the quantiles as for each $1 \leq k \leq n$ we have

$$
\begin{aligned}
P\left(Y \leq X_{(k)}\right) = \frac{k}{n+1} \quad &\Rightarrow \quad G(x_{(k)}) = P\left(Y \leq x_{(k)}\right) \approx \frac{k}{n+1} \\
&\Rightarrow \quad x_{(k)} \approx G^{-1}\left(\frac{k}{n+1}\right)
\end{aligned}
$$

# Families of distributions

Recall: Statistical models are collections of distribution.

Note: Certain sets or "families" of distributions have special characteristics that help in building a statistical model.

Special cases: We study the following two families of distributions:

- The location-scale family: flexible method to define an interpretable collection of distribution.
- The exponential family: simplifies calculations and has nice theoretical properties.

# The location-scale family

Intuition: A location-scale family is created by

1. taking *any* pdf,
2. shifting its graph along the x-axis, and
3. contracting/expanding the graph while retaining its basic shape.

## Definition (3.5.5)

Let $g(x)$ be any pdf. Then,

$$g(x|\mu,\sigma) = \left\{ \frac{1}{\sigma} g\left(\frac{x-\mu}{\sigma}\right) \mid \mu \in \mathbb{R}, \sigma > 0 \right\},$$

is called the *location-scale family* $g$.

# Properties of the location-scale family

Note: A location-scale family can also be characterized in terms of *cumulative* distribution functions.

**Lemma**

*Let* $g(x|\mu,\sigma)$ *be a member of the location-scale family* $g$*. Then, the cdf of* $g(x|\mu,\sigma)$ *satisfies* $G(x|\mu,\sigma) = G\left(\frac{x-\mu}{\sigma}\right)$*, where* $G$ *is the cdf of* $g$*.*

**Proof.**

*Tutorial exercise* $\square$

# Random variables in a location-scale family

## Lemma

*Let $Y$ be a random variable with cdf $H(x)$, let $\mu \in \boldsymbol{R}$ and $\sigma > 0$ and define $Y_{\mu,\sigma} = \mu + \sigma Y$. Then $Y_{\mu,\sigma}$ has cdf $H(x \mid \mu, \sigma)$.*

## Proof.

$$P(Y_{\mu,\sigma} \leq y) = P(\mu + \sigma Y \leq y) = P\left(Y \leq \frac{y-\mu}{\sigma}\right) = H\left(\frac{y-\mu}{\sigma}\right).$$

□

## Example

Suppose that $Y \sim N(0,1)$. Then we know that $\mu + \sigma Y \sim N(\mu, \sigma^2)$ and thus the location-scale family of $N(0,1)$ is the set of all normal distributions.

# QQ-plots and the location-scale family

Important: QQ-plots can be used to check whether the data generating process is a member of a certain location-scale family.

Suppose that the data is a sample from $g(x|\mu, \sigma)$, which is a member of the location-scale family $h$ with CDF $H$.

Then, it follows that

$$\frac{k}{n+1} \approx G(x_{(k)}|\mu, \sigma) = H\left(\frac{x_{(k)} - \mu}{\sigma}\right) \Rightarrow H^{-1}\left(\frac{k}{n+1}\right) \approx -\frac{\mu}{\sigma} + \frac{1}{\sigma} x_{(k)}.$$

Hence, the points $(x_{(k)}, H^{-1}\left(\frac{k}{n+1}\right)$ should follow a straight line with intercept $-\mu/\sigma$ and slope $1/\sigma$.

Conclusion: the location-scale family of $h$ is a correctly specified statistical model!

# The exponential family

Another important family of distributions in statistics is the exponential family.

## Definition (3.4.1)

A family of pdfs or pmfs is called an exponential family if it can be expressed as

$$g(x|\boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta})\exp\left(\sum_{i=1}^{k} w_i(\boldsymbol{\theta})t_i(x)\right), \tag{1}$$

where $h(x) \geq 0$, $c(\boldsymbol{\theta}) \geq 0$, $t_1(x), \ldots, t_k(x)$ are real valued functions of $x$ that *do not depend* on $\boldsymbol{\theta}$, and $w_1(\boldsymbol{\theta}), \ldots, w_k(\boldsymbol{\theta})$ are real-valued functions of the parameter(s) $\boldsymbol{\theta}$.

Important: The definition should hold over the complete real line! Indicator functions may be needed.

# The exponential family: Binomial distribution

## Example (3.4.1)

Let $X \sim \text{Binomial}(n, p)$ with $n$ known and pdf given by

$$g(x|n,p) = \binom{n}{x} p^x (1-p)^{n-x}, \quad 0 < p < 1.$$

Then, $g(x|n,p)$ is a member of the exponential family, which becomes clear upon rewriting

$$\begin{aligned} g(x|n,p) &= \binom{n}{x} p^x (1-p)^{n-x} = \binom{n}{x} (1-p)^n \left(\frac{p}{1-p}\right)^x \\ &= \binom{n}{x} (1-p)^n \exp\left(\log\left(\frac{p}{1-p}\right) x\right). \end{aligned}$$

such that $h(x) = \binom{n}{x}$, $c(\boldsymbol{\theta}) = (1-p)^n$, $w_1(\boldsymbol{\theta}) = \log\left(\frac{p}{1-p}\right)$ and $t_1(x) = x$.

# The exponential family: relevance

Relevance: The exponential family is important, because

- its member distributions are "well-behaved"
- calculating moments is simplified
- parts of the data can be discarded

Sufficiency: The part $h(\boldsymbol{x})$ in the decomposition does not depend on the parameters. It can therefore be safely ignored when estimating parameters (see sufficiency in Week 3).

# Binomial distribution with $n$ and $p$ unknown

Important: when the support of the distribution depends on the parameter, the distribution cannot be a member of the exponential family.

## Example

Let $X \sim \text{Binomial}(k, p)$, with *both* $k$ and $p$ unknown. Then the pmf of $X$ is given by

$$f(x \mid k, p) = \binom{k}{p} p^x (1-p)^{k-x} \mathbb{1}_{\{0,1,\ldots,k\}}(x).$$

Since the indicator function cannot be split into an $h(x)$ and $c(\boldsymbol{\theta})$ function, nor can it be represented by an exponential function, this is not a member of the exponential family.