A Primer in Econometric Theory

Lecture 7: Estimators

John Stachurski
Lectures by Akshay Shanker

March 26, 2017

Probability and Statistics

Probability theory

- aim to deduce the likelihood of different outcomes based on known probability distributions

Statistics

- infer unknown probability distributions from outcomes we observe

Start in setting where successive observations are IID

The fundamental problem of econometrics and statistics:

Problem

We observe independent Z*-valued draws* $\mathbf{z}_1, \ldots, \mathbf{z}_N$ *from a common but unknown distribution* $P \in \mathscr{P}$*, where* $\mathscr{P}$ *is a class of distributions on* Z*. We wish to infer some features of* P *from this sample.*

The set $\mathscr{P}$ can be anything, including the set of all distributions on the outcome space

A task of economic theory is to restrict $\mathscr{P}$ narrow the set of distributions to search over

Example. Benhabib et al. (2015) study the wealth distribution in a model with idiosyncratic capital income risk

The model predicts that the wealth distribution will have a Pareto right tail

Some notation:

- Z is the **outcome space** where each **observation** $\mathbf{z}_n$ takes values
- $\mathbf{z}_{\mathcal{D}}$ denotes the **sample** or **data set** $(\mathbf{z}_1, \ldots, \mathbf{z}_N)$
- $Z_{\mathcal{D}} := \times_{n=1}^{N} Z$ is the **sample space** in which $\mathbf{z}_{\mathcal{D}}$ takes values
- $P_{\mathcal{D}} := \mathcal{L}(\mathbf{z}_{\mathcal{D}})$ is the **joint distribution of the sample**

Note calling $Z_{\mathcal{D}}$ the sample space is overloading terminology used to describe probability spaces

Assume IID data, so the joint distribution $P_{\mathcal{D}}$ will be the Nth product of P

Example. Let $x_1, \ldots, x_N$ be observations of labor income from a given population

We model $x_1, \ldots, x_N$ as IID draws from a common univariate distribution P

In our terminology:

- x_n is an observation
- the outcome space Z is $\mathbb{R}$
- the sample space $Z_{\mathcal{D}}$ is $\mathbb{R}^N$
- the sample $\mathbf{z}_{\mathcal{D}}$ is the vector $(x_1, \ldots, x_N)$

Features of P we might wish to learn about:

- the mean and higher moments of P,
- measures of dispersion, such as the variance, or properties of tails,
- the median and other quantiles, and
- P itself, or the density of P if it exists

Example. Suppose we wish to learn about the relationship between profitability and R&D spending within a group of firms

Let $\mathbf{z}_n = (x_n, y_n)$ be an observation of these two quantities at the nth firm

Treat the observations as IID across firms, with common marginal distribution $P = \mathcal{L}(\mathbf{z}_n)$

The outcome space is $Z = \mathbb{R}^2$, and the sample space:

$$Z_{\mathcal{D}} := \mathbb{R}^2 \times \cdots \times \mathbb{R}^2 = \mathbb{R}^{2 \times N}$$

Marginal distribution P of the observations is now multivariate – new features:

- correlations across coordinates of P,
- the variance–covariance matrix associated with P, and
- parameters controlling dependence when, say, P is modeled via a copula over certain marginals

Features of P

Define a **feature** of P to be an object of the form

$$\gamma(P) \quad \text{for some} \quad \gamma\colon \mathscr{P} \to S$$

When P is understood, we'll write $\gamma(P)$ as γ

Examples for univariate P routinely estimated in econometric studies:

- $\gamma(P) = \int s^k P(\mathrm{d}s)$, the kth moment of P
- $\gamma(P) = \inf\{s \in \mathbb{R} : P(-\infty, s] \geq 1/2\}$, the median of P
- $\gamma(P) = P$, when we want to estimate P itself
- $\gamma(P) =$ the density of P when P is absolutely continuous

If P is multivariate over $\mathbf{z} = (\mathbf{x}, y)$, then a feature of interest is the regression function $f^*(\mathbf{x}) := \mathbb{E}[y \mid \mathbf{x}]$

- the function is uniquely determined by P (recall our discussion in §5.2.5)

Parametric and Nonparametric Classes

We assumed above that the unknown distribution belongs to some class $\mathscr{P}$

A class of distributions is called a **parametric class** if it can be indexed by finitely many parameters

$$\mathscr{P} = \{P_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta}\in\Theta} =: \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\} \quad \text{for some } \Theta \subset \mathbb{R}^K$$

A class of distributions is called **nonparametric** if it is not parametric

Example. Let $\mathscr{P}$ be the set of all univariate normal distributions with positive variance:

$$\mathscr{P} := \left\{ \text{all } p \ \text{ s.t. } \ p(s) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{(s-\mu)^2}{2\sigma^2} \right\} \right.$$

$$\left. \text{for some } \ \mu \in \mathbb{R}, \ \sigma > 0 \right\}$$

The set $\mathscr{P}$ is an example of a parametric class

- the parameters are $\boldsymbol{\theta} = (\mu, \sigma)$
- particular choice of parameters determines (parameterizes) an element of the class

Example. Suppose outcome space Z is finite, containing J elements

Every distribution P on Z can be represented by $J-1$ parameters (the probability p_j of each outcome but the last)

Hence any family $\mathscr{P}$ of distributions on Z is a parametric class

Example. The set of all distributions on $\mathbb{R}$ cannot be parameterized by a finite vector of parameters because the space of distributions is infinite dimensional

Hence $\mathscr{P} :=$ all distributions on $\mathbb{R}$ is a nonparametric class

Example. Let $\mathscr{P}$ be the set of all absolutely continuous distributions on $\mathbb{R}$ with finite second moment:

$$\mathscr{P} := \left\{ \text{all} p \colon \mathbb{R} \to \mathbb{R} \text{ s.t. } p \geq 0, \int p(s)\,\mathrm{d}s = 1, \int s^2 p(s)\,\mathrm{d}s < \infty \right\}$$

The set is ....

Traditional methods of inference:

- assume data generated by an unknown element $P_{\boldsymbol{\theta}}$ of a parametric class $\mathscr{P}$
- estimate $\boldsymbol{\theta}$ using the data – the estimate of $\boldsymbol{\theta}$ is $\hat{\boldsymbol{\theta}}$
- plug $\hat{\boldsymbol{\theta}}$ back into the parametric class to obtain an estimate $P_{\hat{\boldsymbol{\theta}}}$ of $P_{\boldsymbol{\theta}}$

In parametric settings, the feature γ that we are most interested in estimating is the parameter vector $\boldsymbol{\theta}$

If we have a good estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$, we can estimate any feature $\gamma = \gamma(P_{\boldsymbol{\theta}})$ via $\hat{\gamma} = \gamma(P_{\hat{\boldsymbol{\theta}}})$

Common usage – refer to the $\boldsymbol{\theta}$ associated with the $P_{\boldsymbol{\theta}}$ that generates the data as the **true value** of the parameter vector

- an assumption, not a “truth.”!

We *assume* data generated by some member of a parametric class

Assumption can be completely false

Why did we first formulate the problem as estimation of features?

- because it is suboptimal to always restrict ourselves to parametric assumptions on $\mathscr{P}$

Discussion of relative merits of parametric and nonparametric estimation in chapter 14

Statistics and Estimators

A **statistic** is any $\mathscr{B}$-measurable function

$$T\colon Z_{\mathcal{D}} \to S$$

that can be evaluated once the data $\mathbf{z}_{\mathcal{D}}$ are observed

As in (10), the set S is left arbitrary to accommodate all the possible features that we might wish to estimate

Sometimes write T as T_N to emphasize dependence on the sample size

An **estimator** is a statistic used to infer some feature $\gamma(P)$ of an unknown distribution P

- statistic becomes an estimator when paired with and compared to a feature of the distribution

Nothing in the definition of an estimator that implies it will be a good estimator of the target feature

Example. If the feature γ we wish to infer is the mean of the marginal distribution P of IID data $x_1, \ldots, x_N$, then the most common estimator is the **sample mean**

$$\bar{x}_N := \frac{1}{N} \sum_{n=1}^{N} x_n$$

Formally, $\bar{x}_N$ is the mapping from $Z_{\mathcal{D}} = \mathbb{R}^N$ to $S = \mathbb{R}$ defined by

$$\mathbf{z}_{\mathcal{D}} = (x_1, \ldots, x_N) \mapsto T(x_1, \ldots, x_N) = \frac{1}{N} \sum_{n=1}^{N} x_n \in \mathbb{R}$$

This mapping is regarded as an estimator of the unknown mean $\gamma(P) = \int s P(\mathrm{d}s)$

Example. The sample mean is not the only way to estimate the mean. For example, we could also use the so-called **mid-range estimator**

$$m_N := \frac{\min_n x_n + \max_n x_n}{2}$$

Another option is a **truncated sample mean**, where values x_n with $|x_n| \geq r$ are truncated for some specified value of r. The truncated sample mean is often used to estimate location parameters in heavy tailed distributions

Example. Given sample $x_1, \ldots, x_N$, let y_n be the nth largest observation of the sample. If N is the odd number $2m+1$, the **sample median** is defined as y_{m+1}. If $N = 2m$, the sample median is $0.5(y_m + y_{m+1})$.

Example. A common estimator of the kth moment $\int s^k P(\mathrm{d}s)$ of P is the kth **sample moment** $\frac{1}{N}\sum_{n=1}^{N} x_n^k$.

Example. A common estimator of the variance of P is the **sample variance**

$$s_N^2 := \frac{1}{N} \sum_{n=1}^{N} (x_n - \bar{x}_N)^2$$

The standard deviation is usually estimated using **sample standard deviation**

$$s_N := \sqrt{s_N^2} = \left[\frac{1}{N} \sum_{n=1}^{N} (x_n - \bar{x}_N)^2 \right]^{1/2}$$

Example. Given bivariate data $\mathbf{z}_{\mathcal{D}} = ((x_1, y_1), \ldots, (x_N, y_N))$, the **sample covariance** is the statistic

$$\frac{1}{N} \sum_{n=1}^{N} (x_n - \bar{x}_N)(y_n - \bar{y}_N)$$

The **sample correlation** is the sample covariance divided by the product of the two sample standard deviations. With some rearranging, this becomes

$$\frac{\sum_{n=1}^{N} (x_n - \bar{x}_N)(y_n - \bar{y}_N)}{\sqrt{\sum_{n=1}^{N} (x_n - \bar{x}_N)^2 \sum_{n=1}^{N} (y_n - \bar{y}_N)^2}}$$

Example. In the case where our observations are vectors $\mathbf{x}_1, \ldots, \mathbf{x}_N$ in $\mathbb{R}^K$, the sample mean is the random vector defined by

$$\bar{\mathbf{x}}_N := \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n$$

The variance–covariance matrix is most often estimated with the **sample variance–covariance matrix**

$$\hat{\mathbf{\Sigma}}_N := \frac{1}{N} \sum_{n=1}^{N} \left[(\mathbf{x}_n - \bar{\mathbf{x}}_N)(\mathbf{x}_n - \bar{\mathbf{x}}_N)^\top \right]$$

Empirical Distribution

The **empirical distribution** of a given Z-valued sample $\mathbf{z}_{\mathcal{D}} = (\mathbf{z}_1, \ldots, \mathbf{z}_N)$ is the discrete distribution on Z that puts equal probability $1/N$ on each sample point $\mathbf{z}_n$

$\hat{P}_N$ assigns to each Borel set $B \subset Z$ the number

$$\hat{P}_N(B) = \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}\{\mathbf{z}_n \in B\}$$

This is just the fraction of the sample that falls in B

The expectation of a function h with respect to $\hat{P}_N$ is

$$\int h(\mathbf{s}) \hat{P}_N(\mathrm{d}\mathbf{s}) = \frac{1}{N} \sum_{n=1}^{N} h(\mathbf{z}_n)$$

Example. Let $\mathbf{z}_n$ be the scalar x_n. The sample mean can be expressed in terms of the empirical distribution as

$$\bar{x}_N = \frac{1}{N}\sum_{n=1}^{N} x_n = \int s\hat{P}_N(\mathrm{d}s)$$

In other words, the sample mean is the mean of the empirical distribution.

The empirical distribution is a statistic, mapping observations $\mathbf{z}_1, \ldots, \mathbf{z}_N$ into $\hat{P}_N = \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}\{\mathbf{z}_n \in \cdot\}$, a random element of the set of all distributions on Z

If we think of $\mathbf{z}_1, \ldots, \mathbf{z}_N$ as independent draws from common but unknown distribution P, then $\hat{P}_N$ becomes an estimator of P

If the feature of P we want to infer is P itself, the simplest natural estimator is the empirical distribution

Why not always try to infer P itself?

General principle of inference with limited information

- try to avoid first solving a more general problem as an intermediate step

Distributions more complicated objects than real numbers

- if we care only about the median of a distribution, say, it might be best to try to discover this single value first

With scalar data $x_1, \ldots, x_N$ we can visualize the empirical distribution by plotting its CDF

The CDF of $\hat{P}_N$ will be denoted in what follows by $\hat{F}_N$

$$\hat{F}_N(s) = \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}\{x_n \leq s\} \qquad (s \in \mathbb{R})$$

The CDF $\hat{F}_N$ is called the **empirical cumulative distribution function**, or ECDF, corresponding to the sample

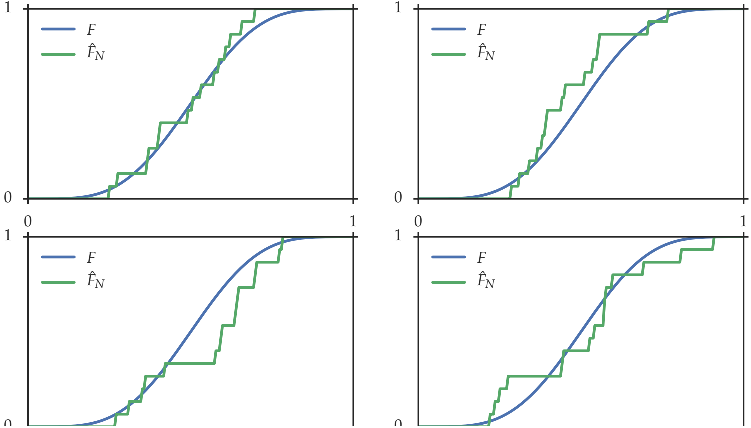

Figure: F and four observations of $\hat{F}_N$ when $N = 15$. Draws from a Beta$(5, 5)$ distribution

Convergence

The empirical distribution is asymptotically an excellent estimator of P

If $\mathbf{z}_1, \ldots, \mathbf{z}_N$ are IID with common distribution P and $\hat{P}_N$ is the empirical distribution, then, by the law of large numbers,

$$\hat{P}_N(B) = \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}\{\mathbf{z}_n \in B\} \xrightarrow{p} \mathbb{P}\{\mathbf{z}_n \in B\} = P(B)$$

for any Borel set B

Specializing to the scalar case with $B := (-\infty, s]$, we have $\hat{F}_N(s) \xrightarrow{p} F(s)$ for any $s \in \mathbb{R}$, where F is the CDF of P

Stronger result:

Theorem. (8.1.1)(**Glivenko–Cantelli**) Let $x_1, \ldots, x_N$ be IID draws from F. If $\hat{F}_N$ is the corresponding ECDF, then

$$\|F - \hat{F}_N\|_\infty \xrightarrow{p} 0 \quad \text{as} \quad N \to \infty$$

a.k.a. fundamental theorem of statistics

The supremum norm:

$$\|F - \hat{F}_N\|_\infty := \sup_{s \in \mathbb{R}} |\hat{F}_N(s) - F(s)|$$

Roughly, the maximal deviation over the domain

In fact the convergence occurs “almost surely,” which is a stronger notion than in probability

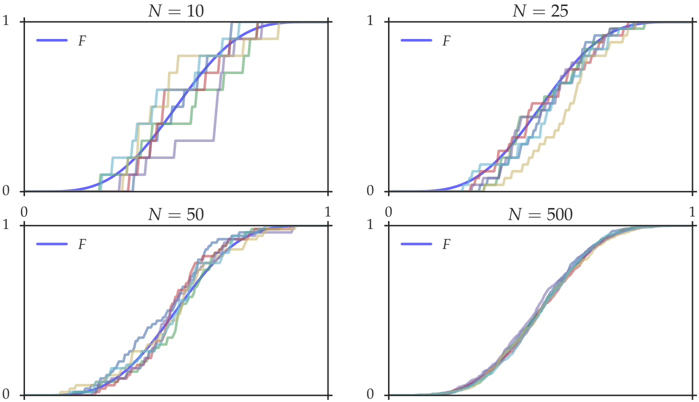

Figure: Realizations of $\hat{F}_N$ with four different sample sizes. Independent draws from Beta(5,5)

Theorem 8.1.1: in the IID setting, with infinite amount of data, we can learn the underlying distribution without any assumptions

Once we know the distribution P, we know any feature $\gamma = \gamma(P)$

Does knowing we can learn the underlying distribution in the limit offer any solution to the problem of estimation? No:

- in practice, we only ever have a finite amount of data
- the estimation problem was about inference – there's no need to generalize if we know the full population
- with finite amount of data, empirical distribution treats the sample like the unknown distribution– extreme form of over-fitting

Identification

A class of distributions $\mathscr{P} = \{P_{\boldsymbol{\theta}}\}$ indexed by $\boldsymbol{\theta} \in \Theta$ is called **identifiable** if the map $\boldsymbol{\theta} \mapsto P_{\boldsymbol{\theta}}$ is one-to-one on Θ

Example. Recall the set of all univariate normal distributions with positive variance:

$$\mathscr{P} := \left\{ \text{all } p \ \text{ s.t. } \ p(s) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{(s-\mu)^2}{2\sigma^2} \right\} \right.$$

$$\left. \text{for some } \ \mu \in \mathbb{R},\ \sigma > 0 \right\}$$

This class is identifiable.

We can show that if (μ_a, σ_a) and (μ_b, σ_b) are distinct vectors, then the distributions $\mathrm{N}(\mu_a, \sigma_a^2)$ and $\mathrm{N}(\mu_b, \sigma_b^2)$ differ at at least one point

Identifiability means the parameter vector associated with the unknown distribution can eventually be distinguished from the data:

- suppose $\{P_{\boldsymbol{\theta}}\}$ is identified on Θ and nature generates an infinite sequence of observations $\{\mathbf{z}_n\}$ from $P = P_{\boldsymbol{\theta}}$
- let $\boldsymbol{\theta}'$ be any other vector in Θ and let $P' = P_{\boldsymbol{\theta}'}$
- by identifiability, there exists at least one Borel set B with $P(B) \neq P'(B)$
- since the empirical distribution $\hat{P}_N(B)$ converges to $P(B)$, we can in the limit conclude that the data are not generated by P'

The Sample Analogue Principle

Most of the estimators defined above can be derived from the **plug-in method**, also reffered to as "analogue estimation" or **sample analogue principle**:

$$\text{to estimate } \gamma(P), \ \text{ use } \ \gamma(\hat{P}_N)$$

$\hat{P}_N$ is the empirical distribution constructed from the sample

We replace the unknown distribution P with the observable distribution $\hat{P}_N$ and then evaluate $\gamma(\hat{P}_N)$

Example. Let $x_1, \ldots, x_N$ be draws from unknown distribution P. Suppose we want to estimate the mean $\gamma(P) := \int s P(\mathrm{d}s)$. The sample analogue principle tells us to replace P with $\hat{P}_N$, which gives

$$\gamma(\hat{P}_N) = \int s \hat{P}_N(\mathrm{d}s) = \bar{x}_N$$

Thus the sample mean is the estimator of the mean produced by the sample analogue principle

The kth sample moment applies the same principle to estimate the kth moment

Example. When it exists, the variance of P can be written as

$$\sigma^2 = \gamma(P) = \int \left[t - \int s P(\mathrm{d}s) \right]^2 P(\mathrm{d}t)$$

Applying the sample analogue principle leads to the estimator

$$\int \left[t - \int s \hat{P}_N(\mathrm{d}s) \right]^2 \hat{P}_N(\mathrm{d}t) = \frac{1}{N} \sum_{n=1}^{N} (x_n - \bar{x}_N)^2$$

This is precisely the sample variance

Other estimation methods that can be obtained as special cases of the sample analogue principle include:

- least squares regression
- maximum likelihood
- the method of moments
- generalized method of moments

Best Linear Prediction

Recall the best linear prediction problem: α and β are chosen to minimize $\mathbb{E}\left[(y-\alpha-\beta x)^2\right]$

Letting P be the distribution of (x, y):

$$(\alpha^*, \beta^*) = \gamma(P) := \underset{\alpha,\beta\in\mathbb{R}}{\operatorname{argmin}} \int [(t-\alpha-\beta s)^2] P(\mathrm{d}s, \mathrm{d}t)$$

Corresponding statistical problem: produce the best linear predictor based only on a sample $\mathbf{z}_{\mathcal{D}} = ((x_1, y_1), \ldots, (x_N, y_N))$ of observations from P

Given $\mathbf{z}_{\mathcal{D}}$, form the empirical distribution:

$$\hat{P}_N(B) = \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}\{(x_n, y_n) \in B\}$$

This is the simple linear bivariate least squares problem. The minimizers are

$$\hat{\beta}_N = \frac{\sum_{n=1}^{N} (x_n - \bar{x}_N)(y_n - \bar{y}_N)}{\sum_{n=1}^{N} (x_n - \bar{x}_N)^2} \quad \text{and} \quad \hat{\alpha}_N = \bar{y}_N - \hat{\beta}_N \bar{x}_N$$

Limitations

Sample analogue principle can fail. For example:

- let $\mathscr{P}$ be the set of absolutely continuous distributions on $\mathbb{R}$
- data generated by P with density

$$\gamma(P) = DP \tag{1}$$

  where $DP :=$ the derivative of the CDF F of P
- let $\hat{P}_N$ be the empirical distribution from a sample

Plug empirical distribution into the right-hand side of (1): since $\hat{F}_N$ is a step function, the derivative is zero everywhere, except at a finite number of jump points where the derivative is undefined

Regularization

One interpretation of limitation: we need to combine the data with some kind of **regularization**

Regularization:

- penalize complexity or impose some kind of “smoothing” a priori
- not grant the empirical distribution equal status with the unknown true distribution
- regard the empirical distribution only as partial information, and seek to combine it with some form of prior information or external theory

Empirical Risk Minimisation

Empirical risk minimization, or ERM: the terminology and main concepts come from the machine learning literature

Nonetheless, many standard estimators in econometrics are special cases of ERM, including maximum likelihood and least squares

We observe an input $\mathbf{x} \in \mathbb{R}^K$ to a system, followed by a scalar output y

Both are random variables and the joint distribution of $\mathbf{z} := (\mathbf{x}, y)$ is P

Our aim is to predict new output values from observed input values

We'll do this by choosing a function f such that $f(\mathbf{x})$ is our prediction of y once $\mathbf{x}$ is observed

In the machine learning literature, f is called a **prediction rule**

In economics, f is called a **strategy** or **policy function**

Incorrect prediction incurs a loss $L(y, f(\mathbf{x}))$

The function L called the **loss function**

Common choices:

- the **quadratic loss function** $L(y, f(\mathbf{x})) = (y - f(\mathbf{x}))^2$
- the **absolute deviation loss function** $L(y, f(\mathbf{x})) = |y - f(\mathbf{x})|$
- the **discrete loss function** $L(y, f(\mathbf{x})) = \mathbb{1}\{y \neq f(\mathbf{x})\}$

Given loss function L, choose f to minimize the **prediction risk** or prediction error, defined as the expected loss

$$R(f) := \mathbb{E}\, L(y, f(\mathbf{x}))$$

Expectation computed using the joint distribution P of $(\mathbf{x}, y)$

Example. Let L be the quadratic loss function

As shown in §5.2.5, the minimizer of $R(f)$ is the regression function, defined at $\mathbf{x}$ by $f^*(\mathbf{x}) = \mathbb{E}\,[y \,|\, \mathbf{x}]$

For any alternative policy g we have

$$R(g) = R(f^*) + \mathbb{E}\,[(f^*(\mathbf{x}) - g(\mathbf{x}))^2] \tag{2}$$

With quadratic loss, good prediction equates to choosing g to be close to the regression function, minimizing the second term on the right-hand side of (2)

The term $R(f^*)$ represents a lower bound for prediction risk

In a statistical setting, we cannot evaluate $\mathbb{E}_P$

We do have access to data $\mathbf{z}_1, \dots, \mathbf{z}_N$, where each pair $\mathbf{z}_n = (\mathbf{x}_n, y_n)$ is an independent draw from P

Apply the sample analogue principle, replace P in the risk function with $\hat{P}_N$:

$$R_{\text{emp}}(f) := \mathbb{E}_{\hat{P}_N} L(y, f(\mathbf{x})) = \frac{1}{N} \sum_{n=1}^{N} L(y_n, f(\mathbf{x}_n))$$

.....this is called empirical risk

The problem to solve:

$$\hat{f} = \operatorname*{argmin}_{f \in \mathcal{H}} R_{\text{emp}}(f) \tag{3}$$

Restrict domain to a class of functions $\mathcal{H}$ called the **hypothesis space**

Should we set $\mathcal{H}$ to be the set of all functions $f\colon \mathbb{R}^K \to \mathbb{R}$?

If the risk-minimizing function $f^* := \operatorname{argmin}_f R(f)$ is not in $\mathcal{H}$, then the solution to (3) is not equal to f^* and we are making a suboptimal choice

More soon, but this reasoning false:

- we are solving a complex problem on the basis of limited information
- better to be restrictive in choice of hypothesis space
- key point: minimising empirical risk not the same as minimising prediction risk

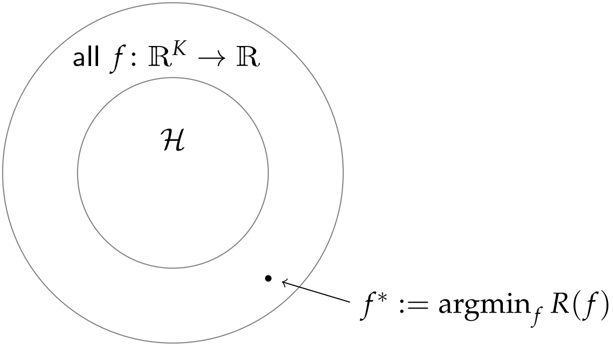
all $f\colon \mathbb{R}^K \to \mathbb{R}$
$\mathcal{H}$
$f^* := \operatorname{argmin}_f R(f)$

Figure: Choosing the hypothesis space

Example. Specializing ERM problem to scalar x and quadratic loss function $L(y, f(x)) = (y - f(x))^2$:

$$\min_{f \in \mathcal{H}} \sum_{n=1}^{N} (y_n - f(x_n))^2$$

This is the **least squares problem**

We can specialize $\mathcal{H}$ to be the set of affine functions

$$\mathcal{H}_\ell := \{ \text{ all functions of the form } \ell(x) = \alpha + \beta x \} \tag{4}$$

Problem becomes the **simple linear least squares problem**

$$\min_{\ell \in \mathcal{H}_\ell} \sum_{n=1}^{N} (y_n - \ell(x_n))^2 = \min_{\alpha, \beta} \sum_{n=1}^{N} (y_n - \alpha - \beta x_n)^2 \qquad (5)$$

Two approaches, same solution

Sample analouge principle

- plug in empirical distribution and solve the best linear preduction problem

Empirical risk minimisation

- trying to find the best predictor
- at the same time, we restrict ourselves to linear approximations in recognition of the limited information we are basing our prediction on

Quantile Regression

Quantile regression: estimate a quantile of a given distribution based on a set of predictive variables

For example, quantile regression to estimate quantiles of CEO compensation as function of the market value of their firms (Koenker and Hallock 2001)

Quantile regression can be regarded as a special case of empirical risk minimization

Let F be a strictly increasing CDF on $\mathbb{R}$ and let $\tau \in (0, 1)$ be given

Recall from §4.2.6 in ET that the τth quantile of F is the ξ that solves $F(\xi) = \tau$

We can define the τth quantile as the solution to:

$$\min_{\xi \in \mathbb{R}} \mathbb{E}\, L_\tau(y, \xi)$$

where y is a random variable with distribution F and

$$L_\tau(y, \xi) := |(y - \xi)(\tau - \mathbb{1}\{y < \xi\})|$$

(See exercise 8.5.6 in ET)

We now want to estimate the τth quantile of F using some input variable x (e.g., a CEO compensation quantile as a function of firm size)

Frame the search for a suitable function as a problem of minimizing the prediction

$$R(f) := \mathbb{E}\, L_\tau(y, f(x))$$

Use the principle of empirical risk minimization with data set $(x_1, y_1), \ldots, (x_N, y_N)$

We get $\min_{f \in \mathcal{H}} \sum_{n=1}^{N} L_\tau(y_n, f(x_n))$, where $\mathcal{H}$ as the hypothesis space

When $\mathcal{H} = \mathcal{H}_\ell$, the set of affine functions, the ERM problem is

$$\min_{\alpha,\beta} \sum_{n=1}^{N} |(y_n - \alpha - \beta x_n)(\tau - \mathbb{1}\{y_n < \alpha + \beta x_n\})|$$

Standard expression for the quantile regression problem

If $\tau = 0.5$, then the objective function is proportional to $\sum_{n=1}^{N} |y_n - \alpha - \beta x_n|$ — called **median regression** or **least absolute deviation regression**

The Choice of Hypothesis Space

Fact. (8.2.1) Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two hypothesis spaces. If $\hat{f}_i$ is the empirical risk minimizer over $\mathcal{H}_i$ as defined in (3), then

$$\mathcal{H}_1 \subset \mathcal{H}_2 \implies R_{\text{emp}}(\hat{f}_1) \geq R_{\text{emp}}(\hat{f}_2)$$

We can always decrease empirical risk by increasing the hypothesis space

Empirical Versus Prediction Risk

Prediction risk of predictor f measures the **out-of-sample fit** of f:

- expected performance of f when confronted with new data
- statistical procedure that produces a predictor f with low prediction risk means that we have succeeded in meeting the goal of *generalizing* from data
- prediction risk unobservable

Empirical risk measures **in-sample fit** of f:

- empirical risk is a downward biased estimator of risk
- as we vary the hypothesis space, prediction risk can rise even when empirical risk is falling

We consider example where empirical risk is minimized over progressively larger hypothesis spaces

Generate input-output pairs via

$$x \sim U[-1,1] \quad \text{and then} \quad y = \cos(\pi x) + u \quad \text{where} \quad u \sim N(0,1)$$

$U[-1,1]$ is the uniform distribution on the interval $[-1,1]$

Hypothesis spaces for predicting y from x will be sets of polynomial functions

Let $\mathscr{P}_d$ be the set of all polynomials of degree d:

$$\mathscr{P}_d := \{ \text{ all functions } f_d(x) = c_0 x^0 + c_1 x^1 + \cdots c_d x^d$$

$$\text{where each } c_i \in \mathbb{R}\}$$

Sequence of hypothesis spaces is increasing:

$$\mathscr{P}_1 \subset \mathscr{P}_2 \subset \mathscr{P}_3 \subset \cdots$$

Using quadratic loss and the set $\mathscr{P}_d$ as our candidate functions, the risk minimization problem is

$$\min_{f \in \mathscr{P}_d} R(f) \quad \text{where} \quad R(f) = \mathbb{E}\left[(y - f(x))^2\right]$$

The empirical risk minimization problem:

$$\min_{f \in \mathscr{P}_d} R_{\text{emp}}(f) \quad \text{where} \quad R_{\text{emp}}(f) = \frac{1}{N} \sum_{n=1}^{N} (y_n - f(x_n))^2 \tag{6}$$

Experiment to demonstrate over-fitting:

1. generate $N = 25$ data points from the model – data plotted as circles
2. taking generated data, solve (6) repeatedly, once for each d in $1, 2, \ldots, 20$
3. solution to the dth minimization problem is denoted $\hat{f}_d$ – plotted in red
4. compare the risk $R(\hat{f}_d)$ and empirical risk $R_{\text{emp}}(\hat{f}_d)$

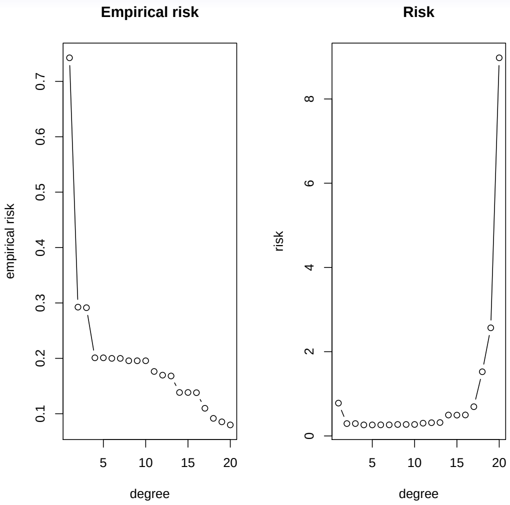

Figure: Risk and empirical risk as a function of d

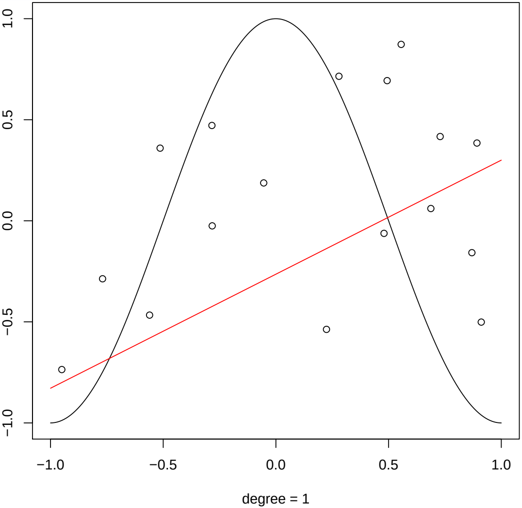

Figure: Fitted polynomial, $d = 1$

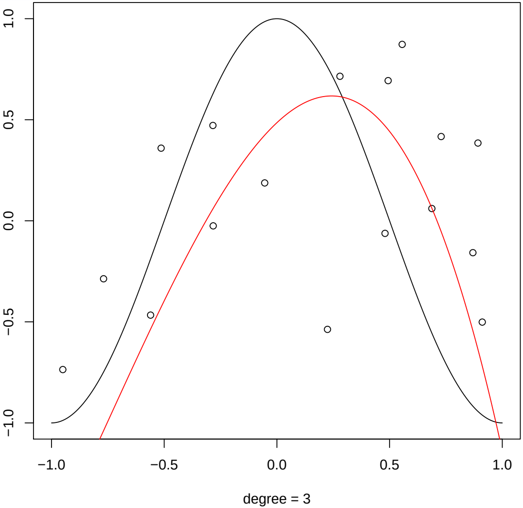

Figure: Fitted polynomial, $d = 3$

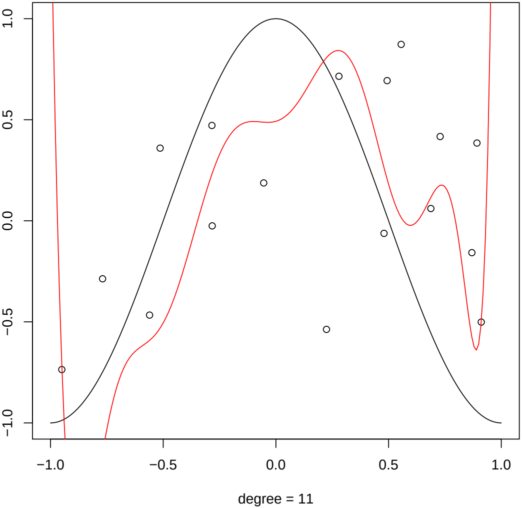

Figure: Fitted polynomial, $d = 11$

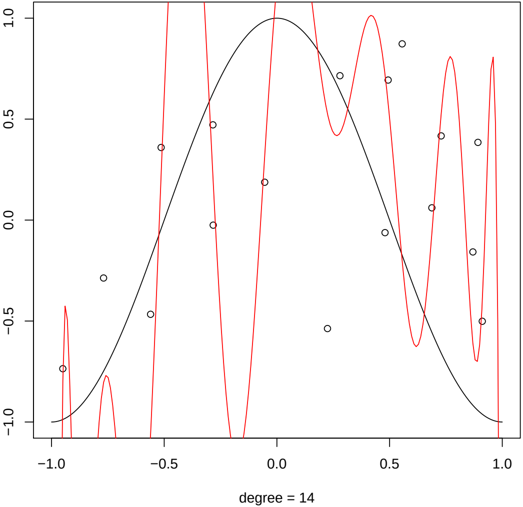

Figure: Fitted polynomial, $d = 14$

Empirical risk falls monotonically with d

But risk decreases slightly and then increases rapidly

- small d (**underfitting**): high empirical risk and high risk
- medium d: risk is minimized
- large d (**overfitting**): small empirical risk and high risk

With large d, too much emphasis has been given to this particular realization of the data

High risk means large expected loss

In applications, we do not know the true data-generating process when we choose $\mathcal{H}$

- best scenario: we have firm theory guiding us to a suitable hypothesis space
- worst scenario: we have no idea and choose blindly

Parametric Methods

Standard parametric estimation methods:

- maximum likelihood
- Bayesian estimation
- generalized method of moments

Maximum Likelihood

Suppose the data $x_1, \ldots, x_N$ has joint density p

Assume $p = p(\cdot\,; \boldsymbol{\theta})$ is a member of a parametric class $\mathscr{P}$ indexed by parameter vector $\boldsymbol{\theta} \in \Theta$

Each choice of $\boldsymbol{\theta}$ pins down a particular density $p = p(\cdot\,; \boldsymbol{\theta})$, but $\boldsymbol{\theta}$ generating the data is unknown

The **likelihood function** is p evaluated at the sample $x_1, \ldots, x_N$

The likelihood function regarded as a function of $\boldsymbol{\theta}$:

$$L(\boldsymbol{\theta}) := p(x_1, \ldots, x_N; \boldsymbol{\theta}) \qquad (\boldsymbol{\theta} \in \Theta)$$

The principle of maximum likelihood: estimate $\boldsymbol{\theta}$ by maximizing $L(\boldsymbol{\theta})$ over $\boldsymbol{\theta} \in \Theta$

A statistic $\hat{\boldsymbol{\theta}}$ is called a **maximum likelihood estimate** (MLE) of $\boldsymbol{\theta}$ if

$$\hat{\boldsymbol{\theta}} \in \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta})$$

Equivalent to maximizing the **log likelihood function**

$$\ell(\boldsymbol{\theta}) := \ln(L(\boldsymbol{\theta})) \qquad (\boldsymbol{\theta} \in \Theta)$$

The set of MLEs can be a singleton, contain multiple elements or be empty

If each x_n is drawn independently from fixed arbitrary (marginal) density $p_n(\cdot\,;\boldsymbol{\theta})$ on $\mathbb{R}$, then

$$L(\boldsymbol{\theta}) = \prod_{n=1}^{N} p_n(x_n;\boldsymbol{\theta}) \quad \text{and} \quad \ell(\boldsymbol{\theta}) = \sum_{n=1}^{N} \ln p_n(x_n;\boldsymbol{\theta})$$

If each data point is multivariate, just replace x_n with $\mathbf{x}_n$

Example. Suppose that $x_1, \dots, x_N$ are IID draws from a normal distribution $\mathrm{N}(\mu, v)$ with $\boldsymbol{\theta} = (\mu, v)$ unknown. The log likelihood function is

$$\ell(\mu, v) = -\frac{N}{2} \ln(2\pi v) - \frac{1}{2} \sum_{n=1}^{N} \frac{(x_n - \mu)^2}{v}$$

Joint maximization over (μ, v) gives the maximum likelihood estimators

$$\hat{\mu} = \frac{1}{N} \sum_{n=1}^{N} x_n \quad \text{and} \quad \hat{v} = \frac{1}{N} \sum_{n=1}^{N} (x_n - \bar{x}_N)^2$$

Thus, the MLEs of μ and v are the sample mean and sample variance

MLE estimators typically have excellent asymptotic properties. However:

- attractive asymptotic theory dependent on correct specification of the parametric class – we must specify the entire joint distribution of the sample
- good finite sample properties are not guaranteed

Conditional Maximum Likelihood

Suppose we observe inputs $\mathbf{x}_1, \ldots, \mathbf{x}_N$ to some system and corresponding outputs $y_1, \ldots, y_N$

Assume $(\mathbf{x}_n, y_n)$ are IID

Aim: estimate $\boldsymbol{\theta}$ in $p(y \mid \mathbf{x}; \boldsymbol{\theta})$

We maximize

$$\ell(\boldsymbol{\theta}) = \sum_{n=1}^{N} \ln p(\mathbf{x}_n, y_n; \boldsymbol{\theta})$$

$$\text{where} \quad p := \text{ the joint density of } (\mathbf{x}_n, y_n)$$

Letting π be the marginal density of $\mathbf{x}$, we write

$$p(\mathbf{x}, y; \boldsymbol{\theta}) = p(y \,|\, \mathbf{x}; \boldsymbol{\theta})\, \pi(\mathbf{x})$$

The density π is unknown but we have not parameterized it because we aren't trying to estimate π

Rewrite the log likelihood as

$$\ell(\boldsymbol{\theta}) = \sum_{n=1}^{N} \ln p(y_n \mid \mathbf{x}_n; \boldsymbol{\theta}) + \sum_{n=1}^{N} \ln \pi(\mathbf{x}_n)$$

Hence the MLE is

$$\operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} \sum_{n=1}^{N} \ln p(y_n \mid \mathbf{x}_n; \boldsymbol{\theta})$$

The objective function here is called the **conditional log likelihood**

Example. Consider a discrete response model with binary output y_n, where $y_n = 1$ indicates that the nth individual in a sample of women participates in the labor force

This decision is influenced by a vector $\mathbf{x}_n$ measuring characteristics such as income from the rest of the household

Let

$$q(\mathbf{s}) := \mathbb{P}\{y = 1 \,|\, \mathbf{x} = \mathbf{s}\} \qquad \left(\mathbf{s} \in \mathbb{R}^K\right)$$

One modeling approach is to take $q(\mathbf{s}) = F(\boldsymbol{\beta}^\top \mathbf{s})$, where $\boldsymbol{\beta}$ is a vector of parameters and F is a specified CDF

Example. (cont.) We can then write

$$\mathbb{P}\{y = i \,|\, \mathbf{x} = \mathbf{s}\} = F(\boldsymbol{\beta}^\mathsf{T}\mathbf{s})^i (1 - F(\boldsymbol{\beta}^\mathsf{T}\mathbf{s}))^{1-i}$$

for $\mathbf{s} \in \mathbb{R}^K$ and $i \in \{0, 1\}$

This is the conditional PMF of y given $\mathbf{x}$, so the conditional log likelihood of the sample is

$$\ell(\boldsymbol{\beta}) = \sum_{n=1}^{N} \ln[F(\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_n)^{y_n} (1 - F(\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_n))^{1-y_n}]$$

$$= \sum_{n=1}^{N} y_n \ln F(\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_n) + \sum_{n=1}^{N} (1 - y_n) \ln(1 - F(\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_n))$$

Example. (cont.) If F is the standard normal CDF Φ, then this model is called the **probit** model

If F is the logistic CDF $F(\mathbf{s}) = 1/(1 + e^{-s})$, then it is called the **logit** model

Method of Moments and GMM

Estimate a vector $\boldsymbol{\theta}$ that solves an equation of the form

$$g(\boldsymbol{\theta}) = \mathbb{E}h(\mathbf{x})$$

The functions g and h are observable and vector-valued

We have observations $\mathbf{x}_1, \ldots, \mathbf{x}_N$ from P

Apply sample analogue principle — **method of moments estimator** is the solution $\hat{\boldsymbol{\theta}}$, if it exists, to the equation

$$g(\hat{\boldsymbol{\theta}}) = \frac{1}{N} \sum_{n=1}^{N} h(\mathbf{x}_n) \tag{7}$$

Example. The mean of the Pareto distribution with scale parameter $s_0 = 1$ and shape parameter α is $\alpha/(\alpha - 1)$. (If $\alpha > 1$.)

Letting $g(\alpha) := \alpha/(\alpha - 1)$:

$$g(\alpha) = \mathbb{E}\, x \quad \text{where} \quad \mathcal{L}(x) = \text{Pareto}(\alpha, 1)$$

To estimate α with observations $x_1, \ldots, x_N$ from a Pareto$(\alpha, 1)$ distribution, solve $g(\hat{\alpha}) = \frac{1}{N} \sum_{n=1}^{N} x_n$ for $\hat{\alpha}$

The result is $\hat{\alpha} := \bar{x}_N / (\bar{x}_N - 1)$

Generalized method of moments (GMM) is a small step from method of moments

If we express (94) above as $\mathbb{E}\left[g(\boldsymbol{\theta}) - h(\mathbf{x})\right] = \mathbf{0}$, then we can consider the more general expression

$$\mathbb{E}\, G(\boldsymbol{\theta}, \mathbf{x}) = \mathbf{0} \tag{8}$$

This expression is called the **orthogonality condition**

The generalized method of moments estimator of $\boldsymbol{\theta}$ is the solution $\hat{\boldsymbol{\theta}}$ to the empirical counterpart, which is

$$\frac{1}{N}\sum_{n=1}^{N} G(\hat{\boldsymbol{\theta}}, \mathbf{x}_n) = \mathbf{0}$$

Of course, no guarantee a solution will exist here

- the function G can be nonlinear
- the number of equations can be greater than the number of unknowns — **overidentified**

When the system is overidentified, our study of overdetermined systems of equations in §3.3.2 suggests:

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \left\| \frac{1}{N} \sum_{n=1}^{N} G(\boldsymbol{\theta}, \mathbf{x}_n) \right\|$$

In practice, replace the Euclidean norm $\|\cdot\|$ with a weighted norm $\|\cdot\|_W$ defined by $\|\mathbf{x}\|_W^2 = \mathbf{x}^\intercal \mathbf{W} \mathbf{x}$, where $\mathbf{W}$ is a positive definite **weighting matrix**

- produce an estimator that has small variance asymptotically

The estimation problem:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmin}} \left[\frac{1}{N} \sum_{n=1}^{N} G(\boldsymbol{\theta}, \mathbf{x}_n) \right]^\intercal \hat{\mathbf{W}} \left[\frac{1}{N} \sum_{n=1}^{N} G(\boldsymbol{\theta}, \mathbf{x}_n) \right]$$

Note $\hat{\mathbf{W}}$ is allowed to depend on the sample

Bayesian Estimation

The main idea: treat parameters as unknown quantities for which we hold subjective beliefs regarding their values

These subjective beliefs are called **priors**

The Bayesian approach to estimation suggests that we take both data and prior knowledge into account when forming an estimate or prediction

A prior can be thought of as a distribution over the set of distributions in play, $\mathscr{P}$

The standard Bayesian approach is parametric — specialize this further to a density over parameter space

Thus the primitives in our analysis are:

- $\boldsymbol{\theta}$, the parameter vector, which takes values in $\Theta \subset \mathbb{R}^J$,
- π, the **prior distribution**, a density over Θ,
- $\mathbf{x}$, the data, and
- $p(\cdot \mid \boldsymbol{\theta})$, the joint density of the data given $\boldsymbol{\theta}$.

Note $L(\boldsymbol{\theta}) := p(\mathbf{x} \mid \boldsymbol{\theta})$ is the likelihood function

Priors are reassessed based on evidence in the data

This process leads to an updated density over parameter space called the **posterior distribution**, which we represent by $\pi(\boldsymbol{\theta} \,|\, \mathbf{x})$

Obtain posterior density via an application of Bayes' law:

$$\pi(\boldsymbol{\theta} \,|\, \mathbf{x}) = \frac{p(\mathbf{x} \,|\, \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{p(\mathbf{x})} = \frac{p(\mathbf{x} \,|\, \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int p(\mathbf{x} \,|\, \boldsymbol{\theta}')\pi(\boldsymbol{\theta}')\, \mathrm{d}\boldsymbol{\theta}'} \tag{9}$$

Here $p(\mathbf{x})$ represents the unconditional density of $\mathbf{x}$ evaluated at the outcome

Example. Consider a one-armed bandit (slot machine) with binary response v indicating a fixed payout ($v = 1$) or nothing ($v = 0$)

We would like to know the probability θ of $v = 1$

Let $v_1, \ldots, v_N$ be a sequence of independent outcomes and let $x := \sum_{n=1}^{N} v_n$ be the total number of payouts

The likelihood for x conditional on θ:

$$p(x \mid \theta) = \binom{N}{x} \theta^x (1 - \theta)^{N-x}$$

Example. (cont.) For our prior we take a Beta(α, β) distribution:

$$\pi(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} \qquad ,\theta \in (0,1) \tag{10}$$

Apply (9):

$$\pi(\theta \,|\, x) = \frac{\theta^{x+\alpha-1}(1-\theta)^{N-x+\beta-1}}{c(x)} \tag{11}$$

where $c(x) := p(x)B(\alpha,\beta)/\binom{N}{x}$

We know (11) is a density in θ given x — $c(x)$ must be the normalizing constant at x

Example. (cont.) Comparing (10) with (11), $\pi(\theta \mid x)$ is a beta density. Thus,

$$\pi(\theta \mid x) = \frac{\theta^{\alpha+x-1}(1-\theta)^{N-x+\beta-1}}{B(x+\alpha, N-x+\beta)}$$

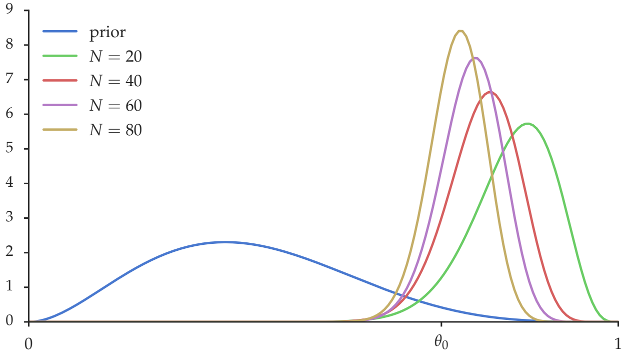

Figure: Evolution of the posterior from Beta$(3, 5)$ prior

Point estimates are extracted from the posterior distribution based on some measure of central tendency, such as the mean, the median, or the **mode** of the posterior

Example. The mean of the posterior in (104) yields the estimator

$$\hat{\theta} := \frac{\alpha + x}{\alpha + \beta + N}$$

More payouts shift our estimator upwards. In the limit, $\hat{\theta}$ is near $\frac{x}{N}$, which is the MLE of θ

This illustrates a common theme: difference between maximum likelihood and Bayesian estimates typically concerns finite sample properties

Priors where the parametric class is preserved under Bayesian updating for a specific likelihood function are called **conjugate**

In applications, integration over the parameter space carried out numerically: standard to use Markov chain Monte Carlo (see §7.4.2 in ET):

- using MCMC, we can reduce complexity by not having to evaluate the integral in the expression for the posterior distribution

Popularity of Bayesian methods:

1. MCMC more successful in practice than the numerical optimization required to obtain maximum likelihood estimates
2. Bayesian estimation provides a form of regularization that stabilizes and typically improves estimation of complex models. See §14.2.3 in ET
3. Bayesian estimation comes with an elegant, unified decision-theoretic approach to inference