Foundations of Statistical Inference

Data 100: Principles and Techniques of Data Science

Sandrine Dudoit

Department of Statistics and Division of Biostatistics, UC Berkeley

Spring 2019

Outline

Outline

Version: 12/03/2019, 17:55

Example: mpg Dataset

- Suppose we are interested in understanding which features of a car are related (to be made more precise) to its fuel consumption, i.e., mileage per gallon (mpg).
- Addressing this rather vague question involves, among other things, identifying relevant data, i.e., features of a car (e.g., number of cylinders, horsepower), collecting these data, and specifying the nature of the function relating mpg to relevant features.
- Here, we start from an already available dataset, the mpg dataset, which provides data on the following 9 variables for a sample of 398 cars: “mpg”, “cylinders”, “displacement”, “horsepower”, “weight”, “acceleration”, “model year”, “origin”, “name”.

Example: mpg Dataset

- The mpg dataset is available from the seaborn data repository (`https://github.com/mwaskom/seaborn-data`) and was originally provided on StatLib (`http://lib.stat.cmu.edu/datasets/cars.desc`).
- We discard any observation with any NA and remove the "name" variable, as it takes on over 300 different values and does not appear useful for predicting mpg.

mpg	cylinders	displacement	horsepower	weight	acceleration	model_year	origin
18.00	8	307.00	130.00	3504	12.00	70	usa
15.00	8	350.00	165.00	3693	11.50	70	usa
18.00	8	318.00	150.00	3436	11.00	70	usa
16.00	8	304.00	150.00	3433	12.00	70	usa
17.00	8	302.00	140.00	3449	10.50	70	usa
15.00	8	429.00	198.00	4341	10.00	70	usa

. . .

Example: mpg Dataset

- A natural and essential first step is exploratory data analysis (EDA), for “getting a feel for the data”, identifying patterns among the different variables, and detecting potential problems with the data.
- There are two main types of variables, quantitative variables (“mpg”, “cylinders”, “displacement”, “horsepower”, “weight”, “acceleration”, “model year”) and qualitative variables (“origin”, “name”), which may need to be handled differently at different stages of the analysis, e.g., visualization, fitting the regression function relating mpg to the other 7 features.

Example: mpg Dataset

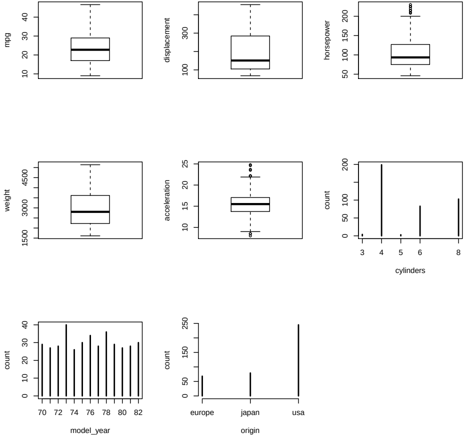


Figure 1: *mpg dataset.* Marginal distributions of mpg and 7 covariates.

Example: mpg Dataset

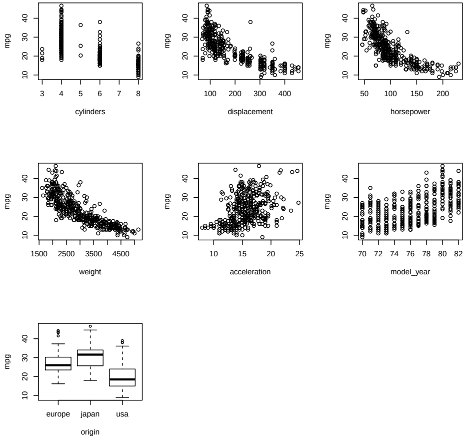


Figure 2: *mpg dataset.* Distribution of mpg vs. each of 7 covariates.

Example: mpg Dataset

- Based on the above plots, it is clear that multiple features of a car affect its mpg. For example, the mpg seems to decrease as horsepower increases and increase with model year.
- How can we use these data to find a function that relates mpg to the other 7 variables?
- A natural function is the conditional mean of mpg given the 7 variables. Such a function is known as a regression function.
- In the regression context, mpg is referred to as an outcome (a.k.a., dependent variable, response) and the other 7 variables as covariates (a.k.a., independent variables, explanatory variables).

Example: mpg Dataset

- We are immediately faced with the following issues.
  - What is an appropriate model for the regression function? There is an infinite number of functions of 7 variables. Should we use all 7 variables? Should we consider polynomial functions? If so, of what degree?
  - The mpg dataset corresponds to a sample of cars from a much larger population of cars (presumably, all cars in the USA for a particular time period). How can we use the sample to accurately infer the regression function for an entire population of cars? This will depend on how the sample was obtained, i.e., whether it was obtained according to a well-defined sampling procedure.

Example: mpg Dataset

- Let Y denote the random variable (random variables are defined precisely below) for mpg and $X = (X_1, \ldots, X_7)$ the random variables for the 7 other features (numbered in the order "cylinders", "displacement", "horsepower", "weight", "acceleration", "model year", "origin"). The data for the ith car are (X_i, Y_i), $i = 1, \ldots, n$, $n = 392$.
- Below, we consider three different types of models for the regression function, i.e., the conditional mean $\psi(X) = \mathrm{E}[Y|X]$ of mpg given the 7 covariates.
- const: Model mpg as a constant.

$$\mathrm{E}[Y|X] = \beta_0.$$

This model completely ignores the obvious association of mpg with features such as horsepower.

Example: mpg Dataset

- hp: Model mpg as a linear[1] function of horsepower.

$$E[Y|X] = \beta_0 + \beta_4 X_4.$$

This model is more informative than the constant model, but doesn't account for the association of mpg with the other 6 covariates or potential non-linear effects of horsepower on mpg (cf. higher order polynomial).

- all: Model mpg as a linear function of all 7 covariates.

$$\begin{aligned} E[Y|X] &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \\ &\quad + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 \\ &\quad + \beta_{7,Japan} \mathrm{I}(X_7 = Japan) + \beta_{7,USA} \mathrm{I}(X_7 = USA). \end{aligned}$$

Example: mpg Dataset

Note that we treat the qualitative variable "origin" differently than the other 6 variables that are quantitative[2]. This model accounts for all 7 covariates, but could miss possible non-linear dependencies of mpg on the 7 features as well as interactions between these features.

- We will discuss how to fit these models, i.e., estimate the regression parameters β of each model, in subsequent lectures.
- For now, let's examine the fitted values from each model, i.e., the mpg values $\hat{Y}_i$ from the estimated regression function $\hat{\psi}(X_i)$ (based on the estimated coefficients $\hat{\beta}$) evaluated for each car in the sample.

Example: mpg Dataset

- Residuals $e_i = Y_i - \hat{Y}_i$, which compare the observed outcomes Y_i to the fitted values $\hat{Y}_i$, can be used to assess the fit of a model.
- A global goodness-of-fit measure is the mean squared error (MSE), i.e., the average of the squared residuals

  $$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

  Small (large) MSE indicates that the fitted and observed outcomes are similar (different).

---

[1] The expression "linear regression" typically refers to linearity in the parameters β. Covariates X can enter the model via arbitrary functions, e.g., polynomial, logarithm, sine functions.

[2] I() denotes the indicator function, equal to one if its argument is true and zero otherwise

Example: mpg Dataset

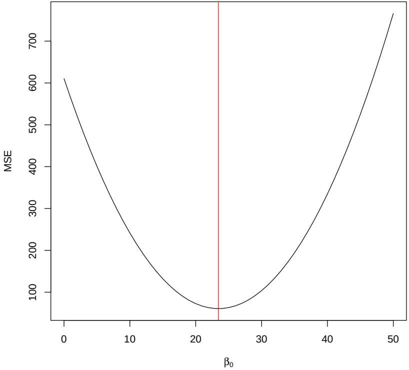


Figure 3: *mpg dataset.* Empirical MSE for constant model $E[Y|X] = \beta_0$: $MSE = \sum_i (Y_i - \beta_0)^2/n$. Red line indicates average mpg.

Example: mpg Dataset

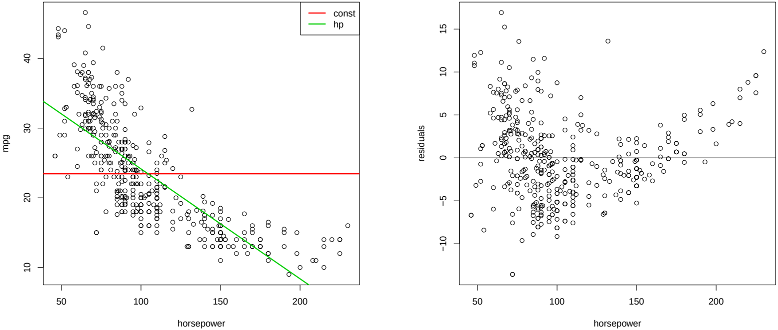


Figure 4: *mpg dataset.* Linear regression of mpg on horsepower.

Example: mpg Dataset

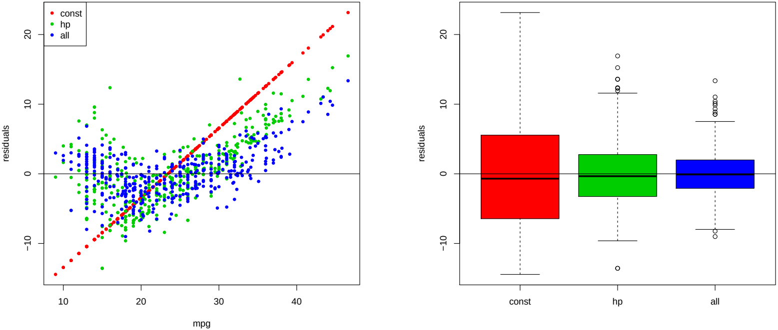


Figure 5: *mpg dataset.* Residuals for three linear regression models.

Example: mpg Dataset

Table 1: *mpg dataset.* Empirical mean squared error for three regression models.

Const	HP	All
60.76	23.94	10.68

Example: mpg Dataset

- A widely used approach for fitting regression models as above is least squares estimation (LSE): The regression parameters are estimated by minimizing MSE with respect to (wrt) β.
- As we will see later, LSE is part of a general inference framework which is based on risk optimization.
- Regression models are typically fit on data from a sample drawn from a population. Important questions therefore include assessing how well the sample-based estimated regression function performs for the population, i.e., assessing the accuracy of the estimated regression coefficients and the prediction error of the regression function.

Example: mpg Dataset

- All of the above issues are part of statistical inference/learning.

Statistical Inference

- The first two and essential aspects of a data-enabled inquiry are
  - framing the question;
  - identifying relevant data, i.e., what to measure.
- This often involves identifying a population of observational units and the variables to measure on each of these units.
- The answer to the question then takes the form of numerical and graphical summaries (i.e., functions) of these data, i.e., statistics.

Statistical Inference

- However, most of the time one cannot collect data for the entire population of interest. Instead, one obtains data for a sample (i.e., subset) of observational units drawn from this population. The sample is, in some sense, a proxy for the population.
- This is where statistical inference/learning comes into play: How to use the sample to infer/learn about the population.
- The sample should be representative of the population and selected according to well-defined probabilistic procedures to allow assessment of the accuracy of the answer, cf. estimator bias and variance.
- With probability sampling one can assign a precise probability to the event that each particular sample is drawn from the population.

Statistical Inference

- This allows us to quantify confidence about an estimator, prediction, or hypothesis test.
- Probability Theory allows us to characterize randomness and quantify uncertainty due to sampling.

Box Model for Sampling

- A useful representation for sampling is a box model, where the population of interest is represented by a box of N tickets, each with values (i.e., variables) written on them.
- A sample is a subset of tickets drawn from the box and the data are the values written on these tickets.
- A simple random sample (SRS) of size n is obtained by drawing n tickets at random **without** replacement from the box.
- For a small sample compared to the population, SRS is very similar to sampling at random **with** replacement.
- As seen in a previous lecture, other forms of probability sampling include cluster sampling and stratified sampling.

Box Model for Sampling

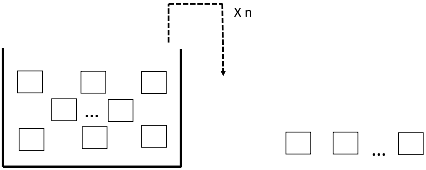


Figure 6: *Box model.* For an SRS, sample *n* tickets are random without replacememt from the box.

Box Model for Sampling

- Sampling at random without replacement. (SRS)
    - How many ways are there so select a sample of size n from a population of size N?

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}.$$

    - What is the chance that a particular element of the population is selected?

$$\frac{\binom{N-1}{n-1}}{\binom{N}{n}}.$$

- Sampling at random with replacement.

Box Model for Sampling

- How many ways are there so select a sample of size n from a population of size N?

$$N^n.$$

- What is the chance that a particular element of the population is selected?

$$1 - \frac{(N-1)^n}{N^n}.$$

- What is a box model for cluster sampling?
- What is a box model for stratified sampling?

Statistical Inference

- The values in the population have a distribution (i.e., frequencies), which we refer to as population distribution or data generating distribution.
- A parameter is a function of the values in the population, i.e., of the data generating distribution.
  E.g. Average of all values in the population.
- In the frequentist inference framework, parameters are typically unknown fixed quantities to be estimated based on data from a sample.
- In the Bayesian inference framework, parameters are viewed as random and having distributions: A prior distribution (before data are collected) and a posterior distribution (conditional on the data collected).

Statistical Inference

- The values in the sample have a distribution which we refer to as data empirical distribution.
- An estimator is a function of the values in the sample, i.e., of the empirical distribution.
  E.g. Average of all values in the sample.
- Estimators are typically known random variables, that is, their values depend on which sample was drawn from the population.
- In the frequentist setting, the sampling distribution of an estimator refers to the different values it takes when repeatedly randomly sampling from the population.

Statistical Inference

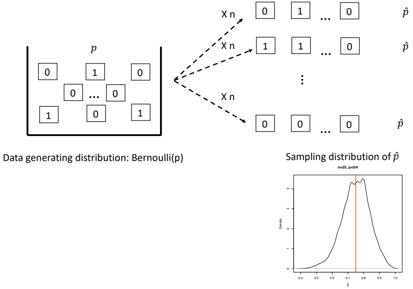


Figure 7: *Box model and sampling distribution.* Sampling distribution of the proportion of "1", $\hat{p}_n$, for $n = 25$ independent draws from a Bernoulli($p = 3/4$) data generating distribution (cf. repeatedly flipping a biased coin 25 times).

Statistical Inference

- A broad range of data-driven inquiries involve statistical inference/learning, i.e., (part of) the question can be framed into estimating or testing hypotheses about a parameter of interest.
- One of the hardest and underestimated aspects of Applied Statistics, as well as Data Science, is to translate, when appropriate, a possibly vague domain question into a statistical inference question, i.e., a parameter to be estimated or for which to test hypotheses.
- Statistical inference/learning involves using the known data empirical distribution to estimate parameters or test hypotheses concerning the unknown data generating distribution.

Statistical Inference

- Statistical inference accounts for randomness/uncertainty due to sampling and involves characterizing the sampling distribution of an estimator of the parameter of interest.
- This includes assessing the bias and variance of an estimator (see below) and the false positive/negative error rates of a testing procedure.
- Optimal statistical inference, i.e., finding an optimal (cf. risk minimization, below) estimator/predictor/test given the question and data, comes at an intermediate stage of the Data Science workflow, after data cleaning and exploratory data analysis.

Statistical Inference

- It is closely connected to previous steps, as the parameter of interest is identified when framing the question and EDA can suggest probabilistic models for the data generating mechanism by revealing patterns in the data and relationships between variables.
- EDA can also suggest a new parameter of interest, reflecting the iterative nature of the workflow.

Statistical Inference

Sample
Data empirical distn.
Estimator
$\hat{\theta}_n = \hat{\Theta}(P_n)$
Known

Statistical
inference/learning
$\Longrightarrow$
Estimation
Hypothesis testing

Population
Data generating distn.
Parameter
$\theta = \Theta(P)$
Unknown

Parameter: Unknown object of interest corresponding to domain question.

Estimator: Data-driven/educated guess at object of interest and answer to domain question.

Statistical Inference: Examples

- Election poll.
    - In a typical poll, the population of interest is the set of all voters in a particular state and a variable of interest is the preferred candidate of each voter.
    - The parameter of interest is the proportion (a mean of binary indicators) of voters intending to vote for each candidate.
    - In practice, one cannot record voting preferences for the entire population. Instead, one estimates the parameter of interest based on voting preferences for a random sample of voters (e.g., SRS, cluster sample).
- Observational case/control study.
    - A typical case/control study concerns the identification of variables (e.g., environmental exposure measures, gene expression measures) associated with a particular disease.

Statistical Inference: Examples

  - The population of interest could be, for instance, the set of all adults living in a particular region.
  - A parameter of interest is the difference in means of an exposure variable between the cases (individuals with disease) and the controls (individuals without the disease).
  - In practice, one cannot measure the variables of interest for the entire population. Instead, one selects random samples of cases and controls, possibly matched on a variety of covariates (e.g., gender, race, age) to avoid confounding.
- A/B testing.
  - In a typical A/B testing problem for conversion rate optimization, the population of interest is a market segment (e.g., members of a certain social network) and a variable of interest is the response to a particular feature of a website.

Statistical Inference: Examples

  - The parameter of interest is the (difference in) proportion of viewers who turn into customers for each feature (A or B) of the website.
  - In practice, one cannot enroll the entire population in the A/B testing trial. Instead, one estimates the parameter of interest based on the responses of a random sample from the market segment.
  - Ideally, the sample is obtained using a randomized design, where viewers are randomly assigned to either the A or B treatment group.

- Regression.

Statistical Inference: Examples

- Suppose one is interested in predicting rent for Berkeley apartments. The population of interest is the set of all rental apartments in Berkeley and variables of interest are the rent, of course, but also features of an apartment such as, square footage, number of bedrooms, number of bathrooms, availability of washer/dryer.
- The overall mean rent of all apartments in Berkeley is not a particularly informative parameter.
- A more interesting parameter is the regression function or conditional mean of the rent given covariates such as square footage, number of bedrooms, number of bathrooms, availability of washer/dryer.
- In practice, one cannot readily collect data on all rental apartments in Berkeley. Instead, one uses a sample of units to estimate the regression function. This sample is typically not a probability sample (e.g., Craigslist data), thus making statistical inference problematic.

Statistical Inference: Examples

What are box models for the above questions?

Statistical Inference: Examples

Measure variables of interest for controls (A)

Measure variables of interest for cases (B)

Figure 8: *Box model: Observational case/control design.*

Statistical Inference: Examples

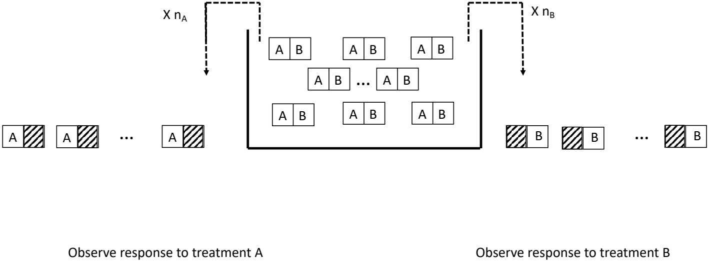


Figure 9: *Box model: Randomized design.*

Random Variables

- A random variable (RV) is a numerial function of a probabilistic event/outcome.
- Random variables are typically real-valued scalars, discrete (e.g., number of heads in ten coin flips) or continuous (e.g., average height for a SRS of one hundred Berkeley students), altough they can take on values in higher dimensions (e.g., values of a stock over time, expression measures for an entire genome).
- In what follows, we will focus mostly on discrete random variables, altough metholology and theory are also available for continuous variables as well as higher-dimensional variables (i.e., random vectors).

Random Variables

- We will denote random variables using upper-case letters and realizations of these variables, i.e., the values they take on, using lower-case letters.

Probability Distributions

- The probability distribution of a random variable specifies probabilities for the values it takes on.
- It is common to distinguish between distributions for
    - discrete random variables, that take on a specified finite or countable list of values,
    - continuous random variables, that take on any numerical value in an interval or collection of intervals.

However, there is theory that provides a unified treatment of both cases.

Probability Distributions

- The cumulative distribution function (CDF) of a random variable X is defined as

$$F_X(x) \equiv \Pr(X \leq x), \qquad \forall x \in \mathbb{R}\,. \tag{1}$$

  The CDF is, by definition, non-decreasing, right-continuous, and its range is $[0, 1]$.

- One can show that, for any $a, b \in \mathbb{R}$,

$$F_X(b) - F_X(a) = \Pr(a < X \leq b). \tag{2}$$

Probability Distributions

- For a discrete random variable X, the probability mass function (PMF) provides the probability that X takes on each of its possible values

$$f_X(x) \equiv \begin{cases} \Pr(X = x), & x \in \mathcal{X} \\ 0, & \text{otherwise} \end{cases}, \qquad (3)$$

where $\mathcal{X}$ denotes the support of X, i.e., the set of all possible values of X.

- A PMF satisfies

$$\begin{aligned} 0 \leq f_X(x) \leq 1, & \qquad \forall x \in \mathcal{X} \qquad (4) \\ \sum_{x \in \mathcal{X}} f_X(x) &= 1. \end{aligned}$$

Probability Distributions

- The CDF of a discrete random variable is given by

$$F_X(x) = \sum_{\{x' \in \mathcal{X} : x' \leq x\}} f_X(x'). \tag{5}$$

  It is a step function, with steps of size $f_X(x)$ for each x in the support of X.

- For a continuous random variable X, the probability density function (PDF) is a non-negative continuous function f_X such that

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt. \tag{6}$$

  Intuitively, one can think of $f_X(x)dx$ as the probability that X falls within the infinitesimal interval $[x, x + dx]$.

Probability Distributions

- The CDF of a continuous random variable is continuous.
- For a continuous random variable,

$$F_X(b) - F_X(a) = \Pr(a < X \leq b) = \int_a^b f_X(x)dx, \qquad (7)$$

  i.e., probabilities correspond to areas under the PDF.
- Distributions are indexed by parameters which are typically unknown and to be inferred from data that ideally correspond to a sample drawn from that distribution. E.g. The (discrete) Binomial distribution has two parameters, the number of Bernoulli trials (with binary outcomes) n and the "success" probability of each trial p (see below).

Probability Distributions

E.g. The (continuous) Gaussian/normal distribution has two parameters, the mean μ and the standard deviation σ (see below).

Gaussian Distribution

- The Gaussian or normal distribution is widely used for real-valued continuous random variables due to its relative simplicity and convenience, as well as probability theoretical results.
- The Gaussian distribution has two parameters, the mean $\mu \in \mathbb{R}$ and the standard deviation $\sigma \in \mathbb{R}^+$, representing, respectively, the center and spread of the distribution.
- A short-hand notation for this distribution is $\mathrm{N}(\mu, \sigma)$.
- The $\mathrm{N}(\mu, \sigma)$ PDF is given by

$$f_N(x; \mu, \sigma) \equiv \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \qquad x \in \mathbb{R}. \tag{8}$$

Gaussian Distribution

- The Gaussian distribution has support on the entire real line $\mathbb{R}$ and is symmetric about the origin.
- There is no closed-form expression for the CDF.
- The Gaussian/normal distribution $\mathrm{N}(0, 1)$ with zero mean and unit standard deviation is referred to as the standard Gaussian/normal distribution.
- One reason for the wide use of the Gaussian distribution is the Central Limit Theorem, which states that the sum of a large number of independent (defined below) random variables is approximately normally distributed, regardless of the distribution of the individual variables.

Gaussian Distribution

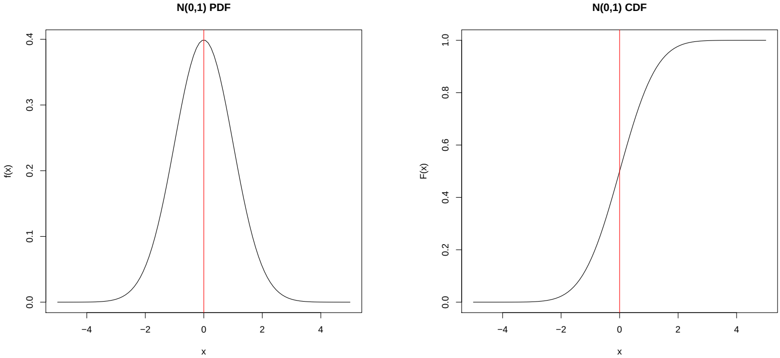


Figure 10: *Gaussian distribution.* Standard Gaussian $N(0, 1)$ PDF (left) and CDF (right).

Gaussian Distribution

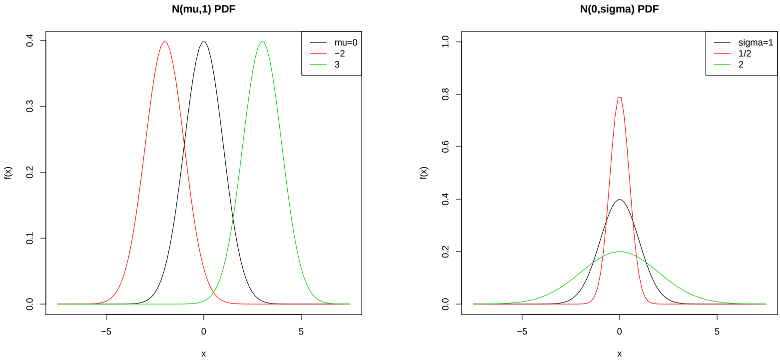


Figure 11: *Gaussian distribution.* Left: N$(\mu, 1)$ PDF for $\mu \in \{0, -2, 3\}$. Right: N$(0, \sigma)$ PDF for $\sigma \in \{1, 1/2, 2\}$.

Binomial Distribution

- Consider flipping a biased coin independently n times and denote by p the probability that the coin lands heads on any given flip.
- Let X_i denote the binary indicator for the outcome of the ith coin flip, equal to 1 if it lands heads ("success") and 0 if it lands tails ("failure").
- The X_i are n independent and identically distributed Bernoulli(p) random variables with "success" probability p, i.e., $\Pr(X_i = 1) = p$.
- Let Y_n denote the sum of the n random variables $\{X_i : i = 1, \ldots, n\}$, i.e., the total number of heads in the n flips.

Binomial Distribution

- Then, the support of Y_n is $\mathcal{Y}_n = \{0, 1, \ldots, n\}$ and Y_n has a Binomial(n, p) distribution with PMF

$$f_{Bin}(y; n, p) \equiv \binom{n}{y} p^y (1-p)^{n-y}, \qquad y \in \mathcal{Y}_n. \quad (9)$$

Binomial Distribution

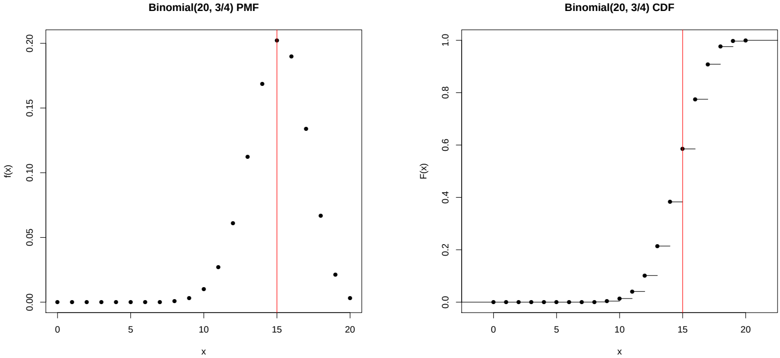
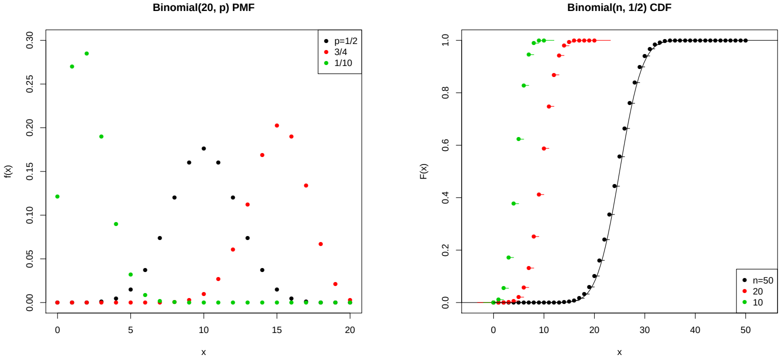


Figure 12: *Binomial distribution.* Binomial(20, 3/4) PMF (left) and CDF (right). Red line indicates mean, $np = 15$.

Binomial Distribution


Binomial(20, p) PMF

Binomial(n, 1/2) CDF

Figure 13: *Binomial distribution.* Left: Binomial$(20, p)$ PMF for $p \in \{1/2, 3/4, 1/10\}$. Right: Binomial$(n, 1/2)$ CDF for $n \in \{50, 20, 10\}$. Black continuous curve is $\mathrm{N}(np, \sqrt{np(1-p)})$ CDF, $n = 50,\ p = 1/2$.

Joint, Conditional, and Marginal Distributions

- Distributions can be defined for multiple random variables.
- The joint distribution of two or more random variables yields the probability that these random variables simultaneously take on a specific set of values.
- For instance, for two discrete random variables X and Y, taking on values in $\mathcal{X}$ and $\mathcal{Y}$, respectively, the joint probability mass function is

$$f_{X,Y}(x,y) \equiv \Pr(X = x, Y = y), \qquad x \in \mathcal{X}, y \in \mathcal{Y}, \quad (10)$$

and

$$\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} f_{X,Y}(x,y) = 1.$$

Joint, Conditional, and Marginal Distributions

- The conditional PMF of Y given $X = x$ is

$$\begin{aligned} f_{Y|X=x}(y) &\equiv \Pr(Y = y|X = x) \\ &= \frac{\Pr(X = x, Y = y)}{\Pr(X = x)} \\ &= \frac{f_{X,Y}(x,y)}{f_X(x)}. \end{aligned} \tag{11}$$

  The conditional PMF of X given $Y = y$ is defined likewise.

- The distribution of each individual variable is referred to as marginal distribution and can be obtained from the joint distribution by adding over all possible values for the other random variable

$$f_X(x) = \sum_{y \in \mathcal{Y}} f_{X,Y}(x,y). \tag{12}$$

Joint, Conditional, and Marginal Distributions

- Two random variables are independent if knowing the value of one variable does not affect the probability of the other one taking on any of its possible values. That is,

$$f_{Y|X=x}(y) = f_Y(y), \qquad \forall x \in \mathcal{X}, y \in \mathcal{Y}.$$

- For independent random variables, the joint PMF is the product of the marginal PMF

$$f_{X,Y}(x, y) = f_X(x) f_Y(y). \tag{13}$$

- Similar results are available for continuous random variables, with PDF replacing PMF and integrals replacing sums. In particular,

$$f_X(x) = \int_{\mathcal{Y}} f_{X,Y}(x, y) dy.$$

Joint, Conditional, and Marginal Distributions: Example

Rolling a fair and a loaded six-sided die.

- Example from `http://prob140.org/textbook/chapters/Chapter_04/02_Examples`.
- Consider rolling (independently) one fair six-sided die and one loaded six-sided die.
- Let X and Y denote, respectively, the number of spots from one roll of the fair and loaded dice, respectively.
- Suppose the loaded die has the following distribution for the number of spots.

$$\begin{aligned}
\Pr(Y=1)=\Pr(Y=2) &= \frac{1}{16}\\
\Pr(Y=3)=\Pr(Y=4) &= \frac{3}{16}\\
\Pr(Y=5)=\Pr(Y=6) &= \frac{4}{16}.
\end{aligned}$$

Joint, Conditional, and Marginal Distributions: Example

- Given the independence of the two rolls, the joint distribution of X and Y is given by

$$\begin{aligned} p_{X,Y}(i,j) &= \Pr(X=i, Y=j) = \Pr(X=i)\Pr(Y=j) \\ &= \frac{1}{6}\Pr(Y=j), \qquad \forall i,j \in \{1,2,\ldots,6\}. \end{aligned}$$

- The joint distribution can be displayed as a matrix, with rows and columns corresponding, respectively, to the fair and loaded dice and with elements equal to the joint PMF.

Joint, Conditional, and Marginal Distributions: Example

Table 2: *Joint distribution of number of spots for one roll of a fair six-sided die and one roll of a loaded six-sided die.* Row and column sums correspond, respectively, to the marginal distributions of the fair and loaded dice.

		Loaded die, Y						
		1	2	3	4	5	6	
Fair die, X	1	$\frac{1}{96}$	$\frac{1}{96}$	$\frac{3}{96}$	$\frac{3}{96}$	$\frac{4}{96}$	$\frac{4}{96}$	$\frac{1}{6}$
	2	$\frac{1}{96}$	$\frac{1}{96}$	$\frac{3}{96}$	$\frac{3}{96}$	$\frac{4}{96}$	$\frac{4}{96}$	$\frac{1}{6}$
	3	$\frac{1}{96}$	$\frac{1}{96}$	$\frac{3}{96}$	$\frac{3}{96}$	$\frac{4}{96}$	$\frac{4}{96}$	$\frac{1}{6}$
	4	$\frac{1}{96}$	$\frac{1}{96}$	$\frac{3}{96}$	$\frac{3}{96}$	$\frac{4}{96}$	$\frac{4}{96}$	$\frac{1}{6}$
	5	$\frac{1}{96}$	$\frac{1}{96}$	$\frac{3}{96}$	$\frac{3}{96}$	$\frac{4}{96}$	$\frac{4}{96}$	$\frac{1}{6}$
	6	$\frac{1}{96}$	$\frac{1}{96}$	$\frac{3}{96}$	$\frac{3}{96}$	$\frac{4}{96}$	$\frac{4}{96}$	$\frac{1}{6}$
		$\frac{1}{16}$	$\frac{1}{16}$	$\frac{3}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{4}{16}$	

Joint, Conditional, and Marginal Distributions: Example

- The chance that we roll the same number for both dice is

$$\begin{aligned}\Pr(X = Y) &= \sum_{i=1}^{6} \Pr(X = Y = i)\\ &= \sum_{i=1}^{6} \Pr(X = i)\Pr(Y = i)\\ &= \frac{1}{6}\sum_{i=1}^{6} \Pr(Y = i) = \frac{1}{6}.\end{aligned}$$

Joint, Conditional, and Marginal Distributions: Example

- The chance that the number on the loaded die exceeds the number on the fair die by more than 2 is

$$\begin{aligned}\Pr(Y - X > 2) &= \sum_{i=1}^{3} \sum_{j=i+3}^{6} p_{X,Y}(i,j) \\ &= \frac{23}{96}.\end{aligned}$$

	Loaded die, Y					
Fair die, X	1	2	3	4	5	6
1	$\frac{1}{96}$	$\frac{1}{96}$	$\frac{3}{96}$	$\frac{3}{96}$	$\frac{4}{96}$	$\frac{4}{96}$
2	$\frac{1}{96}$	$\frac{1}{96}$	$\frac{3}{96}$	$\frac{3}{96}$	$\frac{4}{96}$	$\frac{4}{96}$
3	$\frac{1}{96}$	$\frac{1}{96}$	$\frac{3}{96}$	$\frac{3}{96}$	$\frac{4}{96}$	$\frac{4}{96}$
4	$\frac{1}{96}$	$\frac{1}{96}$	$\frac{3}{96}$	$\frac{3}{96}$	$\frac{4}{96}$	$\frac{4}{96}$
5	$\frac{1}{96}$	$\frac{1}{96}$	$\frac{3}{96}$	$\frac{3}{96}$	$\frac{4}{96}$	$\frac{4}{96}$
6	$\frac{1}{96}$	$\frac{1}{96}$	$\frac{3}{96}$	$\frac{3}{96}$	$\frac{4}{96}$	$\frac{4}{96}$

Joint, Conditional, and Marginal Distributions: Example

- The chance that the numbers on the two dice differ by no more than 1 is

$$
\begin{aligned}
\Pr(|X-Y| \leq 1) &= \sum_{j=1}^{2} p_{X,Y}(1,j) + \sum_{i=2}^{5} \sum_{j=i-1}^{i+1} p_{X,Y}(i,j) \\
&\quad + \sum_{j=5}^{6} p_{X,Y}(6,j) \\
&= \frac{43}{96}.
\end{aligned}
$$

		Loaded die, Y					
		1	2	3	4	5	6
Fair die, X	1	$\frac{1}{96}$	$\frac{1}{96}$	$\frac{3}{96}$	$\frac{3}{96}$	$\frac{4}{96}$	$\frac{4}{96}$
	2	$\frac{1}{96}$	$\frac{1}{96}$	$\frac{3}{96}$	$\frac{3}{96}$	$\frac{4}{96}$	$\frac{4}{96}$
	3	$\frac{1}{96}$	$\frac{1}{96}$	$\frac{3}{96}$	$\frac{3}{96}$	$\frac{4}{96}$	$\frac{4}{96}$
	4	$\frac{1}{96}$	$\frac{1}{96}$	$\frac{3}{96}$	$\frac{3}{96}$	$\frac{4}{96}$	$\frac{4}{96}$
	5	$\frac{1}{96}$	$\frac{1}{96}$	$\frac{3}{96}$	$\frac{3}{96}$	$\frac{4}{96}$	$\frac{4}{96}$
	6	$\frac{1}{96}$	$\frac{1}{96}$	$\frac{3}{96}$	$\frac{3}{96}$	$\frac{4}{96}$	$\frac{4}{96}$

Joint, Conditional, and Marginal Distributions: Example

- The chance that the sum of the numbers on the two dice is 7 is

$$\begin{aligned}\Pr(X+Y=7) &= \frac{1}{6}\sum_{i=1}^{6}\Pr(Y=7-i)\\ &= \frac{1}{6}.\end{aligned}$$

		Loaded die, Y					
		1	2	3	4	5	6
	1	$\frac{1}{96}$	$\frac{1}{96}$	$\frac{3}{96}$	$\frac{3}{96}$	$\frac{4}{96}$	$\frac{4}{96}$
	2	$\frac{1}{96}$	$\frac{1}{96}$	$\frac{3}{96}$	$\frac{3}{96}$	$\frac{4}{96}$	$\frac{4}{96}$
Fair die, X	3	$\frac{1}{96}$	$\frac{1}{96}$	$\frac{3}{96}$	$\frac{3}{96}$	$\frac{4}{96}$	$\frac{4}{96}$
	4	$\frac{1}{96}$	$\frac{1}{96}$	$\frac{3}{96}$	$\frac{3}{96}$	$\frac{4}{96}$	$\frac{4}{96}$
	5	$\frac{1}{96}$	$\frac{1}{96}$	$\frac{3}{96}$	$\frac{3}{96}$	$\frac{4}{96}$	$\frac{4}{96}$
	6	$\frac{1}{96}$	$\frac{1}{96}$	$\frac{3}{96}$	$\frac{3}{96}$	$\frac{4}{96}$	$\frac{4}{96}$

Joint, Conditional, and Marginal Distributions: Example

- The conditional probability that the number on the loaded die is 4 given that the sum of the numbers is 7 is

$$\begin{aligned}
\Pr(Y=4|X+Y=7) &= \frac{\Pr(X+Y=7, Y=4)}{\Pr(X+Y=7)}\\
&= \frac{\Pr(X=3, Y=4)}{\Pr(X+Y=7)}\\
&= \frac{\Pr(X=3)\Pr(Y=4)}{\Pr(X+Y=7)}\\
&= \frac{\frac{1}{6}\frac{3}{16}}{\frac{1}{6}} = \frac{3}{16}.
\end{aligned}$$

Expected Values

- Two useful summaries/parameters of the distribution of a random variable are its expected value and its variance, which pertain, respectively, to its average/center/location and spread/scale.
- The expected value (in short, expectation) or mean value (in short, mean) of a discrete random variable X with PMF f_X is defined as

$$\mathrm{E}[X] \equiv \sum_{x \in \mathcal{X}} x \; f_X(x). \tag{14}$$

For a continuous random variable with PDF f_X,

$$\mathrm{E}[X] \equiv \int_{\mathcal{X}} x f_X(x) dx. \tag{15}$$

Expected Values

- E.g. For the example with the fair and loaded six-sided dice

$$
\begin{aligned}
\mathrm{E}[X] &= \sum_{x=1}^{6} x\frac{1}{6} = \frac{1}{6}\sum_{i=1}^{6} x \\
&= \frac{1}{6}\frac{7\times 6}{2} = \frac{7}{2} = 3.5
\end{aligned}
$$

and

$$
\begin{aligned}
\mathrm{E}[Y] &= 1\times\frac{1}{16} + 2\times\frac{1}{16} + 3\times\frac{3}{16} \\
&\quad + 4\times\frac{3}{16} + 5\times\frac{4}{16} + 6\times\frac{4}{16} \\
&= \frac{17}{4} = 4.25.
\end{aligned}
$$

Expected Values

- Expected values satisfy a linearity property, in the sense that for two (possibly dependent) random variables X and Y and a constant $c \in \mathbb{R}$

$$\begin{aligned} \mathrm{E}[X+Y] &= \mathrm{E}[X] + \mathrm{E}[Y] \qquad (16) \\ \mathrm{E}[cX] &= c\,\mathrm{E}[X]. \end{aligned}$$

- For two independent random variables X and Y,

$$\mathrm{E}[XY] = \mathrm{E}[X]\,\mathrm{E}[Y]. \qquad (17)$$

  Note that this does not hold in general for dependent variables.

Variances

- The variance of a random variable X is its mean squared deviation from its mean

$$\mathrm{Var}[X] \equiv \mathrm{E}[(X - \mathrm{E}[X])^2]. \tag{18}$$

- The square root of the variance is the standard deviation (SD).
- It can easily be shown that

$$\mathrm{Var}[X] = \mathrm{E}[X^2] - (\mathrm{E}[X])^2. \tag{19}$$

- For any two constants $a, b \in \mathbb{R}$

$$\mathrm{Var}[aX + b] = a^2 \, \mathrm{Var}[X]. \tag{20}$$

Variances

- For two independent random variables X and Y,

$$\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]. \tag{21}$$

Unlike with expectations, this linearity property does not hold in general for dependent variables.

Covariances and Correlations

- The covariance of two random variables X and Y is defined as

$$\mathrm{Cov}[X, Y] \equiv \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])]. \tag{22}$$

- When $X = Y$, the covariance reduces to the variance

$$\mathrm{Cov}[X, X] = \mathrm{Var}[X].$$

- It can easily be shown that

$$\mathrm{Cov}[X, Y] = \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y]. \tag{23}$$

- When X and Y are independent

$$\mathrm{Cov}[X, Y] = 0. \tag{24}$$

Covariances and Correlations

- For any two random variables X and Y (dependent or not)

$$\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\,\mathrm{Cov}[X,Y]. \tag{25}$$

- While variances are non-negative, covariances can be negative.
- Covariances are used to measure the linear association or correlation between two random variables and the sign of the covariance reflects whether they are positively or negatively correlated.
- The Pearson correlation coefficient between X and Y is their covariance scaled by the square root of their variances

$$\mathrm{Cor}[X,Y] \equiv \frac{\mathrm{Cov}[X,Y]}{\sqrt{\mathrm{Var}[X]\,\mathrm{Var}[Y]}}. \tag{26}$$

Covariances and Correlations

- It can be shown that

$$-1 \leq \text{Cor}[X, Y] \leq 1. \tag{27}$$

- Interpretation.
  - A correlation of 1 corresponds to a perfect linear relationship between X and Y, where Y increases as X increases: $Y = aX + b$, with $a > 0$.
  - A correlation of -1 corresponds to a perfect linear relationship between X and Y, where Y decreases as X increases: $Y = aX + b$, with $a < 0$.
  - A correlation of 0 implies that there is no linear relationship between X and Y (but possibly a non-linear relationship).

Covariances and Correlations

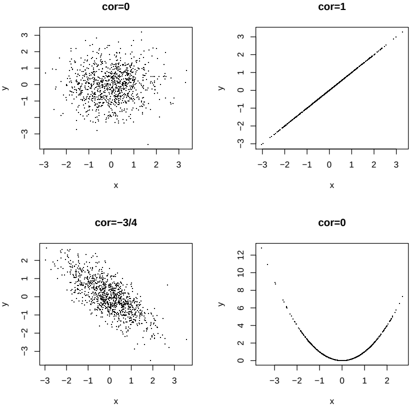


Figure 14: *Correlation coefficient.* First 3 panels: X and Y have same N$(0, 1)$ distributions, but different correlation coefficients. Bottom-right panel: $Y = X^2$, but $\text{Cor}[X, Y] = 0$.

Correlation, Association/Dependence, and Causation

- Uncorrelation does not imply independence.
  E.g. Consider $X \sim \mathrm{N}(0,1)$ and $Y = X^2$. Then, $\mathrm{Cor}[X,Y] = 0$, but X and Y are clearly dependent.
- Association/dependence does not imply causation.
  If two random variables are dependent this does not imply that changes in the value of one cause the other to change.
- The association between two random variables could be due to a confounding variable, i.e., a third variable that influences both variables.
  E.g. Association between murder rate and sale of ice cream, with weather as confounding variable.
- Causation is in general much harder to establish than association and typically requires the use of randomized controlled experiments (vs. observational studies).

Correlation, Association/Dependence, and Causation

- Causation does not imply correlation, as the correlation coefficient is only a measure of linear association between two random variables.

Robust Statistics

- Means, variances, and covariances can be sensitive to outliers, i.e., observations that are distant from most of the other observations.
- By contrast, rank-based statistics, such as the median, i.e., the 50th percentile, are more robust to outliers.
- E.g. Consider two sets of numbers $\mathcal{X} = \{1, 2, \ldots, 10, 11\}$ and $\mathcal{Y} = \{1, 2, \ldots, 10, 110\}$, identical except for one value. Then, the means of $\mathcal{X}$ and $\mathcal{Y}$ are 6 and 15, respectively, while the median is 6 for both datasets.

Robust Statistics

- The median Median$[X]$ of a random variable X is any number m such that

$$\Pr(X \leq m) \geq \frac{1}{2} \qquad \text{and} \qquad \Pr(X \geq m) \geq \frac{1}{2}. \qquad (28)$$

Alternately,

$$F_X(m) \geq \frac{1}{2} \qquad \text{and} \qquad F_X(m^-) \leq \frac{1}{2}.$$

For a continuous random variable with CDF F_X

$$F_X(m) = \Pr(X \leq m) = \Pr(X \geq m) = \frac{1}{2}.$$

Robust Statistics

- A robust measure of the variability of a distribution (to be used in place of the standard deviation) is the median absolute deviation (MAD)

$$\mathrm{MAD}[X] \equiv \mathrm{Median}[|X - \mathrm{Median}[X]|]. \tag{29}$$

- E.g. The SDs of $\mathcal{X}$ and $\mathcal{Y}$ are 3.32 and 31.64, respectively, while the MAD is 4.45 for both datasets.

Notation

- When needed, we may use a subscript to indicate the distribution with respect to which an expected value, variance, covariance, correlation, or median is computed.
- For instance, $\mathrm{E}_P[X]$ and $\mathrm{E}_{P_n}[X]$ refer, respectively, to a population and an empirical/sample mean.

Estimators and Their Sampling Distributions

- In the context of statistical inference/learning, one seeks to infer/learn a population parameter $\theta = \Theta(P)$ based on data from a random sample from the population of interest.
- E.g. In regression, the parameter of interest is the regression function for the population, i.e., the conditional mean $\mathrm{E}_P[Y|X]$ of an outcome Y given covariates X, where the expected value is computed with respect to the unknown data generating distribution P.
- A random sample of size n can be represented by n random variables $\{X_i : i = 1, \ldots, n\}$. These are the data.

Estimators and Their Sampling Distributions

- When sampling at random with replacement from a population, the random variables are independent and identically distributed (IID), with distribution the data generating distribution P. That is, $X_i \sim P$.
- For a given sample, the data empirical distribution is the discrete distribution P_n that places probability $1/n$ on each X_i.
- An estimator $\hat{\theta}_n$ is a function of the data, i.e., a function of $\{X_i : i = 1, \ldots, n\}$ or, equivalently, P_n: $\hat{\theta}_n = \hat{\Theta}(P_n)$.
- An estimator is therefore a random variable.
- In frequentist inference, the sampling distribution of an estimator refers to its distribution over repeated random samples from the population.

Estimators and Their Sampling Distributions

- Useful parameters of the sampling distribution of an estimator, that can be used to assess its performance, are related to its expected value and its variance, computed with respect to the data generating distribution P.
- The bias of an estimator $\hat{\theta}_n$ of the parameter θ is the difference between its expected value and θ

$$\text{Bias}_P[\hat{\theta}_n] \equiv \text{E}_P[\hat{\theta}_n] - \theta. \tag{30}$$

- The estimator is said to be unbiased if its expected value is equal to the parameter it seeks to estimate, i.e., if $\text{Bias}_P[\hat{\theta}_n] = 0$.

Estimators and Their Sampling Distributions

- The standard error (SE) of an estimator is the square root of its variance, i.e., its standard deviation.

$$\mathrm{SE}_P[\hat{\theta}_n] \equiv \sqrt{\mathrm{Var}_P[\hat{\theta}_n]}. \tag{31}$$

- The mean squared error (MSE) of an estimator $\hat{\theta}_n$ of the parameter θ is the expected value of its squared difference with θ

$$\mathrm{MSE}_P[\hat{\theta}_n] \equiv \mathrm{E}_P[(\hat{\theta}_n - \theta)^2]. \tag{32}$$

Estimators and Their Sampling Distributions

- One can show that the MSE is the sum of the variance and of the square of the bias

$$\mathrm{MSE}[\hat{\theta}_n] = \mathrm{Var}[\hat{\theta}_n] + (\mathrm{Bias}[\hat{\theta}_n])^2. \tag{33}$$

  This result holds for expected values computed with respect to any distribution (we therefore did not use a subscript for the distribution).

Estimators and Their Sampling Distributions

Proof.

$$\begin{aligned}
\mathrm{MSE}[\hat{\theta}_n] &= \mathrm{E}[(\hat{\theta}_n - \theta)^2] \\
&= \mathrm{E}[(\hat{\theta}_n - \mathrm{E}[\hat{\theta}_n] + \mathrm{E}[\hat{\theta}_n] - \theta)^2] \\
&= \mathrm{E}[(\hat{\theta}_n - \mathrm{E}[\hat{\theta}_n])^2 + 2(\hat{\theta}_n - \mathrm{E}[\hat{\theta}_n])(\mathrm{E}[\hat{\theta}_n] - \theta) \\
&\quad +(\mathrm{E}[\hat{\theta}_n] - \theta)^2] \\
&= \mathrm{E}[(\hat{\theta}_n - \mathrm{E}[\hat{\theta}_n])^2] + 2(\mathrm{E}[\hat{\theta}_n] - \theta)\,\mathrm{E}[(\hat{\theta}_n - \mathrm{E}[\hat{\theta}_n])] \\
&\quad + \mathrm{E}[(\mathrm{E}[\hat{\theta}_n] - \theta)^2] \\
&= \mathrm{Var}[\hat{\theta}_n] + 2(\mathrm{E}[\hat{\theta}_n] - \theta) \times 0 + (\mathrm{E}[\hat{\theta}_n] - \theta)^2 \\
&= \mathrm{Var}[\hat{\theta}_n] + (\mathrm{Bias}[\hat{\theta}_n])^2.
\end{aligned}$$

□

Estimators and Their Sampling Distributions

- Ideally, we would like estimators to be both unbiased (i.e., on average equal to the parameter of interest) and have low standard error (i.e., low variability around their average). However, as we will discuss later, there is a bias-variance trade-off.

Example: Binomial Distribution

Flipping a biased coin.

- Suppose one has a biased coin and one wishes to estimate the probability p that it lands heads on a any given flip.
- A natural estimator of p is obtained by flipping the coin independently n times and recording the proportion of heads.
- Let X_i denote the binary indicator for the outcome of the ith coin flip, equal to 1 if it lands heads ("success") and 0 if it lands tails ("failure").
- The X_i are n independent and identically distributed Bernoulli(p) random variables with "success" probability p, i.e., $\Pr(X_i = 1) = p$.

Example: Binomial Distribution

- The mean of a Bernoulli(p) random variable X is

$$E[X] = 0 \times (1 - p) + 1 \times p = p$$

and its variance

$$\begin{aligned} \text{Var}[X] &= E[X^2] - (E[X])^2 \\ &= 0^2 \times (1 - p) + 1^2 \times p - p^2 \\ &= p(1 - p). \end{aligned}$$

- Let Y_n denote the sum of the n random variables $\{X_i : i = 1, \ldots, n\}$, i.e., the total number of heads in the n flips.

Example: Binomial Distribution

- Then, the support of Y_n is $\mathcal{Y}_n = \{0, 1, \ldots, n\}$ and Y_n has a Binomial(n, p) distribution with PMF

$$f_{Bin}(y; n, p) = \binom{n}{y} p^y (1-p)^{n-y}, \qquad y \in \mathcal{Y}_n.$$

- By the linearity property of expectations, the mean of Y_n is

$$\text{E}[Y_n] = \text{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \text{E}[X_i] = np.$$

- By the linearity property of variances for independent random variables, the variance of Y_n is

$$\text{Var}[Y_n] = \text{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \text{Var}[X_i] = np(1-p).$$

Example: Binomial Distribution

- The proportion of heads in n flips, $\hat{p}_n = Y_n/n$, is a natural estimator of the parameter p. It is unbiased

$$\mathrm{E}[\hat{p}_n] = \frac{1}{n}\mathrm{E}[Y_n] = p$$

and its variance decreases with the number of flips

$$\mathrm{Var}[\hat{p}_n] = \frac{1}{n^2}\mathrm{Var}[Y_n] = \frac{p(1-p)}{n}.$$

- Furthermore, according to the Central Limit Theorem, $\hat{p}_n$ is approximately normally distributed for large n

$$\hat{p}_n \sim \mathrm{N}\left(p, \sqrt{\frac{p(1-p)}{n}}\right).$$

Example: Binomial Distribution

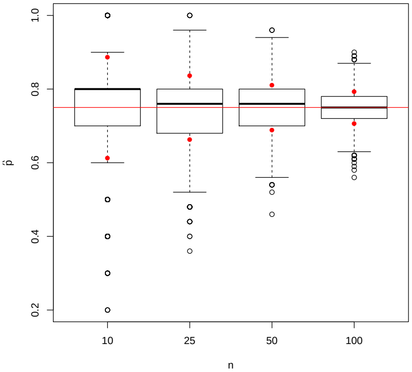


Figure 15: *Binomial distribution.* Sampling distribution of proportion of heads $\hat{p}_n$ for $n \in \{10, 25, 50, 100\}$ (over 10,000 samples). Solid red line indicates mean and red dots plus/min one SE from mean.

Example: Binomial Distribution

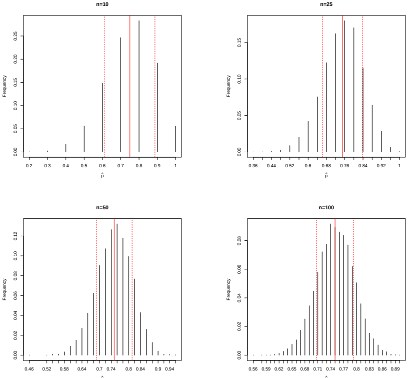


Figure 16: *Binomial distribution.* Sampling distribution of proportion of heads $\hat{p}_n$ for $n \in \{10, 25, 50, 100\}$ (over 10,000 samples). Solid red line indicates mean and dashed red lines plus/min one SE from mean. Note different x-axis scales.

Example: Binomial Distribution

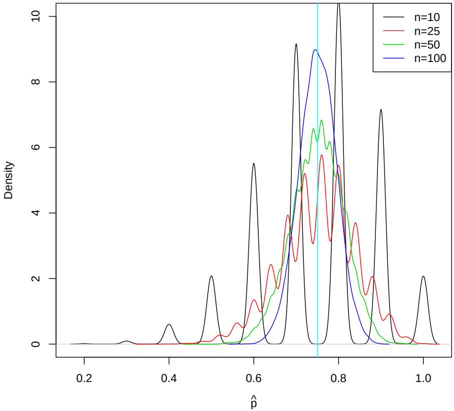


Figure 17: *Binomial distribution.* Sampling distribution of proportion of heads $\hat{p}_n$ for $n \in \{10, 25, 50, 100\}$ (over 10,000 samples).

Statistical Models

"All models are wrong, but some are useful." (Box, 1976)

Statistical Models

- A statistical model is a set of distributions for a random variable of interest.
- When one focuses on certain families of distributions (e.g., Gaussian distributions) or types of parameters of a distribution (e.g., regression function), a model can correspond to a set of parameter values.
- A model is an idealized representation of reality.
- Models involve assumptions about the data generating mechanism and are used to make inference from the sample to the population.

Statistical Models

- E.g. Assumptions about PDF.
  Let Y denote the height of a random Berkeley student. A possible model for Y is the set of all Gaussian distributions $\mathrm{N}(\mu, \sigma)$, with mean $\mu \in \mathbb{R}^+$ and standard deviation $\sigma \in \mathbb{R}^+$.
- E.g. Assumptions about regression function.
  Let Y denote the rent of a random Berkeley apartment and let X_1 denote the number of bedrooms, X_2 the number of bathrooms, X_3 the square footage, X_4 an indicator for washer/dryer availability, and

Statistical Models

$X = (X_1, X_2, X_3, X_4)$. Possible models for the expected rent are

$$\begin{aligned}
\mathrm{E}[Y|X] &= \beta_0 \\
\mathrm{E}[Y|X] &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 \\
\mathrm{E}[Y|X] &= \beta_0 + \sum_{k=1}^{K} \beta_k \, \mathrm{I}(X \in A_k) \\
\mathrm{E}[Y|X] &= f(X),
\end{aligned}$$

where the sets A_k form a partition of the covariate space and f denotes an arbitrary function of the covariates X.

- The first constant model does not account for the dependence of rent on obvious covariates.

Statistical Models

- The second model is a linear regression model. It accounts for obvious covariates, but involves adding incommensurate variables and could miss non-linear relationships and interactions.
- The third model is a regression tree model. It is well-suited for covariates of different types and interactions, but could lead to unstable estimators.
- The fourth general non-parametric model could be fit by, e.g., robust local regression (e.g., loess), which does not provide a simple interpretable regression function f. The function could be very data-adaptive at the risk of overfitting, i.e., fit the sample data very closely but not additional data from the same population.

Statistical Models

- E.g. Independence assumptions.
  A pervasive assumption is that units in the sample are drawn independently from the population, i.e., IID random variables $\{X_1, \ldots, X_n\}$.
  Independence assumptions justify multiplying probabilities for joint distributions and adding variances for sums of random variables.
- Model assumptions are often unrealistic and hard to verify.
  E.g. Independence assumptions.
- Wrong assumptions can lead to wrong inference.
- Inference results should be driven by the data and not by model assumptions. Cf. Robustness.

Statistical Models

- Models are often wrong, i.e., do not accurately represent how the data were generated, but can still be useful, e.g., yield accurate predictions.
- When selecting a regression model, there can be a trade-off between intepretability and predictive accuracy (Breiman, 2001).
  E.g. Simple parametric models (e.g., linear regression model) are easier to interpret but potentially lead to less accurate predictions than more complex or "black box" models (e.g., ensembles of regression trees, neural networks, loess).

Statistical Models

- When selecting a model, there is also a trade-off between bias and variance. More data-adaptive/complex models tend to have less bias but larger variance.
  E.g. Robust local regression (loess) and kernel density estimation with different bandwidths.
- In particular, a model that fits the learning data very closely often "generalizes" poorly on new test data.
- Multiple models can lead to the same fit/result.
  E.g. Linear regression with polynomials of different degrees can lead to similar fitted values and MSE (see below).

Statistical Models

- The distinction between parametric and non-parametric models can be problematic.
- There is no clear dichotomy, but rather a continuum, in the degree of "parametricity" of distributions and methods.
- E.g. In non-parametric density estimation, the parameter can be the entire density function, under some conditions such as smoothness. In non-parametric regression, the parameter can be the entire regression function $\mathrm{E}[Y|X]$.
- The distinction may have been more relevant historically.
- Idem for the terms model-based and model-free.
- Cf. Terence's Stuff: A Rose ... in *IMS Bulletin*, Volume 37, Issue 7, August/September 2008.

Statistical Models: Example

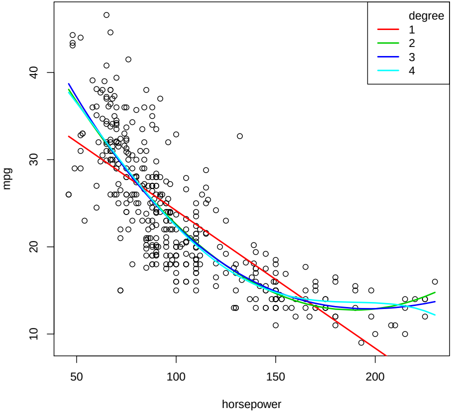


Figure 18: *mpg dataset.* Linear regression of mpg on horsepower, polynomials of degree 1 to 4.

Statistical Models: Example

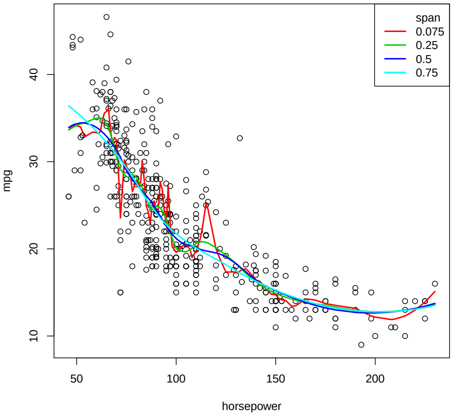


Figure 19: *mpg dataset.* Robust local regression (loess) of mpg on horsepower, 4 different spans.

Statistical Models: Example

Table 3: *mpg dataset.* Empirical mean squared error for linear regression and robust local regression (loess) of mpg on horsepower.

	Linear regression			
Degree	1	2	3	4
MSE	23.94	18.98	18.94	18.88

	Robust local regression			
Span	0.075	0.250	0.500	0.750
MSE	14.99	17.31	18.00	18.38

Statistical Models

Essential aspects of statistical inference include:

- Identifying/selecting an appropriate model.
- Fitting the model to the data, i.e., deriving an "optimal" estimator given the model.
- Assessing the performance of the model. Cf. Goodness-of-fit (e.g., residual analysis), accuracy of estimator/prediction, robustness of the results to modeling assumptions.

Loss functions and risk play an essential role for each of these issues.

Loss Functions and Risk

- A loss function is a real-valued function L of a random variable X and a parameter value θ (not necessarily the true value):

$$L(X, \theta) \in \mathbb{R}\,.$$

- As the name suggests, loss functions are measures of performance, indicating how far a parameter value is from the data.
- Examples of loss functions, used in different inference contexts, are listed in the table below.

Loss Functions and Risk

Table 4: *Loss functions.*

Name	Definition	Application
Squared error, L_2	$(Y - \theta(X))^2$	Regression – least squares estimation
Absolute error, L_1	$\lvert Y - \theta(X) \rvert$	Regression
Indicator, zero-one	$\mathrm{I}(Y, \theta(X))$	Classification
Negative log	$-\log \theta(X)$	Density estimation – maximum likelihood estimation

- Risk is the expected value of a loss function

$$R(P, \theta) \equiv \mathrm{E}_P[L(X, \theta)], \tag{34}$$

where P denotes the distribution of X with respect to which the expected value is computed.

Loss Functions and Risk

- It is essential to note that risk can be computed with respect to different distributions, e.g., the true unknown data generating distribution P, the known data empirical distribution P_n.
- Both parameters and estimators thereof can be defined as risk minimizers for a suitably defined loss function.
    - Parameters minimize risk with respect to the typically unknown data generating distribution.
    - Estimators minimize risk with respect to the known data empirical distribution.
- Optimal statistical inference. A very broad class of statistical inference methods can be framed in terms of risk optimization.
    - Least squares estimation (LSE) involves minimizing risk for the squared error loss function.

Loss Functions and Risk

- Maximum likelihood estimation (MLE) involves minimizing risk for the negative log loss function.

Squared Error Loss Function

- One of the most widely used loss functions is the squared error loss function or L_2 loss function

$$L_2(X, \theta) \equiv (X - \theta)^2. \tag{35}$$

- For the squared error loss function, risk is the mean squared error (MSE),

$$\begin{aligned} R_2(P, \theta) &\equiv E_P[(X - \theta)^2] \qquad \text{population risk} \qquad (36) \\ &= \begin{cases} \sum_{x \in \mathcal{X}} (x - \theta)^2 f_X(x) & \text{(discrete)} \\ \int_{x \in \mathcal{X}} (x - \theta)^2 f_X(x) dx & \text{(continuous)} \end{cases} \\ R_2(P_n, \theta) &= \frac{1}{n} \sum_{i=1}^{n} (X_i - \theta)^2 \qquad \text{empirical risk.} \end{aligned}$$

Squared Error Loss Function

- Expected values minimize risk for the squared error loss function.

  The expected value $\mathrm{E}_P[X]$ minimizes risk for the squared error loss function with respect to the distribution P

  $$\mathrm{argmin}_{\theta \in \mathbb{R}} R_2(P, \theta) = \mathrm{argmin}_{\theta \in \mathbb{R}} \mathrm{E}_P[(X - \theta)^2] = \mathrm{E}_P[X]. \tag{37}$$

  Proof. We proved earlier that the MSE is equal to the variance plus bias squared, thus

  $$\mathrm{E}_P[(X - \theta)^2] = \mathrm{Var}_P[X] + (\mathrm{E}_P[X] - \theta)^2$$

  and risk is minimized wrt θ when bias is equal to zero, that is, $\theta = \mathrm{E}_P[X]$. $\square$

- The result holds for any distribution P.

Squared Error Loss Function

- In particular, the empirical/sample average $\bar{X}_n$ minimizes empirical MSE (known), i.e., MSE with respect to the data empirical distribution P_n

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \mathrm{E}_{P_n}[X] = \mathrm{argmin}_{\theta \in \mathbb{R}} \frac{1}{n}\sum_{i=1}^{n}(X_i - \theta)^2. \tag{38}$$

- The population mean $\mathrm{E}_P[X]$ minimizes the population MSE (unknown), i.e., MSE with respect to the data generating distribution P.

Absolute Error Loss Function

- With the squared error loss function, risk optimization amounts to defining parameters and estimators as population and sample means, respectively.
- However, means can be sensitive to outliers, i.e., observations that are distant from most of the other observations.
- By contrast, medians are more robust to outliers.
- When robustness is an issue, it makes sense to consider a loss function that measures discrepancies between random variables and parameter values in terms of absolute differences rather than squared differences, as squaring tends to amplify differences.

Absolute Error Loss Function

- The absolute error loss function or L_1 loss function is defined as
$$L_1(X, \theta) \equiv |X - \theta|. \tag{39}$$
- Risk for the absolute error loss function is the mean absolute error (MAE)
$$R_1(P, \theta) \equiv \mathrm{E}_P[|X - \theta|]. \tag{40}$$
- The mean absolute error is minimized by the median
$$\mathrm{argmin}_{\theta \in \mathbb{R}}\, R_1(P, \theta) = \mathrm{argmin}_{\theta \in \mathbb{R}}\, \mathrm{E}_P[|X - \theta|] = \mathrm{Median}_P[X]. \tag{41}$$

Absolute Error Loss Function

Proof. Use the short-hand notation $m = \text{Median}_P[X]$ and suppose $\theta > m$. Then,

$$\begin{aligned} |X-\theta| - |X-m| &= \begin{cases} \theta - m, & X \leq m \\ \theta + m - 2X, & m < X \leq \theta \\ m - \theta, & X > \theta \end{cases} \\ &\geq \begin{cases} \theta - m, & X \leq m \\ m - \theta, & X > m \end{cases}. \end{aligned}$$

Absolute Error Loss Function

Hence,

$$\begin{aligned}
&\mathrm{E}[|X-\theta|]-\mathrm{E}[|X-m|]\\
&\qquad \geq \mathrm{E}\left[(\theta-m)\,\mathrm{I}(X\leq m)+(m-\theta)\,\mathrm{I}(X>m)\right]\\
&\qquad = (\theta-m)\,\Pr(X\leq m)+(m-\theta)\,\Pr(X>m)\\
&\qquad = (\theta-m)\,\Pr(X\leq m)+(m-\theta)(1-\Pr(X\leq m))\\
&\qquad = 2(\theta-m)\,\Pr(X\leq m)+(m-\theta)\\
&\qquad \geq 0,
\end{aligned}$$

as $\Pr(X \leq m) \geq 1/2$. The proof for $\theta < m$ is similar. □

Huber Loss Function

- The squared error/L_2 loss function is more convenient mathematically than the absolute error/L_1 loss function.
- Risk for the L_2 loss function, i.e., the MSE, has a unique minimizer (the mean), whereas risk for the L_1 loss function, i.e., the MAE, can have multiple minimizers (non-uniqueness of the median).
- However, the MSE and the mean are more sensitive to outliers than the MAE and the median.
- The Huber loss function is a compromise between the L_1 and L_2 loss functions defined as

$$L_H(X,\theta) \equiv \begin{cases} \frac{1}{2}(X-\theta)^2, & |X-\theta| \leq \delta \\ \delta\left(|X-\theta| - \frac{1}{2}\delta\right), & \text{otherwise} \end{cases}, \tag{42}$$

where $\delta \in \mathbb{R}^+$ is a tuning parameter.

Huber Loss Function

- The Huber loss function is quadratic for small differences and linear for large differences, thus more robust to outliers than the L_2 loss function.
- There are no closed-form expressions for the risk minimizer for the Huber loss function. Instead, one can use optimization methods such as gradient descent.

Example: tips Dataset

- A particular waiter is interested in inferring the tip percentage he could expect. He collected the following data on all $n = 244$ tables he served during a month of employment: Total bill, tip, sex of customer tipping, smoking status of customer, day, time, and size of party.

	total_bill	tip	sex	smoker	day	time	size	tip_percent
1	16.99	1.01	Female	No	Sun	Dinner	2	0.06
2	10.34	1.66	Male	No	Sun	Dinner	3	0.16
3	21.01	3.50	Male	No	Sun	Dinner	3	0.17
4	23.68	3.31	Male	No	Sun	Dinner	2	0.14
5	24.59	3.61	Female	No	Sun	Dinner	4	0.15
6	25.29	4.71	Male	No	Sun	Dinner	4	0.19
					...			

- In the USA, a typical tip is 15% of the total bill. Thus, we expect a linear relationship between tip and total bill.

Example: tips Dataset

- The mean tip percentage is 16.08%, most tips are between 10 and 20%, with a few outlying large tips (maximum of 70%).
- The tip percentage does not appear, however, to vary much with variables such as sex, smoker, day, time, and size.
- We therefore consider a constant model for the tip percentage Y

$$\mathrm{E}[Y] = \theta.$$

- What is a "good" estimator of θ? This can be defined in terms of a loss function.

Example: tips Dataset

- For the squared error loss function, the empirical risk minimizer is the empirical average

$$\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i = 0.16.$$

- For the absolute error loss function, the empirical risk minimizer is the empirical median

$$\tilde{Y}_n = \text{Median}_{P_n}[Y] = 0.15.$$

- Let's consider the following estimates of the mean tip percentage θ and examine how well they fit the data, i.e., how close they are to the observed tip percentages in terms of the empirical risk for both the L_2 and L_1 loss functions

$$\hat{\theta}_n = 0.10, 0.15, \tilde{Y}_n, \bar{Y}_n, 0.20.$$

Example: tips Dataset

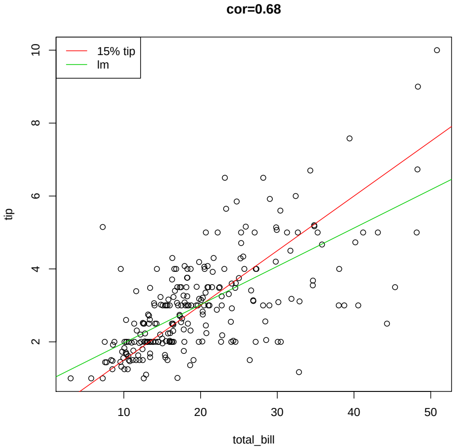


Figure 20: *tips dataset.* Tip vs. total bill.

Example: tips Dataset

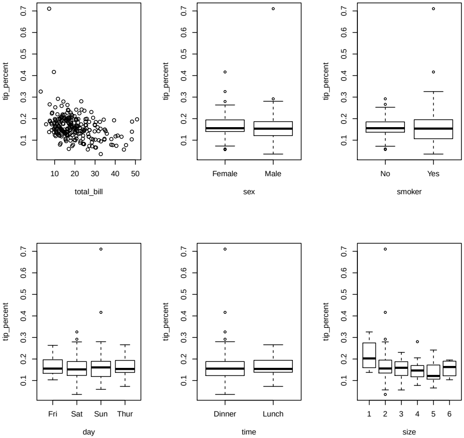


Figure 21: *tips dataset.* Tip percentage vs. covariates.

Example: tips Dataset

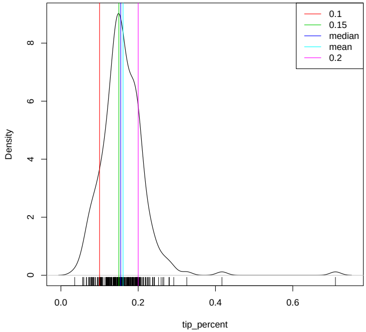


Figure 22: *tips dataset.* Gaussian kernel density estimator for tip percentage θ. Vertical lines indicates different estimates of the mean tip percentage, $\hat{\theta}_n = 0.10, 0.15, \tilde{Y}_n, \bar{Y}_n, 0.20$.

Example: tips Dataset

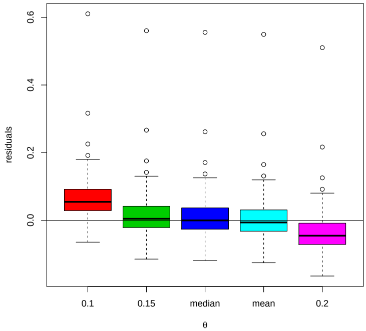


Figure 23: *tips dataset.* Boxplots of residuals for different estimates of the mean tip percentage, $\hat{\theta}_n = 0.10, 0.15, \tilde{Y}_n, \bar{Y}_n, 0.20$.

Example: tips Dataset

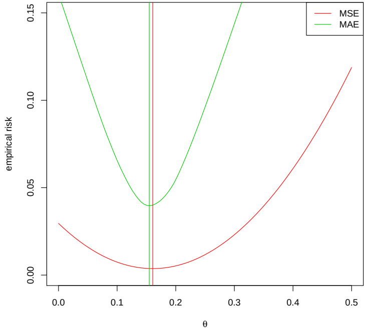


Figure 24: *tips dataset.* Empirical MSE and MAE as a function of mean tip percentage θ. Vertical lines indicates empirical mean (red) and median (green).

Example: tips Dataset

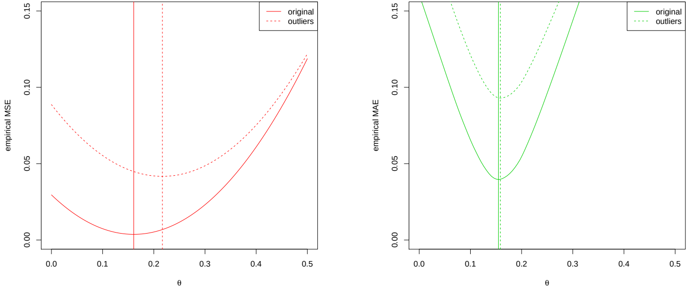


Figure 25: *tips dataset.* Empirical MSE and MAE as a function of mean tip percentage θ, when large outlying tip percentages are included in the dataset. Vertical lines indicates empirical mean (red) and median (green).

Example: tips Dataset

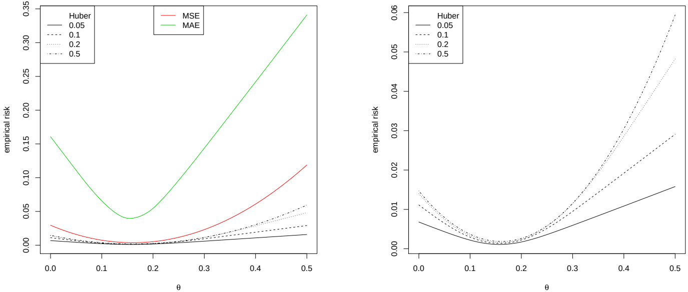


Figure 26: *tips dataset.* Empirical Huber risk, MSE, and MAE as a function of mean tip percentage θ. Right panel is zoom on Huber risk.

Example: tips Dataset

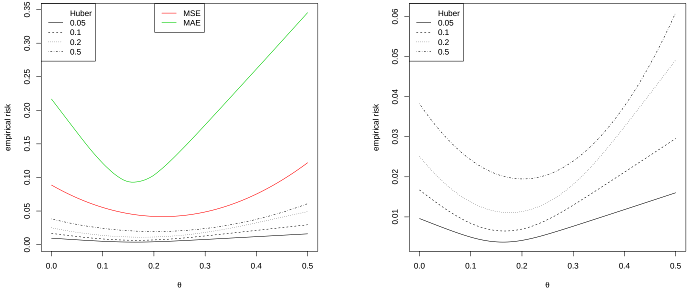


Figure 27: *tips dataset.* Empirical Huber risk, MSE, and MAE as a function of mean tip percentage θ, when large outlying tip percentages are included in the dataset. Right panel is zoom on Huber risk.

References

G. E. P. Box. Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799, 1976.

L. Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16 (3):199–215, 2001.