Multivariate Verfahren

4. Multivariate Distributions

Hannah Schulz-Kümpel

Institut für Statistik, LMU München

Summer Semester 2024

Contents

1. Concepts and examples for RVs with univariate distribution
2. Pairs of random variables
   - Joint, marginal, and conditional distributions
   - Covariance and Correlation
3. Theoretical side-note: Fubini's theorem
4. Multivariate Distributions
   - Extending the concepts to vector notation
   - Relevant examples
5. Estimating distributions and characteristics from data
   - Empirical mean, (co)variance, and correlation
6. Side Note: Categorical variables
   - Simply Estimating the data's distribution "from scratch"
   - More complicated procedures to estimate distributions

Recap: Expectation and Variance I

- The **expected value** indicates the average value of a random variable.
- Given a probability space $(\Omega, \mathcal{F}, \mathrm{P})$ any random variable X that is integrable w.r.t. P, it is defined as $\mathbb{E}[X] = \int_\Omega X(\omega) \mathrm{d}\,\mathrm{P}(\omega)$. (*integrable w.r.t.* P simply means $\mathbb{E}[|X|] = \int_\Omega |X(\omega)| \mathrm{d}\,\mathrm{P}(\omega) < \infty$.)
- In practice, however, corresponding to the probability density/mass function, the expected value is often defined separately for continuous and random variables (in an equivalent but easier to read way):

Recap: Expectation and Variance II

Definition (Expected value)

- For a *continuous random variable* X with distribution defined via density f the expected value is defined as

$$\mathbb{E}[X] = \int_{\mathbb{R}} x \cdot f(x)\mathrm{d}x\,.$$

- For a *discrete random variable* X with distribution defined via probability function p the expected value is defined as

$$\mathbb{E}[X] = \sum_{x \in \mathrm{supp}(p)} x \cdot p(x)\,.$$

Recap: Expectation and Variance III

Some rules that follow directly from the corresponding properties of the integral:

- *Linearity*: For $c \in \mathbb{R}$ and real, integrable random variables X, Y on the probability space $(\Omega, \mathcal{F}, \mathrm{P})$ we have
  - The random variable $Z := cX$ is clearly also an integrable random variable on $(\Omega, \mathcal{F}, \mathrm{P})$ and $\mathbb{E}[Z] = \mathbb{E}[cX] = c\mathbb{E}[X]$.
  - The random variable $Z := X + Y$ is clearly also an integrable random variable on $(\Omega, \mathcal{F}, \mathrm{P})$ and $\mathbb{E}[Z] = \mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.
- *Triangle inequality*: For a real, integrable random variable X on the probability space $(\Omega, \mathcal{F}, \mathrm{P})$ it holds that $|\mathbb{E}[X]| \leq \mathbb{E}[|X|]$.

Recap: Expectation and Variance IV

- The **variance** of a random variable X is denoted by $\mathrm{Var}(X)$, $\mathbb{V}(X)$, or simply σ^2, if the context does not require the RV to be specified.

- Given a probability space $(\Omega, \mathcal{F}, \mathrm{P})$ any random variable X that is square integrable w.r.t. P, it is defined as

$$\mathrm{Var}(X) = \mathbb{E}\big[(X - \mathbb{E}[X])^2\big] \, .$$

  (*Square integrable w.r.t.* P simply means $\mathbb{E}[|X^2|] = \int_\Omega |X(\omega)^2| \mathrm{d}\,\mathrm{P}(\omega) < \infty$.)

- The **standard deviation** of a random variable is *a measure of how dispersed the data is in relation to the mean*. It is often denoted by σ and given by the square root of the variance, i.e. $\sigma = \sqrt{\mathrm{Var}(X)}$.

Recap: Expectation and Variance V

Alternative representation of Variance

Given the Linearity of the expected value, it immediately follows that we can also write the variance of a random variable X as the mean of the square of X minus the square of the mean of X:

$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{E}\left[(X - \mathrm{E}[X])^2\right] \\ &= \mathrm{E}\left[X^2 - 2X\,\mathrm{E}[X] + \mathrm{E}[X]^2\right] \\ &= \mathrm{E}\left[X^2\right] - 2\,\mathrm{E}[X]\,\mathrm{E}[X] + \mathrm{E}[X]^2 \\ &= \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2 .\end{aligned}$$

Recap: Expectation and Variance VI

Some helpful basic properties of the variance of a random variable X are, for some constant $a \in \mathbb{R}$:

- $\mathrm{Var}(X) \geq 0$,
- $\mathrm{Var}(a) = 0$,
- $\mathrm{Var}(X + a) = \mathrm{Var}(X)$,
- $\mathrm{Var}(aX) = a^2 \, \mathrm{Var}(X)$.

Relevant characteristics of distributions

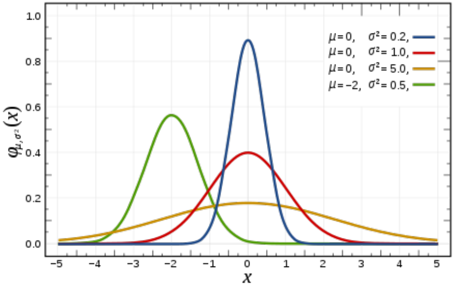
The next slides will summarize some relevant univariate distributions, giving the following characteristics for each:

- **discrete or continous** - i.e. is the distribution defined via a *(probability) density (function)* or a *probability (mass) function*?
- The **probability density/mass function** and its
  - **Parameters**
  - **Support** - i.e. the subset of the domain of the defining probability density/mass function containing those elements that are not mapped to 0.
- The **expected value** and **variance** of any random variable following the distribution.

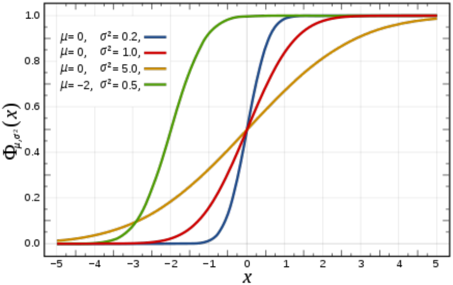
Normal distribution - continuous

- Notation: $X \sim \mathcal{N}(\mu, \sigma^2)$
- Density: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
- Parameters: $\mu \in \mathbb{R}$ (location), $\sigma^2 \in \mathbb{R}_{>0}$ (scale)
- Support: $\mathbb{R}$
- $\mathbb{E}[X] = \mu$; $\mathrm{Var}[X] = \sigma^2$

Density plots

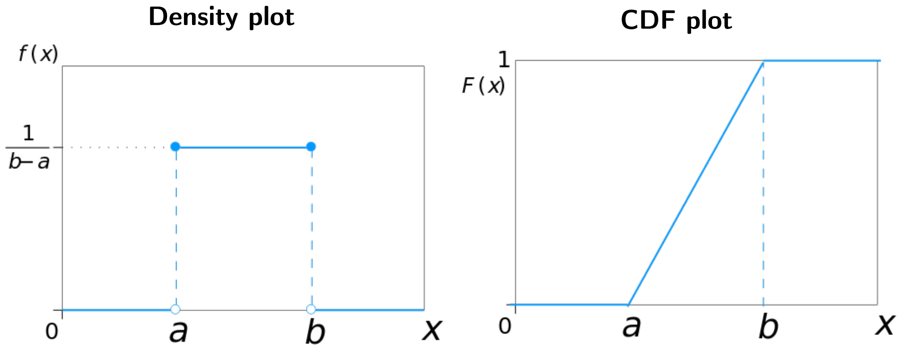


CDF plots



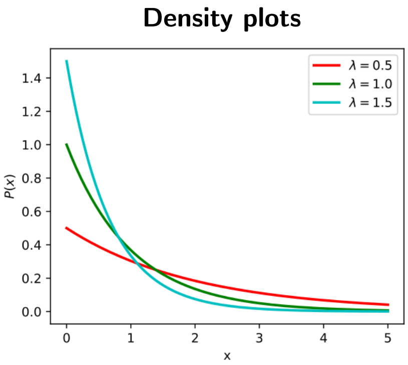
(Continuous) Uniform distribution - continuous

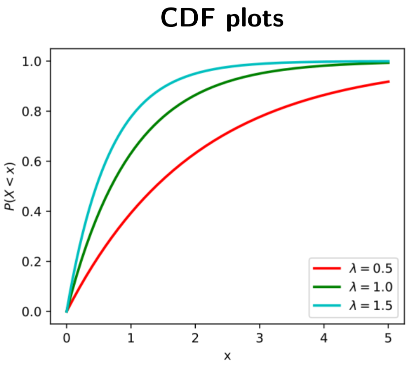
- Notation: $X \sim U(a,b)$
- Density: $f(x) = \begin{cases} \frac{1}{b-a} & \text{for } x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$
- Parameters: $a, b, \in \mathbb{R}$ with $a < b$
- Support: $[a,b]$
- $\mathbb{E}[X] = \frac{1}{2}(a+b)$; $\text{Var}[X] = \frac{1}{12}(b-a)^2$



Exponential distribution - continuous

- Notation: $X \sim \text{Exp}(\lambda)$
- Density: $f(x) = \lambda e^{-\lambda x}$
- Parameters: $\lambda \in \mathbb{R}_{>0}$ (rate)
- Support: $\mathbb{R}_{\geq 0}$
- $\mathbb{E}[X] = \frac{1}{\lambda}$; $\text{Var}[X] = \dfrac{1}{\lambda^2}$

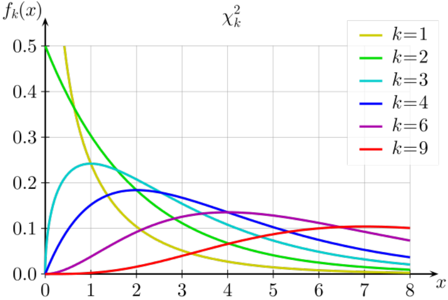




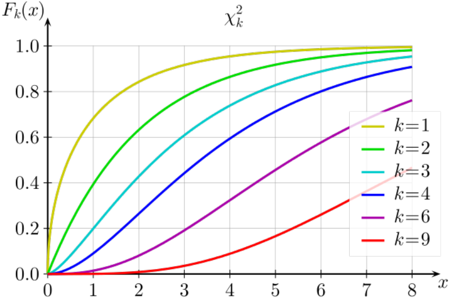
χ^2 distribution - continuous

- Notation: $X \sim \chi^2$ or χ^2_k
- Density: $f(x) = \dfrac{1}{2^{k/2}\Gamma(k/2)}\, x^{k/2-1}e^{-x/2}$
- Parameters: $k \in \mathbb{N}$ (degrees of freedom)
- Support: $\mathbb{R}_{\geq 0}$, or $\mathbb{R}_{>0}$ if $k = 1$
- $\mathbb{E}[X] = k$; $\mathrm{Var}[X] = 2k$

Density plots



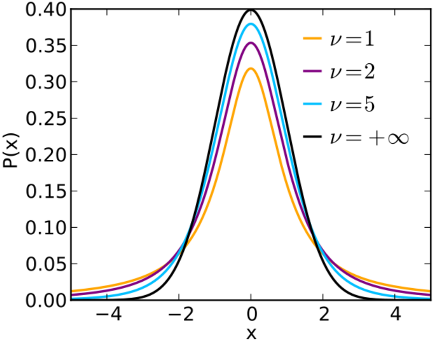
CDF plots



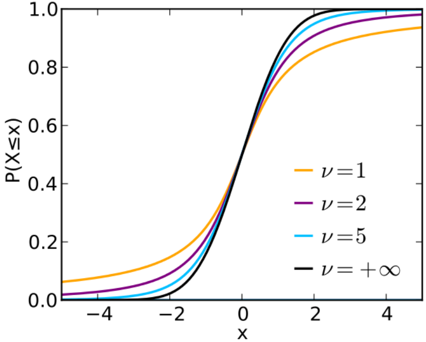
Student's-t distribution - continuous

- Notation: $X \sim t_\nu$
- Density: $f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$
- Parameters: $\nu \in \mathbb{R}_{>0}$ (degrees of freedom)
- Support: $\mathbb{R}$
- $\mathbb{E}[X] = 0$ for $\nu > 1$, else undefined; $\mathrm{Var}[X] = \frac{\nu}{\nu-2}$ for $\nu > 2$, else undefined

Density plots



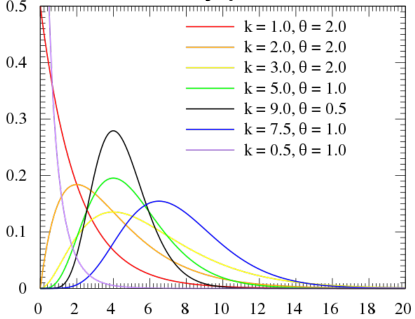
CDF plots



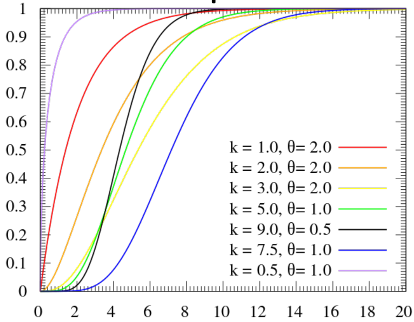
Gamma distribution - continuous

- Notation: $X \sim \Gamma(k, \frac{1}{\theta})$ or $\mathrm{Gamma}(k, \frac{1}{\theta})$
- Density: $f(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}}$
- Parameters: $k, \theta \in \mathbb{R}_{>0}$ (shape, scale)   Note: there is an alternative parametrization
- Support: $\mathbb{R}_{>0}$
- $\mathbb{E}[X] = k\theta$; $\mathrm{Var}[X] = k\theta^2$

Density plots

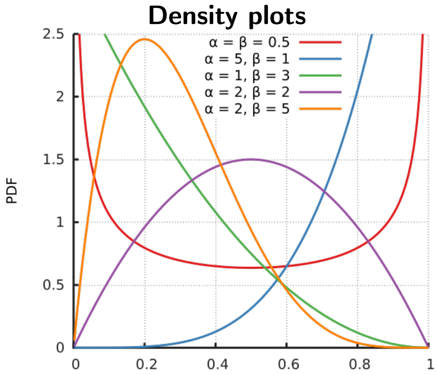


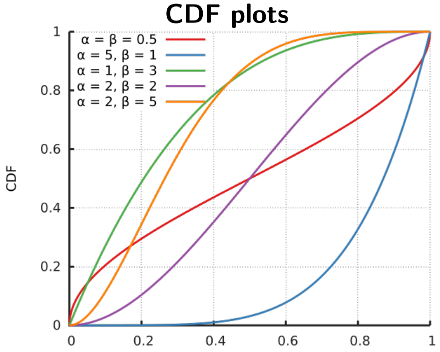
CDF plots



Beta distribution - continuous

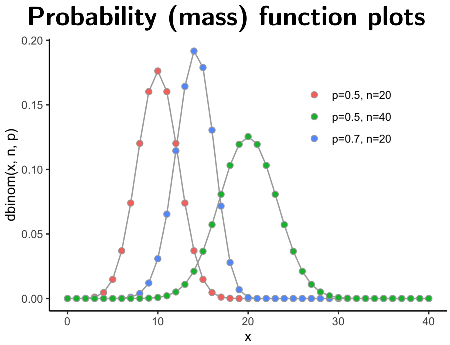
- Notation: $X \sim \mathrm{Beta}(\alpha, \beta)$
- Density: $f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}$ with $\mathrm{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$
- Parameters: $\alpha, \beta \in \mathbb{R}_{>0}$
- Support: $[0, 1]$
- $\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}$; $\mathrm{Var}[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

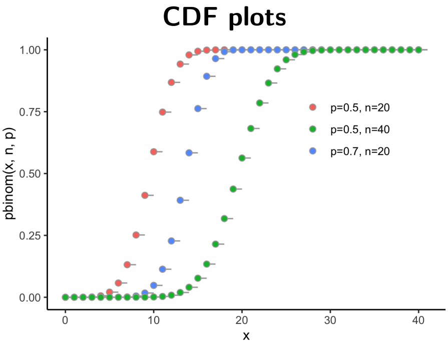




Binomial distribution - discrete

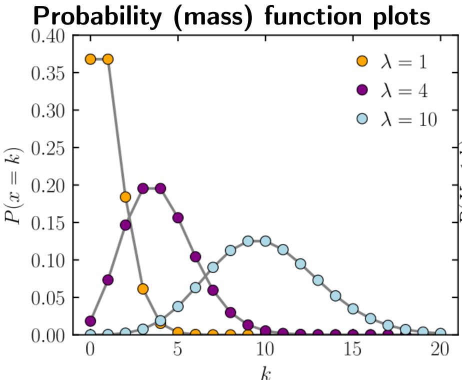
- Notation: $X \sim B(n, p)$
- Probability (mass) function: $p(x) = \binom{n}{x} p^x q^{n-x}$
- Parameters: $n \in \mathbb{N}_0$, $p \in [0, 1]$, $q = 1 - p$
  (number of trials, success probability for each trial, complementary probability)
- Support: $\mathbb{N}_0$
- $\mathbb{E}[X] = np$; $\mathrm{Var}[X] = npq$

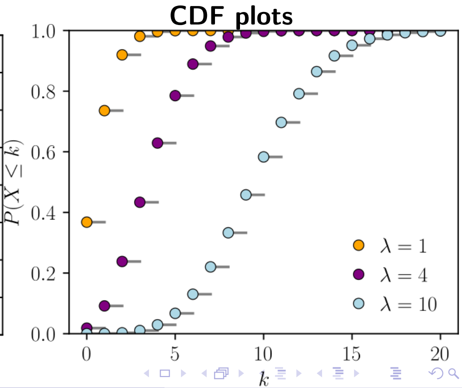




Poisson distribution - discrete

- Notation: $X \sim \mathrm{Pois}(\lambda)$ or $\mathrm{Poi}(\lambda)$
- Probability (mass) function: $p(x) = \frac{\lambda^x e^{-\lambda}}{x!}$
- Parameters: $\lambda \in \mathbb{R}_{\geq 0}$
- Support: $\mathbb{N}_0$
- $\mathbb{E}[X] = \lambda$; $\mathrm{Var}[X] = \lambda$

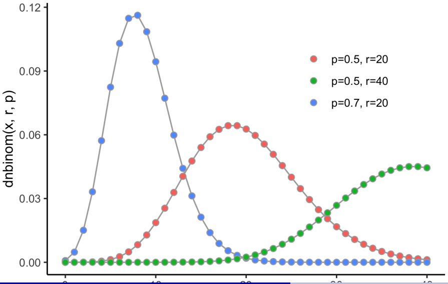




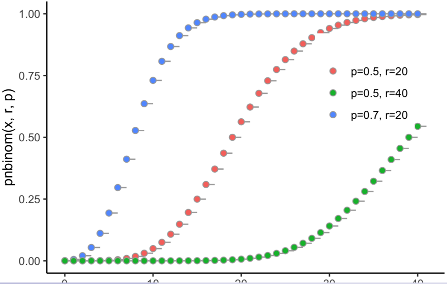
Negative Binomial distribution - discrete

- Notation: $X \sim \mathrm{NB}(r, p)$ or $\mathrm{negBin}(r, p)$
- Probability (mass) function: $p(x) = \binom{x+r-1}{x} \cdot (1-p)^x p^r$,
- Parameters: $r \in \mathbb{N}_0$, $p \in [0, 1]$ (number of successes until the experiment is stopped, success probability in each experiment)
- Support: $\mathbb{N}_0$
- $\mathbb{E}[X] = \dfrac{r(1-p)}{p}$; $\mathrm{Var}[X] = \dfrac{r(1-p)}{p^2}$

Probability (mass) function plots



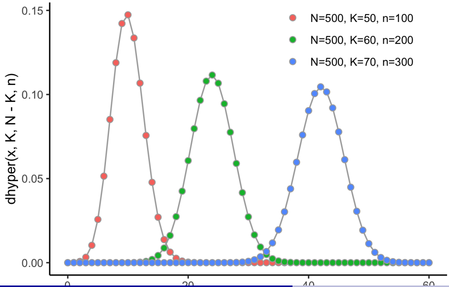
CDF plots



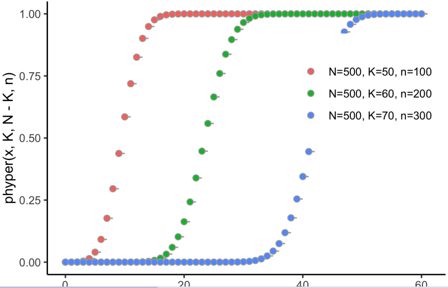
Hypergeometric distribution - discrete

- Notation: varies, sometimes $X \sim \mathrm{H}(N, K, n)$
- Probability (mass) function: $p(x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$
- Parameters: $N \in \mathbb{N}_0$, $K \in \{0, 1, 2, \ldots, N\}$, $n \in \{0, 1, 2, \ldots, N\}$ (population size, number of success states in the population, number of draws)
- Support: $\{\max(0,\, n + K - N),\, \ldots,\, \min(n,\, K)\}$
- $\mathbb{E}[X] = n\frac{K}{N}$; $\mathrm{Var}[X] = n\frac{K}{N}\frac{N-K}{N}\frac{N-n}{N-1}$

Probability (mass) function plots



CDF plots



Contents

1. Concepts and examples for RVs with univariate distribution
2. Pairs of random variables
   - Joint, marginal, and conditional distributions
   - Covariance and Correlation
3. Theoretical side-note: Fubini's theorem
4. Multivariate Distributions
   - Extending the concepts to vector notation
   - Relevant examples
5. Estimating distributions and characteristics from data
   - Empirical mean, (co)variance, and correlation
6. Side Note: Categorical variables
   - Simply Estimating the data's distribution "from scratch"
   - More complicated procedures to estimate distributions

Joint consideration of two random variables X and Y I

- Given two random variables X and Y, a natural quantity of interest is their joint distribution or **joint cumulative distribution function**, given by

$$F_{XY}(x, y) = \mathrm{P}(X \leq x, Y \leq y).$$

- For cases where one of the random variables X and Y is continuous and the other discrete, F_{XY} *can* be easy so define in some cases but rather complicated in others.

- In this lecture, we will focus only on *jointly* continuously/discretely distributed random variables:

Joint consideration of two random variables X and Y II

Definition (joint probability density/mass function)

- Two continuous random variables X and Y are jointly continuous if there exists a nonnegative function $f_{XY} : \mathbb{R}^2 \longrightarrow \mathbb{R}$, so that, for any set $A := [a_X, b_X] \times [a_Y, b_Y]$ with $a_X, a_Y, b_X, b_Y \in \mathbb{R}$, we have

$$P\big((X,Y) \in A\big) = \int_{a_Y}^{b_Y} \int_{a_X}^{b_X} f_{XY}(x,y)\,\mathrm{d}x\mathrm{d}y$$

  The function $f_{XY}(x,y)$ is called the **joint probability density function** of X and Y.

- The **joint probability (mass) function** of two jointly discrete random variables X and Y is defined as

$$p_{XY}(x,y) := \mathrm{P}(X = x, Y = y) \quad \Big(\hat{=}\,\mathrm{P}(X = x \text{ and } Y = y)\Big).$$

Marginal distributions for random variables X and Y I

Next, let $\boldsymbol{p}_X$ and $\boldsymbol{p}_Y$ denote the probability density **OR** mass functions of the random variables X and Y, respectively.

- Clearly, if
  - we start with $\boldsymbol{p}_X$ and $\boldsymbol{p}_Y$ as given and
  - know that X and Y are **independent** and **both** either discretely or continuously distributed

  it immediately follows that the joint probability density/mass function is given by

$$\boldsymbol{p}_{XY}(x, y) = \boldsymbol{p}_X(x) \cdot \boldsymbol{p}_Y(y) .$$

- Conversely, if the joint probability density/mass function of jointly distributed random variables X Y is given, we can deduce the probability density/mass functions **regardless of dependence** of X and Y by calculating the marginal distributions:

Marginal distributions for random variables X and Y II

Definition (Marginal probability density functions)

For two jointly continuous random variables X and Y with joint density f_{XY}, the densities defining the distributions of X and Y, respectively, are given by

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)dy\,, \quad \forall x \in \mathbb{R}, \text{ and}$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)dx\,, \quad \forall y \in \mathbb{R}.$$

Note: The following holds for both jointly discrete and continuous random variables: Given a joint CDF F_{XY}, the marginal CDFs are given by:

$$F_X(x) = F_{XY}(x,\infty) \quad \text{and} \quad F_Y(y) = F_{XY}(\infty,y)\,.$$

Marginal distributions for random variables X and Y III

Definition (Marginal probability mass functions)

For two jointly discrete random variables X and Y with joint probability function p_{XY}, the probability functions defining the distributions of X and Y, respectively, are given by

$$p_X(x) = \sum_{y_j \in \mathrm{supp}(p_Y)} p_{XY}(x, y_j)\,, \qquad \forall x \in \mathrm{supp}(p_X) \text{ and}$$

$$p_Y(y) = \sum_{x_i \in \mathrm{supp}(p_X)} p_{XY}(x_i, y)\,, \qquad \forall y \in \mathrm{supp}(p_Y)\,.$$

Note: The following holds for both jointly discrete and continuous random variables: Given a joint CDF F_{XY}, the marginal CDFs are given by:

$$F_X(x) = F_{XY}(x, \infty) \quad \text{and} \quad F_Y(y) = F_{XY}(\infty, y)\,.$$

Conditional distributions for random variables X and Y I

Next, let $\boldsymbol{p}_X$ and $\boldsymbol{p}_Y$ again denote the probability density **OR** mass functions of the random variables X and Y, respectively, and $\boldsymbol{p}_{XY}$ denote the joint probability density/mass function of X and Y.

Definition (Conditional probability density/mass function)

In the above setting, the **conditional probability density/mass function** of X given Y and vice versa is defined by

$$\boldsymbol{p}_{X|Y}(x, y) = \frac{\boldsymbol{p}_{XY}(x, y)}{\boldsymbol{p}_Y(y)} .$$

Conditional distributions for random variables X and Y II

Given this, note the following:

1. If X and Y are independent,

$$\boldsymbol{p}_{X|Y}(x,y) = \frac{\boldsymbol{p}_{XY}(x,y)}{\boldsymbol{p}_Y(y)} = \frac{\boldsymbol{p}_X(x)\boldsymbol{p}_Y(y)}{\boldsymbol{p}_Y(y)} = \boldsymbol{p}_X(x)\,.$$

2. For some set A, the conditional probability that $X \in A$ given that $Y = a$ for some fixed value a is given by
   - $\mathrm{P}(X \in A|Y = a) = \int_A f_{X|Y}(x,a)\mathrm{d}x$, if X and Y are continuously distributed.
   - $\mathrm{P}(X \in A|Y = a) = \sum\limits_{x_i \in A \cap \mathrm{supp}(p_X)} p_{X|Y}(x_i, a)$, if X and Y are discretely distributed.

Conditional distributions for random variables X and Y III

3. The conditional CDF of X given $Y = a$ for some fixed value a is given by

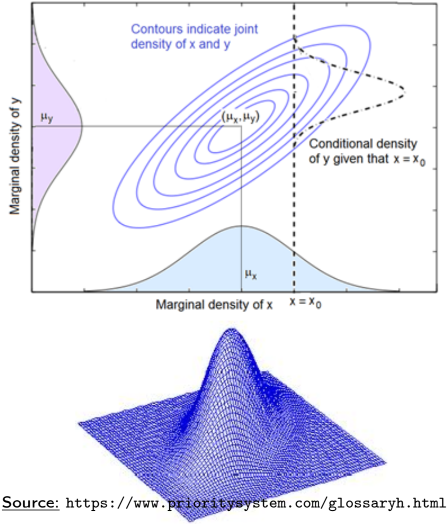
   - If X and Y are continuously distributed:

$$F_{X|Y}(x, a) = \mathrm{P}(X \leq x|Y = a) = \int_{-\infty}^{x} f_{X|Y}(u, a)\mathrm{d}u\,.$$

   - If X and Y are discretely distributed:

$$F_{X|Y}(x, a) = \mathrm{P}(X \leq x|Y = a) = \sum_{x_i \in [-\infty, x] \cap \mathrm{supp}(p_X)} p_{X|Y}(x_i, a)\,.$$

Joint, marginal, and conditional distributions for a bivariate normal probability distribution



Source: `https://www.prioritysystem.com/glossaryh.html`

Contents

1. Concepts and examples for RVs with univariate distribution
2. Pairs of random variables
   - Joint, marginal, and conditional distributions
   - Covariance and Correlation
3. Theoretical side-note: Fubini's theorem
4. Multivariate Distributions
   - Extending the concepts to vector notation
   - Relevant examples
5. Estimating distributions and characteristics from data
   - Empirical mean, (co)variance, and correlation
6. Side Note: Categorical variables
   - Simply Estimating the data's distribution "from scratch"
   - More complicated procedures to estimate distributions

Covariance I

- The covariance quantifies the statistical relation of two random variables by *considering their behavior with respect to their respective expectations*.

Definition (Covariance)

For two random variables X and Y with $\mathbb{E}[X], \mathbb{E}[Y] < \infty$, the covariance of X and Y, denoted by $\mathrm{Cov}(X, Y)$, is defined as

$$\mathrm{Cov}(X, Y) = \mathbb{E}\big[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\big] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]\,.$$

- Note that, by definition,

$$\mathrm{Cov}(X, X) = \mathbb{E}[XX] - \mathbb{E}[X]\mathbb{E}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathrm{Var}(X)\,.$$

Covariance II

- Furthermore, for **independent** random variables X and Y, it immediately follows that

$$\mathrm{Cov}(X, Y) = \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] = 0\,.$$

- Similarly, the following properties are easily proven:

1. $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$.
2. $\mathrm{Cov}(aX, Y) = a\mathrm{Cov}(X, Y)$ for some constant $a \in \mathbb{R}$.
3. $\mathrm{Cov}(X + c, Y) = \mathrm{Cov}(X, Y)$ for some constant $c \in \mathbb{R}$.
4. $\mathrm{Cov}(X + Y, Z) = \mathrm{Cov}(X, Z) + \mathrm{Cov}(Y, Z)$ for some third randon variable Z.

Variance of sums

- In addition to indicating the statistical relationship between random variables, the covariance is helpful for calculating the variance of sums of random variables.

- Specifically, for two random variables X and Y, and a random variable defined as $Z := X + Y$ the following holds:

$$\begin{aligned}\mathrm{Var}(Z) &= \mathrm{Cov}(Z, Z)\\ &= \mathrm{Cov}(X + Y, X + Y)\\ &= \mathrm{Cov}(X, X) + \mathrm{Cov}(X, Y) + \mathrm{Cov}(Y, X) + \mathrm{Cov}(Y, Y)\\ &= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y)\,.\end{aligned}$$

- More generally, for constants $a, b \in \mathbb{R}$, we have

$$\mathrm{Var}(aX + bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y) + 2ab\mathrm{Cov}(X, Y)\,.$$

Correlation I

- While the covariance is already very helpful and central to many methods, its magnitude is always dependent on the range of values the two variables in question take.

- There are many situations where the answer to the question "*How related are two random variables X and Y on a scale from -1 to 1?*" is of interest.

$\longrightarrow$ This question is answered by the correlation, which, for two random variables X and Y, is denoted by ρ_{XY} or $\mathrm{corr}(X, Y)$.

- This is achieved by calculating the covariance of the standardized version of each random variable.

Correlation II

- For a random variable X, the standardized version, with we denote by X_{stand}, is defined as $X_{\text{stand}} := \dfrac{X - \mathbb{E}[X]}{\sqrt{\text{Var}(X)}}$.

Definition (Correlation)

The correlation of two random variables X and Y, is defined as

$$\begin{aligned}\rho_{XY} &= \text{Cov}(X_{\text{stand}}, Y_{\text{stand}}) = \text{Cov}\left(\frac{X - \mathbb{E}[X]}{\sqrt{\text{Var}(X)}}, \frac{Y - \mathbb{E}[Y]}{\sqrt{\text{Var}(Y)}}\right) \\ &= \text{Cov}\left(\frac{X}{\sqrt{\text{Var}(X)}}, \frac{Y}{\sqrt{\text{Var}(Y)}}\right) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}}.\end{aligned}$$

Correlation III

- For two random variables X and Y, we say that
  - X and Y are **uncorrelated**, if $\rho_{XY} = 0$ and
  - X and Y are **positively/negatively correlated**, if $\rho_{XY} > 0$ and $\rho_{XY} < 0$, respectively.
- It clearly holds that $\rho_{XY} = 0 \Leftrightarrow \mathrm{Cov}(X, Y) = 0$ and, therefore, the following holds for **two uncorrelated** random variables X and Y

$$\mathrm{Var}(X, Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2 \cdot 0 = \mathrm{Var}(X) + \mathrm{Var}(Y)\,.$$

Correlation IV

Here are some neat properties of the correlation of two random variables X and Y:

1. $-1 \leq \text{corr}(X, Y) \leq 1$,
2. $\text{corr}(X, Y) = 1 \Rightarrow$ there exist constants $a \in \mathbb{R}_{>0}$ and $b \in \mathbb{R}$ s.t. $Y = aX + b$,
3. $\text{corr}(X, Y) = -1 \Rightarrow$ there exist constants $a \in \mathbb{R}_{<0}$ and $b \in \mathbb{R}$ s.t. $Y = aX + b$,
4. For some constants $a, b \in \mathbb{R}_{>0}$ the following holds: $\text{corr}(aX + b, cY + d) = \text{corr}(X, Y)$.

Contents

1. Concepts and examples for RVs with univariate distribution
2. Pairs of random variables
   - Joint, marginal, and conditional distributions
   - Covariance and Correlation
3. Theoretical side-note: Fubini's theorem
4. Multivariate Distributions
   - Extending the concepts to vector notation
   - Relevant examples
5. Estimating distributions and characteristics from data
   - Empirical mean, (co)variance, and correlation
6. Side Note: Categorical variables
   - Simply Estimating the data's distribution "from scratch"
   - More complicated procedures to estimate distributions

Theoretical side-note

- Next, we will look at **random vectors**, i.e. vectors with random variables as entries.
- Technically, the theoretical foundations (corresponding to what we looked at in the last lecture) of such objects would first require
  - the introduction of Product spaces and Product measures
  - as well as the consideration of measurable functions from Ω to $\mathbb{R}^k$, $k \in \mathbb{N}$.
- These concepts are not really relevant to applied statistics. However, there is one related theorem (versions of) which is (are) very relevant.

Fubini's Theorem

- Fubini's theorem, heuristically, tells us that we can calculate an integral over (a subset of) $\mathbb{R}^k$, $k \in \mathbb{N}$ as an **iterated integral in arbitrary order**, if the integral of the absolute value is finite.

- An example: For some function $h : \mathbb{R}^2 \longrightarrow \mathbb{R}$ and set $A := [a_1, b_1] \times [a_2, b_2]$; $a_1, a_2, b_1, b_2 \in \mathbb{R}$, if we know that

$$\int_A |h(x,y)| \mathrm{d}\lambda(x,y) < \infty$$

it immediately follows that

$$\int_A h(x,y) \mathrm{d}\lambda(x,y) = \int_{a_1}^{b_1} \int_{a_2}^{b_2} h(x,y) \mathrm{d}y \mathrm{d}x = \int_{a_2}^{b_2} \int_{a_1}^{b_1} h(x,y) \mathrm{d}x \mathrm{d}y \,.$$

- For a formal version, see Fubini, G. (1907), 'Sugli integrali multipli.', Rom. Acc. L. Rend. (5) 16(1), 608–614..

Why should we care about this?

- Clearly, we use iterated integrals when calculating probabilities for joint distributions.
- For the common established distributions, you can always assume that Fubini's theorem applies. However, when dealing with complicated and unconventional situations, it's validity might need to be verified!

Example

The function $f : \mathbb{R}^2 \longrightarrow \mathbb{R}$ defined by

$$f(x,y) := \begin{cases} 1, & \text{if } x \geq 0 \text{ and } x \leq y < x+1 \\ -1, & \text{if } x \geq 0 \text{ and } x+1 \leq y < x+2 \\ 0, & \text{otherwise,} \end{cases}$$

cannot be calculated as an iterated integral, since

$$0 = \iint f(x,y)\,dy\,dx \neq \iint f(x,y)\,dx\,dy = 1\,.$$

Contents

1. Concepts and examples for RVs with univariate distribution
2. Pairs of random variables
   - Joint, marginal, and conditional distributions
   - Covariance and Correlation
3. Theoretical side-note: Fubini's theorem
4. Multivariate Distributions
   - Extending the concepts to vector notation
   - Relevant examples
5. Estimating distributions and characteristics from data
   - Empirical mean, (co)variance, and correlation
6. Side Note: Categorical variables
   - Simply Estimating the data's distribution "from scratch"
   - More complicated procedures to estimate distributions

More than two random variables

- All the concepts we just considered for two random variables can be extended to three or more random variables.
- When dealing with multiple ($p \in \mathbb{N}_{>2}$) random variables $X_1, ..., X_p$, it is usually convenient to write them in *vector notation*.
- Specifically, we consider the **random vector**

$$\mathbf{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}$$

with realizations in $\mathbb{R}^p$.

Extending expectation and variance

- The **expected value vector** of a p-dimensional random vector X is defined as

$$\mathbb{E}[\mathbf{X}] = \left(\mathbb{E}[\mathbf{X_1}], \ldots, \mathbb{E}[\mathbf{X_p}]\right)^\top.$$

- The **covariance matrix**, often denoted by $\mathbb{V}(\mathbf{X})$, is defined as $\mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top]$, which is equal to

$$\mathbb{E}\begin{bmatrix} (X_1 - EX_1)^2 & (X_1 - EX_1)(X_2 - EX_2) & \ldots & (X_1 - EX_1)(X_p - EX_p) \\ (X_2 - EX_2)(X_1 - EX_1) & (X_2 - EX_2)^2 & \ldots & (X_2 - EX_2)(X_p - EX_p) \\ \vdots & \vdots & \vdots & \vdots \\ (X_p - EX_p)(X_1 - EX_1) & (X_p - EX_p)(X_2 - EX_2) & \ldots & (X_p - EX_p)^2 \end{bmatrix}$$

$$= \begin{bmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \ldots & \mathrm{Cov}(X_1, X_p) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & \ldots & \mathrm{Cov}(X_2, X_p) \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{Cov}(X_p, X_1) & \mathrm{Cov}(X_p X_2) & \ldots & \mathrm{Var}(X_p) \end{bmatrix}.$$

Which of these matrices is a covariance matrix?

$$\Sigma_1 = \begin{pmatrix} 0.2 & 0.5 \\ 0.2 & 0.3 \\ 0.5 & 0.3 \end{pmatrix} \quad \Sigma_2 = \begin{pmatrix} 0.5 & 0.7 & 0.9 \\ 0.3 & 0.9 & 0.3 \\ 0.9 & 0.7 & 0.5 \end{pmatrix}$$

$$\Sigma_3 = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix} \quad \Sigma_4 = \begin{pmatrix} 0.5 & 0.7 & -0.9 \\ 0.7 & 0.9 & 0.3 \\ -0.9 & 0.3 & -0.5 \end{pmatrix}$$

$\longrightarrow \Sigma_3$ and Σ_4.

Covariance and correlation in multivariate cases (continued)

- By definition, the covariance matrix has the following neat properties: It is
  1. square
  2. symmetric and
  3. positive semi-definite.
- In the context of a random vector $\mathbf{X} = \left(X_1, \ldots, X_p\right)^\top$, the correlation of two random variables that are elements of said vector, i.e. $\rho_{X_i X_j}$, $i, j \in \{1, ..., p\}$, is sometimes called **marginal correlation**.

Extending multivariate distributions from 2 to more dims I

- Equivalently to the case of two random variables, the **joint cumulative distribution function** (joint CDF) of $p \in \mathbb{N}$ random variables $X_1,\ X_2, \dots, X_p$ is given by

$$F_{X_1 \dots X_p}(x_1, x_2, \dots, x_p) = \mathrm{P}(X_1 \leq x_1, X_2 \leq x_2, \dots, X_p \leq x_p)\,.$$

- $p \in \mathbb{N}$ random variables $X_1,\ X_2, \dots, X_p$ are said to be **independent and identically distributed (i.i.d.)** if they are independent, and they have the same marginal distributions:

$$F_{X_1}(x) = F_{X_2}(x) = \dots = F_{X_p}(x) \quad \forall x \in \mathbb{R}\,.$$

Extending multivariate distributions from 2 to more dims II

- Again, equivalently to before, $p \in \mathbb{N}$ random variables $X_1,\ X_2, \ldots, X_p$ are jointly continuous if there exists a nonnegative function $f_{X_1 \ldots X_p} : \mathbb{R}^p \longrightarrow \mathbb{R}$, so that, for any set $A \in \mathcal{B}(\mathbb{R}^p)$ with, we have

$$\mathrm{P}\left((X_1, X_2, \ldots, X_p) \in A\right) = \int \ldots \int_A \ldots \int f_{X_1 \ldots X_p}(x_1, x_2, \ldots, x_p) \mathrm{d}x_1 \mathrm{d}x_2 \ldots \mathrm{d}x_p \,.$$

  Also, the function $f_{X_1 \ldots X_p}(x_1, x_2, \ldots, x_p)$ is called the **joint probability density function** of $X_1,\ X_2, \ldots, X_p$.

- The **joint probability (mass) function** of $p \in \mathbb{N}$ jointly discrete random variables $X_1,\ X_2, \ldots, X_p$ is defined as

$$p_{X_1 \ldots X_p}(x_1, x_2, \ldots, x_p) := \mathrm{P}\left(X_1 = x_1,\ X_2 = x_2, \ldots,\ X_p = x_p\right).$$

Extending multivariate distributions from 2 to more dims III

The conditional and marginal probability density/mass functions for $p \in \mathbb{N}$ random variables X_1, $X_2, \ldots, X_p$ are again defined analogously to the case of two random variables (see slides 25ff. and 29ff.):

- Given the joint CDF $F_{X_1 \ldots X_p}(x_1, x_2, \ldots, x_p)$, the **marginal CDF** F_{X_i} of the random variable X_i for any $i \in \{1, ..., p\}$ is given by the function

$$F_{X_i}(x_i) = F_{X_1 \ldots X_p}(\infty, \ldots, \infty, x_i, \infty, \ldots, \infty)\,.$$

- The **conditional probability density/mass function** of X_i given $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots X_p$ for any $i \in \{1, ..., p\}$ is defined by

$$\boldsymbol{p}_{X_i|X_1,\ldots,X_{i-1},X_{i+1},\ldots X_p}(x_1, x_2, \ldots, x_p) = \frac{\boldsymbol{p}_{X_1 \ldots X_p}(x_1, x_2, \ldots, x_p)}{\boldsymbol{p}_{X_1,\ldots,X_{i-1},X_{i+1},\ldots X_p}(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots x_p)}\,.$$

Extending multivariate distributions from 2 to more dims IV

- The idea of independence is also exactly the same as before: $p \in \mathbb{N}$ random variables $X_1,\, X_2, \ldots, X_p$ are independent, if for all $(x_1, x_2, ..., x_p) \in \mathbb{R}^p$
  - for continuous $X_1,\, X_2, \ldots, X_p$, the joint density is given by

$$f_{X_1 \ldots X_p}(x_1, x_2, \ldots, x_p) = \prod_{i=1}^{p} f_{X_i}(x_i)\,,$$

  - and for discrete $X_1,\, X_2, \ldots, X_p$, the joint probability (mass) function is given by

$$p_{X_1 \ldots X_p}(x_1, x_2, \ldots, x_p) = \prod_{i=1}^{p} p_{X_i}(x_i) \quad \left(= \prod_{i=1}^{p} \mathrm{P}(X_i = x_i) \right).$$

Contents

1. Concepts and examples for RVs with univariate distribution
2. Pairs of random variables
   - Joint, marginal, and conditional distributions
   - Covariance and Correlation
3. Theoretical side-note: Fubini's theorem
4. Multivariate Distributions
   - Extending the concepts to vector notation
   - Relevant examples
5. Estimating distributions and characteristics from data
   - Empirical mean, (co)variance, and correlation
6. Side Note: Categorical variables
   - Simply Estimating the data's distribution "from scratch"
   - More complicated procedures to estimate distributions

Multivariate Normal distribution

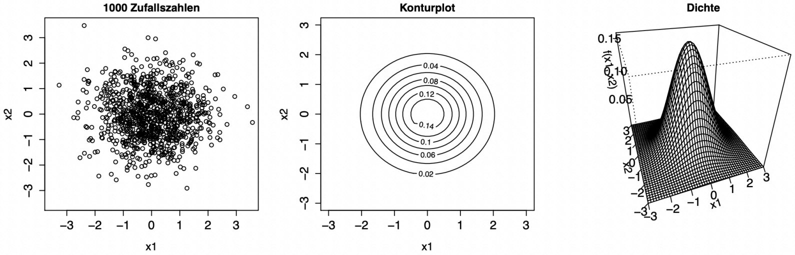
- We denote a p-dimensional random vector that follows the multivariate normal distribution by $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and the density function is given by

$$f : \mathbb{R}^p \longrightarrow \mathbb{R}, \quad x \mapsto \frac{1}{(2\pi)^{p/2} \mid \boldsymbol{\Sigma} \mid^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} .$$

- Parameters:
  - $\boldsymbol{\mu} \in \mathbb{R}^p$: expected value
  - $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$: covariance matrix
- Support: $\mu + \mathrm{span}(\Sigma) \subseteq \mathbb{R}^p$

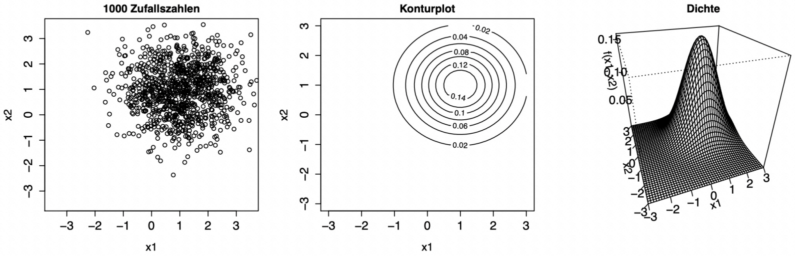
Examples

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$$



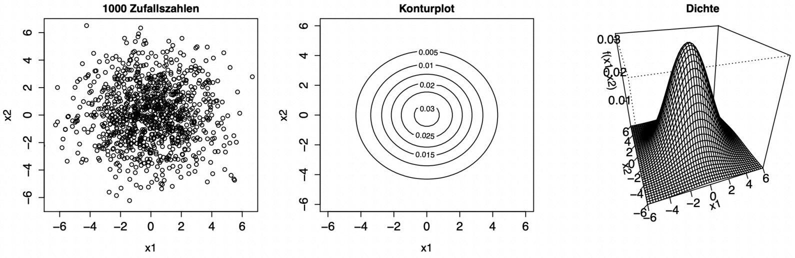
Examples

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$$



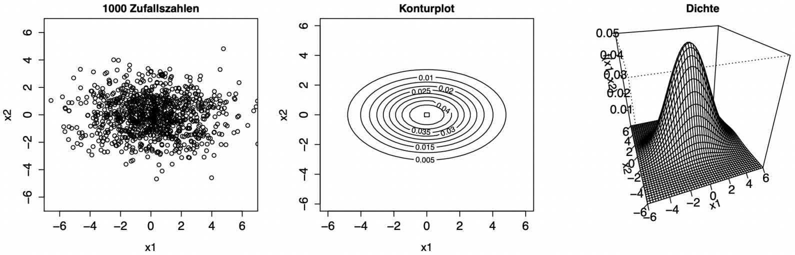
Examples

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 5 & 0 \\ 0 & 5 \end{pmatrix}\right)$$



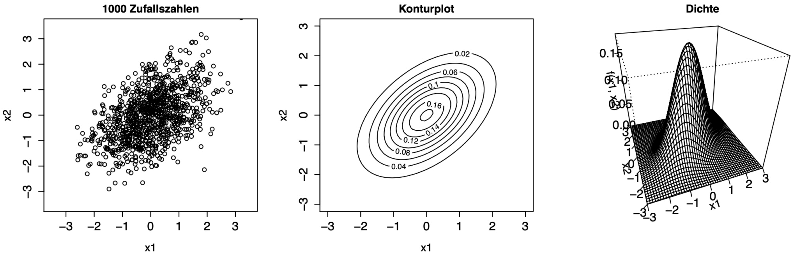
Examples

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 5 & 0 \\ 0 & 2 \end{pmatrix} \right)$$



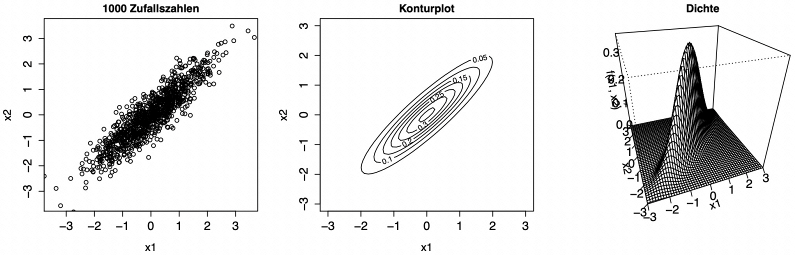
Examples

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \right)$$



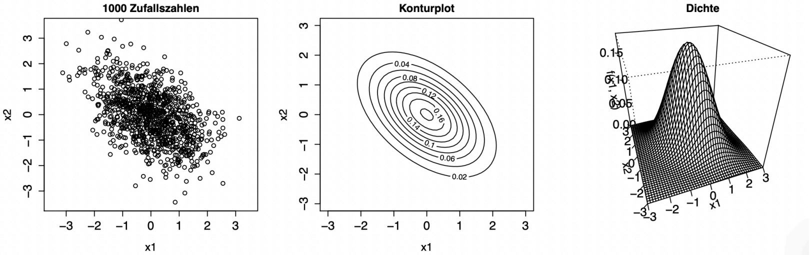
Examples

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix} \right)$$



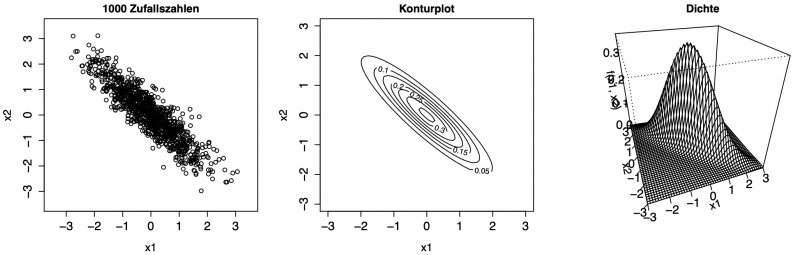
Examples

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}\right)$$



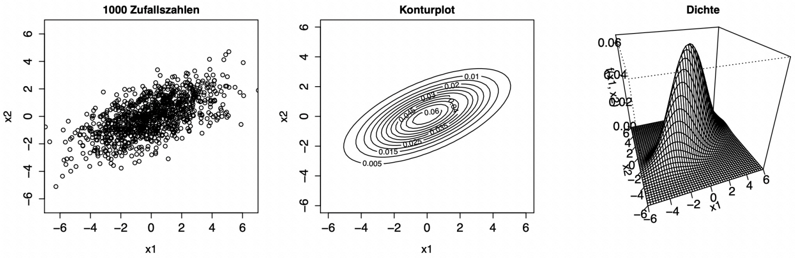
Examples

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & -0.9 \\ -0.9 & 1 \end{pmatrix} \right)$$



Examples

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 5 & 2 \\ 2 & 2 \end{pmatrix}\right)$$



Multivariate normal distribution: special cases

- For $p = 1$ we get the univariate normal distribution with parameters $\mu = \mathbb{E}(X)$ and $\Sigma = \mathrm{Var}(X)$.

- The standard multivariate normal distribution with parameters

$$\boldsymbol{\mu} = \mathbf{0} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \ \boldsymbol{\Sigma} = \mathbf{I} = \begin{pmatrix} 1 & \dots & 0 \\ & \ddots & \\ 0 & \dots & 1 \end{pmatrix},$$

Thusly distributed random vectors are denoted as $\mathbf{X} \sim N_p(\mathbf{0}, \mathbf{1})$.

Some specific properties

- If $\mathbf{X} \sim \mathbf{N_p}(\mu, \mathbf{\Sigma})$ holds, then $\boldsymbol{Y} = A\mathbf{X} + \mathbf{b}$ with $(q \times p)$–matrix A and $(q \times 1)$–vector $\mathbf{b}$ is in turn multivariate normally distributed with

$$\boldsymbol{Y} \sim N_q(A\mu + b, A\Sigma A^T)\,.$$

- If $\mathbf{X} \sim \mathbf{N_p}(\mu, \mathbf{\Sigma})$ holds, then $\boldsymbol{Y} = \Sigma^{-1/2}(X - \mu)$ is multivariate standard normally distributed, i.e. $\boldsymbol{Y} \sim N_p(\mathbf{0}, I)$.
Thus, the quadratic form $(\mathbf{X} - \mu)^\mathbf{T}\mathbf{\Sigma}^{-\mathbf{1}}(\mathbf{X} - \mu)$ is χ^2–distributed:

$$(\mathbf{X} - \mu)^\mathbf{T}\mathbf{\Sigma}^{-\mathbf{1}}(\mathbf{X} - \mu) \sim \chi^\mathbf{2}(\mathbf{p})\ .$$

Conditional normal distribution

- Consider $\mathbf{X} \sim \mathbf{N}(\mu, \mathbf{\Sigma})$ which is partitioned into $\mathbf{X^T} = (\mathbf{X_1^T}, \mathbf{X_2^T})$ as follows:

$$\begin{aligned} \boldsymbol{\mu}^T &= \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \\ \mathbf{\Sigma} &= \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}. \end{aligned}$$

The following then holds:

$$\boldsymbol{X}_1 | \boldsymbol{X}_2 \sim \mathcal{N}(\boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2}),$$

with

$$\begin{aligned} \boldsymbol{\mu}_{1|2} &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \\ \boldsymbol{\Sigma}_{1|2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \,. \end{aligned}$$

See `https://statproofbook.github.io/P/mvn-cond` for a proof.

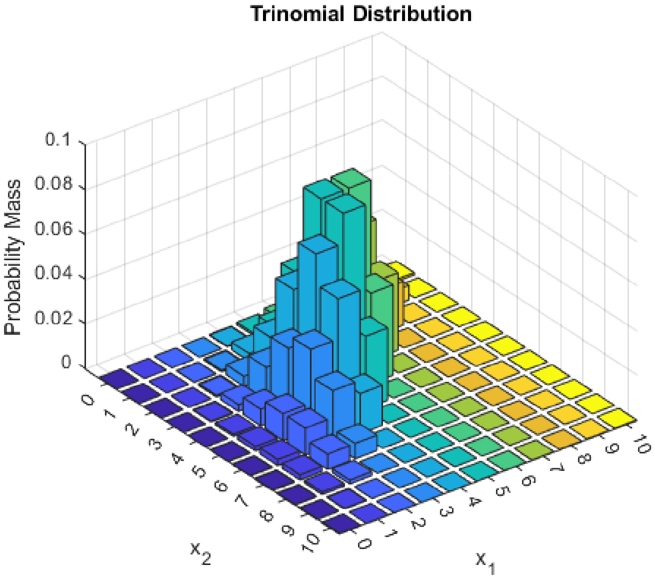
Multinomial distribution

- While the Binomial distribution models n independent trials of an experiment with two possible outcomes, the multinomial distribution is a generalization to n independent trials with k mutually exclusive outcomes.
- Parameters: $n \in \mathbb{N}$, $k \in \mathbb{N}$, $p_i \in [0,1]$ with $\sum_{i=1}^{k} p_i = 1$
- Support:

$$\left\{(x_1, ..., x_k)^\top \middle| x_i \in \{0, \ldots, n\}, \forall i \in \{1, \ldots, k\}, \quad \sum_{i=1}^{k} x_i = n\right\}$$

- Probability (mass) function: $f(x_1, \ldots, x_k) = \dfrac{n!}{x_1! \ldots x_k!}\, p_1^{x_1} \cdot \ldots \cdot p_k^{x_k}$

Multinomial distribution example



Dirichlet distribution I

- The Dirichlet distribution is the multivariate generalization of the Beta distribution.
- Parameter: $K \in \mathbb{N}_{\geq 2}$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)^\top \in \mathbb{R}^K$ with $\alpha_i > 0$
- Support: $\left\{(x_1, \ldots, x_K)^\top \middle| x_i \in [0,1] : \sum_{i=1}^K x_i = 1\right\}$
- Density:

$$f(x) = \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i)} \prod_{i=1}^K x^{\alpha_i - 1}$$

Dirichlet distribution II

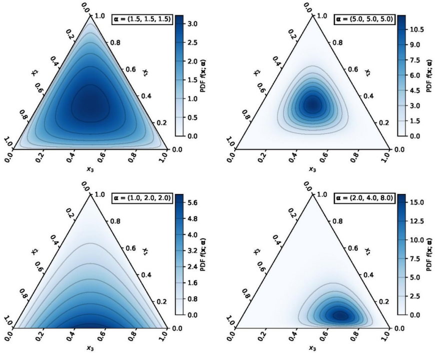
Properties:

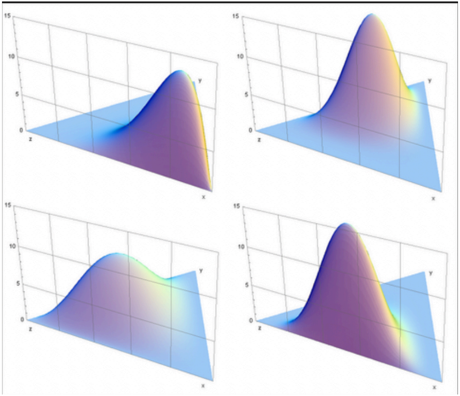
- $(X_1, \ldots, X_i + X_j, \ldots, X_k) \sim Dir(\alpha_1, \ldots, \alpha_i + \alpha_j, \ldots, \alpha_K)$

- For K independent Gamma distributed random variables $Y_1 \sim Gamma(\alpha_1, \theta), \ldots, Y_K \sim Gamma(\alpha_K, \theta)$ with $V = \sum_{i=1}^{K} Y_i \sim Gamma(\sum_{i-1}^{K} \alpha_i, \theta)$ the following holds $X = (X_1, \ldots, X_K) = \left(\frac{Y_1}{V}, \ldots, \frac{Y_K}{V}\right) \sim Dir(\alpha_1, \ldots, \alpha_K)$

- Dirichlet distributions are commonly used as prior distributions. In fact, the Dirichlet distribution is the conjugate prior of the categorical distribution and multinomial distribution.

Dirichlet distribution examples





Multivariate hypergeometric distribution

This distribution corresponds to the generalization of “drawing without replacement”. n elements are drawn from a total of N, grouped into K classes containing $N_1, \ldots, N_K$ elements, respectively.

The probability mass function is given by

$$P(X_1 = n_1, \ldots, X_K = n_K) = \frac{\prod_{k=1}^{K} \binom{N_k}{n_k}}{\binom{N}{n}} \quad \text{with} \quad \sum n_k = n \, .$$

Wishart-Verteilung

Consider the random variables $\mathbf{X}_1, \ldots, \mathbf{X}_m \overset{i.i.d.}{\sim} N_p(\mathbf{0}, \mathbf{\Sigma})$. The following matrix is then Wishart distributed with parameters $\mathbf{\Sigma}$ und $m \in \mathbb{N}$ (i.e. $\mathbf{M} \sim W_p(\mathbf{\Sigma}, m)$)

$$\mathbf{M} = \sum_{i=1}^{m} \mathbf{X}_i \mathbf{X}_i^\top = \mathbf{X}^\top \mathbf{X} \quad \in \mathbb{R}^{p \times p}.$$

- If $p = 1$, then $\boldsymbol{M} = \sum_{i=1}^{m} X_i^2 \sim \chi^2(m)$, with $X_i \sim N(0, \sigma^2)$

$\Rightarrow$ The Wishart distribution is the multivariate generalization of the χ^2- distribution.

Wilks' Λ distribution I

Consider two independent random variables $\mathbf{A} \sim W_p(\mathbf{I}, m)$ and $\mathbf{B} \sim W_p(\mathbf{I}, n)$ then

$$\Lambda = \frac{\det(\mathbf{A})}{\det(\mathbf{A} + \mathbf{B})}$$

is Wilks' Λ-distributed with parameters p, m, and n.

- $\Lambda \sim \Lambda(p, m, n)$
- If $p = 1$, then $A \sim \chi^2(m)$ and $B \sim \chi^2(n)$ and thus we get: $\Lambda \sim B(m/2, n/2)$
- Wilks' Λ-distribution is used for testing in the context of one-way analysis of variance.

Wilks' Λ distribution II

Properties:

1. For the one-dimensional special case $A \sim \chi^2(1),\ B \sim \chi^2(1)$ we get the Beta–distribution $\Lambda(1,1,1) \hat{=} B(0.5, 0.5)$.

2. The distributions $\Lambda(p,m,n)$ and $\Lambda(n, m+n-p, p)$ are identical.

Hotellings T^2 distribution

- Hotellings T^2 distribution is used for multivariate hypothesis testing problems (specifically the multivariate generalization of the t-test).

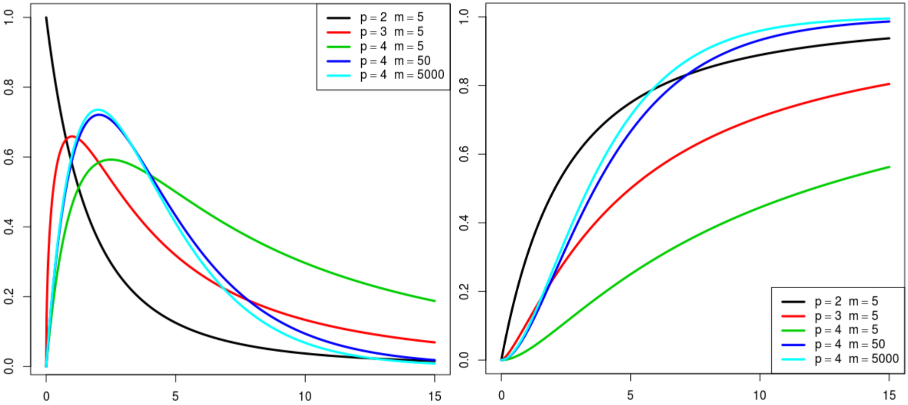
- Consider the independent random vector $\mathbf{d} \sim N_p(\mathbf{0}, \mathbf{I})$ and random matrix $\mathbf{M} \sim W_p(\mathbf{I}, m)$. The quadratic form

$$u = m\mathbf{d}^\top \mathbf{M}^{-1}\mathbf{d} \in \mathbb{R}$$

  is then Hotelings T^2 distributed with parameter p and m (we write $u \sim T^2(p, m)$).

- The support is $\begin{cases} \mathbb{R}_{>0}\,, & \text{if } p = 1, \\ \mathbb{R}_{\geq 0}\,, & \text{otherwise.} \end{cases}$

Hotellings T^2 distribution pdf and cdf plots



Contents

1. Concepts and examples for RVs with univariate distribution
2. Pairs of random variables
   - Joint, marginal, and conditional distributions
   - Covariance and Correlation
3. Theoretical side-note: Fubini's theorem
4. Multivariate Distributions
   - Extending the concepts to vector notation
   - Relevant examples
5. Estimating distributions and characteristics from data
   - Empirical mean, (co)variance, and correlation
6. Side Note: Categorical variables
   - Simply Estimating the data's distribution "from scratch"
   - More complicated procedures to estimate distributions

The Data I

- Let's say we are given a data set with n observations of m variables:

	X_1	X_2	X_3	$\ldots$	X_m
1	x_{11}	x_{12}	x_{13}	$\ldots$	x_{1m}
2	x_{21}	x_{22}	x_{23}	$\ldots$	x_{2m}
3	x_{31}	x_{32}	x_{33}	$\ldots$	x_{3m}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n	x_{n1}	x_{n2}	x_{n3}	$\ldots$	x_{nm}

The Data II

- **Question:** How do we write this data down mathematically?

  Answer: There is no one right answer! *But*, most of the time, we will consider the rows to be random vectors $\mathbf{X_1}, \dots, \mathbf{X_n}$ drawn **i.i.d.**, meaning *independent and identically distributed*.

Definition (i.i.d.)

A collection of $n \in \mathbb{N}_{>0}$ random variables or vectors with realization in $\mathbb{R}^p$ is said to be independent and identically distributed, or i.i.d., iff the following two conditions hold:

$$F_{X_1}(x) = F_{X_k}(x) \quad \forall k \in \{1, \dots, n\} \text{ and } \forall x \in \mathbb{R}^p$$

$$F_{X_1,\dots,X_n}(x_1, \dots, x_n) = F_{X_1}(x_1) \cdot \ldots \cdot F_{X_n}(x_n) \quad \forall x_1, \dots, x_n \in \mathbb{R}^p\,.$$

Empirical mean, variance, and covariance I

- Sometimes, we might just be interested in some characteristics of the distribution defined by the CDF $F_{X_k}(x) \quad \forall k \in \{1, \ldots, n\}$, such as the espected value.
- Other times, we might have made a *distributional assumption*, such as "normal distribution" and just need to estimate the parameters.

Given the sequence of data points $\{\boldsymbol{x}_i\}_{i=1,\ldots,n}$, with $\boldsymbol{x}_i$ representing an observation of a univariate random variable ($\boldsymbol{x}_i \in \mathbb{R}$) or a random vector $p \in \mathbb{N}_{>0}$ ($\boldsymbol{x}_i \in \mathbb{R}^p$) variables, some common empirical estimators of distribution characteristics include the following:

Empirical mean, variance, and covariance II

- The **arithmetic mean** is an intuitive choice for empirically estimating the expected value:

$$\bar{\boldsymbol{x}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \,.$$

- The **sample variance** is used for empirically estimating the variance

$$S^2 = \frac{1}{n-1} \sum (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^2 \,.$$

- Finally, for <u>two variables</u> with realizations $\{x_i^{(1)}\}_{i=1,\ldots,n}$, $\{x_i^{(2)}\}_{i=1,\ldots,n}$ the **sample covariance** is given by

$$cov_{x^{(1)}x^{(2)}} = \frac{1}{n-1} \sum (x_i^{(1)} - \bar{x}^{(1)}) \sum (x_i^{(2)} - \bar{x}^{(2)}) \,.$$

Empirical correlation I

- In statistics, the term "*correlation*" is often used to refer to various measures of the relationship between the two variables.

→ There are different types correlation coefficients, e.g. rank coefficients etc.

- The formal correlation of two random variables X and Y, defined as $\frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$, measures the linear association between variables

  (this is also why $\rho_{XY} = 0$ DOES NOT imply independence, only the other way around).

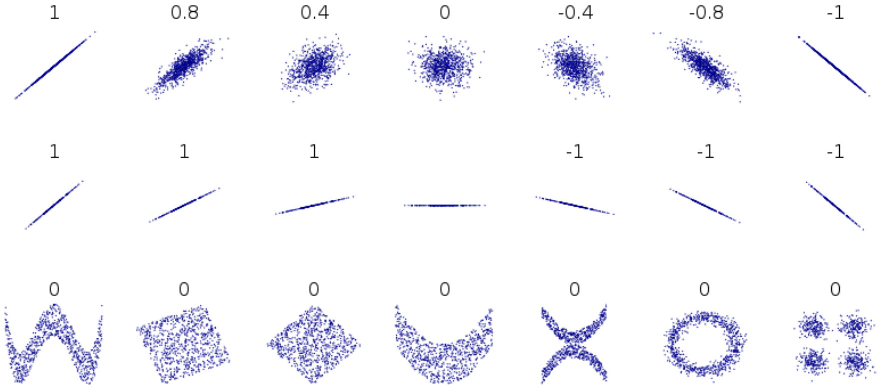
Empirical correlation II

- For <u>two variables</u> with realizations $\{x_i\}_{i=1,\dots,n}$, $\{y_i\}_{i=1,\dots,n}$, this correlation ρ_{XY} can be empirically estimated via the **Pearson correlation coefficient**

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \,.$$

- The following visualizes the Pearson correlation coefficient for different data points. (By Denis Boigelot, original uploader was Imagecreator - Own work, original uploader was Imagecreator, CC0, `https://commons.wikimedia.org/w/index.php?curid=15165296`)

Empirical correlation III



Estimating the data's distribution "from scratch"

- Let's say that we
  - **are** assuming our observations are realizations of i.i.d. random variables/vectors (RVs)
  - but we **do not** have a certain distribution $\mathcal{D}$ in mind to make the assumption

$$\mathbf{X_1}, \ldots, \mathbf{X_n} \overset{\mathbf{i.i.d}}{\sim} \mathcal{D}\Big(\text{some parameters}\Big) .$$

- How can we still make inferences about the distribution from which the data was drawn?

Side-Note: Categorical variables I

- When dealing with data, we often distinguish between *metric/numeric* and *categorical* variables.

- Usually, metric variables take values in $\mathbb{R}$, while a categorical variable C only takes values, of any kind, including text, that are elements of a *finite set* defining the possible values C may take.

- A classical example would be a variable with two possible values, such as "*individual smokes*" and "*individual doesn't smoke*".
  $\longrightarrow$ Of course, if we want this variable to take values in $\mathbb{R}$, we can simply recode it as

$$\tilde{c}_i = \begin{cases} 1, & \text{if } c_i = \{\text{individual smokes}\} \\ 0, & \text{if } c_i = \{\text{individual doesn't smoke}\}\,. \end{cases} \qquad (\star)$$

Side-Note: Categorical variables II

Q1: *What about if C can take more than two, lets say $k \in \mathbb{N}_{>2}$, values?*

A1: We can repeat the procedure of eq.($\star$) $k-1$ times.

Q2: *Why not k times?*

A2: If all $k-1$ new variables representing a possible value of the ith observation of C are equal to 0, this means that c_i is equal to the kth value for which we didn't create a separate column.

$\Longrightarrow$ This is called **dummy coding**.

Side-Note: Dummy coding in R

- In R, we can use the `fastDummies` package to dummy code quickly :)
- Calling

```
library(kableExtra)
fastDummies_example <- data.frame(ID = 1:6,
                                  sex  = c("male", "male", "intersex","intersex","female","female"),
                                  choice = c("YES", "NO", "YES", "NO","YES", "NO"),
                                  DOB   = as.Date(c("1999-01-01", "2003-12-30","2001-05-20",
                                                    "2000-08-17", "1997-12-10","2000-06-27")),
                                  stringsAsFactors = FALSE)
recoded <- fastDummies::dummy_cols(fastDummies_example, select_columns = c("sex","choice"))
kbl(recoded) %>%
  kable_classic_2(full_width = F)
```

gives us

ID	sex	choice	DOB	sex_female	sex_intersex	sex_male	choice_NO	choice_YES
1	male	YES	1999-01-01	0	0	1	0	1
2	male	NO	2003-12-30	0	0	1	1	0
3	intersex	YES	2001-05-20	0	1	0	0	1
4	intersex	NO	2000-08-17	0	1	0	1	0
5	female	YES	1997-12-10	1	0	0	0	1
6	female	NO	2000-06-27	1	0	0	1	0

Relative Frequency

- This can be used for any kind of data, including a mix of metric and categorical variables!

- Given the sequence of data points $\{\boldsymbol{x}_i\}_{i=1,\dots,n}$, with $\boldsymbol{x}_i$ representing an observation of one ($\boldsymbol{x}_i \in \mathbb{R}$) or $p \in \mathbb{N}_{>0}$ ($\boldsymbol{x}_i \in \mathbb{R}^p$) variables, the following is clearly a valid estimation of a **discrete** probability function for a random variable with realizations in $\{\boldsymbol{x}_i\}_{i=1,\dots,n}$:

$$\widehat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{x=\boldsymbol{x}_i\}} .$$

$\rightarrow$ each data point gets assigned the probability

$$\frac{\#\text{data point appears in the data set}}{n} .$$

The CDF of relative frequency I

- You may have already dealt with what is often referred to as the *empirical distribution* defined as $\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{\boldsymbol{x}_i \le x\}}$.

- We can also write it out as follows:

$$\begin{aligned}
\widehat{F}_n(x) = P(X \le x) &= \sum_{\omega \in \operatorname{supp}(\hat{f}_n) \cap [-\infty, x]} P(X = \omega) \\
&= \sum_{\omega \in \operatorname{supp}(\hat{f}_n) \cap [-\infty, x]} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{\omega = \boldsymbol{x}_i\}} \\
&= \frac{1}{n} \sum_{i=1}^{n} \sum_{\omega \in \operatorname{supp}(\hat{f}_n) \cap [-\infty, x]} \mathbb{1}_{\{\omega = \boldsymbol{x}_i\}}.
\end{aligned}$$

The CDF of relative frequency II

Q: Can we just use that formula for our data made up of observations $\{\boldsymbol{x}_i\}_{i=1,\dots,n}$?

The CDF of relative frequency II

Q: Can we just use that formula for our data made up of observations $\{\boldsymbol{x}_i\}_{i=1,\ldots,n}$?

A: Immediately, iff $\boldsymbol{x}_i \in \mathbb{R}$, $\forall i \in \{1, \ldots, n\}$! However,

- **For $\boldsymbol{x}_i \in \mathbb{R}^p$, $p \geq 2$:** the usual preorder (binary relation that is reflexive and transitive) $\leq$ that we use on $\mathbb{R}$ does not extend to $\mathbb{R}^n$, $n \in \mathbb{N}_{>1}$, we would first need to establish a fitting preorder, **if we want to quantify the joint distribution of two or more variables together**.
- If our categories aren't ordered, we have no chance at all.

Still, we can always use the relative frequency! We just won't have a CDF to go with it.

CHALLENGE TIME!

Next, you will all have 20 minutes to (by yourself or in a group) think of ways to define an empirical CDF for data that has more than one column!

You can come up with creative solutions yourself or browse the internet for inspiration - if you find (an) interesting paper(s) that's also great!

The person(s) presenting the result I'm most impressed with will get a small prize :))

Kernel density estimation (KDE) I

- An option designed *purely for density estimation* - i.e. applicable only when all variables are metric - is KDE.

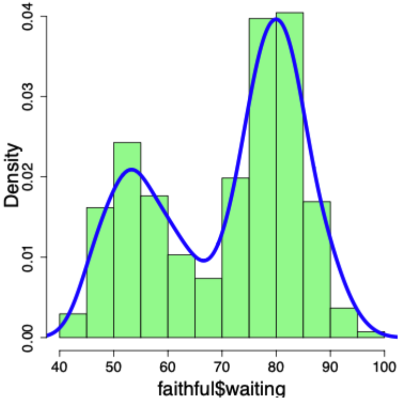
- Given the sequence of data points $\{\boldsymbol{x}_i\}_{i=1,\dots,n}$, with $\boldsymbol{x}_i$ representing an observation of one ($\boldsymbol{x}_i \in \mathbb{R}$) or $p \in \mathbb{N}_{>0}$ ($\boldsymbol{x}_i \in \mathbb{R}^p$) metric variables, the following can be used to estimate the **continuous** density for a random variable with realizations in $\{\boldsymbol{x}_i\}_{i=1,\dots,n}$:

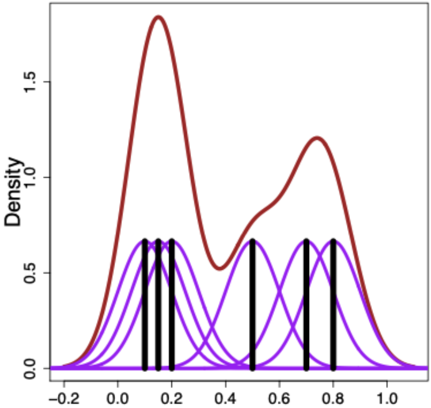
$$\widehat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - \boldsymbol{x}_i) = \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{x - \boldsymbol{x}_i}{h}\Big),$$

where K is the kernel — a non-negative function — and $h > 0$ is a smoothing parameter called the bandwidth.

Kernel density estimation (KDE) II

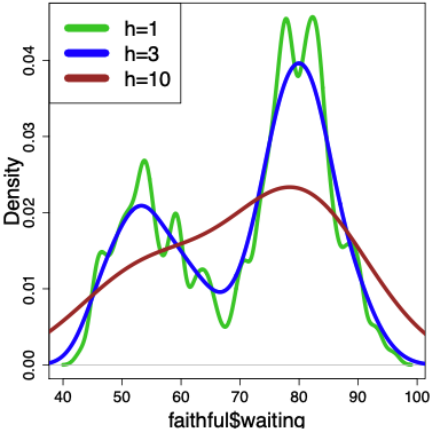
This method is usually how histogram plots are smoothed. The following and all later KDE-graphics are taken from this nice lecture about density estimation.





Kernel density estimation (KDE) III

The smoothing parameter h is key - it should be chosen neither too small nor too large.



Kernel density estimation (KDE) IV

- *How do we choose a kernel?* It should satisfy the following
  1. $K(x)$ is symmetric.
  2. $\int K(x)dx = 1$.
  3. $\lim_{x \to -\infty} K(x) = \lim_{x \to +\infty} K(x) = 0$.

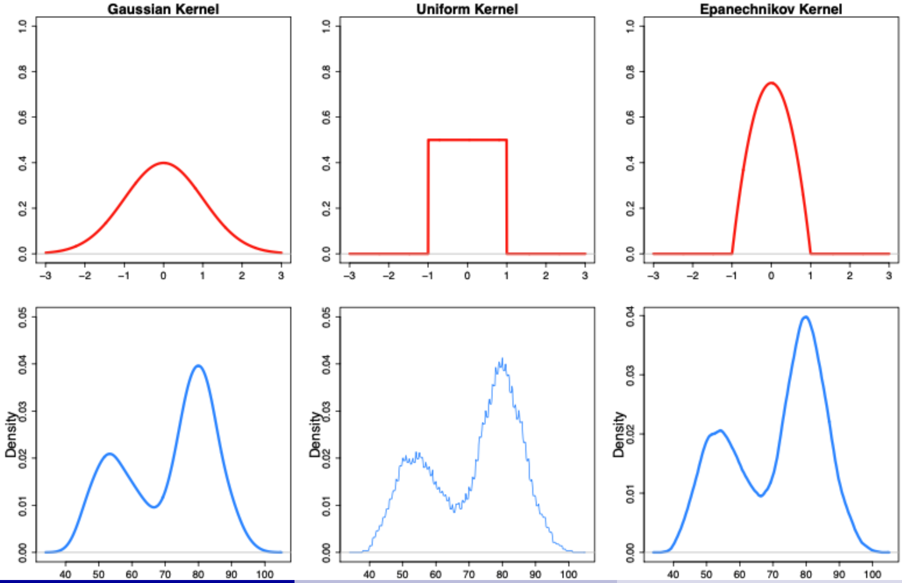
- Some commonly chosen kernels are:

$$\textbf{Gaussian } K(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}},$$

$$\textbf{Uniform } K(x) = \frac{1}{2} \mathbb{1}_{\{-1 \leq x \leq 1\}},$$

$$\textbf{Epanechnikov } K(x) = \frac{3}{4} \cdot \max\left\{1 - x^2, 0\right\}.$$

Kernel density estimation (KDE) V



Data generating processes (DGPs) I

- Of course, smoothing histograms is neat - but can we use KDE for anything else?

- Absolutely! $\hat{f}_h$ defines a data generating process (DGP) - i.e. a way for us to "generate" new data points from the same distribution as the data we have already observed.

Be careful

While being able to generate new data points is great, it will only be as "good" as the data we have already observed.

Data generating processes (DGPs) II

Q: What would we use this for?

A: So much! Just some examples include:

- More complex parameter estimation.
- Calculating probability via Monte Carlo Integration.
- Model validation.
- Posterior predictive checks.

Example: Bootstrap estimation I

Definition (The bootstrap principle, see also this lecture)

1. $x_1, x_2, \ldots, x_n$ is a data sample drawn from a distribution F.
2. u is a statistic computed from the sample.
3. F^* is the empirical distribution of the data (the resampling distribution).
4. $x_1^*, x_2^*, \ldots, x_n^*$ is a resample of the data of the same size as the original sample
5. u^* is the statistic computed from the resample.

Then the bootstrap principle says that

1. F^* is approximately equal to F.
2. The statistic u is well approximated by u^*.
3. The variation of u is well approximated by the variation of u^*.

Example: Bootstrap estimation II

- Here, we are sampling with replacement, so you could say that the *relative frequency* $\hat{p}$ is the probability function that defines our DGP.

- We can of course use bootstrap to estimate our *mean/expected value* and *variance* the same way we would have done on the original data.

- However, we could also use the bootstrap principle as follows:

  1. Calculate the set $\{\delta^*\}_{b=1}^{B}$ with $\delta^* := \bar{x}^* - \bar{x}$.
  2. Calculate the 0.025 and 0.975 quantiles of $\{\delta^*\}_{b=1}^{B}$, denoted by $\delta^*_{0.025}$ and $\delta^*_{0.975}$.
  3. Get a 95% CI for the mean via

$$[\bar{x} - \delta^*_{0.025}, \bar{x} + \delta^*_{0.975}] \, .$$

Example: Monte Carlo Integration I

- **Monte Carlo Integration** is a technique to approximate the integral over a multidimensional function $g : \mathbb{R}^l \longrightarrow \mathbb{R}^m$, $l, m \in \mathbb{N}_{>0}$.

- Specifically, consider a set $M \subset \mathbb{R}^m$ and
  - a sample of n points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ from the *Uniform distribution* on M and
  - V to be the volume of M, i.e. $V := \int_{\mathbb{R}^m} \mathbb{1}_{\{x \in M\}} d(x)$.

- Then, we can approximate

$$\int_M g(x)dx \approx V \cdot \frac{1}{n} \sum_{i=1}^{n} g(\boldsymbol{x}_i) .$$

Example: Monte Carlo Integration II

- A common special case is when $m = 1$. Then, for $a, b \in \mathbb{R}$ we can approximate
$$\int_a^b g(x)dx \approx \frac{b-a}{n}\sum_{i=1}^{n} g(\boldsymbol{x}_i)\,,$$
where $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are sampled from the $\mathcal{U}(a, b)$ distribution.

- Of course, that means that *integrate w.r.t. any distribution we can generate draws from and thereby calculate probability on sets*.

⇒ This is where DGPs become super helpful.

Outlook: Current DGP Research

- Another context in which DGPs are of interest is *privacy concerns* - to preserve privacy, it would be super neat if we could share data with the same characteristics as what we observed without sharing the actual data.

- Two more fancy ways to estimate DGPs:

  1. Fitting bayesian models and using the posterior distribution: `https://www.jmlr.org/papers/volume18/15-257/15-257.pdf`
  2. Large Language Learners: `https://arxiv.org/pdf/2210.06280`