Statistical Data Analysis

Dr. Jana de Wiljes

2. November 2021

Universität Potsdam

Continuous Random Variables

Normal Distribution

A normal or Gaussian distributed random variable $X : \Omega \to \mathbb{R}$ with parameters $\mu \in \mathbb{R}$ and $\sigma > 0$ has the following density

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 \right\}$$

and expected value and variance

$$\mathbb{E}[X] = \mu$$
$$Var(X) = \sigma^2$$

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

Normal Distribution

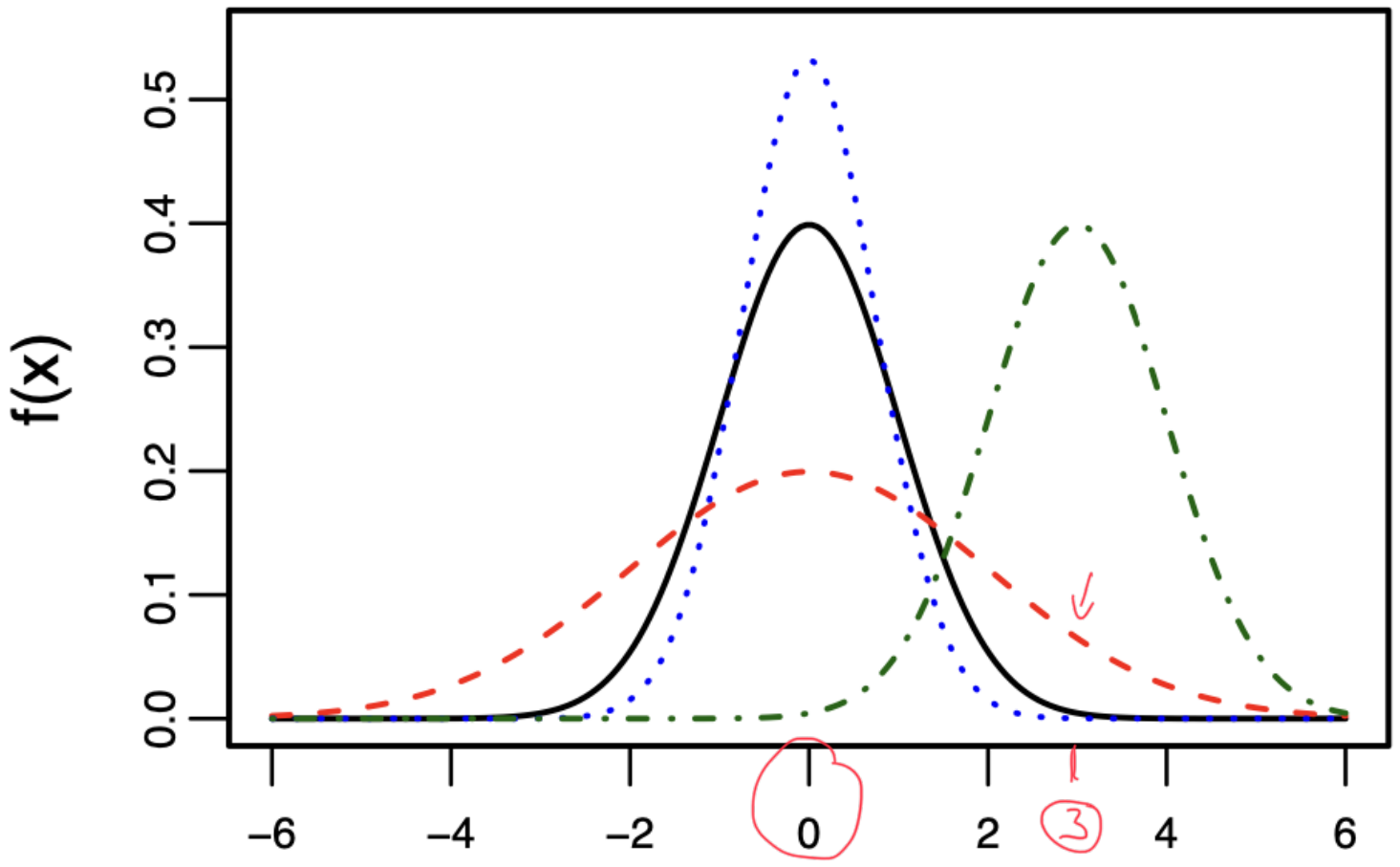


Abbildung 1: $\mu = 0$, $\sigma = 1$ (black), $\mu = 0$, $\sigma = 2$ (red), $\mu = 0$, $\sigma = 0.75$ (blue) and $\mu = 3$, $\sigma = 1$ (green)

Normal Distribution

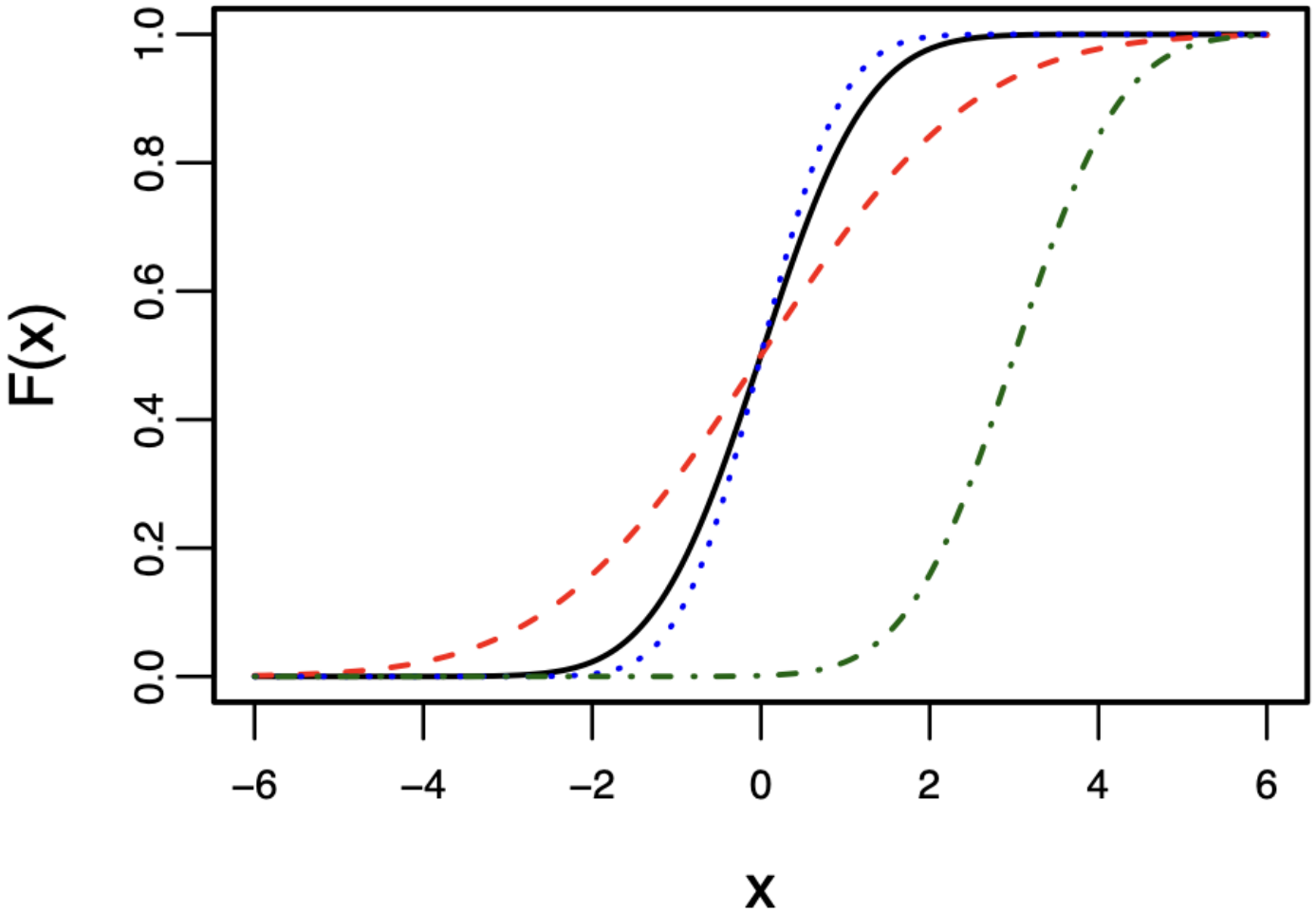


Abbildung 2: $\mu = 0$, $\sigma = 1$ (black), $\mu = 0$, $\sigma = 2$ (red), $\mu = 0$, $\sigma = 0.75$ (blue) and $\mu = 3$, $\sigma = 1$ (green)

Quantile

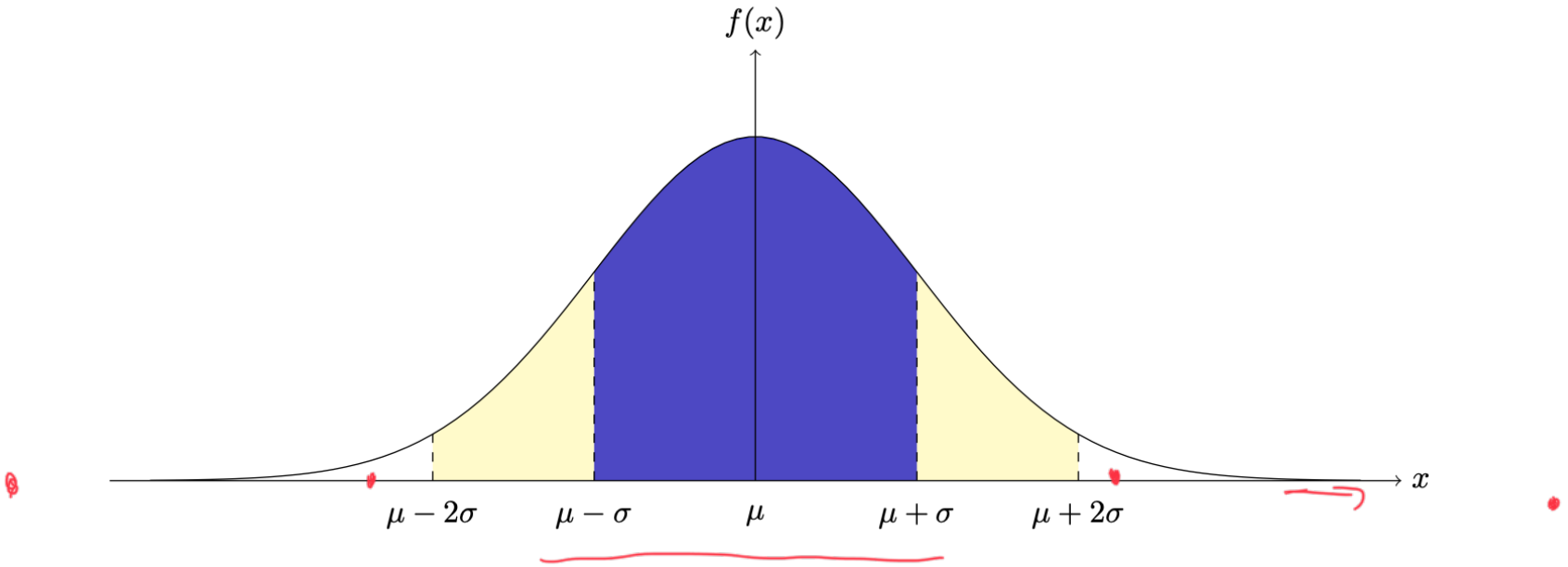


Abbildung 3: 68% of area under the curve (colored in blue) are in the $[\mu - \sigma, \mu + \sigma]$ interval and 95% of the area under the curve are in the interval $[\mu - 2\sigma, \mu + 2\sigma]$.

99.7% in $[\mu - 3\sigma, \mu + 3\sigma]$

Standard normal distribution

A variable $X : \Omega \to \mathbb{R}$ follows a standard normal distribution, i.e., $X \sim \mathcal{N}(0, 1)$ if the associated density has the following form

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{ -\left(\frac{x^2}{2}\right)\right\}$$

with the associate cumulative distribution

$$\Phi(x) = \int_{-\infty}^{x} \phi(u)du \tag{1}$$

and quantile

$$z_\alpha = \Phi^{-1}(\alpha), \quad \alpha \in (0, 1) \tag{2}$$

Relationship between standard normal distribution and Normal distribution

$$F(x) = \Phi\left(\frac{x - \mu}{\sigma}\right) \tag{3}$$

Exponential Distribution

A random variable $X : \Omega \to \mathbb{R}$ follows the exponential distribution with parameters $\lambda > 0$ has the following density and cdf

$$f(x) = \begin{cases} 0 & x < 0 \\ \lambda \exp(-\lambda x) & x \geq 0 \end{cases}$$

$$F(x) = \begin{cases} 0 & x < 0 \\ 1 - \exp(-\lambda x) & x \geq 0 \end{cases}$$

and expected value and variance

$$\mathbb{E}[X] = \frac{1}{\lambda}$$

$$Var(X) = \frac{1}{\lambda^2}$$

Notation: $X \sim \text{Exp}(\lambda)$ (often used for waiting times and lifetimes)

Exponential Distribution

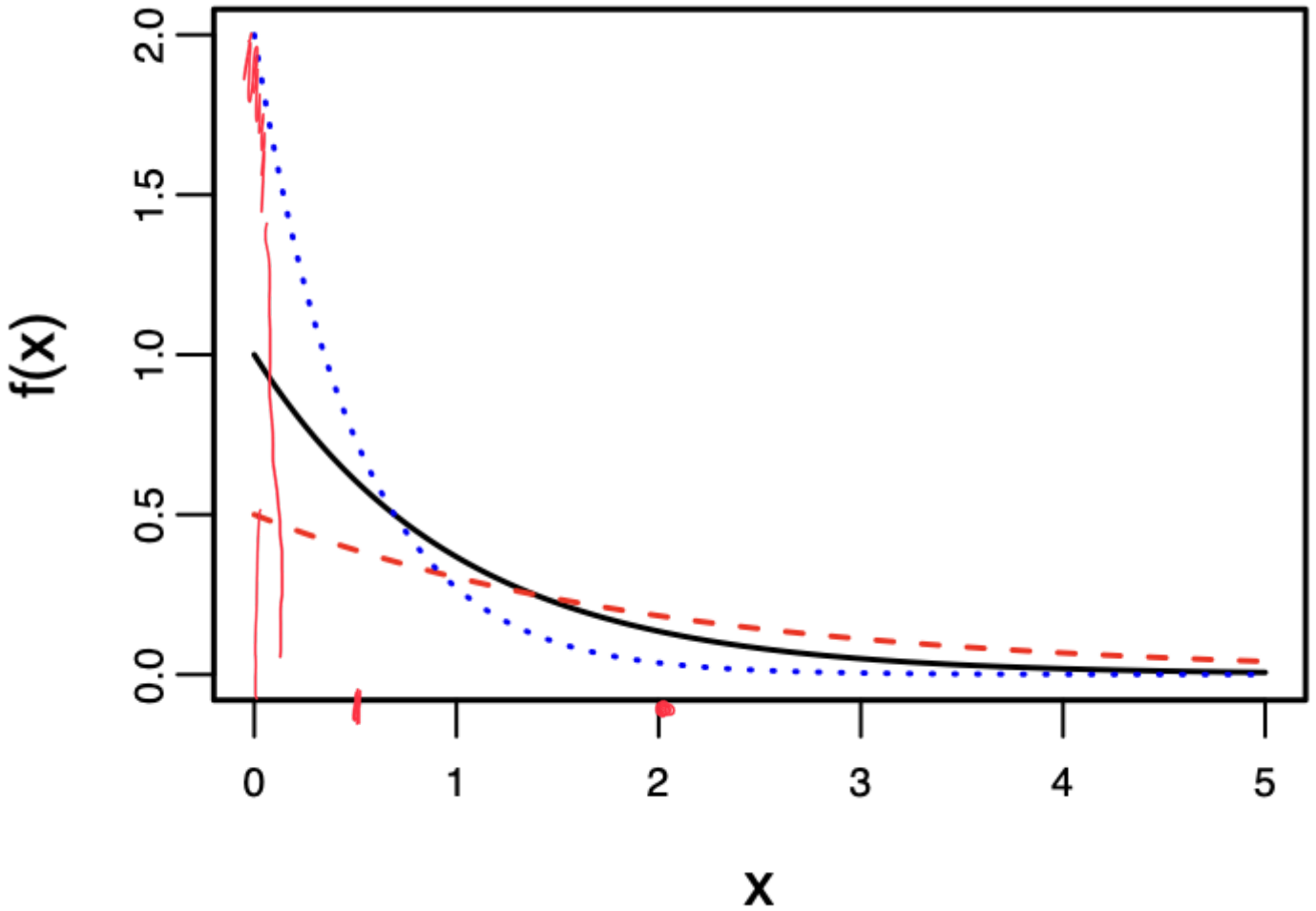


Abbildung 4: $\lambda = 1$ (black), $\lambda = 2$ (blue) and $\lambda = 1/2$ (red).

Exponential Distribution

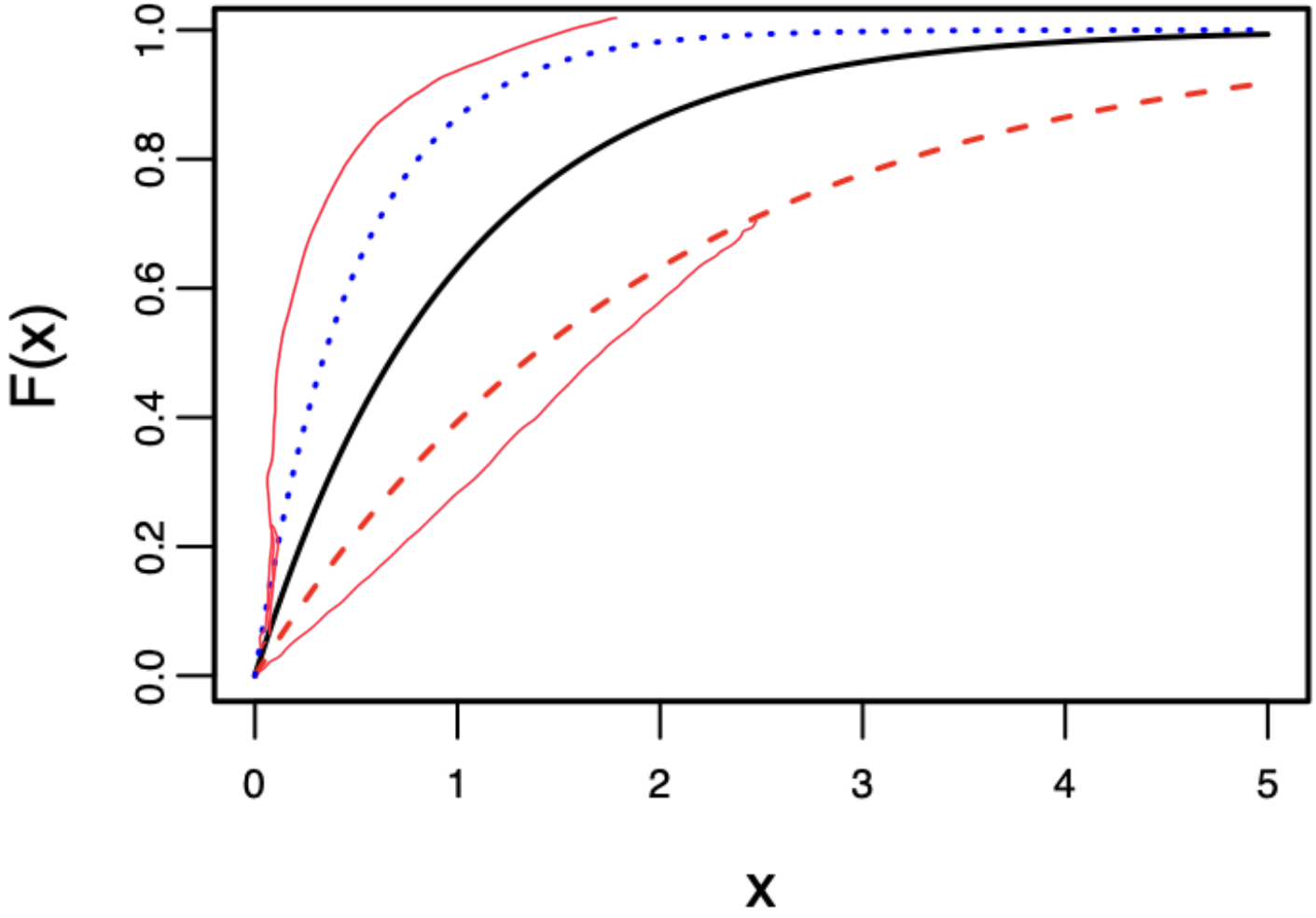


Abbildung 5: $\lambda = 1$ (black), $\lambda = 2$ (blue) and $\lambda = 1/2$ (red).

Example

Setting: The lifetime T of a computer chip is exponentially distributed, i.e., $T \sim \text{Exp}(\lambda)$ with expected lifetime of 15 weeks, i.e., parameter $\lambda = \frac{1}{15}$

Question:

- What is the probability that the computer chip is defect within the first 10 weeks?

$$\mathbb{P}(T \leq 10) = F(10) = 1 - e^{-\lambda \cdot 10} = 1 - e^{-10/15} = 0.487$$

- What is the probability that the computer chip will last at least 20 weeks?

$$\mathbb{P}(T \geq 20) = 1 - \mathbb{P}(T < 20) = 1 - F(20) = e^{-20/15} = 0.264$$

Transformation

Linear case: g linear i.e., $g(x) = bx + a$ for $b \geq 0$, $Y = g(X)$

$$F_Y(y) = P(Y \leq y) = P(bX + a \leq y) = P\left(X \leq \frac{y-a}{b}\right)$$

$$= F_X\left(\frac{y-a}{b}\right)$$

case $b < 0$ we have

$$F_Y(y) = P(Y \leq y) = P(bX \leq y-a) = P\left(X > \frac{y-a}{b}\right)$$

$$1 - F_X\left(\frac{y-a}{b}\right)$$

Transformation

On the level of the densities this boils down to:

for $b \neq 0$ the density of Y with respect X is:

$$f_Y(y) = \frac{1}{|b|} \cdot f_X\left(\frac{y-a}{b}\right)$$

Example: $X \sim N(\mu, \sigma^2)$ $\quad Y = a + bX \quad \Rightarrow \quad Y \sim N(a + b\mu, b^2\sigma^2)$

with the density being

$$f_Y(y) = \underbrace{\frac{1}{\sqrt{2\pi}\sigma|b|} \exp\left\{-\frac{1}{2}\left(\frac{\frac{y-a}{b} - \mu}{\sigma}\right)^2\right\}}_{f_X\left(\frac{y-a}{b}\right)}$$

Transformation

Transformation

Reminder: for arbitrary g the following holds:

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx \tag{4}$$

Proposition: Let g be a differentiable, strictly monoton function and X and random variable. Then $Y = g(X)$ has the following density

$$f_Y(y) = \left| \frac{1}{g'(g^{-1}(y))} \right| f_X(g^{-1}(y)), y \in E_Y \tag{5}$$

E_Y is given by the value space of X via

$$E_Y = g(E_X) = \{g(x) : x \in E_X\} \tag{6}$$

Example: Lognormal distirbution

Example : $X \sim N(\mu, \sigma^2)$ normal distribution

Then we call $Y = e^X$ lognormaldistribution

with

$$f_Y(y) = \begin{cases} 0 & y \leq 0 \\ \frac{1}{\sqrt{2\pi}\,\sigma\, y} \exp\left\{-\frac{1}{2}\left(\frac{\log(y) - \mu}{\sigma}\right)^2\right\} & y > 0 \end{cases}$$

Jensen's inequality

Proposition: Let g be a convex function and X random variable

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X]) \tag{7}$$

Example:

For a lognormal-distribution Y

$$\mathbb{E}[Y] = e^{\mu + \sigma^2/2} > g(\mu) = e^{\mu}$$

Samples

Definition: Let $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space and $X_1, \ldots, X_n$ be associated random variables. Realizations

$$x_1 := X_1(\omega), \ldots, x_n := X_n(\omega) \tag{8}$$

are referred to as *samples* and n the sample size.

Estimator

Definition: A measurable function $\varphi : \mathbb{R}^n \to \mathbb{R}^m$ is referred to as *sample function*, *estimator* or *statistic*.

Note: we will also consider the composition:

$$\varphi(X) : \Omega \to \mathbb{R}^m \tag{9}$$

$$\omega \mapsto \varphi(X_1(\omega), \ldots, X_n(\omega)) \tag{10}$$

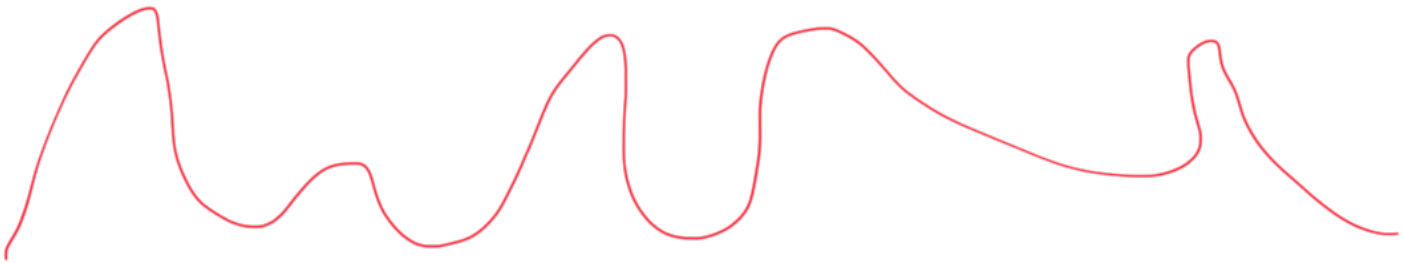
Sample estimation

Given: $(x_1, \ldots, x_n) \in \mathbb{R}^n$ of independent and identical random variables $X_1, \ldots, X_n$ where

$$F(t) = \mathbb{P}[X_i \leq t], \quad t \in \mathbb{R} \tag{11}$$

but **unknown**

Goal: estimate $\mathbb{E}[X_i]$ or $Var[X_i]$



Empirical mean

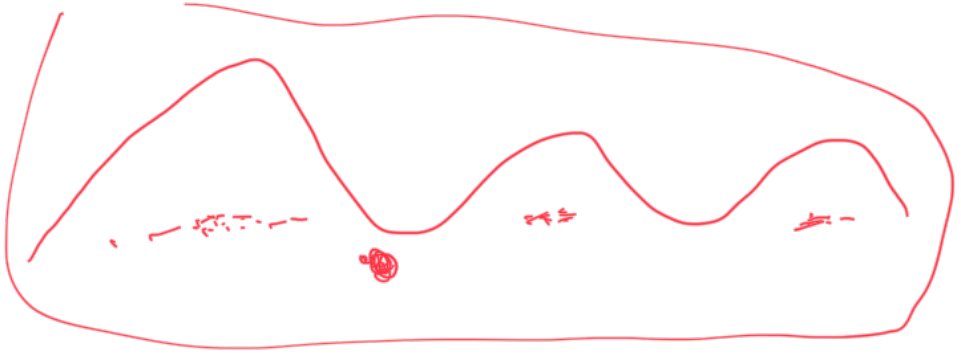
Definition: The empirical mean is defined by

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{12}$$

Note: we will also use an analog notation for the random variables:

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i \tag{13}$$

RV



Random variables

Proposition: Let $X_1, \ldots, X_n$ be independent and identical random variables with $\mathbb{E}[X_i] = \mu$ and $Var[X_i] = \sigma^2$. Then

$$\mathbb{E}[\bar{X}_n] = \mu \text{ and } Var[\bar{X}_n] = \frac{\sigma^2}{n} \tag{14}$$

Proof

- $\mathbb{E}[\bar{X}_n] \overset{\text{Def of } \bar{X}}{=} \mathbb{E}\left[\frac{X_1 + \dots + X_n}{n}\right] = \frac{1}{n} \cdot \mathbb{E}[X_1 + \dots + X_n]$

$= \frac{1}{n} \cdot n \cdot \mathbb{E}[X_i] = \mu$ (where $\mathbb{E}[X_i] = \mu$) □

$\frac{1}{n}\left(\mathbb{E}[X_1] + \dots + \mathbb{E}[X_n]\right)$

- $\mathrm{Var}(\bar{X}_n) = \mathrm{Var}\left(\frac{X_1 + \dots + X_n}{n}\right) = \frac{1}{n^2} \cdot \mathrm{Var}(X_1 + \dots + X_n)$

$= \frac{1}{n^2} \cdot n \cdot \mathrm{Var}(X_i) = \frac{\sigma^2}{n}$ □

Use this because X_i are independent

$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$ if X and Y are independent

Proof

Law of large numbers

Proposition: Let $X_1, \ldots, X_n$ be independent and identical random variables with $\mathbb{E}[X_i] = \mu$. Then

$$\bar{X}_n \to \mu \text{ for } n \to \infty \text{ (almost sure)} \tag{15}$$

Empirical variance

Definition: The empirical variance is defined by

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 \tag{16}$$

Note: we will also use an analog notation for the random variables:

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \tag{17}$$

Empirical variance

Proposition: Let $X_1, \ldots, X_n$ be independent and identical random variables. Then

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i^2 - n\bar{X}_n^2) \tag{18}$$

Proof

Proof

Empirical variance

Proposition: Let $X_1, \ldots, X_n$ be independent and identical random variables with $\mathbb{E}[X_i] = \mu$ and $Var[X_i] = \sigma^2$. Then

$$\mathbb{E}[S_n^2] = \sigma^2 \tag{19}$$

Proof

Proof