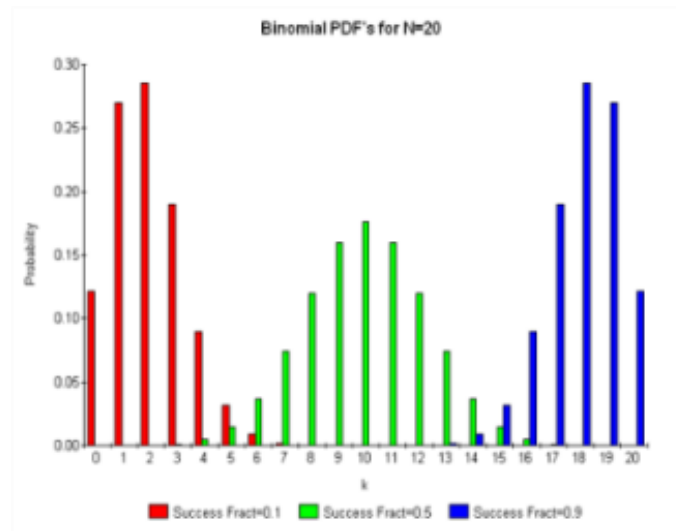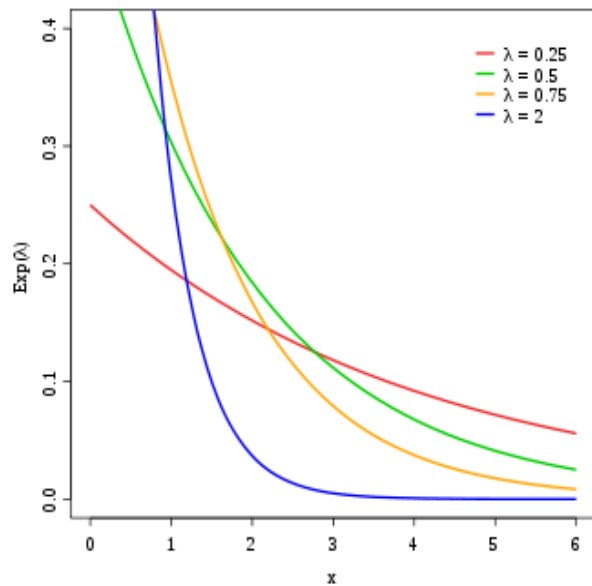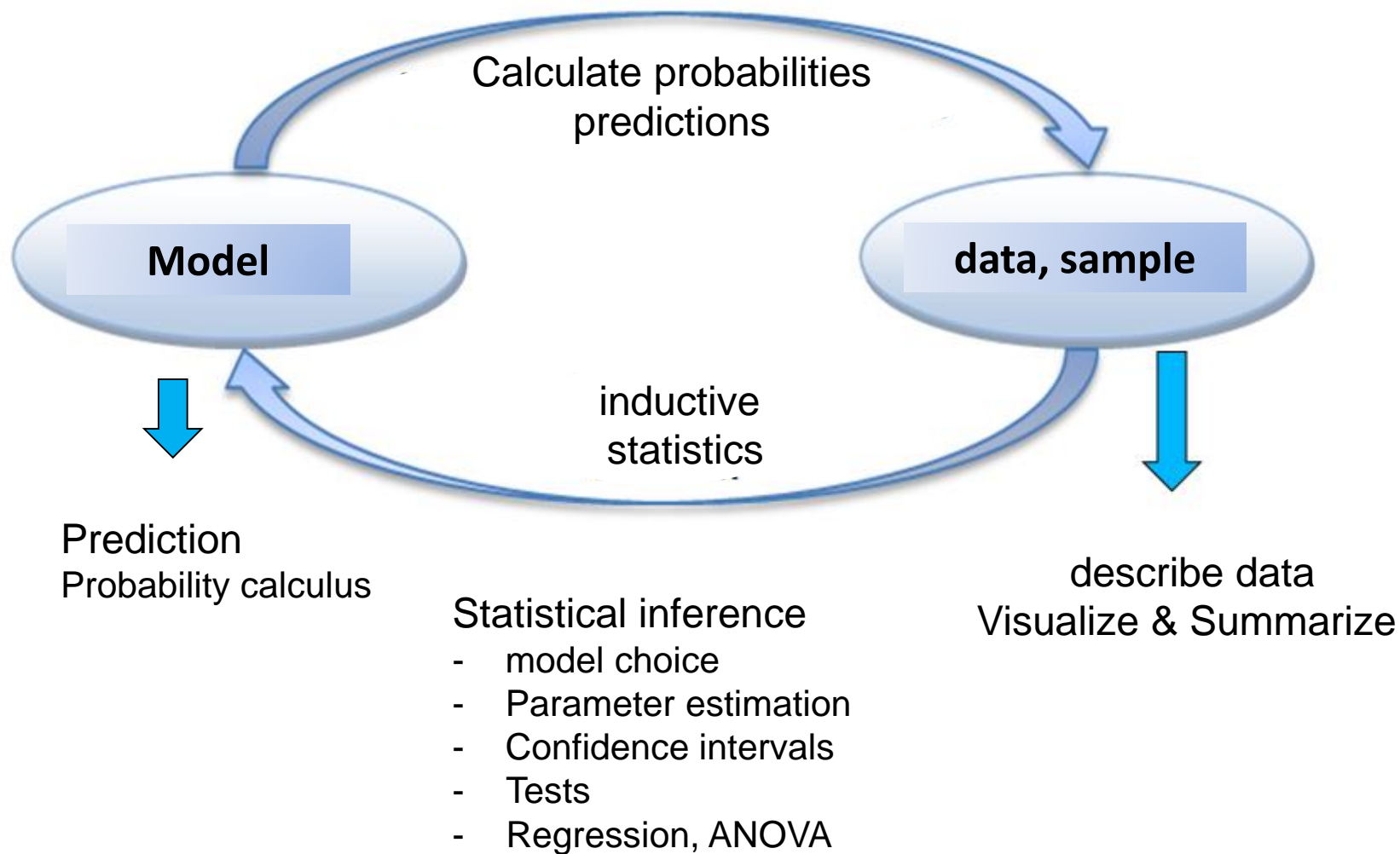# Biostatistics Week 3

Inferential statistics

➢ **What is inferential statistics about?**

➢ **The zoo of distribution models & model choice**

➢ **Expected value and variance and how to estimate them**

Statistics connects data with models

Calculate probabilities
predictions

Model

data, sample

inductive
statistics

Prediction
Probability calculus

Statistical inference
- model choice
- Parameter estimation
- Confidence intervals
- Tests
- Regression, ANOVA

describe data
Visualize & Summarize

What is a model?

Model for the human heart

Model airplane in a wind tunnel

EXPONENTIAL HISTOGRAM

model ($f=e^{-x}$) : probability density
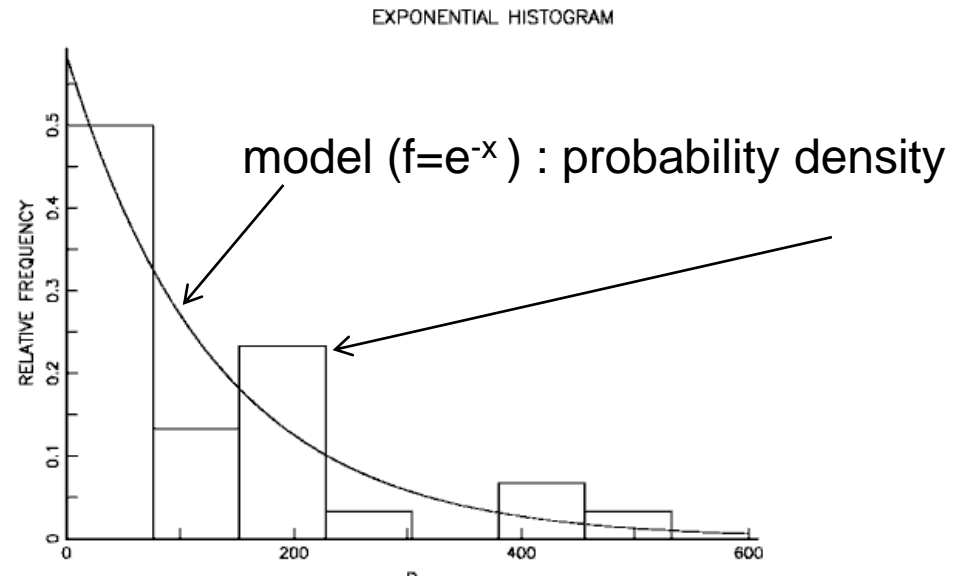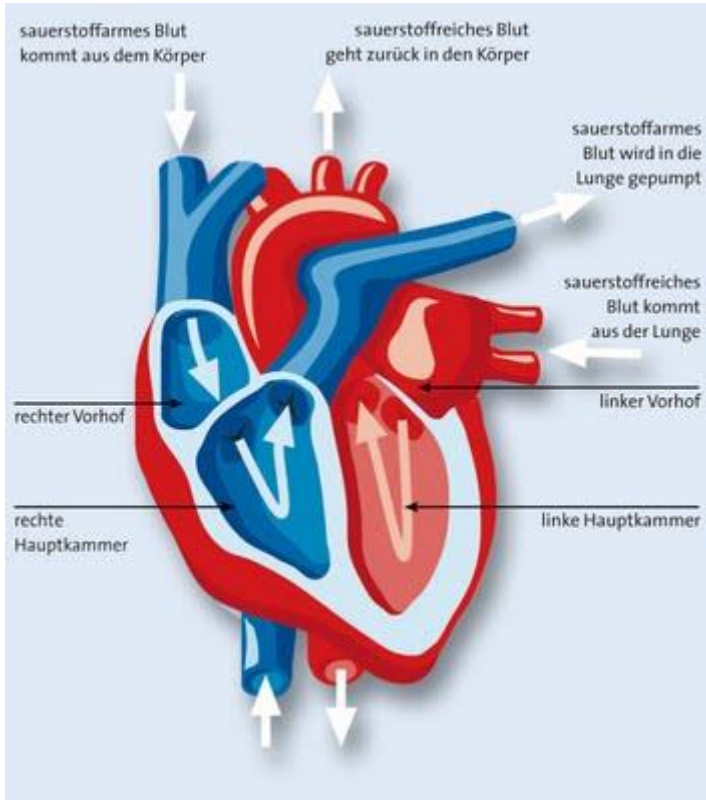
The world of data and the world of models

data/reality	model
sample	population
discrete data/features (numeric or categorical)	discrete random variable (numeric)
continuous data/features (numeric)	continuous random variable (numeric)
observation	Random variable
relative frequency	probability (P)
histogram (scaled)	Density continuous distribution
bar plot of frequencies (scaled) (of rel. frequency at discrete features)	Probability distribution discrete distribution
average $\overline{x}$	expected value μ
sample variance s^2	variance σ^2

What can be meant by probability?

1. The Probability to get a value in a certain range can be given as limit of relative frequencies.

2. Probability can be determined from a probability model – we have a model for dice in our mind which tells us that the probability for each side is equal.

3. An opinion resp. experience or expectation, can be expressed as probability. E.g. the probability of a airplane crashing in a nuclear power station is given by expert opinion.

4. In mathematical statistics probability is given by the Kolmogorov's formulation, sets are interpreted as events and probability itself as a measure (probability) on a class of sets.
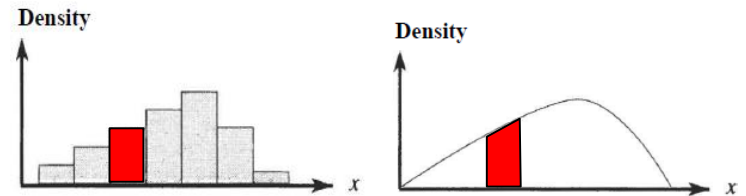
Axiom 1	$0 \leq P(A) \leq 1$
Axiom 2	$P(S) = 1$
Axiom 3	$P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

Example for a distribution model:
the discrete uniform distribution

X: Result when throwing a dice (X is a random variable)

model

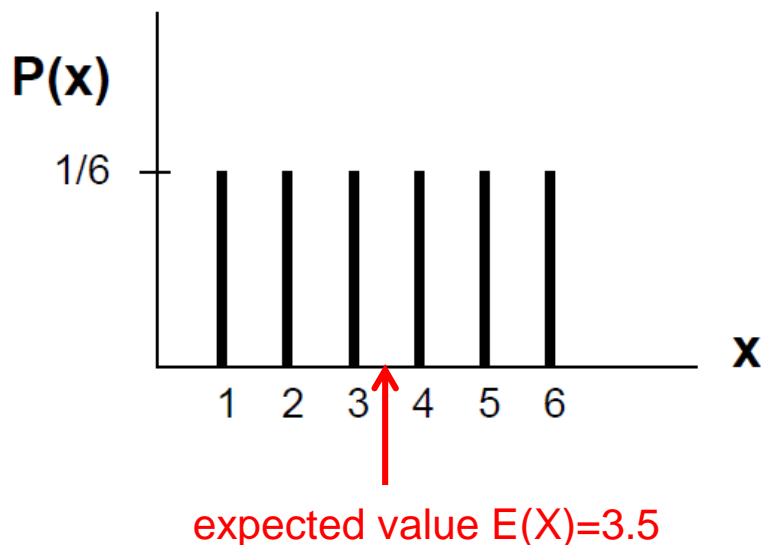Probability distribution in table representation:

X	1	2	3	…	6
p	1/6	1/6	1/6	…	1/6

generates rv X

Probability distribution as graph:

P(x)

1/6

1 2 3 4 5 6
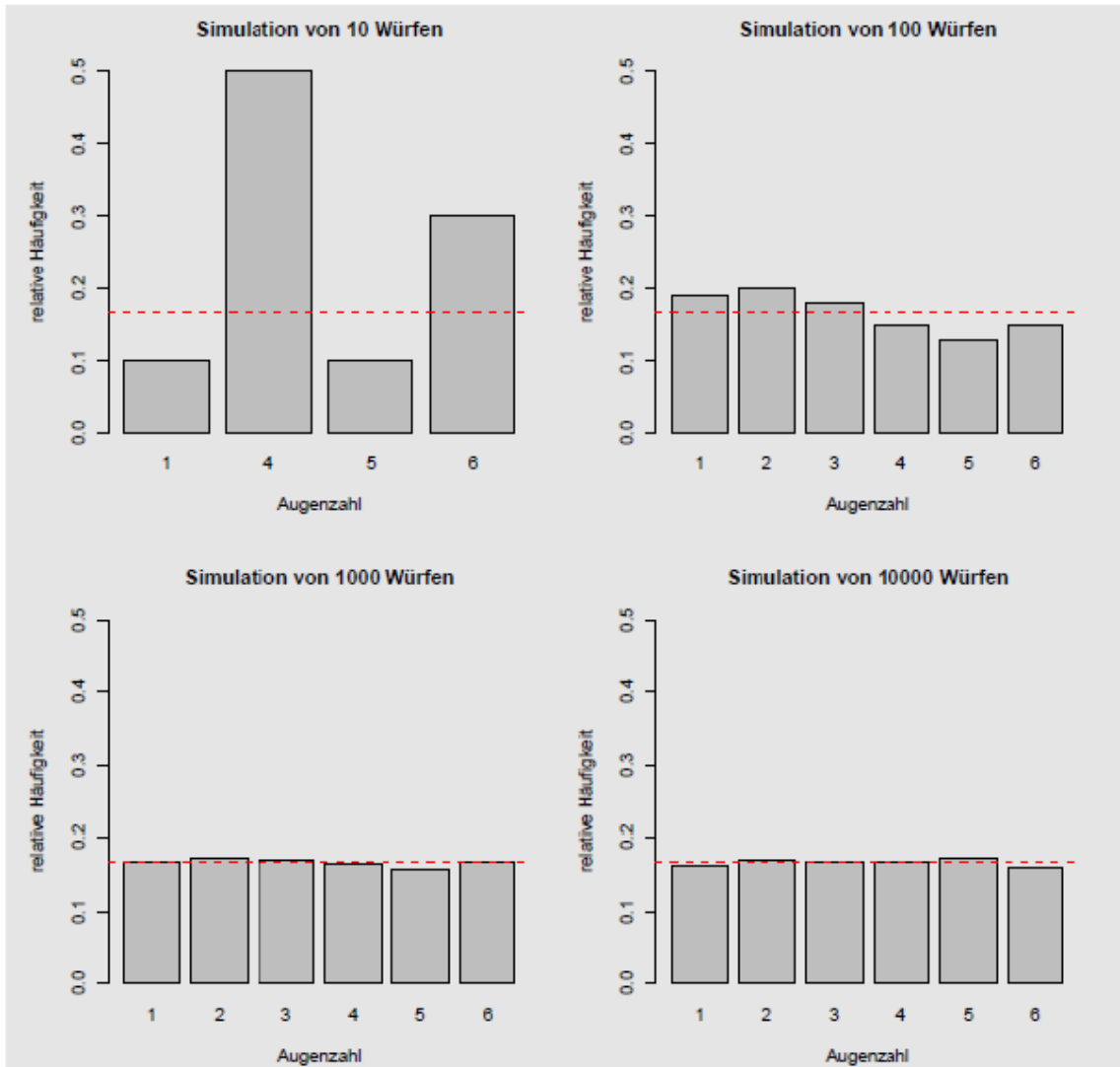
X

expected value E(X)=3.5

Probability distribution as function:

$$P(X = k) = \frac{1}{r} \; , \; k \in \{1,2,...,r\}$$

r: # possibel results

$$\sum_{k=1}^{r} P(X = k) = 1$$

Estimate the probability distribution and expected value via determining the relative frequencies by experiment or simulation

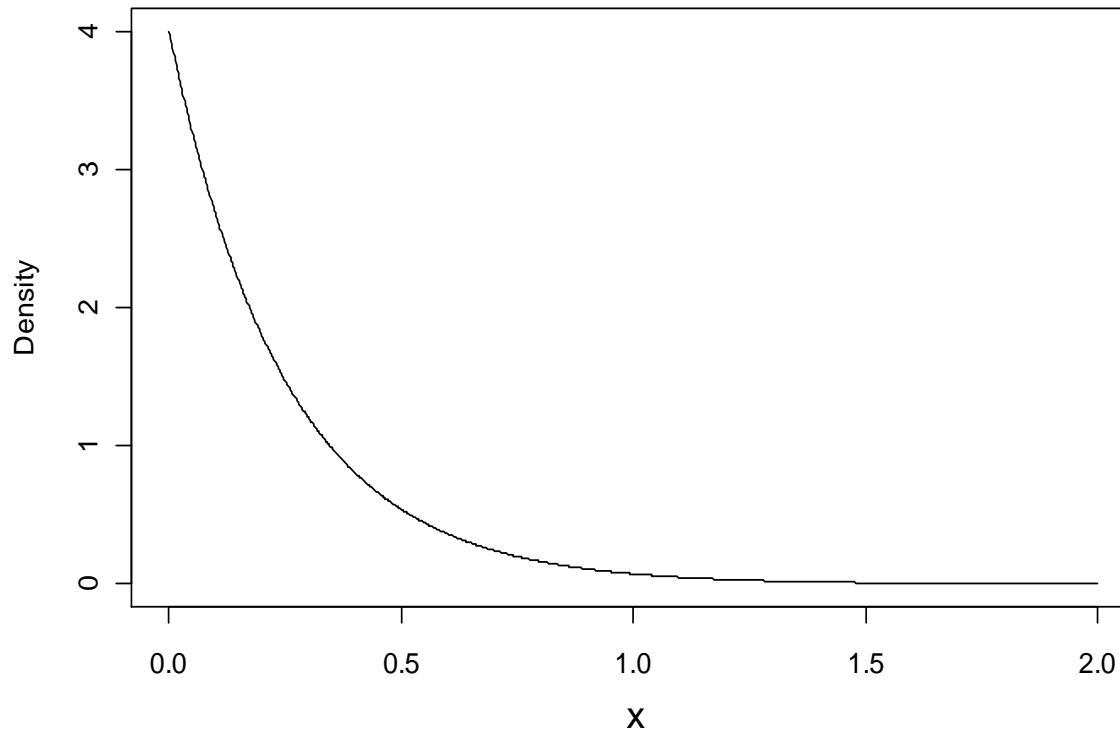We need >10'000 throws of a dice or runs of simulations before getting a reliable estimate of the distribution or the expected value.

The scaled histogram of waiting times for an increasing number of observations
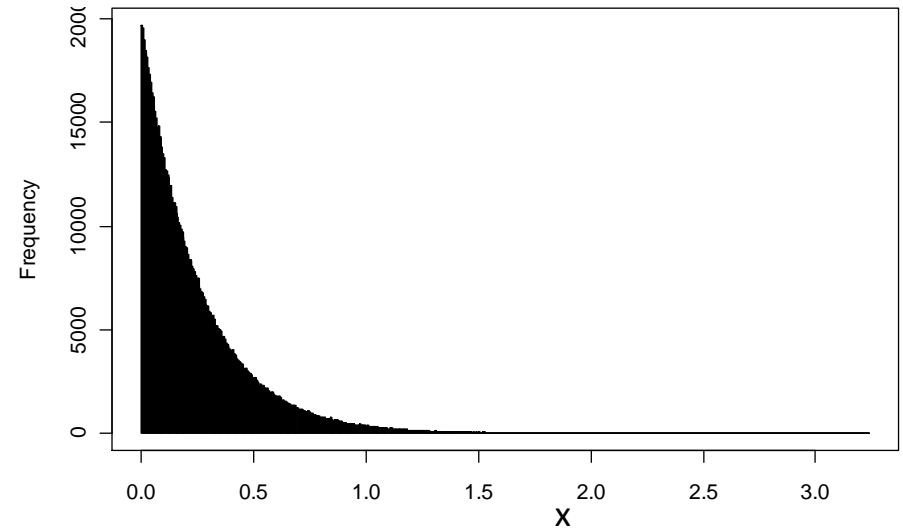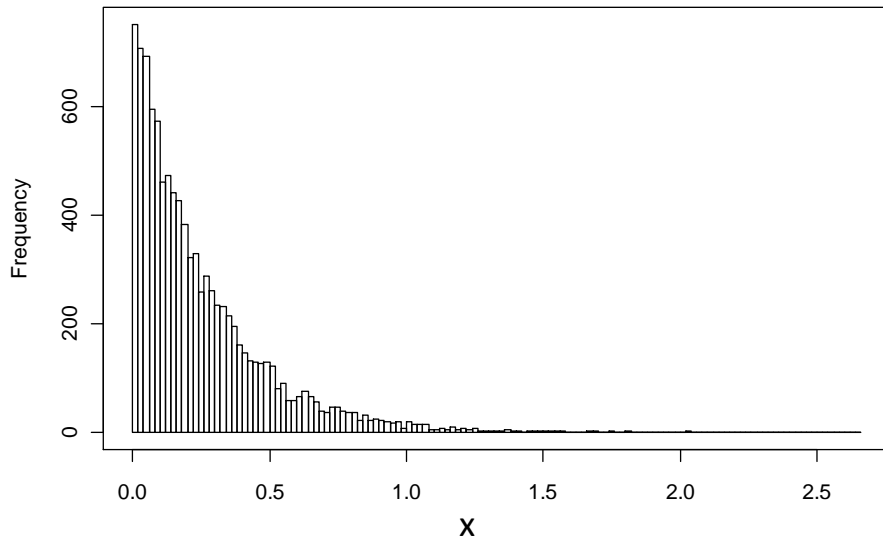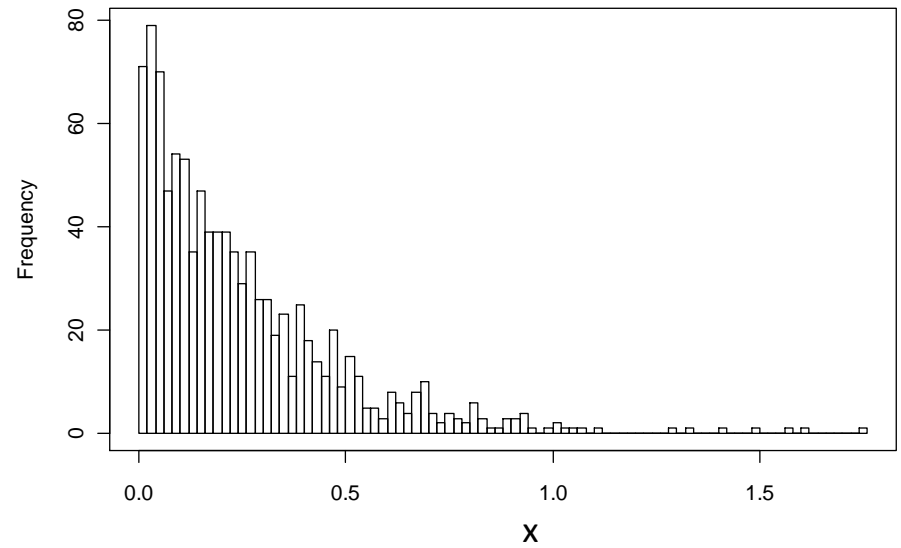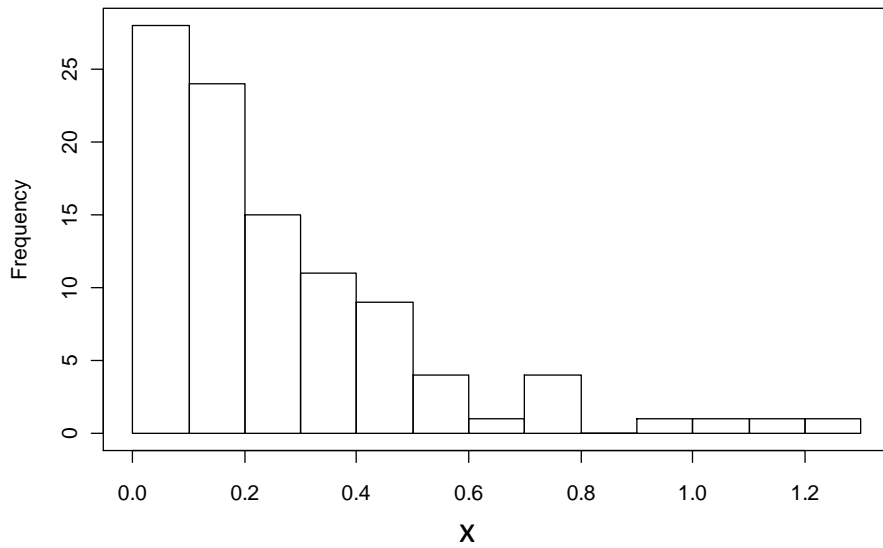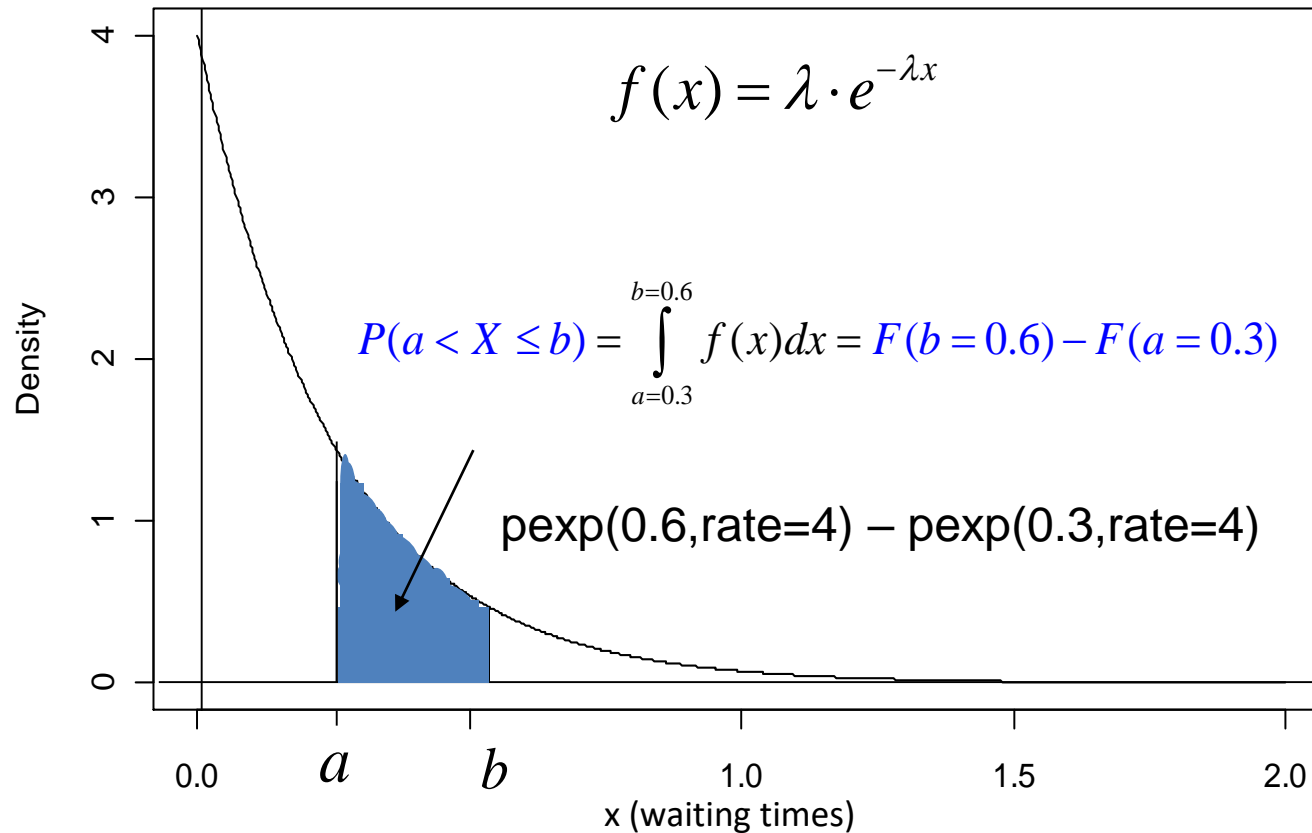
The limit-case of infinity observations

If we have infinity observations the hull of a scaled histogram can be modeled by smooth function, called the density function, with an area under the curve of one.

The probability density function (pdf)

$$f(x) = \lambda \cdot e^{-\lambda x}$$

$$P(a < X \le b) = \int_{a=0.3}^{b=0.6} f(x)dx = F(b=0.6) - F(a=0.3)$$

pexp(0.6,rate=4) − pexp(0.3,rate=4)

Density

x (waiting times)

The probability of getting a result between a and b is equal to the area under the density function above the interval [a,b]. The calculation of the probability is made by integrating the density function in interval [a,b]

Properties of a density function

The density function f of a continuous variable X
is a piece-wise continuous function with

$$f(x) \geq 0$$

and

$$\int_{-\infty}^{+\infty} f(x)dx = 1$$

11

The expected value = population mean

•The expected value of a random variable is the average, which we would get with an infinite big sample

• It measures the location of the random variable

• It corresponds to the center of mass of the density – balance point (see red line)

• It often determines the parameter of the model

• The expected value can also be calculated from the model

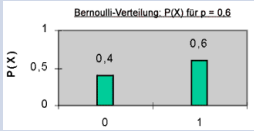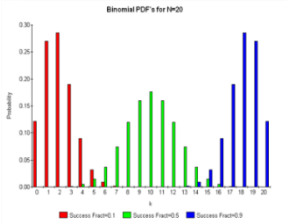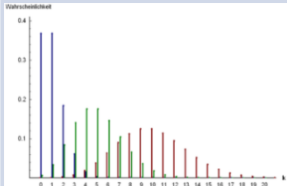$$EW(X) = \sum_{i=1}^{n} x_i \cdot P(X = x_i)$$

$$EW(X) = \int x \cdot f(x) \, dx$$

Po(λ=2.5)

Exponentialverteilung
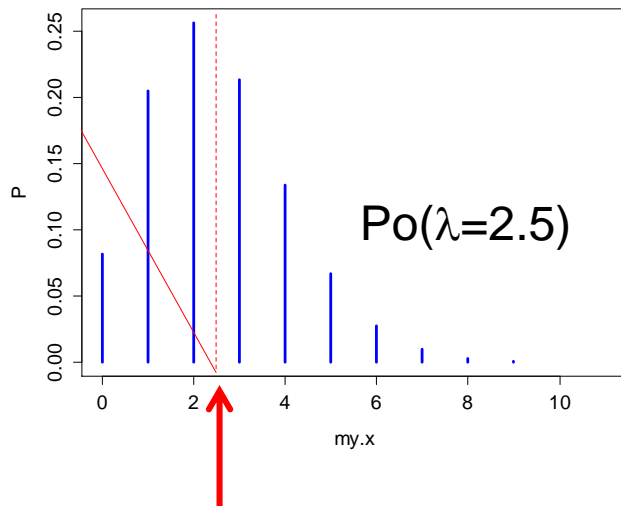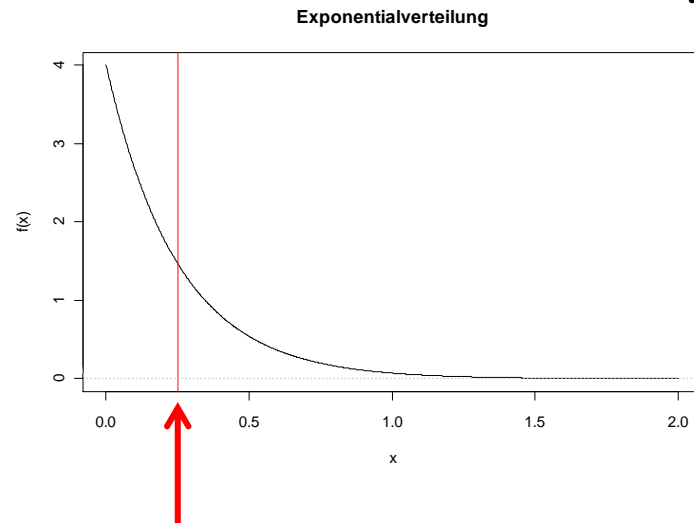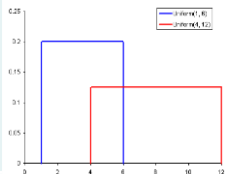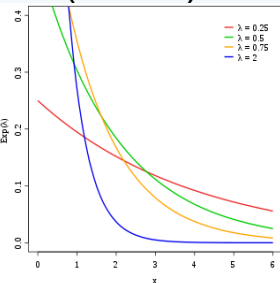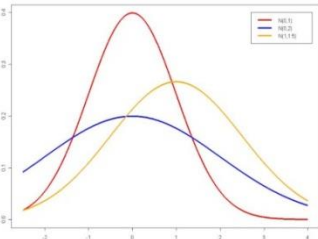
expected value E(X)

expected value E(X)

The most famous discrete distributions/models

name of the distribution	possible values x $P(X=k)$	expected value μ variance σ^2	application
Bernoulli $X \sim Bern(\pi)$	$x \in \{0,1\}$ $$P(X=1)=\pi$$ $$P(X=0)=1-\pi$$	$\mu=E(X)=\pi$ $\sigma^2=Var(X)=\pi*(1-\pi)$	X: indicates if an event occurs or not
Binomial $X \sim B(n,\pi)$	$x \in \{0,1,...,n\}$ $$P(X=k)=\binom{n}{k} \cdot \pi^k \cdot (1-\pi)^{n-k}$$	$\mu=E(X)=n*\pi$ $\sigma^2=Var(X)=n*\pi*(1-\pi)$	X: number of successes in n independent Bernoulli trials
Poisson $X \sim Po(\lambda)$	$x \in \{0,1,...\}$ $$P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda}$$	$\mu=E(X)=\lambda$ $\sigma^2=Var(X)=\lambda$	X: number of events in a certain interval or time-bin

The most important continuous distributions/models

Name of the distribution (parameter)	domain density f distribution F	expected value variance	application
Uniform V. $X \sim U(a,b)$	$\mathbb{R}$ $f(x) = \dfrac{1}{b-a}$, $\quad$ *for* $a \pounds x \pounds b$, $\quad$ *otherwise* $f(x) = 0$ $F(x) = \dfrac{x-a}{b-a}$ *für* $a \le x \le b$	$E(X) = \dfrac{a+b}{2}$ $Var(X) = \dfrac{(b-a)^2}{12}$	if all events have the same probability or if the probability is not known at all
Exponential V. (rate λ)	$\mathbb{R}_0^+$ $f(x) = \lambda \cdot e^{-\lambda x}$ $F(x) = 1 - e^{-\lambda x'}$	$E(X) = \dfrac{1}{\lambda}$ $Var(X) = \dfrac{1}{\lambda^2}$	waiting times, time to fail (or decay)
Normal V. $X \sim N(\mu, \sigma^2)$	$\mathbb{R}$ $f(x) = \dfrac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ $F(x) = \displaystyle\int_{-\infty}^{x} \dfrac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x'-\mu)^2}{2\sigma^2}} \, dx'$	$E(X) = \mu$ $Var(X) = \sigma^2$	typical measurements (affected symmetrically by various factors), Asymptotic approximation for other distributions

Model choice

Which model fits the data better?
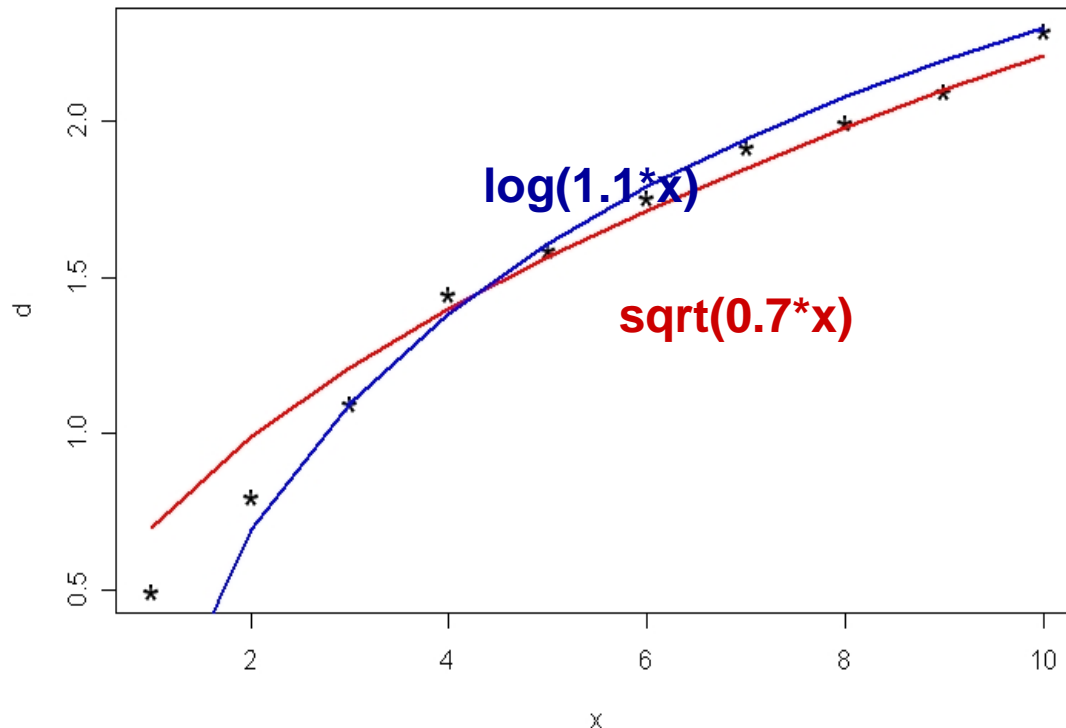
Logarithms or Square-Root function?

We should always check if the model is appropriate and fits the data(visually, by residual analysis, qq-Plot…)

The Poisson model is appropriate when the measure of interest is the number of events per unit

Poisson Distribution for count data:

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda} \; , \; k = 0, 1, 2, ...$$

$E(X) = \text{Var}(X) = \lambda = $ mean number of events per unit (e.g. time bin)

Counts in PET measurement
Counts in Geiger-Mueller counter
Counts in a photo diode
Counts of phone calls in USZ

...

index of 5minute time bin
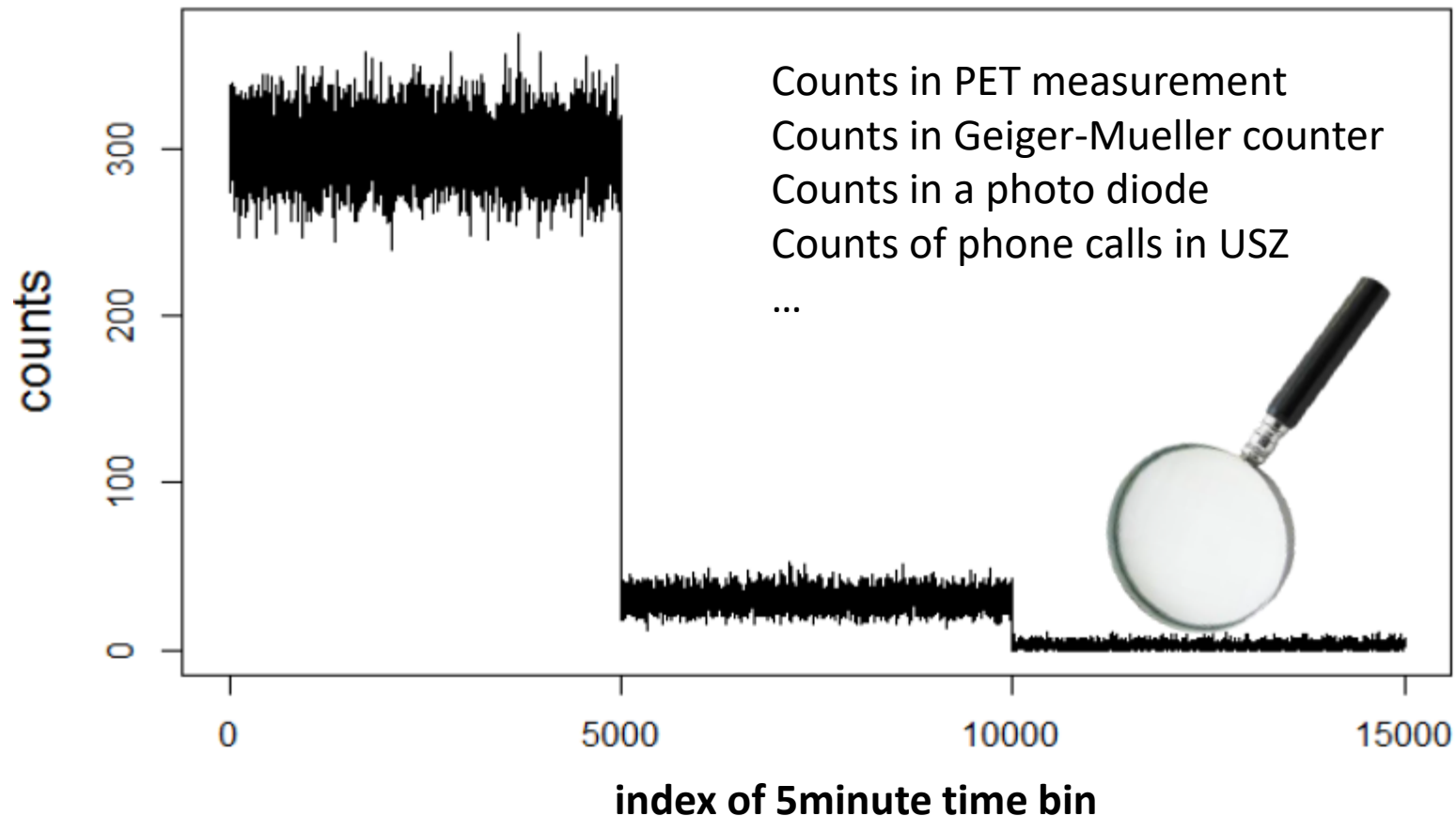
The Exponential model is often appropriate when we measure time to event, e.g. the time between 2 Poisson-events

$$f(x) = \lambda \cdot e^{-\lambda x}$$

$$\lambda = \frac{1}{\text{mean time lag}}$$
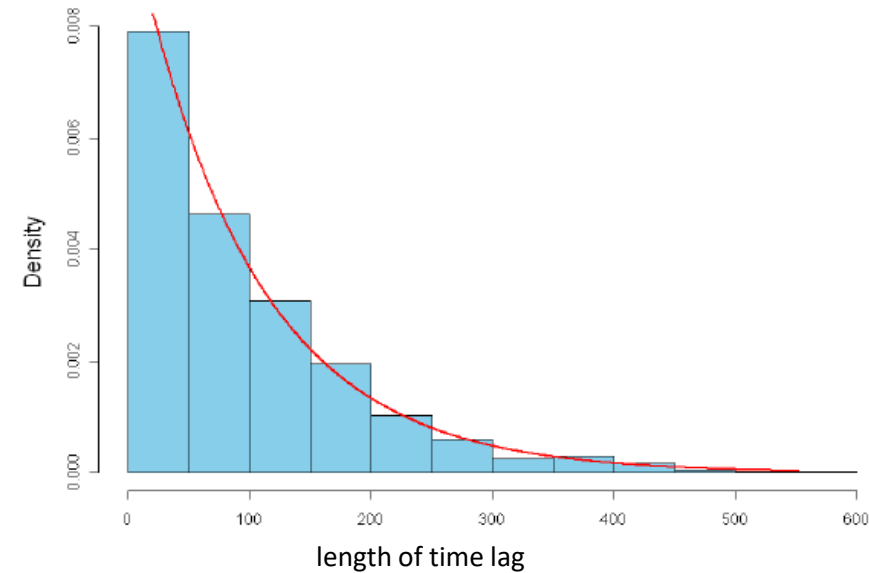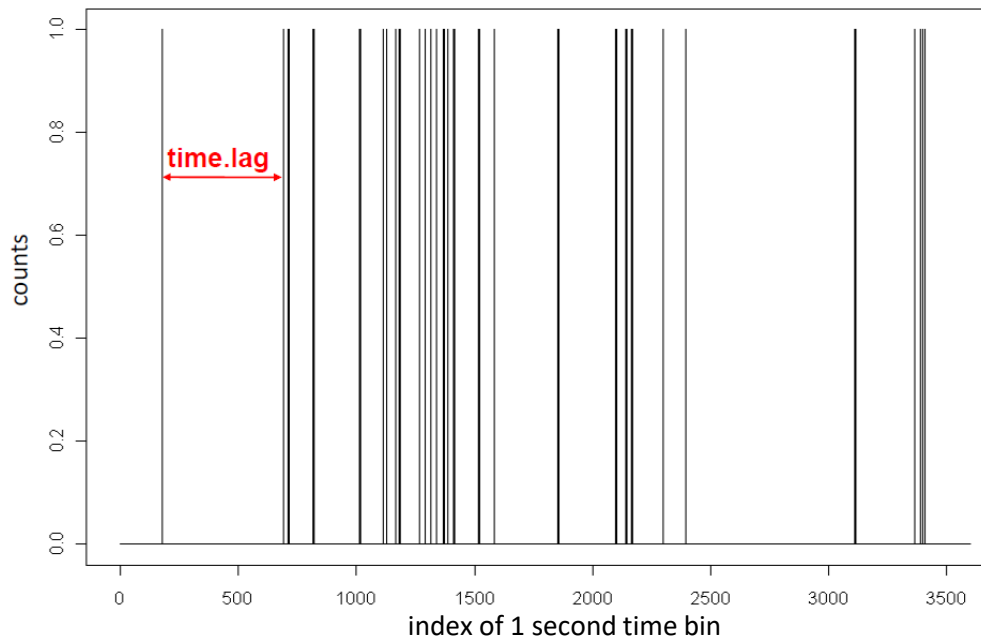
The parameter estimation problem

Location and the shape of a distribution is determined by the values of its parameters

We assume we know the appropriate model class.

How to estimate the parameter value from the data?
Which parameter value range is plausible?

Is a given parameter value compatible with the data?
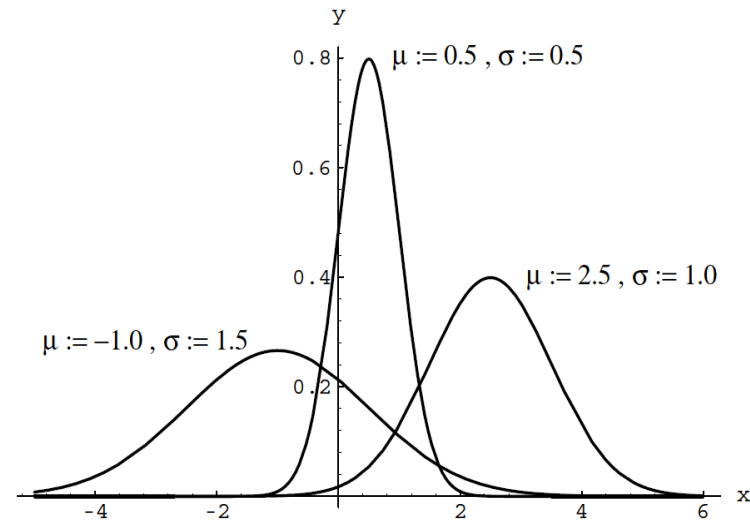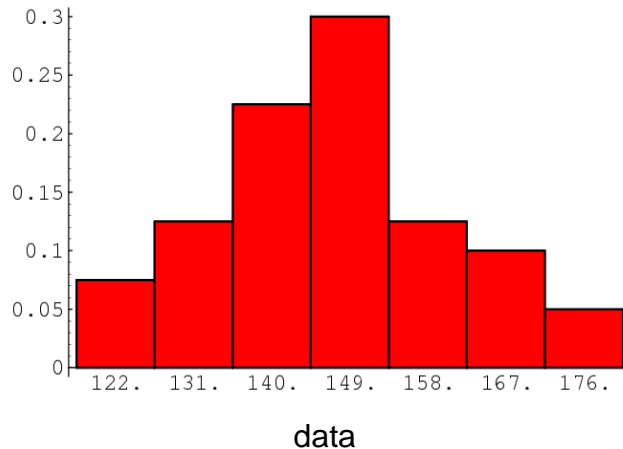
data

Paramter estimation for the most important distributions

Distributionfamily V X~V(Parameter-Set)	Relation Parameter-E(X)-Var(X)	Parameter-estimator as function of the data
Exponential X~Exp(λ)	$E(X) = \dfrac{1}{\lambda}$ $Var(X) = \dfrac{1}{\lambda^2}$	$\hat{\lambda} = \dfrac{1}{\hat{E}(X)} = \dfrac{1}{\bar{x}}$
Normal X~N(μ,σ^2)	$E(X) = \mu$ $Var(X) = \sigma^2$	$\hat{\mu} = \hat{E}(X) = \bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$ $\hat{\sigma}^2 = \hat{var}(X) = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
Binomial X~B(n,p)	$E(X) = n \cdot p$ $Var(X) = n \cdot p \cdot (1-p)$	$\hat{p} =$ average #successes per n trials
Poisson X~Po(λ)	$E(x) = \lambda$ $Var(X) = \lambda$	$\hat{\lambda} = \hat{E}(X) = \bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$

19

The expected value is often a parameter of the model

Poisson Distribution for count data:

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda} \ , \ \ k = 0, 1, 2, \ldots$$

$$E(X) = \sum_{k=1}^{\infty} k \cdot P(X = k) = \lambda$$

Poisson Distribution: Mean = 0.5

λ**=0.5**
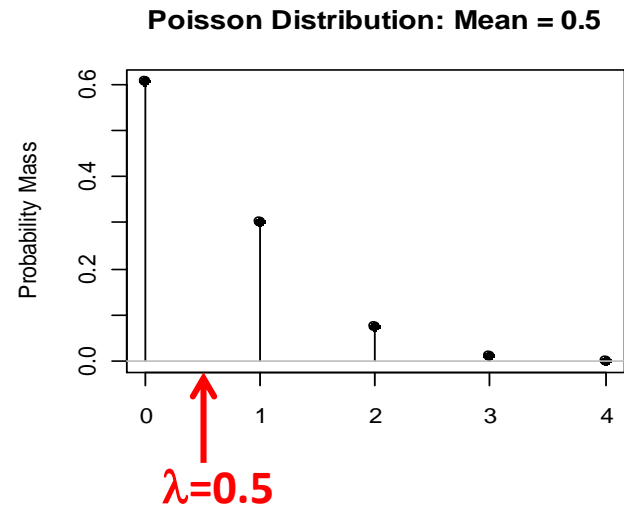
We use the sample mean to estimate the population mean = expected value (the central limit theorem (CLT) gives justification – see later in this lecture).

The ^ (hat) above the parameter indicates, that the parameter is not known but estimated.

$$\bar{X} \xrightarrow{\ n \to \infty\ } E(X)$$

$$\hat{E}(X) = \bar{X}$$

The average of the results obtained from a large number of trials should be close to the expected value, and will tend to become closer as more trials are performed.:
$$\overline{X} \xrightarrow{\ n\to\infty\ } E(X)$$
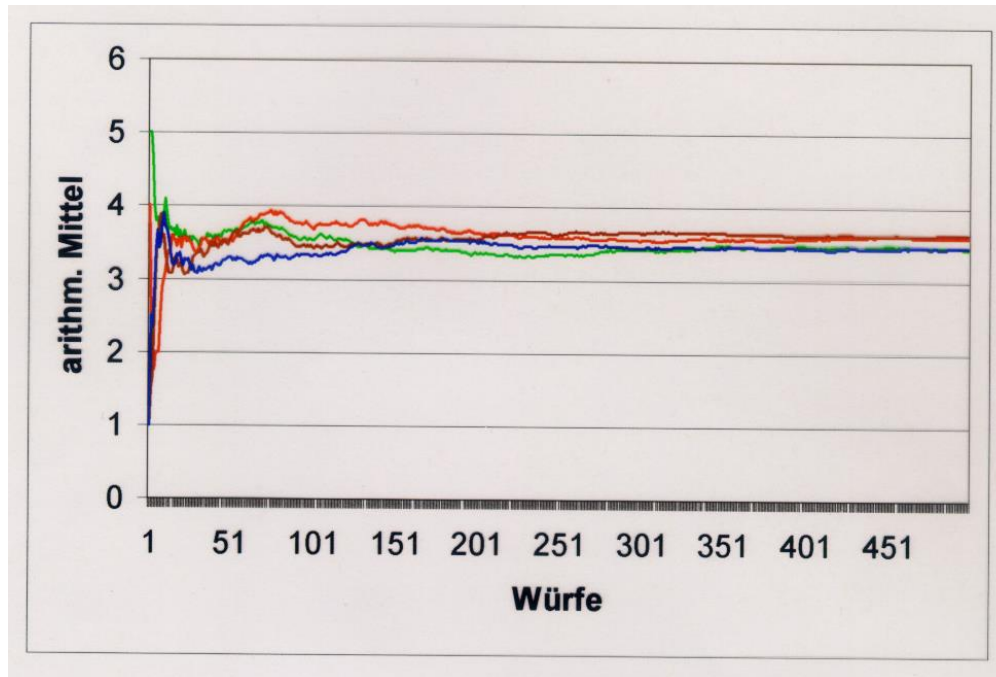
The expected value a dice throw is 3.5.

The estimated expected value is random since it is based on a random sample Sample Variation & Central Limit Theorem

➢ Because of the sample variation also derived statistics like the mean value varies from sample to sample

➢ The sample mean is an unbiased estimator for the population mean E(X).

➢ CLT: The sample mean is normaly distributed around the population mean and the variation decreases with increasing sample size.
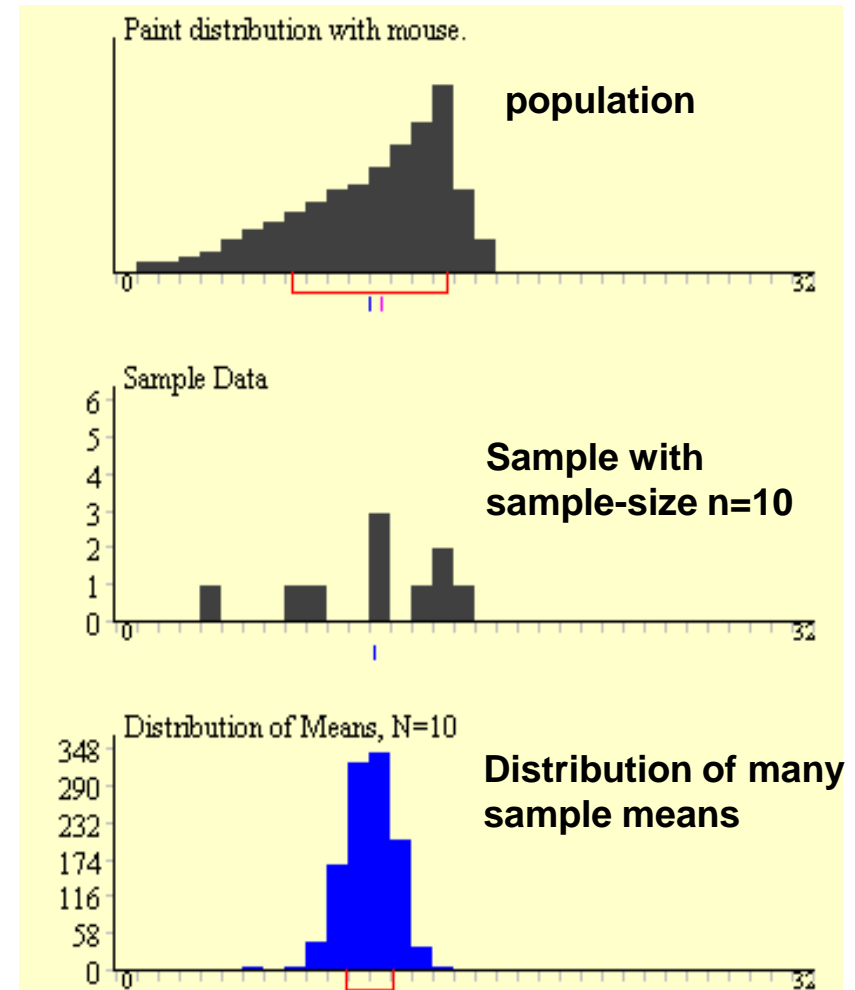
$$\bar{X} \overset{a}{\sim} N\left(\mu_x , \frac{\sigma_x^2}{n} \right) = N\left(E(X) , \frac{Var(X)}{n} \right)$$

Paint distribution with mouse.

population

Sample Data

Sample with sample-size n=10

Distribution of Means, N=10

Distribution of many sample means

Warning: There are exceptions to the CLT, e.g. t_1 also called Cauchy distribution or Lorenz curve in Spectroscopy which "has no mean".

Animation: http://onlinestatbook.com/stat_sim/sampling_dist/index.html

The distribution of the mean-> Central Limit Theorem

The arithmetic mean of a sufficiently large number of independent random variables, will be approximately normally distributed

← distribution of the population

← distribution of the mean (n=2)

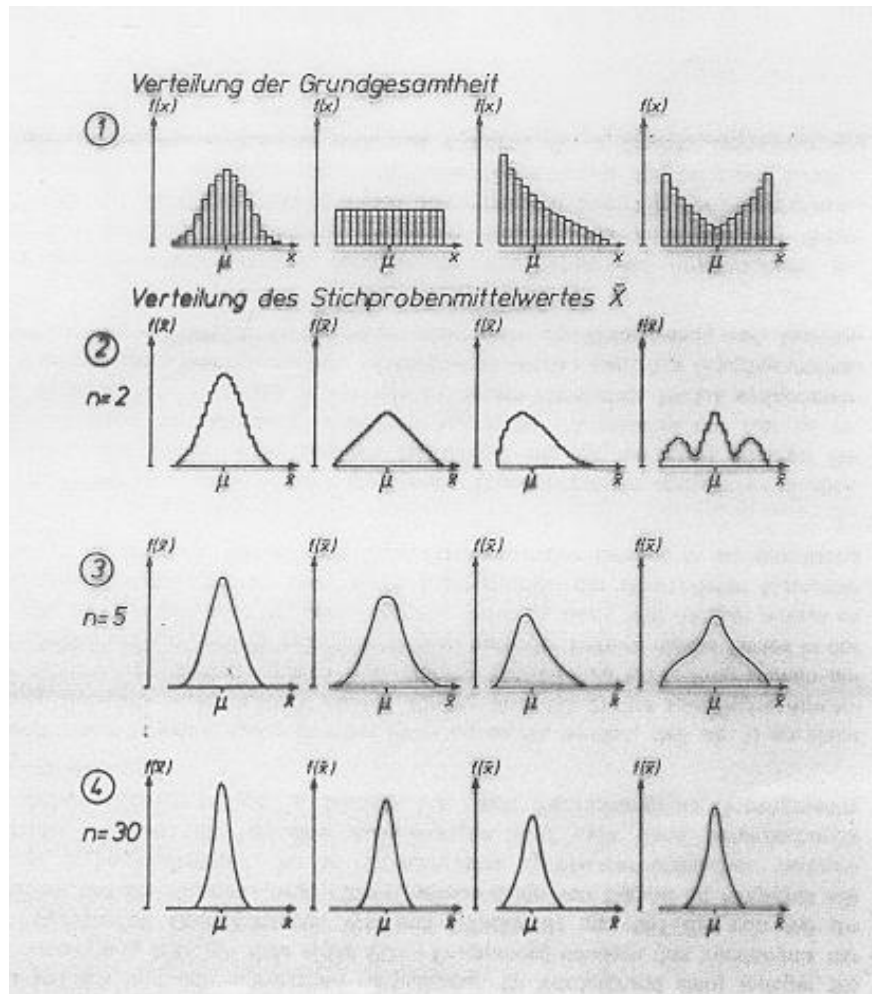← distribution of the mean (n=5)

← distribution of the mean (n=30)

Warning: There are exceptions to the CLT, e.g. t_1 also called Cauchy distribution or Lorenz curve in Spectroscopy which "has no mean".

Density of the Normal distribution

$$f(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \cdot e^{-\frac{1}{2} \cdot \frac{(x-\mu)^2}{\sigma^2}}$$

inflection point ⟷ P=0.68 ⟷ inflection point

P=0.95

P=0.997

μ−3σ   μ−2σ   μ−σ   μ   μ+σ   μ+2σ   μ+3σ
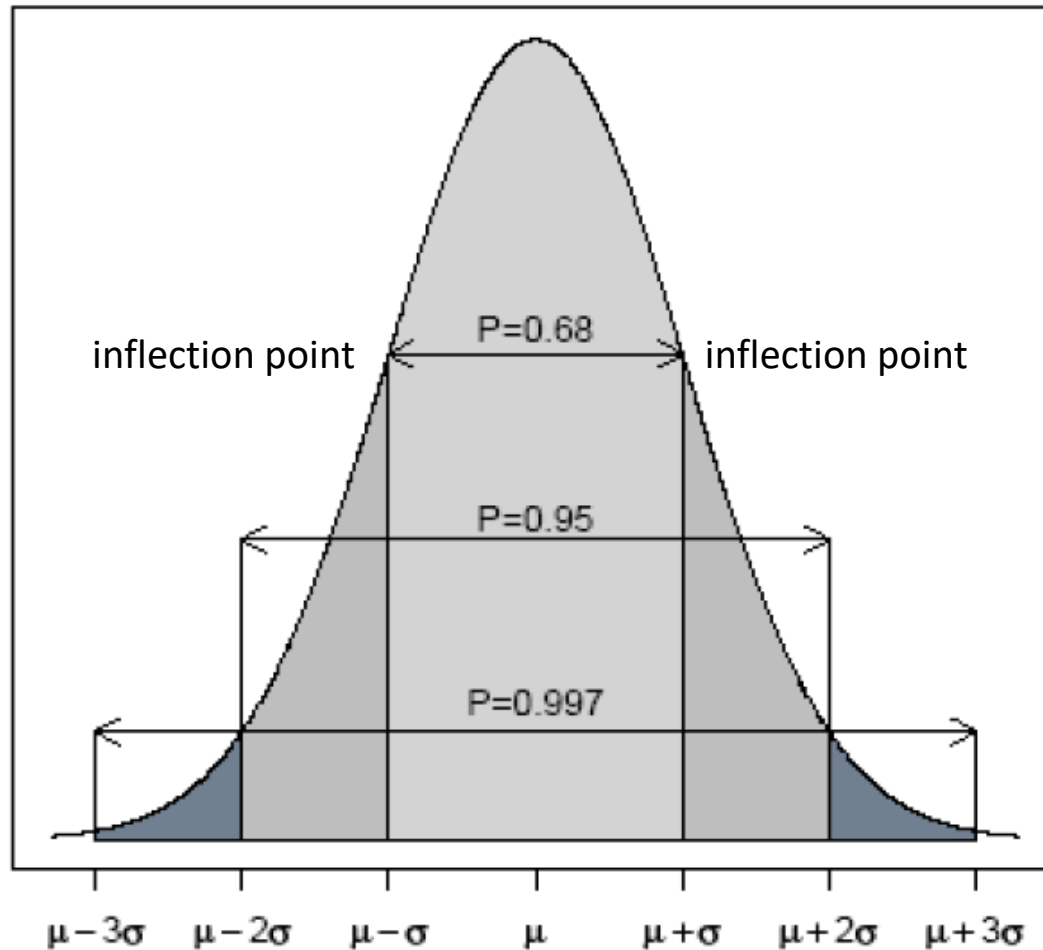
Rule of thumb:

A random value X~N(μ,σ²) has 95% of its probability mass within the following interval:

$$[\mu - 2\sigma, \mu + 2\sigma]$$

Or equivalently:

$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 95\%$$

At which intervals do people usually look?

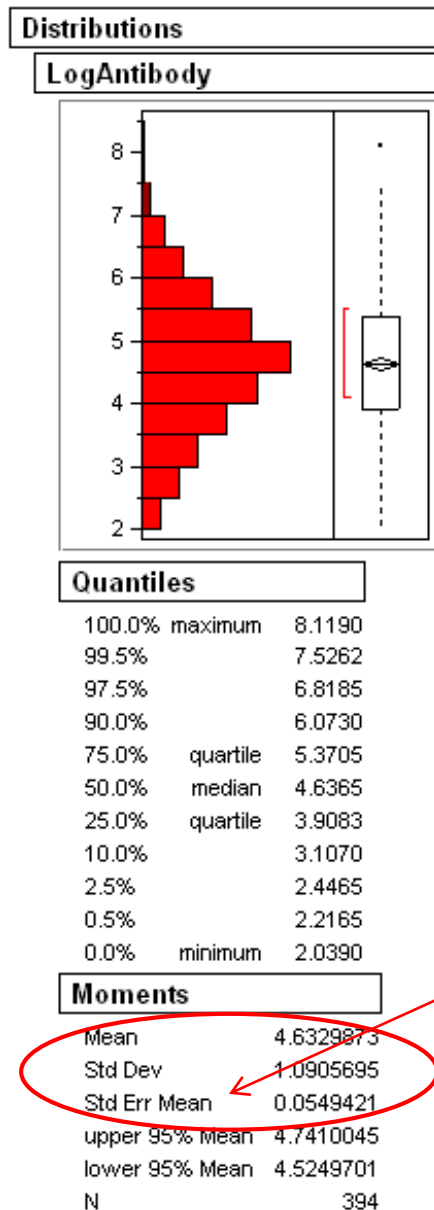**Distributions**
LogAntibody

Quantiles

100.0%	maximum	8.1190
99.5%		7.5262
97.5%		6.8185
90.0%		6.0730
75.0%	quartile	5.3705
50.0%	median	4.6365
25.0%	quartile	3.9083
10.0%		3.1070
2.5%		2.4465
0.5%		2.2165
0.0%	minimum	2.0390

Moments

Mean	4.6329873
Std Dev	1.0905695
Std Err Mean	0.0549421
upper 95% Mean	4.7410045
lower 95% Mean	4.5249701
N	394

The expression of antibodies is measured on a log scale in n=394 cells – at intervals for the expression do we look?

PI (prediction interval) or reference range, which covers 95% of the individual observations

$$\bar{x} \pm 2 \cdot sd(x)$$

Reference range: [102.8;909.6]

Confidence Interval (CI) for the expected value E(X) covers with 95% probability the true value E(X).

$$\bar{x} \pm 2 \cdot \frac{sd(x)}{\sqrt{n}} = \bar{x} \pm 2 \cdot se(\bar{x})$$

$$se(\bar{x}) = \frac{sd(x)}{\sqrt{n}}$$

standard error

CI: [92.1, 114.8]

To get to the original scale, we can exp-transform the limits of the intervals