Statistical distributions

Theory

Sarah Kaspar

Biostatistical Basics 2021

EMBL

EMBL

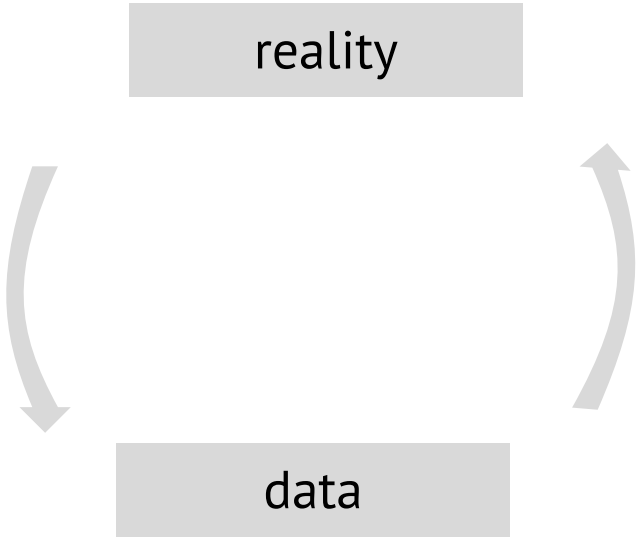
Goals for this lecture:
- know some common distributions of biological data
- be able to decide which distribution might fit your data

Content:
- what is a distribution
- common distributions
- tools for comparing distributions

Exercises:
- apply the tools on example data
- decide for a distribution

Sources:
- “Introduction to statistical thought” (Michael Lavine)
- “Modern Statistics for Modern Biologists” (Wolfgang Huber and Susan Holmes)

Motivation

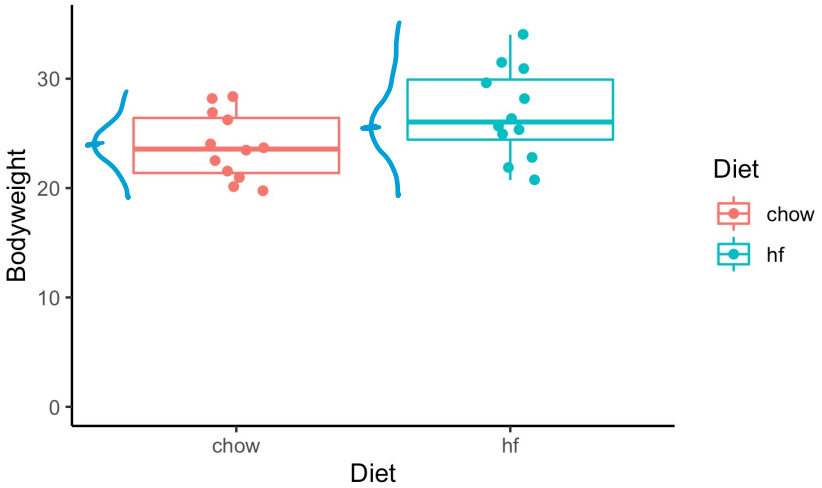


sampling /
experiment

inference:

- find **rules** that the data follow
- model them with a suitable **distribution**
- this helps understanding reality
- underlying distributions are important for **statistical tests**

Motivation



Question: Is there a difference in weight between mice with control vs. high-fat diet?

Problem: The difference in means could be by chance:

- we only have a sample of mice for each diet
- there is variation in the weights

Knowing the rules for randomness / variation will tell us how likely it is to see this difference by chance.

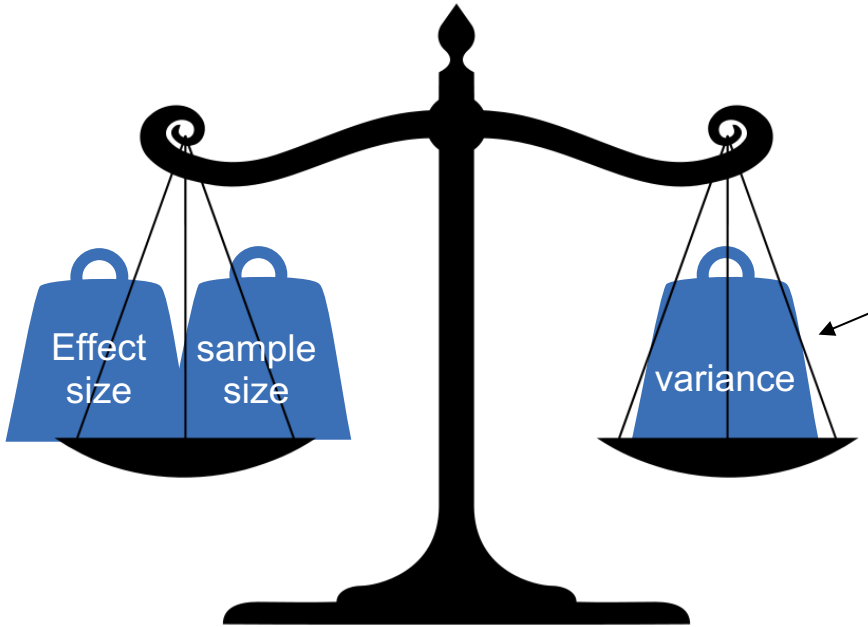
Statistical model:

$$weight = diet + residuals$$

group means — *follow a statistical distribution*

data from Söderhede and Ahrén (2004)

Motivation

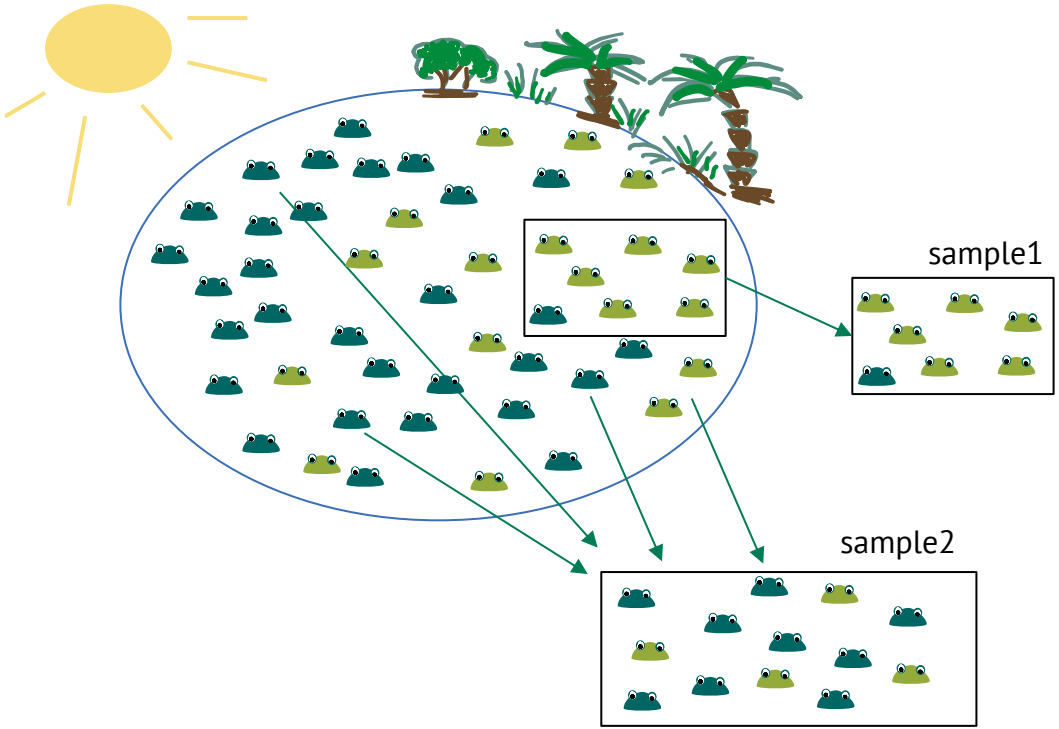


The interpretation of variance depends on the distribution

Image: https://de.m.wikipedia.org/wiki/Datei:Balanced_scale_of_Justice.svg

Sampling

Example: estimate the fraction of light green frogs



Sample: randomly and idenpendently drawn events from the population of interest

Population: the population/process you are interested in.

Randomness: You hope that these represent the population/reality well, but it is not guaranteed.

Independence: The observations don't depend on each other.

Sample size: Number of observations in your sample

Distribution: Set of rules that the random numbers follow.

Random doesn't mean there are no rules!

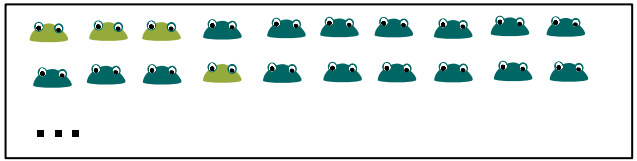
What is a probability distribution?

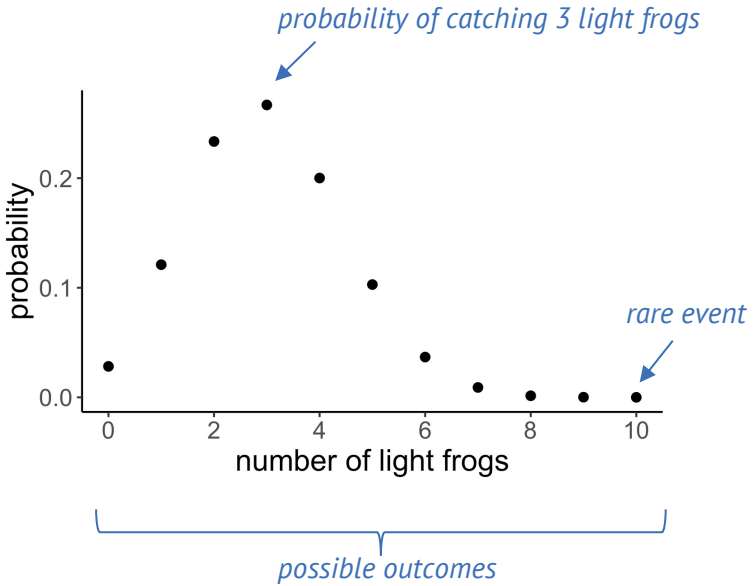
EMBL

→ It assigns probabilities to possible outcomes of an experiment.
→ Rules for randomness

Example:
Number of light frogs within 10 catches.
True fraction: 1/3

possible outcomes:

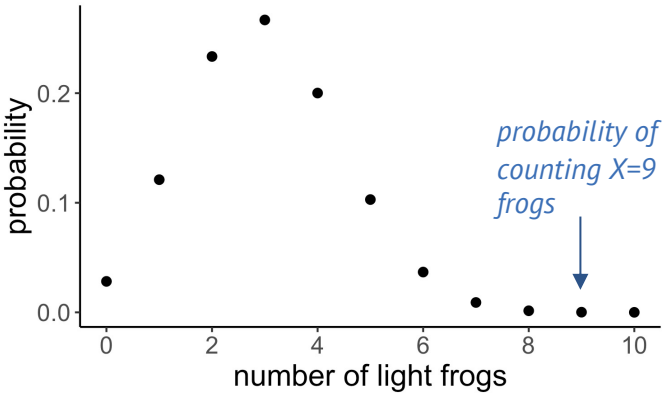




Two types of probability distribution

discrete case:
observations can take only integer values (e.g. counts)

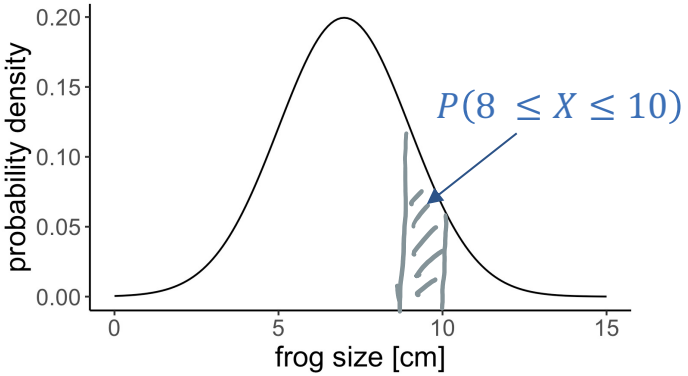
probability mass function



continuous case:

$P(X = 9) = 0$

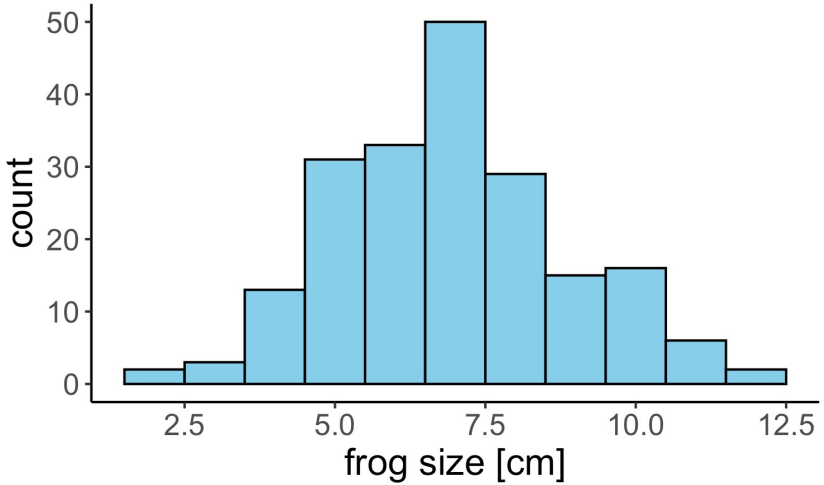
probability density function



Sampling distribution

EMBL

→ might follow the rules given by a theoretical distribution
→ but as you're sampling, it includes randomness



A few common distributions

Binomial distribution

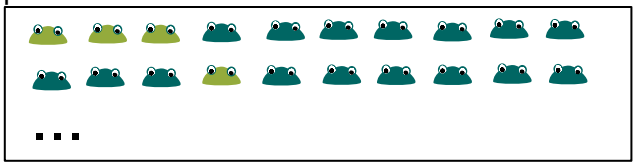
→ models the number of sucesses in a series of trials

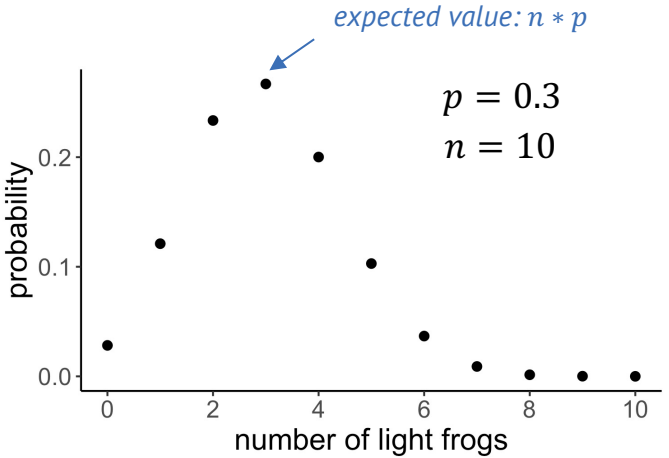
Examples:
- prevalence of a disease within a fixed number of patients
- counting mutations in a genome
- how many of the caught frogs are green?

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Parameters:
n: # caught frogs
p: probability of a caught frog being light green

possible outcomes:





Approximating the binomial distribution

EMBL

<u>Special case: large *n* and small *p*</u>
How many light frogs can we expect to catch within 1h?

- catch about 100 frogs per hour: $n = 100$
- The fraction of light frogs is low: $p = 0.02$

→ Expected number of light frogs per hour:

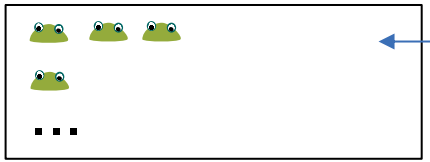
$$\lambda = n * p = 2$$

The number of light frogs caught within one hour can be approximated with a Poisson distribution with just one parameter λ.

fill the net for 1h

$n \approx 100$

possible outcomes:



we're not interested in the dark frogs, or how many frogs we caught in total!

Poisson

counting events over a fixed domain (period of time, space)
events have an underlying rate

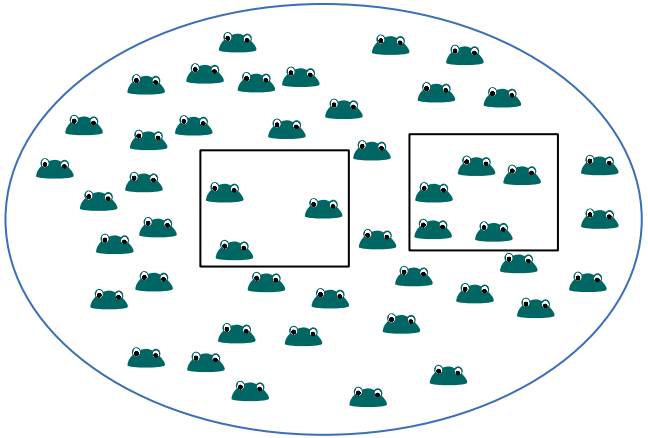
$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

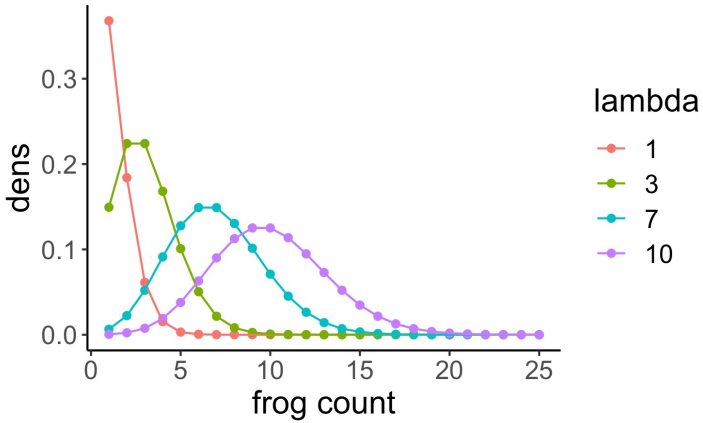
Examples:

- counted frogs over 1 hour
- counting cells in a fixed volume
- mutations in genome

Parameters:

- estimate for λ: sample mean
- variance = λ





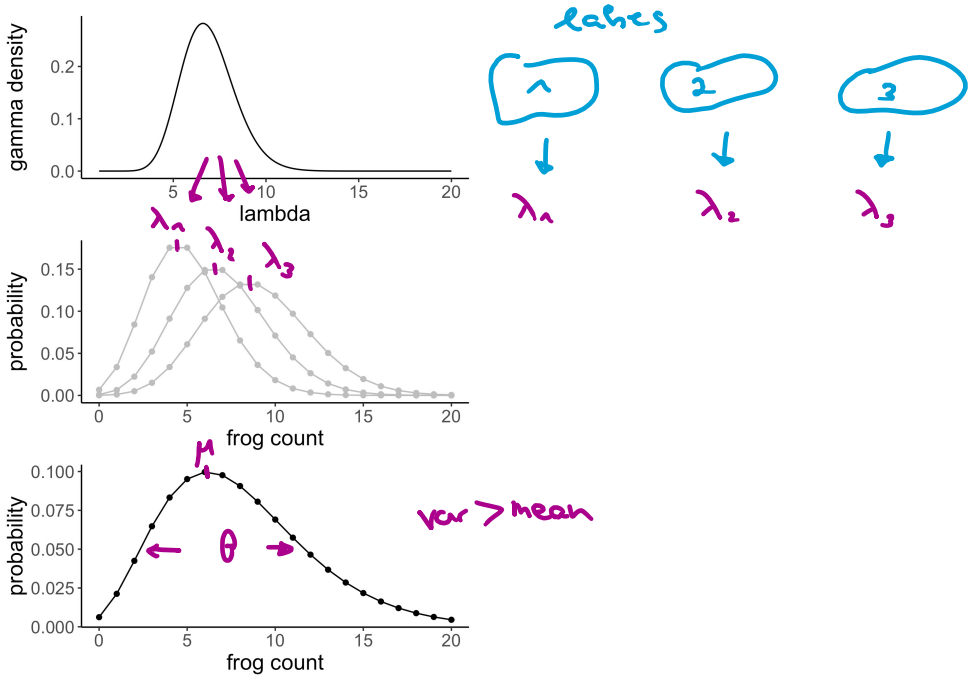
Gamma-Poisson

→ the overdispersed (= more spread out) version of Poisson

Examples:
- Counting frogs in different lakes
- read counts of a gene (difference between samples)
- cell counts in different volumes / individuals

Parameters:
mean: the average Poisson rate
scale: how much the lambdas spread

$$X \sim Pois(gamma(\mu, \theta))$$



Gaussian

EMBL

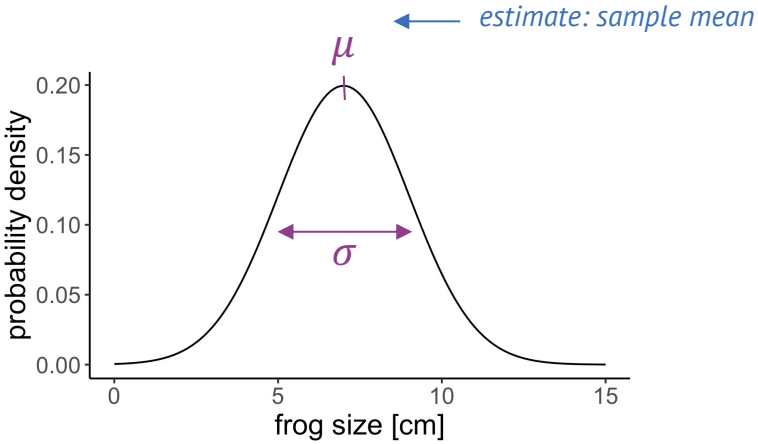
Examples:
- frog / cell sizes
- temperatures
- pixel intensities

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

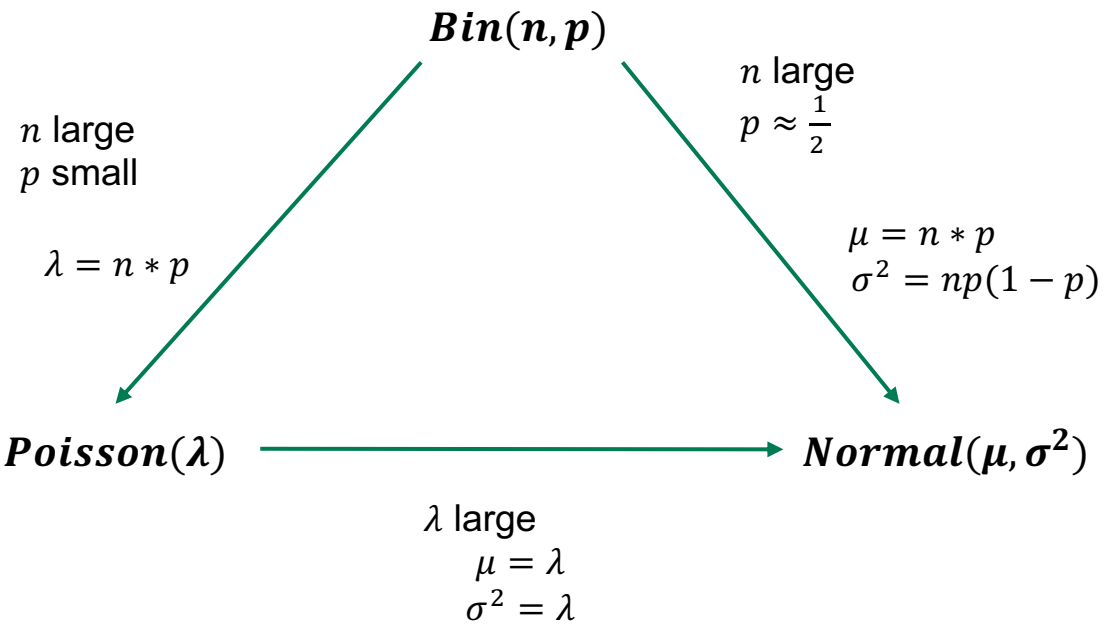
Parameters:

μ mean

σ variance



How are the distributions related?



Graphic adapted from
https://www.youtube.com/watch?v=u9onO78hDlw

What does ”fitting a distribution” mean?

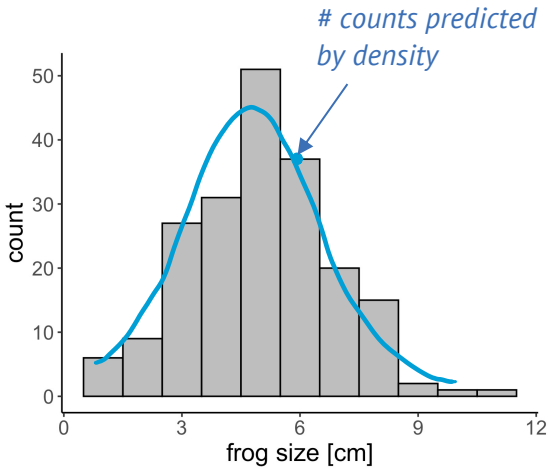
EMBL

- most common: maximum likelihood approach
- find the parameters for which it is most likely to see the given data (optimization problem)
- minimize the deviance (i.e. the distance between the “line” and the “points”)
- in Gaussian case: minimize sum of squares

Tools for comparing distributions

EMBL

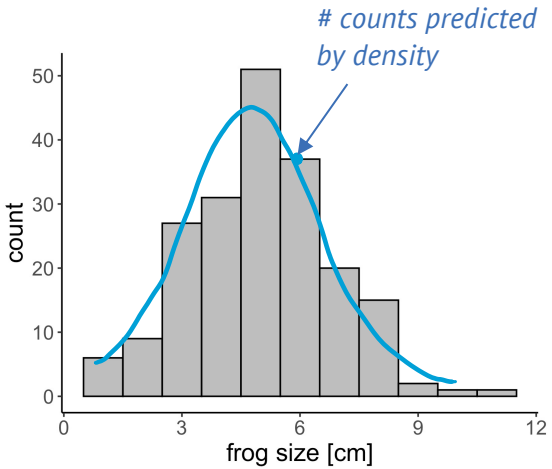
Histogram



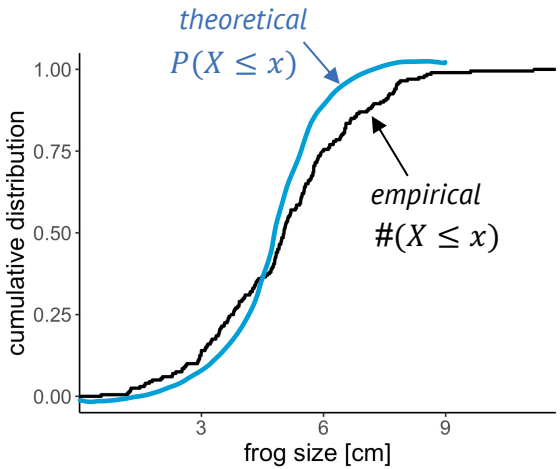
Tools for comparing distributions

EMBL

Histogram



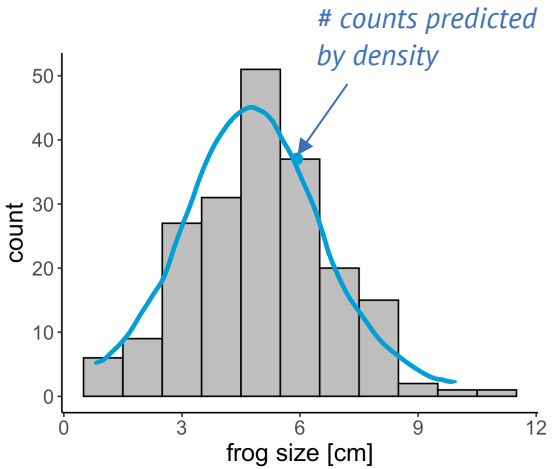
Cumulative distribution



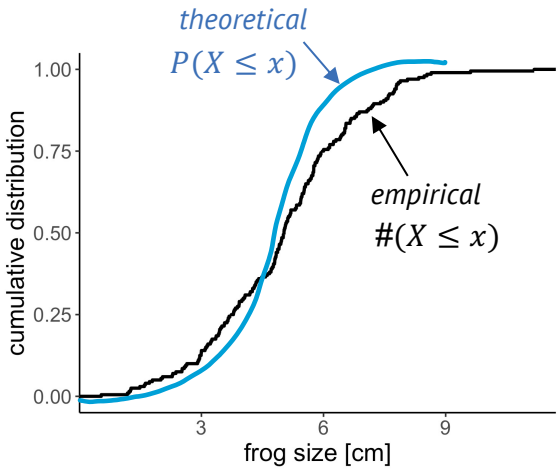
Tools for comparing distributions

EMBL

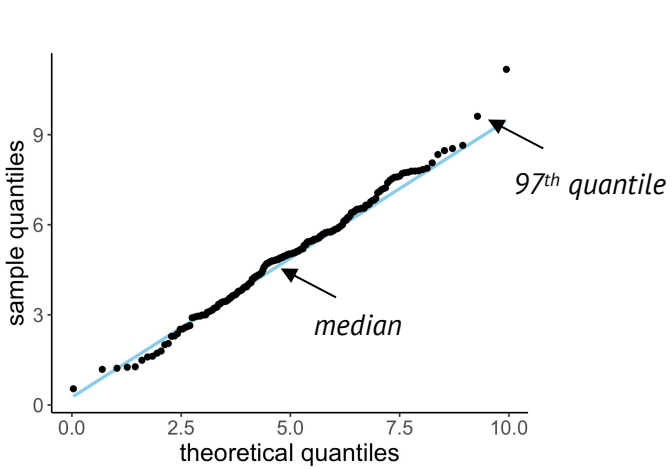
Histogram



Cumulative distribution



QQ-plot



The *25th* quantile is the value *k* at which *25%* of the data points are smaller than *k*.

How to find the right distribution for your data

EMBL

1. Fit your data to a distribution that you consider plausible
   → You get the best parameters for this distribution given your data

2. Visually compare the theory (=fitted distribution) to your data points
   → A good fit doesn't show systematic deviations of the data points from theory

3. Do the same with other plausible distributions

4. Decide which of the fits looks best to you (not always obvious)!

References

- Sörhede Winzell, M., & Ahrén, B. (2004). *The High-Fat Diet-Fed Mouse – A Model for Studying Mechanisms and Treatment of Impaired Glucose Tolerance and Type 2 Diabetes*.