Introduction to statistics

Computational biology week - London, April 2022

Jesús Urtasun Elizari, PhD

Outline

Chapter I
Probability in a nutshell

Probability in a nutshell

Discrete and continuous probability

- Discrete output $\longrightarrow$ probability well defined, compute by counting
- Continuous output $\longrightarrow$ need for probability distributions
- Definition of probability in both scenarios
- Computing probabilities in both scenarios

Probability in a nutshell

Discrete probability

Compute probability of getting 4 heads in a row

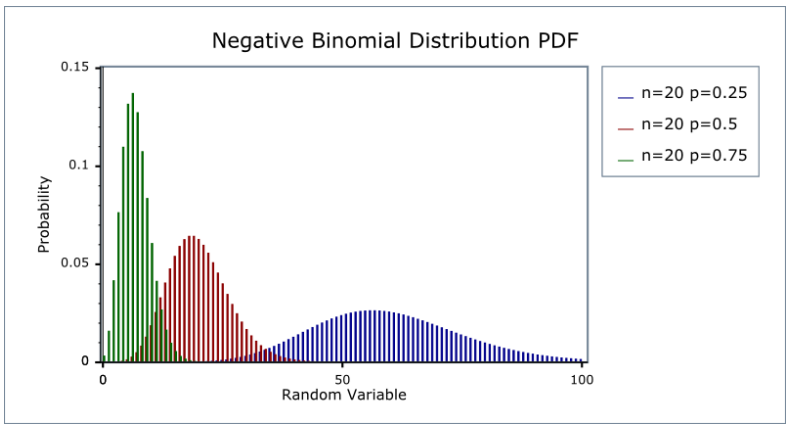
Probability in a nutshell

Discrete probability

Compute probability of getting a specific genotype

Probability in a nutshell

Binomial distribution



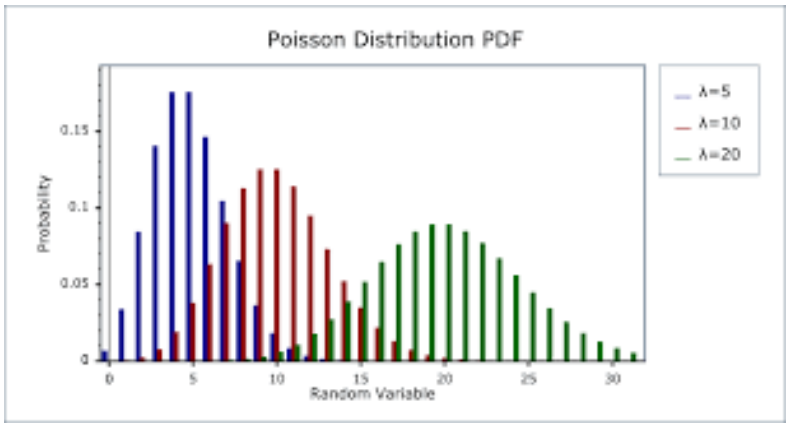
- Binomial distribution: probability of getting k events in n trials

$$B(x = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- Getting k heads in n coins
- Getting k successes randomly filling a test

Probability in a nutshell

Poisson distribution



- Poisson distribution: probability of counting k events in a given interval

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

- Counting k new cancer patients per day
- Counting k car accidents per year
- Counting k reads from RNAseq exp

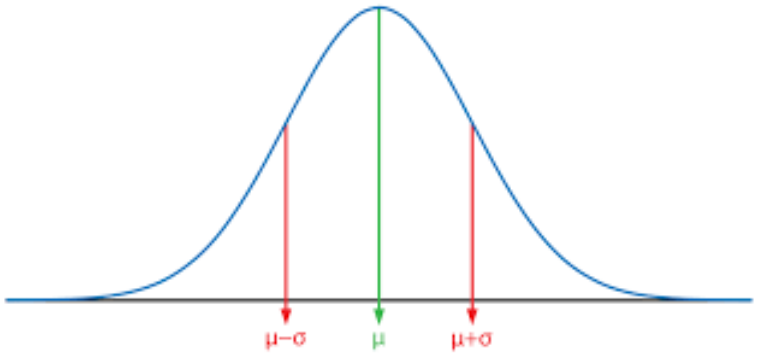
Probability in a nutshell

Continuous probability

Compute probability of measuring a particular height, weight (...)
Used in statistical analysis when (comparing means, errors, p-values, ...)

Probability in a nutshell

Gaussian distribution

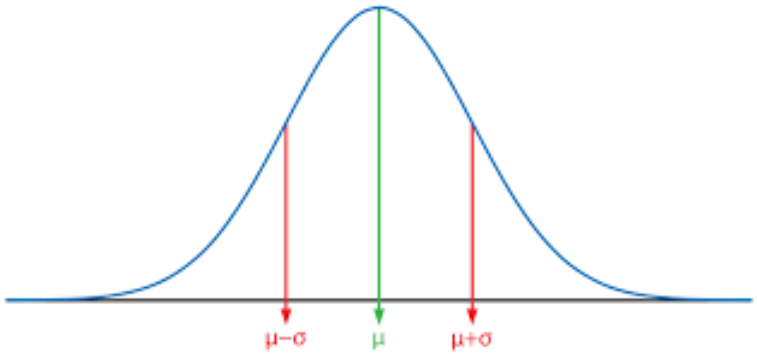


- Gaussian distribution: continuous distribution for a real-valued random variable

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

Probability in a nutshell

Mean and variance



- Momenta of a distribution
- Mean, variance, skewness, kurtosis

Probability in a nutshell

Standard Deviation vs Standard Error

Performing multiple sets of measurements and compare means

- Standard deviation (σ) quantifies variation over one set of measurements (second momentum of the probability distribution)
- Standard error (SE) quantifies variation of means from multiple sets!

$$\mathrm{SE} = \frac{\sigma}{\sqrt{N}}$$

- Source of confusion: SE can be estimated from single set of measurements, even though it describes variation of means

Probability in a nutshell

Exercises in R

https://lmsbioinformatics.github.io/LMS_StatisticsInR/course/CBW_StatisticsInR_course

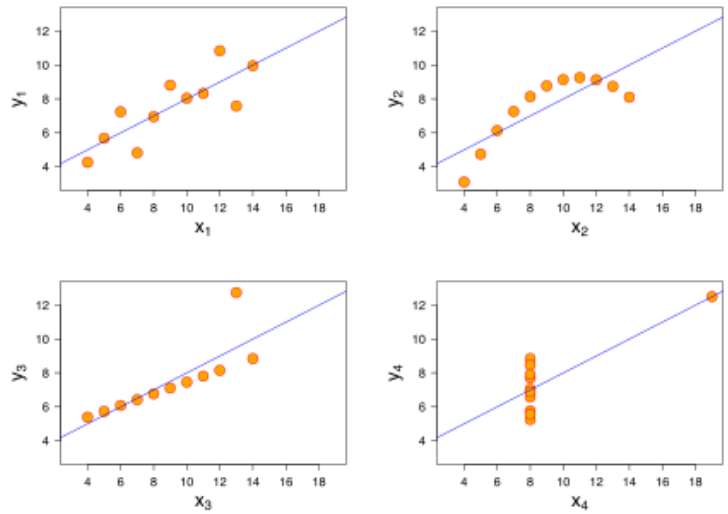
Chapter II
Linear models

Linear models

What is a linear model

1. Find a function that describes a set of observations

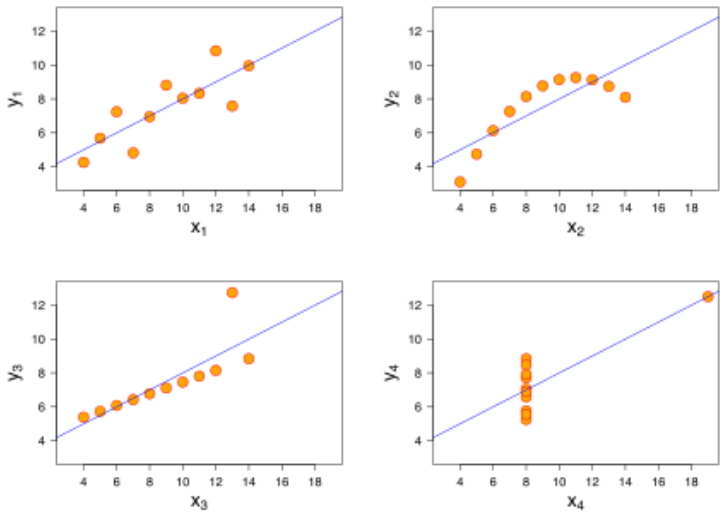
$$y(x) = b_0 \cdot x + b_1$$



Linear models

Fitting a linear model

1. Get the overall mean
2. Compute sum of residuals (SS) over the mean
3. Find tine with the smallest SS $\rightarrow$ fit
4. Evaluate fit - correlation coefficient R^2



Linear models

Generalized linear models

1. Fit count data (non-linearly distributed)
2. Exponential regression (problem with negative values)

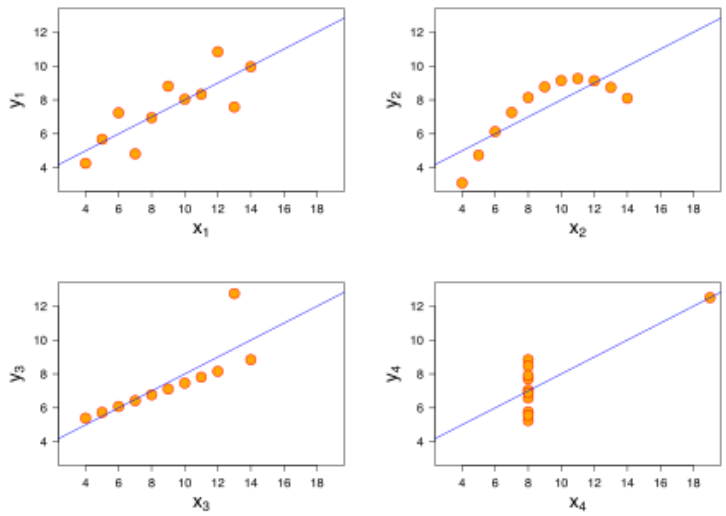
$$y(x) = e^{b_0 + b_1 x} = y_0 e^{b_1 x}$$

3. Poisson regression (assume variance $\sim$ mean)

$$y(x) = P(x)$$

4. Negative binomial

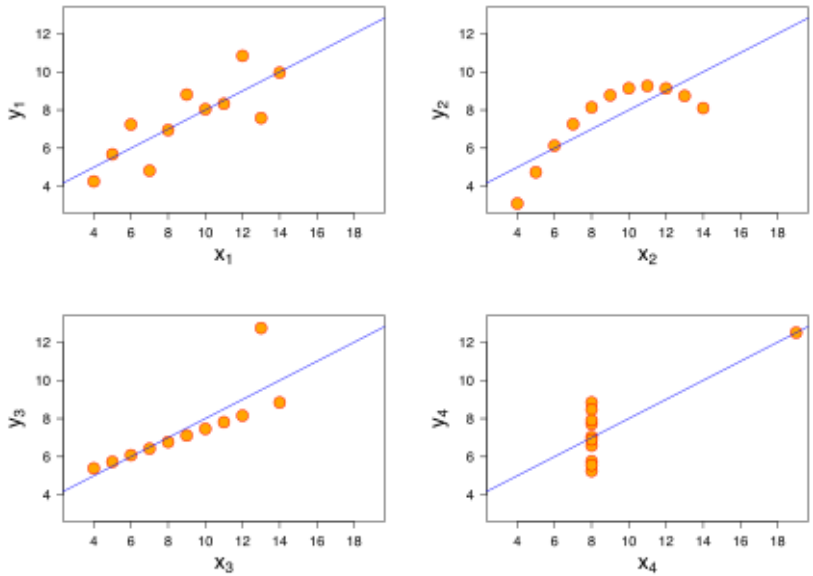
$$y(x) = NB(x)$$



Linear models

Advanced topics

1. ANOVA
2. (...)



Linear models

Exercises in R

https://lmsbioinformatics.github.io/LMS_StatisticsInR/course/CBW_StatisticsInR_course

Chapter III
Hypothesis testing

Hypothesis testing

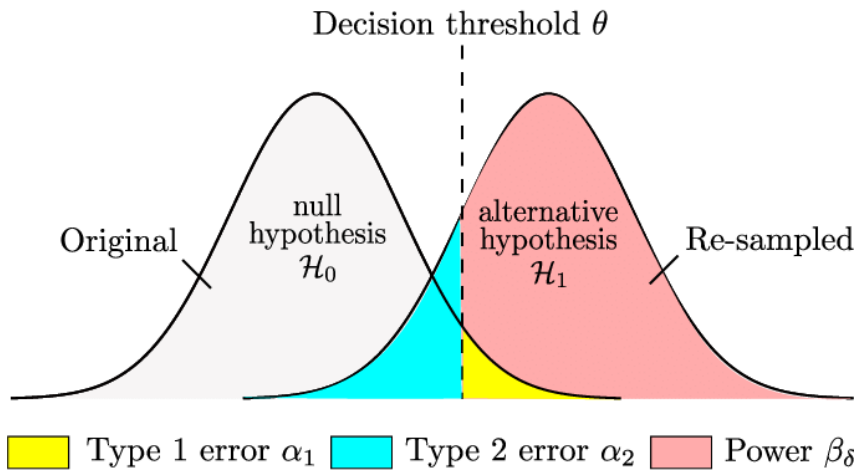
Introduction

- Null hypothesis and alternative hypothesis
- Statistic tests and p-values
- χ^2-test, t-test, Wald test
- Exercises in R

Hypothesis testing

Compare different models

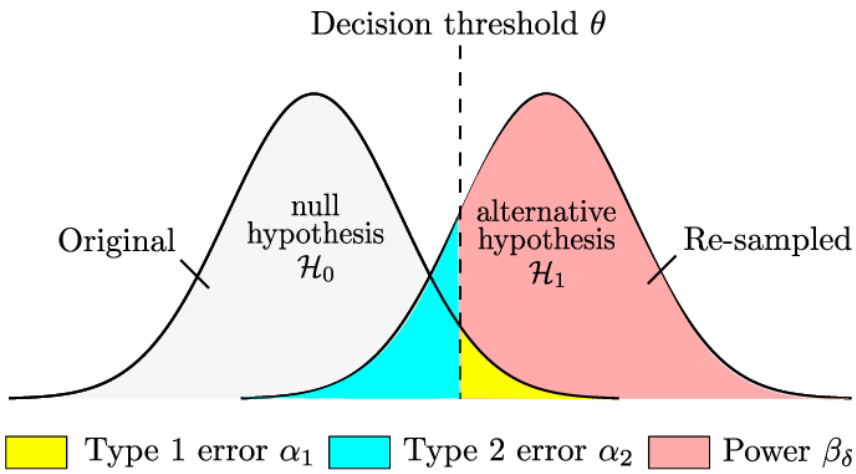
1. Null and alternative hypothesis



Hypothesis testing

Statistic test

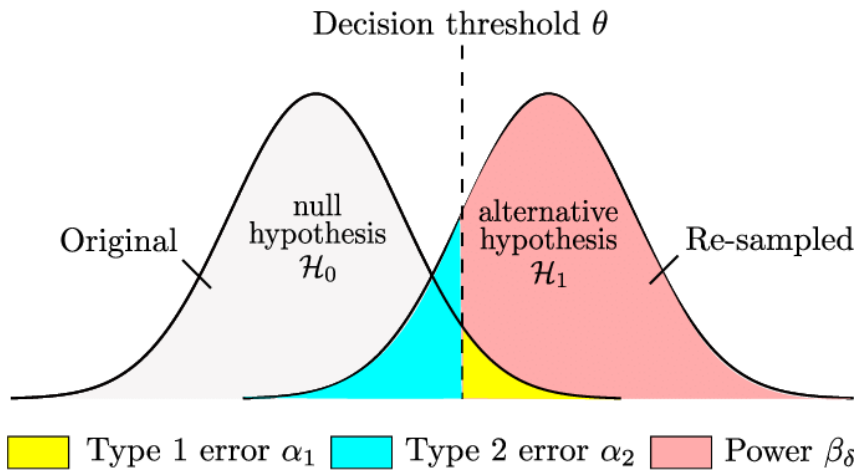
1. Quantify the significance of an observation
2. Certainty when accepting / rejecting a hypothesis



Hypothesis testing

p-values

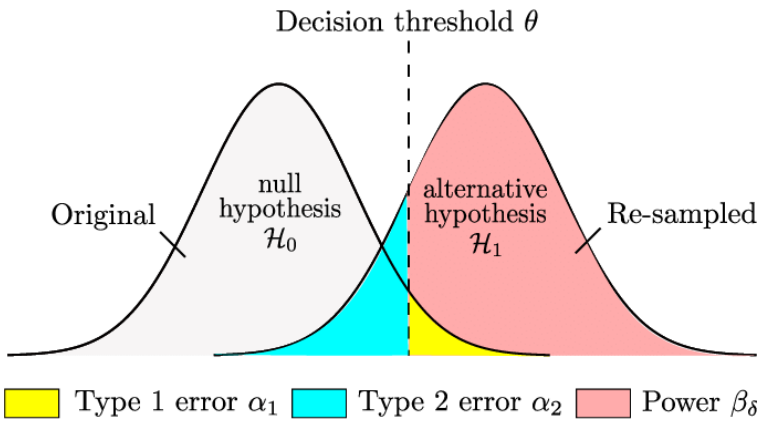
1. p-values



Hypothesis testing

Power calculations

1. Do not just add samples until getting a good p-value $\rightarrow$ increases chance of reporting a false positive
2. "Power": probability that a test will correctly give a small p-value
3. Factors that affect power
    1. Effect size
    2. Variation in data
    3. Sample size
    4. Test used



Hypothesis testing

Exercises in R

https://lmsbioinformatics.github.io/LMS_StatisticsInR/course/CBW_StatisticsInR_course