

# **Biostatistics**

## **Applications in Medicine**

**Nuno Sepúlveda, 25.11.2024**

# Syllabus

## 1. General review

- a. What is Biostatistics?
- b. Population/Sample/Sample size
- c. Type of Data – quantitative and qualitative variables
- d. Common probability distributions
- e. Work example – Malaria in Tanzania

## 2. Applications in Medicine

- a. Construction and analysis of diagnostic tools – Binomial distribution, sensitivity, specificity, ROC curve, Rogal-Gladen estimator
- b. Estimation of treatment effects - generalized linear models
- c. Survival analysis - Weibull regression, Kaplan-Meier curve, log-rank test, Cox's proportional hazards model

## 3. Applications in Genetics, Genomics, and other 'omics data

- a. Genetic association studies – Hardy-Weinberg test, homozygosity, minor allele frequencies, additive model, multiple testing correction
- b. Methylation association studies – M versus beta values, estimation of biological age
- c. Gene expression studies based on RNA-seq experiments – Tests based on Poisson and Negative-Binomial

## 4. Other Topics

- a. Estimation of Species diversity – Diversity indexes, Poisson mixture models
- b. Serological analysis – Gaussian (skew-normal) mixture models
- c. Advanced sample size and power calculations

**Parametric analysis**

**versus**

**Non-parametric analysis**

# Parametric analysis

**Complete  
data**



**Incomplete  
data**

# Non-parametric analysis



# Non-parametric methods

## Comparison of different survival curves

Log-rank test  
Peto-Peto test

## Semi-parametric regression

Cox's proportional hazard model

# Comparison of different survival curves

Two treatments under comparison

Time to clinical response

$$H_0 : S_1(t) = S_2(t) \text{ versus } H_0 : S_1(t) \neq S_2(t)$$

Log-rank test as a Mantel-Haenszel test for categorical data

# Mantel-Haenszel test

Analysis of the association in  $K \times 2 \times 2$  contingency tables (an extension of Fisher's exact test to  $K$  tables  $2 \times 2$ ).

Stratum	Treatment	Responded	Not Responded
1	A		
	B		
2	A		
	B		
3	A		
	B		

In stratum  $i$

$$\Delta_i = \frac{\pi_{1i}(1 - \pi_{2i})}{(1 - \pi_{1i})\pi_{2i}}$$

$\pi_{1i}$  = prob. of response to treatment 1

$\pi_{2i}$  = prob. of response to treatment 2

$H_0 : \Delta_1 = \dots = \Delta_K = 1$  (t) versus  $H_1 : \exists_{i,j} \Delta_i \neq \Delta_j = 1$

under the assumption of  $\Delta_1 = \dots = \Delta_K = \Delta$



# Log-rank test

Adaptation of the classical Mantel-Haenszel test for  $k \times 2 \times 2$  contingency tables where  $k$  is the number of different timepoints in which it was observed the event of interest



## Basic idea

There are  $k$  2 x 2 tables like this one

Group	Number of “deaths” at $t_{(i)}$	Number of “survivors” beyond $t_{(i)}$	Total
1	$d_{1i}$	$n_{1i} - d_{1i}$	$n_{1i}$
2	$d_{2i}$	$n_{2i} - d_{2i}$	$n_{2i}$
Total	$d_i$	$n_i - d_i$	$n_i$

## Conditional probability (see Fisher's exact test)

$$H_0 : S_1(t) = S_2(t) \text{ versus } H_1 : S_1(t) \neq S_2(t)$$

$$H_0 : \pi_{1i} = \pi_{2i} = \pi \text{ versus } H_1 : \pi_{1i} \neq \pi_{2i}$$

$\pi_{1i}$  = probability of "death" at time  $t_{(i)}$  in group 1

$\pi_{2i}$  = probability of "death" at time  $t_{(i)}$  in group 2

$$d_{li} \mid \pi_{li}, n_{li} \rightsquigarrow \text{Binomial}(n = n_{li}, \pi = \pi_{li}), l = 1, 2$$

$$d_i \mid \pi_{li}, n_{li}, H_0 \rightsquigarrow \text{Binomial}(n = n_i, \pi = \pi_i)$$

## Basic idea

Calculate the distribution of  $d_{1i}$  conditional to the total marginals

Group	Number of “deaths” at $t_{(i)}$	Number of “survivors” beyond $t_{(i)}$	Total
1	$d_{1i}$	$n_{1i} - d_{1i}$	$n_{1i}$
2	$d_{2i}$	$n_{2i} - d_{2i}$	$n_{2i}$
Total	$d_i$	$n_i - d_i$	$n_i$

## Conditional probability (see Fisher's exact test)

$$d_{1i} | d_i, n_i, n_{1i}, H_0 \rightsquigarrow \text{Hypergeometric}(N = n_i, M = d_i, n = n_{1i})$$

$$P [d_{1i} = d | d_i, n_i, n_{1i}, H_0] = \frac{\binom{d_i}{d} \binom{n_i - d_i}{n_{1i} - d}}{\binom{n_i}{n_{1i}}}$$

$$E [d_{1i} | d_i, n_i, n_{1i}, H_0] = n_{1i} \frac{d_i}{n_i} \qquad \text{Var} [d_{1i} | d_i, n_i, n_{1i}, H_0] = n_{1i} \frac{d_i}{n_i} \left(1 - \frac{d_i}{n_i}\right) \frac{n_i - n_{1i}}{n_i - 1}$$

## Test statistic

Incorporating information from  $k$  2 x 2 contingency tables

$$U = \sum_{i=1}^k (d_{1i} - e_{1i})$$

$$e_{1i} = E [d_{1i} | d_i, n_i, n_{1i}, H_0] = n_{1i} \frac{d_i}{n_i}$$

$$E [U | H_0] = 0$$

$$v_{1i} = Var [d_{1i} | d_i, n_i, n_{1i}, H_0]$$

$$Var [U | H_0] = \sum_{i=1}^k v_{1i}$$

$$= n_{1i} \frac{d_i}{n_i} \left( 1 - \frac{d_i}{n_i} \right) \frac{n_i - n_{1i}}{n_i - 1}$$

# Log-rank test

For large samples

$$Q = \frac{U - \overbrace{E(U)}^{=0}}{\sqrt{\text{var}(U)}} \mid H_0 \rightsquigarrow \text{Normal}(\mu = 0, \sigma = 1)$$

$$Q^* = \frac{U^2}{\text{var}(U)} \mid H_0 \rightsquigarrow \chi^2_{(1)}$$

Decision rule

$$p = P [Q^* > q_{obs} \mid H_0]$$

$\begin{cases} \text{do not reject } H_0, & \text{if } p > \alpha \\ \text{reject } H_0, & \text{otherwise} \end{cases}$

## A general class of non-parametric tests

$$Q^* = \frac{\left[ \sum_{i=1}^k w_i (d_{1i} - e_{1i}) \right]^2}{\sum_{i=1}^k w_i^2 v_{1i}} \mid H_0 \rightsquigarrow \chi_{(1)}^2 \quad \text{for large samples}$$

### Choices of the weights

$w_i = 1$ , log-rank test

$w_i = \sqrt{n_i}$ , Tarone-Ware test

$w_i = n_i$ , Gehan test

$w_i = \prod_{j:t(j) \leq t(i)} \left( 1 - \frac{d_j}{n_j + 1} \right)$ , Peto-Peto test



## A general statistic for non-parametric tests

$$Q^* = \frac{\left[ \sum_{i=1}^k w_i (d_{1i} - e_{1i}) \right]^2}{\sum_{i=1}^k w_i^2 v_{1i}} \mid H_0 \rightsquigarrow \chi_{(1)}^2 \quad \text{for large samples}$$

### Choices of the weights

$w_i = 1$ , log-rank test

$w_i = \sqrt{n_i}$ , Tarone-Ware test

$w_i = n_i$ , Gehan test

$w_i = \prod_{j:t(j) \leq t(i)} \left( 1 - \frac{d_j}{n_j + 1} \right)$ , Peto-Peto test

## Harrington-Fleming class of tests (Survival package)

$$Q^* = \frac{\left[ \sum_{i=1}^k w_i (d_{1i} - e_{1i}) \right]^2}{\sum_{i=1}^k w_i^2 v_{1i}} \mid H_0 \rightsquigarrow \chi_{(1)}^2 \quad \text{for large samples}$$

$$w_i = \hat{S}_{t(i)}^\rho (1 - \hat{S}_{t(i)})^\gamma,$$

$\hat{S}_{t(i)}$  = Kaplan-Meier estimates for the common survival function

$$\rho = 0 \Rightarrow w_i = 1 \text{ - log-rank test}$$

$$\gamma = 0 \Rightarrow w_i = \hat{S}_{t(i)}^\rho$$

$$\rho = 1 \Rightarrow w_i = \hat{S}_{t(i)} \text{ - Peto-Peto test}$$

## Exercise 1: rituximab clinical trial data

survival package (analysis)

survminer package (plotting)

Plot survival curves (surfit command) of time to treatment response for:

- (i) males versus females
- (ii) patients with and without an infection disease trigger
- (iii) patients with and without family history of autoimmune diseases

Compared with the curves for each case using log-rank and Peto-Peto tests  
(survdifff function from survival package)

Draw your conclusions.

# Non-parametric methods

## Comparison of different survival curves

Log-rank test  
Peto-Peto test

## Semi-parametric regression

Cox's proportional hazard model

# Cox's proportional hazard model

$$h_{x_{ij}}(t) = h_0(t) e^{\sum_{j=1}^p \beta_j x_{ij}}$$

“All models are wrong, some are useful.”

George Box (1976)



John Wiley

<https://rss.onlinelibrary.wiley.com/doi/j.2517-6161.1...>

## Regression Models and Life-Tables - Cox - 1972

by DR Cox · 1972 · Cited by 60930 — **Cox, D. R.** (1959). The **analysis** of exponentially distributed **life-times** with two types of failure. **J. R. Statist. Soc. B**, 21, 411–421. **Cox, D. R.**....

## Cox's proportional hazard model

$$\log \frac{h_{x_{ij}}(t)}{h_0(t)} = \sum_{j=1}^p \beta_j x_{ij}$$

Let be two individuals  $i$  and  $k$  with covariates  $\{x_{ij}\}$  and  $\{x_{kj}\}$

$$\frac{h_{x_{ij}}(t)}{h_{x_{kj}}(t)} = e^{\sum_{j=1}^p \beta_j (x_{ij} - x_{kj})}$$

## Interpretation of the coefficients

Let be two individuals  $i$  and  $k$  with covariates  $\{x_{ij}\}$  and  $\{x_{kj}\}$

$\{x_{ij}\}$  and  $\{x_{kj}\}$  are only different at  $x_{1k}$  and  $x_{2k}$  in one unit

$$\frac{h_{x_{ij}}(t)}{h_{x_{kj}}(t)} = e^{\beta_j}$$

Relative risk when one changes one unit in covariate  $k$  while maintaining the remaining covariates fixed

## Estimation

$$t_{(1)} < \dots < t_{(k)}, k < n$$

$$R_i = R(t_{(i)}) = \left\{ j : t_j \geq t_{(i)} \right\}$$

$$D = \left\{ i : t_{(i)} \right\}$$

Maximisation of the following function

$$L(\beta_1, \dots, \beta_p) = \prod_{i \in D} \frac{e^{\sum_{j=1}^p \beta_j x_{(i)j}}}{\sum_{l \in R_i} e^{\sum_{j=1}^p \beta_j x_{lj}}}$$

(numerical methods)



## A theoretical note

$$L(\beta_1, \dots, \beta_p) = \prod_{i \in D} \frac{e^{\sum_{j=1}^p \beta_j x_{(i)j}}}{\sum_{l \in R_i} e^{\sum_{j=1}^p \beta_j x_{lj}}}$$

Partial likelihood (because it is independent of the baseline hazard function)

This is not a likelihood function in the strict sense as it does not represent the probability of a given event.

## A theoretical note

True likelihood

$$h(t) = \frac{f(t)}{S(t)} \Leftrightarrow f(t) = S(t) \times h(t)$$

$$L(\beta_1, \dots, \beta_p, h_0(t)) = \prod_i (h_0(t) e^{\sum_j \beta_j x_{ij}} S_0(t))$$

$$L(\beta_1, \dots, \beta_p; h_0(t)) = L(\beta_1, \dots, \beta_p) \times \prod_{i \in D} \left( h_0(t) \sum_{l \in R_i} e^{\sum_j \beta_j x_{lj}} \right) \times \prod_{i=1}^n S_0(t)^{\exp\left\{ \sum_j x_{ij} \right\}}$$

## Exercise 2: rituximab clinical trial data

Fit a Cox's proportional hazard model including the following covariates

- (i) gender
- (ii) age
- (iii) history of autoimmune diseases
- (iv) disease duration

Draw your conclusions.

## Model selection and comparison

$$M_1 \subset M_2$$

$$M_1 : \log \frac{h_{x_{ij}}(t)}{h_0(t)} = \beta_1 x_{i1} + \cdots + \beta_1 x_{ip}$$

$$M_2 : \log \frac{h_{x_{ij}}(t)}{h_0(t)} = \beta_1 x_{i1} + \cdots + \beta_1 x_{ip} + \beta_1 x_{i,p+1} + \cdots + \beta_1 x_{i,p+m}$$

Are AIC or BIC applicable?

# Analysis of Residuals

## Cox-Snel Residual

$$r_i = e^{\hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}} \hat{H}_0(t)$$

$$r_i \rightsquigarrow \text{Exponential}(1)$$

$\hat{H}_0(t)$  is the estimated cumulative baseline hazard

$$H(t) = \int_0^t h(u) du$$

$$\hat{H}_0(t) = -\log \hat{S}_0(t)$$

$$H(t) = -\log S(t)$$

# Syllabus

## 1. General review

- a. What is Biostatistics?
- b. Population/Sample/Sample size
- c. Type of Data – quantitative and qualitative variables
- d. Common probability distributions
- e. Work example – Malaria in Tanzania

## 2. Applications in Medicine

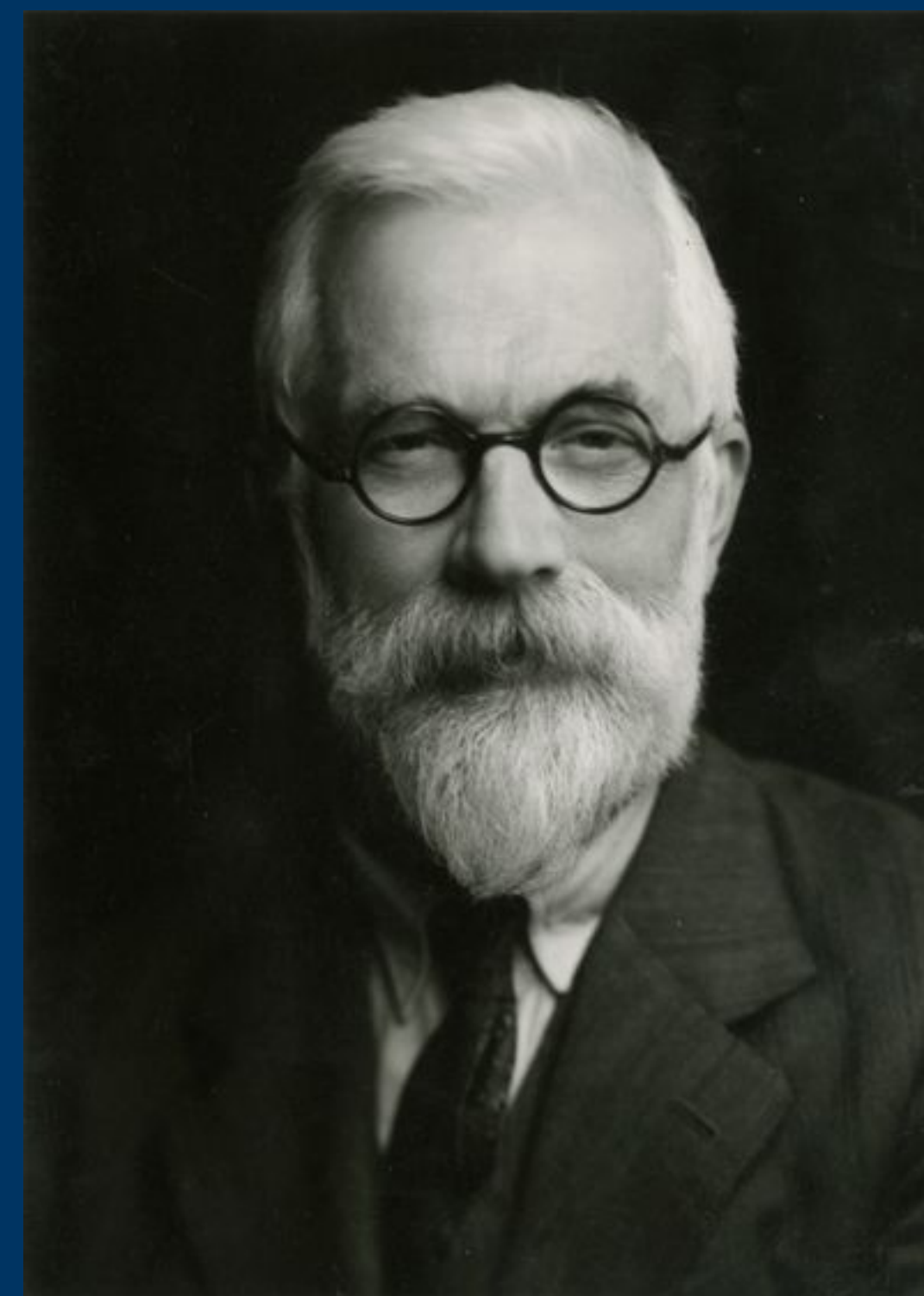
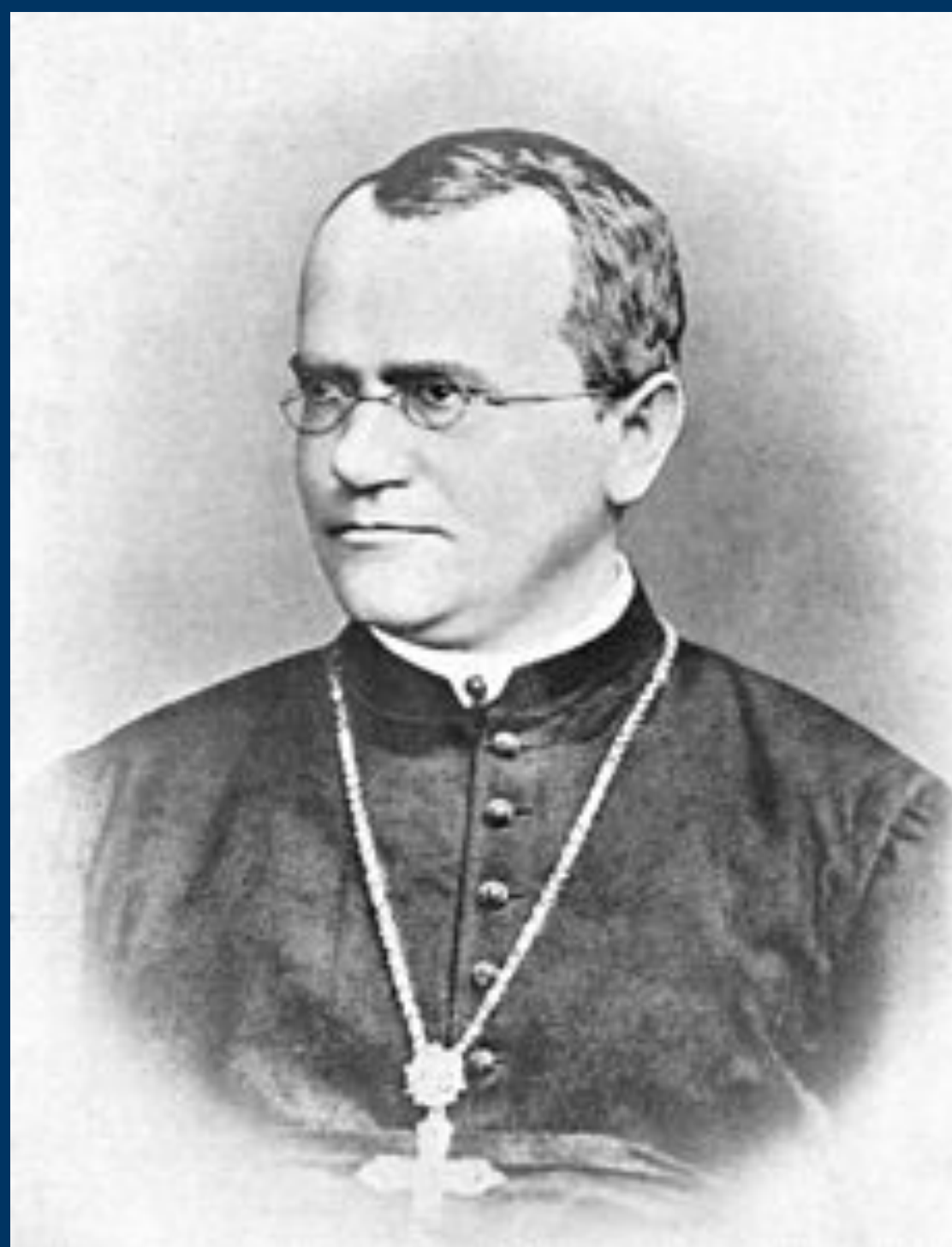
- a. Construction and analysis of diagnostic tools – Binomial distribution, sensitivity, specificity, ROC curve, Rogal-Gladen estimator
- b. Estimation of treatment effects - generalized linear models
- c. Survival analysis - Weibull regression, Kaplan-Meier curve, log-rank test, Cox's proportional hazards model

## 3. Applications in Genetics, Genomics, and other 'omics data

- a. Genetic association studies – Hardy-Weinberg test, homozygosity, minor allele frequencies, additive model, multiple testing correction
- b. Methylation association studies – M versus beta values, estimation of biological age
- c. Gene expression studies based on RNA-seq experiments – Tests based on Poisson and Negative-Binomial

## 4. Other Topics

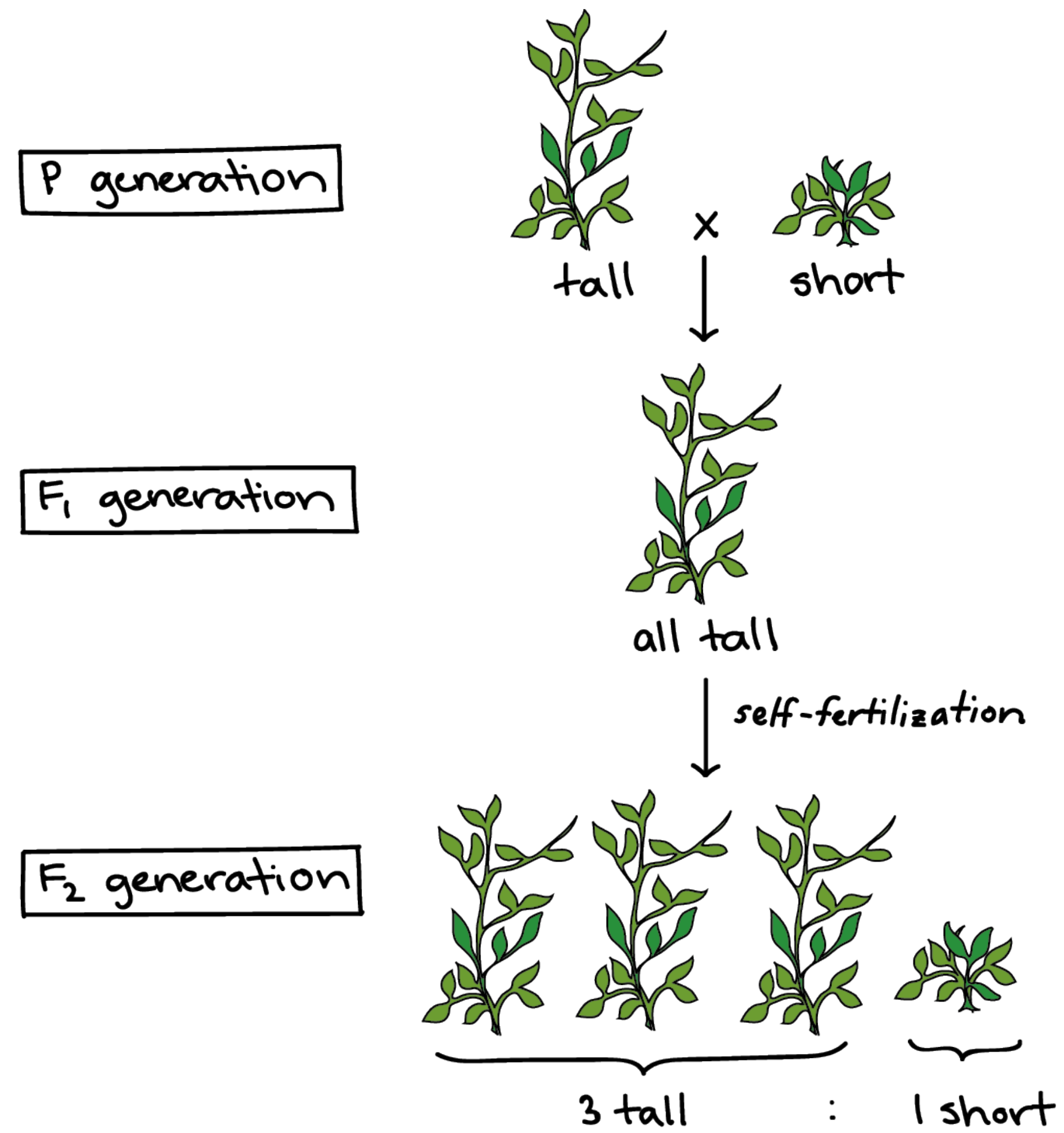
- a. Estimation of Species diversity – Diversity indexes, Poisson mixture models
- b. Serological analysis – Gaussian (skew-normal) mixture models
- c. Advanced sample size and power calculations



Do you know these people?



# Mendelian genetics





# Mendelian genetics

Phenotype /Trait = Biological Characteristic Under study (categorical)

Gene = Unit of Inheritance

Genotype = Composition of gene in terms of alleles

Allele = Variant of a gene

AA

# Mendel's idea/interpretation

Generation F0

Phenotype A x Phenotype a



Generation F1

100% Phenotype A

F1 x F1



75%

Phenotype A



25%

Phenotype A

Generation F2

AA x aa



Aa

Aa x Aa



AA



Aa or aA



aa

# First two Mendel's laws

## **The law of Dominance and Uniformity**

Some alleles are dominant over the other alleles for a given gene

## **The law of Segregation**

Two alleles for each gene separate from each other during gametogenesis so that the parent may only pass off one allele; thus, the offspring can only inherit one allele from each parent

# Exercise 1: data\_mendel\_single\_trait.csv

TABLE 1  
*Data given in Mendel (1866) for the single trait experiments. “A” (“a”) denotes the dominant (recessive) phenotype; A (a) denotes the dominant (recessive) allele; n is the total number of observations per experiment (that is, seeds for the seed trait experiments and plants otherwise);  $n_{“A”}$ ,  $n_{“a”}$ ,  $n_{Aa}$  and  $n_{AA}$  denote observed frequencies*

	Trait	“A”	“a”	n	Obs. freq.		Theor. ratio
					$n_{“A”}$	$n_{“a”}$	“A” : “a”
$F_2$	Seed shape	round	wrinkled	7324	5474	1850	3 : 1
	Seed color	yellow	green	8023	6022	2001	3 : 1
	Flower color	purple	white	929	705	224	3 : 1
	Pod shape	inflated	constricted	1181	882	299	3 : 1
	Pod color	yellow	green	580	428	152	3 : 1
	Flower position	axial	terminal	858	651	207	3 : 1
	Stem length	long	short	1064	787	277	3 : 1

Test Mendel’s predictions for each trait using an appropriate statistical test.  
Draw your conclusions.

# Third Mendel's law

## **The law of Independent Assortment (law of reassortment)**

Alleles of different genes segregate independently of one another during gametogenesis

# Bifactorial experiments

Generation F0

Phenotypes A/B x Phenotype a/b



Generation F1

100% Phenotypes A/B

F1 x F1



9:16 Phenotype  
A/B

3:16 Phenotype  
A/b

3:16 Phenotype  
a/B

1:16 Phenotype  
a/b

# Bifactorial experiments

## Combined genotypes

<b>x</b>	<b>BB</b>	<b>Bb</b>	<b>bb</b>
<b>AA</b>	AA/BB	AA/Bb	AA/bb
<b>Aa</b>	Aa/BB	Aa/Bb	Aa/bb
<b>aa</b>	aa/BB	aa/Bb	aa/bb

# Bifactorial experiments

Possibilities (n=16)

Cross	BB	Bb	bb
AA	1	2	1
Aa	2	4	2
aa	1	2	1



# Bifactorial experiments

Possibilities (n=16)

Cross	BB	Bb	bb
AA	1 Phenotype A/B	2	1 Phenotype A/b
Aa	2	4	2
aa	1 Phenotype a/B	2	1 Phenotype a/b

# Bifactorial experiments

Possibilities (n=16)

Cross	BB	Bb	bb
AA	1 Phenotypes A/B	2	1 Phenotypes A/b
Aa	2	4	2
aa	1 Phenotypes a/B	2	1 Phenotypes a/b

## Exercise 2:

TABLE 2 <i>Data from the bifactorial experiment [as organized by Fisher (1936)]</i>				
	<i>AA</i>	<i>Aa</i>	<i>aa</i>	<b>Total</b>
<i>BB</i>	38	60	28	126
<i>Bb</i>	65	138	68	271
<i>bb</i>	35	67	30	132
Total	138	265	126	529

Test the third Mendel's law predictions for each trait using an appropriate statistical test.

Draw your conclusions.