GWAS model extensions

- categorical and longitudinal data -

Christian Werner

(Quantitative geneticist and biostatistician) **EiB, CIMMYT**, Texcoco (Mexico)

Filippo Biscarini HerrFalloppio

(Biostatistician, bioinformatician and quantitative geneticist) **CNR-IBBA**, Milan (Italy)

Oscar González-Recio OscarGenomics

(Computational biologist and quantitative geneticist) **INIA-UPM**, Madrid (Spain)

A primer on GWAS for categorical traits

categorical traits

1. **Unordered** (nominal) categorical traits
    - → `multinomial logistic regression / softmax regression`
2. **Ordered** (ordinal) categorical traits
    - → `ordered logistic regression`

unordered categorical traits

- breeds
- shapes and colors (e.g. fruit, flowers, eyes, coats)
- blood type

BLACK
BAY
RED
GREY
SILVER BLACK
SILVER BAY
FLEXAN
DAPPLE GREY
BLUE DUN
BAY DUN
RED DUN
PALOMINO
BLUE ROAN
BAY ROAN
RED ROAN
PINTO
SEAL BROWN
BLANKET
LEOPARD
FLEA BITTEN
VectorStock®
VectorStock.com/27785634

From: https://www.vectorstock.com/royalty-free-vector/common-horse-coat-colors-vector-27785634

unordered categorical traits

- breeds
- shapes and colors (e.g. fruit, flowers, eyes, coats)
- blood type
- type of diet
- etc.

From: https://treatsforchickens.com/blogs/treats-for-chickens-blog/the-ultimate-guide-to-chicken-feed-for-laying-hens-types-of-chicken-food-101

unordered categorical traits

- breeds
- shapes and colors (e.g. fruit, flowers, eyes, coats)
- blood type
- type of diet
- etc.

- no high/low
- no better/worse
- etc.

From: https://treatsforchickens.com/blogs/treats-for-chickens-blog/the-ultimate-guide-to-chicken-feed-for-laying-hens-types-of-chicken-food-101

multinomial logistic regression

- binary logistic regression is used when the dependent variable is categorical (nominal/ordinal) and has two classes (e.g. cases/controls).
- when there are more than two nominal (unordered) classes for the categorical dependent variable, the model can be extended to **multinomial logistic regression**

multinomial logistic regression

- binary logistic regression:

$$log\left(\frac{Pr(y=1|x)}{1-Pr(y=1|x)}\right) = \beta_0 + \beta_1 x$$

[to recap]

multinomial logistic regression

- **multinomial logistic regression**: the analysis breaks down into a series of comparisons between two categories (e.g. if you have three outcome categories (A, B and C), then the analysis will consist of two comparisons against an arbitrary reference category)

$$log\left(\frac{Pr(y{=}1|x)}{Pr(y{=}K|x)}\right) = \beta_{10} + \beta_{11}x$$

1.

$$log\left(\frac{Pr(y{=}2|x)}{Pr(y{=}K|x)}\right) = \beta_{20} + \beta_{21}x$$

2.

$$\vdots$$

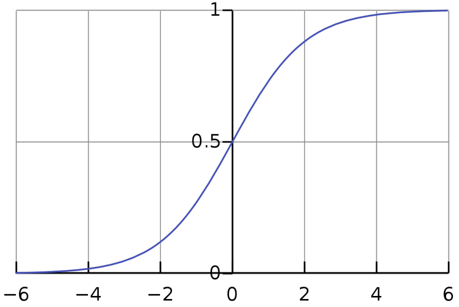
$$log\left(\frac{Pr(y{=}K{-}1|x)}{Pr(y{=}K|x)}\right) = \beta_{(K-1)0} + \beta_{(K-1)1}x$$

K-1.

the softmax function

generalization of the logistic function to multiple dimensions

$$\sigma(x) = \frac{exp(x)}{1+exp(x)}$$

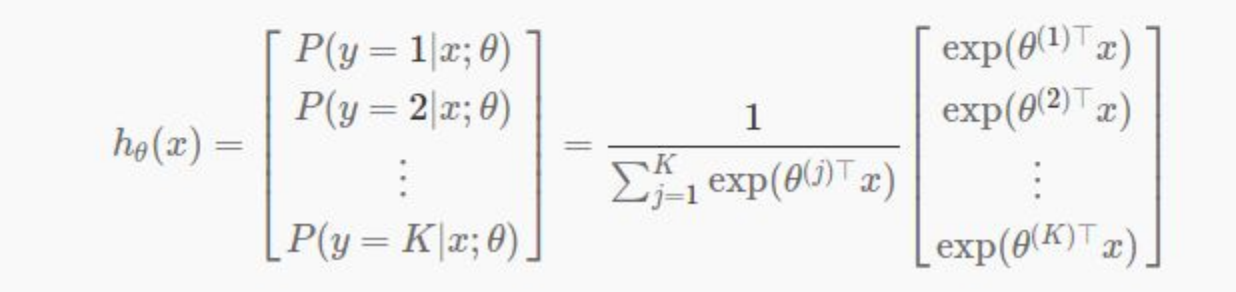


$$\mathrm{softmax}(x)_i = \frac{exp(x_i)}{\sum_{j=1}^{k} exp(x_j)}$$

Vector of k probabilities for each observations

the softmax function

generalization of the logistic function to multiple dimensions



$$h_\theta(x) = \begin{bmatrix} P(y=1|x;\theta) \\ P(y=2|x;\theta) \\ \vdots \\ P(y=K|x;\theta) \end{bmatrix} = \frac{1}{\sum_{j=1}^{K} \exp(\theta^{(j)\top} x)} \begin{bmatrix} \exp(\theta^{(1)\top} x) \\ \exp(\theta^{(2)\top} x) \\ \vdots \\ \exp(\theta^{(K)\top} x) \end{bmatrix}$$

From: http://deeplearning.stanford.edu/tutorial/supervised/SoftmaxRegression/

multinomial logistic regression: p-values

How do we get p-values from multinomial logistic regression models?

- linear regression: t-test (single coefficients) or F-test (model comparisons)

Logistic regression

1. **Wald test**
2. **Likelihood ratio test**

multinomial logistic regression: p-values

Wald test

$$W = \frac{(\hat{\beta} - \beta_0)^2}{\sigma^2(\hat{\beta})}$$

- difference between estimated coefficients and null hypothesis (e.g. β = 0)
- W is distributed as a **chi-square random variable** (1 d.f.)
- equivalent to the square of a N(0,1) random variable

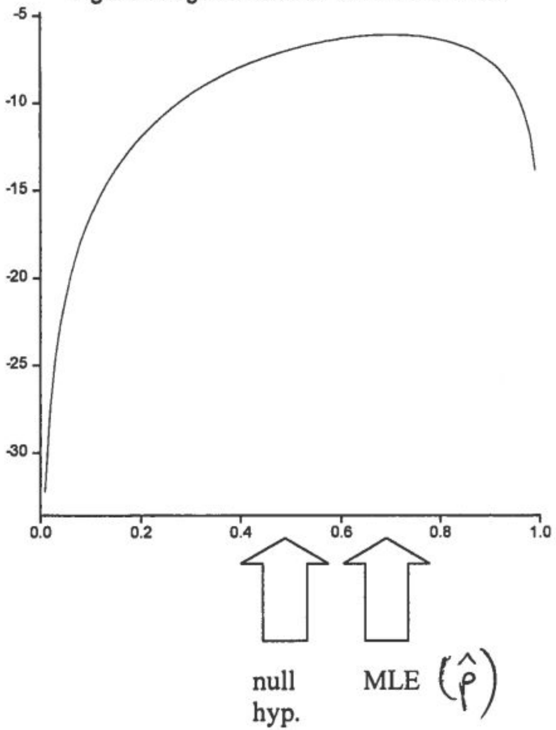
multinomial logistic regression: p-values

Likelihood ratio test

$$LR = 2\{logL(y,\hat{\beta}) - logL(y,\beta_0)\}$$

- likelihood ratio → difference between log(likelihood)
- compare **full vs reduced model**
- Under H_0 LR follows a chi-square distribution (1 d.f.)

Figure 2 Log likelihood for insemination data



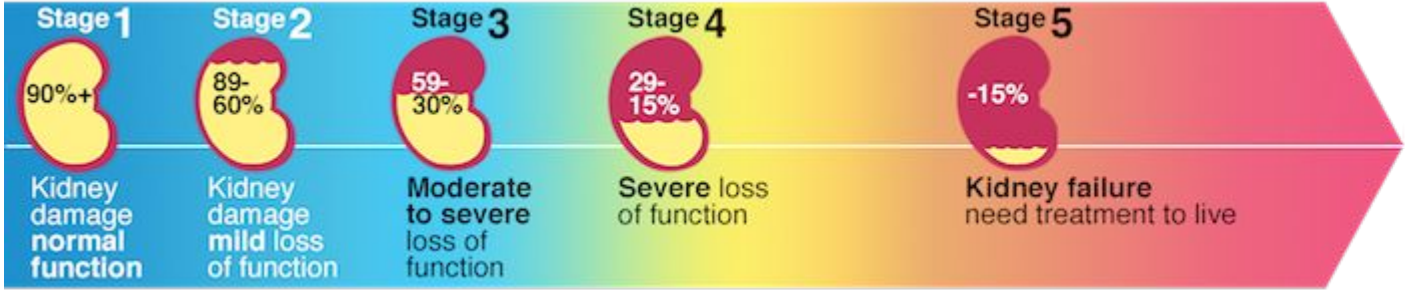
Practical session

Rstudio:

→ **1.categorical_gwas.Rmd** (part 1: multinomial logistic regression)

ordered categorical traits

- calving ease/difficulty:
  - A = easy
  - B = assisted
  - C = cesarean
  - D = difficult
  - E = embryotomy
- litter size
- diseases can be graded on scales from least severe to most severe
  - e.g. COPD (chronic obstructive pulmonary disease): **stages 1 - 4** (least to most severe)
  - CDK (chronic kidney disease): stages 1 - 5


Stage 1
90%+
Kidney damage normal function
Stage 2
89-60%
Kidney damage mild loss of function
Stage 3
59-30%
Moderate to severe loss of function
Stage 4
29-15%
Severe loss of function
Stage 5
-15%
Kidney failure need treatment to live

From: https://lifeoptions.org/learn-about-kidney-disease/causes-and-stages/

ordered categorical traits

- natural ordering (ranking) between categories
- intervals are not necessarily equally spaced (this may have consequences on modeling and interpretation):
    - disagree → no opinion → agree [equally spaced]
    - 4 seasons [equally spaced]
    - primary school → high-school - BS → MSc → PhD [uneven spaces]

ordered categorical traits - analysis options

- **linear regression**: maybe problematic because some of the assumptions are violated (especially when categories are not evenly spaced)
- **ANOVA**: if you have only one continuous predictor, you could “flip” the model around so that the categorical variable becomes the outcome variable (special case of linear regression model)
- **multinomial logistic regression**: it assumes that there is no order to the categories of the outcome variable (i.e., the categories are nominal). The downside of this approach is that the information contained in the ordering is lost.

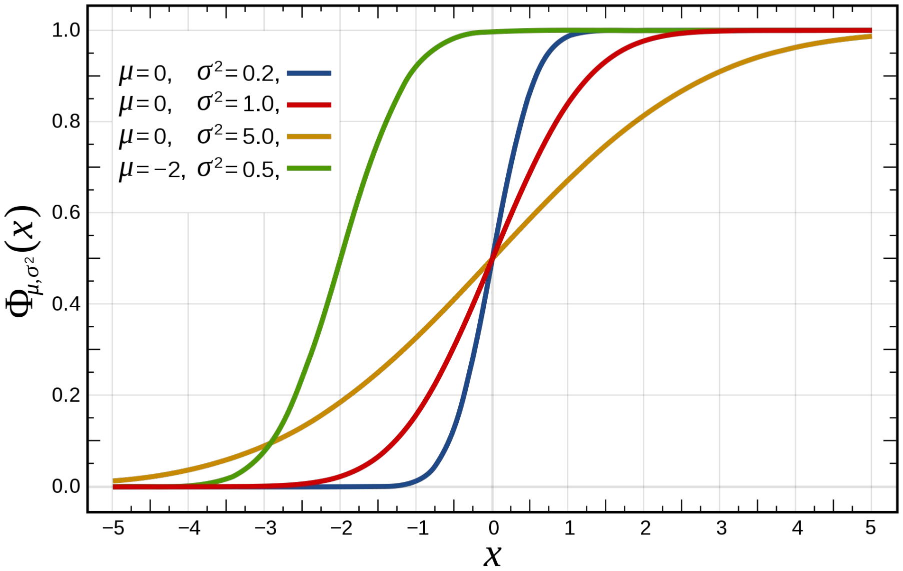
ordered categorical traits - analysis options

- ~~**linear regression**: maybe problematic because some of the assumptions are violated (especially when categories are not evenly spaced)~~
- ~~**ANOVA**: if you have only one continuous predictor, you could “flip” the model around so that the categorical variable becomes the outcome variable (special case of linear regression model)~~
- ~~**multinomial logistic regression**: it assumes that there is no order to the categories of the outcome variable (i.e., the categories are nominal). The downside of this approach is that the information contained in the ordering is lost.~~
- **ordered logistic regression**: based on **cumulative probabilities** (rather than the probability of individual events, as in multinomial logistic regression).
- **ordered probit regression**: very (very!) similar to ordered logistic regression (the main difference is in the interpretation of the coefficients → no longer log(odds))

ordered logistic regression

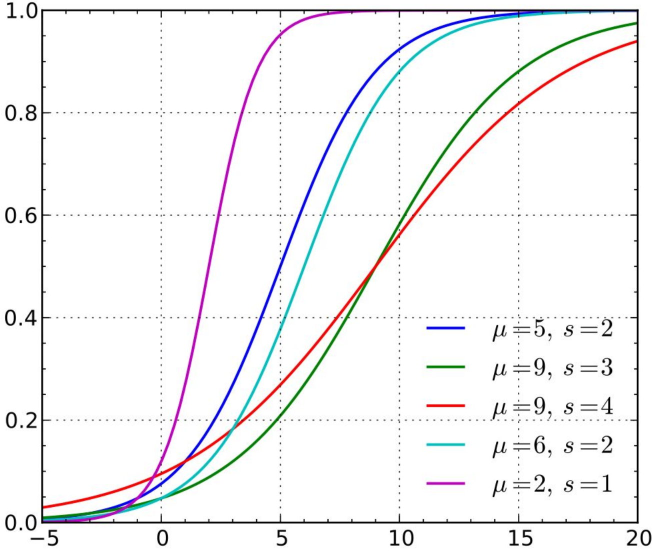
$$P(y \leq j) = \pi_1 + \pi_2 + \ldots + \pi_j, [j = 1, 2, \ldots, J]$$

Normal CDF



From: https://en.wikipedia.org/wiki/Cumulative_distribution_function

Logistic CDF



ordered logistic regression

- **y**: ordinal variable with *J categories*
- **P(y <= j):** cumulative probability of y less than (or equal) a specific category *j*
- we can then express the odds

$$\frac{P(y \leq j)}{P(y > j)}$$

ordered logistic regression

- **y**: ordinal variable with *J categories*
- **P(y <= j):** cumulative probability of y less than (or equal) a specific category *j*
- we can then express the odds:

$$\frac{P(y \leq j)}{P(y > j)}$$

- and then the log(odds) [the logit!]:

$$log\left(\frac{P(y \leq j)}{P(y > j)}\right) = logit(P(y \leq j)) = \beta_{j0} + \beta_1 x_1 + \ldots + \beta_p x_p$$

ordered logistic regression

- **y**: ordinal variable with *J categories*
- **P(y <= j):** cumulative probability of y less than (or equal) a specific category *j*
- we can then express the odds:

$$\frac{P(y\leq j)}{P(y>j)}$$

- and then the log(odds) [the logit!]:

$$log\left(\frac{P(y\leq j)}{P(y>j)}\right) = logit(P(y \leq j)) = \beta_{j0} + \beta_1 x_1 + \ldots + \beta_p x_p$$

different intercept for each category!

ordered logistic regression: p-values

Like in multinomial logistic regression (and GLMs in general), we again have (at least) two ways to obtain the p-values for the effects in the model:

1. Wald test
2. Likelihood ratio test

ordered logistic regression: probit vs logit

- the **probit function** is the inverse of the normal CDF (just like the logit function is the inverse of the binomial CDF → the logistic function!)

- $\boldsymbol{\varphi}(x) = p \rightarrow \text{probit}(p) = x$
- $\boldsymbol{\sigma}(x) = p \rightarrow \text{logit}(p) = x$

Practical session

Rstudio:

→ **1.categorical_gwas.Rmd** (part 2: ordered logistic regression)
→ **2.categorical_gwas_example.Rmd**

A primer on GWAS for longitudinal traits

time-to-event data

Examples of types of events:

- relapse
- progression
- death
- in cows (livestock) also longevity

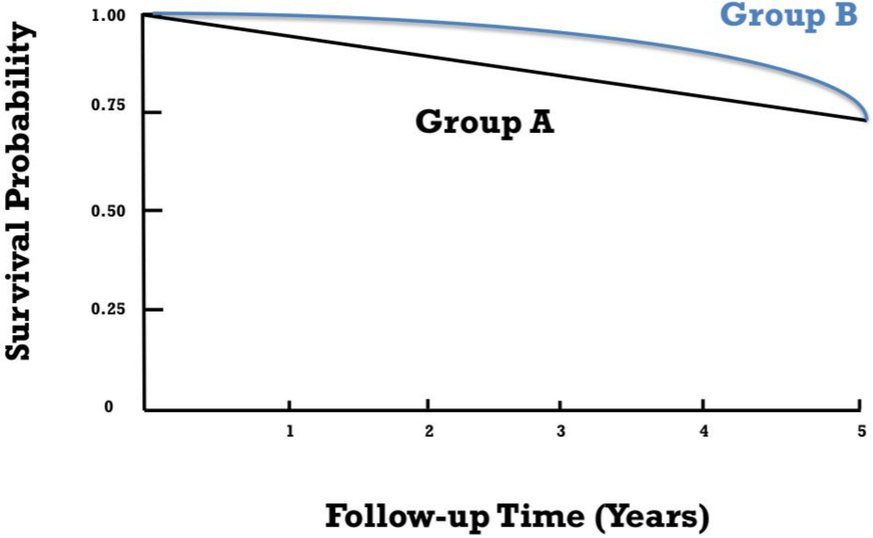
time-to-event data

Characteristics of time events:

- subjects enter at different times and have different duration of follow-up

- entire survival experience, not just the percentages who remain alive at the end of the study

- the survival distributions may differ even though the five-year survival rates are similar

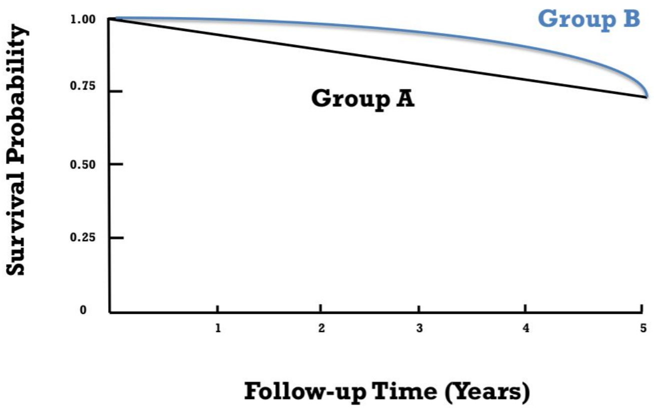


time-to-event data

Quantities of interest:

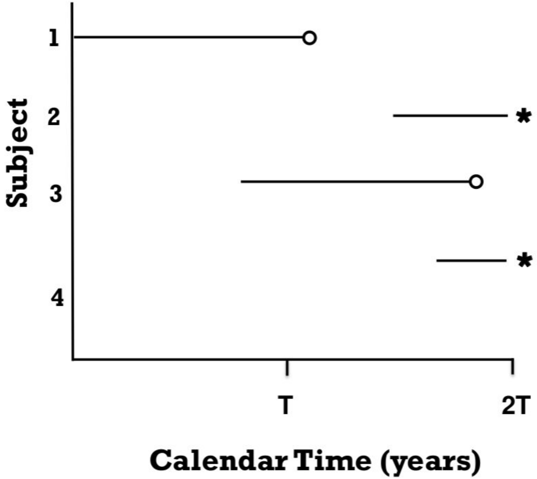
- **survival probability**, also known as the survival function ***S(t)***, is the probability that an individual survives from the time origin (e.g. diagnosis of cancer) to a specified **future time "t"**

- **hazard**, denoted by ***h(t)*** (or ***λ(t)***), is the probability that an individual who is under observation at a time "t" **has an event** at that time



survival data

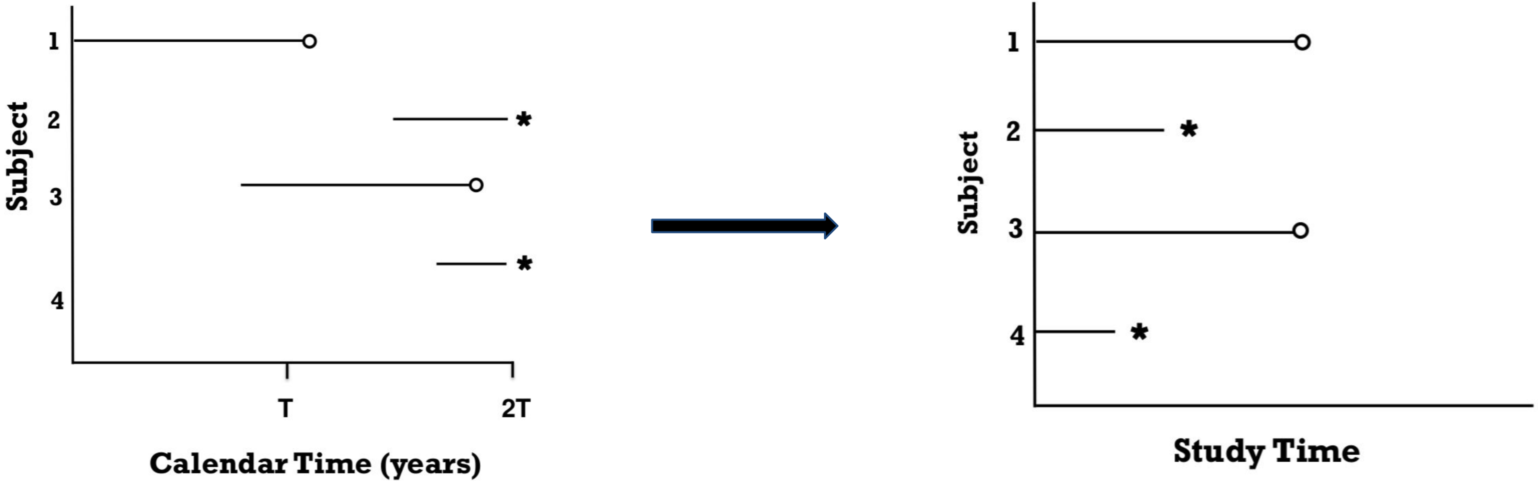
- we can’t assume normality
- **censored data** (*right censoring*):
  - study follow-up ends before a participant has experienced the event
  - participants withdraw or are lost to follow-up, again prior to observing the event



Two patients have events (circles), two are censored (asterisks) because the study ended

survival data

Time to event data are **normalised** by representing each **time record relative to admission date/enrollment**



survival data

- *T*: random variable representing **time to event** (e.g. death) for a subject
- *F(t)*: the **probability** that the event (e.g. death) occurs before time ***t*** (end of study): **cumulative risk**, or **distribution function for time-to-event** (T)

$$F(t) = Pr(T < t)$$

- survival is the **complement of *F(t)***, defined as the probability that the subject has not had the event by time t

$$S(t) = 1 - F(t)$$

Kaplan-Meier estimate of S(t)

- **S(t)** would be easy to estimate if there were no censoring: however, we almost always have censored data → **Kaplan-Meier estimate of S(t)**
- K-M **updates S(t)** (**step function**) **when events occur** based on the proportion of study participants followed to that time point who have an event

Kaplan-Meier estimate of S(t)

group	year	deaths	survivors
1	1	20	80
2	1	25	75
1	2	20	60
2	2	NA	NA

group	deaths	survivors	year_1	year_2
1	20	80	0.8	NA
1	20	60	NA	0.75
2	25	75	0.75	NA
2	NA	NA	NA	NA

Kaplan-Meier estimate of S(t)

group	year	deaths	survivors
1	1	20	80
2	1	25	75
1	2	20	60
2	2	NA	NA

group	deaths	survivors	year_1	year_2
1	20	80	0.8	NA
1	20	60	NA	0.75
2	25	75	0.75	NA
2	NA	NA	NA	NA

year	Pr(S)
1	0.775
2	0.75

Kaplan-Meier estimate of S(t)

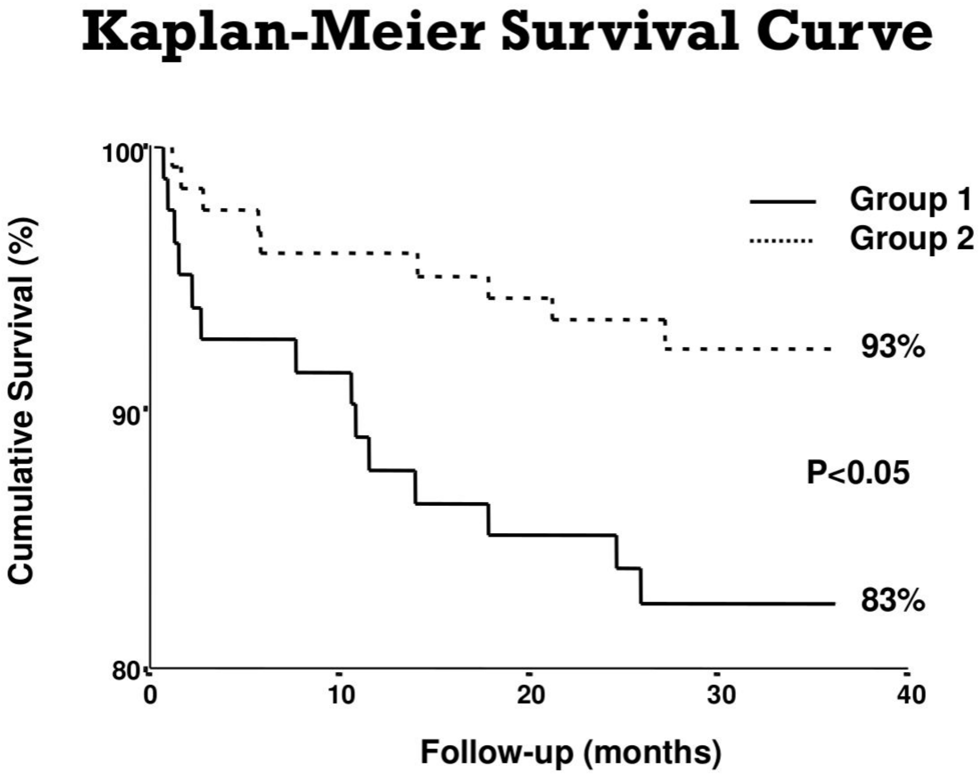
year	Pr(S)
1	0.775
2	0.75

Conditional probabilities:

- **S(year1) = 0.775**
- **S(year2|year1) = 0.75** (alive at year 2 given they're alive at year 1)

Then, **S(year2) = S(year1)*S(year2|year1) = 0.775*0.75 = 0.58**

Kaplan-Meier curves



comparing survival curves

- once we have constructed our survival curves, we usually want to know if they differ between groups (e.g. treatments, sexes, breeds etc.)

Many ways to do this:

1. Mantel-Haenszel test
2. Log-rank test

Both are based on multi-dimensional contingency (frequency) tables, comparing observed and expected frequencies accounting for stratification (e.g. treatment or sex or breed)

from Kaplan-Meier curves to Cox models

- Kaplan-Meier curves and log-rank tests are useful for univariate analysis, describing survival in terms of one factor under investigation, and typically work only with categorical predictors (e.g. sex, treatment A vs treatment B etc.)

- this is where **Cox proportional hazards regression analysis** comes in handy: it works for **both quantitative predictor** variables **and for categorical variables**. Furthermore, the Cox regression model extends survival analysis methods to **assess simultaneously the effect of several risk factors** on survival time

- Cox models examine how specified factors (covariates) influence the rate (hazard rate) of a particular event happening (e.g. infection, death) at a particular point in time

from Kaplan-Meier curves to Cox models

The hazard function (**λ()** or **h()**) is defined as the **event rate at time *t* conditional on survival until time *t*** (or later, T ≥ t) → suppose a subject has survived for a time *t* and we want the probability that it will not survive for an additional time dt:

$$h(t) = \frac{P(t<T<(t+dt))}{P(T>t)dt} = h_0(t)exp(\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)$$

where $h_0(t)$ is some baseline or reference hazard

from Kaplan-Meier curves to Cox models

we can then express h(t) relative to h_0(t) and take the logarithm (rings a bell?):

$$\ln\left(\frac{h(t)}{h_0(t)}\right) = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

integrating over the elapsed time *t*, we obtain the cumulative risk (**F(t)**), which is related to the survival function (**S(t)**)

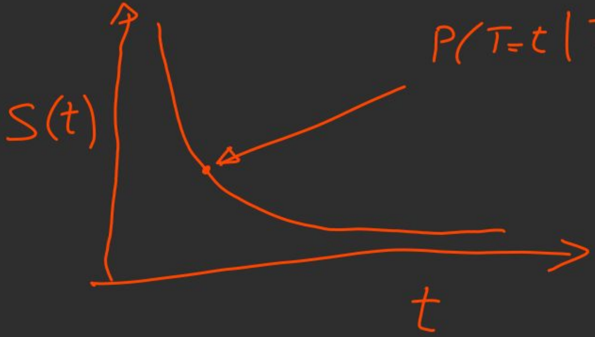
$$S(t) = P(T \geq t) = 1 - P(T < t) = 1 - F(t)$$

survival at time t

adverse event "death"

cumulative distribution of T at time "t"

S(t) follows a distribution, e.g. exponential



$$P(T=t \mid T \geq t) = \underline{h(t)} : \underline{\text{hazard function}}$$

conditional probability

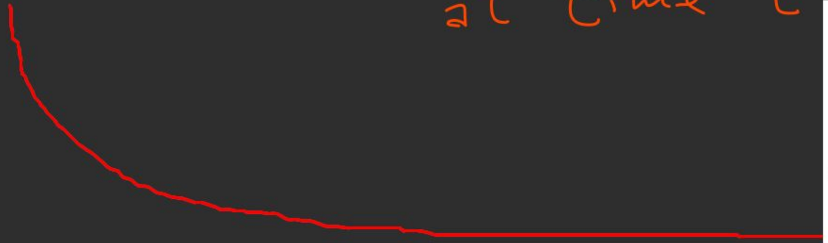
$$\frac{P(T=t)}{S(t)}$$

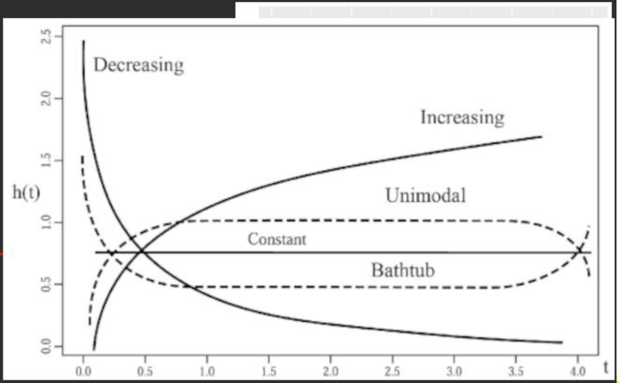
hazard function

$$h(t) = \frac{P(T=t)}{S(t)} \leftarrow \text{derivative of } S(t) \text{ at "}t\text{"}$$

$$= \frac{\frac{d(S(t))}{dt}}{S(t)} \longleftarrow \left[\frac{f'(x)}{f(x)} = \frac{d \ln(f(x))}{dx} \right]$$

$$= \frac{d}{dt} \ln(S(t))$$: probability of "death" at time "t"





$h(t) = P(T=t \mid T \geq t)$

↳ can be modelled as function of covariates (risk factors)

COX REGRESSION (proportional hazard model)

$$\boxed{h(t) = h_0(t) \cdot e^{\beta x}}$$

covariates: additional risk given by risk factors (on top of $h_0(t)$)

baseline hazard: risk of "death" at time "t" (from the S(t) distribution function)

$$h(t) = h_0(t) \cdot e^{\beta x}$$

$$\frac{h(t)}{h_0(t)} = e^{\beta x} \longrightarrow \ln\left(\frac{h(t)}{h_0(t)}\right) = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

Interpretation of coefficients (β's) $\left[\ln(h'(t)) = \beta x\right]$

$[x = 1]$

$$\boxed{e^{\beta} - 1} \rightarrow e^{0.1} - 1 = 1.10 - 1 = 0.10 \rightarrow$$ 10% increased risk for unit of the "x" variable

$\rightarrow [e^{0} = 1]$

Practical session

Rstudio:

→ **3.longitudinal_gwas.Rmd**
→ **4.r_packages_for_longitudinal_gwas.Rmd (??)**

NEXT LECTURE

GWASpoly example for polyploid species