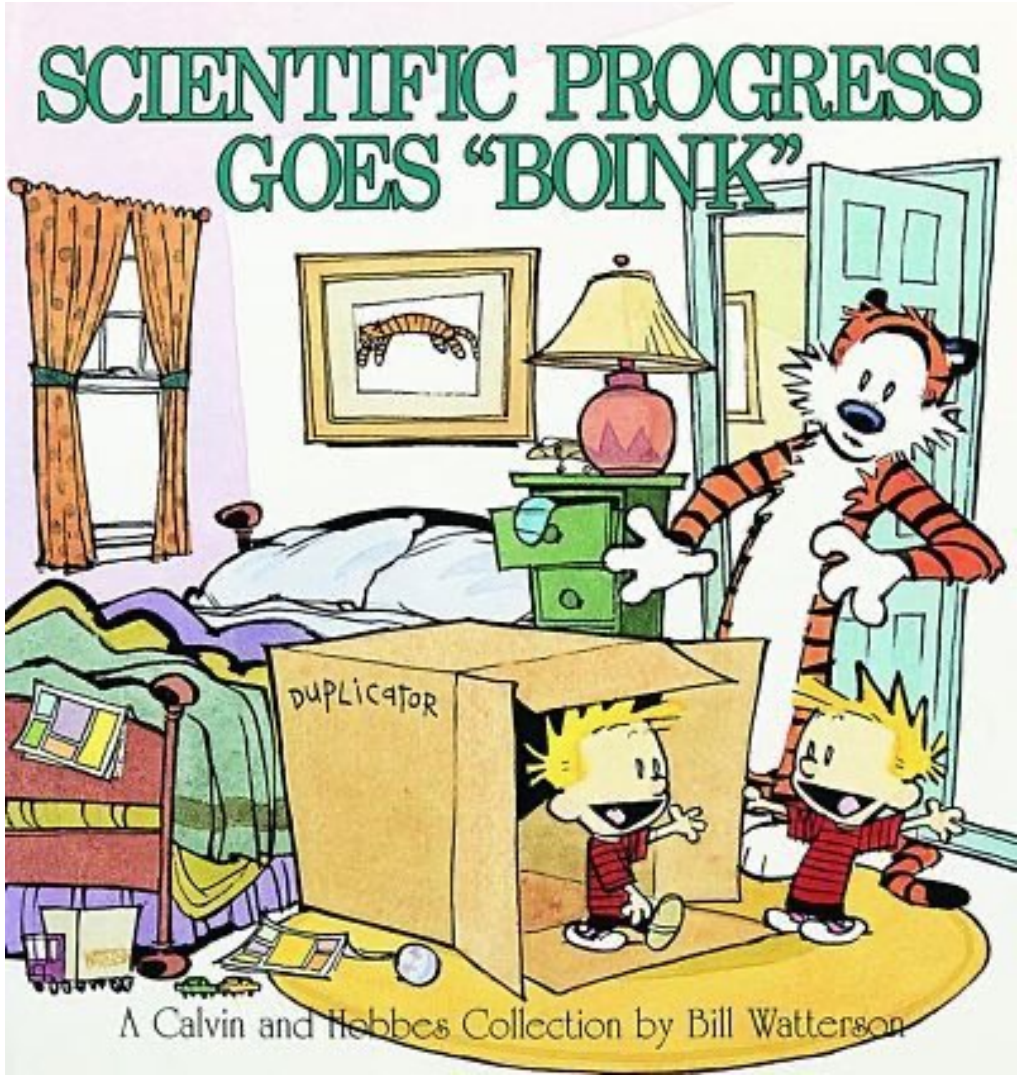
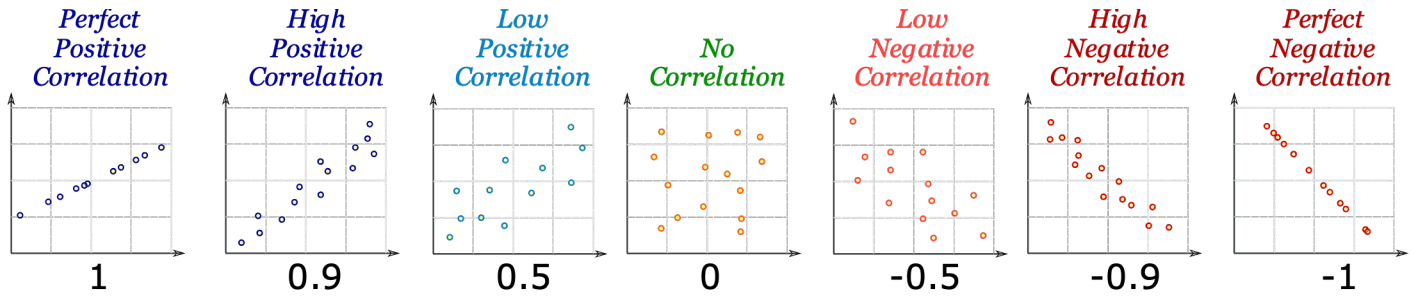
Testing your hypotheses in R!



Brendan Reid
Philippines PIRE post-Omics workshop
7/12/22

A refresher on frequentist statistics

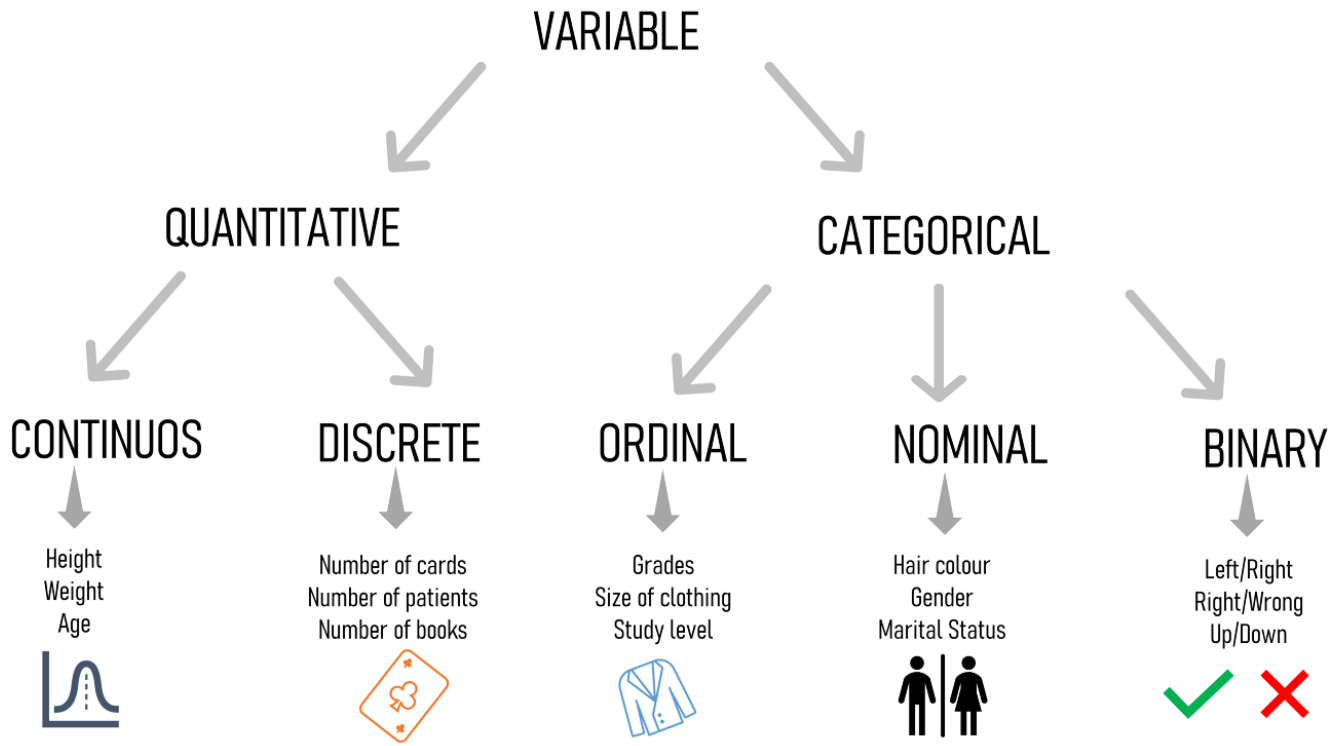
- We have our observed data
  - Response/dependent variable
    - Variable we are interested
    - Usually plotted on y axis
  - Predictor/independent variable(s)
    - Variables that we think might be associated with or have an effect on the response variable
    - Usually plotted on x axis



- There will usually be some correlation between response and predictor
- **How often would would we observe a similar relationship between a random predictor variable and our response?**

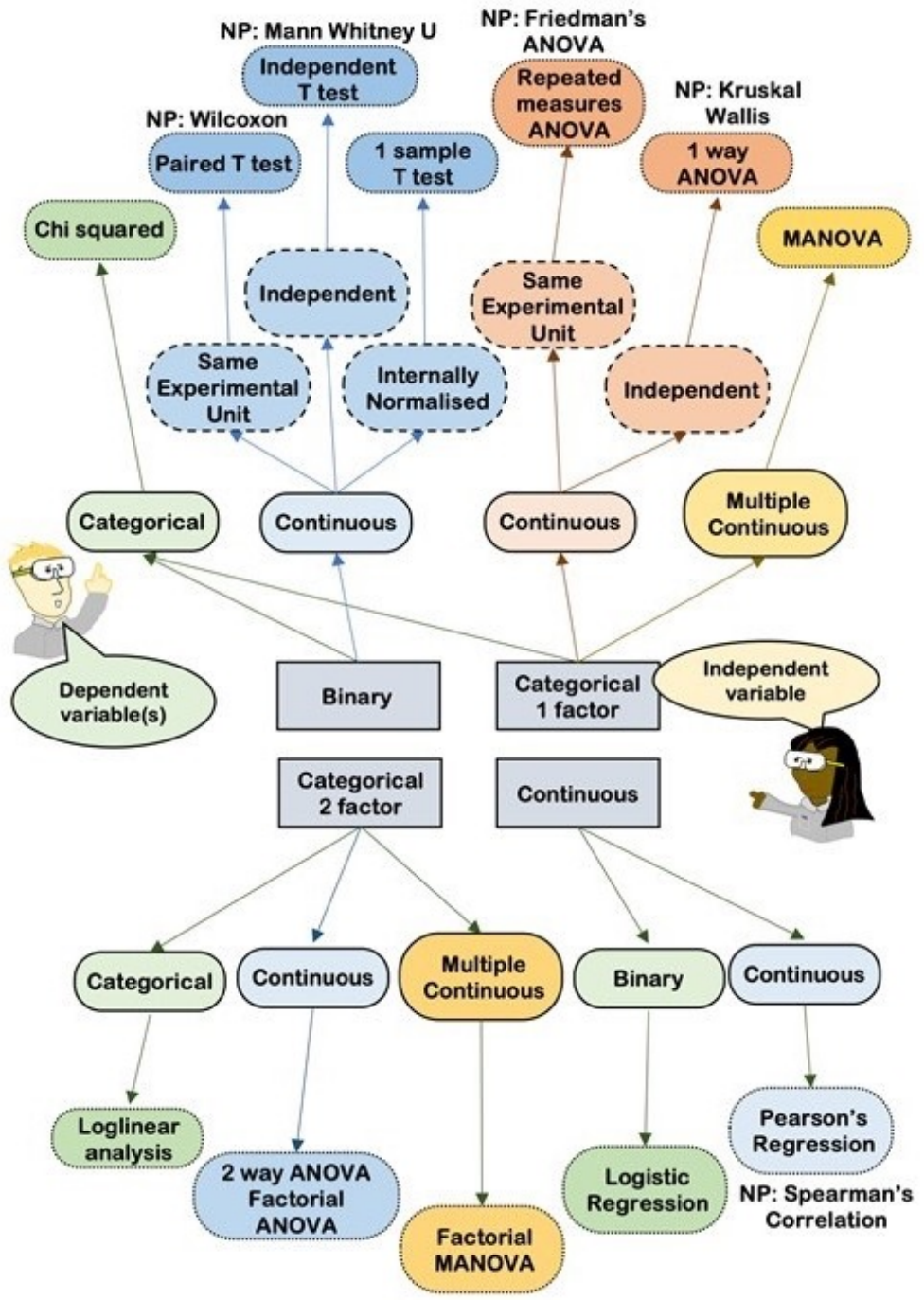
Types of variables

- Quantitative
  - Continuous or discrete
  - Ordered
- Categorical
  - Binary (true/false or 0/1)
  - Nonimal (unordered)
  - Ordinal (ordered!)



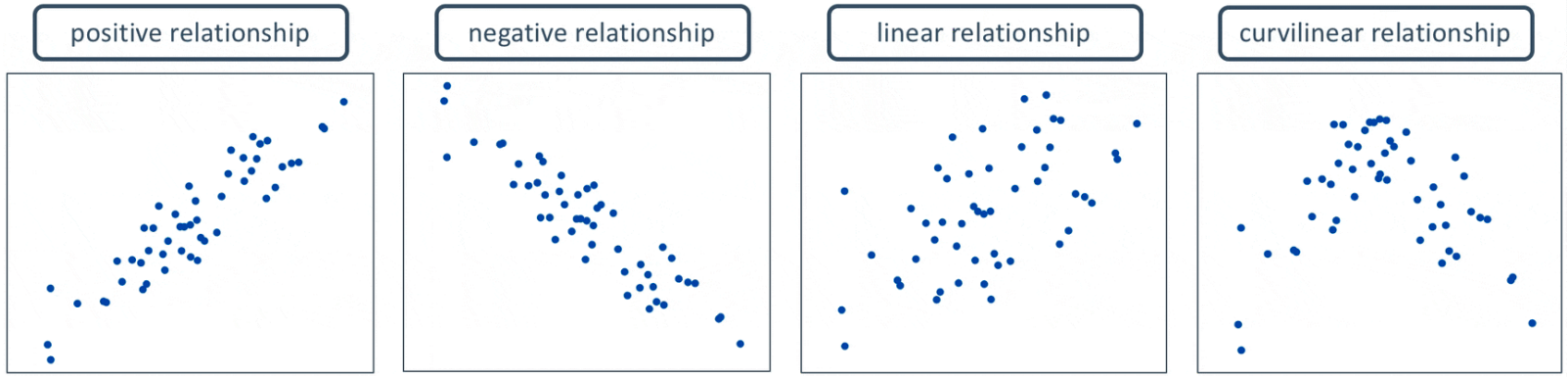
towarddatascience.com

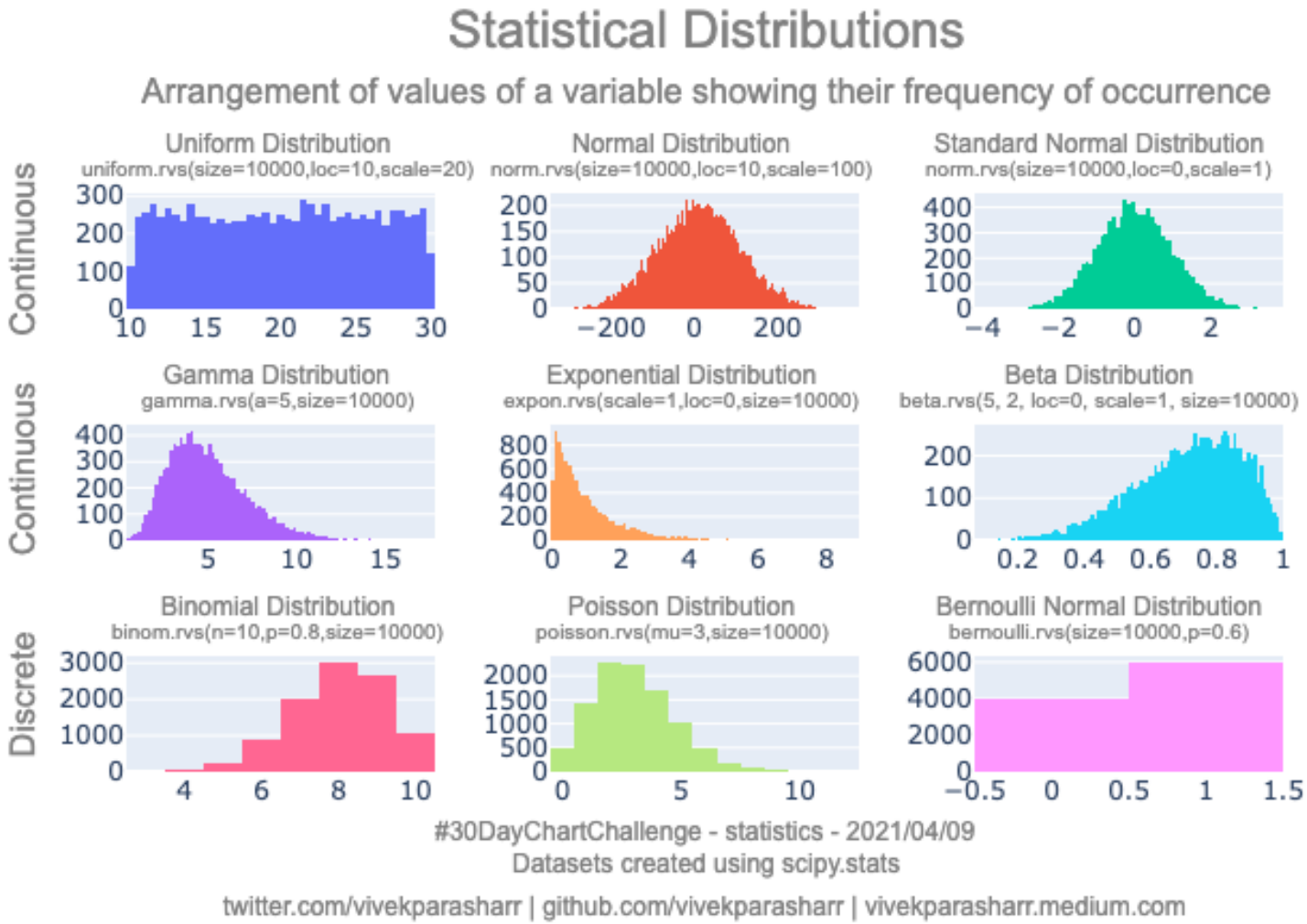
Statistical tests!

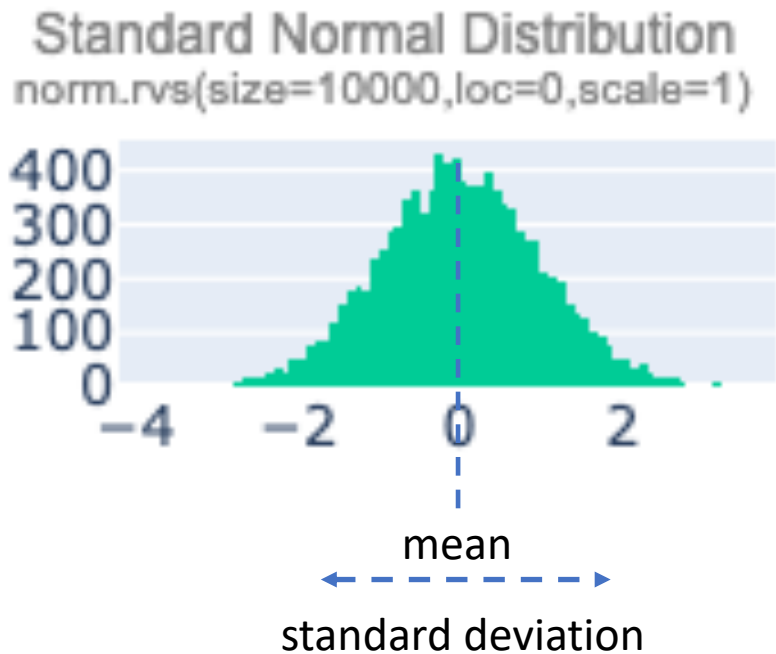


Linear regression

- Response/dependent variable and predictor/independent variables are both continuous
- Find the **linear model** that fits the data best
  - Remember y = mx + b …
  - $Y = \beta_0 + \beta_1 x + \varepsilon$
  - β_0 is the intercept
  - β_1 is the coefficient associated with x
  - ε is error

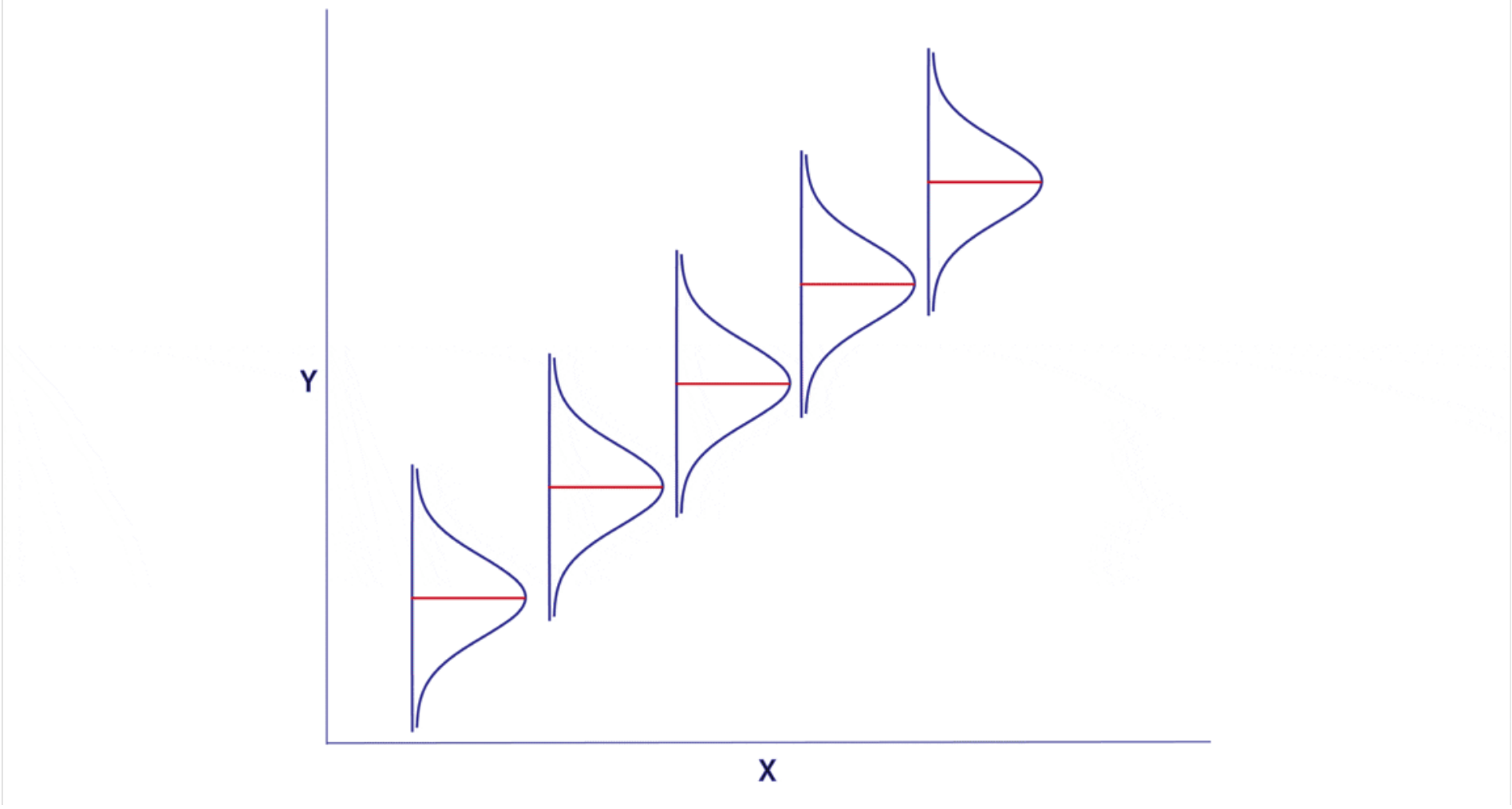






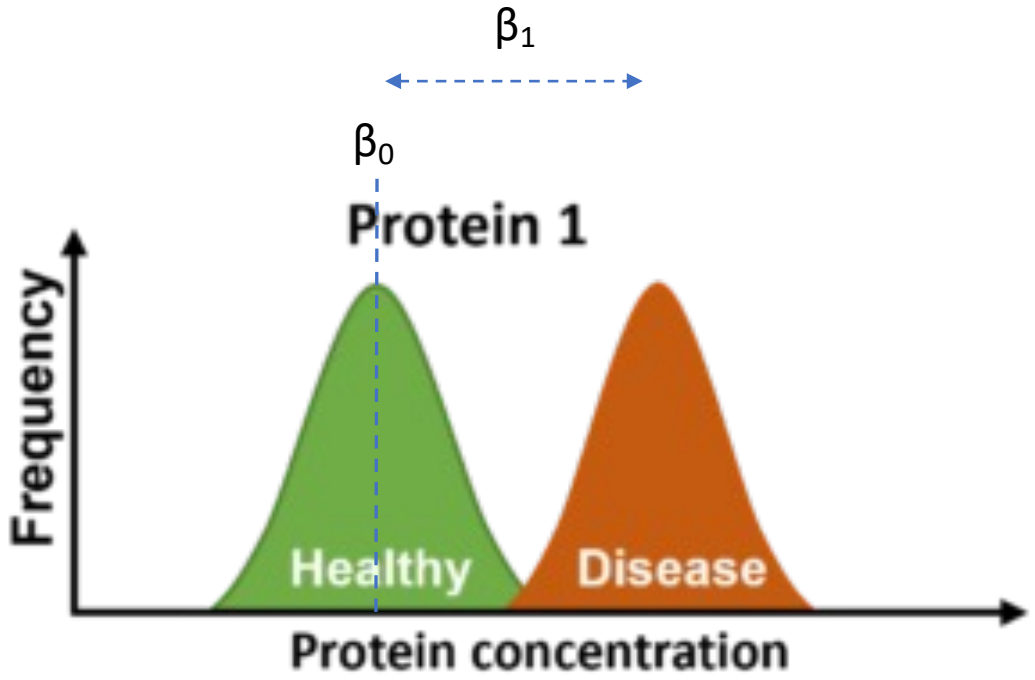
becominghuman.ai

Fitting a linear regression



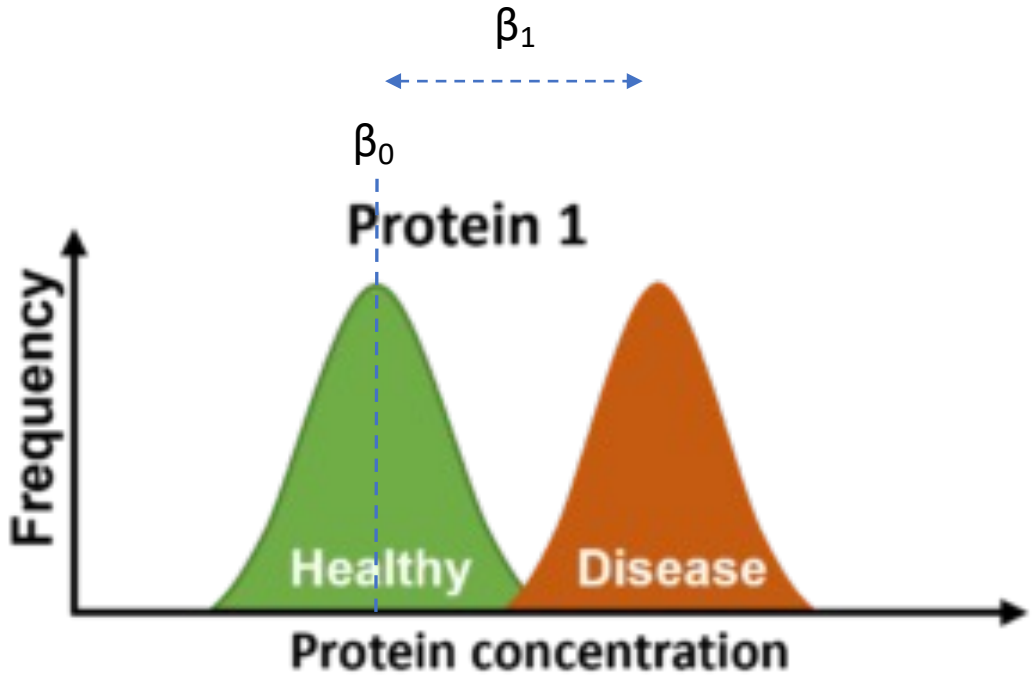
ANOVA

- Response/independent variable is continuous, predictor/dependent variable is categorical



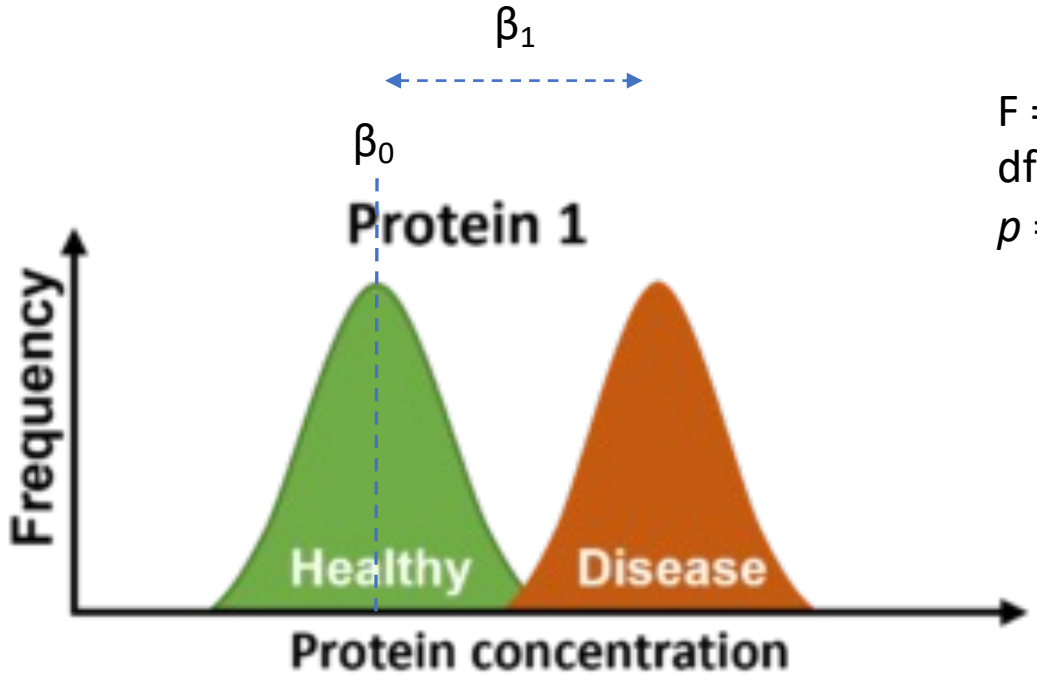
Hypotheses

- Null hypothesis
- Alternate hypothesis (or hypotheses)



Statistical tests and *p*-values

- F-statistic: based on difference between means and the variances
- F can be used with degrees of freedom to calculate a p-value (probability of obtaining similar results if the null hypothesis is true)
- Statistical significance: $p < \alpha$ (usually 0.05)



F = 13.8
df = (1,13)
p = 0.002596

Running linear regression and ANOVA in R

- lm(<formula>,<data>,…)
- Formula syntax = dependent variable ~ independent variable(s)
- Both dependent and independent variables should be columns in a data frame <data>

Fitting Linear Models

Description

`lm` is used to fit linear models, including multivariate ones. It can be used to carry out regression, single stratum analysis of variance and analysis of covariance (although aov may provide a more convenient interface for these).

Usage

```
lm(formula, data, subset, weights, na.action,
   method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE,
   singular.ok = TRUE, contrasts = NULL, offset, ...)
```

Assumptions and diagnostics

- Assumptions
  - Normality
  - Equal variance among groups
  - Independence
  - Linearity (for regression
- How do we test these assumptions?
  - Q-Q plot

Plotting regression results

- Scatter plots
- Confidence bands

R exercise

PSMC data