Statistics for Life Sciences

Nathan Harmston

2022-04-11

Problem set 2

So what have we done ?

Probability is the basis for all statistics

- **describe** and **understand** randomness

Statistics

- we want to ask whether what we observe / end up with is different to what we'd expect by chance (**observed** vs. **expected**)
- make inferences and decisions about a population from a sample

Laws of probability

- $P(A) \geq 0$
- $P(\Omega) = 1$
- additivity = if $A \cap B = \emptyset$, then $P(A \cup B) = P(A) + P(B)$

Conditional probability

- $P(A|B)$ - probability that event A will occur given that event B has occurred

$$P(A \cap B) = P(B)P(A|B)$$

$$P(A \cap B) = P(A)P(B|A)$$

$$\mathrm{P(B)} > 0$$

Bayes Theorem

- $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

...

- prevalence of a disease affects the probability of having the disease given a positive test
- Monty Hall problem

Independence

statistical independence refers to the case where one cannot predict anything about one variable from the value of another variable

$$P(A \cap B) = P(A)P(B)$$

odds ratios

the odds of the event occurring in condition A versus the odds of the event occurring in a not condition A

odds

$$odds = \frac{p}{1-p} \quad (1)$$

$$\text{OR} = \frac{\frac{p_{C|A}}{1-p_{C|A}}}{\frac{p_{C|!A}}{1-p_{C|!A}}} \quad (2)$$

often we talk about the log(odds)

Probability distributions

- Bernoulli
- Binomial
- Poisson
- Normal (and standard Normal)
- t
- F
- χ^2
- beta

functions in R

- d
- p
- q
- r

Probability distributions

- Probability mass function
- Probability density function
- Cumulative distribution function

characterised by parameters....

- mean μ
- standard deviation σ
- λ
- number of trials, successes and θ
- degrees of freedom

Central limit theorem

if your sample size is large enough, the distribution of means will approximate a Gaussian distribution, even if the population is not Gaussian

Regression to the mean

in any series of random events an extraordinary event is more likely to be followed, due to chance, by an more ordinary one

Law of large numbers

empirical probability will approach the true probability as the sample size increases

Standard Error

$$\frac{\sigma}{\sqrt{n}} \tag{3}$$

Confidence intervals

$$\mu = x + t^*(\frac{s}{\sqrt{n}}) \tag{4}$$

$$\mu = x - t^*(\frac{s}{\sqrt{n}}) \tag{5}$$

Hypothesis testing

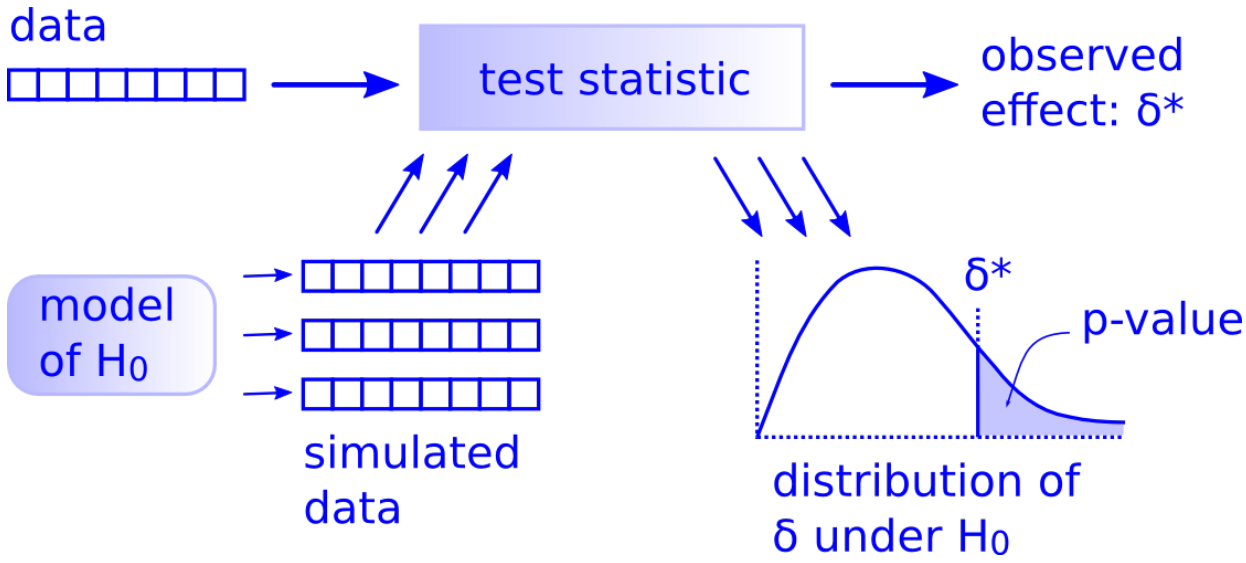
Hypotheses

- null - there is no difference
- alternative - there is a difference

Hypotheses

- one-tailed
- two-tailed

Hypothesis testing



Hypothesis testing

Type I errors

- false positives
- α

Type II errors

- false negatives
- β

Hypothesis testing

lots of different tests

- Binomial test
- ....

comparing means

- t-tests
- Mann-Whitney
- Permutation tests

Hypothesis testing

Testing for independence

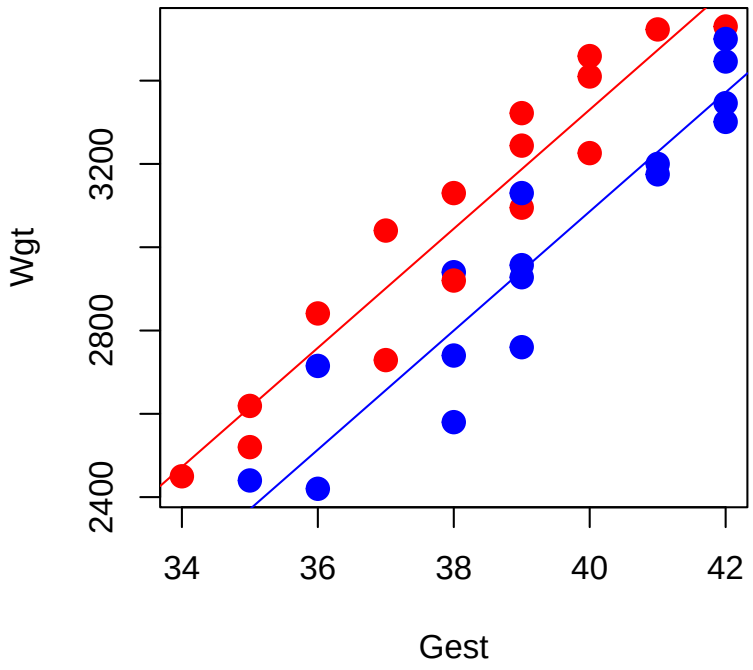
- χ^2 test
  - what do we observe compared to what we expect under a model independence?
- Fishers exact test

Goodness of fit

- χ^2 test
  - what do we observe compared to what we expect under our model?

Regression

- Linear regression
- Multiple linear regression
- ....



ANOVA

- partitioning variance into its different components
- between group variability compared to within group variability
- F-distribution

interactions

- interaction effects represent the combined effects of factors on the dependent variable.
- the effect of one factor depends on the level of the other factor

F-test

test for equality of variances (?var.test)

so what is the p-value of the differences we observe?

F-statistic

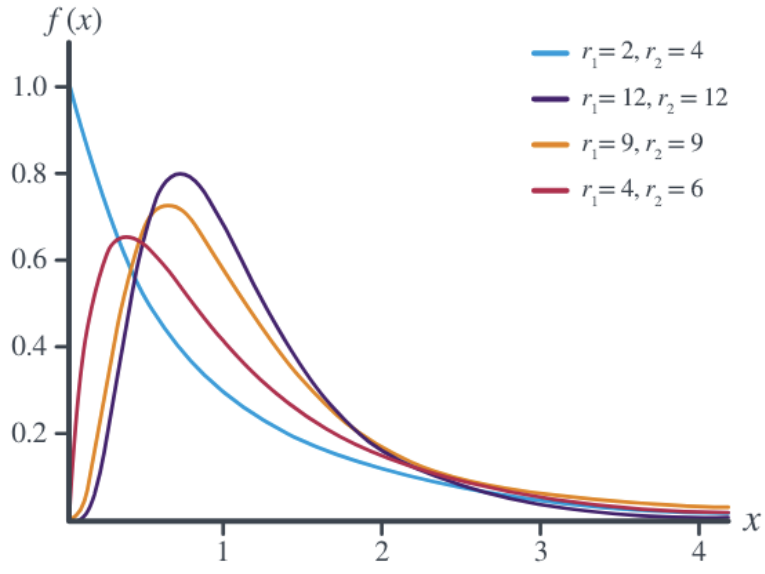
$$F = \frac{\text{between-group variability}}{\text{within-group variability}} \tag{6}$$

F-distribution

when two random samples are taken from two normal distributions with equal variance the ratio of those variances follows an F-distribution

defined by degrees on freedom in sample one and degrees of freedom in sample two

- ?df
- ?qf
- ?pf
- ?rf



Multiple hypothesis testing

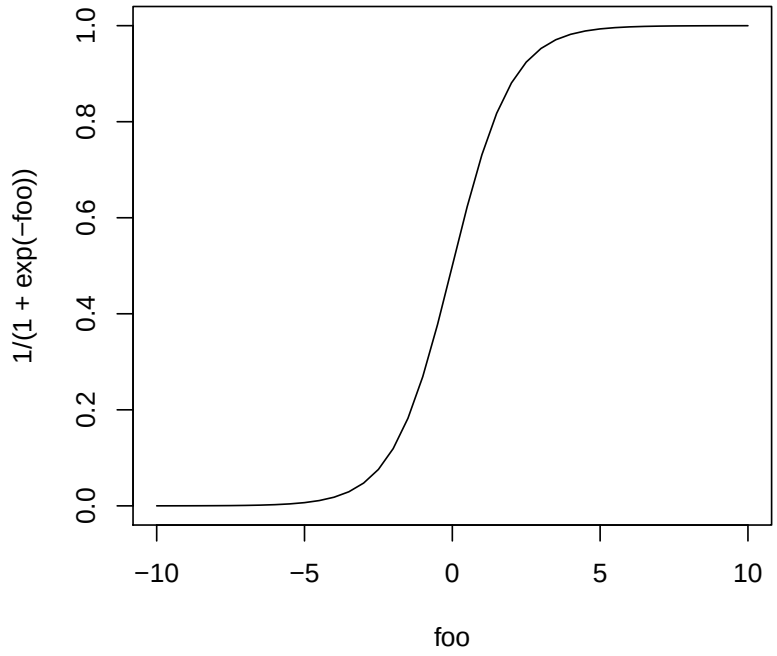
- Type I errors are bad
- control either the family wise error rate - Bonferroni
- or control the False discovery rate - Benjami-Hochberg

Power

- the ability of a test to correctly reject a false null hypothesis
- multiple factors affecting power

Logistic regression

- logit function
- binary outcome variable - continuous or discrete predictors



$$log(\frac{p(\text{residual disease})}{1-p(\text{residual disease})}) = \beta_0 + \beta_1 age$$

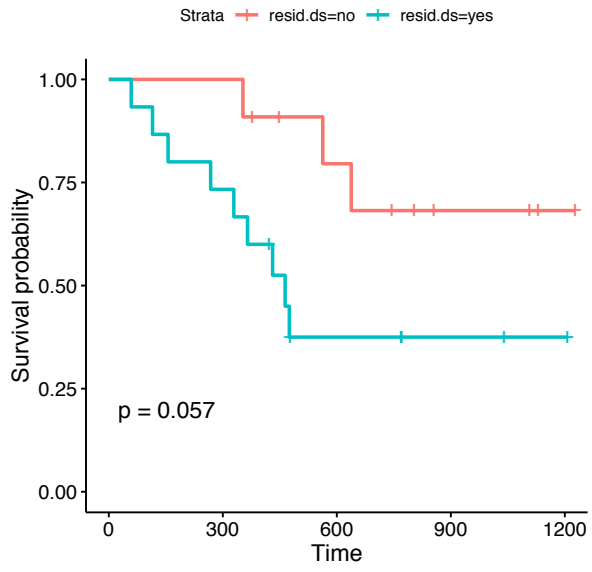
Time-to-Event Analysis

Censoring

- lots of reasons for individuals to get censored during the course of a study

Kaplan-Meier

- survival function
- log-rank test



Hazard

instantaneous risk that the event of interest happens, within a very narrow time frame.

$$HR = \frac{HAZ(X = 1)}{HAZ(X = 0)}$$

Cox proportional hazards regression

- very similar to logistic regression
- coefficients give you information on the hazard

Bayesian statistics

- posterior is proportional to the likelihood times the prior
- Monty Hall problem
- Ajay beats Nathan in the casino - similar to the problem of the points
- balance between data and prior knowledge

Model building

all models are wrong, some are useful

predict

overfitting

cross-validation

picking the best model

AIC, BIC, (adjusted) R-squared

forwards, backwards, stepwise, all subsets

Most Published Research Findings Are False

the probability that a research finding is indeed true depends on the prior probability of it being true (before doing the study), the statistical power of the study, and the level of statistical significance

Ioannidis 2005

Most Published Research Findings Are False

Bias

Ioannidis 2005

Most Published Research Findings Are False

The smaller the studies conducted in a scientific field, the less likely the research findings are to be true.

- small sample size means lower power
- PPV decreases as power decreases

Ioannidis 2005

Most Published Research Findings Are False

The smaller the effect sizes in a scientific field, the less likely the research findings are to be true.

- small effect size means lower power
- PPV decreases as power decreases

Ioannidis 2005

Most Published Research Findings Are False

The greater the number and the lesser the selection of tested relationships in a scientific field, the less likely the research findings are to be true.

- the probability that a finding is true after you've done the study depends on the pre-study odds it being true

Ioannidis 2005

Most Published Research Findings Are False

The greater the flexibility in designs, definitions, outcomes, and analytical modes in a scientific fi eld, the less likely the research findings are to be true.

- the more flexible you are with how you analyse the data, the more opportunities for bias occur

Ioannidis 2005

Most Published Research Findings Are False

The greater the financial and other interests and prejudices in a scientific field, the less likely the research findings are to be true.

- Conflicts of interest and prejudice may increase bias
- Prestigious investigators may suppress via the peer review process the appearance and dissemination of findings that refute their findings, thus condemning their field to perpetuate false dogma

Ioannidis 2005

The hotter a scientific field (with more scientific teams involved), the less likely the research findings are to be true.

- Proteus phenomenon - alternating extreme research claim and extremely opposite refutations

Ioannidis 2005

- always plot your data
- statistics means never having to say you're certain
- all statistical tests are based on assumptions
- statistics is there to make inferences from limited data
- sample size has a massive effect on the p-value
- every p-value tests a null hypothesis
- statistically significant does not mean large effect of biologically important
- multiple comparisons is a problem
- correlation does not imply causation (but sometimes its indicative)
- computers, R etc. is not a replacement for your ability to think and reason about your data/analysis/results
- Most Published Research Findings Are False