Statistical testing

3502-440 Methods of Scientific Working for Crop Science

WS 2024/2025

Prof. Dr. Karl Schmid
Institute of Plant Breeding, Seed Science and Population Genetics
University of Hohenheim

Outline

Statistics as inductive logic

Bayesian statistics

Frequentist Hypothesis Testing

Learning goals

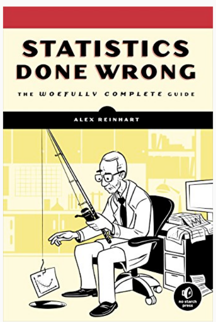
- Understand why statistical testing is central to scientific work
- Key problems in statistical analysis
- Main differences in Bayesian and frequentist statistical analysis

Motivation for this topic

Reasons for replication crisis

- HARKing (Hypothesizing After Results are Known)
- Low power
- *p*-hacking
- Publication bias

Know the rules of the game



Recapitulation of probability lecture

- Probability is an abstract concept
- Probability theory is purely deductive (deducted from Kolmogoroff's axioms)
- Bayes' Theorem
  - $P(H|D) \propto P(D|H) \times P(H)$ or
    Posterior $\propto$ Likelihood $\times$ Prior
  - $P(H|D) = \frac{P(D|H) \times P(H)}{P(D)}$
  - Formula for multiple hypotheses:

$$P(H_1|D) = \frac{P(D|H_1) \times P(H_1)}{\sum_{i=1}^{n}(P(D|H_i) \times P(H_i))}$$

  - Ratio form:

$$\frac{P(H_1|D)}{P(H_2|D)} = \frac{P(D|H_1)}{P(D|H_2)} \times \frac{P(H_1)}{P(H_2)}$$

Key concepts of today's lecture

- Probability theory versus statistical inference
- Bayesian versus frequentist hypothesis testing
- Measures of confidence: Posterior probability versus P-value

Independent experiments and sampling

Sampling: Let Ω be a finite set objects and A a subset. If we sample one object at random, we sample an object from A with probability

$$P(A) = \frac{\text{Number of elements in } A}{\text{Number of elements in } \Omega}$$

We denote the number of elements in a set A with $|A|$.

Independent experiments and sampling

Independence: Two experiments are independent if for each sets of outcomes A from Experiment 1 and B from Experiment 2 we have

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

Independent experiments and sampling

Sampling with replacement: Let Ω be a finite set of objects. If we sample one element with replacement, we draw one object at random from Ω and replace it with an identical object ('put' it back) $\Rightarrow$ Probability of sampling objects $1, ..., k$ with replacement from sets $A_1, \ldots, A_k$ is

$$P(A_1) \cdots P(A_k)$$

Binomial distribution

Consider an urn with red and green balls (= set of objects Ω with either trait A or B).

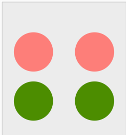
The frequency of red balls is $p(= \frac{|A|}{|\Omega|})$, the frequency of green balls is $(1-p)(= \frac{|B|}{|\Omega|})$.

We draw n times with replacement. Then the probability of drawing k red balls and $n-k$ green balls (in any order) is

$$P(k \text{ red balls}) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}.$$

Example calculation

Setup of urn: $p(\text{red ball}) = p(\text{green ball}) = 0.5$



Outcome of an experiment ($n = 5$)

Statistics as inductive logic

Bayesian statistics

Frequentist Hypothesis Testing

Statistics as inductive logic

Bayesian statistics

Frequentist Hypothesis Testing

- Statistics is applied inductive logic
- Statistics infers from a sample to the population
- Two goals parameter estimation of model and hypothesis testing
- Two main schools of testing procedures: Bayesian and frequentist testing
- Issues: Small sample size, correct data for hypothesis testing

Statistics as inductive logic

Bayesian statistics

Frequentist Hypothesis Testing

Example of a marble experiment

Setup:

- Flip a fair coin.
- If heads, place in an urn 1 white and 3 blue marbles
- If tails, place in an urn 3 white and 1 blue marbles

Hypothesis:

- H_B: 1 white and 3 blue marbles (WBBB)
- H_W: 3 white and 1 blue marbles (WWWB)

Purpose: To determine which hypothesis, H_B or H_W, is probably true.

Experiment: Mix the marbles, draw a marble, observe its color, and replace it, repeating this procedure as necessary.

Stopping rule: Stop when a hypothesis reaches a posterior probability of 0.999

Analysing the experiment with a Bayesian approach

Which hypothesis is more likely? $\rightarrow$ Use ratio form of Bayes's rule:

$$\frac{P(H_B|D)}{P(H_W|D)} = \frac{P(D|H_B)}{P(D|H_W)} \times \frac{P(H_B)}{P(H_W)}$$

Prior probabilities: $P(H_B) = P(H_W) = 0.5$

Likelihood function for one draw:

- H_B (WBBB): $P(blue|H_B) = 3/4 = 0.75$
- H_W (WWWB): $P(blue|H_W) = 1/4 = 0.25$

Likelihood for two draws (sampling with replacement):

- H_B (WBBB): $P(blue, blue|H_B) = (3/4)^2 = 0.5625$
- H_W (WWWB): $P(blue, blue|H_W) = (1/4)^2 = 0.0625$

In a $(n+1)$ draw experiment the posterior odds of the n-draw experiment is the prior to the $(n+1)$ draw experiment

Table 7.1. Bayesian analysis for a marble experiment, assuming prior odds for H_B:H_W of 1:1

Draw	Result	Posterior H_B:H_W	Posterior $P(H_B|D)$
	(Prior)	1:1	0.500000
1	White	1:3	0.250000
2	Blue	1:1	0.500000
3	White	1:3	0.250000
4	Blue	1:1	0.500000
5	Blue	3:1	0.750000
6	Blue	9:1	0.900000
7	Blue	27:1	0.964286
8	Blue	81:1	0.987805
9	Blue	243:1	0.995902
10	White	81:1	0.987805
11	Blue	243:1	0.995902
12	White	81:1	0.987805
13	Blue	243:1	0.995902
14	Blue	729:1	0.998630
15	Blue	2187:1	0.999543

Note: The final conclusion, at 15 draws, is that the probability that H_B is true is 0.999543, and that H_W is true is 0.000457. These values mean that H_B can be accepted with 99.9543% confidence of truth, or that the conclusion that H_B is true has a chance of error of only 1 in about 2,200.

How many trials before accepting a hypothesis?

- Quantity M: number of blue draws - number of white draws
- Since probabilities of H_B and H_W are 3:1, posterior odds of $H_B : H_W$ are $3^M : 1$
- Odds are $>$999:1 in favor of H_B when $M = 7$ (1:999$<$ when $M = -7$)
- $L \approx 2M$ trials are necessary for a given confidence level (on average).

Terminology for Bayesian statistics

$$\underbrace{\frac{P(H_B|D)}{P(H_W|D)}}_{\text{odds ratio}} = \underbrace{\frac{P(D|H_B)}{P(D|H_W)}}_{\text{Bayes factor}} \times \underbrace{\frac{P(H_B)}{P(H_W)}}_{\text{prior odds}}$$

- Bayes factor: Ratio of the likelihoods. How many times more probable is it to see the data under H_B than under H_W?
- Odds ratio: Ratio of the posterior probabilities. How many times more probable is H_B compared with H_W given the observed data
- If the odds ratio is smaller than 1 (e.g., 0.1), it is convenient to say that H_W is $\frac{1}{\text{odds ratio}}$ times more probable than H_B (e.g., $(0.1)^{-1}=10$ times more probable) given the data

Possible problems with the Bayesian approach

- Controversial background information
- Messy data
- Wrong hypotheses
- Different statistical methods

Statistics as inductive logic

Bayesian statistics

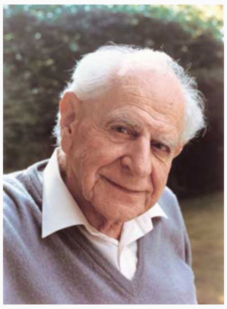
Frequentist Hypothesis Testing

The Frequentist Paradigm

- Bayesian paradigm preceded frequentist paradigm by ca. 200 years
- **Goal:** Get rid of Bayesian prior, make statistics more objective
- R. A. Fisher one of the key proponents of Frequentists
- Jerzy Neyman and E. S. Pearson developed frequentist hypothesis testing with an emphasis on falsification
- Karl Popper developed falsification theory
- Most scientists started to use frequentist paradigm

Pioneers of Frequentist analysis

R. A. Fisher (1890-1962)



Karl Popper (1902-1994)

Basics of frequentist hypothesis testing

Null Hypothesis (H_0): e.g.: There is no effect of various treatments, locus is in Hardy-Weinberg equilibrium,...

Alternative Hypothesis (H_1): e.g.: There is a treatment effect, locus is not in Hardy-Weinberg equilibrium

Null hypothesis is true or false

Statistical tests accept or reject the null hypothesis

To accept means to not reject

Hypothesis testing takes place in an explicit setup (e.g. fixed stopping rule, fixed hypotheses, ...)

Errors in testing

A frequentist's test chooses between H_0 and H_1. The decision is based on the sampled data and may not be correct. Two types of errors are possible:

	H_0	
Decision	True	False
Accept	Success	Type II Error
Reject	Type I Error	Success

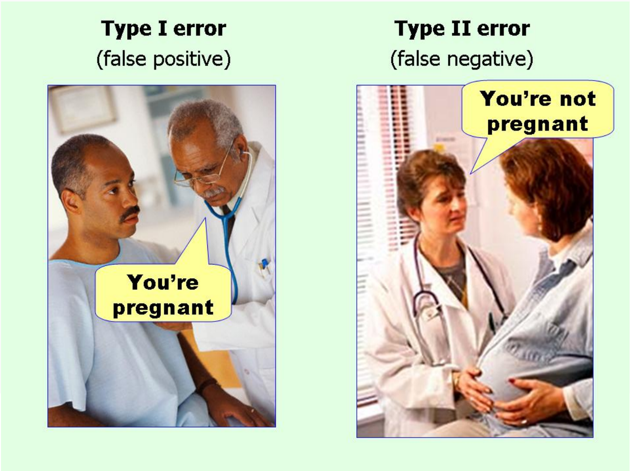
- False Positive (Type I Error) → Reject a null hypothesis even it is true.
- False Negative (Type II Error) → Accept a null hypothesis even it is false.

Example of false positives and negatives

H_0: A fungicide does not protect a plant (→ Treatment has no effect)

False positive error (Type I): Conclude that a fungicide protects a plant even it does not (H_0 is true)

False negative error (Type II: Conclude that a fungicide does not protect a plant even though it does (H_0 is false)



Source: Unknown

Example of false positives and negatives

H_0: Person is not pregnant

False positive error (Type I): Conclude that person is pregnant even so he is not (H_0 is true)

False negative error (Type II: Conclude that person is not pregnant even though she is (H_0 is false)

Trade-Off between Type I and Type II Error

- Avoid any Type I Error: Accept null hypothesis in any case
- Avoid any type II error: Reject null hypothesis in any case

$\rightarrow$ Both approaches are not useful

- Solution: Accept certain probability α of Type I error (significance level)
- Assign p value: Probability of a Type I error for the results a given experiment
- Small p - value: Strong rejection of null hypothesis.

How to calculate the p value: Repeat the experiment infinite times under the assumption that H_0 is true and find the probability of getting an outcome as extreme or more extreme than actual experimental outcome

A frequentist analysis of the urn experiment

- Null hypothesis H_W: Three white and one blue marble
- Alternative hypothesis H_B
- Arbitrary selection of a null hypothesis! But: Choice of null hypothesis influences test interpretation
- No assumption about prior
- Choose significance level α
- Describe outcome of experiments by test statistic $T =$ number of blue draws
- Hypothesis test: Calculate p-value as $P(T \geq t \mid H_W)$, where t is the actual value of T in the experiment. Reject H_W if $p \leq \alpha$

A frequentist analysis of the urn experiment

- 15 draws, the table shows the possible 16 outcomes of T
- Average of $15 \times 0.25 = 3.75$ draws of blue is expected given H_W
- Exact calculation with b, w and $n = b + w$ draws (given H_W):

$$p_b = \frac{n!}{b! \times w!} \times 0.25^b \times 0.75^w$$

$\Rightarrow$ Binomial distribution

- Probability for 5 blue and 10 white draws (given H_W):

$$\frac{15!}{5! \times 10!} \times 0.25^5 \times 0.75^{10}$$

- If t blue draws are observed, p-value is $p = \sum_{b \geq t} p_b$
- This test paradigm is called binomial test

A frequentist analysis of the urn experiment

Table 7.2. Frequentist analysis for a marble experiment, assuming that the null hypothesis H_W is true and the experiment stops at 15 draws

Blue draws	Probability	*p*-value
0	0.01336346101016	1.00000000000000
1	0.06681730505079	0.98663653898984
2	0.15590704511851	0.91981923393905
3	0.22519906517118	0.76391218882054
4	0.22519906517118	0.53871312364936
5	0.16514598112553	0.31351405847818
6	0.09174776729196	0.14836807735264
7	0.03932047169656	0.05662031006068
8	0.01310682389885	0.01729983836412
9	0.00339806545526	0.00419301446527
10	0.00067961309105	0.00079494901001
11	0.00010297168046	0.00011533591896*
12	0.00001144129783	0.00001236423850
13	0.00000088009983	0.00000092294067
14	0.00000004190952	0.00000004284084
15	0.00000000093132	0.00000000093132

Note: The conclusion, marked by an asterisk for the experimental outcome of 11 blue draws, is to reject H_W at a highly significant *p*-value of 0.000115. This *p*-value means that if experiments with H_W true were to be conducted numerous times, a result as extreme as 11 or more blue draws would occur with a frequency of only 0.000115, or only 1 in about 8,700 experiments.

Interpretation of the p-value table

- Assume in 15 draws we have 8 draws of the blue marble
- The exact probability to draw 8 marbles under H_W is 0.013106
- The probability to draw 8 marbles or more is 0.01729, calculated as the sum of the individual probabilities for 8 to 15 blue draws.
- This sum is the p-value
- If we assign $\alpha = 0.05$ we can state that if we draw 8 or more blue marbles, we reject H_W because p-value $< \alpha$

p-values vs. effect sizes

- A highly significant test result (=very low p-value) just gives the information that we can distinguish very well between the null and alternative hypotheses using our inference rule (=test) and our data
- It does not state whether the difference between the hypotheses is very big or not
- Given enough data, even very minor differences between hypotheses can be highly significant

p-values vs. effect sizes

Example: Consider two sets of two urns:

xperiment 1: Urn A1 has 9 blue marbles and 1 white marble (H_1), Urn A2 has 1 blue marble and 9 white marbles (H_0). In 3 draws three 3 blue marbles were obtained.
$\Rightarrow$ p-value: $(0.1)^3 = 0.001$

xperiment 2: Urn B1 has 6 blue and 4 white marbles (H_1), Urn B2 has 5 blue and 5 white marbles (H_0). In 12 draws 12 blue marbles were obtained.
$\Rightarrow$ p-value: $(0.5)^{12} \approx 0.00024$

- Chance to make a Type I error is lower in the second experiment
- But: Difference between the hypotheses is much smaller in the second experiment!

Issues: 'Bayesian' misreading of frequentist hypothesis test

- Bayesian testing: $P(H|D)$
- Frequentist testing: $P(D|H)$
- Frequentist testing does not say which hypothesis is more probable

Issues: The influencal stopping rule

- A *p* value depends on the data *and* the stopping rule
- Example:
  - Stop at 15 draws $\rightarrow$ 16 outcomes
  - Stop at 11 blue draws $\rightarrow \infty$ outcomes
  - Stop at 4 white draws $\rightarrow \infty$ outcomes
- Same data but different outcomes!
- Computer software never asks for the stopping rule (assumptions are usually implemented)

Comparison of Bayesian and frequentist approaches

- **Bayesian statistics:** What is the probability that a hypothesis is true, given the data and any prior knowledge?
- **Frequentist statistics:** How reliable is an inference procedure, by virtue of not rejecting a true hypothesis or accepting a false hypothesis?

Bayesian vs. frequentist statistics

Bayesian statistics:

1. Answers the question about probability of hypotheses given data directly
2. Compares several hypotheses
3. Needs a prior distribution
4. Does not need a stopping rule

Frequentist statistics:

1. An inference scheme which has a certain quality: Controlled type I error
2. Can compare hypotheses, but also test whether a null hypothesis can be rejected without specifying an alternative ($H = H_0$ vs. $H \neq H_0$)
3. Does not need a prior
4. Needs a stopping rule

Summary

- Induction ≈ statistics
- Bayesian statistics
- Frequentist statistics
- Frequentist vs. Bayesian statistics

Further reading

- Hugh G. Gauch (2012) - Scientific Method in Brief. Cambridge University Press, Chapter 9
- Colquhoun (2017) An investigation of the fale discovery rate and the misinterpretation of p-values. R. Soc. Open Sci. 1:140216
- Nuzzo (2014) Statistical Errors. Nature 506:150-152
- Lakens (2022) Why P values are not measures of evidence. Trends in Ecology and Evolution 37:289-290
- Nature Methods: Statistics for biologists. Excellent collection of short essays on different aspects of statistics. `http://www.nature.com/collections/qghhqm`

Study questions i

- List the pros and cons for Bayesian and frequentist statistics
- We redo the marble experiment. This time, we only draw 4 times and we see the following sample: blue,blue,blue,white. Do a Bayesian and a frequentist's test for deciding whether this sample gives evidence for H_W or H_B. For the Bayesian test, just give the posterior probability ratio, for the frequentist's test compute the p-value

Study questions ii

- We want to test (in the frequentist's way) whether a locus is in Hardy-Weinberg equilibrium and formulate the hypotheses

  H_0: Locus is in HWE

  H_1: Locus is not in HWE

  What are the possible Type I and Type II errors for these hypotheses? Remark: You don't have to formulate a test, just describe the possible errors.

References i

Colquhoun, D. (2017). The Reproducibility Of Research And The Misinterpretation Of P Values. *bioRxiv*, page 144337. 00000.

Hugh G. Gauch, J. (2012). *Scientific Method in Brief*. Cambridge University Press.

Lakens, D. (2022). Why P values are not measures of evidence. *Trends in Ecology & Evolution*, 37(4):289–290.

Nuzzo, R. (2014). Scientific method: Statistical errors. *Nature News*, 506(7487):150.