Statistical concepts in a nutshell

Statistics: dream job of next decade

Content

Approaches to statistics

- Frequentist
  - Neyman, Pearson, Wald, etc.
  - Sampling is infinite and decision rules can be sharp
- Bayesian
  - Bayes, Laplace, de Finetti, etc.
  - Unknown quantities are treated probabilistically and the state of the world can always be updated
- Likelihood (Bayesian who do not know it)
  - Fisher, etc.
  - Single sample inference based on maximizing the likelihood function

"Statistical concepts in a nutshell"

Frequentist | What is fixed?

- Data are a repeatable random sample
- Studies are repeatable and allow for frequency interpretation
- Underlying parameters remain fixed during repeatable sampling
- **Parameters are fixed**

Bayesian | What is fixed?

- Data are observed from realized sample
- Studies are fixed
- Parameters are unknown and described probabilistically
- **Data are fixed**

Frequentist | Inference

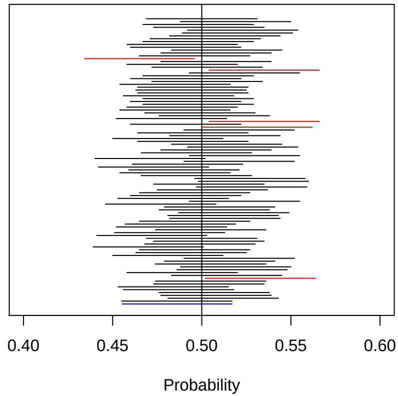
- Point estimates and standard error or 95% confidence intervals are a repeatable random sample
- Deduction from sampling distribution of the data $\Pr(\mathcal{D} \mid \mathcal{H}_0)$ with prefixed significance level α
- Accept $\mathcal{H}_1$ if $\Pr(\mathcal{D} \mid \mathcal{H}_0) < \alpha$
- Accept $\mathcal{H}_0$ if $\Pr(\mathcal{D} \mid \mathcal{H}_0) \geq \alpha$

Bayesian | Inference

- Induction from posterior distribution $p(\theta \mid \mathcal{D})$ starting with prior $p(\theta)$
- Summarize posterior distribution (e.g. using means and quantiles)
- Highest posterior density interval indicates region of highest posterior probability

95% confidence intervals

- Frequentist
  - In repeated sampling, 95% of the realized confidence intervals cover the true parameter value
- Bayesian
  - For these data, the parameter is in the interval with 95% probability
- Example: Coin tossing



Why is frequentist statistics so popular?

- The beginning
  - From 1950–1990 nobody did Bayesian statistics (only a small minority)
  - Automated 'cookbook' type procedures were in demand or it is what statistician sold
  - Computers were slow and relatively unavailable
  - Bayesian statistics need lots of computation
- Bayesian revolution
  - Gelfand and Smith (1990) reveal the potential of Gibbs sampling for conventional statistical problems
  - Development of computational algorithms for complex models and practical problems
  - Cracked open countless problems with an avalanche of articles during 1990–2005

Use the complete statistical toolbox

- Frequentist
  - Assumption that data come from a controlled repeatable experiment
  - Nothing is more important than repeatability, no matter what
  - Repeatability can be important
- Bayesian
  - Very careful about assumption and willing to defend it
  - Statistical models are subjective, not a problem to admit that
  - Inference can be appropriate

Content

Fisher's p-value

- Single hypothesis $\mathcal{H}$ assumed to be true
- Test statistic is used to measure the discrepancy between what is observed from data and what is expected under $\mathcal{H}$
  - Equal population means: $\bar{X} - \bar{Y}$

    Large values indicate a discrepancy between the observed data and the hypothesis that the population means are equal
  - Equal population variances: S_X^2 / S_Y^2

    Large or small values indicate a discrepancy between the observed data and the hypothesis that the population variances are equal

Fisher's p-value | Example

- Hypothesis $\mathcal{H}$: average height is 165cm in population
- Collected data $\mathcal{D}$: average height is 169cm

Is something wrong with the hypothesis or were we just lucky to get such an extreme data set?

- Summarize the data into a single number

$$t = \frac{169 - 165}{10/\sqrt{50}} = 3.05$$

- A large absolute value $|t|$ indicates a discrepancy between what we observed in the data set (average height 169cm) and what we expected to see in the data set if the average height were 165cm in the population
- Compute the p-value

  Small p-value means that it is unlikely to get a data set in which the average height is 169cm if the average height were truly 165cm

Fisher's p-value | Example

1. Identify the null hypothesis $\mathcal{H}$ and a test statistic
2. Calculate the value of test test statistic from the observed data
3. Determine the p-value
4. Reject the null hypothesis $\mathcal{H}$ if the p-value is sufficiently small (otherwise reach no conclusion)

$\rightarrow$ The hypothesis $\mathcal{H}$ is not rejected if it is compatible with observed data (but that does not proof that $\mathcal{H}$ is true)

Neyman-Pearson approach

- Two hypotheses
  - Null hypothesis $\mathcal{H}_0$
  - Complementary hypothesis $\mathcal{H}_1$
- A decision is prescribed
  - Rejection of $\mathcal{H}_0$ or $\mathcal{H}_1$ and acceptance of the other
- Errors occur if $\mathcal{H}_0$ is rejected or accepted when it is actually true or false

	$\mathcal{H}_0$ **is true**	$\boldsymbol{\mathcal{H}_0}$ **is false (** $\boldsymbol{\mathcal{H}_1}$ **is true)**
Reject $\boldsymbol{\mathcal{H}_0}$	Type I error (α)	Correct
Accept $\boldsymbol{\mathcal{H}_0}$ **(Reject** $\boldsymbol{\mathcal{H}_1}$**)**	Correct	Type II error (β)

Neyman-Pearson approach | Example

- Null hypothesis $\mathcal{H}_0$: average height is 165cm in population
- Collected data $\mathcal{D}$: average height is 169cm

Is something wrong with the hypothesis or were we just lucky to get such an extreme data set?

- Summarize the data into a single number

$$t = \frac{169 - 165}{10/\sqrt{50}} = 3.05$$

- A large absolute value $|t|$ indicates a discrepancy between what we observed in the data set (average height 169cm) and what we expected to see in the data set if the average height were 165cm in the population
- Fix the number of times we allow the null hypothesis $\mathcal{H}_0$ to be rejected if it is actually true (e.g. 5%).
- Compute the critical value

  Reject the null hypothesis if the value of test statistic is larger than critical value and accept the complementary hypothesis

Neyman-Pearson's recipe

1. Identify a null hypothesis $\mathcal{H}_0$ and its complement $\mathcal{H}_1$
2. Determine a test statistic
3. Specify the significance level α and compute the respective critical value c under $\mathcal{H}_0$
4. Calculate the value of the test statistic from the observed data
5. Reject null hypothesis $\mathcal{H}_0$ if the test statistic is larger than critical value (accept $\mathcal{H}_0$ otherwise)

$\rightarrow$ There exists no summary of the evidence provided by the data with respect to the hypotheses

Application of Neyman-Pearson's approach results in a decision about rejection or acceptance of $\mathcal{H}_0$ such that the number of wrong decisions in repeated experiments is controlled for

Jeffreys' Bayes factor

- Two hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$
- Hypotheses are treated probabilistically
  - Prior probabilities $\Pr(\mathcal{H}_0)$ and $\Pr(\mathcal{H}_1) = 1 - \Pr(\mathcal{H}_0)$
- Posterior probability of $\mathcal{H}_0$ and $\mathcal{H}_1$ conditional on observed data

$$\Pr(\mathcal{H}_0 \,|\, \mathcal{D}) = \frac{p(\mathcal{D} \,|\, \mathcal{H}_0)\Pr(\mathcal{H}_0)}{p(\mathcal{D})} \quad \text{and} \quad \Pr(\mathcal{H}_1 \,|\, \mathcal{D}) = \frac{p(\mathcal{D} \,|\, \mathcal{H}_1)\Pr(\mathcal{H}_1)}{p(\mathcal{D})}$$

- Ratio of the posterior probabilities (posterior odds)

$$\frac{\Pr(\mathcal{H}_1 \,|\, \mathcal{D})}{\Pr(\mathcal{H}_0 \,|\, \mathcal{D})} = \underbrace{\frac{p(\mathcal{D} \,|\, \mathcal{H}_1)}{p(\mathcal{D} \,|\, \mathcal{H}_0)}}_{\text{Bayes factor } B_{10}} \times \underbrace{\frac{\Pr(\mathcal{H}_1)}{\Pr(\mathcal{H}_0)}}_{\text{Prior odds}}$$

- Posterior odds can be converted to probability scale

$$p(\mathcal{H}_1 \,|\, \mathcal{D}) = \frac{PO_{10}}{1 + PO_{10}} \quad \text{and} \quad p(\mathcal{H}_0 \,|\, \mathcal{D}) = \frac{1}{1 + PO_{10}}$$

Jeffreys' scale

- Bayes factor quantifies the update from the prior odds to the posterior odds in the light of observed data
- Hypothesis $\mathcal{H}_1$ is rejected if $B_{10} \leq 1$ (and accepted, otherwise)
- Jeffreys' scale gives strength of the evidence provided by observed data in favor of $\mathcal{H}_1$ compared to that of $\mathcal{H}_0$

		B_{10}	$\leq$	1/10	Strong evidence for $\mathcal{H}_0$
1/10	$<$	B_{10}	$\leq$	1/3	Moderate evidence for $\mathcal{H}_0$
1/3	$<$	B_{10}	$\leq$	1	Weak evidence for $\mathcal{H}_0$
1	$<$	B_{10}	$\leq$	3	Weak evidence for $\mathcal{H}_1$
3	$<$	B_{10}	$\leq$	10	Moderate evidence for $\mathcal{H}_1$
10	$<$	B_{10}			Strong evidence for $\mathcal{H}_1$

Jeffreys' scale

- Bayes factor is the ratio of prior predictive distributions
- Prior predictive distribution under $\mathcal{H}_i$

  $$p(\mathcal{D} \,|\, \mathcal{H}_i) = \int p(\mathcal{D} \,|\, \theta_i, \mathcal{H}_i) p(\theta_i \,|\, \mathcal{H}_i) \,\mathrm{d}\theta_i$$

  Interpretation:

  Averaged likelihood function $p(\mathcal{D} \,|\, \theta_i, \mathcal{H}_i)$ over all possible parameter choices weighted by the prior $p(\theta_i \,|\, \mathcal{H}_i)$
- Prior predictive distribution indicates how likely data is to be observed before collecting it
- Closed-form expression are not always possible, but Monte Carlo methods and asymptotic approximations (e.g. Laplace's method or BIC) are helpful

Summary

- Fisher's p-value
  - Summarizes the evidence provided by observed data against the hypothesis $\mathcal{H}$
  - Does not yield a probabilistic statement about whether hypothesis $\mathcal{H}$ is true
- Neyman-Pearson approach
  - Results in a decision about rejection or acceptance of null hypothesis $\mathcal{H}_0$ such that the number of wrong decisions in repeated experiments is controlled for
  - Does not contain a summary of the evidence provided by observed data for or against the hypotheses
- Jeffreys' Bayes factor
  - Allows probabilistic statements about the truth of hypotheses
  - Required computations and prior specifications may be difficult though