Statistical genetics:
statistical concepts in a nutshell

Content

1. Motivation
2. Approaches to statistics
3. Hypothesis testing
4. Model comparison
5. SNP association studies

Motivation

Statistics: dream job of the next decade (Val Harian from Google)

- Massive data is generated for a relatively low price
- Statisticians help all sorts of people with data analyses
- Requires analytic talent and extensive knowledge of statistical tools

Available courses

- Computational statistics
- Markovian modelling and Bayesian learning
- Bayesian theory with applications
- Statistical software tools: high performance computing
- Probabilistic Models

Content

1. Motivation
2. **Approaches to statistics**
3. Hypothesis testing
4. Model comparison
5. SNP association studies

Approaches to statistics

What is statistics?

- Statistics is the study of uncertainty
- There is uncertainty in data and parameters
- Statisticians help people deal with uncertainty

Frequentist

- Probability is interpreted as a long-run expected frequency
- Only probabilistic statements about repeatable events with uncertainty induced by randomness are permitted

Bayesian

- Probability is interpreted as a personal degree of belief
- Probabilistic statements about events with uncertainty induced by lack of knowledge are also permitted

Approaches to statistics

Frequentist

- Parameter θ is assumed to be fixed to some unknown value
- Observed data $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ is a repeatable random sample from a sampling distribution
- Confidence interval estimation and hypothesis tests have long-run expected frequency interpretation

Bayesian

- Parameter θ is a random variable with prior $p(\theta)$
- Probabilistic statements about θ are conditional on observed data $\boldsymbol{y}$
- Bayes' theorem updates the prior to the posterior in the light of observed data $\boldsymbol{y}$
- Credible intervals and hypothesis test are interpreted as desired

Approaches to statistics

Confidence interval for a population mean μ

- $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$ are iid Normal(μ, σ^2)

Solution 1: Frequentist confidence interval

- The test statistic $t(\boldsymbol{Y}) = (\bar{Y} - \mu_0)/(S/\sqrt{n})$ follows a Student's-t distribution with $n-1$ degrees of freedom
- $1-\alpha$ confidence interval for μ is obtained by

$$\begin{aligned} 1-\alpha &= \mathcal{P}\left(-t_{1-\alpha/2} \leq \frac{\bar{Y}-\mu}{S/\sqrt{n}} \leq t_{1-\alpha/2}\right) \\ &= \mathcal{P}\left(\bar{Y} - t_{1-\alpha/2}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{Y} + t_{1-\alpha/2}\frac{S}{\sqrt{n}}\right) \end{aligned} \tag{1}$$

- The long-run expected frequency that the confidence interval covers μ is $1-\alpha$

Approaches to statistics

Solution 1: Frequentist confidence interval cont'd

- For observed data $\boldsymbol{y}$, the $1 - \alpha$ confidence interval either covers the unknown value of μ or not
- So what? At least there is a $1 - \alpha$ probability of getting data $\boldsymbol{y}$ for which the confidence interval covers μ, but you will never know

```
freqInterval <- function(D, mean) {
   int <- mean(D) + qt(c(0.05 / 2, 1 - 0.05 / 2), length(D) - 1) *
      sd(D) / sqrt(length(D));
   return(int[1] < mean & int[2] > mean)
}
nData <- 10000;
data <- matrix(rnorm(nData * 50, mean = 1, sd = 0.5), nData, 50);
1 - mean(apply(data, 1, freqInterval, mean = 1))

# [1] 0.0465
```

Approaches to statistics

Solution 2: Bayesian credible interval

- The likelihood is

$$p(\boldsymbol{y} \,|\, \mu, \sigma^2) \propto \sigma^{-n} \exp\left\{ -\frac{(n-1)s^2 + n(\bar{y} - \mu)^2}{2\sigma^2} \right\}$$

- A convenient prior is

$$p(\mu, \sigma^2) \propto \sigma^{-2}$$

- The posterior is then

$$\begin{aligned} p(\mu, \sigma^2 \,|\, \boldsymbol{y}) &\propto p(\boldsymbol{y} \,|\, \mu, \sigma^2) p(\mu, \sigma^2) \\ &= \sigma^{-n-2} \exp\left\{ -\frac{(n-1)s^2 + n(\bar{y} - \mu)^2}{2\sigma^2} \right\} \end{aligned}$$

Approaches to statistics

Solution 2: Bayesian credible interval cont'd

- The marginal of μ is

$$
\begin{aligned}
p(\mu \,|\, \boldsymbol{y}) &\propto \int_0^\infty p(\mu, \sigma^2 \,|\, \boldsymbol{y}) \, \mathrm{d}\sigma^2 \\
&\propto \int_0^\infty \frac{1}{(\sigma^2)^{n/2+1}} \exp\left\{ -\frac{(n-1)s^2 + n(\mu - \bar{y})^2}{2\sigma^2} \right\} \mathrm{d}\sigma^2
\end{aligned}
$$

- Change of variable:

$$
x = \frac{(n-1)s^2 + n(\mu - \bar{y})^2}{\sigma^2} \Leftrightarrow \sigma^2 = \frac{(n-1)s^2 + n(\mu - \bar{y})^2}{x}
$$

Approaches to statistics

Solution 2: Bayesian credible interval cont'd

- The marginal of μ is then

$$
\begin{aligned}
p(\mu \,|\, \boldsymbol{y}) \propto & \int_0^\infty \left[\frac{x}{(n-1)s^2 + n(\mu-\bar{y})^2}\right]^{n/2+1} \times \\
& \qquad \exp\left\{-\frac{x}{2}\right\} \left|\frac{\mathrm{d}\sigma^2}{\mathrm{d}x}\right| \mathrm{d}x \\
& = \left[(n-1)s^2 + n(\mu-\bar{y})^2\right]^{-n/2} \underbrace{\int_0^\infty x^{n/2-1} e^{-x/2}\,\mathrm{d}x}_{\text{Kernel of a } \chi^2_n \text{ distribution}} \\
& \propto \left[1 + \frac{1}{n-1}\frac{(\mu-\bar{y})^2}{s^2/n}\right]^{-n/2}
\end{aligned}
$$

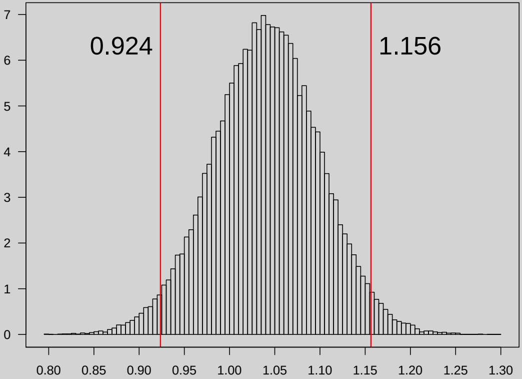
a Student's-t distribution with $n-1$ degrees of freedom, location parameter $\bar{y}$ and scale parameter s^2/n

Approaches to statistics

Solution 2: Bayesian credible interval cont'd

- For observed data $\boldsymbol{y}$, the $1 - \alpha$ credible interval contains the parameter μ with probability $1 - \alpha$

```
D <- rnorm(n = 50, mean = 1, sd = 0.5);
Dbar <- mean(D); Dsd <- sd(D); n <- length(D);
u <- rt(50000, n - 1) * Dsd / sqrt(n) + Dbar;
credInt <- quantile(u, prob = c(0.025, 0.975));
```



Approaches to statistics

Bayesian modeling recipe

1. Chose a parametric model indexed by parameters $\boldsymbol{\Theta}$ from which the observed data $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ are thought to originate
2. Chose priors for the parameters $\boldsymbol{\Theta}$
3. Update the prior to the posterior by means of Bayes' rule

$$p(\boldsymbol{\theta} \,|\, \boldsymbol{y}) = \frac{p(\boldsymbol{y}, \boldsymbol{\theta})}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{y} \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} p(\boldsymbol{y} \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}}$$

4. Check the model fit and sensitivity of the prior

Where do priors come from?

- Previous studies and published research
- Expert knowledge or researcher's intuition
- Conjugacy or uninformative priors

Approaches to statistics

If Bayesian statistics is...

- "What you think that classical statistics is"
- "The only good statistics"

Why has not everybody used it by now?

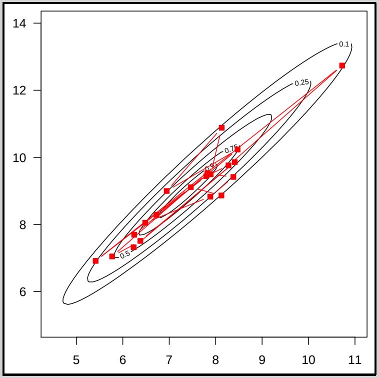
A sketch of the history of statistics

19th century	Astronomy/physics	Bernoulli, Laplace, Poisson, Legendre, Gauss, Gibbs
20th century	Biology	Venn, Fisher, Neyman, Wald
		Cookbook statistics and slow computers
1990 - today	Everything changed...	

Approaches to statistics

The Bayesian revolution

1950's	Metropolis sampler
1970's	Metropolis Hastings sampler
1980's	Gibbs sampler
1990's - today	Computationally intensive methods for practical problems with complex models



Approaches to statistics

Summary

- There exists frequentist and Bayesian statistics
- The results from these approaches are interpreted differently
- Powerful enough computers made Bayesian computations feasible
- In the end, use all statistical tools that help solve a problem

Content

1. Motivation
2. Approaches to statistics
3. **Hypothesis testing**
4. Model comparison
5. SNP association studies

Hypothesis testing

NEYMAN PEARSON FISHER JEFFREYS

(1930's) (1920's) (1960's)

Hypothesis testing

Fisher's p-value approach

- There is only a single hypothesis $\mathcal{H}$, which is assumed to be true, and a test statistic $t(\boldsymbol{Y})$
- The test statistic is used to measure the discrepancy between what is observed from collected data and what is expected under $\mathcal{H}$
    - Equal population means: $t(\boldsymbol{X}, \boldsymbol{Y}) = \bar{X} - \bar{Y}$. Large values of $|t(\boldsymbol{x}, \boldsymbol{y})|$ represent a discrepancy between the observed data and $\mathcal{H}$ that the population means are equal
    - Equal population variances: $t(\boldsymbol{X}, \boldsymbol{Y}) = S_X^2 / S_Y^2$. Large or small values of $t(\boldsymbol{x}, \boldsymbol{y})$ represent a discrepancy between the observed data and $\mathcal{H}$ that the population variances are equal

Hypothesis testing

Fisher's p-value approach cont'd

- The test statistic $t(\boldsymbol{Y})$ is a function of the random sample and therefore a random variable
- The distribution of the test statistic $t(\boldsymbol{Y})$ may be determined under the assumption that $\mathcal{H}$ is true
- The p-value

  $$\mathcal{P}(t(\boldsymbol{Y} \geq t(\boldsymbol{y})\,|\,\mathcal{H} \text{ is true})$$

  is the probability of obtaining a discrepancy as represented by $t(\boldsymbol{Y})$ greater than or equal to the observed value $t(\boldsymbol{y})$ if $\mathcal{H}$ is true
- Small p-values are equivalent with large discrepancies between what is observed from collected data and what is expected under $\mathcal{H}$

Hypothesis testing

Cohan J.: The earth is round ($p < .05$)

”[Null hypothesis significance testing] does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does! What we want to know is ’Given these data, what is the probability that $\mathcal{H}$ is true?’ But as most of us know, what it tells us is ’Given that $\mathcal{H}$ is true, what is the probability of these (or more extreme) data?’”

Hypothesis testing

Cohan J.: The earth is round ($p < .05$)

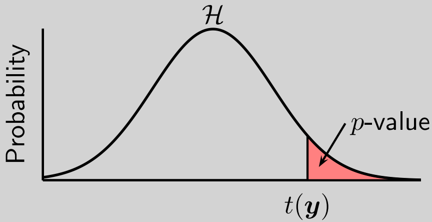
”[Null hypothesis significance testing] does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does! What we want to know is 'Given these data, what is the probability that $\mathcal{H}$ is true?' But as most of us know, what it tells us is 'Given that $\mathcal{H}$ is true, what is the probability of these (or more extreme) data?'”

Hypothesis testing

FISHER

Fisher's recipe

1. Identify the null hypothesis $\mathcal{H}$ and a test statistic $t(\boldsymbol{Y})$ with its distribution assuming that $\mathcal{H}$ is true
2. Calculate $t(\boldsymbol{y})$ from the observed data $\boldsymbol{y}$
3. Determine the p-value $\mathcal{P}(t(\boldsymbol{Y}) \geq t(\boldsymbol{y}) \mid \mathcal{H}$ is true)



4. Reject $\mathcal{H}$ if the p-value is sufficiently small and otherwise reach no conclusion

→ The hypothesis $\mathcal{H}$ is not rejected if it is compatible with observed data $\boldsymbol{y}$, but that does not proof that $\mathcal{H}$ is true

Hypothesis testing

Neyman-Pearson decision-theoretic aproach

- There are two hypotheses, namely a null hypothesis $\mathcal{H}_0$ and its complement $\mathcal{H}_1$
- A decision is prescribed, that is, rejection of either of the two hypotheses and acceptance of the other
- Errors occur if $\mathcal{H}_0$ is respectively rejected and accepted when it is actually true and false

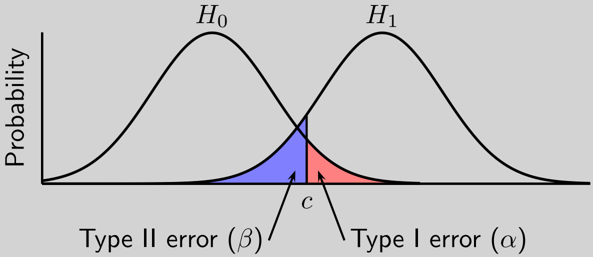
	$\mathcal{H}_0$ is true	**$\mathcal{H}_0$ is false ($\mathcal{H}_1$ is true)**
Reject $\mathcal{H}_0$	Type I error (α)	Correct
Accept $\mathcal{H}_0$ **(Reject $\mathcal{H}_1$)**	Correct	Type II error (β)

Hypothesis testing

NEYMAN PEARSON

Neyman-Pearson decision-theoretic approach cont'd

- Similar to Fisher's p-value approach, a test statistic $t(\boldsymbol{Y})$ is chosen and its distribution determined under the assumption that $\mathcal{H}_0$ is true
- The probability α of rejecting $\mathcal{H}_0$ when it is actually true is specified and the respective critical value c calculated
- The critical value c determines for which values of the test statistic $t(\boldsymbol{Y})$ the null hypothesis $\mathcal{H}_0$ is rejected or accepted



Hypothesis testing

Rozeboom W.: The fallacy of the null-hypothesis significance test

”The null-hypothesis significance test treats ’acceptance’ or ’rejection’ of a hypothesis as though these were decisions one makes. But a hypothesis is not something, like a piece of pie offered for dessert, which can be accepted or rejected by a voluntary physical action. Acceptance or rejection of a hypothesis is a cognitive process, a degree of believing or disbelieving which, if rational, is not a matter of choice but determined solely by how likely it is, given the evidence, that the hypothesis is true.”

Hypothesis testing

Rozeboom W.: The fallacy of the null-hypothesis significance test

”The null-hypothesis significance test treats ’acceptance’ or ’rejection’ of a hypothesis as though these were decisions one makes. But a hypothesis is not something, like a piece of pie offered for dessert, which can be accepted or rejected by a voluntary physical action. Acceptance or rejection of a hypothesis is a cognitive process, a degree of believing or disbelieving which, if rational, is not a matter of choice but determined solely by how likely it is, given the evidence, that the hypothesis is true.”

Hypothesis testing

NEYMAN

PEARSON

Neyman-Pearson's recipe

1. Identify the null hypothesis $\mathcal{H}_0$ and its complement $\mathcal{H}_1$
2. Determine a test statistic $t(\boldsymbol{Y})$ and its distribution assuming that $\mathcal{H}_0$ is true
3. Specify the significance level α and compute the respective critical value c under $\mathcal{H}_0$
4. Calculate $t(\boldsymbol{y})$ from the observed data $\boldsymbol{y}$
5. Reject $\mathcal{H}_0$ if the test statistic $t(\boldsymbol{y}) > c$ and accept $\mathcal{H}_0$ otherwise

→ There exists no summary of the evidence provided by the data with respect to the hypotheses at hand. Application of Neyman-Pearson's decision-theoretic 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.

Hypothesis testing

Neyman-Pearson's frequentist approach

```
freqTest <- function(D, mu0, alpha, c) {
   t <- (mean(D) - mu0) / sd(D) * sqrt(length(D))
   return(-c < t & c > t)
}
nData <- 10000;
data <- matrix(rnorm(nData * 50, mean = 1, sd = 0.5), nData, 50);
c <- qt(1 - 0.05 / 2, n - 1)
1 - mean(apply(data, 1, freqTest, mu0 = 1, alpha = 0.05, c = c))

# [1] 0.0527
```

Hypothesis testing

Jeffreys' Bayes factor approach

- There are two hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$
- The fundamental difference to everything before is that prior probabilities $p(\mathcal{H}_0)$ and $p(\mathcal{H}_1) = 1 - p(\mathcal{H}_0)$ are respectively assigned to $\mathcal{H}_0$ and $\mathcal{H}_1$
- The posterior probability of $\mathcal{H}_0$ and $\mathcal{H}_1$ conditional on observed data $\boldsymbol{y}$ is respectively

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

- The ratio of the posterior probabilities, also called posterior odds, is

$$\frac{p(\mathcal{H}_1 \,|\, \boldsymbol{y})}{p(\mathcal{H}_0 \,|\, \boldsymbol{y})} = \frac{p(\boldsymbol{y} \,|\, \mathcal{H}_1)}{p(\boldsymbol{y} \,|\, \mathcal{H}_0)} \times \frac{p(\mathcal{H}_1)}{p(\mathcal{H}_0)}$$

Jeffreys' Bayes factor approach cont'd

- The ratio of the posterior probabilities, also called posterior odds, is

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

  where the first term on the right hand side denotes the Bayes factor B_{10} and the second term designates the prior odds
- The posterior odds can be converted to probability scale by computing

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

Hypothesis testing

Jeffreys' Bayes factor approach cont'd

- The Bayes factor $B_{10} = p(\boldsymbol{y} \,|\, \mathcal{H}_1)/p(\boldsymbol{y} \,|\, \mathcal{H}_0)$ represents the update from the prior odds of the two hypotheses to the posterior odds in the light of observed data $\boldsymbol{y}$
- Hypothesis $\mathcal{H}_1$ is rejected if $B_{10} \leq 1$ and accepted, otherwise
- The strength of the evidence provided by observed data $\boldsymbol{y}$ in favor of $\mathcal{H}_1$ compared to that of $\mathcal{H}_0$ is evaluated by means of Jeffreys' scale

		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 $\mathcal{H}_1$

Hypothesis testing

Prior predictive distribution

- The Bayes factor B_{10} is the ratio of the prior predictive distribution under $\mathcal{H}_1$ and $\mathcal{H}_0$
- The prior predictive distribution under $\mathcal{H}_i$

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

  is obtained by averaging the likelihood function $p(\boldsymbol{y} \,|\, \theta_i, \mathcal{H}_i)$ over all possible parameter choices weighted by the prior $p(\theta_i \,|\, \mathcal{H}_i)$
- The prior predictive distribution indicates how likely data $\boldsymbol{Y}$ is to be observed prior to collecting it
- Obtaining a closed-form expression is not always possible, but Monte Carlo methods and asymptotic approximations such as Laplace's method or the Bayesian Information Criterion are a remedy to this problem

Hypothesis testing

Summary

- Fisher's p-value summarizes the evidence provided by observed data $\boldsymbol{y}$ against $\mathcal{H}$. It does not result in a probabilistic statement about whether $\mathcal{H}$ is true
- Neyman-Pearson decision-theoretic approach does not contain a summary of the evidence provided by the data with respect to the hypotheses at hand. Their 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 approach allows probabilistic statements about the truth of hypotheses. The required computations and prior specifications may be difficult though

Content

1. Motivation
2. Approaches to statistics
3. Hypothesis testing
4. **Model comparison**
5. SNP association studies

Model comparison

Likelihood ratio test

- The null hypothesis $\mathcal{H}_0$ states that model $\mathcal{M}_0$ fits the observed data $\boldsymbol{y}$ as well as $\mathcal{M}_1$ does, while the alternative hypothesis $\mathcal{H}_1$ states the opposite
- The test statistic is defined as

$$t(\boldsymbol{y}) = -2\log\left\{\frac{p(\boldsymbol{y}\,|\,\hat{\boldsymbol{\theta}}_0, \mathcal{M}_0)}{p(\boldsymbol{y}\,|\,\hat{\boldsymbol{\theta}}_1, \mathcal{M}_1)}\right\},$$

where $\mathcal{M}_0$ and $\mathcal{M}_1$ are respectively the simpler (null) and more complex (alternative) model and $\hat{\boldsymbol{\theta}}_0$ and $\hat{\boldsymbol{\theta}}_1$ the corresponding maximum likelihood estimates

Model comparison

Likelihood ratio test

- The distribution of the test statistic

$$t(\boldsymbol{y}) = -2\log\left\{\frac{p(\boldsymbol{y}\,|\,\hat{\boldsymbol{\theta}}_0, \mathcal{M}_0)}{p(\boldsymbol{y}\,|\,\hat{\boldsymbol{\theta}}_1, \mathcal{M}_1)}\right\},$$

  under the assumption that $\mathcal{H}_0$ is true, converges against a $\chi^2(\nu)$ distribution as n approaches ∞ with ν being the difference in the number of parameters between $\mathcal{M}_1$ and $\mathcal{M}_0$
- The observed value $t(\boldsymbol{y})$ of the test statistic is calculated and the hypothesis $\mathcal{H}_0$ that model $\mathcal{M}_0$ fits the data as well as $\mathcal{M}_1$ does is rejected if $t(\boldsymbol{y}) > c$

Model comparison

χ^2 approximation to the likelihood ratio for a simple $\mathcal{H}_0$

- $\mathcal{H}_0 : \theta = \theta_0$ and $\mathcal{H}_1 : \theta \neq \theta_0$
- Expand the log-likelihood $\ell(\theta_0 \,|\, \boldsymbol{y}) = p(\boldsymbol{y} \,|\, \theta_0)$ as a second-order Taylor series around the maximum likelihood estimate $\hat{\theta}$

$$\ell(\theta_0 \,|\, \boldsymbol{y}) \approx \ell(\hat{\theta} \,|\, \boldsymbol{y}) + \ell'(\hat{\theta} \,|\, \boldsymbol{y})(\theta_0 - \hat{\theta}) + \ell''(\hat{\theta} \,|\, \boldsymbol{y})(\theta_0 - \hat{\theta})^2/2$$

- Plug the expansion into $t(\boldsymbol{y}) = -2\ell(\theta_0 \,|\, \boldsymbol{y}) + 2\ell(\hat{\theta} \,|\, \boldsymbol{y})$

$$\begin{aligned} t(\boldsymbol{y}) &\approx -2\ell(\hat{\theta} \,|\, \boldsymbol{y}) - \ell''(\hat{\theta} \,|\, \boldsymbol{y})(\theta_0 - \hat{\theta})^2 + 2\ell(\hat{\theta} \,|\, \boldsymbol{y}) \\ &= -\ell''(\hat{\theta} \,|\, \boldsymbol{y})(\theta_0 - \hat{\theta})^2 \end{aligned}$$

- By the LLN and since $\hat{\theta}$ is a consistent estimator

$$-\frac{1}{n}\ell''(\hat{\theta} \,|\, \boldsymbol{y}) \xrightarrow{\mathcal{P}} -\mathbb{E}[\ell''(\theta_0 \,|\, \boldsymbol{y})] = \mathcal{I}(\theta_0)$$

Model comparison

χ^2 approximation to the likelihood ratio for a simple $\mathcal{H}_0$ cont'd

- The maximum likelihood estimator $\hat{\theta}$ is asymptotically normal:

$$\sqrt{n\mathcal{I}(\theta_0)}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{D}} \mathrm{Normal}(0, 1)$$

- The distribution of the test statistic $t(\boldsymbol{y})$ converges against a $\chi^2(1)$ distribution as n approaches ∞

$$n\mathcal{I}(\theta_0)(\hat{\theta} - \theta_0)^2 \xrightarrow{\mathcal{D}} \chi^2(1)$$

because the square of a standard normal random variable is χ^2 distributed with 1 degree of freedom

Model comparison

Bayesian information criterion (BIC)

- Approximate the prior predictive distribution under $\mathcal{M}_i$ with Laplace's method

$$p(\boldsymbol{y}\,|\,\mathcal{M}_i) \approx (2\pi)^{d/2}\det(Q)^{-1/2}p(\boldsymbol{y}\,|\,\hat{\theta},\mathcal{M}_i)p(\hat{\theta}\,|\,\mathcal{M}_i)\,,$$

  where $\hat{\theta}$ is the maximum likelihood estimate and Q the negative Hessian of the log-likelihood evaluated at $\hat{\theta}$

- The BIC is derived by writing

$$\begin{aligned}-2\log p(\boldsymbol{y}\,|\,\mathcal{M}_i) \approx -d\log(2\pi) + \log\det(Q) - 2\log p(\hat{\theta}\,|\,\mathcal{M}_i)\\ - 2\log p(\boldsymbol{y}\,|\,\hat{\theta},\mathcal{M}_i)\end{aligned}$$

- By the LLN and since $\hat{\theta}$ is a consistent estimator

$$Q = -\frac{n}{n}\ell''(\hat{\theta}\,|\,\boldsymbol{y}) \xrightarrow{\mathcal{P}} -n\mathbb{E}[\ell''(\theta_0\,|\,\boldsymbol{y})] = n\mathcal{I}(\theta_0)$$

Model comparison

Bayesian information criterion (BIC) cont'd

- -2 times the log prior predictive distribution is thus

$$\begin{aligned}-2\log p(\boldsymbol{y}\,|\,\mathcal{M}_i) \approx& -d\log(2\pi) + d\log n + \log\det\{\mathcal{I}(\theta_0)\}\\ &- 2\log p(\hat{\theta}\,|\,\mathcal{M}_i) - 2\log p(\boldsymbol{y}\,|\,\hat{\theta},\mathcal{M}_i)\end{aligned}$$

- Dropping all terms that remain fixed as the sample size approaches ∞ results in

$$-2\log p(\boldsymbol{y}\,|\,\mathcal{M}_i) \approx \mathrm{BIC}_i = -2\log p(\boldsymbol{y}\,|\,\hat{\theta},\mathcal{M}_i) + d\log n$$

- Small BIC values correspond to better models
- The Akaike Information Criterion is equal to

$$\mathrm{AIC}_i = -2\log p(\boldsymbol{y}\,|\,\hat{\theta},\mathcal{M}_i) + 2d$$

- Small AIC values also correspond to better models

Model comparison

Bayesian model averaging

- Considering that some models perform equally well, it seems reasonable to base inference on several models by using Bayesian model averaging
- Model uncertainty is then accounted for by including information from all models weighted by their posterior model probability
- The model-averaged posterior of some quantity of interest Δ with the same interpretation across models is given by

$$p(\Delta \,|\, \boldsymbol{y}) = \sum_{k=1}^{K} p(\Delta \,|\, \mathcal{M}_k, \boldsymbol{y}) p(\mathcal{M}_k \,|\, \boldsymbol{y})$$

Model comparison

Bayesian model averaging cont'd

- The posterior probability of model $\mathcal{M}_k$ is given by

$$p(\mathcal{M}_k \,|\, \boldsymbol{y}) = \frac{p(\boldsymbol{y} \,|\, \mathcal{M}_k) p(\mathcal{M}_k)}{\sum_{l=1}^{K} p(\boldsymbol{y} \,|\, \mathcal{M}_l) p(\mathcal{M}_l)} \,,$$

  where $p(\boldsymbol{y} \,|\, \mathcal{M}_k)$ denotes the marginal likelihood of model $\mathcal{M}_k$ and $p(\mathcal{M}_k)$ the prior probability that model $\mathcal{M}_k$ is true
- Assuming uniform prior model probabilities, the approximate posterior model probability of model $\mathcal{M}_k$ is using the BIC given by

$$\hat{p}(\mathcal{M}_k \,|\, \mathcal{D}) = \frac{\exp\left\{-\frac{1}{2}\mathsf{BIC}_k\right\}}{\sum_{l=1}^{K} \exp\left\{-\frac{1}{2}\mathsf{BIC}_l\right\}}$$

Model comparison

Bayesian model averaging for logistic regression

- The model-averaged posterior inclusion probability of a predictor $\boldsymbol{x}_i$ is given by

$$p(\beta_i \neq 0 \,|\, \mathcal{D}) = \sum_{k=1}^{K} \mathbb{1}_{\mathcal{M}_k}(\beta_i) p(\mathcal{M}_k \,|\, \mathcal{D})$$

The model-averaged posterior mean of β_i is given by

$$\mathbb{E}[\beta_i \,|\, \mathcal{D}] = \sum_{k=1}^{K} \mathbb{E}[\beta_i \,|\, \mathcal{M}_k, \mathcal{D}] p(\mathcal{M}_k \,|\, \mathcal{D})$$

Model comparison

Bayesian model averaging for logistic regression cont'd

- The model-averaged posterior variance of β_i is given by

$$\mathbb{V}[\beta_i \,|\, \mathcal{D}] = \sum_{k=1}^{K} \left\{ \left(\mathbb{V}[\beta_i \,|\, \mathcal{M}_k, \mathcal{D}] + \mathbb{E}[\beta_i \,|\, \mathcal{M}_k, \mathcal{D}]^2 \right) \times \right. \\ \left. p(\mathcal{M}_k \,|\, \mathcal{D}) \right\} - \mathbb{E}[\beta_i \,|\, \mathcal{D}]^2$$

- If the sample size of the observed data $\boldsymbol{y}$ is large, then the posterior $p(\beta_i \,|\, \boldsymbol{y}, \mathcal{M}_k)$ of β_i under model $\mathcal{M}_k$ is asymptotically normal
- The mean is equal to the maximum likelihood estimator and variance equal to respective diagonal element of the inverse of the observed information matrix evaluated at the maximum likelihood estimator

Model comparison

Bayesian model averaging for logistic regression cont'd

- Dobutamine stress echocardiography study at the UCLA School of Medicine from 1991 until it closed in 1996
- The aim of the study was to assess if measurements taken during the stress echocardiography may be used to predict cardiac death, heart attack or coronary heart disease

Top 5 posterior model probabilities

Rank	Model	Posterior model probability†	Cumulative posterior model probability	Posterior model odds
01	$\mathcal{M}_1$: posSE, dobEF, hxofHT, restwma	0.0948	0.0948	1.00
02	$\mathcal{M}_2$: posSE, dobEF	0.0864	0.1812	1.10
03	$\mathcal{M}_3$: posSE, dobEF, restwma	0.0818	0.2631	1.16
04	$\mathcal{M}_4$: posSE, dobEF, hxofHT	0.0797	0.3427	1.19
05	$\mathcal{M}_5$: posSE, dobEF, hxofHT, ecg	0.0719	0.4147	1.32

Model comparison

Bayesian model averaging estimate

Predictor	Posterior inclusion probability	Posterior mean	Posterior standard deviation	95% equal tail interval	
				lower	upper
intercept	1.000	-0.34999	1.1031	-2.5120	1.8120
posSE	0.978	1.1126	0.2975	0.5295	1.6957
dobEF	0.882	-0.03617	0.0177	-0.0655	-0.0166
hxofHT	0.546	0.42712	0.4550	0.1586	1.4061
restwma	0.492	0.43209	0.5078	0.1647	1.5915
ecg	0.403	0.31211	0.4315	0.1419	1.4064
hxofMI	0.208	0.11074	0.2502	-0.0121	1.0750
hxofDM	0.147	0.06297	0.1811	-0.0753	0.9344
baseEF	0.132	-0.00277	0.0132	-0.0807	0.0388

Model comparison

Summary

- Likelihood ratio tests require nested models and rely on asymptotic approximations
- The BIC and AIC are tools that help balances model complexity and fit, which is evaluated through the maximized likelihood. The BIC prefers simpler models with small amounts of data, but becomes willing to accept more complex ones with increasing amount of data.
- Bayesian model averaging is a powerful tool to account for model uncertainty. The BIC may be used to approximate posterior model probabilities

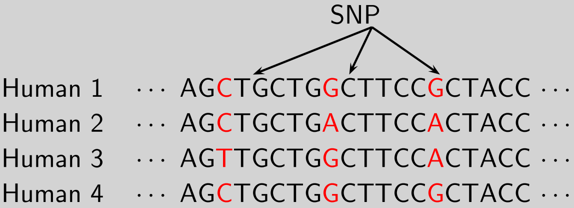
Content

1. Motivation
2. Approaches to statistics
3. Hypothesis testing
4. Model comparison
5. **SNP association studies**

SNP association studies

Relationship between SNP genotype and phenotype

- Genetic information is stored in the DNA in form of 4 nucleotide bases
- The human reference genome is approximately 3 giga bases long and any 2 humans differ in their genetic code by a small fraction
- Single Nucleotide Polymorphism (SNP) is a form of genetic variation at a genetic site at which the nucleotide base between 2 humans differs



SNP

Human 1 ··· AGCTGCTGGCTTCCGCTACC ···

Human 2 ··· AGCTGCTGACTTCCACTACC ···

Human 3 ··· AGTTGCTGGCTTCCACTACC ···

Human 4 ··· AGCTGCTGGCTTCCGCTACC ···

SNP association studies

Relationship between SNP genotype and phenotype cont'd

- The genotype is, among other factors, a strong influence on the phenotype
- An association between genotype and phenotype may be presumed for disease susceptibility, drug treatment or crop yields
- In case of drug treatments, some people react normally to the treatment, whereas others show none or life-threatening effects
- A particular set of SNPs may be characteristic for these phenotypes
- The goal of genome-wide association studies is then to reveal SNP patterns that permit disease susceptibility screens or personalized drug treatments

Relationship between SNP genotype and phenotype cont'd

- DNA carried by the chromosomes is present in 2 copies
- Without considering DNA copy number variation, the genotype at a biallelic SNP is either
  - AA - 2 copies of the common allele
  - AB - 1 copy of each allele
  - BB - 2 copies of the rare allele

  where allele refers to the particular nucleotide base.
- The frequency distribution of a SNP genotype and phenotype can be visualized in a contingency table

	AA	**AB**	**BB**	
Case	$n_{\text{AA}}^{\text{Case}}$	$n_{\text{AB}}^{\text{Case}}$	$n_{\text{BB}}^{\text{Case}}$	n^{Case}
Control	$n_{\text{AA}}^{\text{Control}}$	$n_{\text{AB}}^{\text{Control}}$	$n_{\text{BB}}^{\text{Control}}$	n^{Control}
	n_{AA}	n_{AB}	n_{BB}	n

SNP association studies

General genotype count model: prospective model

- The counts may be modeled directly

	AA	**AB**	**BB**	
Case	$n_{\text{AA}}^{\text{Case}}$	$n_{\text{AB}}^{\text{Case}}$	$n_{\text{BB}}^{\text{Case}}$	n^{Case}
Control	$n_{\text{AA}}^{\text{Control}}$	$n_{\text{AB}}^{\text{Control}}$	$n_{\text{BB}}^{\text{Control}}$	n^{Control}
	n_{AA}	n_{AB}	n_{BB}	n

- In a prospective model, the phenotype is the random variable, whereas the genotype variable is supposed to be known
- Under the null hypothesis $\mathcal{H}_0$, there exists no association between both variables and thus

$$p(\boldsymbol{y} \,|\, \theta, \mathcal{H}_0) = \binom{n}{n^{\text{Case}}} \theta^{n^{\text{Case}}} (1-\theta)^{n^{\text{Control}}}$$

SNP association studies

General genotype count model: prospective model cont'd

- Assume that the prior of θ, which represents the probability of being a case, is a Beta distribution

$$p(\theta) = \frac{1}{\mathcal{B}(\alpha,\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1},$$

where the Beta function is equal to

$$1/\mathcal{B}(\alpha,\beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$$

- The prior predictive distribution is then

$$\begin{aligned} p(\boldsymbol{y} \,|\, \mathcal{H}_0) &= \binom{n}{n^{\text{Case}}} \frac{1}{\mathcal{B}(\alpha,\beta)} \int_0^1 \theta^{n^{\text{Case}}+\alpha-1}(1-\theta)^{n^{\text{Control}}+\beta-1}\,\mathrm{d}\theta \\ &= \binom{n}{n^{\text{Case}}} \frac{\mathcal{B}(n^{\text{Case}}+\alpha, n^{\text{Control}}+\beta)}{\mathcal{B}(\alpha,\beta)} \end{aligned}$$

SNP association studies

General genotype count model: prospective model cont'd

- Under the alternative hypothesis $\mathcal{H}_1$, the 3 genotypes are assumed to be independent and thus

$$p(\boldsymbol{y} \,|\, \tau_{AA}, \tau_{AB}, \tau_{BB}, \mathcal{H}_1) = \binom{n}{n^{\text{Case}}} \prod_{i \in \{AA,AB,BB\}} \tau_i^{n_i^{\text{Case}}} \times (1-\tau_i)^{n_i^{\text{Control}}}$$

- Assume that the prior of τ_i, which represents the probability of being a case given that the genotype is i, has also a Beta distribution

General genotype count model: prospective model cont'd

- The prior predictive distribution is then

$$
\begin{aligned}
p(\boldsymbol{y} \,|\, \mathcal{H}_1) = \binom{n}{n^{\mathsf{Case}}} \prod_{i \in \{AA,AB,BB\}} \frac{1}{\mathcal{B}(\alpha,\beta)} \times \\
\int_0^1 \tau_i^{n_i^{\mathsf{Case}}+\alpha-1}(1-\tau_i)^{n_i^{\mathsf{Control}}+\beta-1}\,\mathrm{d}\tau_i \\
= \binom{n}{n^{\mathsf{Case}}} \prod_{i \in \{AA,AB,BB\}} \frac{\mathcal{B}(n_i^{\mathsf{Case}}+\alpha, n_i^{\mathsf{Control}}+\beta)}{\mathcal{B}(\alpha,\beta)}
\end{aligned}
$$

General genotype count model: prospective model cont'd

- The prospective Bayes factor is

$$B_{10} = \frac{p(\boldsymbol{y} \,|\, \mathcal{H}_1)}{p(\boldsymbol{y} \,|\, \mathcal{H}_0)}$$
$$= \frac{\mathcal{B}(\alpha, \beta)}{\mathcal{B}(n^{\text{Case}} + \alpha, n^{\text{Control}} + \beta)} \times \prod_{i \in \{AA, AB, BB\}} \frac{\mathcal{B}(n_i^{\text{Case}} + \alpha, n_i^{\text{Control}} + \beta)}{\mathcal{B}(\alpha, \beta)}$$

- Data-dependent hyperparameters may be used:

$$(\alpha, \beta) = \lambda \left(n^{\text{Case}}/n, n^{\text{Control}}/n \right) \,,$$

which are uninformative in distinguishing $\mathcal{H}_0$ from $\mathcal{H}_1$ and where λ is used to scale the effect size.

Thank you for your attention