CHAPTER 13: HYPOTHESIS TESTING

In this chapter we discuss:

- Frequentist and Bayesian approaches to testing a single hypothesis.
- Multiple hypothesis testing, including variable selection.

Testing hypotheses is very context specific; we consider three distinct scenarios:

1. Confirmatory analyses in which an a priori hypothesis concerning a particular response/covariate relationship is of interest, and other variables have been measured and we wish to know which to adjust for.
2. Exploratory analyses where the aim is to gain clues as to structure in the data. For example, which covariates are causally related to a response, or characterizing sources of variability.
3. Prediction in which we are not concerned with causality, but merely with predicting a response given a set of variables.

Frequentist hypothesis testing

Suppose we are interested in the null hypothesis:

$$H_0 : \beta = 0$$

and we have a statistic T, with large values being increasingly unlikely under the null. The observed value is t_{obs}.

Possibilities for T:

- Squared Wald statistic from a regression analysis.
- Score statistic.

Under regularity conditions, $T \rightarrow_n \chi^2_1$ under the null, as $n \rightarrow \infty$. If n is not large then permutation/Monte Carlo can be used to give the distibution of the statistic under the null.

Interpetation of p-values

The $\Pr(T > t|H_0) = 1 - F(T)$ is uniform under the null. Let

$$p = \Pr(T > t_{\text{obs}}|H_0)$$

denote the observed p-value — the probability of observing t_{obs}, or a more extreme value under the null.

Historically there has two approaches to the use of the above approach.

- Fisher and the pure test of significance. Quote the observed p-value as the measure of evidence against the null. No concept of rejecting the null in terms of the alternative, as there is no alternative!

  No long-run frequency control of the type I error — $\alpha = \Pr(T > t_{\text{fix}})$, the probability of rejection of the null when it is true.

- Neyman-Pearson: specify an alternative hypothesis (with H_0 nested in H_1), and then evaluate the likelihood ratio statistic.

  Use of the Neyman-Pearson lemma to find, for fixed α, the most powerful test. Decision rule: if $p < \alpha$ reject the null; under a fixed threshold this procedure controls the type I error.

Critique of Fisherian approach

- How do we decide on whether p is small or not, i.e. how to decide on a threshold for significance?
- With large sample sizes we will always obtain small p-values because although the p-value is still uniform under the null, we will be able to detect very subtle departures from the null with a large sample size:
  - To rectify this a confidence interval for β is often given along with the p-value, so that the scientific significance of the departure can be determined.
  - Suggests that the p-value threshold should decrease with increasing n.
- What is the prescription for finding a suitable T, particularly when we have nuisance parameters? Intuitively we want a statistic that takes large values (has high power) for realistic alternatives.
- Another aspect of p-values is that they use a tail area and

$$\Pr(H_0|T > t_{\text{obs}}) < \Pr(H_0|T = t_{\text{obs}}).$$

Critique of Neyman-Pearson approach

- How do we decide on a size α?
- The emphasis is on fixing α, but this should also decrease with increasing n.
- Just stating that a p-value less than α was achieved is throwing away information — but if we give an observed p-value, how do we interpret.
- What if H_0 and H_1 are both wrong (always true in practice!).

Under both approaches, wouldn't we rather know $\Pr(H_0|\boldsymbol{y})$? This requires a Bayesian approach.

Bayesian decision theory approach to hypothesis testing

Under a Bayes decision theory approach to hypothesis testing the "decision" δ is taken that minimizes the posterior expected loss.

$L(\delta, H)$		Decision	
		$\delta = 0$	$\delta = 1$
Truth H	H_0	0	L_α
	H_1	L_β	0

Table 14: Losses corresponding to the decision δ, when the truth is H, L_α is the loss associated with a type I error and L_β with a type II error.

For hypothesis testing we have two possible data generating mechanisms:

$$\begin{aligned} H_0 &\rightarrow \boldsymbol{\beta}_0|H_0 \rightarrow \boldsymbol{y}|\boldsymbol{\beta}_0 \\ H_1 &\rightarrow \boldsymbol{\beta}_1|H_1 \rightarrow \boldsymbol{y}|\boldsymbol{\beta}_1 \end{aligned}$$

As always in the Bayesian approach in which unknowns are treated as random, the true hypothesis H is viewed as an unknown parameter for which the posterior may be derived.

We have the posterior probability of H_j:

$$\Pr(H_j|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|H_j) \times \pi_j}{p(\boldsymbol{y})}$$

with π_j the prior probability of hypothesis H_j and

$$p(\boldsymbol{y}|H_j) = \int p(\boldsymbol{y}|\boldsymbol{\beta}_j) f(\boldsymbol{\beta}_j|H_j)\ d\boldsymbol{\beta}_j$$

where $f(\boldsymbol{\beta}_j|H_j)$ is the prior distribution over the parameters associated with hypothesis H_j, $j = 1, 2$.

With respect to Table 14, the posterior expected loss associated with the decision δ is

$$\mathrm{E}[L(\delta, H)] = L(\delta, H_0) \Pr(H_0|\boldsymbol{y}) + L(\delta, H_1) \Pr(H_1|\boldsymbol{y})$$

so that for the two possible decisions (accept/reject H_0) the expected losses are:

$$\begin{aligned} \mathrm{E}[L(\delta = 0, H)] &= 0 \times \Pr(H_0|\boldsymbol{y}) + L_\beta \Pr(H_1|\boldsymbol{y}) \\ \mathrm{E}[L(\delta = 1, H)] &= L_\alpha \Pr(H_0|\boldsymbol{y}) + 0 \times \Pr(H_1|\boldsymbol{y}) \end{aligned}$$

To find the decision that minimizes posterior expected loss, we therefore need to take the smaller of:

$$\begin{aligned} \mathrm{E}[L(\delta = 0, H)] &= L_\beta \times \Pr(H_1|\boldsymbol{y}) \\ \mathrm{E}[L(\delta = 1, H)] &= L_\alpha \times \Pr(H_0|\boldsymbol{y}). \end{aligned}$$

We should choose $\delta = 1$ if

$$L_\beta \times \Pr(H_1|\boldsymbol{y}) \geq L_\alpha \Pr(H_0|\boldsymbol{y})$$

i.e. if

$$\frac{\Pr(H_1|\boldsymbol{y})}{1 - \Pr(H_1|\boldsymbol{y})} \geq \frac{L_\alpha}{L_\beta}$$

or

$$\Pr(H_1|\boldsymbol{y}) \geq \frac{L_\alpha/L_\beta}{1 + L_\alpha/L_\beta}.$$

Hence we only need to specify the ratio of losses.

If equal errors choose H_1 if $\Pr(H_1|\boldsymbol{y}) > \Pr(H_0|\boldsymbol{y})$.

Rearranging we can say that we should decide on H_1 if

$$\frac{\Pr(H_0|\boldsymbol{y})}{\Pr(H_1|\boldsymbol{y})} = \text{Bayes factor} \times \frac{\pi_0}{\pi_1} < \frac{L_\beta}{L_\alpha}$$

where

$$\text{Bayes factor} = \frac{p(\boldsymbol{y}|H_0)}{p(\boldsymbol{y}|H_1)} = \frac{\text{Posterior Odds}}{\text{Prior Odds}}.$$

So if a type I error is 4 times as bad as a type II error we should choose H_1 if the posterior odds on H_0 drop below 0.25.

Critique of Bayes Approach

- Need to specify prior distributions on all parameters under null and alternative, and on the hypotheses. In general cannot get away with improper priors when hypothesis testing is considered (unlike estimation).
- All of the calculations above should be conditioned on $H_0 \cup H_1$ — we are really obtaining the posterior probability of the null given one of the null or alternatives is true, and under our assumed data and prior models.
- The calculation of the Bayes factor requires integals, which will usually be intractable.
- How to decide on the ratio of losses?

Calibrating p-values

Let data $= T > t_{\text{obs}}$ — we want $\Pr(H_0|$ data $)$ but to obtain this we must specify alternatives – consider a simple alternative, say $H_1 : \beta = \beta_1$.

Then

$$\Pr(H_0|\text{ data }) = \frac{\Pr(\text{ data }|H_0)\pi_0}{\Pr(\text{ data }|H_0)\pi_0 + \Pr(\text{ data }|H_1)\pi_1}$$

where $\pi_j = \Pr(H_j)$, j=0, 1. Dividing by $\Pr(H_1|$ data $)$:

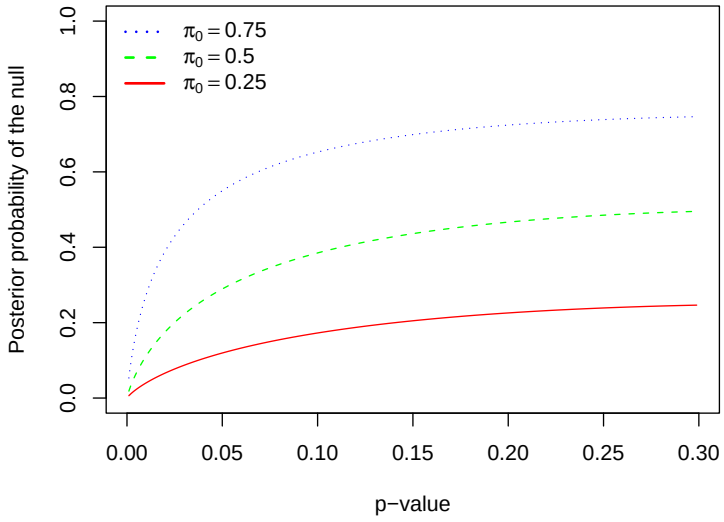
$$\begin{aligned}\text{Posterior Odds of } H_0 &= \frac{\Pr(\text{ data }|H_0)}{\Pr(\text{ data }|H_1)} \times \text{ Prior Odds of } H_0 \\ &= \frac{p\text{-value}}{\text{power at } H_1} \times \text{ Prior Odds of } H_0\end{aligned}$$

which depends on:

- The prior on H_0, π_0.
- The power, $\Pr(\text{ data }|H_1)$ — greater power, more "evidence" in p-value.

Sellke, Bayarri and Berger (2001) show that for a p-value $p < 1/e$:

$$\Pr(H_0|\text{ data }) \geq \left\{1 - \frac{1}{e\ p \log p} \times \frac{\pi_1}{\pi_0}\right\}^{-1}$$



The interpretation of p-values

Sanity check! Why does anyone use p-values?

Historically it was usual to carry out single experiments and the prior on the alternative was not tiny.

With $\pi_0 = 0.5$:

- p-value $= 0.05$ gives $\Pr(H_0|\text{ data }) > 0.29$.
- p-value $= 0.01$ gives $\Pr(H_0|\text{ data }) > 0.11$.

But perhaps this is one of the reasons so many "findings" are not reproducible (along with confounding, multiple testing, errors-in-variables,...).

Variable Selection

We define the *null* model as that which contains an intercept only, the *minimal* model as the smallest model which is consistent with prior information.

So for example, in an epidemiological investigation we would almost always want to include terms for age and gender.

The minimal model may also be a function of the design so in matched case-control studies we include a term for each of the matching sets. Similarly in clinical trials in which treatments are randomized within a priori chosen strata, we again will include a term for strata.

Selection of Regressors

Trade-off: as we include more covariates, bias is reduced, but variability may be increased, dependent on how strong a predictor the covariate is (and its association with other covariates) — this is why we don't fit the full model.

We now describe some of the approaches to subset selection that have been proposed in the literature:

Forward selection. Begins with the simplest model. At each stage the 'best' unselected variable that satisfies the selection criterion is added. Best here is defined to be that variable whose deviance (or Wald or score statistic) is largest. This variable is added to the regression if its statistic is greater than a threshold of a specified significance level. This value, is contentious. Note that a maximum of p models will be considered in this procedure (out of 2^p).

Backward elimination. Begins with the full model. At each stage the covariate with the smallest deviance value that is less than a specified value is removed.

Stepwise regression (Efroymson's algorithm). Follows forward selection with backward elimination.

Difficulties

There are a number of problems with selection methods (Miller, 1990). In the conventional use of a hypothesis test, for the correct interpretation of significance levels the hypotheses must be specified before the data are examined. Note: same problem with transformations of y and/or x, choice of variance models and error distributions.

As a Bayesian all of the possible models that will be fitted should be assigned, priorm with model averaging producing the summaries, averaging over all models.

Similarly for interval estimates to be valid the model must be specified *a priori*.

There now exists great potential for over-fitting in which models become too dataset-specific as they are refined on the basis of the examination of diagnostics.

In practice, if refinement is carried out through the fitting of alternative models (e.g. transformation of covariates, choice of distribution for the responses), then interval estimates will often be too narrow since they are produced by conditioning on the final model, and hence do not reflect the mechanism by which the model was selected.

We illustrate some of the difficulties of model selection with two simple examples.

Example 1: If we carry out a *single* hypothesis test and report the estimate of β_1 in a simple linear regression *only if* the null hypothesis of $\beta_1 = 0$ is rejected.

Figure 53 results – the bias is clear.

There are close links with publication bias in meta-analysis.

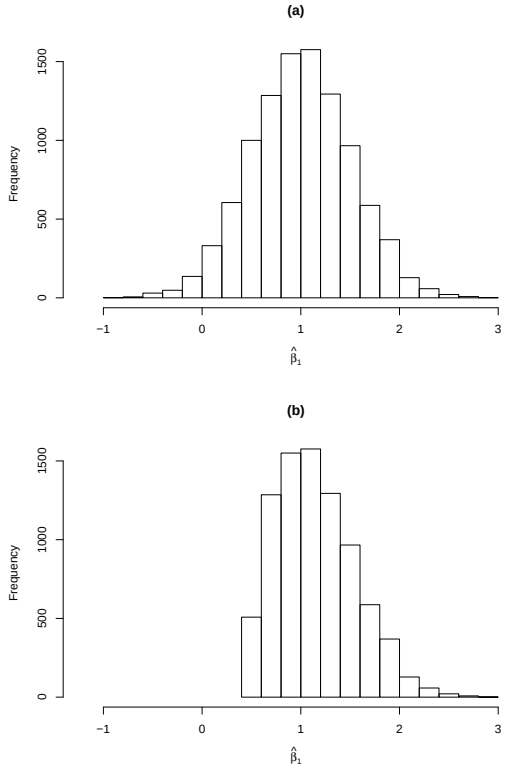


Figure 53: $\mathrm{E}[\widehat{\beta}_1] = 1.00$, while $\mathrm{E}[\widehat{\beta}_1|$ rejection of $H_0] = 1.27$.

Example 2: Suppose we are interested in β_1 but we wish to "control" for β_2 by testing whether the latter is significant.

In the following simulation, a multiple linear regression in X_1 and X_2 was carried out. The true values were $\beta_1 = \beta_2 = 1$ and X_1, X_2 were simulated from a bivariate normal with means zero, variances one, and correlation 0.7.

Figure 54 shows the results, in (a) we display the sampling distributions of $\widehat{\beta}_1$ from the adjusted model. The mean and standard deviation of the distribution of $\widehat{\beta}_1$ are 1.00 and 1.23.

Panel (b) displays the sampling distribution of the *reported* estimator.

The mean and standard deviation of the distribution of the reported estimate of β_1 are 1.23 and 1.01, respectively, showing positive bias and a reduced variance.

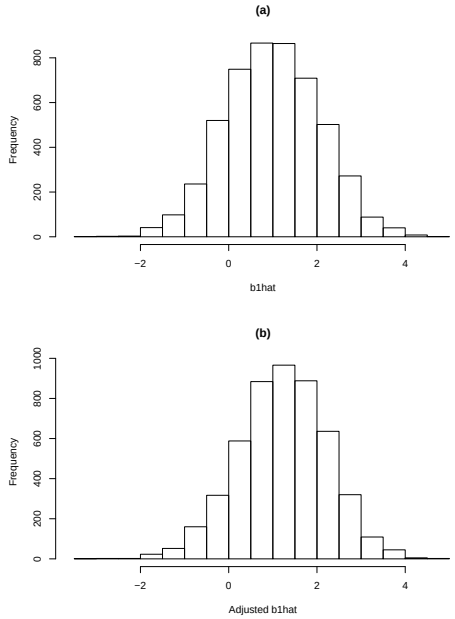


Figure 54: (a) Sampling distribution of $\widehat{\beta}_1$, (b) sampling distribution of $\widehat{\beta}_1$ given "control" for the possibility that $\beta_2 \neq 0$.

Frequentist model selection difficulties

From a frequentist standpoint estimators and test statistics should be examined via their long-run behaviour *given* the model-fitting process, including refinement. To be more explicit, let P denote the procedure by which a final model M is decided upon. Then suppose it is of interest to examine the bias of a statistic T,

$$\mathrm{E}[T|P] = \mathrm{E}_{M|P}\{\mathrm{E}[T|M]\}. \tag{50}$$

In general it will be incorrect to report $T \mid \widehat{M}$ where $\widehat{M}$ is the final model chosen, since this does not reflect the procedure by which $\widehat{M}$ was chosen, but rather acts as if the final model is the "truth".

We know that

$$\mathrm{var}(T|P) = \mathrm{E}_{M|P}[\mathrm{var}(T|M)] + \mathrm{var}_{M|P}(\mathrm{E}[T|M]).$$

but $\mathrm{var}(T|\widehat{M})$ is reported (which approximates the first term only).

Under a frequentist approach inference follows from the behaviour of an estimator under repeated sampling.

Bayesian model selection difficulties

From a Bayesian standpoint the same problem of dredging exists because the posterior distribution should reflect all sources of uncertainty and *a priori* all possible models that may be entertained should be explicitly stated, with prior distributions being placed upon different likelihoods and the parameters of these likelihoods.

Model averaging (see later) should then be carried out across the different possibilities, a process which is fraught with difficulties not least in placing "comparable" priors over what may be fundamentally different objects.

(One solution is to place prior on "model-free" quantities.)

Model Averaging

Suppose the action concerns a parameter of interest T (which for simplicity we assume is univariate) that is well-defined for all models.

We have

$$\mathrm{E}[T|\boldsymbol{y}] = \sum_{j=1}^{J} \mathrm{E}[T|\boldsymbol{y}, M_j] \times \Pr(M_j|\boldsymbol{y}),$$

and

$$\begin{aligned} \mathrm{var}(T|\boldsymbol{y}) &= \sum_{j=1}^{J} \mathrm{var}(T|\boldsymbol{y}, M_j) \times \Pr(M_j|\boldsymbol{y}) \\ &+ \sum_{j=1}^{J} \{\mathrm{E}[T|\boldsymbol{y}, M_j] - \mathrm{E}[T|\boldsymbol{y}]\}^2 \times \Pr(M_j|\boldsymbol{y}). \end{aligned}$$

This latter term shows how not only parameter uncertainty but *model* uncertainty is accounted for.

- Specification of priors is not trivial.
- Interpretation.
- Continuous model expansion.

A possible compromise

One solution is to never refine the model for a given data set. This approach is operationally pure but pragmatically dubious (unless one is in the context of a randomized experiment) since we may obtain appropriate inference for a model that is a very poor description of the phenomenon under study.

The philosophy suggested here is to think as carefully as possible about the initial model class before the analysis proceeds, but after fitting to carry out model checking and refine the model in the face of *clear* model misspecification, with refinement ideally being carried out within distinct *a priori* known classes.

With reference to (50), if a model is chosen because it is clearly superior to the alternatives, then it may be reasonable to assume that $\mathrm{E}[T \mid P] \approx \mathrm{E}[T \mid \widehat{M}]$, because $\widehat{M}$ would be consistently chosen in repeated sampling under these circumstances.

So, for example, examining quantile-quantile plots for different t distributions and picking the one that produces the straightest line would not be a good idea.

Inference then proceeds as if the final model were the one that were chosen initially. This is clearly a subjective procedure but can be informally justified via either philosophical approaches.

In a similar vein, under a Bayesian approach the above procedure is consistent with model-averaging but with the posterior model weight being concentrated upon the chosen model (since alternative models are only rejected on the basis of clear inadequacy).

The aim is to provide probability statements, from either philosophical standpoints that are "honest" representations of uncertainty. The above approach is relevant to analyses that are more confirmatory in their outlook, as opposed to being used for prediction, or for more exploratory purposes (for example, to gain clues to models that may be appropriate for future data analyses).

Conclusions

- For Confirmatory studies – try to avoid any model selection. Use background context to specify model.
- Exploratory studies – stepwise and all subsets may point to important variables, but attaching a p-value is difficult. Model averaging is another possibility.
- Prediction – some form of shrinkage should be used. Cross-validation may be used to choose a model. Bayesian model averaging also useful.

Lindley's Paradox

Lindley (1957) gave a much-quoted example in which Bayesian and frequentist inference differ.

We assume that $\overline{Y}|\theta \sim \mathrm{N}(\theta, \sigma^2/n)$ with σ^2 known and θ unknown.

We first state Lindley's paradox, without loss of generality suppose the null is $H_0 : \theta = 0$, with alternative $H_1 : \theta \neq 0$.

Let

$$\overline{y}_n = Z_p \sigma/\sqrt{n}$$

where p is the p-value and Z_p is the corresponding quantile of the normal distribution, i.e. that Z_p such that $\Pr(Z > Z_p) = p/2 = \Pr(Z < -Z_p)$.

We define $\overline{y}_n$ in this way so that as n increases the p-value remains constant.

For a Bayesian analysis assume that we assign π_0 to H_0, and under the alternative $\theta \sim \mathrm{N}(0, \tau^2)$. Then we have

$$\Pr(H_0|\overline{y}_n) = \frac{\mathrm{BF} \times \mathrm{PO}}{1 + \mathrm{BF} \times \mathrm{PO}}$$

where the Bayes factor is given by

$$\mathrm{BF} = \frac{p(\overline{y}_n|H_0)}{p(\overline{y}_n|H_1)}$$

and $\mathrm{PO} = \pi_0/(1-\pi_0)$ is the prior odds.

We have

$$\begin{aligned} \overline{y}_n|H_0 &\sim \mathrm{N}(0, \sigma^2/n) \\ \overline{y}_n|H_1 &\sim \mathrm{N}(0, \sigma^2/n + \tau^2) \end{aligned}$$

so that

$$\mathrm{BF} = \frac{(2\pi\sigma^2/n)^{-1/2} \exp\left[-\frac{\overline{y}_n^2}{2\sigma^2/n}\right]}{(2\pi[\sigma^2/n + \tau^2])^{-1/2} \exp\left[-\frac{\overline{y}_n^2}{2(\sigma^2/n+\tau^2)}\right]}$$

For $\sigma^2 = 1, \tau^2 = 0.2^2, \pi_0 = 0.5, p = 0.05$ and $n = 1, ..., 100,000$, Figure 55 shows the posterior probability of the null as a function of n.

From the starting position of $\Pr(H_0|\overline{y}_n) = 0.5$ the curve initially falls, reaching a minimum at around $n = 100$, and then increases towards 1, illustrating the "paradox".

For large values of n, $\overline{y}_n$ is very close to the null value of 0, but there is high power to detect any difference from 0, and so a p-value of 0.05 is not difficult to achieve.

The Bayes factor calculates the density under the alternative also, and values close to 0 are more likely under the null, Figure 56 illustrates for $n = 1000$, the green vertical line denotes $\overline{y}_n$.

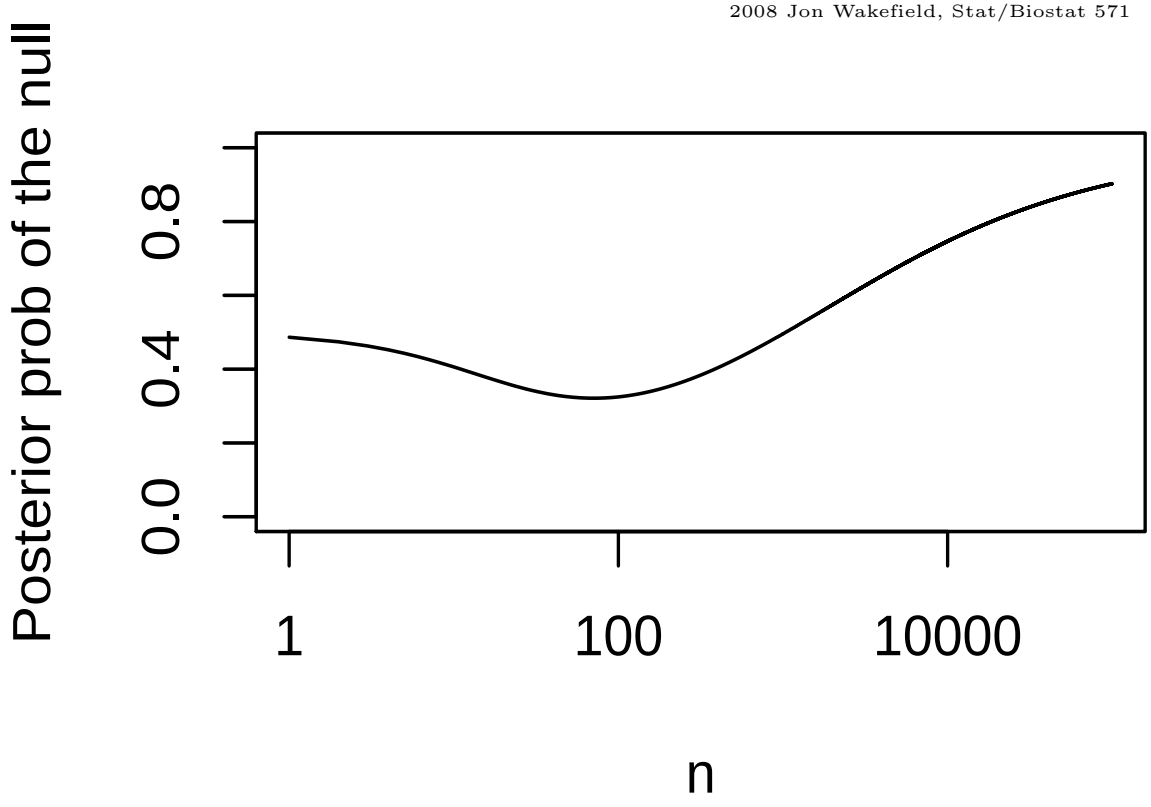


Figure 55: Posterior probability of the null for a fixed p-value of 0.05.

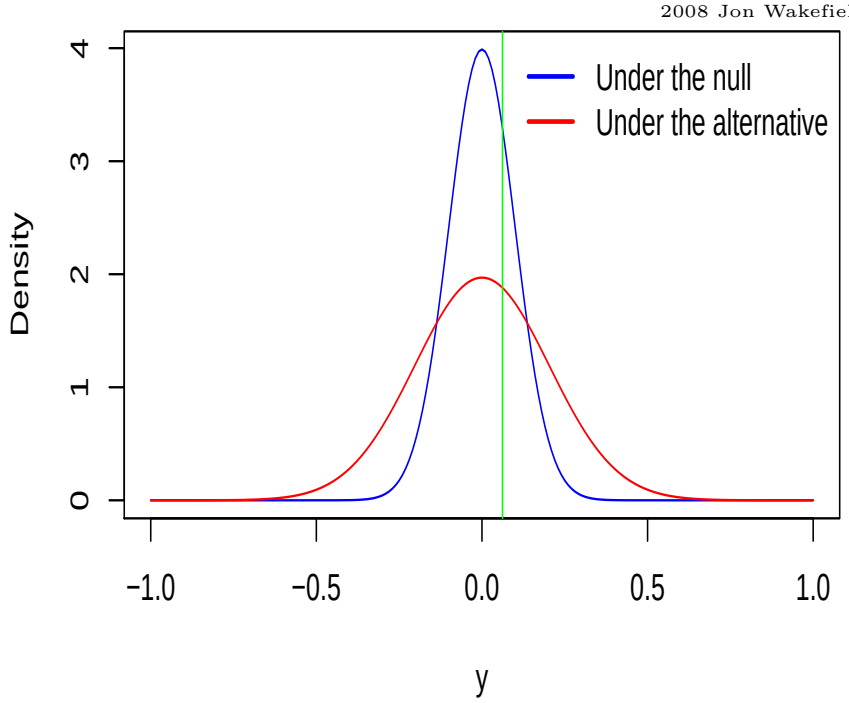


Figure 56: Numerator (blue) and denominator (red) of the Bayes factor for $n = 1000$. The green line represents $\overline{y}_n$ for this n.

Bayesian analysis with a tail-area

We now consider a Bayesian analysis in which the data appear in the form of knowing that $|\overline{Y}_n| \geq \overline{y}_n$, a censored observation. This is clearly not realistic since a Bayesian would condition on the *actual* value observed, but it does allow considerations of statements about power made by several authors.

We have

$$\text{BF} = \frac{\Pr(|\overline{Y}_n| \geq \overline{y}_n|H_0)}{\Pr(|\overline{Y}_n| \geq \overline{y}_n|H_1)}$$

so that the numerator is the p-value, and the denominator is given by

$$\begin{aligned}
\Pr(|\overline{Y}_n| \geq \overline{y}_n|H_1) &= \int \Pr(|\overline{Y}_n| \geq \overline{y}_n|\theta)\pi(\theta)d\theta \\
&= \int \left\{\Pr(\overline{Y}_n \geq \overline{y}_n|\theta) + \Pr(\overline{Y}_n \leq -\overline{y}_n|\theta)\right\}\pi(\theta)d\theta \\
&= \int \left\{\Phi\left(Z \geq \frac{\sqrt{n}(\overline{y}-\theta)}{\sigma}\right) + \Phi\left(Z \leq \frac{\sqrt{n}(-\overline{y}-\theta)}{\sigma}\right)\right\}\pi(\theta)d\theta
\end{aligned}$$

where $Z \sim \text{N}(0,1)$, so that we are evaluating the average of the power over the prior $\pi(\theta)$.

We emphasize that it is the post-data power that is being evaluated, i.e. it is based on the observed value of the statistic.

Figure 57 gives the average power as a function of n, and we see a monotonic increase with sample size towards the value 1.

Hence, as seen in Figure 58 the Bayes factor based on the tail-area information is monotonic decreasing towards the p-value as n increases (with $\pi_0 = 0.5$ this gives the posterior probability of the null also). Hence this justifies the claims of a number of authors that greater credence should be given to p-values based on large sample sizes/power.

The difference in behavior between a genuine Bayesian analysis that conditions on the actual statistic and that based on the tail area is apparent.

As noted by Lindley (1957, p. 189–190), "...the paradox arises because the significance level argument is based on the area under a curve and the Bayesian argument is based on the ordinate of the curve".

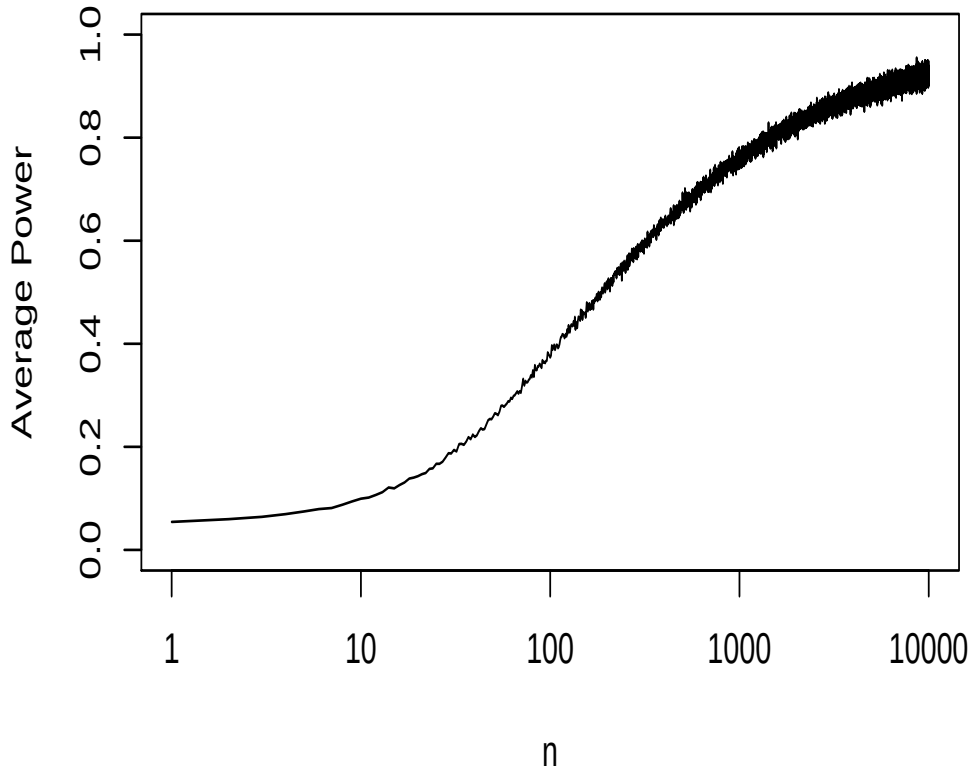


Figure 57: Average power under the $N(0, 0.2^2)$ prior for a fixed p-value of 0.05.

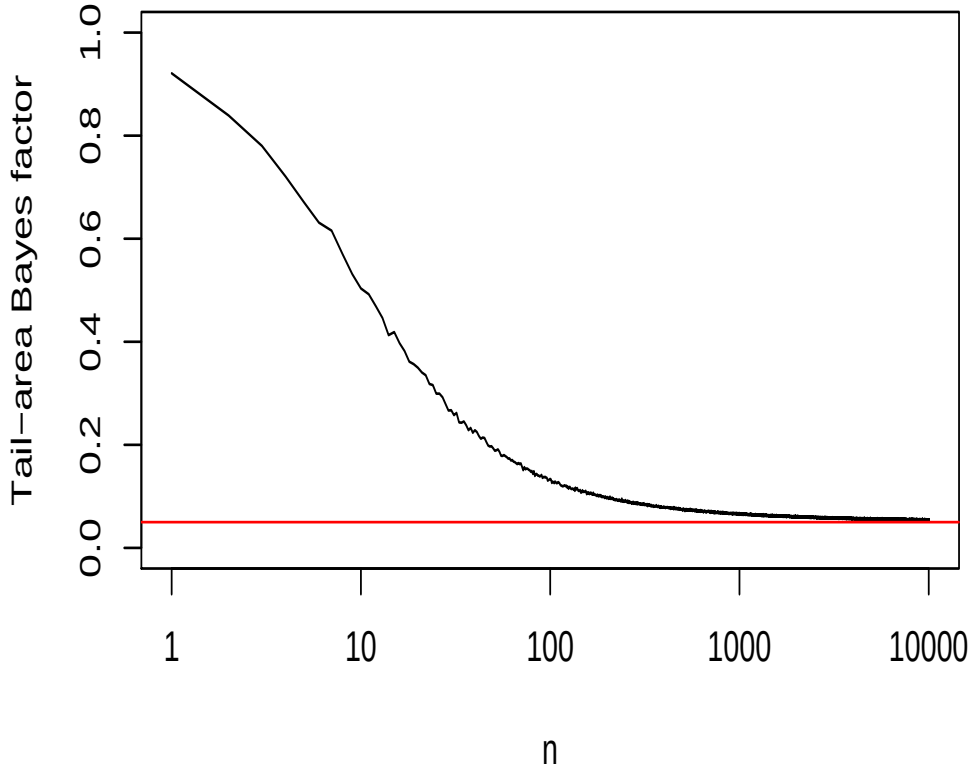


Figure 58: Bayes factor based on the observed tail area.

Multiple Hypothesis Testing

In subgroup analyses we want to examine the data for effects in different sub-groups.

Possibilities when m tests are performed and K are flagged as requiring further attention:

	Non-Flagged	Flagged	
H_0	A	B	m_0
H_1	C	D	m_1
	$m-K$	K	m

- m_0 is the number of true nulls.
- B is the number of type I errors.
- C is the number of type II errors.

How do we select a rule that will determine K?

The Statistical Set-Up

The p-value was introduced in a single test situation to control the Type I error rate – historically the emphasis was placed on avoiding a Type I error.

The family-wise error rate (FWER) is the probability of making at least one Type I error, i.e. $\Pr(B \geq 1|$ all H_0 true $)$.

More recently there has been more interest in the false discovery rate — the expected proportion of rejected nulls that are actually true, see Benjamini and Hochberg (1995) and Storey (2002).

Let B_i be the event that the i-th null is incorrectly rejected, so that $B = \cup_{i=1}^m B_i$ is the total number of incorrectly rejected nulls.

The FWER is given by:

$$\begin{aligned} \alpha = \Pr(B \geq 1| \text{ all } H_0 \text{ true }) &= \Pr\left(\cup_{i=1}^m B_i| \text{ all } H_0 \text{ true }\right) \\ &\leq \sum_{i=1}^m \Pr(B_i| \text{ all } H_0 \text{ true }) \\ &= m\alpha^\star \end{aligned}$$

where $\alpha^\star$ is the level for each test.

Bonferroni takes $\alpha^\star = \alpha/m$ to give FWER $\leq \alpha$.

For control at $\alpha = 0.05$ with $m = 10$ tests take $\alpha^\star = 0.05/10 = 0.005$.

Such stringent rules lead to a loss of power, but not ridiculous if you think there is a reasonable chance that all nulls could be true

Bayesian Bonferroni

If we have prior probability of the null $\pi_{0i} = \pi_0$ for $i = 1, ..., m$ then

$$\Pi_0 = \Pr(\text{ prior probability that all nulls are true}) = \pi_0^m$$

For example, if $\pi_0 = 0.5$ and $m = 10$, $\Pi_0 = 0.00098$, which may be deemed too small.

Westfall et al. (1997) show that if we take $\pi_{0i} = \Pi_0^{1/m}$ (so that the probability that all nulls are true is Π_o) — call this prior 2 — then for independent tests we have approximately

$$\alpha_B = \Pr(H_{0i}|\boldsymbol{y}_i, \text{ prior 2 }) \approx m \times \Pr(H_{0i}|\boldsymbol{y}_i, \text{ prior 1 }) = m \times \alpha_B^\star$$

where prior 1 is $\pi_{0i} = \Pi_0$.

CHAPTER 14: MISSING DATA

A serious problem in data analysis is the existence of missing data. We concentrate on missing responses in a dependent data situation.

Implications of missing data:

1. Data are unbalanced – not a problem given modern regression techniques.
2. Information loss.
3. Depending on the mechanism of missingness, bias in estimation may result.

Missing data can arise in numerous ways, and understanding the mechanism is crucial to appropriate modeling assumptions.

In a longitudinal study, if *drop-out* occurs at a certain time then no additional data are observed after that point.

Examples:

1. In a health-air pollution study an individual may be unavailable for measurement because he/she took a job in another area.
2. In a clinical trial, patients may be removed from the study if their longitudinal measurements are below/above some limit.
3. Censoring – measurement instruments may be inaccurate below a lower limit of detection, this limit is then reported.
4. The value of the outcome may itself determine the missingness, but the outcome is unobserved.

In 1, the missingness will not be a problem unless the person moved area because of health problems. In 2, the missingness will be a function of the responses on previous occasions, while in 3 and 4 it depends on the actual measurement that would have been recorded.

Example: Simulated Data

Data were simulated in which the data ($m = 200, n_i = 10, i = 1, ..., m$) were generated from a linear mixed model in which intercepts and slopes are random (and independent), with measurement error and $\beta_0 = 100, \beta_1 = -5$.

Figure 59 shows the resultant data.

We then simulated drop-out by a mechanism in which if the outcome falls below 65, the subsequent observations are lost (but we retain the initial one below 65).

Figure 60 shows the resultant data (509 data points were lost).

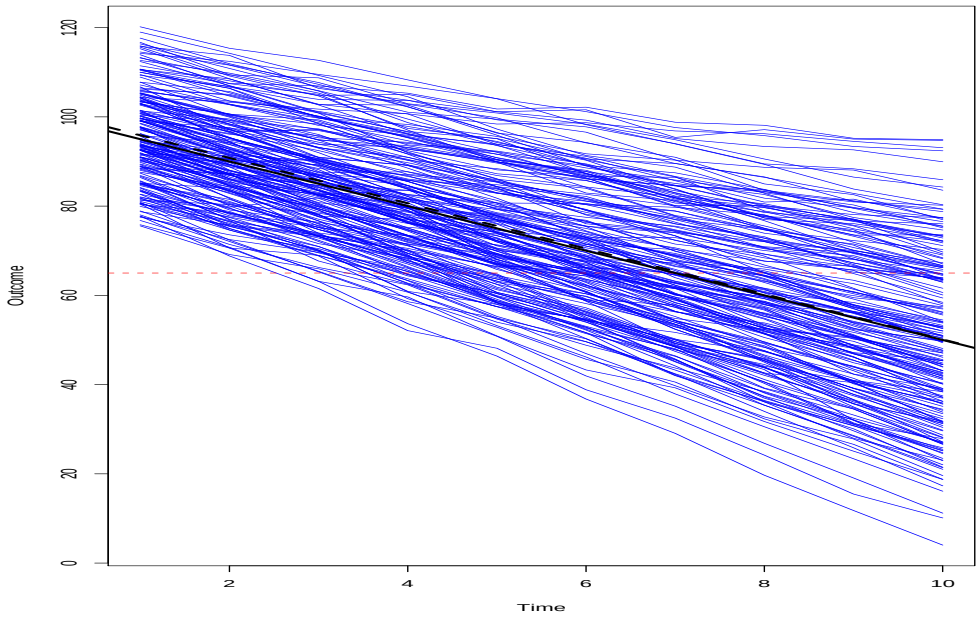


Figure 59: Full simulated data set: solid line is truth and dashed the LS line.

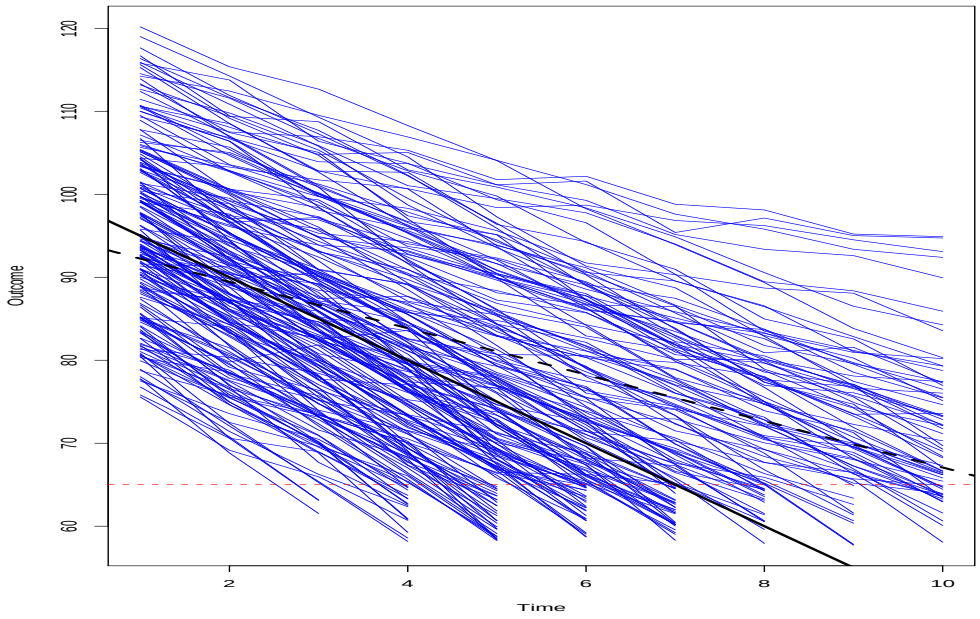


Figure 60: Simulated data set with drop-out: solid line is truth and dashed the LS line.

Mechanisms of Missingness

The impact of missing data depends crucially on the mechanism of missingness, that is the probability model for missingness.

We let $\boldsymbol{R}_i$ be a vector of response indicators for the i-th units so that

$$R_{ij} = \begin{cases} 1 & \text{if } Y_{ij} \text{ is observed} \\ 0 & \text{if } Y_{ij} \text{ is missing} \end{cases}$$

We partition the complete data vector $\boldsymbol{Y}_i = (\boldsymbol{Y}_i^O, \boldsymbol{Y}_i^M)$ into those components that are observed, $\boldsymbol{Y}_i^O$, and those that are missing $\boldsymbol{Y}_i^M$.

There are two ways of factroring the data:

$$\begin{aligned} p(\boldsymbol{Y}, \boldsymbol{R} \mid \boldsymbol{x}) &= p(\boldsymbol{Y} \mid \boldsymbol{x}) \times p(\boldsymbol{R} \mid \boldsymbol{Y}, \boldsymbol{x}) \\ p(\boldsymbol{Y}, \boldsymbol{R} \mid \boldsymbol{x}) &= p(\boldsymbol{Y} \mid \boldsymbol{R}, \boldsymbol{x}) \times p(\boldsymbol{R} \mid \boldsymbol{x}) \end{aligned}$$

The first is known as a selection model (individuals are selected according to their outcome), and the second as a pattern mixture model (we "mix" pattern specific models). We concentrate on the former.

Three situations are distinguished:

1. Missing completely at random (MCAR).
2. Missing at random (MAR).
3. Not missing at random (NMAR).

each of which we now discuss in detail.

Unfortunately the terminology is confusing!

Missing Completely at Random (MCAR)

Data are MCAR if

$$\Pr(R_{ij} = 1 \mid \boldsymbol{Y}^O, \boldsymbol{Y}^M, \boldsymbol{x}) = \Pr(R_{ij} \mid \boldsymbol{x}),$$

so that the missingness does not depend on the response data, observed or unobserved.

This implies that

$$\mathrm{E}[Y_{ij}|R_{ij} = 1, \boldsymbol{X}_i] = \mathrm{E}[Y_{ij}|\boldsymbol{X}_i]$$

No selection bias.

Missing at Random (MAR)

Data are MCAR if

$$\Pr(R_{ij} = 1 \mid \boldsymbol{Y}^O, \boldsymbol{Y}^M, \boldsymbol{x}) = \Pr(R_{ij} \mid \boldsymbol{Y}^O, \boldsymbol{x}),$$

so that the missingness may depend on observed values.

This implies that

$$\mathrm{E}[Y_{ij}|R_{ij} = 1, \boldsymbol{X}_i] \neq \mathrm{E}[Y_{ij}|\boldsymbol{X}_i]$$

which suggests that the GEE approach might be in trouble in terms of biased parameter estimates.

Not Missing at Random (NMAR)

If the missingness depends on $\boldsymbol{Y}^M$, i.e.

$$\Pr(R_{ij} = 1 \mid \boldsymbol{Y}^O, \boldsymbol{Y}^M, \boldsymbol{x}) = \Pr(R_{ij} \mid \boldsymbol{Y}^O, \boldsymbol{Y}^M, \boldsymbol{x}).$$

In this case the mechanism is also sometimes referred to as non-ignorable.

This selection bias is not fixable, since we don't know the outcomes that caused the problems. Models can be postulated, but are not checkable from the observed data alone.

In general it is obviously best if we know why the data are missing.

Approaches

Complete-case analysis

A simple approach is to exclude units that did not provide data at all intended occasions. Clearly there is a loss of information in this process, and bias will result unless the data are MCAR. Not to be recommended.

Available-case analysis

This approach uses the largest set of available data for estimating parameters. Will provide biased estimates unless the data are MCAR.

Last observation carried forward

In a longitudinal setting we could simply "fill-in" the missing values, extrapolating from the last observed value. As a general method not to be recommended.

Imputation

An appealing approach is to "fill-in", or impute, the missing values and then carry out a conventional analysis. Complex models for the missingness can be incorporated (closely related to *data augmentation* which we describe later).

Likelihood-based approach

Let $\boldsymbol{\theta}$ be the parameters of the model for $\boldsymbol{Y}$, and $\boldsymbol{\phi}$ the parameters for $\boldsymbol{R}$.

In general, a natural way to decompose the data as

$$\begin{aligned} p(\boldsymbol{Y}^O, \boldsymbol{Y}^M, \boldsymbol{R} \mid \boldsymbol{\theta}, \boldsymbol{\phi}) &= p(\boldsymbol{Y}^O, \boldsymbol{Y}^M \mid \boldsymbol{\theta}, \boldsymbol{\phi}) \times \Pr(\boldsymbol{R} \mid \boldsymbol{Y}^O, \boldsymbol{Y}^M, \boldsymbol{\theta}, \boldsymbol{\phi}) \\ &= p(\boldsymbol{Y}^O, \boldsymbol{Y}^M \mid \boldsymbol{\theta}) \times \Pr(\boldsymbol{R} \mid \boldsymbol{Y}^O, \boldsymbol{Y}^M, \boldsymbol{\phi}) \end{aligned}$$

where we have also assumed that the data and missingness models have distinct parameters.

We require a distribution for the observed data, $\boldsymbol{Y}^O, \boldsymbol{R}$:

$$p(\boldsymbol{Y}^O, \boldsymbol{R} \mid \boldsymbol{\theta}, \boldsymbol{\phi}) = \int p(\boldsymbol{Y}^O, \boldsymbol{Y}^M \mid \boldsymbol{\theta}) \times \Pr(\boldsymbol{R} \mid \boldsymbol{Y}^O, \boldsymbol{Y}^M, \boldsymbol{\phi}) \, d\boldsymbol{Y}^M.$$

This is an example of a selection model.

Suppose we are in the MAR situation so that

$$\Pr(\boldsymbol{R} \mid \boldsymbol{Y}^O, \boldsymbol{Y}^M, \boldsymbol{\phi}) = \Pr(\boldsymbol{R} \mid \boldsymbol{Y}^O, \boldsymbol{\phi}).$$

In this situation the likelihood is given by

$$\begin{aligned} p(\boldsymbol{Y}^O, \boldsymbol{R} \mid \boldsymbol{\theta}, \boldsymbol{\phi}) &= \int p(\boldsymbol{Y}^O, \boldsymbol{Y}^M \mid \boldsymbol{\theta}) \, d\boldsymbol{Y}^M \times \Pr(\boldsymbol{R} \mid \boldsymbol{Y}^O, \boldsymbol{\phi}) \\ &= p(\boldsymbol{Y}^0 \mid \boldsymbol{\theta}) \times \Pr(\boldsymbol{R} \mid \boldsymbol{Y}^O, \boldsymbol{\phi}) \end{aligned}$$

Hence we have the log-likelihood

$$\log p(\boldsymbol{Y}^0 \mid \boldsymbol{\theta}) + \log \Pr(\boldsymbol{R} \mid \boldsymbol{Y}^O, \boldsymbol{\phi})$$

and can ignore the second term and don't have to model the missingness mechanism.

Important Point: We need to get the model right!!!

Simulation Study

Model	$\widehat{\beta}_1$	(s.e.)	$\widehat{\beta}_2$	(s.e.)
GEE ind	101.0	0.715	−5.084	0.109
GEE exch	101.0	0.715	−5.084	0.109
LMEM1	101.0	0.981	−5.084	0.037
LMEM2	101.0	0.720	−5.084	0.109
GEEiDO ind	95.0	0.894	−2.796	0.134
GEEDO exch	98.8	0.787	−4.304	0.114
LMEM1DO	98.8	0.837	−4.282	0.041
LMEM2DO	100.1	0.722	−5.097	0.112

Table 15: Results of GEE and LMEM analyses of full and drop-out simulated data, LMEM1 is random intercepts only, LMEM2 is random intercepts and slopes.

- Bias for GEE is bad (particularly working independence).
- Bias for LMEM if we only assume random intercepts (terrible se on $\widehat{\beta}_2$).
- LMEM with random intercepts and slopes recovers the truth.

Models for Drop-out

If the missingness is monotone, in the sense that if $R_{ij} = 0$ then $R_{ik} = 0$ for all $k > j$, then we define the drop-out time as

$$D_i = \min_k \{R_{ik} = 0\}.$$

Hence $2 \leq D_i \leq n_i + 1$, with $D_i = n_i + 1$ for an individual that does not drop out.

The reason for drop-out may be that the individual was not responding well, and their outcomes reflect this.

To examine this possibility we could fit logistic models of the form:

$$\log\left(\frac{\Pr(D_i = k|D_i \geq k, Y_{i1}, ..., Y_{ik})}{\Pr(D_i > k|D_i \geq k, Y_{i1}, ..., Y_{ik})}\right) = \phi_0 + \phi_1 Y_{ik-1}$$

and look for evidence that $\phi_1 \neq 0$.

Bayesian Inference via Data Augmentation

Data augmentation is a auxiliary variable method that treats the missing observations as unknown parameters – this can lead to simple MCMC schemes.

General formulation: we have posterior

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{Y}^M \mid \boldsymbol{Y}^O) &= p(\boldsymbol{\theta} \mid \boldsymbol{Y}^M, \boldsymbol{Y}^O) p(\boldsymbol{Y}^M \mid \boldsymbol{Y}^O) \\ &= p(\boldsymbol{Y}^M \mid \boldsymbol{\theta}, \boldsymbol{Y}^O) p(\boldsymbol{\theta} \mid \boldsymbol{Y}^O) \end{aligned}$$

MCMC scheme:

1. Auxiliary variables:
$$\boldsymbol{Y}^M \sim p(\boldsymbol{Y}^M \mid \boldsymbol{Y}^O, \boldsymbol{\theta}).$$
2. Model parameters:
$$\boldsymbol{\theta} \sim p(\boldsymbol{\theta} \mid \boldsymbol{Y}^O, \boldsymbol{Y}^M).$$

The auxiliary variable scheme may be modified to $p(\boldsymbol{Y}^M \mid \boldsymbol{Y}^O, \boldsymbol{\theta}) \sim p(\boldsymbol{Y}^M \mid \boldsymbol{\theta})$, depending on the missing data model, as we now illustrate.

Example: Censoring Model

Suppose we have data Y_i measured at times t_i, $j = 1, ..., n$, but measurements *below the lower limit of detection*, D (assumed known) are not recorded. Also suppose that the data generating model (likelihood) is:

$$Y \mid \boldsymbol{\beta}, \sigma \sim_{ind} \mathrm{N}(\eta(\boldsymbol{\beta}, t), \sigma^2).$$

Clearly setting such measurements to zero or ignoring the measurements will lead to bias in estimation.

Figure 61 illustrates for a set of simulated data in which the true slope was -0.01; the slope estimates are -0.0099, -0.0095 and -0.0087 for the full data, set equal to D and ignored schemes, respectively.

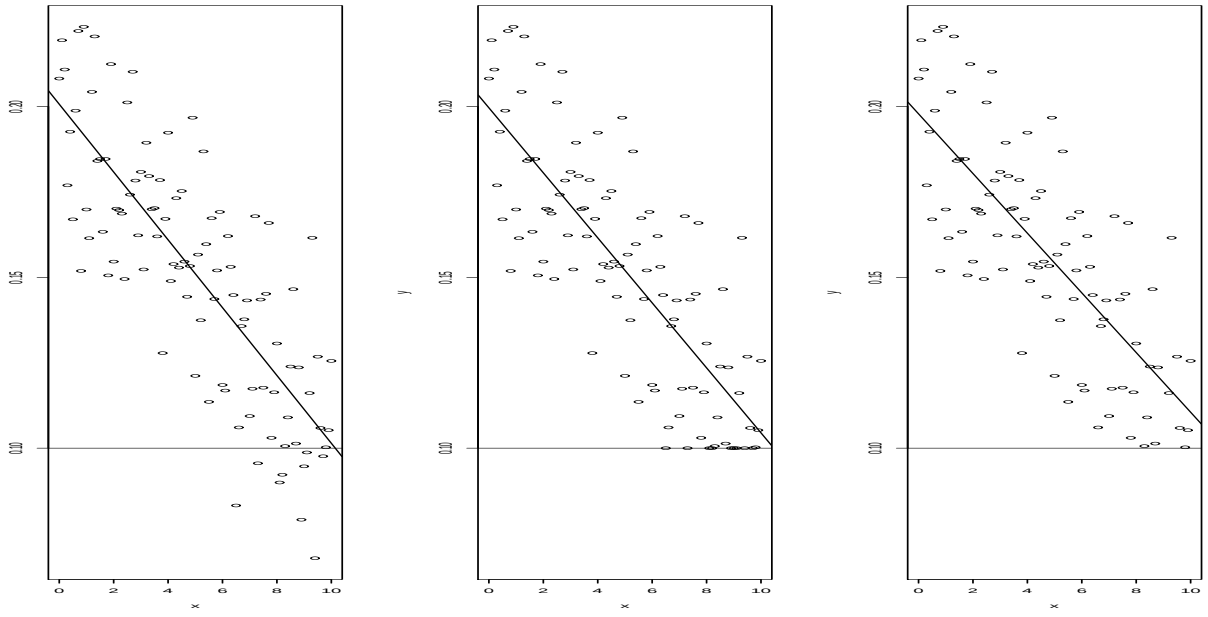


Figure 61: All data (left), assigned to lower limit (middle), ignored (right). Horizontal line is the lower limit of detection.

Suppose that the last c measurements are censored, the remaining $n-c$ being uncensored. Then

$$\begin{aligned} p(\boldsymbol{y} \mid \theta) &= \prod_{i=1}^{n-c} p(y_i \mid \boldsymbol{\beta}, \sigma^2) \prod_{i=c+1}^{n} \Pr(Y_i < D \mid \boldsymbol{\beta}, \sigma^2) \\ &= \prod_{i=1}^{n-c} \phi\left(\frac{y_i - \eta(\boldsymbol{\beta}, t_i)}{\sigma}\right) \prod_{i=c+1}^{n} \Phi\left(\frac{D - \eta(\boldsymbol{\beta}, t_i)}{\sigma}\right) \end{aligned}$$

where

$$\phi(z) = (2\pi)^{-1/2} \exp(-z^2/2)$$

and

$$\Phi(z_0) = \Pr(Z < z_0) = \int_{-\infty}^{z_0} \phi(z)\, dz$$

where Z is an $\mathrm{N}(0,1)$ random variable.

To perform likelihood or Bayesian inference we need to numerically evaluate the distribution function of a normal distribution for each likelihood calculation.

Data Augmentation Scheme

Letting $\boldsymbol{Y}^O = \{Y_i, i = 1, ..., n-c\}$ and $\boldsymbol{Y}^M = \{Y_i, i = n-c+1, ..., n\}$, we iterate between

1. $y_i \mid \boldsymbol{\beta}, \sigma \sim \text{TruncNorm}(\eta(\boldsymbol{\beta}, t_i), \sigma^2)$, on $(-\infty, D)$, $i = n-c+1, ..., n$.
2. $\boldsymbol{\beta} \mid y_1, ..., y_n, \sigma^2 \propto \prod_{i=1}^n p(y_i \mid \boldsymbol{\beta}, \sigma^2)\pi(\boldsymbol{\beta})$. Usual (uncensored) posterior.
3. $\sigma^2 \mid y_1, ..., y_n, \boldsymbol{\beta} \propto \prod_{i=1}^n p(y_i \mid \boldsymbol{\beta}, \sigma^2)\pi(\sigma^2)$. Usual (uncensored) posterior.

We give an example in a different context — survival analysis with censored data.

Related Example: Survival Analysis with Censored Data

From the `WinBUGS` manual: *Mice: Weibull regression*

Dellaportas and Smith (1993) analyse data from Grieve (1987) on photocarcinogenicity in four groups, each containing 20 mice, who have recorded a survival time and whether they died or were censored at that time.

A portion of the data, giving survival times in weeks, are shown below.

A * indicates censoring.

Mouse	Irradia control	Vehicle control	Test substan	Positive control
1	12	32	22	27
.....				
18	*40	30	24	12
19	31	37	37	17
20	36	27	29	26

The survival distribution is assumed to be Weibull. That is

$$p(t_i, z_i) = r \exp(\boldsymbol{\beta} \boldsymbol{z}_i) t_i^{r-1} \exp\{-\exp(\boldsymbol{\beta} \boldsymbol{z}_i t_i r)\}$$

where t_i is the failure time of an individual with covariate vector $\boldsymbol{z}_i$ and $\boldsymbol{\beta}$ is a vector of unknown regression coefficients. The baseline hazard is given by

$$\lambda_0(t_i) = r t_i^{r-1}.$$

Setting $\mu_i = \exp(\boldsymbol{\beta} \boldsymbol{z}_i)$ gives the parameterization

$$t_i \sim \text{Weibull}(r, \mu_i)$$

For censored observations, the survival distribution is a truncated Weibull, with lower bound corresponding to the censoring time. The regression coefficients $\boldsymbol{\beta}$ are assumed a priori to follow independent Normal distributions with zero mean and "vague" precision 0.0001. The shape parameter r for the survival distribution was given a Gamma(1, 0.0001) prior, which is slowly decreasing on the positive real line.

Median survival for individuals with covariate vector $\boldsymbol{z}_i$ is given by

$$m_i = (\log 2 \exp(-\boldsymbol{\beta} \boldsymbol{z}_i))^{1/r}.$$

WinBUGS code

```
model
{
        for(i in 1 : M) {
                for(j in 1 : N) {
                        t[i, j] ~ dweib(r, mu[i])I(t.cen[i, j],)
                }
                mu[i] <- exp(beta[i])
                beta[i] ~ dnorm(0.0, 0.001)
                median[i] <- pow(log(2) * exp(-beta[i]), 1/r)
        }
        r ~ dexp(0.001)
        veh.control <- beta[2] - beta[1]
        test.sub <- beta[3] - beta[1]
        pos.control <- beta[4] - beta[1]
}
```

```
list(t = structure(.Data = c(12, 1, 21, 25, 11, 26, 27, 30, 13, 12,
21, 20, 23, 25, 23, 29, 35, NA, 31, 36, 32, 27, 23, 12, 18, NA, NA,
38, 29, 30, NA, 32, NA, NA, NA, NA, 25, 30, 37, 27, 22, 26, NA, 28,
19, 15, 12, 35, 35, 10, 22, 18, NA, 12, NA, NA, 31, 24, 37, 29, 27,
18, 22, 13, 18, 29, 28, NA, 16, 22, 26, 19, NA, NA, 17, 28, 26, 12,
17, 26), .Dim = c(4, 20)), t.cen = structure(.Data = c( 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 40, 0, 0, 0, 0, 0, 0, 0, 40, 40,
0, 0, 0, 40, 0, 40, 40, 40, 40, 0, 0, 0, 0, 0, 0, 10, 0, 0, 0, 0, 0,
0, 0, 0, 0, 24, 0, 40, 40, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 20, 0, 0,
0, 0, 29, 10, 0, 0, 0, 0, 0, 0), .Dim = c(4, 20)), M = 4, N = 20)
```

We note a number of tricks in setting up this model.

First, individuals who are censored are given a missing value in the vector of failure times t, whilst individuals who fail are given a zero in the censoring time vector t.cen.

The truncated Weibull is modelled using I(t.cen[i],) to set a lower bound. Second, we set a parameter beta[j] for each treatment group j. The contrasts beta[j] with group 1 (the irradiated control) are calculated at the end. Alternatively, we could have included a grand mean term in the relative risk model and constrained beta[1] to be zero.

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
median[1]	23.9	1.967	0.05889	20.3	23.89	28.09	1001	10000
median[2]	35.2	3.359	0.04757	29.46	34.93	42.64	1001	10000
median[3]	26.9	2.383	0.0582	22.62	26.79	31.91	1001	10000
median[4]	21.4	1.799	0.03362	18.2	21.32	25.36	1001	10000
pos.control	0.3409	.3457	0.00723	-0.327	0.3429	1.009	1001	10000
r	3.03	0.3182	0.02749	2.388	3.045	3.64	1001	10000
test.sub	-0.351	0.3459	0.004433	-1.035	-0.3541	0.3303	1001	10000
veh.control	-1.16	0.3679	0.005974	-1.893	-1.156	-0.444	1001	10000

GEE Approaches

Suppose that if the full data had been observed there would have been n_i observations on each individual, $i = 1, ..., m$.

We write the usual estimating equation as

$$\boldsymbol{G}(\boldsymbol{\beta}) = \sum_{i=1}^{m} \boldsymbol{D}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} \boldsymbol{R}_i (\boldsymbol{Y}_i - \boldsymbol{\mu}_i)$$

where $\boldsymbol{R}_i$ is the diagonal matrix with elements R_{ij}, $j = 1, ..., n_i$.

For the estimator, $\widehat{\boldsymbol{\beta}}$ to be consistent we require $\boldsymbol{G}$ to be unbiased. The random variables are now $\boldsymbol{Y}, \boldsymbol{R}$ and so we have

$$\begin{aligned}
\mathrm{E}_{Y,R}[\boldsymbol{G}(\boldsymbol{\beta})] &= \mathrm{E}_R\{\mathrm{E}_{Y|R}[\boldsymbol{G}(\boldsymbol{\beta})]\} \\
&= \sum_{i=1}^{m} \mathrm{E}_{R_i}\{\mathrm{E}_{Y_i|R_i}[\boldsymbol{D}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} \boldsymbol{R}_i (\boldsymbol{Y}_i - \boldsymbol{\mu}_i)]\} \\
&= \sum_{i=1}^{m} \mathrm{E}_{R_i}\{\boldsymbol{D}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} \boldsymbol{R}_i \mathrm{E}_{Y_i|R_i}[\boldsymbol{Y}_i - \boldsymbol{\mu}_i]\} \\
&= \sum_{i=1}^{m} \mathrm{E}_{R_i}\{\boldsymbol{D}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} \boldsymbol{R}_i \left(\mathrm{E}_{Y_i|R_i}[\boldsymbol{Y}_i] - \boldsymbol{\mu}_i\right)\}
\end{aligned}$$

Hence, to obtain an unbiased estimating equation we require

$$\mathrm{E}[\boldsymbol{Y}_i \mid \boldsymbol{R}_i, \boldsymbol{x}_i] = \mathrm{E}[\boldsymbol{Y}_i \mid \boldsymbol{x}_i] = \boldsymbol{\mu}_i$$

so that we are fine under MCAR but not under MAR, since the distribution of $\boldsymbol{Y}_i \mid \boldsymbol{x}_i, \boldsymbol{R}_i$ is different from that of $\boldsymbol{Y}_i \mid \boldsymbol{x}_i$ under MAR.

To rectify the situation we need to modify the usual estimating equation.

Let

$$\pi_{ij} = \mathrm{E}[R_{ij} \mid \boldsymbol{x}_i, \boldsymbol{H}_{i,j-1}]$$

where $\boldsymbol{H}_{i,j-1} = (Y_{i1}, ..., Y_{i,j-1})$ contains the "history" of responses.

Consider the estimating equation:

$$\sum_{i=1}^{m} \boldsymbol{D}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} \boldsymbol{P}_i (\boldsymbol{Y}_i - \boldsymbol{\mu}_i)$$

where $\boldsymbol{P}_i$ is a diagonal matrix which contains terms R_{ij}/π_{ij}, for $j = 1, ..., n_i$.

We have

$$\mathrm{E}_Y \left\{ \sum_{i=1}^{m} \mathrm{E}_{R|Y} \left[\boldsymbol{D}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} \boldsymbol{P}_i (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) \right] \right\} = \mathrm{E}_Y \left\{ \sum_{i=1}^{m} \boldsymbol{D}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} \mathrm{E}_{R|Y}[\boldsymbol{P}_i](\boldsymbol{Y}_i - \boldsymbol{\mu}_i) \right\} = 0$$

since $\mathrm{E}[\boldsymbol{P}_i] = \boldsymbol{I}$ if π_{ij} is correctly specified.

In both GEE and likelihood we are basically accounting for the biased sampling scheme of MAR; likelihood does this by assuming a model, while GEE adjusts by modeling the probabilities of seeing the data.