
THE
LADDER
OF
INFERENCE
I TAKE
ACTION
(BASED ON MY
BELIEFS)
I ADOPT
BELIEFS
(ABOUT THE
WORLD)
I DRAW
CONCLUSION
(BASED ON
ASSUMPTION)
I MAKE
ASSUMPTION
(BASED ON
MEANINGS)
I ADD
MEANINGS
(CULTURE &
PERSONAL)
I SELECT
DATA
FROM WHAT
I OBSERVE
OBSERVABLE
"DATA" AND
EXPERIENCES
(AS A MIGHT
CAPTURE IT)
THE REFLEXIVE LOOP
(OUR BELIEFS AT LEAST WHAT DATA WE SELECT NEXT TIME)
IKIN
INFORMATIKA KARYA
INDAH NUANSA

Inference on the basis of a single model fit

```
summary(model1) # a poisson glm model with one covariate

             Estimate  Std. Error       z value     Pr(>|z|)

(Intercept) 0.186865       0.188728 0.990       0.3221

Distance   -0.006138       0.003667 -1.674      0.0941 .


(Dispersion parameter for poisson family taken to be 1)


    Null deviance: 149.48  on 93  degrees of freedom

Residual deviance: 146.64  on 92  degrees of freedom
```

Inference based on model comparison (pairwise – many modes)

```
drop1(model,test="F") # F test for slope, glm where dispersion is estimated

Single term deletions

Model:
y ~ x
       Df  Deviance    AIC     F value    Pr(F)
<none>      0.236    -26.071
x      1    58.795    82.324   4474.4     < 2.2e-16 ***
```

Statistical Analysis is used for **inductive inference**:

From a particular dataset to a more general population

The "process of drawing conclusions from data that are subject to random variation"

- 3 schools (Frequentist / Bayesian / Likelihood)

- 2 cultures (Breiman 2001)

Two cultures:

- Predicting other/new events as accurately as possible

- **Discovering mechanisms that generated the data**

either assuming that the truth is among the candidate models or not

The third way (the third school)

(Direct) **Likelihood inference**

- All conclusions within the **modeling** exercise and the **data**, no reference to external probability or replication
- Weaker than frequentist inference which should be preferred if possible

Sir R Fisher

Example: direct likelihood Interval

likelihood of a parameter value: probability that the data x occur, given a parameter value θ and model f

$$L(\theta \mid x) = f(x \mid \theta)$$

$L(\hat{\theta})$ is the largest possible likelihood over the range of θ and $\hat{\theta}$ is the maximum likelihood value of θ

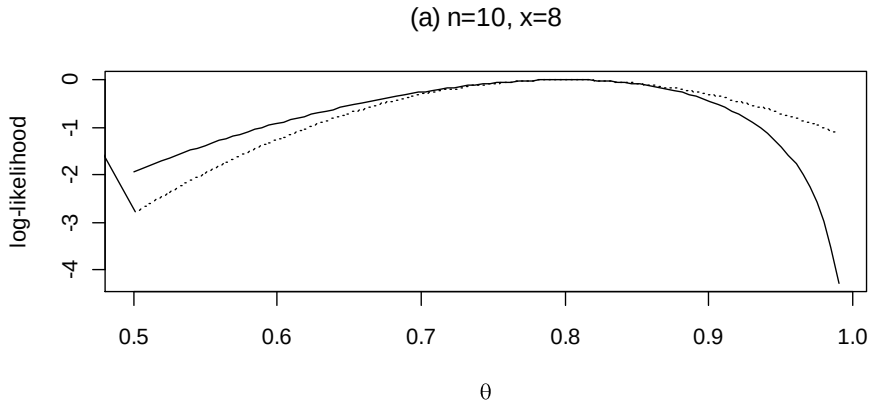
A likelihood interval are values of θ for which the relative/normalized likelihood is larger than some cutoff value c ($0 < c < 1$)

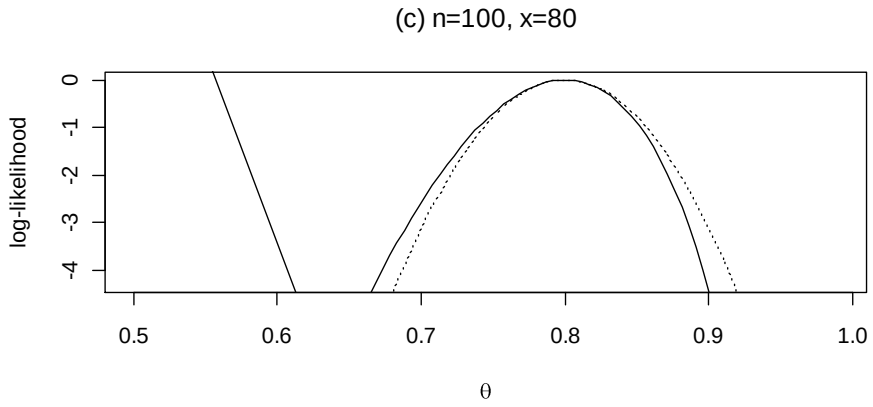
$$\left\{\theta, \frac{L(\theta)}{L(\hat{\theta})} > c\right\}$$

10 trials with 8 successes, binomial parameter $\hat{\theta} = 0.8$ and $I(\hat{\theta}) = 62.5$

100 trials with 80 successes, binomial parameter $\hat{\theta} = 0.8$ and $I(\hat{\theta}) = 625$

write the R code to estimate θ and calculate the likelihood intervals.





Note the dotted lines

Dotted lines: Quadratic approximation at the ML parameter values
Often used to obtain confidence intervals

$$\ln L(\theta) = \ln L(\hat{\theta}) + \left.\frac{\partial \ln L}{\partial \theta}\right|_{\hat{\theta}} (\theta - \hat{\theta}) + \frac{1}{2}\left.\frac{\partial^2 \ln L}{\partial \theta^2}\right|_{\hat{\theta}} (\theta - \hat{\theta})^2 + h.o.t$$

$$= \ln L(\hat{\theta}) + \frac{1}{2}\left.\frac{\partial^2 \ln L}{\partial \theta^2}\right|_{\hat{\theta}} (\theta - \hat{\theta})^2 + h.o.t$$

$$\approx \ln L(\hat{\theta}) - \frac{1}{2} I(\hat{\theta})(\theta - \hat{\theta})^2$$

- I is the observed Fisher information ~ the curvature of the likelihood at the MLE estimate of the parameter

With *several* parameters in a model, I becomes a matrix

Frequentist inference

Example

The data are 11 heavy sleepers, 16 others are not heavy sleeping

If the proportion of heavy sleepers is 50 % in the population, then the probability to find these data, or more extreme proportions is twice a sum of probabilities:

```
pbinom(11,27,0.5) # probability of 11 or fewer successes

[1] 0.2210342

 binom.test(c(11,16))
        Exact binomial test
number of successes = 11, number of trials = 27, p-value = 0.4421
alternative hypothesis: true probability of success is not equal to 0.5
```

Frequentist inference

- Probability is a long-run frequency based on repeated sampling/experiments

- The **repeated sampling principle**: procedures should be evaluated on the basis of repeated experimentation in the same conditions

Sampling distribution theory

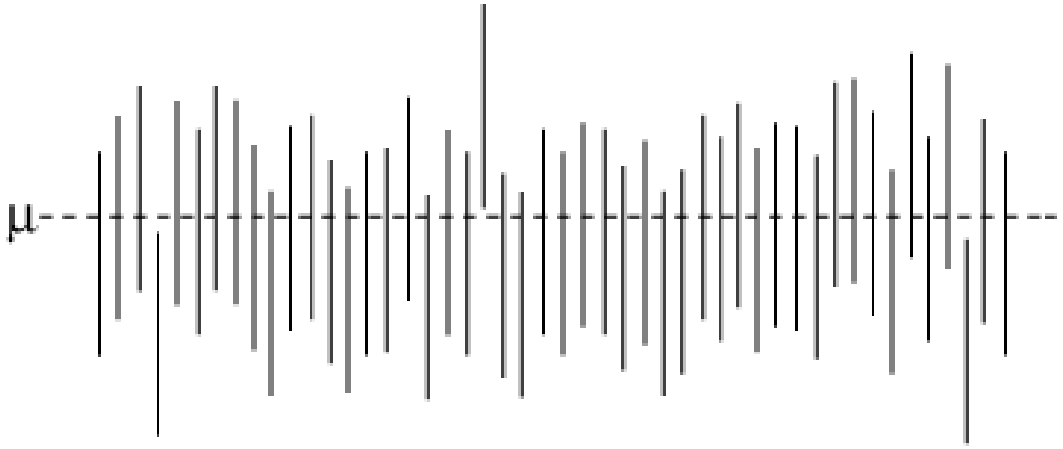
Jerzy Neyman 

- The repeated sampling principle: procedures should be evaluated on the basis of repeated experimentation in the same conditions

- **Hypothetical requirement** – replication can be impossible

- Sometimes concepts not in line with intuitive expectation of uncertainty. e.g. : confidence intervals

• Confidence interval: if the procedure is repeated and a 95% interval is constructed each time, it will contain the true value in 95% of the experiments.



Assume data x_1, ..., x_n are a sample from a normal distribution with known standard deviation σ and unknown mean θ

Approximation of likelihood of a parameter value θ given the MLE estimate

$$\ln L(\theta) = \ln L(\hat{\theta}) - \frac{1}{2} I(\hat{\theta})(\theta - \hat{\theta})^2$$ [for the normal distribution the equality holds]

Now θ is the true value, what is the relative likelihood of the MLE given the true value?

$$\ln L(\theta) - \ln L(\hat{\theta}) = -\frac{1}{2} I(\hat{\theta})(\theta - \hat{\theta})^2 = -\frac{n}{2\sigma^2}(\theta - \bar{x})^2$$

$$\ln L(\hat{\theta}) - \ln L(\theta) = \frac{1}{2} I(\hat{\theta})(\theta - \hat{\theta})^2 = \frac{n}{2\sigma^2}(\bar{x} - \theta)^2$$

It has been determined that a squared standardized normal random variable has a chi-squared distribution, so considering repeated sampling

$$\frac{n}{\sigma^2}\left(\bar{x}-\theta\right)^2 \sim \chi_1^2$$

and $2\left(\ln L\left(\hat{\theta}\right)-\ln L\left(\theta\right)\right)\sim \chi_1^2$

if we draw repeated samples from the same population, with parameters σ (known) and θ (unknown), the distribution of this statistic will be chi-squared with 1 d.f.

$$2\left(\ln L\left(\hat{\theta}\right) - \ln L(\theta)\right) \sim \chi_1^2$$

Repeated sampling We don't know what the true value is. If we calculate direct likelihood intervals on each sample:

The **probability** that the likelihood interval covers the true value θ is determined by choosing *c*

i.e. choose *c* such that the true value has high enough likelihood to be in the interval sufficiently often

$$P\left(\frac{L(\theta)}{L\left(\hat{\theta}\right)} > c\right) = P\left(2\ln L\left(\hat{\theta}\right) - 2\ln L(\theta) < -2\ln c\right)$$
$$= P\left(\chi_1^2 < -2\ln c\right)$$

if we choose *c* = 0.1465 this probability will be 95% pchisq(-2*log(0.1465),1)

```
confint(model1) # using the diagonal of the information matrix
# when model1 is fitted using lm()
```

For a multi-parameter model which includes a parameter θ,
the profile likelihood of θ is a likelihood function of θ alone.

The profile likelihood at θ is the likelihood of the parameter value θ while all other parameters are fixed at their **ML** values for that value of θ.
All other parameters are seen as functions of θ.

Profile likelihood confidence intervals
There is a glm-specific confidence interval method:

```
library(MASS)
confint(model1) # profile likelihood confidence intervals
```

Hypothesis testing

Null Hypothesis - **Alternative** Hypothesis

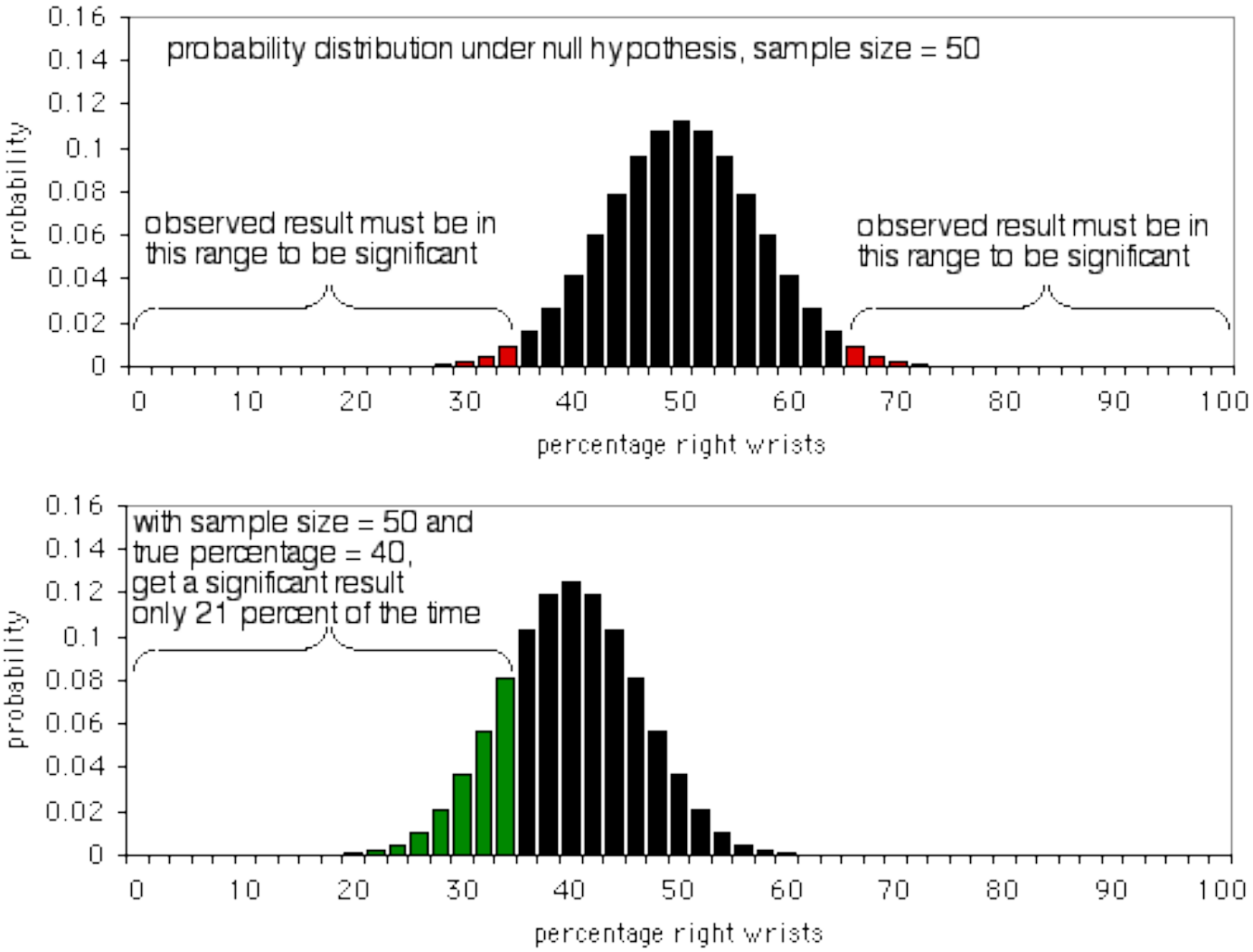
General procedure: assuming the Null is true, what is the probability that the value of the statistic or a more extreme value occurs?

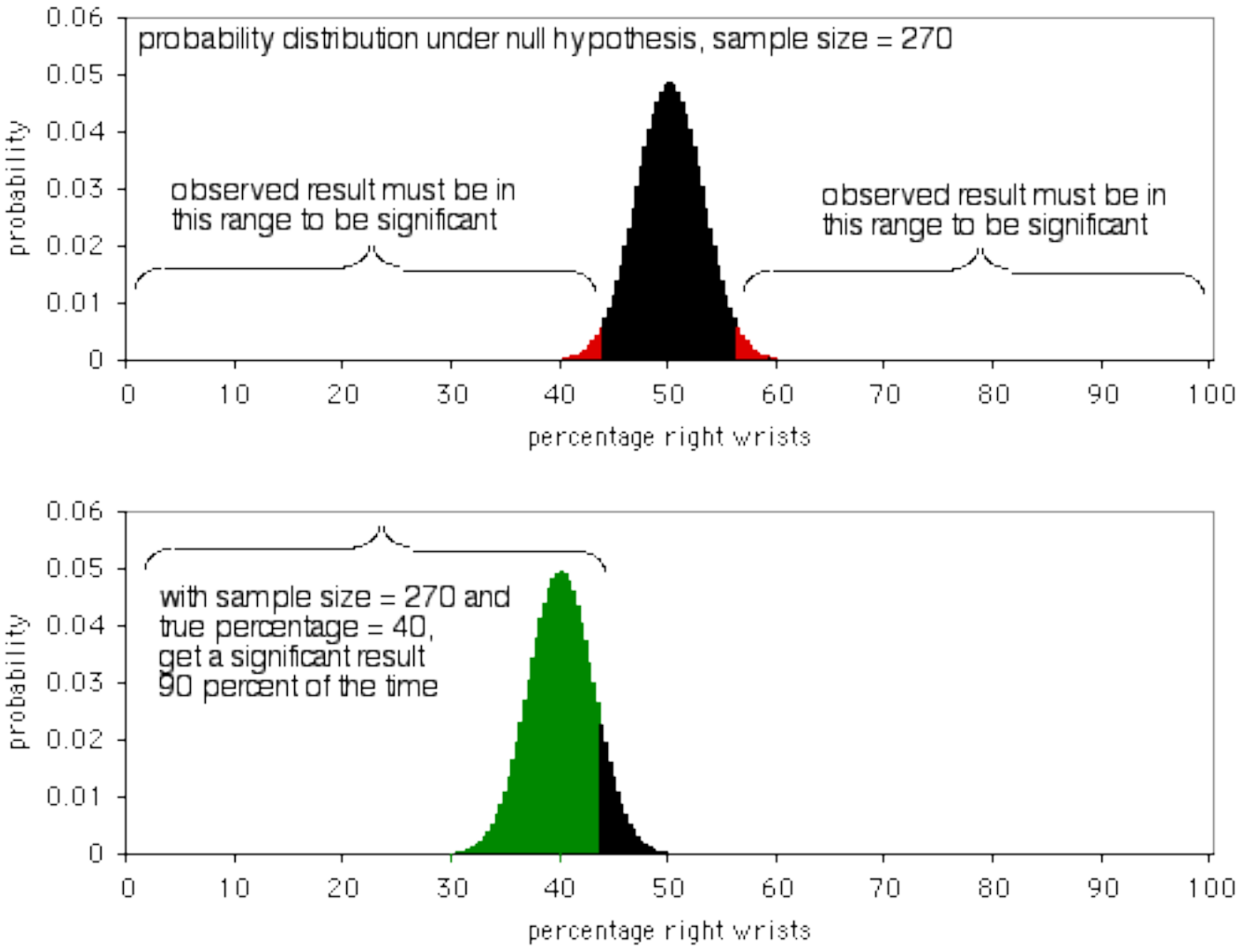
Tail probability of a test (*p*-value). → reject the Null when the value is improbable ($p < 0.05$ for $\alpha = 0.05$).

Type I error: The Null hypothesis is true, but we reject it in favour of the alternative. The frequency at which this happens is controlled to be **not more** than α. (if we can't control exactly, we will be **conservative**).

Type II error: The alternative hypothesis is true, but we don't reject the null hypothesis. This occurs with probability β. Power: 1 - β

Frequentist approach - **Power analysis –** based on a statistic





Usually, a power analysis has to be carried out conditional **on effect sizes** and **sample sizes**.

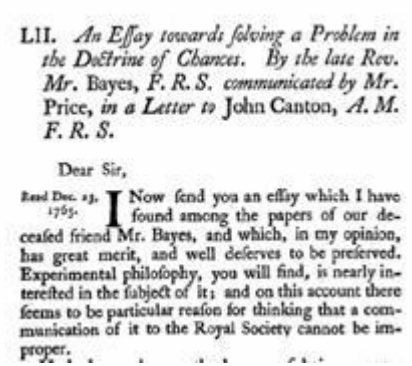
The chi-squared statistic also appears in likelihood ratio tests, it is only exact for data from a normal distribution with known standard deviation.

How can we check whether an approximate hypothesis testing procedure that we use performs reasonably well?

(hint: simulate appropriate data, do the test, then...)

Can become involved for a model selection procedure with multiple model comparisons

- **Bayesian** approach



LII. *An Essay towards solving a Problem in the Doctrine of Chances. By the late Rev. Mr.* Bayes, *F. R. S. communicated by Mr.* Price, *in a Letter to* John Canton, *A. M. F. R. S.*

Dear Sir,

Read Dec. 23, 1763. I Now send you an essay which I have found among the papers of our deceased friend Mr. Bayes, and which, in my opinion, has great merit, and well deserves to be preserved. Experimental philosophy, you will find, is nearly interested in the subject of it; and on this account there seems to be particular reason for thinking that a communication of it to the Royal Society cannot be improper.

Bayesian inference: unified approach to all problems of uncertainty

The notion of probability includes the **subjective** notion of uncertainty

- Model structure is fixed, all parameters are **random**

- All parameters θ are samples from (prior) probability distributions $f(\theta)$

- We start with a **prior** distribution, collect data *x*, use that to obtain the

posterior distribution $f(\theta \mid x) = \dfrac{f(x \mid \theta) f(\theta)}{f(x)}$ of the parameters

- These distributions express our (changing) uncertainty about θ

- Prior distributions can be chosen axiomatically (by thinking alone) or not and be processed to produce a posterior distribution

Example from Albert (2009)

Investigating the proportion p of heavy sleepers: who sleeps eight hours or more?

data: s success, f failure; total number of observations: $f + s$

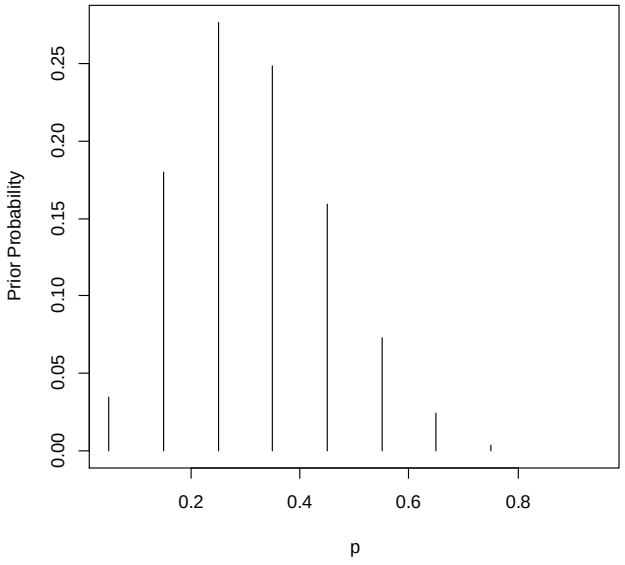
probability of the data ~ $p^s (1 - p)^f$

given proportion p heavy sleepers in the population

Assuming a discrete prior probability distribution

```
library(LearnBayes)

p = seq(0.05, 0.95, by = 0.1)
prior = c(1, 5.2, 8, 7.2, 4.6, 2.1, 0.7, 0.1, 0, 0)
prior = prior/sum(prior)
plot(p, prior, type = "h", ylab="Prior Probability")
```

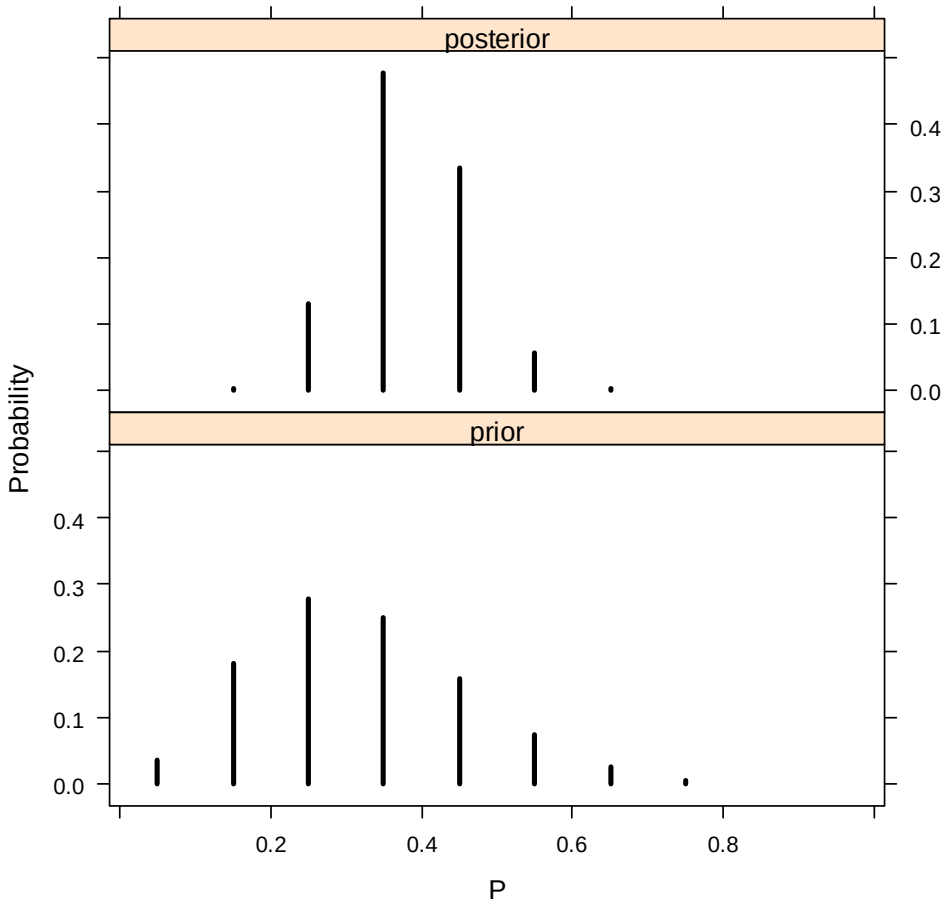


```
data = c(11, 16)# 11 heavy sleepers, 16 not heavy
post = pdisc(p, prior, data)# calculate posterior for a
list of proportions p

round(cbind(p, prior, post),2)

 library(lattice)# prepare to make a graph of the posterior
 PRIOR=data.frame("prior",p,prior)
 POST=data.frame("posterior",p,post)
 names(PRIOR)=c("Type","P","Probability")
 names(POST)=c("Type","P","Probability")
 data=rbind(PRIOR,POST)

xyplot(Probability~P|Type,data=data,layout=c(1,2),type="h",
lwd=3,col="black") # graph
```


posterior
prior
Probability
P
0.0
0.1
0.2
0.3
0.4
0.2
0.4
0.6
0.8

Empirical Bayes approach

see the background files

"Empirical Bayes methods are procedures for statistical inference in which the prior distribution is estimated from the data."


Fender
www.fender.com
Fender
Custom Shop

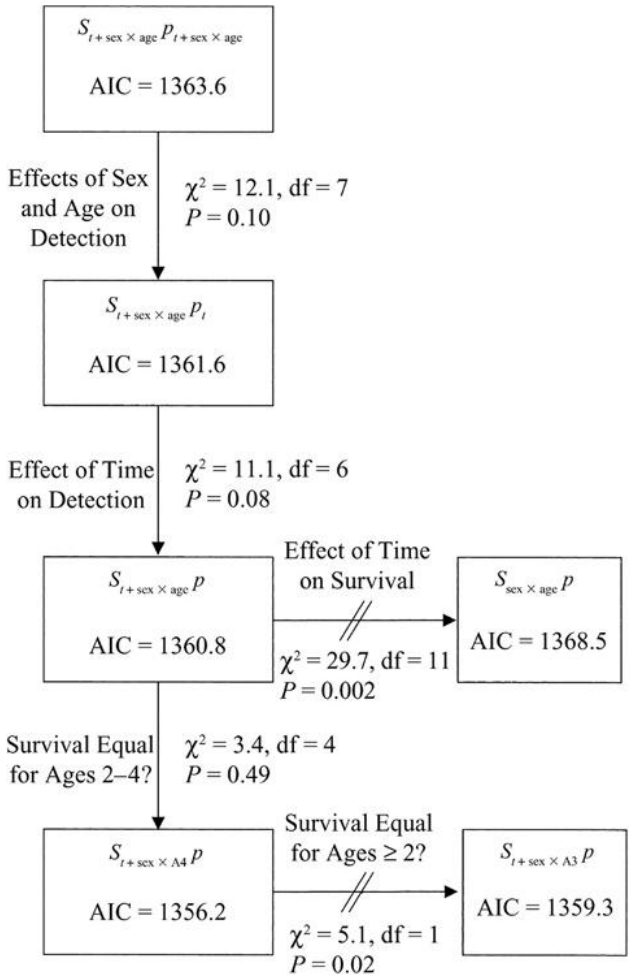
Model comparisons and model selection

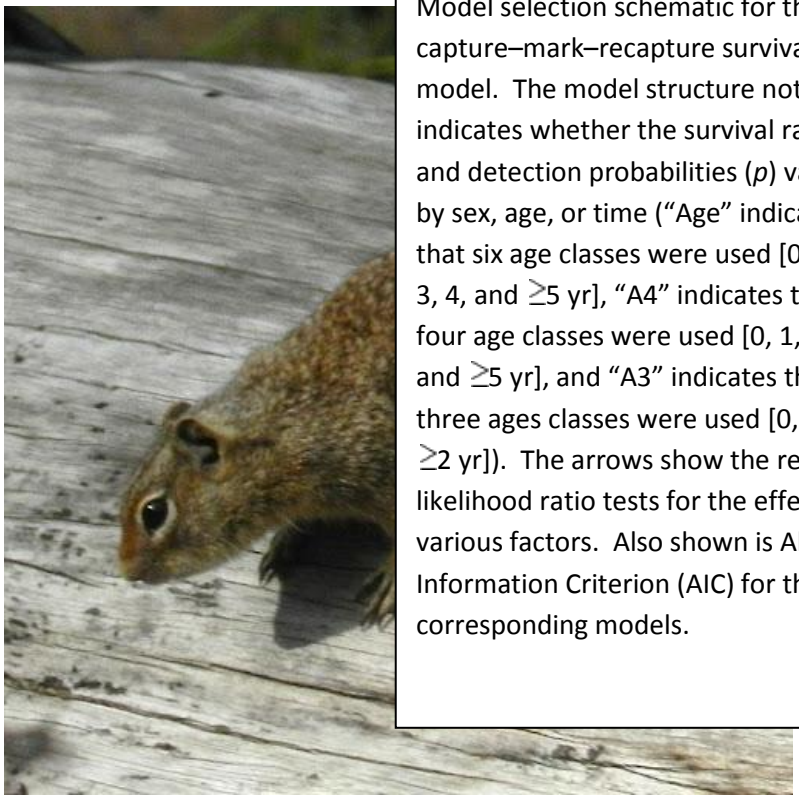
We often don't just have one null hypothesis to test on the same data

We can fit many models to the same data, from very simple to very complex, which model(s) to use for hypothesis testing?

Are *p*-values obtained from a maximal or MAM model I prefer still valid for hypothesis testing, after my model-selection procedure, which can have sampling effects too?

What if my model is very likely (a little bit) wrong?





Model selection schematic for the capture–mark–recapture survival model. The model structure notation indicates whether the survival rates (*S*) and detection probabilities (*p*) varied by sex, age, or time ("Age" indicates that six age classes were used [0, 1, 2, 3, 4, and ≥5 yr], "A4" indicates that four age classes were used [0, 1, 2–4, and ≥5 yr], and "A3" indicates that three ages classes were used [0, 1, and ≥2 yr]). The arrows show the results of likelihood ratio tests for the effects of various factors. Also shown is Akaike's Information Criterion (AIC) for the corresponding models.

Paul W. Sherman and Michael C. Runge. 2002. Demography of a population collapse: the northern Idaho ground squirrel (*Spermophilus brunneus brunneus*). *Ecology* 83:2816–2831.

Model Selection Principles

AIC and deviance comparisons
Hypothesis testing
Confidence intervals

Parsimony

Bias and Accuracy

Distance from the truth (AIC)

Consistency

Parsimony

Occam's Razor:

entities must not be multiplied beyond necessity

as written down by John Punch from Cork in 1639

- **Parsimony** heuristic: when AIC of two models is equal take the one with least parameters

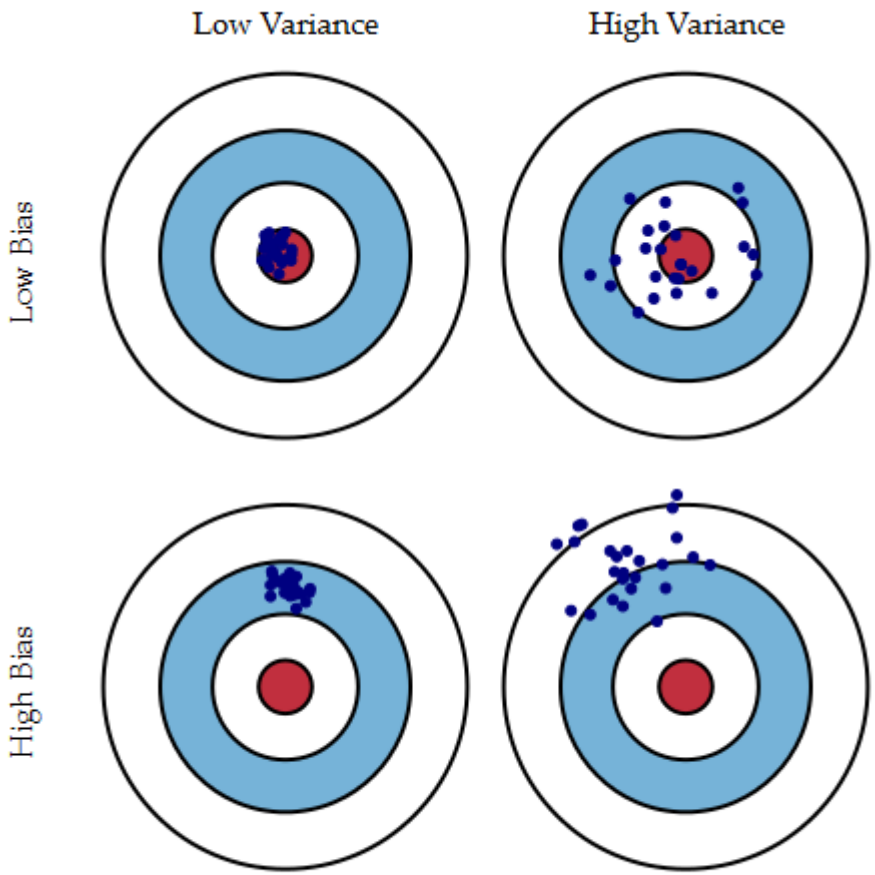


Fig. 1 Graphical illustration of bias and variance.

Observations

$$y_i = f(x_i) + \epsilon$$

Expected error at a point *x*

$$\begin{aligned} \mathrm{E}[(y - \hat{f}(x))^2] &= \mathrm{E}[(f(x) - \mathrm{E}[\hat{f}(x)])^2] + \mathrm{E}[(\hat{f}(x) - \mathrm{E}[\hat{f}(x)])^2] + \mathrm{E}[\epsilon^2] \\ &= \mathrm{Bias}(\hat{f}(x))^2 + \mathrm{Var}(\hat{f}(x)) + \sigma^2 \end{aligned}$$

Maximum accuracy = minimum expected error

Expected error is the sum of **bias** and **variance of the prediction** plus the variance of the residual ε

Accuracy depends on the value(s) of *x*

This idea generalizes to model comparisons
Simpler models: more bias. Many parameters: less precision

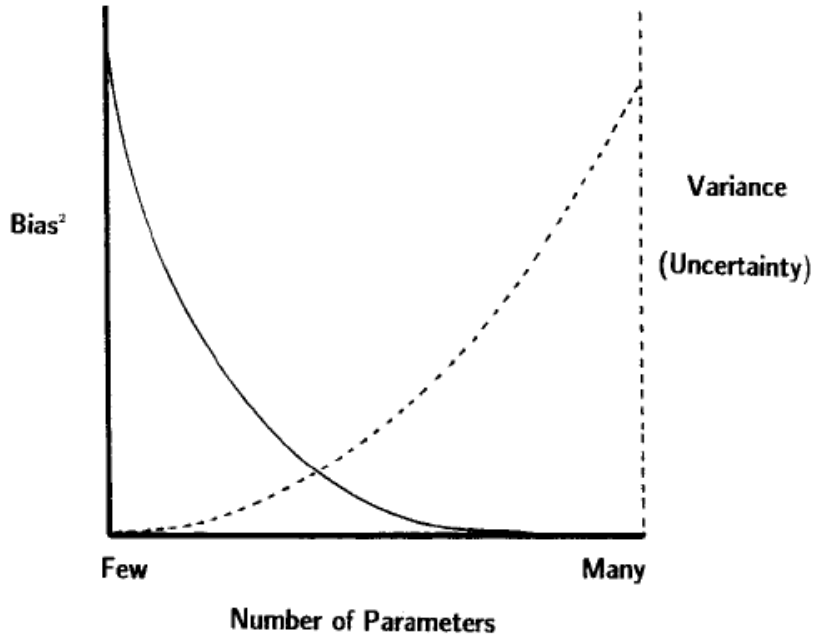


Fig. 1. The principle of parsimony: the conceptual trade-off between squared bias (solid line) and variance (i.e. uncertainty) versus the number of estimable parameters in the model. The best model has dimension (K_0) near the intersection of the two lines, while full reality lies far to the right of trade-off region.

AIC : an estimate of the distance from the truth

```
Call:
glm(formula = y ~ x)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.991912   0.047108  148.42   <2e-16 ***
x           0.299751   0.003932   76.22   <2e-16 ***
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.0102839)

    Null deviance: 59.93563  on 19  degrees of freedom
Residual deviance:  0.18511  on 18  degrees of freedom
AIC: -30.893

Number of Fisher Scoring iterations: 2
```

Kullback Leibler Information

Is a measure of the distance between the true data generating distribution *g* and the distribution implied by a specific approximate model *f*.

$$I(f,g) = \int dy g(y) \ln \frac{g(y)}{f(y|\theta)}$$

Advance 2 – Estimation of Kullback–Leibler information (AIC)

Akaike (1973, 1974) found a formal relationship between K–L information (a dominant paradigm in information and coding theory) and maximum likelihood (the dominant paradigm in statistics) (see deLeeuw 1992). This finding makes it possible to combine estimation (e.g. maximum likelihood or least squares) and model selection under a single theoretical framework – optimisation. Akaike's breakthrough was the finding of an estimator of the expected, relative K–L information, based on the maximised log-likelihood function. Akaike's derivation (which is for large samples) relied on K–L information as averaged entropy and this lead to 'Akaike's information criterion' (AIC),

$$\mathrm{AIC} = -2\log_e(L(\hat{\theta} \mid data)) + 2K,$$

where $\log_e(L(\hat{\theta} \mid data))$ is the value of the maximised log-likelihood over the unknown parameters (θ), given the data and the model, and K is the number of estimable parameters

Kullback - Leibler distance

$$\int dyg(y)(lng(y) - lnf(y|\theta))$$

log likelihood for a sample of size *n* is $l_n(\theta) = \sum_{i=1}^{n} lnf(y_i|\theta)$

strong law of large numbers:

$$n^{-1} l_n(\theta) \xrightarrow{a.s.} \int dyg(y) lnf(y|\theta)$$

for a choice of *f* the MLE estimate of θ will be the value that minimizes the KL distance. The first term of KL does not depend on the model chosen.

Kullback - Leibler distance for a ML fitted model

$$\int dyg(y)\left(lng(y) - lnf(y|\hat{\theta})\right)$$

The expectation of the second term, where MLE $\hat{\theta}$ is seen as a random variable that depends on *g*

$$E_g \int dyg(y)lnf(y|\hat{\theta})$$

can be estimated by $n^{-1}l_n(\hat{\theta})$. This estimate is **biased**.

Different bias corrections have been proposed. The simplest proposed estimate of bias is the number of estimated parameters in the model (i.e. the length of vector $\hat{\theta}$) divided by sample size *n*. This is used in the AIC.

AIC $-2l_n(\hat{\theta}) + 2length(\theta)$ smaller AIC is better

-2 log likelihood of the data given the ML model plus number of parameters, times two

BIC $-2l_n(\hat{\theta}) + \ln(n)length(\theta)$ smaller BIC is better

Bayesian information criterion: to compare posterior probabilities of models, essential quantity to select the model with largest posterior probability

and many more **IC** exist (ad-hoc, partly general):

AICc for small sample sizes
CAIC which is the AIC made consistent (see below)

DIC deviance information criterion

- **Consistency : a property of estimators and procedures. For IC:**

Strong: when a most parsimonious model with minimum KL distance is selected almost surely when n goes to ∞

Weak: when the prob. to select a most parsimonious model with minimum KL distance tends to 1 when n goes to ∞

BIC for n going to ∞ selects the true parsimonious model

AIC is consistent only when there is a single model among the candidates which is minimizing KL (not when several models can minimize KL). **This is the reason AIC is not the preferred method to compare nested models**

Also

- **Efficiency**

Minimizes

MSE mean squared estimation error = bias^2 + variance of parameter

AIC for n → ∞ selects the most efficient model with probability one

BIC not

Model selection

Consider different inference goals:

- one parameter has more interest. How test for it best? **A**
- We just want the most appropriate model for prediction. **B**
- We want a minimum adequate model, it's not such a problem if we don't demonstrate the weakest effects. We want to stay conservative. **C**

AIC, other criteria or

hypothesis testing are used for model selection **C**

Parameter of interest: frequentist test in maximal model and in MAM to assess model selection bias, power changes **A**

AIC: prediction accuracy **B** - **all models may be wrong C**

AIC: non-nested models **C**

BIC: find the true model among the candidates **C**

• Model selection improves the chance of finding the model which generated the data relevant for **A and C**

• Model selection can confuse inference if we are after frequentist inference **A**

frequentist inference often does not consider that we test w.r.t. a model obtained after comparing several model pairs.

Model selection uncertainty (sampling effects) is often forgotten

If we simplify a model, it can be seen that parameter estimates become seemingly more precise**. A, B**
If we simplify, bias is also expected to increase. **A, B**

Hybridization of the different schools: increasingly often

Frequentist model selection, IC, Bayesian model fitting methods all in one analysis

We use expectations over repeated experiments to motivate the use of AIC or of alternative IC, we look at effects of sampling variation to construct appropriate model selection strategies **(Frequentist**)

can involve simulation of model selection procedures

We use principles from **Bayesian** statistics to improve model fitting of complicated models and to obtain better estimates

Some advice out of the box:

- The **context**: all modeling is rooted in a scientific context, and is for a certain purpose, be efficient!

- The **purpose** of models is not to fit the data, but to sharpen the questions (S Karlin)

- Box: Most models are wrong and some are useful

Maybe the most important slide:

Make better use of your data, try to make sure any outcome of an experiment is useful:

The Method of Multiple Working Hypotheses

With this method the dangers of parental affection for a favorite theory can be circumvented.

T. C. Chamberlin

References

http://www.math.umt.edu/patterson/ProfileLikelihoodCI.pdf

Albert, J (2009) Bayesian Computation with R (Use R! Series). Springer Verlag.

http://www.d.umn.edu/~mille066/Teaching/3000/Chamberlin-MWH.pdf

KP Burnham, DR Anderson (2001) Kullback-Leibler information as a basis for strong inference in ecological studies. Wildlife research 28: 111-119

Claeskens, G and NL Hjort (2007) Model Selection and Model Averaging. Cambridge University Press

Pawitan, Y (2001) In all Likelihood. Oxford University Press