V. Introduction to Likelihood

1 Likelihood basics

1.1 Under assumed model, let $X$ denote data and $\Theta$ the parameter(s) of the model.

1.2 Probability: $P(X|\Theta)$

Just to make this concrete, suppose the assumed model is a Poisson process (events occurring at a constant instantaneous rate), the parameter is $\lambda$ (the product of the instantaneous rate and the length of time over which the process acts), and the data consist of the observed number of events, $k$. Then this would be $P(k|\lambda)$.

1.3 Likelihood ("inverse probability"): $L(\Theta|X) = c \cdot P(X|\Theta)$, where $c$ is a constant of proportionality.

We generally speak of the likelihood of a hypothesis, where the hypothesis is the particular parameter values (with model assumed).

1.4 Combining data

1.4.1 Let $X_1$ and $X_2$ denote two independent sets of data.

1.4.2 We know, if $X_1$ and $X_2$ are independent, that $P(X_1\&X_2|\Theta) = P(X_1|\Theta) \cdot P(X_2|\Theta)$

1.4.3 Therefore, $L(\Theta|X_1\&X_2) = L(\Theta|X_1) \cdot L(\Theta|X_2)$

1.5 Likelihood ratio of two hypotheses:

$$LR = L(\Theta_1 \text{ vs. } \Theta_2|X) = \frac{L(\Theta_1|X)}{L(\Theta_2|X)}$$

1.5.1 If $LR > 1$, then hypothesis $\Theta_1$ better supported by the data.

1.6 Support: $S(\Theta|X) = \ln L(\Theta|X)$

As we’ll see, it is often more convenient to work with support than with likelihood.
1.6.1 Properties:

- Transitivity: If \( S(\Theta_1) > S(\Theta_2) \) and \( S(\Theta_2) > S(\Theta_3) \), then \( S(\Theta_1) > S(\Theta_3) \).
- Additivity (see combining data for likelihoods): \( S(\Theta|X_1 \& X_2) = S(\Theta|X_1) + S(\Theta|X_2) \)
- Invariance under transformation of data (e.g. \( k \) vs. \( k/n \) in binomial).
- Invariance under transformation of parameters (e.g. \( \sigma \) vs. \( \sigma^2 \) in normal).

1.6.2 Difference in support between two hypotheses:

\[
\Delta S = \ln(\text{likelihood ratio}) = S(\Theta_1|X) - S(\Theta_2|X)
\]

If \( \Delta S \) is positive, hypothesis \( \Theta_1 \) has stronger support than hypothesis \( \Theta_2 \).

1.7 Example: Binomial, with probability of success \( p \), \( n \) trials, \( k \) successes:

\[
P(k|n, p) = \binom{n}{k} p^k (1 - p)^{n-k}
\]

\[
L(p|n, k) = cp^k (1 - p)^{n-k},
\]

where the binomial coefficient is sunk into the constant \( c \).

\[
S(p|n, k) = \ln(c) + k \ln(p) + (n - k) \ln(1 - p)
\]

1.8 NB: Hereinafter we will generally ignore the constant \( c \), since we are generally interested in likelihood ratios, differences in support, and derivatives of support functions, where it cancels out.

1.9 For continuous function with density \( f(x|\Theta) \) and observations \( x_i, i = 1, \ldots \)

\[
L(\Theta|x_i) = \prod_i f(x_i|\Theta) \quad \text{or}
\]

\[
S(\Theta|x_i) = \sum_i \ln[f(x_i|\Theta)],
\]

where \( f(x_i) \) is the density evaluated at the observation \( x_i \).
Binomial likelihood and support functions (from Edwards). These plots show $L/L_{\text{max}}$ and $\ln(L/L_{\text{max}})$.

2 Method of maximum likelihood (max. support)

2.1 Develop likelihood or support function, usually the latter.

2.2 Take first derivative with respect to $\Theta$.

This first derivative is called the score.

2.3 Maximize support by setting score to zero and solving for $\Theta$.

Value of $\Theta$ that maximizes likelihood (support) is called the evaluate or the ML estimator, and denoted $\hat{\Theta}$.

2.4 ML solution may be exact, or may need to be found numerically (q.v.)

2.5 Examples of maximum-likelihood solutions

2.5.1 Binomial with probability of success $p$

- Data consist of $n$ trials and $k$ successes.
- $L(p|n,k) = p^k(1-p)^{n-k}$
• $S(p|n, k) = k \ln(p) + (n - k) \ln(1 - p)$

• $\frac{dS}{dp} = \frac{k}{p} - \frac{n-k}{1-p}$

• $\frac{dS}{dp} = 0 \Rightarrow \hat{p} = \frac{k}{n}$

### 2.5.2 Poisson with parameter $\lambda$

• Data consist of $k$ events

• $L(\lambda|k) = e^{-\lambda} \frac{\lambda^k}{k!}$

• $S(\lambda|k) = -\lambda + k \ln(\lambda) - \ln(k!)$

• $\frac{dS}{d\lambda} = -1 + \frac{k}{\lambda}$

• $\frac{dS}{d\lambda} = 0 \Rightarrow \hat{\lambda} = k$

### 2.5.3 Exponential with parameter $r$

• Data consist of $n$ observed waiting times $x_1, ..., x_n$.

• $L(r|x_i) = \prod_i e^{-rx_i}$

• $S(r|x_i) = \sum_i [\ln(r) - rx_i] = n \ln(r) - \sum_i rx_i$

• $\frac{dS}{dr} = \frac{n}{r} - \sum_i x_i$

• $\frac{dS}{dr} = 0 \Rightarrow \hat{r} = \frac{n}{\sum_i x_i}$

### 2.6 Maximum likelihood with multiple parameters

#### 2.6.1 Determine score vector of partial derivatives with respect to each parameter in turn:

$$
\begin{pmatrix}
\frac{\partial S}{\partial \Theta_1} \\
\frac{\partial S}{\partial \Theta_2} \\
\vdots
\end{pmatrix}
$$

#### 2.6.2 Set each element of vector to zero and solve.

#### 2.6.3 Example: normal with mean $\mu$ and variance $\sigma^2$

• Data are $n$ observed values $x_1, ..., x_n$

• $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2} \left( \frac{x-x}{\sigma} \right)^2}$
• Likelihood function:

\[ L(\mu, \sigma^2 | x_i) = c \prod_i \frac{1}{\sigma} e^{-\frac{1}{2}(\frac{x_i - \mu}{\sigma})^2} \]

\[ = \frac{1}{\sigma^n} e^{-\frac{\sum(x_i - \mu)^2}{2\sigma^2}}, \]

where \( c \) absorbs \( \frac{1}{\sqrt{2\pi}} \).

• Note that \( \sum (x_i - \mu)^2 = \sum_i (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2 \), where \( \bar{x} \) is the sample mean.

• Note also that \( \frac{\sum (x_i - \bar{x})^2}{n} \) is the sample variance, \( s^2 \).

• Then \( \sum (x_i - \mu)^2 = n[s^2 + (\bar{x} - \mu)^2] \).

• Thus

\[ L(\mu, \sigma^2 | x_i) = c\sigma^{-n} e^{-\frac{n(s^2 + (\bar{x} - \mu)^2)}{2\sigma^2}} \]

Note that this contains both the parameters and sample statistics.

• Now, forgetting about the constant \( c \), we have

\[ S(\mu, \sigma^2 | x_i) = -n \ln(\sigma) - \frac{n}{2\sigma^2} [s^2 + (\bar{x} - \mu)^2] \]

• Taking partial derivatives:

\[ \frac{\partial S}{\partial \mu} = -\frac{n}{2\sigma^2} [2(\bar{x} - \mu)] \]

\[ \frac{\partial S}{\partial \mu} = 0 \Rightarrow \hat{\mu} = \bar{x} \]

\[ \frac{\partial S}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{n}{2\sigma^4} [s^2 + (\bar{x} - \mu)^2] \]

\[ \frac{\partial S}{\partial \sigma^2} = 0 \Rightarrow \hat{s}^2 + (\bar{x} - \mu)^2 \]

If we take \( \mu \) at the evaluate \( \hat{\mu} \), then \( (\bar{x} - \mu) = 0 \) and \( \hat{s}^2 = s^2 \)

• Thus, \( (\hat{\mu}, \hat{s}^2) = (\bar{x}, s^2) \).

Note that this requires joint parameter estimation; the estimation of \( \hat{s}^2 \) depends on the sample mean \( \bar{x} \). This is a common situation in multi-parameter models.
2.6.4 Example: parameters of generalized random walk (Hunt [2006], Paleobiology 32:578-601)

- Model I. Ignore sampling error within populations.

- Then

\[ \Delta X = \sum_{i=1}^{t} s_i, \]

where \( \Delta X \) is the net evolutionary difference between starting and ending points (what we have been calling \( S_n \)), \( t \) is the number of time steps (what we have been calling \( n \)), and \( s_i \) are the step sizes, drawn independently and randomly from a common distribution with mean \( \mu_{\text{step}} \) and variance \( \sigma_{\text{step}}^2 \).

- \( E(\Delta X) = t\mu_{\text{step}} \)

- \( V(\Delta X) = t\sigma_{\text{step}}^2 \)

- Model II. Include sampling error.

- Then

\[ \Delta X = \sum_{i=1}^{t} s_i + \epsilon_0 + \epsilon_t, \]

where \( \epsilon_0 \) and \( \epsilon_t \) are error terms at time 0 and time \( t \).

- These error terms are assumed to be normally distributed with zero mean and variance equal to \( V_0/n_0 \) and \( V_t/n_t \), where \( V_0 \) and \( V_t \) are the population variances at the beginning and end of the sequence, and \( n_0 \) and \( n_t \) are the corresponding sample sizes.

- This then yields:

\[ V(\Delta X) = t\sigma_{\text{step}}^2 + V_0/n_0 + V_t/n_t \]

- We then have a normal density function with mean \( E(\Delta X) \) and variance \( V(\Delta X) \), and, given this mean and variance, we can obtain the relative probability that the observed net difference \( \Delta X \) will take on any particular value.

- Given a number of observed \( \Delta X \) values and accepting observed values of \( V_0 \) and \( V_t \) (or using the pooled within-population variance as an estimate), we can obtain maximum-likelihood estimates of \( \mu_{\text{step}} \) and \( \sigma_{\text{step}}^2 \).

- NB: Hunt finds this solution numerically.
2.7 Aside: Comments on some approaches to numerical optimization (based on my own experience rather than any computer-science expertise):

2.7.1 General goals (which tend to work against each other) are speed and avoidance of local optima.

2.7.2 Exhaustive evaluation of points a grid
   • Should avoid local optima if grid spacing fine relative to scale length of objective-function surface
   • Prohibitively slow with many parameters
   • Easy to program

2.7.3 Gradient methods
   • Lots of existing code to call upon
   • Best if objective-function surface smooth

2.7.4 Downhill simplex method (aka “amoeba”)
   • “Oozes” into well of minimum, while stretching and bouncing around to look for other minima that may be even lower.
   • Seems fairly good at avoiding local optima
   • Canned code readily available
   • Best if objective-function surface smooth

2.7.5 Simulated annealing
   • Highly flexible and adaptable
   • Extremely easy to code
   • Works with rough or even discontinuous objective-function surface
   • Requires lots of effort to “tune” cooling schedule
   • May be slow
   • Core of annealing routine
     - go through “temperature” \((T)\) steps with exponentially declining temperature
– at each temperature step, perturb best-fitting solution
– compute $D$, difference between minimum function value and function value corresponding to candidate parameters
– if $D < 0$ accept new solution (objective function reduced)
– if $D > 0$ accept new solution with probability $\exp(-D/T)$
– Thus, worse-fitting solutions are accepted with high probability early on, with lower probability later. The worse the fit, the lower the probability of accepting it.
– This approach, given slow enough annealing schedule, tends to avoid local optima.

2.7.6 Sample R code for calling intrinsic `optim()` function:

```r
#Sample code to estimate mean and variance from sample, assuming normal.
data<-c(0.32,-1.34,0.10,-0.08,0.86,-1.28,0.69,-0.38,-1.11,-0.92)
support.function<-function(pars){
  mu<-pars[1]
  var<-pars[2]
  density<-dnorm(data,mean=mu,sd=sqrt(var))
  support<-sum(log(density))
  return(-support)
  #optim() minimizes by default; return -S to maximize S
}
p<-runif(2) #assign initial parameter values at random
mins<-c(-1,0.0001) #assign parameter range if parameters bounded
maxs<-c(1,2)
result<-optim(p,support.function,method="L-BFGS-B",lower=mins,upper=maxs)
mu.est<-result$par[1]
var.est<-result$par[2]
```
2.7.7 Sample R code for calling intrinsic one-dimensional \texttt{optimize()} function:

```r
#Sample code estimates parameter of exponential distribution,
#given a sample of waiting times.
data<-c(0.3,0.9,2.3,1.1,0.5,0.3,0.4,1.7,0.2,2.8)
support.function<-function(r,x) {
density<-dexp(x,rate=r)
support<-sum(log(density))
return(support)
}
result<-optimize(f=support.function,interval=c(0,100),maximum=TRUE,x=data)
#First argument to function is the parameter to be optimized.
r.est<-result$maximum
```

3 Evaluation of alternative parameter values (model fixed)

3.1 Support region (analogue of confidence interval):
How does support change as we go away from $\hat{\Theta}$?

3.2 Develop with binomial as example (but more general)

3.3 Information (negative second derivative of support w.r.t. parameter):

$$-\frac{d^2S}{d\Theta^2}$$

3.3.1 For binomial:

$$-\frac{d^2S}{dp^2} = \frac{k}{p^2} + \frac{n-k}{(1-p)^2}$$
3.4 Observed information: information evaluated at $\hat{\Theta}$

3.4.1 For binomial this is $\frac{n}{\hat{p}(1-\hat{p})}$

3.5 Formation: reciprocal of information

3.6 Observed formation: reciprocal of observed information

3.6.1 This is the radius of curvature of the support function at its maximum value

3.6.2 For binomial this is $\frac{\hat{p}(1-\hat{p})}{n}$, which you’ll recognize as the variance of the binomial.

3.7 Span: square root of observed formation (measure of width of support function near maximum)

3.8 Relation to difference in support

3.8.1 Denote span by $w$. Then

$$w^2 = -\frac{1}{d^2S} \text{ at } \hat{\Theta}$$

3.8.2 Using Taylor series expansion of $S(\Theta)$ around $\hat{\Theta}$ we can approximate this by a parabola:

$$S(\Theta) \approx S(\hat{\Theta}) - \frac{(\Theta - \hat{\Theta})^2}{2w^2}$$

3.8.3 $\Delta S \approx S(\hat{\Theta}) - S(\Theta) \approx \frac{(\Theta - \hat{\Theta})^2}{2w^2}$

3.8.4 I.e. $2\Delta S \approx \frac{\Delta \Theta^2}{w^2}$

3.8.5 For binomial, this is:

$$2\Delta S \approx \frac{(p - k/n)^2}{p(1-p)/n}$$
3.8.6 Note that a squared deviation from expected value, scaled to variance, is distributed as $\chi^2$. The expression $\Delta \Theta^2 w^2$ follows this form.

3.8.7 For a one-parameter model, twice the difference in support is approximately distributed as $\chi^2$ with one degree of freedom.

3.8.8 $\chi^2_{0.05,1} = 3.84$. Rounding 3.84 up to 4, this leads to the conventional use of a 2-unit support region as approximately analogous to a 95% confidence interval.

3.8.9 Note relationship between $\chi^2$ values and deviations sampled from a normal curve

- $\chi^2_{0.05,1 \ d.f.} = 3.84 = (1.96)^2$
- $z_{0.05,\infty} = 1.96 = \sqrt{3.84}$
4 Model comparison

In general, a model with more parameters will fit the data better (i.e., yield a higher likelihood). How do we determine whether more complex model—with more parameters—is justified?

4.1 Nested models: $2\Delta S \sim \chi^2$ with degrees of freedom equal to difference in number of parameters.

4.1.1 Example of single vs. multiple extinction probabilities

- Given two groups with different proportions of taxa that go extinct in some event, what are the relative merits of a one-parameter model (both groups have same underlying extinction probability) vs. a two-parameter model (groups have different probabilities)?

<table>
<thead>
<tr>
<th>Group</th>
<th>$N_{total}$</th>
<th>$N_{extinct}$</th>
<th>$\hat{p}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>100</td>
<td>50</td>
<td>0.5</td>
</tr>
<tr>
<td>B</td>
<td>150</td>
<td>50</td>
<td>0.333</td>
</tr>
<tr>
<td>A+B</td>
<td>250</td>
<td>100</td>
<td>0.4</td>
</tr>
</tbody>
</table>

- Model 1: $p_1 = p_2 = 0.4$
  $S = 50 \ln(0.4) + 50 \ln(0.6) + 50 \ln(0.4) + 100 \ln(0.6) = -168.3$

- Model 2: $p_1 = 0.5; p_2 = 0.333$
  $S = 50 \ln(0.5) + 50 \ln(0.5) + 50 \ln(0.333) + 100 \ln(0.667) = -164.8$

- $2\Delta S = 6.92; p < 0.001 \left( \chi^2_{0.01,1} \ a.f. = 6.64 \right)$

- Thus two-parameter model would be favored.

4.2 Akaike Information Criterion (and variants) (for nested or non-nested models)


4.2.1 $AIC = -2S + 2k$, where $k$ is the number of free parameters being fitted to the data.

(Multiplication by 2 is for convenience in comparing to $2\Delta S$.)
4.2.2 Lower \( AIC \) implies better fit.

4.2.3 Example of single vs. multiple extinction probabilities

- See previous data table
  \[
  AIC_1 = -2(-168.3) + 2(1) = 338.6
  \]
  \[
  AIC_2 = -2(-164.8) + 2(2) = 333.6
  \]
- So two-paramter model would be favored.

4.2.4 Practically speaking, to justify adding a parameter to a model, support should be increased by at least one unit.

4.2.5 Thus, in choosing between simpler and more complex models, use of \( \Delta AIC \) is less stringent than use of \( 2\Delta S \sim \chi^2 \) for small difference in number of parameters, but more stringent for larger difference in number of parameters.
4.2.6 *AIC* looks like a simple “adjustment” of support $S$, but actually has a more formal justification (via what is known as the Kullback-Leibler divergence between model and “truth”):

- Let $g(x)$ be the true density function
- $f(x)$ is a statistical model that approximates $g(x)$
- Define $I(g : f)$ as the following expectation:
  \[
  I(g : f) = E\left[ \log \frac{g(x)}{f(x)} \right] = \int_{-\infty}^{\infty} \log \frac{g(x)}{f(x)} \cdot g(x) \, dx
  \]

- $I(g : f) \geq 0$
- $I(g : f) = 0 \Rightarrow g = f$ and vice versa
- Therefore it is desirable to minimize $I$.
- But how to minimize $I$ when we have $f$ and data $x_i$, but we don’t know what $g$ is?
- $I(g : f) = E\{\log[g(x)]\} - E\{\log[f(x)]\}$. The first term is constant (albeit unknown), regardless of the statistical model $f$.
- So all we care about is the second term. This is:
  \[
  E\{\log[f(x)]\} = \int_{-\infty}^{\infty} g(y) \cdot \log[f(y)] \, dy,
  \]
  where $g$ is still unknown.
- If $x_i$ are our observed data ($i = 1, ..., N$), then, as $N \to \infty$,
  \[
  E\{\log[f(x)]\} \to \frac{\sum_{i=1}^{N} \log[f(x_i)]}{N}
  \]
- Thus, $I$ is minimized when $\frac{\sum_{i=1}^{N} \log[f(x_i)]}{N}$ is maximized.
- $\frac{\sum_{i=1}^{N} \log[f(x_i)]}{N}$ is clearly just the support or log-likelihood (divided by $N$).
- As it turns out, using $\frac{\sum_{i=1}^{N} \log[f(x_i)]}{N}$ to estimate $E\{\log[f(x)]\}$ is biased, because the same data are used to calculate the support and to estimate the parameters.
• The magnitude of the bias is:

\[
BIAS = E\{\log[f(x)]\} - \frac{\sum_{i=1}^{N} \log[f(x_i)]}{N}
\]

• What Akaike showed is that this bias is approximately equal to \(-k/N\), where \(k\) is the number of parameters (see his paper for the gory details).

4.2.7 Because AIC is based on large-\(N\) approximation, a correction is often used to make the expression more accurate for smaller \(N\):

\[
AIC_c = AIC + \frac{2k(k+1)}{N-k-1},
\]

where \(N\) is sample size (number of observations).

• \(AIC_c \approx AIC\) when \(N\) is large relative to \(k\).

• It is therefore a good idea to use \(AIC_c\) in general. (The correction has an effect when it is called for, and does no harm when it is not needed.)

• See Connolly and Miller (2001a, *Paleobiology* 27:751) for further discussion.

4.2.8 Example of instar determination via mixture models (Hunt and Chapman 2001, *Paleobiology* 27:466)

• Model: Assume population is a mixture of normal distributions.

• We want to estimate number of groups (normal distributions) that added together to form the mixture, as well as the mixing proportions and the mean and variance of each group.

• Density:

\[
f_{mix} = \sum_{i=1}^{g} \pi_i f(\mu_i, \sigma^2_i),
\]

where \(g\) is the number of groups, \(\pi_i\) is the proportion of individuals coming from group \(i\), and \(\mu_i\) and \(\sigma^2_i\) are the mean and variance of group \(i\).

• Support: If we observe a trait \(x_j\) on \(n\) individuals, then overall support for a candidate set of parameters is:

\[
S = \sum_{i=1}^{g} \sum_{j=1}^{n} \tau_{ij} \ln[\pi_i f(\mu_i, \sigma^2_i|x_j)],
\]

where \(\tau_{ij}\) is the probability that individual \(j\) belongs to group \(i\).
Figure 1. Hypothetical size distribution, with a one-group (A) and two-group (B) model fit to the data. In (B), the dotted curve is the probability density function of the mixture model, and the solid curves are the probability density functions of the two component distributions. In this example, the one-group model is rejected in favor of the two-group model.
• All the τ’s, π’s, μ’s and σ²’s must be estimated from the data (via numerical optimization).

• Start with simple model, estimate parameters, record support.

• Then add one group to model, estimate parameters, assess whether increase in support warrants addition of parameters.

• For mixture models, −2ΔS not distributed as χ², so use parametric bootstrap as alternative:

Details of step-wise model comparison

\( I_2 \) or \( S_1 - S_2 \). Under certain conditions, −2λ is distributed asymptotically as χ², allowing one to judge the significance of the support gain of a complex model relative to a simpler model. Unfortunately, these conditions do not hold for mixture models (McLachlan and Basford 1988).

As an alternative, we compared hypotheses with differing numbers of groups using parametric bootstrapping (McLachlan and Basford 1988). The following procedure is used to determine if the support advantage of the hypothesis with more groups (H₂) is great enough to reject the hypothesis with fewer groups (H₁):

1. Generate a sample based on the parameter estimates of H₁. If the original data set has \( n \) individuals, then the sample will consist of \( n \) random draws from the population defined by the means, variances, and mixing proportions according to H₁.
2. Calculate the support for the maximum likelihood estimates for this simulated data set assuming the number of groups implied by H₁.
3. Calculate the support for the maximum likelihood estimates for the same simulated data, this time assuming the number of groups implied by H₂.
4. Calculate and save −2λ = −2(S₁ − S₂). It is conventional (although not necessary) to multiply the support difference by −2 to parallel the procedure for the χ² approximation.
5. Repeat the above steps \( K \) times, where \( K \) is a large number (e.g., 1000). This generates a distribution of the amount of support gained by considering the number or groups implied by H₂, even though H₁ generated the underlying distribution. Set a significance level, \( \alpha \). If the observed value for −2λ is greater than the \( (1 - \alpha)K^{th} \) largest simulated value, then H₁ is rejected in favor of H₂. In such cases, the gain in support for H₂ over H₁ is greater than reasonably can be expected from the addition of parameters to the model.
• Assess statistical power (ability to reject simpler hypothesis when more complex hypothesis is true):

Details of power analysis

specimens per group than simpler models. We used the following procedure to assess the power of a bootstrap test to correctly reject a simpler model \(H_1\) in favor of the more complex model \(H_2\) at significance level \(\alpha\):

1. Create a bootstrap distribution of \(-2\lambda\) as described above, using \(H_1\) to generate the data. Record the critical value of \(-2\lambda\) at the significance level \(\alpha\).

2. Create a second distribution of \(-2\lambda\), this time using the parameters of \(H_2\) (rather than \(H_1\)) to generate the distribution.

3. Calculate the proportion of the values from the second bootstrap distribution of \(-2\lambda\) that are larger than the critical value recorded in Step 1. This proportion estimates the power of the test, i.e., the probability that \(H_1\) will be rejected in favor of \(H_2\), when \(H_2\) is actually true.
• Raw data

![Graph](image_url)

**Figure 3.** Scatter plot of cephalic dimensions for *Amphicoma biliaris*. The eight putative clusters apparent from visual inspection are labeled I-VIII. The line is the major axis of the data and has the equation \( y = 0.573x + 2.02 \).

• Stepwise model comparison

<table>
<thead>
<tr>
<th>No. of groups</th>
<th>Support</th>
<th>Bootstrap test (( \hat{H}_1 ) vs. ( \hat{H}_2 ))</th>
<th>Observed (-2\lambda)</th>
<th>Bootstrap results</th>
<th>Bootstrap ( p )-value</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>2</td>
<td>10.60</td>
<td>1 vs. 2 groups</td>
<td>13.12</td>
<td>0.086*</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>3</td>
<td>18.30</td>
<td>2 vs. 3 groups</td>
<td>15.38</td>
<td>0.065</td>
<td>0.84</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>19.55</td>
<td>2 vs. 4 groups</td>
<td>17.88</td>
<td>0.363</td>
<td>0.67</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>33.35</td>
<td>2 vs. 5 groups</td>
<td>15.50</td>
<td>0.092*</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>6</td>
<td>47.54</td>
<td>3 vs. 6 groups</td>
<td>28.37</td>
<td>0.092*</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>7</td>
<td>50.39</td>
<td>6 vs. 7 groups</td>
<td>26.10</td>
<td>0.003*</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>8</td>
<td>99.39</td>
<td>7 vs. 8 groups</td>
<td>17.59</td>
<td>0.084</td>
<td>0.57</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>72.52</td>
<td>7 vs. 9 groups</td>
<td>23.86</td>
<td>0.218</td>
<td>0.59</td>
<td></td>
</tr>
</tbody>
</table>

• Interpretation: Evidence for several instar groupings. Tests in which simpler model (fewer instars) cannot be rejected generally have low statistical power.
5 Bayesian Inference (ultra-rudimentary treatment)

5.1 Likelihood analysis led to ML estimate of parameter and estimate of tightness of support.

5.1.1 A somewhat philosophical question is whether this should be interpreted in terms of our strength of belief in the hypothesis.

5.1.2 A more pragmatic question is whether—and how—to modify likelihoods to take prior evidence or belief into consideration.

5.2 Recall the basics of conditional probability:

\[ P(A|H) = \frac{P(AH)}{P(H)} \]

Likewise

\[ P(H|A) = \frac{P(AH)}{P(A)} \]

From these relationships we can derive the following:

\[ P(H|A) = \frac{P(A|H)P(H)}{P(A)} \]

5.3 To put this is terms that are relevant to inference:

5.3.1 \( P(A|H) \) is the conditional probability of observing data \( A \), given hypothesis \( H \).

5.3.2 \( P(H) \) is the probability of hypothesis \( H \).

This is called the prior probability.

5.3.3 \( P(A) \) is the unconditional probability of observing data \( A \) (regardless of any prior hypothesis).

It is also known as the marginal probability. It is simply the conditional probability of \( A \) given \( H \), summed over all possible hypotheses \( H \). I.e.

\[ P(A) = \sum_i P(A|H_i)P(H_i) \]

5.3.4 \( P(H|A) \) is the probability of \( H \) given the data \( A \).

This is referred to as the posterior probability.
5.4 You can probably see that what we have been doing so far is
to ignore the term $P(H)$, or tacitly to assume that it is the
same for all hypotheses $H_1, H_2, \ldots$.

5.4.1 But if we have some prior reason to assign differential probabilities to
the various hypotheses, this can be taken into consideration.

5.5 Example with outcomes of binomial experiment

5.5.1 Suppose our data consist of nothing more than that there were $k$
successes. We assume $p = 0.5$. We do not know the number of trials $n$.
We want to evaluate the relative probability of the possible values of $n$.

5.5.2 Case 1: Uniform prior

- Suppose $k = 5$. Then $n$ can be anything from 5 to $\infty$.
- Let’s limit ourselves to values of $n$ up to 30 (without loss of generality). We do this
because the probability of getting only 5 successes in 30 or more trials is extremely
small.
- Denote the candidate values of $n$ by $n_i$.
- There are 26 different values of $n_i$ (from 5 to 30). So our prior probabilities are $1/26$ for all values of $n_i$.
- Thus

$$P(n_i|k) = \frac{(n_i)^{k} \cdot 0.5^{n_i} \cdot 1}{26} \sum_j \frac{(n_j)^{k} \cdot 0.5^{n_j} \cdot 1}{26}$$

These posterior probabilities are plotted, along with the uniform prior, in panel A of the
figure. Clearly, the prior has no influence on the posterior probabilities, and we find (as we
would with the ML approach), that the highest probabilities of getting $k = 5$ occur when
$n = 9$ and $n = 10$.

5.5.3 Case 2: One example of non-uniform priors

- Suppose you are given the additional data that the trials consist of the tosses of all
the pennies found in some random person’s pocket.
- This might lead you to adjust your prior probabilities, for you don’t think it is
equally probable that a randomly chosen person is walking around with, say, 10 vs.
30 pennies.
- So you do a brief survey of the people in the classroom and find out that the average
number of pennies per pocket is 7.
Red line shows posterior probability with uniform prior.
Posterior probabilities scaled to maximum of unity.

- This leads you to suggest a distribution of prior probabilities that is Poisson with parameter $\lambda = 7$.
- Now
  \[ P(n_i|k) = \frac{\binom{n_i}{k} 0.5^n e^{-\lambda} \frac{\lambda^n}{n_i!}}{\sum_j \binom{n_i}{j} 0.5^j e^{-\lambda} \frac{\lambda^j}{j!}} \]
- The prior and posterior probabilities are plotted in panel B.
- The maximum posterior probability falls at $n_i = 8$, which is less than the value of $n_i = 9$ or $n_i = 10$ in the case of uniform priors.
- Thus, the prior probability has had some influence on the posterior distribution.

5.5.4 Case 3: Another example of non-uniform priors
- Assume that you have a slightly different prior than in Case 2. This time you suppose that the prior probability of having $n_i$ pennies also follows a Poisson
distribution, but this time with parameter $\lambda = 10$.

- The results in panel C show the maximum posterior probability in the same place as in the case of uniform priors. This makes sense since the central tendency of the prior distribution is at $n = 10$.

- At the same time, the incorporation of variable priors (which support the same conclusion as the data) leads to a tighter distribution of posterior probabilities.

5.5.5 Case 4: An example of non-uniform priors with additional data

- Suppose now that we have six sets of observations, with $k = 5, k = 5, k = 4, k = 6, k = 7$, and $k = 3$. Denote these values $k_m (m = 1, \ldots, 6)$.

- Suppose we use the same prior as in Case 2: Poisson with $\lambda = 7$.

- Because the $k_m$ are independent, the probabilities are multiplicative:

$$P(n_i|k_m) = \frac{\left[ \prod_{m=1}^{6} \begin{pmatrix} n_i \\ k_m \end{pmatrix} 0.5^{n_i} \right] e^{-\lambda} \lambda^{n_i} / n_i!}{\sum_j \left[ \prod_{m=1}^{6} \begin{pmatrix} n_j \\ k_m \end{pmatrix} 0.5^{n_j} \right] e^{-\lambda} \lambda^{n_j} / n_j!}$$

- The posterior probabilities are not so strongly “pulled” by the prior probabilities as in Case 2. This makes sense, since there are many more observations pointing toward $n = 10$ than in Case 2.

- This situation, in which the data “swamp” the priors, is fairly common in practice.

5.6 NB: The prior probabilities may (and often do) reflect previous observations rather than subjective beliefs.

5.6.1 In this respect, combining likelihoods from independent observations can be seen as a special case of conditioning on prior probabilities.

5.6.2 Example with two outcomes from our binomial sampling experiment: $k_1 = 5$ and $k_2 = 6$

- For the $k_1 = 5$ case, assume uniform prior and get

$$P(n_i|k_1) = \frac{\begin{pmatrix} n_i \\ k_1 \end{pmatrix} 0.5^{n_i} \cdot \frac{1}{26}}{\sum_j \begin{pmatrix} n_j \\ k_1 \end{pmatrix} 0.5^{n_j} \cdot \frac{1}{26}}$$

- For the $k_2 = 6$ case, take the posterior probabilities just calculated and use them as prior probabilities:

$$P(n_i|k_2) = \frac{\begin{pmatrix} n_i \\ k_2 \end{pmatrix} 0.5^{n_i} \cdot P(n_i|k_1)}{\sum_j \begin{pmatrix} n_j \\ k_2 \end{pmatrix} 0.5^{n_j} \cdot P(n_i|k_1)}$$
• This second set of posterior probabilities is equal to:

\[
P(n_i|k_m) = \frac{\prod_{m=1}^{2} \binom{n_i}{k_m} 0.5^{n_i} \frac{1}{26}}{\sum_j \prod_{m=1}^{2} \binom{n_j}{k_m} 0.5^{n_j} \frac{1}{26}}
\]

5.6.3 In other words, using previous observations as prior probabilities has the same effect as combining data and assuming a uniform prior.