Teaching Foundational Statistical Topics to Biologists: Package asbio

by Ken Aho

Abstract The package **asbio** (Aho, 2010) is designed to facilitate the teaching of statistics to biology students. It contains simple graphical functions for conceptualizing sampling distributions, likelihood, and other foundational topics. Animation is often used in these functions to emphasize central ideas.

As a teacher of classes in biostatistics I have found that fundamental concepts (e.g. sampling distributions, power, hypothesis testing) are often poorly understood by graduate students, even those with several stats courses behind them. Even less well understood are more complex but foundational topics like probability density functions and likelihood (Horgan et al., 1999).

Statistical concepts can be elucidated with R-programing (Fox, 2002) and computer-generated graphics including animation (Xie and Chang, 2008). In this paper I demonstrate applications in the R-package **asbio** (Aho, 2010) that facilitate understanding of two fundamental topics: sampling distributions and likelihood. This is accomplished with animated graphical functions that can be easily customized by teachers and students.

Sampling distributions

The function `samp.dist` from **asbio** samples without replacement from up to two parental distributions with up to two distinct sample sizes. The default statistic calculated at each sample iteration is the sample mean although any statistic can be specified. Indeed, up to four distinct distributions of statistics can be assembled, and these can be combined in infinite ways by calling an auxiliary function. The resulting distribution is displayed in an animated histogram. The package **animation** allows R-animations to be saved as movie files, and provides useful animations, including depictions of the central limit theorem (function `clt.ani`). The function `samp.dist`, however, provides additional flexibility (in demonstrating sampling distributions) by allowing alternative statistics, multiple statistics, and multiple simultaneous parent distributions.

Sampling distribution of $\bar{X}$

If a parental distribution can be described as $X \sim (\mu,\sigma^2)$, then repeated random sampling, using a sample size n, will result in the distribution: $\bar{X} \xrightarrow{d} N\left(\mu, \sigma^2/n\right)$, as $n \to \infty$.

I will graphically demonstrate this idea (using `samp.dist`) by sampling from an exponential parent distribution using four different sample sizes (Fig. 1).

```
require(asbio)
exp.parent<-rexp(10000)
par(mfrow=c(2,2),mar=c(4.4,4.5,1,0.5))
samp.dist(parent=exp.parent,s.size=1,anim=FALSE)
samp.dist(parent=exp.parent,s.size=5,anim=FALSE)
samp.dist(parent=exp.parent,s.size=10,anim=FALSE)
samp.dist(parent=exp.parent,s.size=50,anim=FALSE)
```

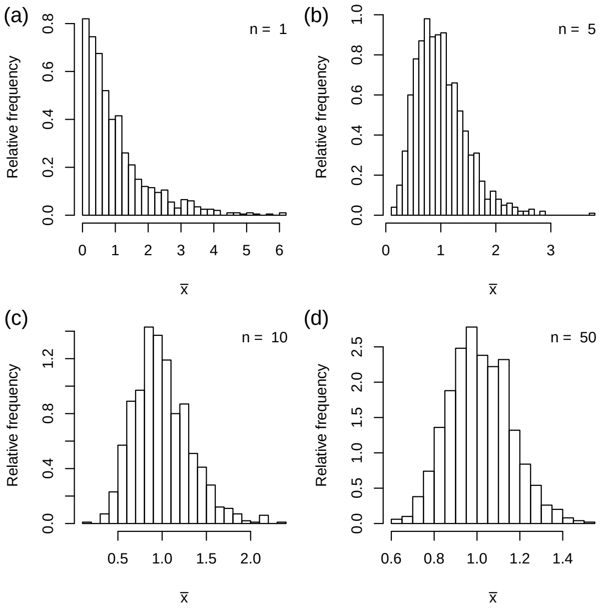

Figure 1: Sampling distribution of $\bar{X}$ for four sample sizes, given a parental distribution $EXP(1)$. As sample size increases the distribution becomes increasingly normal.

The process of sample statistic accumulation can be animated by allowing the default `anim=TRUE`. For instance, the code:

```
samp.dist(parent=exp.parent,s.size=5)
```

depicts the accumulation of means for the distribution in Fig. 1b. This animation is pedagogically important because it shows that a true sampling distribution requires an infinite number of estimates. Clearly a collection of ten means will not resemble a normal distribution regardless of their sample size. Only when the number of means grows large does this distribution *begin* to approximate "true" frequentist characteristics.

The function `samp.dist` also provides seamless depictions of the effect of changing sample sizes. For instance, the code:

```
samp.dist(parent=exp.parent,fix.n=FALSE,interval=.3)
```

shows sampling distributions of $\bar{X}$ based on sample sizes from 1 to 30, with an animation speed of one frame/0.3 second.

This sort of animation can be combined with estimates of the standard error, by specifying `show.SE = TRUE`. This presentation can then be used to demonstrate the consistency and efficiency of an estimator. For example, the following code shows that while both $\bar{X}$ and the sample median are consistent for μ, $\bar{X}$ is more efficient, given a normal parent distribution.

```
parent<-rnorm(10000)
samp.dist(parent, fix.n=FALSE, interval=.3,n.seq=
seq(1,100),show.SE=TRUE);dev.new()
samp.dist(parent, fix.n=FALSE,interval=.3,stat=
median,xlab="Median",n.seq=seq(1,100),show.SE=TRUE)
```

Variance estimates for these sampling distributions (i.e. SE^2's) show that the median is only about 64% as efficient as the mean (Fig. 2).

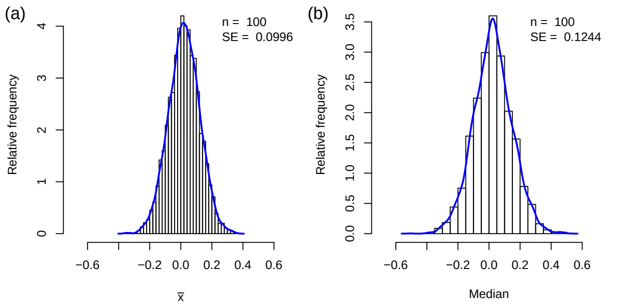

Figure 2: Sampling distributions of (a) $\bar{X}$, and (b) the median, given a standard normal parent distribution.

It is also possible to demonstrate the effect of sample contamination on estimators. For example, we can rerun the code above, but now specify a standard normal parent distribution with 10% contamination from $N(10,1)$.

```
parent<-c(rnorm(9000),rnorm(1000,mean=10))
```

In this situation the median will be nearly six times *more* efficient than the mean.

Sampling distribution of S^2

Unlike the sampling distribution of $\bar{X}$, the sampling distribution of S^2 is rarely graphically demonstrated. This is unfortunate because this distribution allows computation of asymmetric confidence intervals for mixed models parameters, vital to contemporary analyses.

If we assume a normal parent distribution, then the sampling distribution of S^2 can be related to a χ^2 distribution. In particular, if $X_1, \ldots, X_n$ represent random samples from $N(\mu, \sigma^2)$, then:

$$(n-1)S^2/\sigma^2 \sim \chi^2(n-1). \qquad (1)$$

To demonstrate this concept we require a normal parent distribution.

```
parent<-rnorm(10000)
```

We are not interested in the sampling distribution of S^2, but the sampling distribution of $(n-1)S^2/\sigma^2$. As a result we create an auxiliary function representing Eq. 1.

```
eq1.dist<-function(s.dist,sigma.sq=1,s.size=8){
func.res<-((s.size-1)*s.dist)/sigma.sq;func.res}
```

We will call this function using the `func` argument from `samp.dist`. The `func` argument is used to incorporate a function (e.g. `eq1.dist`) that manipulates and combines one or more sampling distributions.[1]

Finally we run `samp.dist`. By default `samp.dist` calculates one thousand statistics for a single sample size. We increase this to 10000 to get a concise view of the sampling distribution for $n = 8$.

```
xlabel<-expression(paste("(n - 1)",sˆ2,"/",sigmaˆ2))
samp.dist(parent, s.size=8,stat=var,xlab=xlabel,
R=10000,func=eq1.dist)
```

We see that this sampling distribution is indeed described by $\chi^2(n-1)$ (Fig. 3).

```
curve(dchisq(x,7),from=0,to=30,add=TRUE,
lwd=2,col=1)
```

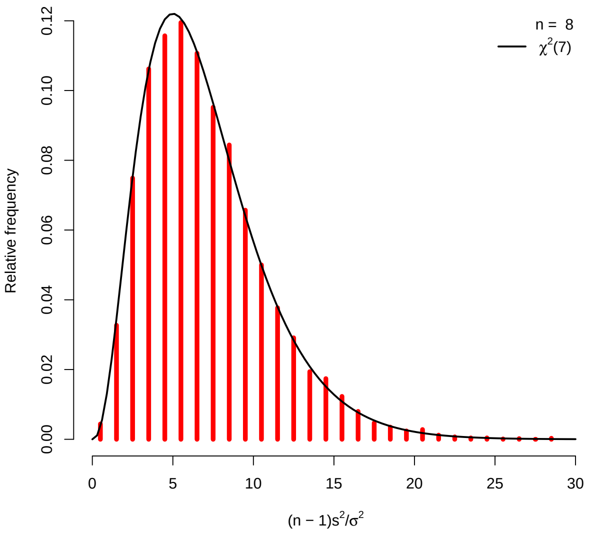

Figure 3: Sampling distribution of $(n-1)S^2/\sigma^2$, for a sample size of 8, and a standard normal parent distribution. The histogram is based on 10000 estimates.

[1] The non-fixed arguments of a function called by `func` must include, and be limited to, sampling distributions (e.g. `s.dist`).

Sampling distributions of test statistics

Imperative to frequentist procedures are sampling distributions of test statistics. The test statistic, t*, is often used to quantify evidence against a null hypothesis (H_0) that the mean of a normal population with an unknown variance is equal to the value, μ_0

$$t^* = \frac{\sqrt{n}(\bar{x} - \mu_0)}{s}. \qquad (2)$$

If H_0 is true, then t*'s can be represented with a random variable T that follows a t-distribution with n - 1 degrees of freedom. In particular, if X_1 ,..., X_n represent a random sample from $N(\mu,\sigma^2)$ then:

$$T = \frac{Z}{\sqrt{V/(n-1)}} \sim t(n-1), \qquad (3)$$

where $Z \sim N(0,1)$ and $V \sim \chi^2(n-1)$.

To demonstrate this we again require a normal parent distribution.

```
parent<-rnorm(100000)
```

We use a standard normal parent distribution because it depicts a true null hypothesis (if $\mu_0 = 0$), since $E(\bar{X}) = \mu_0 = 0$.

Once again we will create an auxiliary function to manipulate and combine sampling distributions. This time, however, we require two sampling distributions, one for$\bar{X}$ and one for S^2. These will be run through a function, representing Eq. 2, to find outcomes for the random variable T.

```
t.star<-function(s.dist1,s.dist3,s.size=8){
func.res<-s.dist1/(sqrt(s.dist3/s.size));func.res}
```

We use the terms `s.dist1` and `s.dist3` to indicate that the same sample from the same parental distribution will be to be used to calculate $\bar{x}$ and s^2 for a particular t*. These statistics will be specified as `stat` and `stat3`. Conversely, `s.dist2` and `s.dist4` will be sampling distributions for statistics from a second parental distribution. These statistics are, if necessary, specified with the arguments `stat2` and `stat4`. This application will be demonstrated shortly.

```
samp.dist(parent,s.size=8,stat=mean,stat3=var,
xlab="t*",func=t.star,col.anim="gray",ylim=c(0,.4))
```

We see that the thicker tails of the t-distribution provide a better representation of the sampling distribution of T than the standard normal distribution (Fig. 4).

```
curve(dt(x,7),from=-6,to=6,add=TRUE,lwd=2,col=4)
curve(dnorm(x),from=-6,to=6,add=TRUE,lwd=2,col=2,
lty=2)
legend("topleft",lwd=c(2,2),lty=c(1,2),col=c(4,2),
legend=c("t(7)","N(0,1)"),bty="n")
```

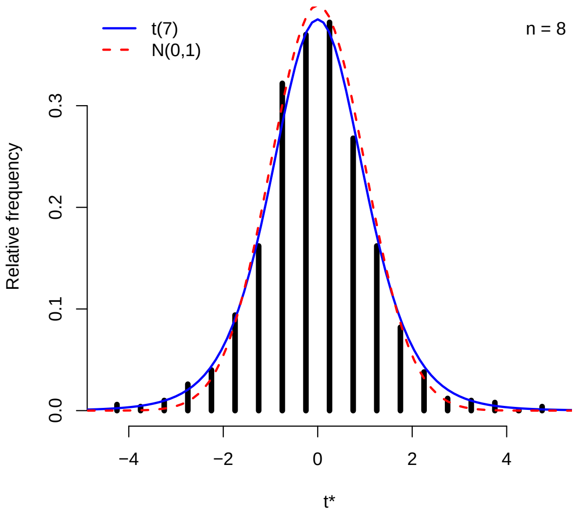

Figure 4: Sampling distribution of T for a one sample t-test given a sample size of 8 and a parent population N(0,1).

The function `samp.dist` is also helpful in illustrating what happens when statistical assumptions are violated. To demonstrate we will use the pooled variance t-test procedure.

The family of t-procedures allow inferential comparisons of two normal parent populations, X_1 and X_2. If the parent populations can be assumed to have equal variances then t* can be calculated as:

$$t^* = \frac{(\bar{x}_1 - \bar{x}_2) - \mu_0}{\sqrt{MSE\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \qquad (4)$$

where MSE is a pooled variance estimator of the joint distribution, X_1 - X_2:

$$MSE = \frac{\sum_i^c \sum_j^{n_c} (x_{ij} - \bar{x}_i)^2}{\sum_i^c n_i - c},$$

and c is the number of populations being compared (here c = 2).

We wish to test the null hypothesis that the difference in population means is equal to μ_0. If H_0 is true, then t*'s can be represented by a random variable, T, where $T \sim t(n_1 + n_2 - 2)$.

We will create two parent populations with equal means. That is, we will create a scenario where the null is true (if μ_0= 0). However we will specify non-normal parent populations for X_1 and X_2 with unequal variances. As a result our example will violate the requirements of the pooled variance t-test procedure.

```
parent<-rexp(100000,1)
parent2<-runif(100000,min=0,max=2)
```

We now require four sampling distributions: $\bar{X}_1$, S_1^2, $\bar{X}_2$, and S_2^2. These will be run through an auxiliary function representing Eq. 4 to create a conceptual sampling distribution for T. Although this is not required by `samp.dist`, we specify the same sample sizes for X_1 and X_2; $n_1 = n_2 = 6$.

```
t.star<-function(s.dist1,s.dist2,s.dist3,s.dist4,
s.size=6,s.size2=s.size){
MSE<-(((s.size-1)*s.dist3)+((s.size2-1)*s.dist4))/
(s.size+s.size2-2)
func.res<-(s.dist1-s.dist2)/(sqrt(MSE)*
sqrt((1/s.size)+(1/s.size2)));func.res}
```

Again, we call `t.star` in the `func` argument for `samp.dist`.

```
samp.dist(parent,parent2=parent2,s.size=6,s.size2=
s.size,stat=mean,stat2=mean,stat3=var,stat4=var,
xlab="t*",func=t.star,show.n=FALSE)
curve(dt(x,10),from=-6,to=6,add=TRUE,lwd=2)
legend("topleft",lwd=2,col=1,legend="t(10)",bty="n")
```

The sampling distribution is strongly negatively skewed (Fig. 5). A t-distribution is clearly inappropriate here.

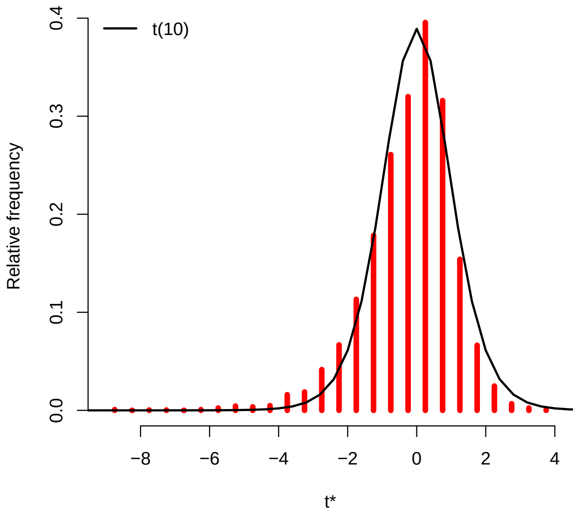

Figure 5: Sampling distribution for T for a pooled variance t procedure, given uniform and exponential parent distributions, and sample sizes: $n_1 = n_2 = 6$.

Similar approaches to these can be used to explore other test statistics, i.e. F^*, X^2, etc.

Likelihood

Vital to modern statistical procedures is the topic of likelihood. Likelihood functions underlie information-theoretic and Bayesian approaches embraced by biologists, but are often poorly understood (Burnham and Anderson , 2002).

The function `loglik.plot` depicts log-likelihood for both important probability density functions (e.g. normal, exponential) and customized likelihood functions. It creates two animated plots. The first shows the log-likelihood function, and demonstrates the derivation of a maximum likelihood estimate (MLE) for a specified parameter. The second shows a probability density function (pdf) that uses MLEs for parameters, and demonstrates how likelihood is calculated.

With likelihood we must first assume a probability term or a pdf that describes a process of interest. The likelihood function is:

$$\mathcal{L}(\theta|\text{data}) = \prod_{i=1}^{n} f(x_i|\theta), \tag{5}$$

where $f(x_i|\theta)$ represents a pdf applied to the ith observation in a sample of size n, given a parameter, θ, required by the pdf. The log-likelihood function is generally more straightforward to apply.

$$\ell(\theta|\text{data}) = \ln\left[\prod_{i=1}^{n} f(x_i|\theta)\right] = \sum_{i=1}^{n} \ln f(x_i|\theta) \tag{6}$$

The MLE for θ is an estimate, $\hat{\theta}$, that maximizes the likelihood or log-likelihood function.

If we assume that a process is normally distributed, then it will have the following log-likelihood function:

$$\begin{aligned} \ell &= \ell(\mu,\sigma^2|\text{data}) \\ &= -n\left(\ln\sigma + \frac{1}{2}\ln 2\pi\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2. \end{aligned} \tag{7}$$

Conceptually the MLE for μ is found by holding σ and the data constant, varying possible estimates for μ, and finding the value that maximizes the function.

Optimization can often be used to derive a ML estimator. For instance, to find the ML estimator for μ we first take the derivative of ℓ with respect to μ:

$$\begin{aligned} \frac{d\ell}{d\mu} &= 0 + \frac{d\ell}{d\mu}\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right] \\ &= (-1)\left(-\frac{2}{2\sigma^2}\right)\sum_{i=1}^{n}(x_i-\mu) \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu). \end{aligned}$$

We then substitute $\hat{\mu}$ for μ, take the derivative equal to 0, and solve for $\hat{\mu}$.

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\hat{\mu}) = 0 \qquad \sum_{i=1}^{n}(x_i-\hat{\mu}) = 0$$

$$\sum_{i=1}^{n}x_i - n\hat{\mu} = 0 \qquad \frac{\sum_{i=1}^{n}x_i}{n} = \hat{\mu}$$

A second derivative test assures us that this solution is a maximum and that the maximum likelihood estimator of μ is the sample mean.

We can demonstrate these important ideas using `loglik.plot`. Consider 10 plant heights (in cm) taken for the forb *Pedicularis oederi* in random sampling of an alpine meadow. Based on previous work we assume that the underlying population of plant heights is normal. Fig. 6a shows the normal log-likelihood function for μ given these data. The function is maximized at 10.79 which is the value of the arithmetic mean. Figure 6b shows a normal distribution with ML estimates for parameters, i.e. the normal distribution whose mean fit the data best. Animation clearly shows that the product of the densities (lengths of lines in Fig. 6b) provides likelihood, while the sum of the logged densities provides log-likelihood. The log-likelihood given MLEs (-14.85; Fig. 6b) is, of course, the height of the log-likelihood function at its maximum (Fig. 6a).

```
X<-c(11.2,10.8,9.0,12.1,10.3,12.4,10.4,10.6,9.3,
11.8);p<-seq(6,15,.01)
loglik.plot(X,parameter="mu",dist="norm",poss=p,
ylim=c(-60,-7),xlim=c(6.5,15))
```

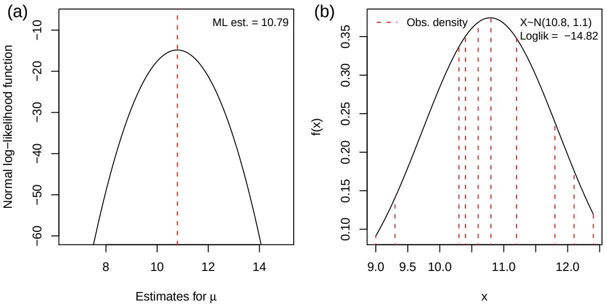

Figure 6: Output from the function `loglik.plot` for the *P. oederi* data (a) Normal log-likelihood function, (b) normal distribution using MLEs with data densities superimposed.

It is interesting to observe the effect of sample size on the likelihood function. For instance if we delete the first five observations from the *Pedicularis oederi* dataset we have:

```
X2<-c(12.4,10.4,10.6,9.3,11.8)
loglik.plot(X2,dist="norm",parameter="mu",poss=p,
ylim=c(-60,-7),xlim=c(6.5,15))
```

As we decrease sample size, less information is contained in the likelihood function and the function attains a platykurtotic (flat) appearance (Fig. 7).

Figures 6 and 7 also demonstrate that, unlike valid pdfs, likelihood functions need not integrate to unity or any other any particular value. The area under the likelihood curve shown in Fig. 6a will be less than the area under the curve in Fig. 7a, and the area of neither will equal one.

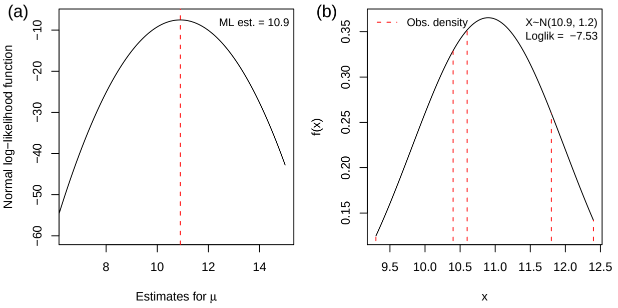

Figure 7: Output from the function `loglik.plot` for a reduced *P. oederi* dataset ($n = 5$) (a) Normal log-likelihood function, (b) normal distribution using MLEs with data densities superimposed.

The same concepts hold for discrete pdfs (e.g. binomial, Poisson). As a biological example of Poisson likelihood we use Dobson (2001) who described the number of chronic medical conditions for women visiting general practitioners in New South Wales. All of the women lived urban locations, were age 70-75, had the same socioeconomic status, and reported to general practitioners three or fewer times in 1996.

```
X<-c(2,0,3,0,0,1,1,1,1,0,0,2,2,1,2,0,0,1,1,1,0
,2,2)
loglik.plot(X,"poi")
```

The MLE for λ is equal to the sample mean (Fig. 8a).

```
mean(X)
[1] 1
```

As with continuous pdfs, likelihood is the product of densities (lengths of lines in Fig. 8b) while log-likelihood is the sum of the log densities.

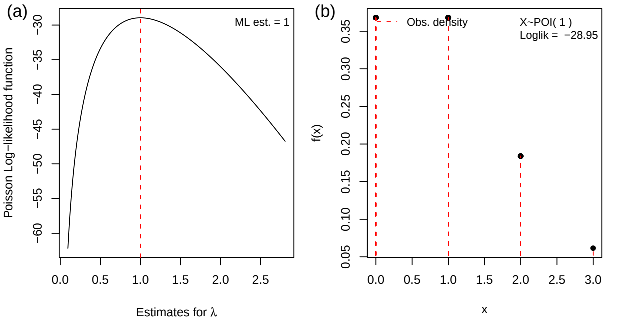

Figure 8: Output from the function `loglik.plot` for the New South Wales dataset (a) Poisson log-likelihood function, (b) data densities for a Poisson distribution using the MLE for λ.

A Bayesian view of likelihood

The goal of most Bayesian computations is to find the posterior distribution. That is, to find $P(\theta|\text{data})$ in:

$$P(\theta|\text{data}) = \frac{P(\text{data}|\theta)P(\theta)}{P(\text{data})}. \quad (8)$$

The denominator in Eq. 8 is a normalizing constant that scales the posterior distribution to the range [0, 1]. Dropping the denominator simplifies Eq. 8 to:

$$P(\theta|\text{data}) \propto P(\text{data}|\theta)P(\theta),$$

or, more explicitly:

$$P(\theta|\text{data}) = cP(\text{data}|\theta)P(\theta), \tag{9}$$

where c is an arbitrary constant of proportionality that absorbs the denominator, P(data), and those parts of the sample and prior distributions that do not involve θ. In Bayesian analyses, the term $cP(\text{data} \mid \theta)$ is called the likelihood function (Gelman et al., 2003). Specifically, $\mathcal{L}(\theta \mid \text{data})$ is proportional to $P(\text{data} \mid \theta)$ up to an arbitrary constant (Edwards, 1972).

An additional perspective into likelihood is provided by this definition. Consider a pdf that describes the probability of sample data given the parameter θ, i.e. $P(\text{data} \mid \theta)$. A likelihood statement concerning θ provided by the entire pdf will be identical (in information content) to one provided by only that part of the pdf involving θ. The rest of the pdf can be apportioned to the arbitrary constant c in the likelihood function, $cP(\text{data} \mid \theta)$.

For instance, consider a binomial variable in which 10 successes and 5 failures are observed. Likelihood information about θ gained by varying θ in $f_1(\theta) = \begin{pmatrix} 15 \\ 5 \end{pmatrix} \theta^5 (1-\theta)^{10}$ is identical to that provided by varying θ in $f_2(\theta) = \theta^5 (1-\theta)^{10}$.

This can be illustrated in `loglik.plot` by specifying `dist="custom"`, and calling a function representing $f_2(\theta)$ in the `func` argument. This can then be be compared to the full binomial pdf, $f_1(\theta)$ (Fig. 9).

```
X<-c(rep(1,5),rep(0,10))
loglik.plot(X,poss=seq(0,1,.01),dist="bin",
plot.density=FALSE,xlab=expression(theta),
ylab=expression(paste(f[1],"(",theta,")")))
f2<-function(X=NULL,p)p^5*(1-p)^10
loglik.plot(X=NULL,func=f2,seq(0,1,.01),
dist="custom",xlab=expression(theta),
ylab=expression(paste(f[2],"(",theta,")")))
```

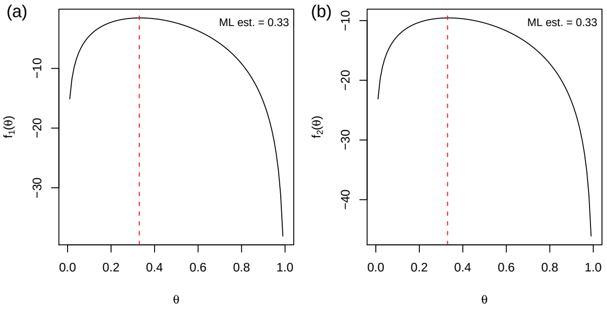

Figure 9: Likelihood functions using a) the entire binomial pdf, and b) only that part of the pdf that includes the binomial parameter, θ.

Changing the value of the constant, c, results only in a change of the units of measurement in the ordinate (Fig. 9). Both likelihood functions are maximized at the ML estimator for θ, which once again is $\bar{x}$.

While it is not obvious at first glance, a Bayesian conception of likelihood corresponds to the frequentist description given in earlier examples. The two approaches provide the same information about θ which is derived in the same way, by varying θ (or $\hat{\theta}$ in a frequentist paradigm), while holding the data constant. The distinction is that Bayesian methods then require that the likelihood function be multiplied by priors, while non-Bayesian likelihood-based methods (e.g. information theoretic approaches) do not.

Summary

This paper presents simple animated functions from package **asbio** for conceptualizing sampling distributions, and illustrating ML estimation, and likelihood calculation. Along with demonstrations of functions I have tried to impart some of my own approaches for teaching these concepts. I hope that this information will be useful to both educators and students of statistics.

Acknowledgements

This work was made possible, in part, by grants from the National Science Foundation, and the National Institute of Standards and Technology. Thanks to T. Bowyer and D. Derryberry (ISU), and R. Remington (Quantified Inc.) for helpful reviews.

Bibliography

Aho, K. *asbio*; a collection of statistical tools for biologists. R package version 0.3-1, 2010. URL http://cran.r-project.org/web/packages/asbio/index.html.

Burnham, K. P., and Anderson, D. R. *Model Selection and Multimodel Inference, A Practical Information-Theoretic Approach, 2nd Edition*. Springer, NY, 2002.

Dobson, A. J. *An introduction to generalized linear models, 2nd edition*. London: Chapman and Hall, 2001.

Edwards, A. W. F. *Likelihood, Expanded Edition*. Johns Hopkins University Press, 1972.

Fox, J. *An R and S-Plus Companion to Applied Regression* Sage Publications, 2002.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. *Bayesian data analysis, 2nd edition*. Chapman and Hall/CRC, 2003.

Horgan, G. W., Elston, D. A., Franklin, M. F. Glasbey, C. A., Hunter, E. A., Talbot, M., Kempton, R. A., McNichol, J. W., and Wright, F. Teaching statistics to biological research scientists. *the Statistician* 48(3): 393-400, 1999.

Xie, Y., and Chang, X. animation: a package for statistical animations *the R Journal* 8(2): 23:27, 2008.

Ken Aho
Idaho State Univeristy
921 S. 8th, Pocatello, ID., 83209-8007
USA
ahoken@isu.edu