FISH 559: Numerical Computing for the Natural Resources
Homework 2 : Estimation of proportions

One of the inputs to a stock assessment is the “mean” sex-ratio over the period of years. The objective of this assignment is to evaluate methods for estimating this sex-ratio and its standard error when the data available are numbers of males and females by year. The tasks to be undertaken are:

1. Show that the conventional estimator $\sum_y M_y / \sum_y (M_y + F_y)$ where M_y is the number of males in year *y*, and F_y is the number of females in year *y* is the maximum likelihood estimate when it is assumed that the sex-ratio is independent of time and each year’s data is the result of a binomial sample, i.e.

$$M_y \sim Bi(\bar{p}, M_y + F_y) \qquad (1)$$

where $\bar{p}$ is the mean sex-ratio.

$$n_y = M_y + F_y$$

$$L(n_1, \ldots n_y \mid p) = \prod_{i=1}^{y} \left(\frac{(M_i + F_i)!}{M_i!\,(F)} \right) p^{M_i} (1-p)^{F_i}$$

$$\ln L(n_y \mid p) = \sum_{i=1}^{y} \ln\left(\frac{(M_i+F_i)!}{M_i!\,(F)} \right) + M_i \ln p + F_i \ln(1-p)$$

$$\frac{d \ln L(n_y \mid p)}{dp} = 0 + \frac{\sum_{i=1}^{y} M_i}{p} - \frac{\sum_{i=1}^{y} F_i}{1-p}$$

Set derivative equal to 0

$$p(1-p)\left[\frac{\sum_{i=1}^{y} M_i}{p} - \frac{\sum_{i=1}^{y} F_i}{1-p} \right] = 0 \; (p(1-p)$$

$$(1-p) \sum_{i=1}^{y} M_i - p \sum_{i=1}^{y} F_i = 0$$

$$-1 \left(\sum_{i=1}^{y} M_i - p \sum_{i=1}^{y} M_i - p \sum_{i=1}^{y} F_i \right) = (0) - 1$$

$$- \sum_{i=1}^{y} M_i + p \sum_{i=1}^{y} M_i + p \sum_{i=1}^{y} F_i = 0$$

$$- \sum_{i=1}^{y} M_i + p \sum_{i=1}^{y} (M_i + F_i) = 0$$

$$\boxed{p = \frac{\sum_i^y M_i}{\sum_i^y (M_i + F_i)}}$$

2. Use TMB to fit model (1) to the data set in HOME2.DAT and report the estimate of $\bar{p}$ and its standard error. Only use data for years for which the sample size for both males and females is at least 1.

Parameter	Estimate	Standard error
$\bar{p}$	0.3536	0.0027

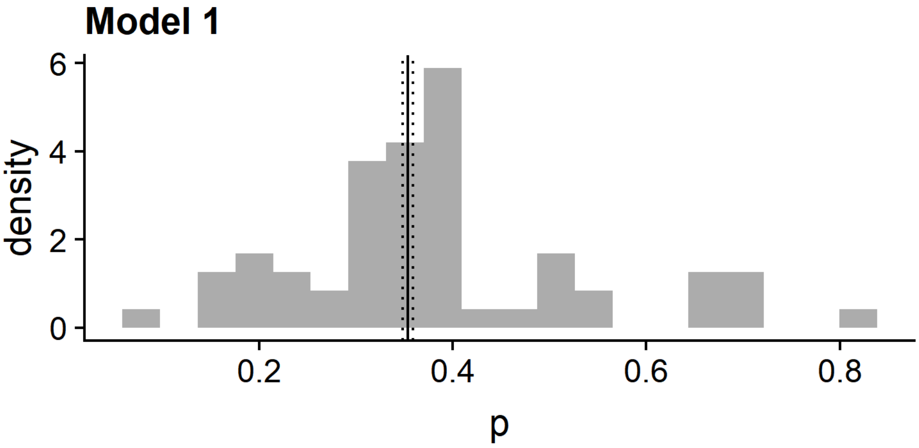


Figure 1. The observed annual sex ratios (p) are shown in grey bars, and the estimated $\bar{p}$ is the black line with +/- 1.96 * s.e. shown as dashed lines.

3. The assumption that p is constant over time is unrealistic. Use simulation to examine whether the coverage probability of this estimator equals the nominal 95% level when the annual sample size is 100 and $p \sim Beta(2,1)$, i.e. how often the estimated 95% confidence interval contains the true value. Base your simulations on 1,000 replicates and 25 years of data.

The coverage probability is relatively poor:

95% CI contains true p?	Count	Percent
Yes	6,975	27.9%
No	18,025	72.1%

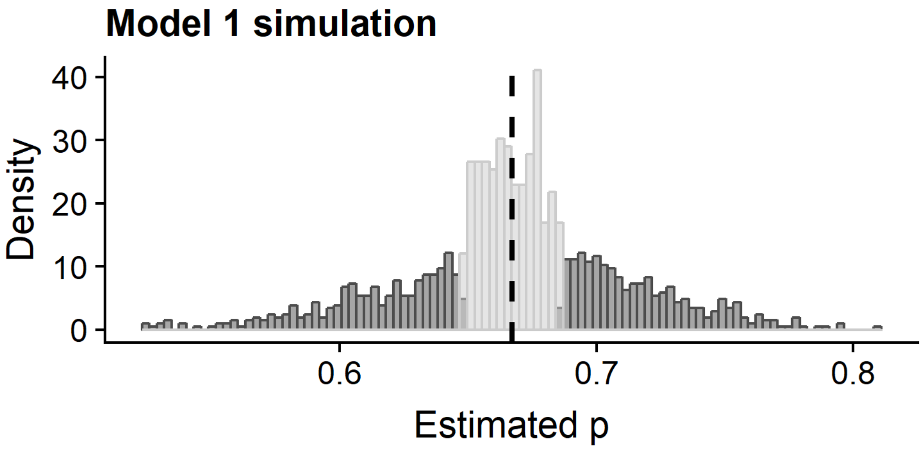


Figure 2. Histogram of estimated *p*s using Model 1 compared to the true *p* (dashed line). Lighter bars are simulations for which the 95% confidence interval contained the true *p*, while darker bars did not.

4. One way to account for the overdispersion caused by *p* not being constant over time is to allow for "process error", using the following negative log-likelihood:

$$-\ell nL = \sum_y \left(\ell\text{n}\sigma_y + \tfrac{1}{2(\sigma_y)^2}\left[p_y - \bar{p} \right]^2 \right) \tag{2}$$

where p_y is the observed sex-ratio for year *y*, and σ_y is standard error of the sex-ratio for year *y*, accounting for over-dispersion, i.e.:

$$\sigma_y = \sqrt{\tau^2 + p_y(1-p_y)/n_y} \tag{3}$$

where n_y is the number of animals sexed during year *y* ($=M_y + F_y$) and τ is the standard deviation of the process error.

4.a Apply this estimator to the data in HOME2.DAT and report the estimates of $\bar{p}$ and τ and their standard errors.

Parameter	Estimate	Standard error
$\bar{p}$	0.3629	0.0171
τ	0.1233	0.0137

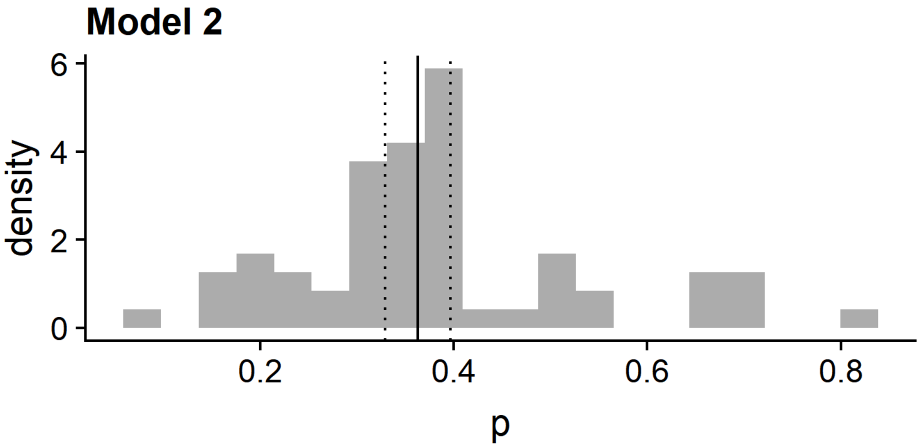


Figure 3. The observed annual sex ratios (*p*) are shown in grey bars, and the estimated $\bar{p}$ is the black line with +/- 1.96 * s.e. shown as dashed lines. Model 2 has a higher variance than Model 1, which is more realistic considering the range of observed *p*'s.

4.b. Apply this equation to the simulated data (task 3) to evaluate how the coverage probability is improved.

The coverage probability is improved, which was expected because the confidence interval is broader in Model 2. However, the model still poorly estimates the *p*.

95% CI contains true *p*?	Count	Percent
Yes	22,850	91.4%
No	2,150	8.6%

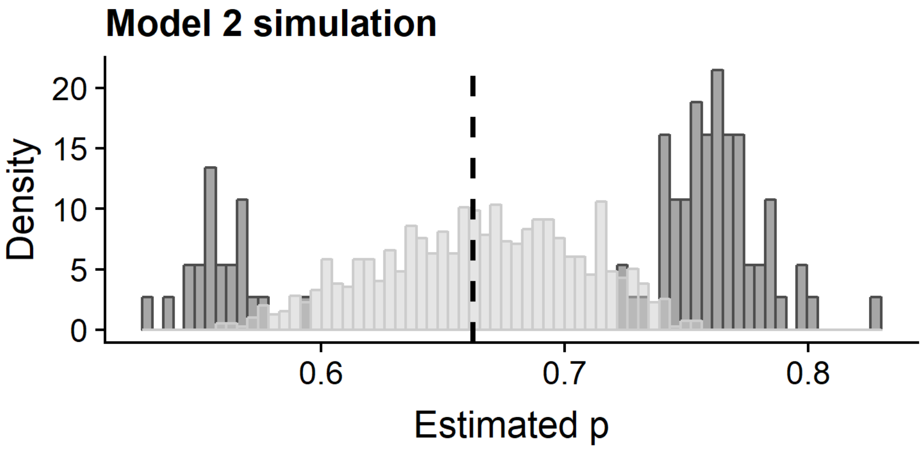
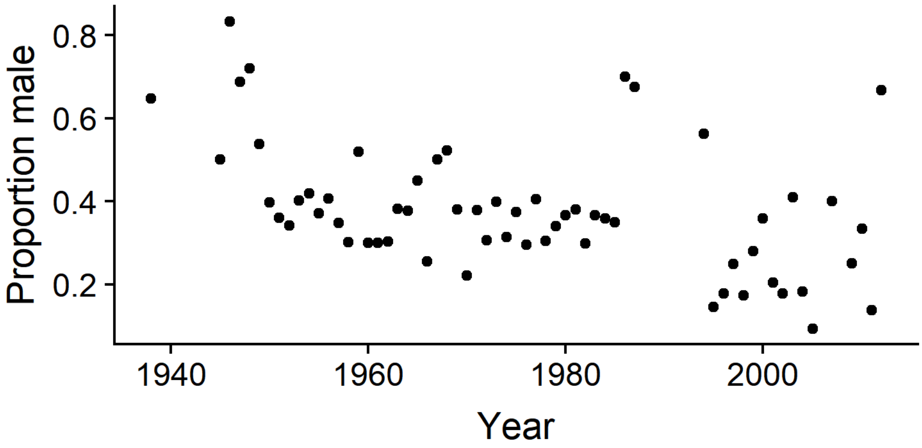


Figure 4. Histogram of estimated *p*s using Model 2 compared to the true *p* (dashed line). Lighter bars are simulations for which the 95% confidence interval contained the true *p*, while darker bars did not.

Equation 2 is still subject to criticism. Comment on whether you think the assumption of a binomial variance for an individual year is valid and whether the assumption that the *p*s are normal is valid. If not, suggest alternatives which you believe are more appropriate.

The binomial variance for an individual year ${\sigma_y}^2$ modelled as a function of process error τ appears valid, because it accounts for sample size and increases the variance estimate of *p*. The assumption that the *p*'s come from a normal distribution is violated. The *p*s are overdispersed. Alternative models that account for dispersion include the negative binomial, the beta-binomial, or Neyman type A.

Another issue with applying these models to this data is that there appears to be a decreasing trend in observed sex-ratio over the time period (i.e. they do not appear to be independent):



I think a model that somehow captured this trend would be appropriate. For example, one could model p_y as a function of an environmental covariates or account for temporal autocorrelation in the σ_ys.