Announcements

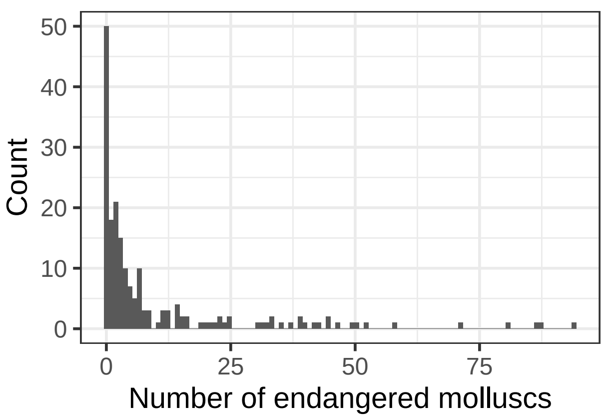
Make sure to sign in on the google form.

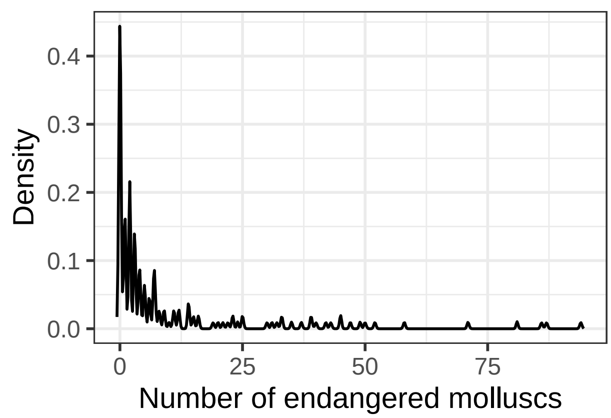
Research in education suggests that when students learn material, take a test, see the material presented again, and take another test, they improve very little from the first to second test. However, when students learn material, take a test, try to solve examples of the material and see how they and others get it wrong, and take the second test, they improve considerably. Thus, my sections will only consist of practice problems, not lecturing, but a compilation of course notes is available from the final review session I led last year.

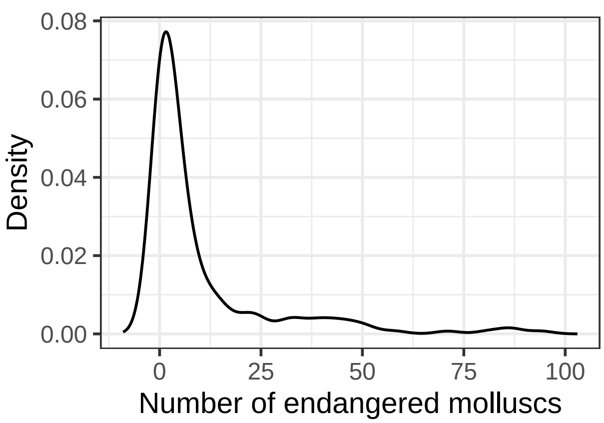
Pset 1 due Friday 2/2

Molluscs

This question will deal with a data set of country-level statistics from this source with an explanation of the data encoding found here. In particular, we'll be looking at the number of threatened species of molluscs (snails, clams, etc.).







1. What distribution does this seem to follow? What are some advantages and disadvantages to each data visualization?

The data seem roughly geometric since it is count data (though with an even larger right tail than usual). The narrow bandwidth is clearly too narrow, but the wide bandwidth suggests there is density less than 0. The histogram is probably the best way to go.

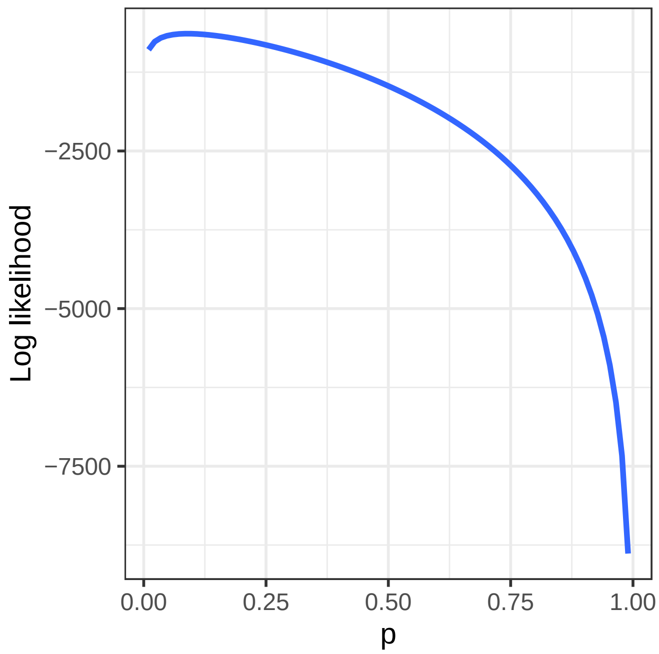
2. Let Y_i be the number of endangered mollusk species in country i for $i \in \{1, ..., 190\}$ and suppose $Y_i \sim \text{Geom}(p)$. Find the log likelihood function for p given $y_1, ..., y_{190}$. (Note that PMFs and PDFs for all major distributions can be found in Appendix C of the Stat 110 book.)

The likelihood function is:

$$L(p; y_1, ..., y_{190}) = \prod_{i=1}^{190} p(1-p)^{y_i} = p^{190}(1-p)^{\sum_{i=1}^{190} y_i}$$

Taking the log gives:

$$l(p; y_1, ..., y_{190}) = 190\log(p) + \left(\sum_{i=1}^{190} y_i\right)\log(1-p)$$



3. Find the $\hat{p}$ that maximizes your log likelihood function for general y_i. In the dataset, $\sum_{i=1}^{190} y_i = 1929$. Is this consistent with the plot above?

We can take the derivative and set it to 0:

$$\begin{aligned} l'(p; y_1, ..., y_{190}) &= \frac{190}{\hat{p}} - \frac{\left(\sum_{i=1}^{190} y_i\right)}{1-\hat{p}} = 0 \\ &\implies 190 - 190\hat{p} = \left(\sum_{i=1}^{190} y_i\right)\hat{p} \\ &\implies \hat{p} = \frac{190}{\left(\sum_{i=1}^{190} y_i\right) + 190} \end{aligned}$$

The estimate from the data is:

```
## [1] 0.08966494
```

This looks consistent with the plot above.

4. Express your $\hat{p}$ in terms of the sample mean $\bar{y}$ and relate this to the mean of a geometric distribution: $(1-p)/p$.

$$\hat{p} = \frac{1}{\bar{y}+1} \implies \frac{1-\hat{p}}{\hat{p}} = \bar{y}$$

5. The result above implies that $\bar{y}$ contains as much information about $\hat{p}$ as all the $y_1, ...y_{190}$ together. However, intuitively, it seems like the standard deviation, the kurtosis, and all sorts of other features from the data might carry useful information. How can this be?

Because we've specified that the data follow a geometric distribution, we've made all of this other information irrelevant for maximum likelihood maximization. In a geometric distribution, p carries all the information there is about the distribution, and it is fundamentally tied to the mean of the distribution.

6. In the process above, what was our estimand, what was our estimator, and what was our estimate?

Our estimand, the object to infer, is the underlying parameter p. Our estimator, the statistic we're going to use to estimate p, is $\frac{1}{\bar{Y}+1}$. Our estimate, the crystallized value of the estimator, is $\frac{1}{\bar{y}+1}$.

7. Suppose a new country is taking an endangered mollusk census. Their initial data show that the country has at least 15 endangered mollusk species. Given this information, find the expected number of endangered mollusk species in the country. Do this in two ways: first, calculate the expected value using a sum with conditioning; second, use a trick with a name I can't remember.

Let X be the number of mollusk species in the country. We want $E(X|X \geq 15)$. Using the definition of conditional expectation,

$$\begin{aligned}
E(X|X \geq 15) &= \sum_{k=15}^{\infty} kP(X = k|X \geq 15) \\
&= \sum_{k=15}^{\infty} kpq^{k-15} \\
&= \sum_{k=0}^{\infty} (15 + k)pq^k \\
&= 15 \sum_{k=0}^{\infty} pq^k + \sum_{k=0}^{\infty} kpq^k \\
&= 15 + \frac{q}{p}
\end{aligned}$$

where the third equality comes from reparameterizing the sum and the fifth equality comes from the fact that the geometric PMF sums to 1 and the geometric expectation is p/q.

Notably, this is the same result we get when applying the memoryless property of the geometric distribution.

Random ~~steps~~ slides

We've all heard of random walks, but who really only steps on integers? In this problem, we'll be exploring random slides in which a person, at time step t, slides to the left or the right (but does not take it back now y'all) and ends up at a position $Y_t|Y_{t-1} \sim \mathcal{N}(Y_{t-1}, \sigma^2)$ on the real number line.

1. Suppose the person starts at $y_0 = 0$ and takes a series of n slides. Find the likelihood and log likelihood function for σ^2. Which terms of the normal density can be dropped?

$$\begin{aligned}
L(\sigma^2; \vec{y}) &= f_{Y_1}(y_1) f_{Y_2}(y_2|Y_1 = y_1) \cdot \ldots \cdot f_{Y_n}(y_n|Y_1 = y_1, \ldots, Y_{n-1} = y_{n-1}) \\
&= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_1 - 0}{\sigma}\right)^2} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_2 - y_1}{\sigma}\right)^2} \ldots \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_n - y_{n-1}}{\sigma}\right)^2} \\
&= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^n \left(\frac{y_i - y_{i-1}}{\sigma}\right)^2} \\
&\propto \left(\frac{1}{\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - y_{i-1})^2}
\end{aligned}$$

Taking the log, the log likelihood is:

$$-n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - y_{i-1})^2$$

2. Find the value of σ that maximizes the likelihood.

Taking the derivative and setting it to 0 gives

$$\begin{aligned}
l'(\sigma; \vec{y}) &= -\frac{n}{\hat{\sigma}} \cdot + \frac{1}{\hat{\sigma}^3}\sum_{i=1}^n (y_i - y_{i-1})^2 = 0 \\
&\implies \hat{\sigma^2} = \frac{\sum_{i=1}^n (y_i - y_{i-1})^2}{n}
\end{aligned}$$

3. What are the estimand, estimator, and estimate here?

The estimand is σ^2, the true variance. The estimator is $\frac{\sum_{i=1}^n (Y_i - Y_{i-1})^2}{n}$. The estimate is $\frac{\sum_{i=1}^n (y_i - y_{i-1})^2}{n}$.

4. Find the bias of the estimator with an explicit calculation.

Let $X_i = Y_i - Y_{i-1} \sim \mathcal{N}(0, \sigma^2)$.

$$\begin{aligned}
E(\hat{\sigma^2}) &= \frac{\sum_{i=1}^n E\left(Y_i - Y_{i-1}\right)^2}{n} \\
&= \frac{\sum_{i=1}^n E\left(X_i^2\right)}{n} \\
&= \frac{\sum_{i=1}^n \mathrm{Var}\left(X_i\right) + (E(X_i))^2}{n} \\
&= \frac{n\sigma^2}{n} \\
&= \sigma^2
\end{aligned}$$

Thus,

$$\mathrm{Bias}(\hat{\sigma}^2) = E(\hat{\sigma^2}) - \sigma^2 = 0$$

5. Use the law of large numbers to argue that this estimator converges towards σ^2 as $n \to \infty$.

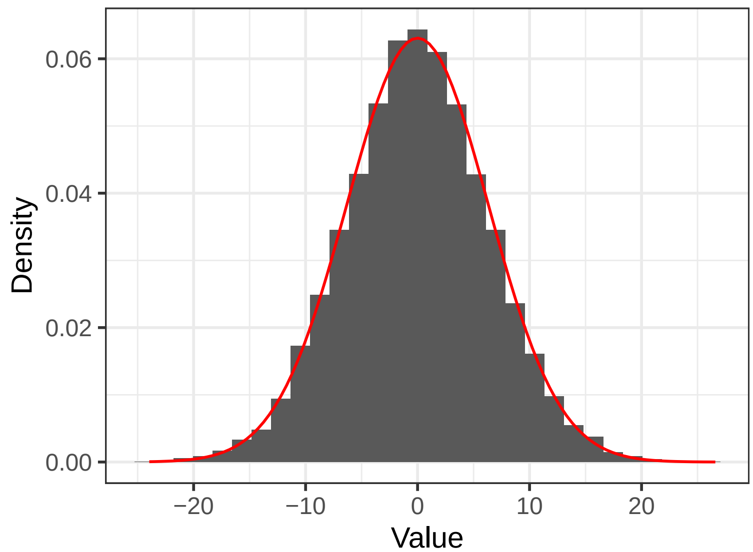
Our estimator $\frac{\sum_{i=1}^{n}(Y_i - Y_{i-1})^2}{n}$ is a mean of squared differences, and we can write $X_i = Y_i - Y_{i=1}$ with $X_i \sim \mathcal{N}(0, \sigma^2)$. Thus, by the law of large numbers,

$$\frac{\sum_{i=1}^{n} X_i^2}{n} \to E(X_i^2) = \sigma^2$$

6. Find the marginal distribution of Y_n.

Let X_i be the difference between the person's location at time i and time $i-1$. Then, $X_i \sim \mathcal{N}(0, \sigma^2)$. $Y_n = \sum_{i=1}^{n} X_i$, and the sum of independent Normals is Normal with a mean as the sum of means and a variance as the sum of variances. Thus, $Y_n \sim \mathcal{N}(0, n\sigma^2)$.

7. We can write a simulation with $n = 10$ and $\sigma = 2$ to verify that the marginal distribution is correct. We'll draw Normal random variables according to the model with `rnorm` and compare them to the true marginal distribution from `dnorm`.



8. Find a maximum likelihood estimator for σ^2 from this distribution.

The likelihood function (dropping multiplicative constants) is

$$L(\sigma^2; \vec{y}) = \frac{1}{\sigma} e^{-\frac{y_n^2}{2n\sigma^2}}$$

Therefore, the log likelihood function is

$$l(\sigma^2; \vec{y}) = -\log(\sigma) - \frac{y_n^2}{2n\sigma^2}$$

Taking the derivative and setting it to 0 gives

$$l'(\sigma^2; \vec{y}) = -\frac{1}{\hat{\sigma}} + \frac{y_n^2}{n\hat{\sigma}^3} = 0 \implies \hat{\sigma}^2 = y_n^2/n$$

9. What is the bias of this estimator?

$$\sigma^2 - E(\hat{\sigma}^2) = \sigma^2 - E(Y_n^2)/n = 0$$

10. What is the standard error of this estimator? (You may find it useful to reparameterize Y_n as $\sqrt{n\sigma^2}Z$ with $Z \sim \mathcal{N}(0, 1)$. The Normal moments from 6.5.2 may also be useful.)

The variance of the estimator is

$$\begin{aligned}
\mathrm{Var}(\hat{\sigma}^2) &= \frac{1}{n^2}\mathrm{Var}(Y_n^2) \\
&= \frac{1}{n^2}[E(Y_n^4) - (E(Y_n^2))^2] \\
&= \frac{1}{n^2}[n^2\sigma^4 E(Z^4) - (n\sigma^2 E(Z^2))^2] \\
&= \frac{1}{n^2}[3n^2\sigma^4 - n^2\sigma^4] \\
&= 2\sigma^4
\end{aligned}$$

Equality 3 uses the reparameterization, and equality 4 uses the standard Normal moments. Thus, the standard error is $\sqrt{2\sigma^4}$.

11. We now have two estimators for the same estimand. Describe when each might be preferable.

In the first estimator, as we add more observations, our estimator converges towards our estimand. However, in our second estimator, as n grows, the standard error doesn't shrink at all! Thus, as we add more observations, our estimate of σ^2 is just as precise as when we started. Probably the only time to use the second estimator is when it is too difficult to store all of the y_i. Otherwise, the first estimator is better.