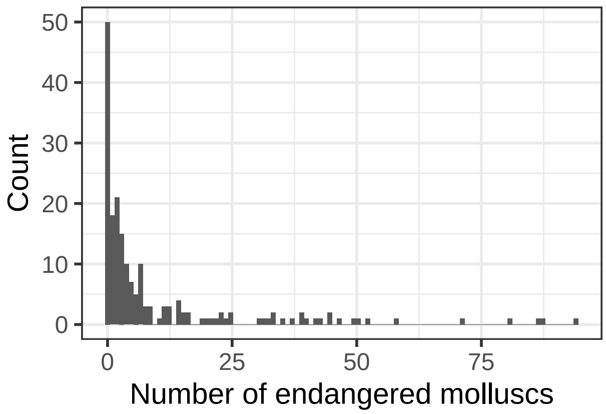
## Announcements

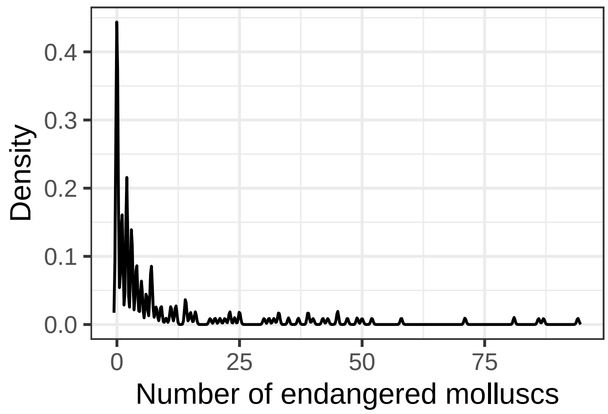
Make sure to sign in on the google form.

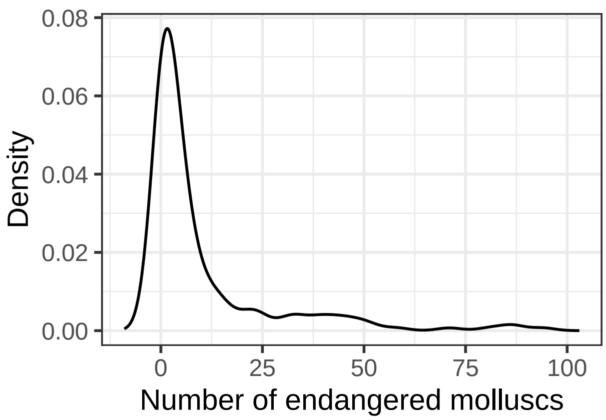
Pset 1 due. . .

## 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.).

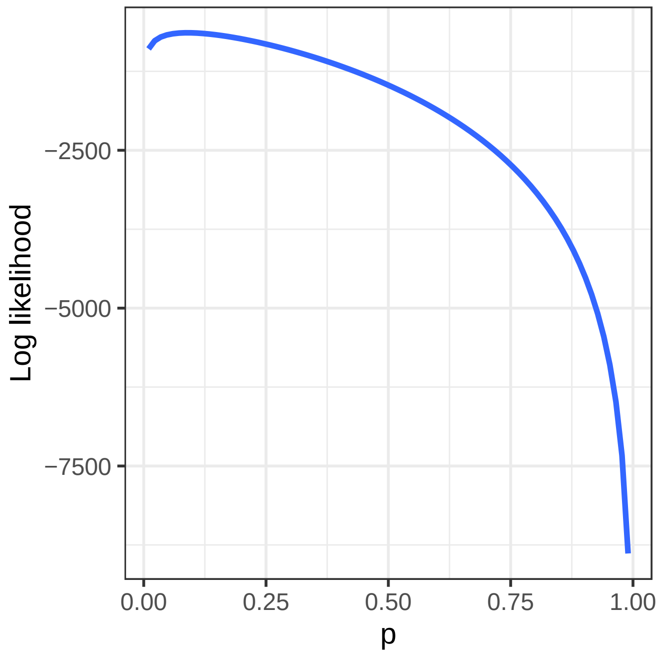



1. What distribution does this seem to follow? What are some advantages and disadvantages to each data visualization?

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.)




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?

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

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?

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

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.

## 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 $\sigma^2$. Which terms of the normal density can be dropped?

2. Find the value of $\sigma$ that maximizes the likelihood.

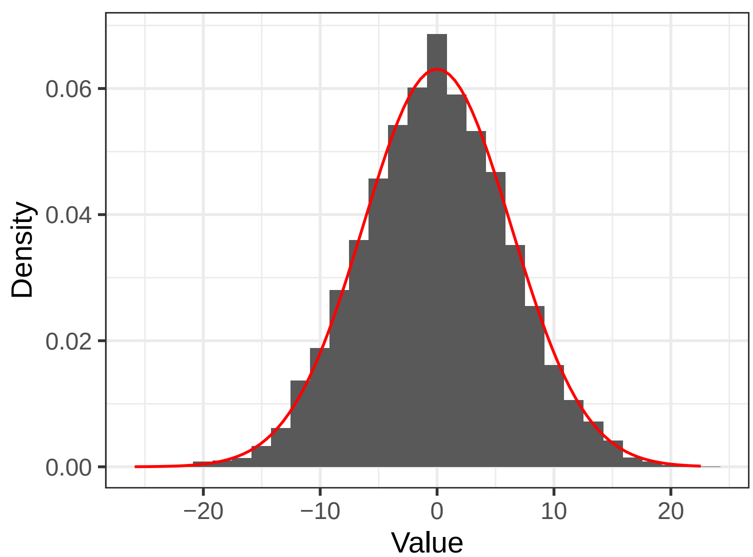
3. What are the estimand, estimator, and estimate here?

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

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

6. Find the marginal distribution of $Y_n$.

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 $\sigma^2$ from this distribution.

9. What is the bias of this estimator?

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.)

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