Announcements

- Make sure to sign in on the google form
- Pset 1 due Friday 2/3

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.

A few useful columns:

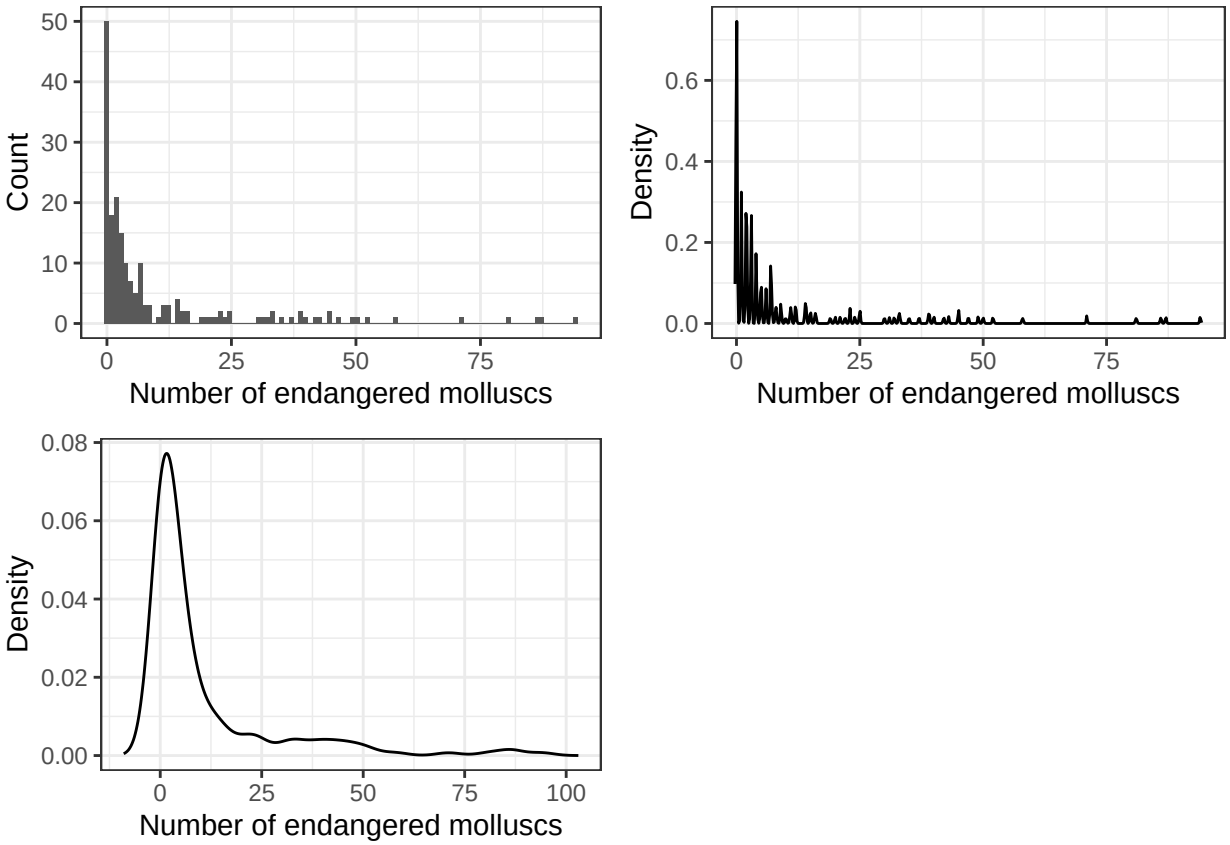
- `bi_molluscs`: Number of threatened species of molluscs (snails, clams, etc.)

```
p1 <- ggplot(countries, aes(x = bi_molluscs)) +
  geom_histogram(bins = 100) +
  theme_bw() +
  xlab("Number of endangered molluscs") +
  ylab("Count")

# Narrow bandwidth
dens <- density(countries$bi_molluscs, bw = 0.1)
p2 <- ggplot(data.frame(x = dens$x, y = dens$y), aes(x = x, y = y)) +
  geom_line() +
  theme_bw() +
  xlab("Number of endangered molluscs") +
  ylab("Density")

# Wide bandwidth
dens <- density(countries$bi_molluscs, bw = 3)
p3 <- ggplot(data.frame(x = dens$x, y = dens$y), aes(x = x, y = y)) +
  geom_line() +
  theme_bw() +
  xlab("Number of endangered molluscs") +
  ylab("Density")

grid.arrange(p1, p2, p3, ncol = 2, nrow=2)
```



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 and plot 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.)

```
p <- seq(0.01, 1-0.01, 0.00005)

logliks <- NA

ggplot(data.frame(p=p, logliks = logliks), aes(x=p, y=logliks)) +
  geom_smooth(method='loess', span=0.001, formula = y~x) +
  ylab("Log likelihood") +
  xlab("p") +
  theme_bw()
```

3. Find the $\hat{p}$ that maximizes your log likelihood function for general y_i and for the data here. Is this consistent with your plot above?

```
# TODO: Find specfic \hat{p} for the data
```

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 is our estimand, what is our estimator, and what is our estimate?

7. Suppose a new country is taking an endangered mollusk census. Their initial data shows 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 whose name I can't remember.

Random ~~walks~~ hops

We've all heard of random walks, but who really only steps on integers? In this problem, we'll be exploring random hops in which a person, at time step t, takes a hop 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 hops. Find the likelihood and log likelihood function for σ^2. Which terms of the Normal density can be dropped?

2. Find the value of σ 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 σ^2 as $n \to \infty$.

6. Find the marginal distribution of Y_n.

7. Write a simulation with $n = 10$ and $\sigma = 2$ to verify that the marginal distribution is correct. Draw Normal random variables according to the model with `rnorm` and compare it to the true marginal distribution from `dnorm`.

```
n <- 10
sigma <- 2
nsims <- 10^4
```

```r
run_simulation <- function() {
  # TODO: Write a function to generate a single simulation draw
  return (NA)
}

# TODO: Repeat the simulation nsims times and store it in 'sim_out'
sim_out <- NA

ggplot(data.frame(sim_out), aes(x = sim_out)) +
  geom_histogram(bins = 30, aes(y = after_stat(density))) +
  stat_function(fun = dnorm, args = c(0, sqrt(n) * sigma), n = 100, col = "red") +
  theme_bw() +
  ylab("Density") +
  xlab("Value")
```

8. Find a maximum likelihood estimator for σ^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.