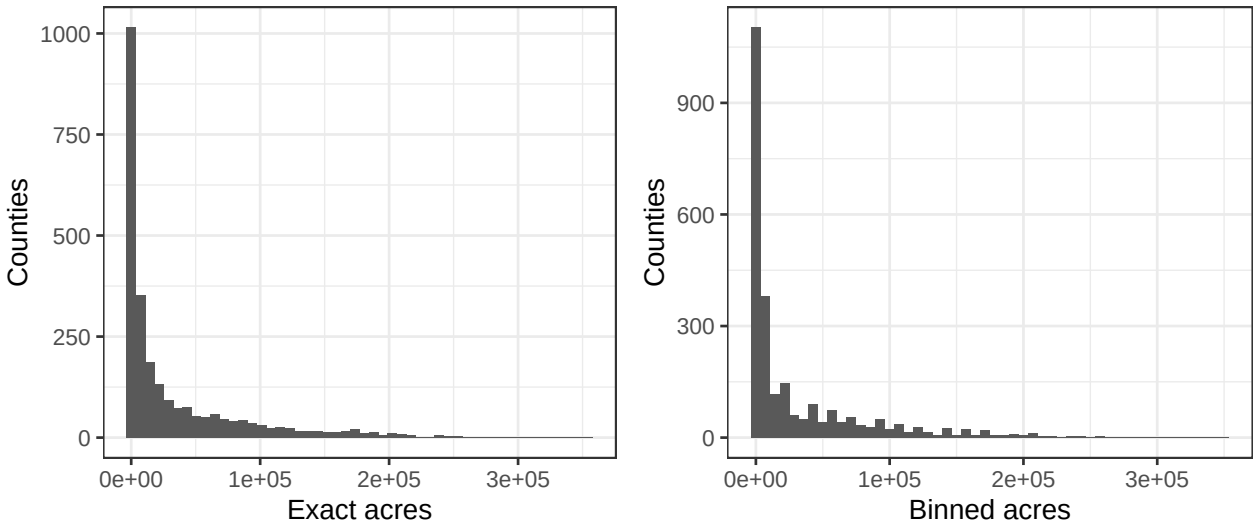
Announcements

- Make sure to sign in on the google form (I send a list of what section questions are useful for what pset questions afterwards)
- Pset 5 due Friday 3/3
- Midterm Tuesday 3/7 in class. Make sure to fill out the conflict form by today if you can't make it.
- In preparation for the midterm, I'll hold office hours rather than section next week at this time

It's Corn - A Big Lump with Knobs

The following questions deal with data on how many acres of corn are planted in each U.S. county available here.



Continuous data is often binned into intervals due to privacy, imprecision, or other reasons. In this set of problems, we'll be modeling corn acreage for U.S. counties under various assumptions.

1. First, assume there has been no binning. Therefore, the acreage devoted to corn farming in a county is i.i.d. $Y_i \sim \text{Expo}(\lambda)$. Show that this model is a Natural Exponential Family by finding θ, $\Psi(\theta)$, and $h(y)$ such that $f_{Y_i}(y) = e^{\theta y - \Psi(\theta)} h(y)$.

2. Use properties of NEFs to find the MLE $\hat{\lambda}$. Also, find the mean and variance of Y_i and the Fisher information of λ using properties of the NEF.

Now, suppose the data has been binned into intervals of τ such that $X_i = \lfloor Y_i/\tau \rfloor \cdot \tau$ where τ is the bin width.

3. Intuitively, will the likelihood function for λ be the same as if we had all the Y_i?

4. Write the likelihood function for this model in terms of the x_i.

5. Find a sufficient statistic for the data by invoking the factorization criterion. (Recall that the factorization criterion says $T(\vec{Y})$ is a sufficient statistic iff we can factor $P(\vec{X} = \vec{x}|\lambda) = g(T(\vec{X}), \lambda)h(\vec{x})$.)

6. As the bin width decreases ($\tau \to 0$), the likelihood function should converge to the likelihood function of i.i.d. exponentials:

$$\lambda^n \exp(-\lambda n \bar{y})$$

   By Taylor expanding the likelihood from (4) with respect to τ, show this is the case.

7. Do the X_i follow a Natural Exponential Family? If so, use $\frac{1}{2^{x+1}}$ as $h(x)$.

8. Find the mean and variance of X_i.

9. Find the MLE for λ.

10. What is the MSE of $\hat{\lambda}_{\text{MLE}}$?

11. Would Rao-Blackwellization improve the MLE for λ? If not, suggest your own improvement.

12. Using the real data, how does the MLE compare to (1) using the points as they are to estimate λ and (2) using the midpoint of each interval to estimate λ? Assess this by calculating the "true" λ from the exact data as one over the sample mean. In the real data, $\tau = 5000$.

```
# TODO: Compare estimators
```

Geometric Underpinnings

It turns out that the distribution above in (7) is just a special form of the Geometric with a support restricted to multiples of τ. In this question, we'll explore the Geometric futher. Let $Y_1, ..., Y_n$ be i.i.d. Geom(p) with p unknown.

1. Show that the model is a Natural Exponential Family by finding θ, $\Psi(\theta)$, and $h(y)$.

2. Use results about NEFs to derive the mean and variance.

3. It follows from (1) that $\bar{Y}$ is a sufficient statistic. Check this directly using the factorization criterion.

4. Consider the estimator of the mean (estimating q/p) Y_1. Use Rao-Blackwellization to improve this estimator.

5. Consider the estimator of the mean (estimating q/p) $\vec{w} \cdot \vec{Y}$ where $\vec{w}$ is a vector of weights that sums to 1. Use Rao-Blackwellization to show that the best $\vec{w}$ is $(1/n, ..., 1/n)$.