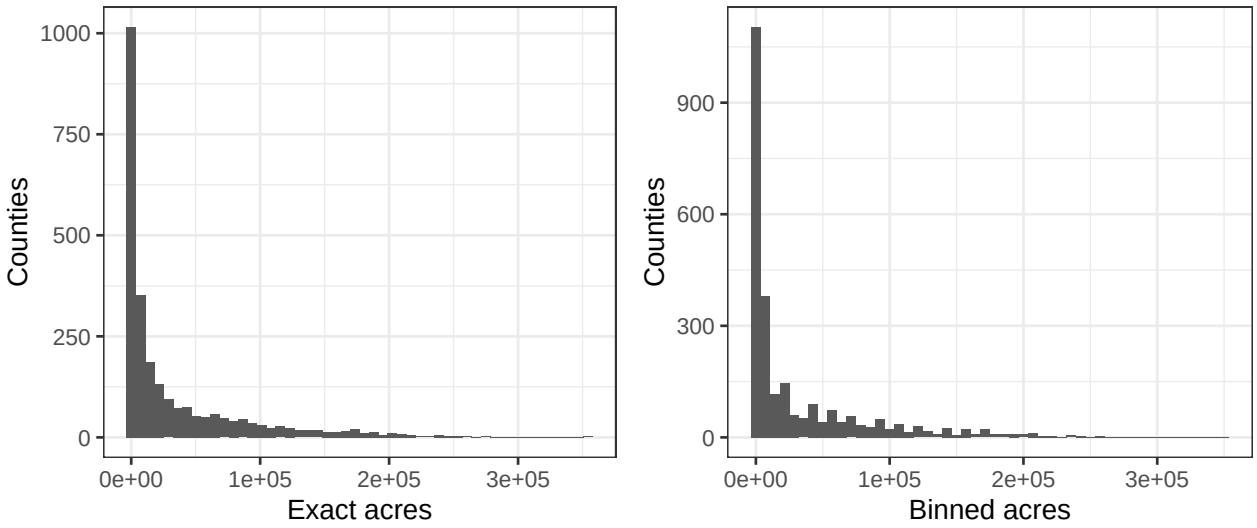
Announcements

- 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 $f_{Y_i}(y) = e^{\theta y - \Psi(\theta)} h(y)$ where $h(y)$ is a PDF or PMF.

$$f_{Y_i}(y) = \lambda e^{-\lambda y} = e^{-\lambda y + \log(\lambda)}$$

Since we want $h(y)$ to be a PDF, we multiply and divide by e^{-y}:

$$f_{Y_i}(y) = \lambda e^{-\lambda y} = e^{(-\lambda+1)y + \log(\lambda)} e^{-y}$$

Thus, $\theta = 1 - \lambda$, $\Psi(\theta) = -\log(1-\theta)$, $h(y) = e^{-y}$. Note that the natural parameter is $1 - \lambda$ rather than λ.

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.

Since the Exponential is a NEF, the MLE for the mean is $\bar{Y}$, so $\hat{\lambda}_{\text{MLE}} = 1/\bar{Y}$ by invariance. $E(Y_i) = \Psi'(\theta) = \frac{1}{1-\theta} = \frac{1}{\lambda}$ as expected.

$$\mathcal{I}_{Y_1}(\theta) = \text{Var}(Y_i) = \Psi''(\theta) = \frac{1}{(1-\theta)^2} = \frac{1}{\lambda^2}$$

$\lambda = g(\theta) = 1 - \theta$, so $g'(\theta) = -1$, so a transformation of Fisher information gives $\mathcal{I}(\lambda) = \mathcal{I}(\theta)/(-1)^2 = \frac{1}{\lambda^2}$.

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?

No: we are losing information when we bin the data, so the likelihood functions will not be the same.

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

$$\begin{aligned} L(\lambda; \vec{x}) &= \prod_{i=1}^{n} (F_{Y_i}(x_i + \tau) - F_{Y_i}(x_i)) \\ &= \prod_{i=1}^{n} (1 - e^{-\lambda(x_i+\tau)} - (1 - e^{-\lambda x_i})) \\ &= \prod_{i=1}^{n} (-e^{-\lambda(x_i+\tau)} + e^{-\lambda x_i}) \\ &= \prod_{i=1}^{n} e^{-\lambda x_i}(-e^{-\lambda\tau} + 1) \\ &= (1 - e^{-\lambda\tau})^n \exp(-\lambda n \bar{x}) \end{aligned}$$

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 $P(\vec{X} = \vec{x}|\lambda) = g(T(\vec{X}), \lambda)h(\vec{x})$.)

The sufficient statistic is $\bar{X}$. The joint PMF above can be written as $g(\bar{X}, \lambda)h(\vec{x})$ where

$$g(\bar{X}) = (1 - e^{-\lambda\tau})^n \exp\left(-\lambda n \bar{X}\right), \qquad h(\vec{x}) = 1$$

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.

A first order Taylor expansion for e^x is $1 + x$ when x is near 0. Therefore,

$$(1 - e^{-\lambda\tau})^n \exp(-\lambda n \bar{x}) \approx (1 - (1 - \lambda\tau))^n \exp(-\lambda n \bar{x}) = (\lambda\tau)^n \exp(-\lambda n \bar{x})$$

τ^n is a multiplicative constant for λ, so we can drop it to get $\lambda^n \exp(-\lambda n \bar{x})$ as desired. Note that the approximation is best when $\lambda\tau$ is small, which can either come from a small τ (small bins) or a small λ. A small λ corresponds to large observations, so the approximation will still be good if we have large observations and τ is moderate. This is what we would expect: the approximation is good when the bins are small relative to the actual data, not just when they are small absolutely.

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

$$P(X_i = x|\lambda) = (1 - e^{-\lambda\tau}) \exp(-\lambda x)$$

Since we want $h(x)$ to be a PMF, we can multiply and divide by $\frac{1}{2^{x+1}}$:

$$P(X_i = x|\lambda) = (1 - e^{-\lambda\tau}) \exp(-\lambda x + (x+1)\log(2)) \frac{1}{2^{x+1}} = \exp(x(-\lambda + \log(2)) + \log\left(2(1 - e^{-\lambda\tau})\right)) \frac{1}{2^{x+1}}$$

Thus, $\theta = -\lambda + \log(2)$, $\Psi(\theta) = -\log\left(2(1 - e^{\tau(\theta - \log(2))})\right) = -\log(2 - \frac{e^{\tau\theta}}{2^{\tau-1}})$, $h(x) = \frac{1}{2^{x+1}}$, and X_i do follow a Natural Exponential Family!

8. Find the mean and variance of X_i.

$$E(X_i) = \Psi'(\theta) = \frac{\tau \frac{e^{\tau\theta}}{2^\tau - 1}}{2 - \frac{e^{\tau\theta}}{2^\tau - 1}} = \frac{\tau}{2^\tau e^{-\tau\theta} - 1} = \frac{\tau}{e^{\tau\lambda} - 1}$$

$$\text{Var}(X_i) = \Psi''(\theta) = \frac{2^\tau \tau^2 e^{-\tau\theta}}{\left(2^\tau e^{-\tau\theta} - 1\right)^2} = \frac{\tau^2 e^{\tau\lambda}}{(e^{\tau\lambda} - 1)^2}$$

We can check these are right with simulation:

```
set.seed(111)
lambda <- 1/5
n <- 1000000
x <- rexp(n, lambda)
tau <- 5
y <- floor(x / tau) * tau

# Mean
c("Predicted" = tau / (exp(tau * lambda) - 1), "Observed" = mean(y))
```

```
## Predicted  Observed
##  2.909884  2.907590
```

```
# Variance
c("Predicted" = tau^2 * exp(tau * lambda) / (exp(tau * lambda) - 1)^2, "Observed" = var(y))
```

```
## Predicted  Observed
##  23.01684  23.04224
```

9. Find the MLE for λ.

Better answer by Beatriz Terres:

We know that the MLE for the mean parameter in a Natural Exponential Family is $\hat{\mu} = \bar{x}$. We can rearrange the result above to show

$$\frac{\log\left(\frac{\tau}{E(X_i)} + 1\right)}{\tau} = \lambda$$

so invariance of the MLE gives

$$\hat{\lambda} = \frac{\log\left(\frac{\tau}{\bar{x}} + 1\right)}{\tau}$$

Original answer:

$$L(\lambda; \vec{x}) = (1 - e^{-\lambda\tau})^n \exp\left(-\lambda n \bar{x}\right)$$

$$\ell(\lambda; \vec{x}) = n \log(1 - e^{-\lambda\tau}) - \lambda n \bar{x}$$

$$s(\lambda; \vec{x}) = \frac{n\tau e^{-\hat{\lambda}\tau}}{1 - e^{-\hat{\lambda}\tau}} - n\bar{x} = 0$$

$$\implies \frac{\tau}{e^{\hat{\lambda}\tau} - 1} = \bar{x} \implies \frac{\log\left(\frac{\tau}{\bar{x}} + 1\right)}{\tau} = \hat{\lambda}$$

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

There is a positive probability of $\bar{X}$ being 0 because we are binning and using the minimum of the bin. Therefore, the expectation is infinity, so the bias is infinity, and the MSE is infinite.

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

No, the MLE is a function of the sufficient statistic because the likelihood only depends on the sufficient statistic. Therefore, conditioning on a sufficient statistic will not change the MLE. We could avoid the divide-by-zero issue by adding a small value like 10^{-3} to $\bar{X}$ so it will always be positive.

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, $\tau = 5000$.

```
true_lambda <- 1/mean(crops_real$Acres)
tau <- 5000
mle = log(tau/mean(crops_binned$Acres) + 1) / tau
left_bin <- 1/mean(crops_binned$Acres)
mid_bin <- 1/mean(crops_binned$Acres + tau / 2)

df <- round(rbind(c(true_lambda, mle, left_bin, mid_bin), true_lambda -
                    c(true_lambda, mle, left_bin, mid_bin)), 10)
colnames(df) <- c("True lambda", "MLE", "Bin left", "Bin center")
rownames(df) <- c("Value", "Difference from true lambda")
df
```

```
##                              True lambda         MLE     Bin left  Bin center
## Value                        2.91777e-05 2.87297e-05  3.08957e-05 2.86804e-05
## Difference from true lambda 0.00000e+00 4.48000e-07 -1.71800e-06 4.97300e-07
```

The MLE does considerably better than using the bin's minimum value and slightly better than using the bin's center value. See below that the MLE does slightly better when the data is exactly exponential and much better when τ is large. This bin's left endpoint clearly underestimates the data, and using the bin's center assumes that each value in the bin is equally likely, but smaller values in the bin are actually more likely.

```
lambda <- 1/5
n <- 1000000
x <- rexp(n, lambda)
tau <- 1
y <- floor(x / tau) * tau

mle = log(tau/mean(y) + 1) / tau
left_bin <- 1/mean(y)
mid_bin <- 1/mean(y + tau / 2)

df <- round(rbind(c(lambda, mle, left_bin, mid_bin), lambda -
                    c(lambda, mle, left_bin, mid_bin)), 5)

tau <- 5
y <- floor(x / tau) * tau

mle = log(tau/mean(y) + 1) / tau
left_bin <- 1/mean(y)
mid_bin <- 1/mean(y + tau / 2)

df <- rbind(df, round(rbind(c(lambda, mle, left_bin, mid_bin),
                            lambda - c(lambda, mle, left_bin, mid_bin)), 5))

colnames(df) <- c("True lambda", "MLE", "Bin left", "Bin center")
rownames(df) <- c("Value (tau = 1)", "Difference from true lambda (tau = 1)",
                  "Value (tau = 5)", "Difference from true lambda (tau = 5)")
```

```
df
```

```
##                                       True lambda     MLE Bin left Bin center
## Value (tau = 1)                               0.2 0.19979  0.22114    0.19913
## Difference from true lambda (tau = 1)         0.0 0.00021 -0.02114    0.00087
## Value (tau = 5)                               0.2 0.19982  0.34317    0.18471
## Difference from true lambda (tau = 5)         0.0 0.00018 -0.14317    0.01529
```

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 $h(y)$.

The PMF is

$$P(Y_i = y) = pq^y = \exp(y\log(q) + \log(1-q))$$

Multiplying and dividing by $h(y) = \frac{1}{2^{y+1}}$ gives

$$\exp(y\log(q) + \log(1-q) + (y+1)\log(2))\frac{1}{2^{y+1}}$$

so we have an NEF with $\theta = \log(2q)$, $\Psi(\theta) = -\log(2 - e^\theta)$, and $h(y) = \frac{1}{2^{y+1}}$.

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

$$E(Y_i) = \Psi'(\theta) = \frac{e^\theta}{2 - e^\theta} = \frac{1}{2e^{-\theta} - 1} = \frac{1}{\frac{1}{q} - 1} = q/(1-q) = q/p$$

$$\text{Var}(Y_i) = \Psi''(\theta) = \frac{2e^{-\theta}}{(2e^{-\theta} - 1)^2} = \frac{1/q}{(1/q - 1)^2} = \frac{q}{(1-q)^2} = q/p^2$$

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

The joint PMF of the data is

$$P(\vec{Y} = \vec{y}) = \prod_{i=1}^{n} pq^{y_i} = p^n q^{n\bar{y}} = g(\bar{y}, p)h(\vec{y})$$

where $g(\bar{y}, p) = p^n q^{n\bar{y}}$ and $h(\vec{y}) = 1$. Note that we don't need to write $h(\vec{y})$ as a PDF or PMF of anything. (What would it be the PDF or PMF of anyway? We're decomposing a PDF or PMF.)

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

$E(Y_1|\bar{Y}) = \frac{1}{n}nE(Y_1|\bar{Y}) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i|\bar{Y}) = \frac{1}{n}E(n\bar{Y}) = \bar{Y}$ where the second equality used the symmetry of i.i.d. draws.

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 $(1/n, ..., 1/n)$.

Using linearity of conditional expectation and the result above,

$$E\left(\sum_{i=1}^{n} w_i Y_i | \bar{Y}\right) = \sum_{i=1}^{n} w_i E\left(Y_i|\bar{Y}\right) = \sum_{i=1}^{n} w_i \bar{Y} = \bar{Y}$$

Since $\sum_{i=1}^{n} w_i Y_i = \bar{Y}$ when $w_i = 1/n$, the best $\vec{w}$ is $(1/n, ..., 1/n)$.

Alternatively, we can show this without linearity: Starting with $\hat{\mu} = \sum_{i=1}^{n} w_i Y_i$, we can apply Rao-Blackwell to get $E(\hat{\mu}|\bar{Y})$. First, note that

$$E(\hat{\mu}|\bar{Y}) = E\left(\sum_{i=1}^{n} w_i Y_i | \sum_{i=1}^{n} Y_i\right)$$

By symmetry, this conditional expectation is the same as if we rotated all the weights to the right: for j in $\{0, ..., n-1\}$,

$$E\left(\sum_{i=1}^{n} w_i Y_i | \sum_{i=1}^{n} Y_i\right) = E\left(\sum_{i=1}^{n-j} w_{i+j} Y_i + \sum_{i=n-j+1}^{n} w_{i+j-n} Y_i | \sum_{i=1}^{n} Y_i\right)$$

Next, since every Y_i is paired with every w_j once and the w_j sum to 1,

$$\sum_{j=0}^{n-1}\left(\sum_{i=1}^{n-j} w_{i+j} Y_i + \sum_{i=n-j+1}^{n} w_{i+j-n} Y_i\right) = \sum_{i=1}^{n}\left(Y_i \sum_{j=1}^{n} w_i\right) = n\bar{Y}$$

Therefore,

$$nE\left(\sum_{i=1}^{n} w_i Y_i | \sum_{i=1}^{n} Y_i\right) = \sum_{j=0}^{n-1} E\left(\sum_{i=1}^{n-j} w_{i+j} Y_i + \sum_{i=n-j+1}^{n} w_{i+j-n} Y_i | \sum_{i=1}^{n} Y_i\right) = n\bar{Y}$$

Thus, $\hat{\mu}_{RB} = E(\hat{\mu}|\bar{Y}) = \bar{Y}$, which means we need w_i such that given $\bar{Y}$, $\sum_{i=1}^{n} w_i Y_i = \bar{Y}$. The only $\vec{w}$ that satisfies this is $(1/n, ..., 1/n)$.