Theory

Baseline Model: No Censoring, No Covariates, Single Sequence

Draw $y_1, y_2, \ldots$ from $y_i \sim \text{Bernoulli}(\pi)$ until $y_i = 1$. This sampling procedure produces a sequence of $n-1$ zeros and a single one, where n is a random variable. In fact, $n \sim \text{geometric}(\pi)$.

Use the sample average $\bar{y} = \frac{1}{n}$ to estimate π. Is $\bar{y}$ an unbiased estimator of π, so that $E(\bar{y}) = \pi$?

Result: $\bar{y}$ is biased upward by a factor of $\frac{-\log(\pi)}{1-\pi}$.

First, notice that, by construction, $\bar{y} = \frac{1}{n}$ (remember that n is a random variable).

But what is the distribution of $\frac{1}{n}$?

$$
\begin{aligned}
P\left(\frac{1}{n} = 1\right) &= \pi \\
P\left(\frac{1}{n} = \frac{1}{2}\right) &= (1-\pi)\pi \\
P\left(\frac{1}{n} = \frac{1}{3}\right) &= (1-\pi)^2\pi \\
&\vdots \\
P\left(\frac{1}{n} = \frac{1}{k}\right) &= (1-\pi)^{k-1}\pi \\
&\vdots
\end{aligned}
$$

Then

$$E\left(\frac{1}{n}\right) = \frac{\pi}{1-\pi} \sum_{i=1}^{\infty} \frac{1}{i}(1-\pi)^i$$

.

The series $\frac{1}{i}(1-\pi)^i$ converges because $(1-\pi) \leq 1 \leq |\frac{1}{i}|^{-\frac{1}{i}}$ for all i. (See radius of convergence for a power series.)

We now need the sum $\sum_{i=1}^{\infty} \frac{1}{i}(1-\pi)^i$. For simplicity, let $q = 1 - \pi$, so that we need the sum $\sum_{i=1}^{\infty} \frac{1}{i} q^i$.

$$
\begin{aligned}
\sum_{i=1}^{\infty} \frac{1}{i} q^i &= \sum_{i=1}^{\infty} \int_0^q x^{(i-1)} dx \text{ (by the F.T.C.)} \\
&= \int_0^q \sum_{i=1}^{\infty} x^{(i-1)} dx \text{ (sum of integrals equals the integral of sums)} \\
&= \int_0^q \frac{1}{1-x} dx \text{ (sum of a geometric series with } -1 < r < 1) \\
&= \int_1^{1-q} \frac{1}{u}(-du) \text{ (integration by substitution; let } u = 1-x,\ du = -dx, \text{ adjust limits)} \\
&= -\log(u)\Big|_1^{1-q} \quad \left(\int \frac{1}{z} dz = \log(z) + C \text{ for } z > 0\right) \\
&= -[\log(1-q) - \log(1)] \text{ (evaluate integral at limits)} \\
&= -\log(1-q) \text{ (simplify)}
\end{aligned}
$$

Substituting this result in, we have $\sum_{i=1}^{\infty} \frac{1}{i}(1-\pi)^i = -\log(\pi)$. Then we have that

$$
E(\bar{y}) = E\left(\frac{1}{n}\right) = \frac{\pi}{1-\pi} \sum_{i=1}^{\infty} \frac{1}{i}(1-\pi)^i = \frac{-\pi \log(\pi)}{1-\pi}
$$

For an unbiased $\hat{y}$, we would have $E(\bar{y}) = \pi$, but we have $E(\bar{y}) = \pi \left[\frac{-\log(\pi)}{1-\pi}\right]$. We can see that $\frac{-\log(\pi)}{1-\pi} = 1$ when $\pi = 1$. Further, we can see that it decreases with π for $0 \leq \pi \leq 1$. This implies that $\bar{y}$ is biased upward by a factor of $\frac{-\log(\pi)}{1-\pi}$.

The simulation below confirms the result.

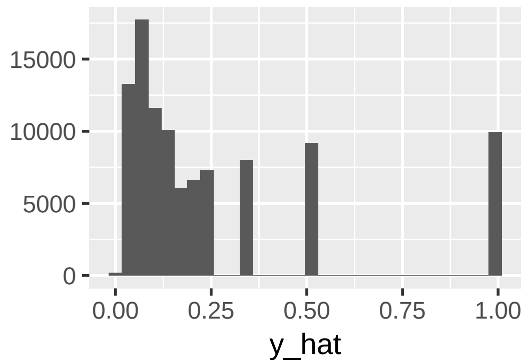
```r
# number of mc simulations
n_sims <- 100000
pi <- 0.1
# draw y from geometric distribution
## note: this is the number of failures (0s)
## before a success
x <- rgeom(n_sims, prob = pi)

# the mean of the series with x 0s and a single 1
y_hat <- 1/(x + 1)

# estimate E(y-hat)
mean(y_hat)
```

```
## [1] 0.2560564
```

```r
# theoretical solution
-pi*log(pi)/(1-pi)
```

```
## [1] 0.2558428
```

```r
# plot sampling distribution
library(ggplot2)
qplot(y_hat)
```



The figure below shows the adjustment factor $\frac{-\log(\pi)}{1-\pi}$.

```
# adjustment factor
adj_fn <- function(p) {
  -log(p)/(1 - p)
}

# plot adjustment factor
library(ggplot2)
ggplot(data.frame(x = c(0.01, 1)), aes(x)) +
  stat_function(fun = adj_fn, geom = "line")
```