

## Announcements

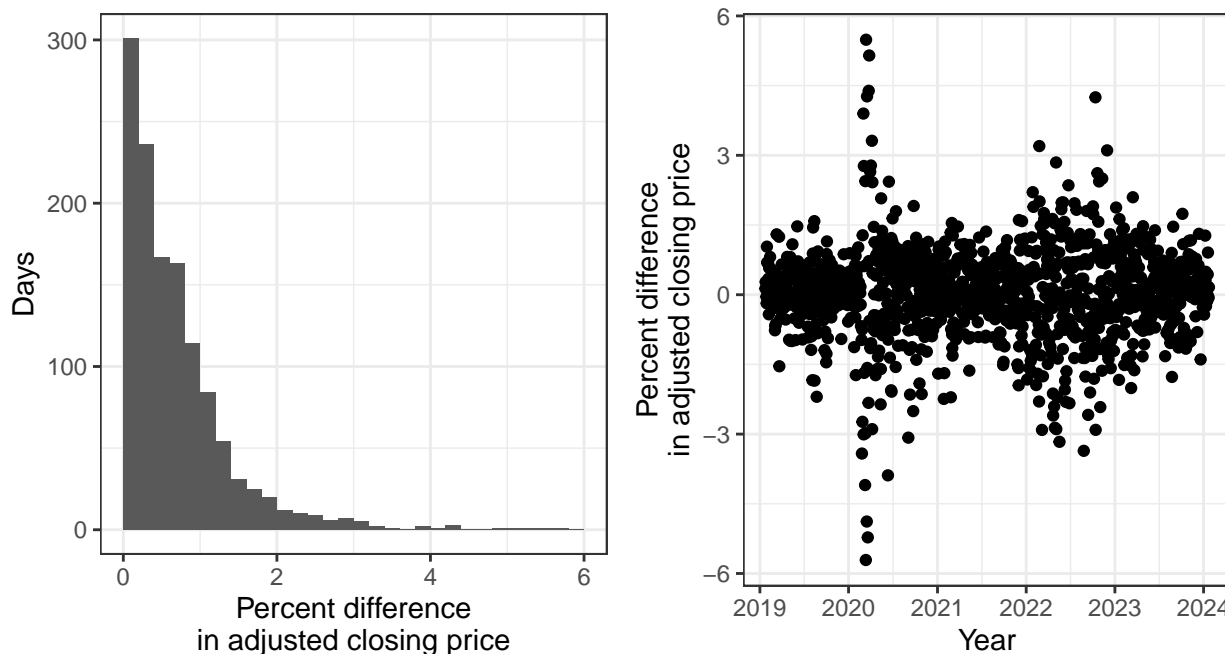
Make sure to sign in on the [google form](#).

Pset 3 due 2/16.



## Stonks

The following questions deal with the past 5 years of S&P 500 adjusted closing prices [available here](#).



In this section, we will be modeling the day-to-day percent differences in the adjusted closing price of the S&P 500 as  $Y_1, \dots, Y_n \sim \text{Expo}(\lambda)$ .

1. Find the score of  $\lambda$  (the score is  $\frac{\partial}{\partial \lambda} \ell(\lambda; \vec{y})$ ) in terms of the sample mean and verify that  $E(s(\lambda^*; \vec{Y})) = 0$ . (This equality was the last part of Neil's Thursday lecture.)

The likelihood function is

$$L(\lambda; \vec{y}) = \prod_{i=1}^n \lambda e^{-\lambda y_i} = \lambda^n e^{-\lambda \sum_{i=1}^n y_i}$$

The log likelihood function is

$$\ell(\lambda; \vec{y}) = n \log(\lambda) - \lambda \sum_{i=1}^n y_i$$

The score is

$$s(\lambda; \vec{y}) = n/\lambda - n\bar{y} \implies E(s(\lambda^*; \vec{Y})) = n/\lambda^* - nE(\bar{Y}) = 0$$

2. Show that the MLE of  $\hat{\lambda}$  is consistent for  $\lambda$ . That is, show that  $\hat{\lambda} \rightarrow \lambda$  as  $n \rightarrow \infty$  by showing the MSE goes to 0, a LLN holds, making a claim using the CMT, or showing convergence directly.

Setting the score to 0 gives  $\hat{\lambda} = 1/\bar{y}$ . We will show consistency with the continuous mapping theorem. Because  $\bar{Y}_n \rightarrow 1/\lambda$  by the LLN,  $\hat{\lambda} = 1/\bar{Y}_n \rightarrow \lambda$  by the CMT since  $1/x$  is a continuous function.

3. In his book *The Black Swan*, Nassim Taleb argues that part of the reason for the 2008 financial crisis was a failure to model market fluctuations and assign sufficient probability to extreme events. Let us consider daily absolute differences above  $\tau$  to be extreme events. Let  $X_i = I(Y_i > \tau)$ . Show that  $\bar{X}$  is consistent for  $p = P(Y_i > \tau)$  first by using MSE and then by using the law of large numbers.

Since the  $Y_i$  are i.i.d., by the story of the binomial  $n\bar{X} \sim \text{Bin}(n, p)$ . Thus,  $E(\bar{X}) = p$  and  $\text{Var}(\bar{X}) = p(1-p)/n \rightarrow 0$  as  $n \rightarrow \infty$ . Since  $\text{MSE}(\bar{X}, p) = \text{Bias}(\bar{X})^2 + \text{Var}(\bar{X}) \rightarrow 0$ ,  $\bar{X}$  is a consistent estimator. This can be seen more easily from the fact that  $\bar{X} \rightarrow E(X_i) = E(I(Y_i > \tau)) = p$  by the law of large numbers and fundamental bridge.

4. Find the asymptotic distribution of  $\bar{X}$  and its approximate distribution for large  $n$  in terms of  $\lambda$ .

Using the Exponential CDF,  $p = P(Y_i > \tau) = e^{-\lambda\tau}$ . By the CLT,

$$\frac{\sqrt{n}(\bar{X} - e^{-\lambda\tau})}{\sqrt{e^{-\lambda\tau}(1 - e^{-\lambda\tau})}} \rightarrow \mathcal{N}(0, 1) \iff \bar{X} \sim \mathcal{N}(e^{-\lambda\tau}, \frac{e^{-\lambda\tau}(1 - e^{-\lambda\tau})}{n})$$

5. Now, suppose we estimate  $P(Y_i > \tau)$  with  $\hat{p} = e^{-\hat{\lambda}\tau}$ . Given that

$$\sqrt{n}(\hat{\lambda} - \lambda^*) \rightarrow \mathcal{N}(0, \lambda^{*2})$$

(we'll see how to find this in future weeks), find the asymptotic distribution of  $\hat{p}$  and its approximate distribution for large  $n$ .

We will use the Delta Method with  $g(\lambda) = e^{-\lambda\tau}$ . Differentiating gives  $g'(\lambda) = -\tau e^{-\lambda\tau}$ . Using the asymptotic distribution and the Delta Method,

$$\sqrt{n}(\hat{p} - e^{-\lambda^*\tau}) \rightarrow \mathcal{N}(0, \lambda^{*2}) \implies \frac{\sqrt{n}(\hat{p} - e^{-\lambda^*\tau})}{|-\tau e^{-\lambda^*\tau}|} \rightarrow \mathcal{N}(0, \lambda^{*2})$$

This gives the approximate distribution

$$\hat{p} \sim \mathcal{N}\left(e^{-\lambda^*\tau}, \frac{\tau^2 e^{-2\lambda^*\tau} \lambda^{*2}}{n}\right)$$

6. The distributions in 4 and 5 should have the same mean. However, the variances are different. The following are the estimated standard errors of each estimator with  $\tau = 5$ . Explain why the results are what they are.

```
## Xbar SE phat SE
## 0.000759 0.000004
```

We've made a distributional assumption about  $Y$ , so observing values even if they aren't over the threshold provides information, lowering the standard error of the estimate.

7. Though  $\hat{p}$  is more efficient, it might be less robust. The following shows the MSEs of the two estimators for estimating  $P(Y > \tau)$  when  $Y \sim \text{Expo}(0.5)$  (the correct model) and  $Y \sim \text{Log-Normal}(0.1, 1)$  (an incorrect model). Again, we have used  $\tau = 5$  and  $n = 100$  with  $10^5$  simulations.

```
##           Xbar      phat
## Log Normal 0.00061 0.00054
## Expo       0.00075 0.00041
```

It turns out that  $\hat{p}$  has lower MSE on both the correct model and a misspecified model, so it is actually both more efficient and more robust. However, this is not always the case!

## Ty Mup

Ty Mup is taking an exam with  $n$  equally hard questions. He has a probability  $p_2$  of getting each question right independently. However, there is also a  $0 < p_1 < 1$  probability he sleeps through his alarm and misses the exam entirely. Let  $Y$  be the number of questions he gets right on his exam. (This distribution is called the zero-inflated binomial.)

1. Find  $E(Y|Y > 0)$ .

If  $Y > 0$ , he must have made it to the exam on time. Call this event  $A$ . From the problem description,  $Y|A \sim \text{Bin}(n, p_2)$ . Then, by the law of total expectation,

$$E(Y|A) = E(Y|A, Y > 0)P(Y > 0|A) + E(Y|A, Y = 0)P(Y = 0|A)$$

Since  $Y > 0 \implies A$ ,

$$E(Y|Y > 0) = E(Y|A, Y > 0) = \frac{E(Y|A)}{P(Y > 0|A)} = \frac{np_2}{1 - (1 - p_2)^n}$$

from the binomial PMF. This is slightly above  $np_2$  as expected.

2. Unfortunately, [the day is February 2nd in Punxsutawney](#) and Ty is destined to repeat this day  $d$  times, scoring i.i.d  $Y_i$  on the exams. Find the likelihood function, the log-likelihood function, and the score for  $p_1$  and  $p_2$ . (Hint: Let  $m$  be the number of 0s.)

The likelihood can be written as the probability of observing the 0 values times the probability of observing everything else:

$$L(p_1, p_2; \vec{y}) = (p_1 + (1 - p_1)(1 - p_2)^n)^m \prod_{i=1, y_i \neq 0}^d (1 - p_1)p_2^{y_i}(1 - p_2)^{n-y_i}$$

Taking the log and using the fact that  $\sum_{i=1, y_i \neq 0}^d y_i = d\bar{y}$  (where the mean is over all  $y_i$ , not just the positive ones),

$$\begin{aligned} \ell(p_1, p_2; \vec{y}) &= m \log(p_1 + (1 - p_1)(1 - p_2)^n) + \sum_{i=1, y_i \neq 0}^d \log(1 - p_1) + y_i \log(p_2) + (n - y_i) \log(1 - p_2) \\ &= m \log(p_1 + (1 - p_1)(1 - p_2)^n) + (d - m) \log(1 - p_1) + d\bar{y} \log(p_2) + \log(1 - p_2)(n(d - m) - d\bar{y}) \end{aligned}$$

$$\begin{aligned} s(p_1; \vec{y}) &= \frac{\partial}{\partial p_1} \ell(p_1, p_2; \vec{y}) = \frac{m(1 - (1 - p_2)^n)}{p_1 + (1 - p_1)(1 - p_2)^n} - \frac{d - m}{1 - p_1} \\ s(p_2; \vec{y}) &= \frac{\partial}{\partial p_2} \ell(p_1, p_2; \vec{y}) = -\frac{nm(1 - p_1)(1 - p_2)^{n-1}}{p_1 + (1 - p_1)(1 - p_2)^n} + \frac{d\bar{y}}{p_2} - \frac{n(d - m) - d\bar{y}}{1 - p_2} \end{aligned}$$

3. Find a two-dimensional sufficient statistic (a two dimensional statistic that contains all the information about the likelihood).

$$(M, \bar{Y})$$

4. Let  $B$  be the event that Ty sleeps through the alarm at least once. Show that as  $d \rightarrow \infty$ ,

$$I_B \frac{d^{1/2}(\bar{Y} - (1 - p_1)np_2)}{\sqrt{np_2(1 - p_2)(1 - p_1) + (np_2)^2 p_1(1 - p_1)}} \xrightarrow{d} \mathcal{N}(0, 1)$$

Let  $A$  be the event Ty made it to the exam on time. By Law of Total Expectation

$$E(\bar{Y}) = E(Y) = E(Y|A)P(A) + E(Y|A^c)P(A^c) = np_2(1 - p_1)$$

Likewise, by Eve's law,

$$\text{Var}(Y) = E(\text{Var}(Y|I_A)) + \text{Var}(E(Y|I_A)) = E(np_2(1-p_2)I_A) + \text{Var}(np_2I_A) = np_2(1-p_2)(1-p_1) + (np_2)^2p_1(1-p_1)$$

Then,  $\text{Var}(\bar{Y}) = \text{Var}(Y)/d$ . By the CLT,  $\frac{d^{1/2}(\bar{Y} - (1-p_1)dp_2)}{\sqrt{np_2(1-p_2)(1-p_1) + (np_2)^2p_1(1-p_1)}} \rightarrow Z$  with  $Z \sim \mathcal{N}(0, 1)$ . Also,  $P(|I_B - 1| > \epsilon)$  is 0 if  $\epsilon \geq 1$  and it is  $(1-p_1)^d$  if  $0 < \epsilon < 1$ . Since  $p_1 < 1$ ,  $P(|I_B - 1| > \epsilon) \rightarrow 0$ , so  $I_B \xrightarrow{p} 1$ . Thus, by Slutsky's theorem,

$$I_B \frac{d^{1/2}(\bar{Y} - (1-p_1)dp_2)}{\sqrt{np_2(1-p_2)(1-p_1) + (np_2)^2p_1(1-p_1)}} \xrightarrow{d} 1 \cdot Z \sim \mathcal{N}(0, 1)$$