ECO 394D: Probability and Statistics

Homework 4

1. Suppose that $X_1, \ldots, X_N \sim$ Bernoulli(p) and that $Y_1, \ldots, Y_M \sim$ Bernoulli(q) (all independent). We will consider $\hat{p} = \bar{X}_N$ and $\hat{q} = \bar{Y}_M$ as estimators of p and q, respectively. We have shown already that $E(\hat{p}) = p$. This implies by extension that $E(\hat{q}) = q$ and therefore that $E(\hat{p} - \hat{q}) = p - q$, the true difference in success probabilities.

   (A) Find the standard error of $\hat{p}$.

   (B) Compute the standard error of $\hat{p} - \hat{q}$ as an estimator of the difference $p - q$.

   (C) Now return to the data from class on the Predimed study, in `predimed.csv`. The trial has three groups: Control, Med Diet + Nuts, and Med Diet + VOO (virgin olive oil). For this problem, please first combine the two Med Diet groups into a single "Med Diet Any" group.

   Now let p be the probability of a cardiac event (event = yes) for those on the control diet, and let q be the probability of a cardiac event for those on any Mediterranean diet (whether combined with nuts or VOO). Use the data, together with your derivations above, to calculate the plug-in standard error for the difference in proportions, $\hat{p} - \hat{q}$.

   (D) Use this to find the 95% normal-based confidence interval for $p - q$ for the predimed data, and compare it with the confidence interval you get using the bootstrap.

   (E) Suppose our null hypothesis is that $\Delta = p - q = 0$: that there is no average difference in the rate of cardiac events between control and Med diet groups. You use the test statistic $\hat{\Delta} = \hat{p} - \hat{q}$ to measure the evidence in your sample against the null hypothesis. Use the assumption that $\hat{\Delta}$ is asymptotically normal to derive the (one-sided) p-value of $\hat{\Delta}$ under the null hypothesis.

2. Suppose we have data on some numerical attribute from two groups: $X_1, \ldots, X_N$ from group 1, and $Y_1, \ldots, Y_M$ from group 2. (Notice the unequal sample sizes N and M.) Suppose that:

   - $E(X_i) = \mu_X$ and $\text{var}(X_i) = \sigma_X^2$, both unknown
   - $E(Y_i) = \mu_Y$ and $\text{var}(Y_i) = \sigma_Y^2$, again both unknown

   Suppose we're interested in the population-level difference in means between the two groups: $\Delta = \mu_X - \mu_Y$. We use the difference in sample means to estimate this quantity:

$$\hat{\Delta} = \bar{X}_N - \bar{Y}_M$$

   (A) Use your knowledge of probability theory to calculate the true (theoretical) standard error of $\hat{\Delta}$. Note: this is step 1 in the process for calculating a plug-in standard error.

   (B) Now download GasPrices.csv data set from the class website. This data set came from a student project in my data-mining class. You'll see that the data set contains many variables, but we'll focus on only two:

   - Price: Price of regular unleaded gas in dollars per gallon
   - Highway: Is the gas station accessible from either a highway or a highway access road?

   Suppose that group 1 is gas stations accessible from a highway, while group 2 is gas stations *not* accessible from a highway. We're interested in Δ, the difference in average price between these two groups, which represents the "highway premium" customers are paying for the privilege of filling up their gas tanks right off the highway.

   Use the data, together with your derivations in part A, to calculate a 95% normal-based confidence interval for Δ. Compare it to the confidence interval you get from bootstrapping the data.

   (C) Under the assumption that $\hat{\Delta}$ is asymptotically normal, calculate a p-value for your observed difference $\hat{\Delta}$ under the null hypothesis that the true difference is $\Delta = 0$.

3. Supose that $X_1, \ldots, X_N$ are independent samples from a Poisson distribution with rate parameter λ, which has PMF

$$P(X = x) = f(x; \lambda) = \frac{\lambda^x}{x!} e^{-\lambda}$$

(A) Show that $E(X_i) = \lambda$. Hint: the exponential function e^y is can be expressed as an infinite series:

$$e^y = \sum_{k=0}^{\infty} \frac{y^k}{k!} .$$

Conclude that $E(\bar{X}_N) = \lambda$, i.e. that the sample mean is an unbiased estimator of the Poisson rate parameter.

(B) Using a similar argument, one can show that $\text{var}(X_i) = \lambda$. You don't need to show this. Rather, use fact this to calculate the theoretical standard error of $\hat{\lambda} = \bar{X}_N$.

(C) Show that $\bar{X}_N$ is also the maximum-likelihood estimate for the rate parameter of the Poisson distribution.

(D) You may have seen the "Popular Times" feature on Google Maps, which gives you an estimate of how busy a restaurant is over the course of the day. For example, here's the Google Maps entry for Odd Duck, a restaurant in Austin: https://goo.gl/maps/QAw6mkZUYZG2. If you scroll down on the panel having information about the restaurant, you'll see a bar chart showing when it's typically busy. This is estimated from the locations of people using Google services on their smartphones.

Suppose that you're working behind the scenes at Google and trying to build the algorithm that displays an estimate of how busy a restaurant is. A raw observation for a particular restaurant is of the form $X_{d,h,i}$: the number of unique people (X) using Google services on day d at hour h, observed repeatedly over multiple weeks $i = 1, \ldots, N$. You decide that a reasonable assumption is that $X_{d,h,i} \sim \text{Poisson}(\lambda_{d,h})$. In other words, each day/hour combination has its own average rate, and each week's observation $X_{d,h,i}$ for that time slot is an independent sample from a Poisson with rate $\lambda_{d,h}$.

Obviously Google must estimate 7×24 rate parameters for every hour of every day. But for now, let's take a particular time period: say, d=Friday, and h = from 5 PM to 6 PM. Suppose you observed 30 different counts $X_{d,h,i}$ over 30 weeks for this time period. These counts were:

```
12 18 15  8 17 13 22 13 13 13 12 11 15 15 12
```

```
 8 20 12 14 11  9 15 16 20  9 15 13 19 18 14
```

Use the results of Part B to construct a 95% normal-based confidence interval for $\lambda_{d,h}$. Compare it to the 95% confidence interval you get from bootstrapping the data.

Note: this is a pretty simple model for demand at a business, since it ignores all sorts of stuff like seasonal variation, whether the Formula 1 race is happening that weekend, etc. But it's a decent starting point.

4. Let $X_1, \ldots, X_n$ be independent and identically distributed (IID) samples from some distribution whose CDF is $F(x)$. Since $F(x) = P(X_i \leq x)$, a natural estimate of $F(x)$ is the **empirical CDF** $\hat{F}_n$, defined as follows:

$$\begin{aligned}\hat{F}_n(x) &= \frac{\text{Number of } X_i \text{ that are less than or equal to } x}{n} \\ &= \frac{\sum_{i=1}^{n} I(X_i \leq x)}{n},\end{aligned}$$

where $I(X_i \leq x)$ is the indicator function that takes the value 1 if $X_i \leq x$ and the value 0 otherwise.

(A) For any fixed x, the empirical CDF $\hat{F}_n(x)$ is a random variable. What are its mean and variance? Note: this was a midterm practice question.

(B) Recall that a sequence of random variables $Y_1, Y_2, \ldots$ converges in probability to some other random variable Y if for every $\epsilon > 0$,

$$P(|Y_n - Y| > \epsilon) \longrightarrow 0 \quad \text{as } n \longrightarrow \infty\,.$$

Use your result in Part A, together with your knowledge of asymptotic theory, to show that for any fixed x, the empirical CDF $\hat{F}_n(x)$ converges in probability to the true CDF $F(x)$.

(C) Go to the `stocks_bonds.csv` data on the class website, which gives annual returns on the S&P 500 going back to 1928. Let X be the return on the S&P 500 in a random year, and let $F(x)$ be the unknown CDF describing the probability distribution of X. Use this data, together with your derivations above, to construct a 95% normal-based confidence interval for $F(-0.1)$: that is, the probability that the S&P 500 will lose 10% or more in a given year.