Statistical Inference

B. Statistical Data Science 2nd Year Indian Statistical Institute

Teacher: Soham Sarkar

Exercise Series 1 (Solutions)

Scenario 1. This is a classic case of Bernoulli modeling. The items can be either *defective* or *non-defective*, giving only two possible (or binary) outcomes. For a randomly selected item, let X be the random variable that takes the value 1 if the item is defective and 0 if it is non-defective. Then $X \sim \text{Bernoulli}(p)$, where $p = \mathbb{P}(X = 1) = \mathbb{P}(\text{an item is defective})$. We can assume that the status of the items are independent, meaning that whether an item is defective or not does not affect the status of the other items (and vice verse). Then, we have i.i.d. observations $X_1, \ldots, X_{50} \sim \text{Bernoulli}(p)$. The unknown parameter of interest is p. The parameter space is $[0, 1]$ (or $(0, 1)$ if you want to avoid the trivial cases).

The company would halt production if $p > 0.1$. Therefore, we are facing a testing problem $\mathcal{H}_0 : p \leq 0.1$ vs. $\mathcal{H}_a : p > 0.1$ based on $X_1, \ldots, X_{50} \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$.

Scenario 2. In this case, we are not merely observing a specific number of Bernoulli random variables. What we observe here is "the number of trials required to get 5 successes". If we define X to be this number, then $X - 5$ is the number of "failures" before 5 successes occur. Therefore $X - 5 \sim \text{NegativeBinomial}(5, p)$, with pmf

$$\mathbb{P}(X - 5 = t) = \binom{t+4}{4} p^5 (1-p)^t, \qquad t = 0, 1, \ldots.$$

Alternatively, we can find the pmf of X directly as

$$\mathbb{P}(X = x) = \binom{x-1}{4} p^5 (1-p)^{x-5}, \qquad x = 5, 6, \ldots.$$

The parameter of the model is $p \in [0, 1]$. Our goal is to test $\mathcal{H}_0 : p \leq 0.1$ vs. $\mathcal{H}_a : p > 0.1$ based on X.

Scenario 3. In this case, we have the counts (numbers) of accidents on a day. This number can be $0, 1, 2, \ldots$, without any known bound. Therefore, a Poisson model seems appropriate here. We can assume that the number of accidents per day over the previous month are $X_1, \ldots, X_{30} \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda)$, with $\lambda > 0$ being the unknown parameter. The parameter space is $(0, \infty)$.

Note: The i.i.d. assumption here may be difficult to fathom, especially after experiencing the roads during monsoon. If a road is in a poor condition today, and if it is not repaired, then the chances of accidents remain high. Once it gets repaired, the chances are lowered. However, while the condition (poor road/repaired) persists, the observations can be thought of as i.i.d., unless we have some further information (e.g., traffic congestion due to festivities). Since we do not have any such information, it is reasonable to assume the observations to be i.i.d..

With $X \sim \text{Poisson}(\lambda)$, $\mathbb{P}(\text{at least one accident}) = \mathbb{P}(X \geq 1) = 1 - \mathbb{P}(X = 0) = 1 - e^{-\lambda}$. A traffic control station will be positioned if this probability is more than 50%. That is, $1 - e^{-\lambda} > 0.5 \Leftrightarrow \lambda > \ln(2)$. So, our goal here is to test $\mathcal{H}_0 : \lambda \leq \ln(2)$ vs. $\mathcal{H}_a : \lambda > \ln(2)$ based on i.i.d. observations $X_1, \ldots, X_{30} \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda)$.

Scenario 4. Here we have two possible outcomes (binary outcomes) *cured* and *not cured*. We may assume that $X_1, \ldots, X_{10000} \overset{\text{i.i.d.}}{\sim} \mathsf{Bernoulli}(p)$ is our random sample, where $X_i = 1$ if the i-th patient was cured and 0 otherwise. The unknown parameter is $p = \mathbb{P}(X = 1) = \mathbb{P}(\text{a patient is cured})$. So, the efficacy of the new vaccine would be $100p\%$. The parameter space here is $[0, 1]$.

The government may start using the new vaccine if it has more efficacy than the previous one (which has 75% efficacy). So, we are trying to check whether $\mathcal{H}_0 : p \leq 0.75$ vs. $\mathcal{H}_a : p > 0.75$. It is also possible that merely being better than the previous one is not good enough for the government to change its course of action. In such situations, it may be more appropriate to construct a confidence interval for p. For example, if a 95% confidence interval for p has the form $(0.82 - 0.91)$, that may be more convincing for the government to make a decision.

Note: The above is described in a very simplistic way. Vaccine trials are not conducted such simplistically. There are several phases of these trials, and some very important questions on ethics are related to it. These are part of a topic called "clinical trials".

Scenario 5. Let X be the household income of a randomly selected household from the district. Our observations are $X_1, \ldots, X_{500}$, which we can assume to be i.i.d. with the same distribution as X. We need to find a suitable statistical model for X. Keep in mind:

- Household income cannot be negative.
- Household income can be any positive number (we don't have a known upper bound).

Therefore, we should look for a continuous distribution, possibly supported on $(0, \infty)$. There are several choices for this, e.g., exponential, truncated/folded normal. It is well-established that income distributions are positively skewed (based on empirical evidence). Therefore a right-skewed distribution supported on $(0, \infty)$ is more appropriate here, e.g., lognormal distribution or noncentral χ^2 distribution. The unknown parameters depend on the choice of distribution. For lognormal, it would be (μ, σ^2). For noncentral χ^2, it would be the "non-centrality parameter" (ncp) δ and "degrees of freedom" (df) ν.

Suppose that our chosen model is $X_1, \ldots, X_{500} \overset{\text{i.i.d.}}{\sim} f_\theta$, where $\theta \in \Theta$. The income profile consists of the income threshold t with $100q\%$ households below income t. That is, t satisfies $\mathbb{P}(X \leq t) = q$. So, $t = t_q$ should be the q-th quantile of the distribution f_θ. Therefore, our goal here is to estimate the q-quantiles of the distribution for $q = 0.1, 0.2, \ldots, 0.9$.

Scenario 6. Here also, we notice that:

- The lifetime of an LED buld is non-negative.
- The lifetime can be any positive value (no known upper bound).

Therefore, we can assume $X \sim f_\theta$ for some continuous distribution supported on $(0, \infty)$. We know the lifetimes of 1000 bulbs. So, our observations are $X_1, \ldots, X_{1000} \overset{\text{i.i.d.}}{\sim} f_\theta$. Traditionally, for this type of lifetime modeling, the Exponential distribution is used.

With $\mathsf{Exponential}(\text{mean} = \lambda)$, for any fixed $T > 0$, $\mathbb{P}(X < T) = 1 - e^{-T/\lambda}$. In one year, there are 365 days. We can assume that a bulb runs for 8 hours on average every day, giving a total of 2920 working hours per year. A replacement needs to be issued if the bulb stops working before this time. So, the probability of issuing a replacement is $\mathbb{P}(X < 2920) = 1 - e^{-2920/\lambda}$. The company wants to check whether this number is more than 0.05, equivalently $\lambda < -2920/\ln(0.95)$. This is a hypothesis testing problem between $\mathcal{H}_0 : \lambda \geq -2920/\ln(0.95)$ and $\mathcal{H}_a : \lambda < -2920/\ln(0.95)$.

Note: The problem would change if a different statistical model (distribution) is used. In any case, it remains a hypothesis testing problem. In this scenario, it may also be interesting to provide an estimate (possibly interval) for the probability of issuing a replacement (i.e., an estimate for $1 - e^{-2920/\lambda}$).

Scenario 7. Let X and Y denote the longivities (in kms) for the tires of Brand A and Brand B, respectively. We have observations $X_1, \ldots, X_{75}$ and $Y_1, \ldots, Y_{60}$. We want to formulate a statistical model (suitable distributions) for X and Y. As before, we need distributions which are continuous and positively supported. Interestingly, for this type of data, where we have no reason to believe that the underlying distributions are skewed, a normal distribution works just fine. After all, for $X \sim \mathsf{N}(\mu, \sigma^2)$, $\mathbb{P}(X \in [\mu - 3\sigma, \mu + 3\sigma]) > 0.997$. So, if μ is a large positive number and σ is relatively small compared to μ, then the corresponding Normal distribution will be highly concentrated on $(0, \infty)$. Therefore, we can assume $X_1, \ldots, X_{75} \overset{\text{i.i.d.}}{\sim} \mathsf{N}(\mu_1, \sigma_1^2)$ and $Y_1, \ldots, Y_{60} \overset{\text{i.i.d.}}{\sim} \mathsf{N}(\mu_2, \sigma_2^2)$ and the two samples are independent (i.e., X and Y are independent). Our goal here is to test which of μ_1 and μ_2 is larger. We can formulate other testing problems for the same question. E.g., whether X is *stochastically larger than* Y, i.e., $\mathbb{P}(X > t) \geq \mathbb{P}(Y > t) \forall t \in \mathbb{R}$. If σ_1^2 and σ_2^2 are assumed to be different, then this will never be the case under the normal model. But, if $\sigma_1^2 = \sigma_2^2$, then this is same as testing $\mu_1 \geq \mu_2$.

Scenario 8. Let the weights at the beginning of the diet be X (in kg) and the weight after six months into the diet be Y (in kg). We are interested in $Z = X - Y$, the change in the weight. The change can be any real value (positive or negative), so we use a continuous distribution supported on $\mathbb{R}$. A convenient choice is $Z \sim \mathsf{N}(\mu, \sigma^2)$. We have observations $(X_1, Y_1), \ldots, (X_{100}, Y_{100})$, from which we can get $Z_1, \ldots, Z_{100}$. The dietitian would recommend the diet if $\mu > 0$. Thus our goal here is to test $\mathcal{H}_0 : \mu \leq 0$ vs. $\mathcal{H}_a : \mu > 0$ based on $Z_1, \ldots, Z_{100} \overset{\text{i.i.d.}}{\sim} \mathsf{N}(\mu, \sigma^2)$. Similar to Scenario 4, we may construct a confidence interval for μ, which can be more informative for the dietitian.

Note: The data here corresponds to *paired samples*, as opposed to independent samples in the previous scenario. The longevity of different tires from two different brands can be assumed to be independent. But, the weight of the same person before and after the diet cannot be assumed independent. Therefore, in this example, X and Y are not independent, as opposed to the previous scenario. In this scenario, we could have assumed $(X, Y) \sim \mathsf{BVN}(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$, with five unknown parameters. But, even then $Z = X - Y$ would follow a normal distribution. For testing purpose, it is enough to estimate the two parameters of the later distribution. So, for our end goal, we already do a transformation of the data before analyses, which reduces the number of unknown parameters.

Scenario 9. We have count data, so a Poisson model seems appropriate. We know the number of calls per 60-seconds for the past 30 days. Therefore, we have $43,200$ observations $X_1, \ldots, X_{43200} \overset{\text{i.i.d.}}{\sim}$ $\mathsf{Poisson}(\lambda)$, $\lambda > 0$, where X_i denotes the number of calls in a 60-second interval. It seems reasonable to believe that the number of calls in a second is $X/60$. Essentially, if at any second the number of calls is X, and the capacity (the threshold of calls) of the tower is C, then $(X - C)_+$ calls are dropped. Here, $(t)_+ = t$ if $t > 0$ and 0 otherwise. The company wants to ensure that $(X - C)/X \leq 0.001$ with high probability. That is $1 - C/X \leq 0.001$ with high probability. There are several ways to find C here. For instance, we can replace X by $\mathbb{E}(X) = \lambda$ and get C by solving $1 - C/\lambda \leq 0.001$, giving $C \geq 0.999\lambda$. This, however, is a crude approximate. Alternatively, we can try to find C such that $\mathbb{P}(1 - C/X \leq 0.001) \geq 0.99$. That is $\mathbb{P}(X \leq C/0.999) \geq 0.99$. Therefore, $C/0.999$ must be larger than or equal to the 0.99-quantile of X. This quantile can be obtained after estimating λ from the observed data.

Note: During the whole day, there would be several patches of time with no calls (e.g., during sleep hours). Thus, the data may have an excessive number of zero calls, which may affect the final estimate. In such cases, it may be more appropriate to use a *zero-inflated* model, which takes into account the excessive occurrence of zeros.

Scenario 10. Let N be the total number of tanks. Our goal is to find (estimate) N. Suppose that we spot n tanks, which are numbered $T_1, T_2, \ldots, T_n$. Since the tanks are similar, except for their

numbers, we can assume that spotting one of them is as likely as spotting any other. Also, spotting one of these tanks has nothing to do with the spotting of another. Moreover, it is quite possible that we spot the same tank twice. With these assumptions, we have $T_1, \ldots, T_n \overset{\text{i.i.d.}}{\sim} \mathsf{Uniform}\{1, 2, \ldots, N\}$. Our goal here is to estimate N based on this sample.

Note: If we spot the same tank twice, then there would be dependence in the sample. However, for the sake of simplification, we can assume that the observations are i.i.d. Alternatively, we could model $T_1, \ldots, T_n$ as a *simple random sample with replacement* from $\{1, 2, \ldots, N\}$. A different approach would be to only consider the distinct numbers $\widetilde{T}_1, \ldots, \widetilde{T}_{\widetilde{n}}$ and assume that they form a *simple random sample without replacement* from $\{1, 2, \ldots, N\}$. In both the cases, the exact distribution of the sample (involving the unknown parameter N) can be derived.

Scenario 11. Let the number of elephants in the jungle be N. After our first capture, there would be 10 elephants which are tagged and $N - 10$ which are untagged. Let the random variable X denote the number of tagged elephants out of the 15 captured in the second stage. Then, X would follow a $\mathsf{Hypergeometric}$ distribution with pmf

$$\mathbb{P}(X = x) = \frac{\binom{10}{x}\binom{N-10}{15-x}}{\binom{N}{15}}, \qquad x = 0, 1, \ldots, 10.$$

Our goal here is to estimate N based on X.

Note: One simple estimate can be obtained by equating $10/N$ with $X/15$ (think why!). This would be the method-of-moments estimator in our model.

Scenario 12. Let X be the (random) time taken in hours for the team to complete a project. The number of projects that the team can complete would be $\lfloor 176/X \rfloor$ (assuming 22 working days in a month with 8 hours per day). We have observations $X_1, \ldots, X_5 \overset{\text{i.i.d.}}{\sim} X$. For X, we should use a distribution which is continuous and supported on $(0, \infty)$. Here, we are looking at time to completion, so the $\mathsf{Exponential}$ distribution is a natural choice.

To decide the number of projects to be assigned, we may use $\lfloor 176/\mathbb{E}(X) \rfloor$. From the CEO's perspective (in terms of the company's reputation), it may be more appropriate to find C such that $\mathbb{P}(CX \leq 176) \geq 0.99$. From this, C can be obtained in terms of the quantile of X.