STA 250: Statistics

Notes 2. Statistical Models and Inference Formulation

Book chapters: 7.1

1 Describing Data & Statistical Models

A physicist has a theory that makes a precise prediction of what's to be observed in data. If the data doesn't match the prediction, then the theory is "falsified". A statistician only has an imprecise description. This could be either because the theory is imprecise, or because random errors are introduced in collecting the data, or a combination of the two.

Therefore a statistician's data, from the perspective of her theory + data collection method, is an "uncertain" quantity X. Any uncertain quantity can be best described by a set of values S the quantity may assume, with a pdf/pmf $f(x)$ on S. The pdf/pmf is to be interpreted as follows: $f(x_1)/f(x_2) = r$ means that $X = x_1$ is r-times as plausible as $X = x_2$.

If the data can be described by a single pmf/pdf then there is no need of statistical analysis. Statistics is needed when a multitude of competing theories lead to a multitude of pmfs/pdfs. When all these pmfs/pdfs are collected together, we have a **statistical model** for our analysis. If θ denotes the quantity by which the constituent pmfs/pdfs of the model differ from each other, then we can write each pmf/pdf as $f(x|\theta)$. The quantity θ is a "parameter" of this model. The set Θ of all possible values of θ is called the parameter space of the model.

Example (Opinion Poll)**.** Take for example a study where one wants to know what percentage of students in a certain university are in favor of a recent government policy. For a large university, soliciting every student's opinion is impossible. The researcher may want to draw a random list of $n = 500$ students and quiz them on their opinion regarding the policy. A random list gives the best chance of guarding against systematic biases in obtaining a representative sample of students.

The data here is the number X of students in the sample who are in favor. If the researcher thinks that a fraction p of the students, among a total of N university students are in favor of the policy, then X can be described as hyper-geometric pmf $f(x|p)$ given by

$$f(x|p) = \begin{cases} \frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}} & \text{for } x = 0, 1, 2, \cdots, \min(n, m) \\ 0 & \text{otherwise} \end{cases}$$

where $m = Np$ is the total number of students in the university who are in favor of the policy. The fraction p represents the researcher's theory about the popularity of the policy among college students. If she considers all possibilities $0 \leq p \leq 1$, then here statistical model for X is $\{f(x|p) : p \in [0, 1]\}$ with $f(x|p)$ given as above.

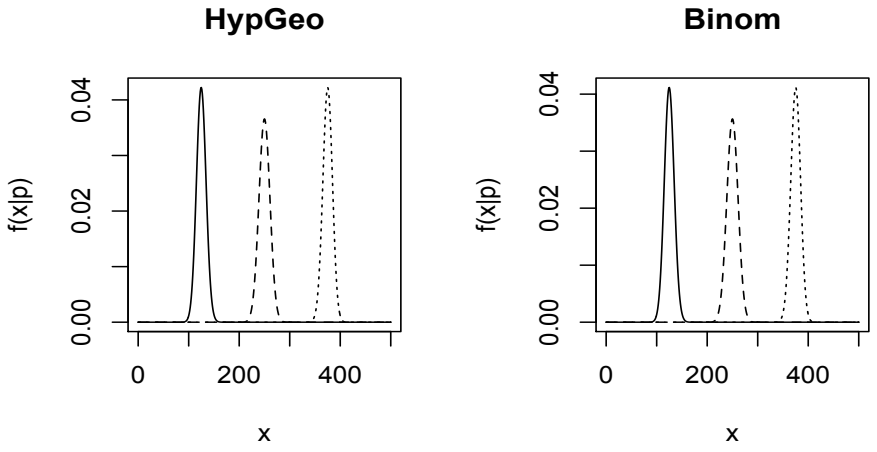

Figure 1: X = number of students favoring the policy in a sample of 500 students. Description of X under hypergeometric (left) and binomial distributions (right) for three possible values of $p = 0.25, 0.5, 0.75$.

When N is very large compared to n, we can also represent X by the binomial pmf

$$f(x|p) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{for } x = 0, 1, 2, \cdots, n \\ 0 & \text{otherwise} \end{cases}$$

Now the researcher's model is $\{f(x|p) : p \in [0, 1]\}$ with $f(x|p)$ given by the binomial pmf above. Figure 1 below shows what the researcher expects to see as data X under the hypergeometric or the binomial distribution for three possible values of p, namely, $p = 1/4$ (solid line), $p = 1/2$ (broken line) and $p = 3/4$ (dotted line).

Example (Trend of TC counts)**.** A climate researcher wants to study whether hurricane activity is intensifying with time. One way to do it is to study the annual counts of tropical cyclones (TC) in an ocean basin, say the north Atlantic basin, for the past 100 years. The data is then of the form $X = (X_1, X_2, \cdots, X_{100})$, with X_t giving the TC count in year t. To describe this data, we can first focus on describing one X_t. Since X_t is a count, we can describe it by a Poisson pmf:

$$f_t(x_t|\mu_t) = \begin{cases} \frac{e^{-\mu_t} \mu_t^{x_t}}{x_t!} & \text{for } x_t = 0, 1, 2, \cdots \\ 0 & \text{otherwise} \end{cases}$$

where μ_t represents the expected count for year t. Now to describe, $X = (X_1, X_2, \cdots, X_{100})$ we can treat the component X_t's as independent and write

$$f(x|\{\mu_t\}) = f_1(x_1|\mu_1) \times f_2(x_2|\mu_2) \times \cdots \times f_{100}(x_{100}|\mu_{100})$$

which gives the joint pmf of X at $x = (x_1, x_2, \cdots, x_{100})$.

Although the above gives a description of X, it is not clear how to study the climate researcher's question within this framework. To achieve this, we now need to say something

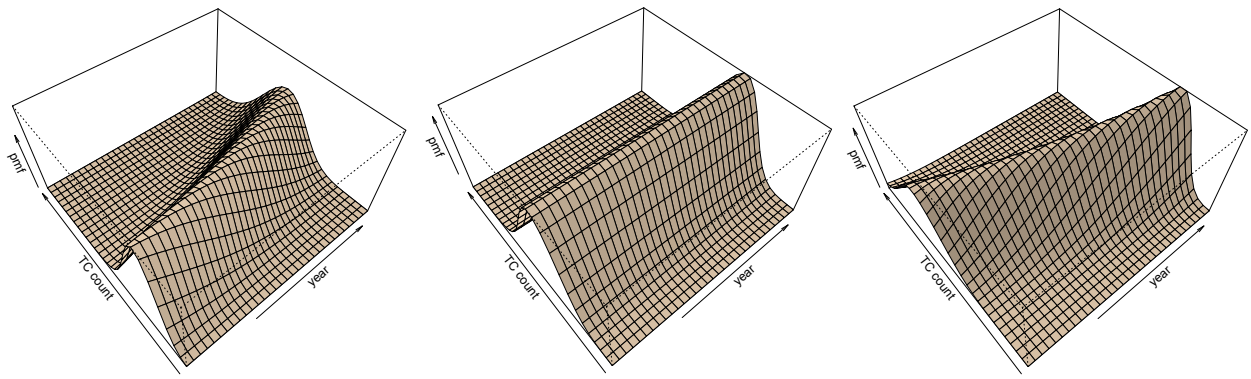

Figure 2: X = annual TC counts for 100 consecutive years. Description of X under Poisson distributions with mean μ_t in year t. Three possible specifications $\mu_t = \alpha\beta^{t-1}$ are considered; $(\alpha, \beta) = (7, 1.005)$, $(12, 1)$ and $(20, 0.995)$.

about how the different μ_t compare to each other, and in particular, how they evolve over time. One possible description is the following:

$$\mu_t = \alpha\beta^{t-1}, \quad t = 1, 2, \cdots, 100$$

which says that the expected annual counts are evolving over time as $\mu_t = \beta\mu_{t-1}$, with a growth factor β.

The research question of whether TC activity is increasing can now be represented by various values of (α, β). In particular, $\beta > 1$ means that TC counts have an upward trend, with larger β indicating faster growth. On the other hand, any $\beta \leq 1$ indicates no or downward trend. Therefore a statistical model for X is given by $\{f(x|\mu_0, \beta) : \alpha \in (0, \infty), \beta \in (0, \infty)\}$ where

$$f(x|\alpha, \beta) = f_1(x_1|\alpha) \times f_2(x_2|\alpha\beta) \times \cdots \times f_{100}(x_{100}|\alpha\beta^{99}).$$

Figure 2 shows the description of X under three choices of (α, β): (7, 1.005), (12, 1) and (20, 0.995).

Note that unlike the previous example, the the choice of model for this example was a lot less obvious. Indeed, one could use many distributions, instead of a Poisson pmf, to describe each X_t. Furthermore, the evolution of μ_t over time t, could also be described in many different ways. What we have built here is "a" description of the data, whether there is a better description can always be debated.

2 Formulating Statistical Inference Problems

Once data has been adequately described by a statistical model, the next task is to formulate the research question in terms of the model. There are a couple of ways this can be done.

Hypothesis testing: First, we could present the research question as trying to decide between two competing "hypotheses" or statements about the model parameter θ: $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$ where Θ_0 and Θ_1 form a partition of the parameter space, i.e., Θ_0 and Θ_1 are disjoint and $\Theta = \Theta_0 \cup \Theta_1$.

For the TC count example, one could set up the research question about trend in terms of the hypotheses $H_0 : \beta = 1$ (no trend) against $H_1 : \beta \neq 1$ (some trend). Note that here $\Theta_0 = \{(\alpha, \beta) : \alpha > 0, \beta = 1\}$ and $\Theta_1 = \{(\alpha, \beta) : \alpha > 0, \beta > 0, \beta \neq 1\}$. For the opinion poll example, the question of whether the policy has majority support can be presented as $H_0 : p \leq 0.5$ (no majority support) against $H_1 : p > 0.5$ (majority support). Here $\Theta_0 = [0, 0.5]$ and $\Theta_1 = (0.5, 1]$.

Once the research question has been represented by two competing hypotheses about the model parameter, the inference task is to decide which statement should be accepted in light of the data that we observe. Such hypotheses testing forms the mainstay of classical statistics and has played an immense role in both development of statistics as a subject as well as its acceptance by the wider scientific community.

Predicting future data: For the TC trend data, it might be more interesting to assess whether we are going to get a higher number of tropical cyclones in the future, and if so then by what extent? For example, we could think of the count X^* in the 10th year from now, and continuing with our model for TC counts, describe X^* by $\mathsf{Poisson}(\alpha\beta^{109})$ pmf. The inference question can be formulated in terms of statements relating to X^*, such as "are we going to see $X^* > 30$?" or "what is an interval of likely values of X^*?". We try to answer these questions about X^* given all the information we gather on (α, β) from data on $X_1, \cdots, X_n$.

Example (Drug Efficacy)**.** A standard application of statistics is in clinical trials for determining effectiveness of a new drug. A pool of subjects are recruited in the trial and each subject is randomly assigned to get either the new drug (treatment) or a placebo (control). Effectiveness measurements $X_1, \cdots, X_n$ and $Y_1, \cdots, Y_m$ are recorded from the two groups. A standard model for such data is: $X_i \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu_1, \sigma^2)$, $Y_j \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu_2, \sigma^2)$, $-\infty < \mu_1, \mu_2 < \infty$, $\sigma > 0$. Usually interest focus on testing $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$. However, one could also think of two future subjects one assigned to treatment and the other to control and speculate about how their effectiveness measurements $X^* \sim \mathsf{Normal}(\mu_1, \sigma^2)$ and $Y^* \sim \mathsf{Normal}(\mu_2, \sigma^2)$ will compare against each other. We could be interested in quantifying the chances of $X^* > Y^*$ or simply reporting an interval for $X^* - Y^*$.

Reporting an Interval for a Parameter: In the drug efficacy study, if we thought of a large number of future subjects assigned to treatment and another large number assigned to control and looked at the difference between their average effectiveness measures, then this quantity will be approximately $\mu_1 - \mu_2$ under our model. And so an interval of likely values for $\mu_1 - \mu_2$ can be interpreted as an interval of likely values for the average treatment effect. In cases like this, reporting an interval for $\mu_1 - \mu_2$ becomes of direct research interest. We will informally refer to such a task as parameter estimation.

3 Classical & Bayesian Paradigms of Inference

Going forward we will see two major paradigms of carrying out statistical inference. We will see how the paradigm of classical statistics is driven by performance guarantees associated with hypotheses testing. The framework also extends to parameter estimation. We will also see how prediction and parameter estimation fit in more naturally within the Bayesian paradigm which focuses more on quantifying uncertainties about quantities at hand.