STA 114: Statistics

Notes 2. Statistical Models and the Likelihood Function

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.

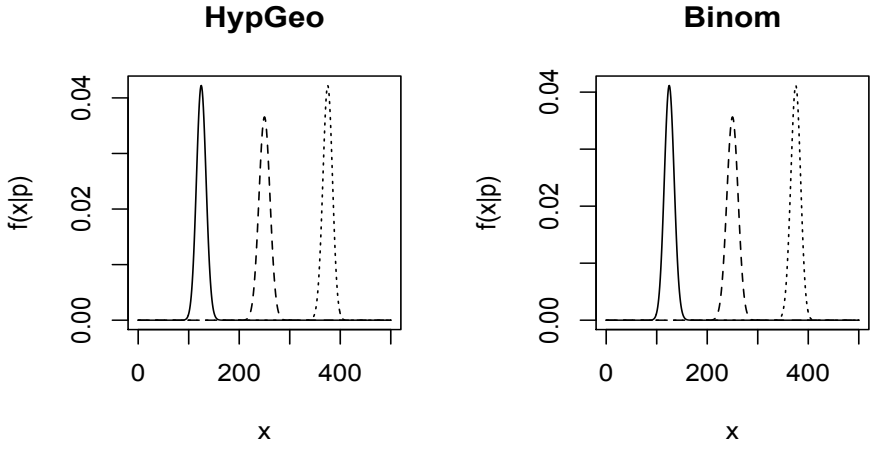

Figure 1: X = number of students favoring the policy in a sample of 500 students. Description of X under hypergeometric (left) and binomial distributions (right) for three possible values of $p = 0.25, 0.5, 0.75$.

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

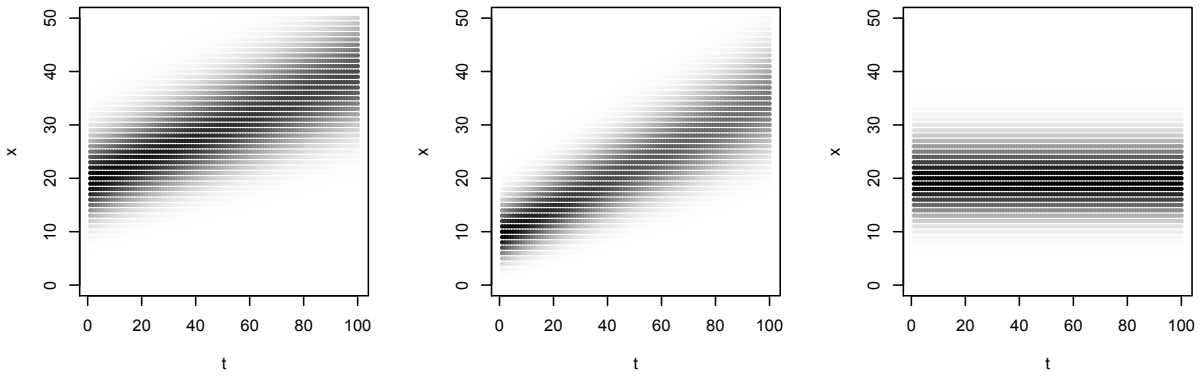

Figure 2: X = annual TC counts for 100 consecutive years. Description of X under Poisson distributions with mean μ_t in year t. Three possible linear specifications $\mu_t = \mu_0 + \beta(t-1)$ are considered.

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 increasing linearly in time, with slope β.

The research question of whether TC activity is increasing can now be represented by various values of (α, β). In particular, a positive sign of β means that TC counts have an upward trend, with larger β indicating faster growth. On the other hand, a zero or a negative value of β indicates no or downward trend. Therefore a statistical model for X is given by $\{f(x|\mu_0, \beta) : \alpha \in (a, b), \beta_0 \in (c, d)\}$ for well chosen limits a, b, c, d, where

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

Figure 2 shows the description of X under three choices of (α, β): (20, 0.2), (10, 0.25) and (20, 0).

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.

The Likelihood Function

Suppose a statistical model $\{f(x|\theta) : \theta \in \Theta\}$ has been constructed for data X, with each θ representing a different theory. When we observed data $X = x$, we can compare two parameter values (i.e., two theories) $\theta = \theta_1$ and $\theta = \theta_2$ by looking at the ratio $f(x|\theta_1)/f(x|\theta_2)$. If this ratio equals 2, then the data $X = x$ is twice as likely to be observed under $\theta = \theta_1$ than it is under $\theta = \theta_2$. Such comparisons can be done based on the **likelihood function**

$$L_x(\theta) := f(x|\theta), \theta \in \Theta.$$

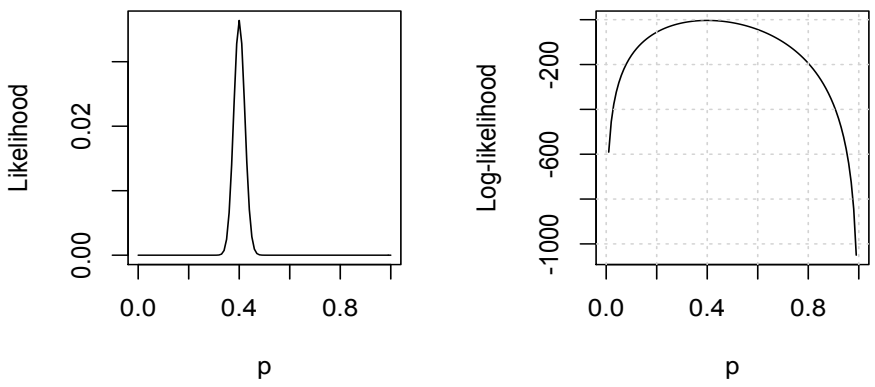

Figure 3: Likelihood and log-likelihood functions in the opinion poll example. The observed data is $X = 200$.

Note that $L_x(\theta)$ is a function over the variable θ taking values in the set Θ.

For all technical purposes, one can work with $L_x(\theta)$ in the log-scale. That is, define the log-likelihood function

$$\ell_x(\theta) = \log L_x(\theta) = \log f(x \mid \theta).$$

Log-scale comparisons between theories are then done by differences $\ell_x(\theta_1) - \ell_x(\theta_2)$.

Example (Opinion Poll, Contd)**.** For the opinion poll example with the statistical model $\{\mathsf{Binomial}(n, p) : p \in [0.1]\}$, the likelihood function in the parameter p is given by

$$L_x(p) = \binom{n}{x} p^x (1-p)^{n-x}, \quad p \in [0, 1]$$

and the log-likelihood function is

$$\ell_x(p) = \log L_x(p) = \log \binom{n}{x} + x \log p + (n-x) \log(1-p), \quad p \in [0, 1]$$

Note that the first term on the right hand side does not involve the function argument p. So we can write

$$\ell_x(p) = \text{const} + x \log p + (n-x) \log(1-p),$$

not caring about the exact value of this additive constant. Indeed, the constant disappears when we look at differences $\ell_x(p_1) - \ell_x(p_2)$.

For data $X = 200$ the theories $p = 0.25$, $p = 0.50$ and $p = 0.75$ receive likelihood scores 6.45×10^{-14}, 1.54×10^{-6} and 1.25×10^{-61}. Figure 3 shows the likelihood function $L_{200}(p)$ and the log-likelihood function $\ell_{200}(p)$ over the grid $p \in \{0.00, 0.01, \cdots, 1.00\}$. These functions indicate that theories with p close to 0.4 fare well in explaining the data $X = 200$. The theory $p = 0.4$ explains the data nearly 10^{80} times better than the theory $p = 0.8$.

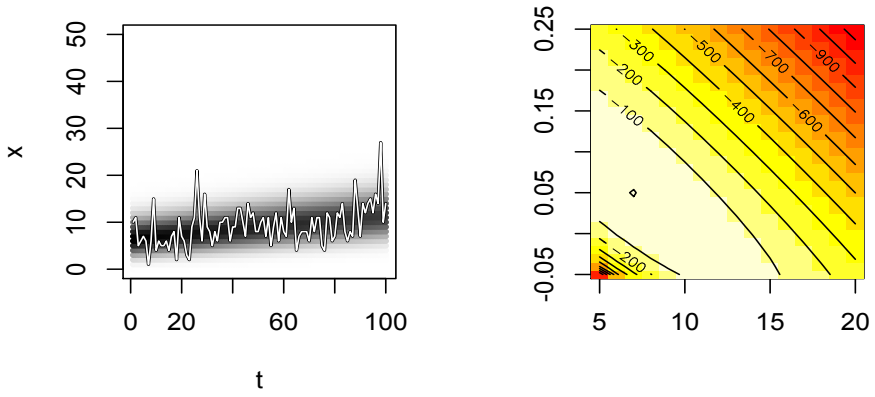

Figure 4: TC counts between 1908 and 2007 and the corresponding log-likelihood function shown as a contour.

Example (TC counts, Contd.)**.** For the statistical model we discussed before, the log-likelihood function is given by:

$$\ell_x(\alpha, \beta) = \text{const} - \sum_{t=1}^{100}(\alpha + \beta(t-1)) + \sum_{t=1}^{100} x_t \log(\alpha + \beta(t-1)).$$

Figure 4 shows the observed annual TC counts between 1908 and 2007 (on the left superimposed on $f(x|7, 0.05)$). A contour plot of the the log-likelihood function over (α, β) is shown on the right. Positive slope values ($\beta > 0$) fare better in explaining the data than negative slopes.

A Word of Caution

The likelihood function gives a numerical comparison of the postulated theories once data $X = x$ is observed. But be clear on what $L_x(\theta_1)/L_x(\theta_2) = 2$ means. It does NOT mean that given the observed data, theory θ_1 is twice more likely than theory θ_1. We don't yet have a platform for discussing likeliness or relative plausibility of theories. Formalizing this concept is the focus of statistical inference.