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 \( \theta \) 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 \( \theta \) is a “parameter” of this model. The set \( \Theta \) of all possible values of \( \theta \) 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

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

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

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

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 $X = (X_1, X_2, \ldots, X_{100})$, 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:

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

where $\mu_t$ represents the expected count for year $t$. Now to describe, $X = (X_1, X_2, \ldots, X_{100})$ 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 $x = (x_1, x_2, \ldots, x_{100})$.

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 \( \mu_t \) 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 \( \mu_t \) compare to each other, and in particular, how they evolve over time. One possible description is the following:

\[
\mu_t = \alpha \beta^{t-1}, \quad t = 1, 2, \ldots, 100
\]

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

The research question of whether TC activity is increasing can now be represented by various values of \( (\alpha, \beta) \). In particular, \( \beta > 1 \) means that TC counts have an upward trend, with larger \( \beta \) 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 \( \{f(x|\alpha, \beta) : \alpha \in (0, \infty), \beta \in (0, \infty)\} \) 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 \( (\alpha, \beta) \): \( (7, 1.005), (12, 1) \) 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 \( \mu_t \) 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 \( \theta \): \( H_0 : \theta \in \Theta_0 \) and \( H_1 : \theta \in \Theta_1 \) where \( \Theta_0 \) and \( \Theta_1 \) form a partition of the parameter space, i.e., \( \Theta_0 \) and \( \Theta_1 \) 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 $\text{Poisson}(\alpha \beta^{109})$ 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 $(\alpha, \beta)$ from data on $X_1, \ldots, X_n$.

**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 $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are recorded from the two groups. A standard model for such data is: $X_i \sim \text{Normal}(\mu_1, \sigma^2)$, $Y_j \sim \text{Normal}(\mu_2, \sigma^2)$, $-\infty < \mu_1, \mu_2 < \infty$, $\sigma > 0$. 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 $X^* \sim \text{Normal}(\mu_1, \sigma^2)$ and $Y^* \sim \text{Normal}(\mu_2, \sigma^2)$ 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: A Very Brief History

There are two major paradigms of carrying out statistical inference: the classical and the Bayesian paradigms. Classical statistics is driven by performance guarantees associated with hypotheses testing. The framework also extends to parameter estimation. On the other hand, prediction and parameter estimation fit in more naturally within the Bayesian paradigm which focuses more on quantifying uncertainties about quantities at hand.
Historically, earliest formal approaches to statistical inference were of the Bayesian flavor, initiated by Thomas Bayes and independently discovered and championed by Pierre-Simon Laplace. Bayes’ now famous paper “An Essay towards solving a Problem in the Doctrine of Chances” was read to the Royal Society in 1763 (after Bayes’ death in 1761). The paper proposed a solution to the “inverse probability” problem, an 18th century speak for “statistical inference”. “Direct probability” problems are what you routinely see in an introductory probability course, e.g., given 10 white and 15 black balls in an urn, what is the probability of drawing a black ball? An inverse probability problem asks the converse: if the urn composition is not known but the first 3 balls drawn from it show 2 whites and one black, what can we say about the number of white and black balls in the urn?.

Both Bayes and Laplace carried out ‘Bayesian’ calculations by using what we now call ‘flat priors’: a prior belief about the unknown model parameter expressed through the uniform distribution on the model space. By late 19th century, this flat prior approach came under severe criticism from several famous statisticians, which culminated in the complete denouncement of this approach by Sir Ronald A Fisher, on the grounds that the flat prior belief and the underlying principle of ‘insufficient reasoning’ (due to Laplace but NOT Bayes) were sensitive to “reparametrization. For the opinion poll example, we could index the model space by a different quantity, say the log-odds: $\theta = \log \left( \frac{p}{1-p} \right)$. The reparametrized model is then given by: $X \sim \text{Binomial}(n, \{1 + \exp(\theta)\}^{-1})$, $-\infty < \theta < \infty$. A Bayesian analysis with a flat prior on $\theta$ leads to different answers than a Bayesian analysis with a flat prior on $p$. Even though the two models are identical (same set of binomial pdfs), the analysis is sensitive to the parametrization.

In a series of work since 1916 and culminating in the classic 1925 text “Statistical Methods for Research Workers”, Fisher developed an alternative ‘Fisherian’ formalization of statistical inference, which eventually led to the modern classical statistics founded on the solid mathematical treatment given by Jerzey Neyman and Egon Pearson in early 1930s. The classical paradigm we see today is mostly the Nyeman-Pearsonian paradigm which is based upon the philosophy (expressed in the context of hypotheses testing):

“Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behavior with regard to them, in following which we insure that, in the long run of experience, we shall not be too often wrong.”

Aided by the Neyman-Pearson formalization, classical statistics underwent rapid advancement and wide acceptance in sciences, industries and governance. Bayesian statistics made a revival in mid 20th century with the works of Jeffreys, Jaynes, Savage, Lindley and later saw a dramatic growth in the 1990s with the advent of Markov chain Monte Carlo computing.

4 The Likelihood Function

A common feature of both the Bayes-Laplace approach and Fisher’s formalization was the use of “the likelihood function”. Suppose a statistical model $\{f(x|\theta) : \theta \in \Theta\}$ has been constructed for data $X$, with each $\theta$ representing a different theory. If we observed data $X = x$, we could 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 are akin to assigning relative scores to parameter values given the observed data, and can be carried out more conveniently by using the “likelihood (score) function”

\[
L_x(\theta) \propto f(x|\theta), \quad \theta \in \Theta,
\]

where any proportionality constant could be used without altering any of the relative scores. Note that \( L_x(\theta) \) is a function of variable \( \theta \in \Theta \) and the whole function depends on the particular observation \( X = x \) as indicated by the subscript. We may drop the subscript from the notation if no risk of ambiguity arises.

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

\[
\ell_x(\theta) = \log L_x(\theta) = \text{const} + \log f(x | \theta)
\]

up to any arbitrary additive constant. Log-scale comparisons between theories are then done by differences \( \ell_x(\theta_1) - \ell_x(\theta_2) \); the additive constant disappears when such differences are taken.

**Example** (Opinion Poll, Contd). For the opinion poll example with the statistical model \( \{\text{Binomial}(n, p) : p \in [0,1]\} \), the likelihood function in the parameter \( p \) is given by any of

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

and the log-likelihood function is

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

For data \( X = 270 \) the theories \( p = 0.25, p = 0.50 \) and \( p = 0.75 \) receive likelihood scores \( 1.2 \times 10^{-43}, 0.0072 \) and \( 1.46 \times 10^{-24} \). Figure 3 shows the likelihood curve \( L_{270}(p) \) and the log-likelihood curve \( \ell_{270}(p) \). These curves indicate that theories with \( p \) close to 0.54 fare well in explaining the data \( X = 270 \). The theory \( p = 0.54 \) explains the data nearly 5 times better than the theory \( p = 0.5 \) and nearly \( 10^{22} \) times better than the theory \( p = 0.75 \).

### 4.1 Bayes-Laplace Approach

Both Bayes and Laplace advocated expressing uncertainty about the unknown parameter \( \theta \) (given observation \( X = x \)) with a pdf proportional to \( L_x(\theta) \). For the opinion poll example with observed data \( X = 270 \), we have \( L(p) \propto p^{270}(1-p)^{230} \) and the normalized pdf on \( p \) would equal \( \text{Beta}(271, 231) \) which assigns only 0.037 probability to \( \Theta_0 = [0, 0.5] \) and 0.963 probability to \( \Theta_1 = (0.5, 1] \).
Figure 3: The opinion poll example with the binomial model. Left panel shows some of the binomial pmfs; and the observed data is \( X = 270 \). Red indicates a good match between theory and observation, blue indicates a poor match. Middle and right panels show the likelihood and log-likelihood functions.

### 4.2 Fisher’s Approach

To test the hypothesis \( H_0 : \theta \in \Theta_0 = [0, 0.5] \), Fisher’s significance testing approach was to first define a measure of evidence in data against \( H_0 \) and then to quantify how extreme the observed evidence is if \( H_0 \) were indeed true. Fisher introduced the theory of maximum likelihood, in which measure of evidence against \( H_0 \) is defined by the “likelihood ratio statistic”

\[
LR(X) = \frac{\max_{\theta \in \Theta} L_X(\theta)}{\max_{\theta \in \Theta_0} L_X(\theta)}
\]

and the extremeness of the observed evidence is quantified by the “p-value”

\[
p = \max_{\theta \in \Theta_0} P\{LR(X) \geq LR(x) \mid \theta\}.
\]

A very small p-value indicates strong or “significant” evidence against \( H_0 \). For the opinion poll example with \( X = 270 \) as the observation \( LR(X) \geq LR(270) \iff X \geq 270 \) due to a special property of the binomial pmfs (monotone likelihood ratio). And so p-value = \( \max_{p \leq 0.5} P(X \geq 270|p) = P(X \geq 270|p = 0.5) = 0.04 \).

### 4.3 Interpretation

With p-value = 0.04 for \( H_0 : p \in [0, 0.5] \), Fisher would argue: *either \( H_0 \) is false or something fairly rare (only 4% chance) has happened*. This statement is logically correct but quite misleading. It appears to give an impression that there is only 4% chance of \( H_0 \) being true – but that is not the correct way to interpret it. If you look carefully, the statement does not provide any sorts of odds for its “either” and “or” parts. In other words, it does not really address the inverse problem, but merely creates an impression of doing so (not so much to statisticians, but to many practitioners using statistics in real world problems). Neyman and Pearson sought to find a logically coherent interpretation of Fisher’s approach and developed the modern classical paradigm based on “frequentist guarantees of statistical rules” (much to Fisher’s chagrin).
The Bayes-Laplace approach, on the other hand, has much more clear interpretation: if your a-priori (i.e., before seeing observed data) about $p$ is best expressed by the \textbf{Uniform}(0, 1) pdf, then by laws of conditional probability (Bayes theorem) you should assign exactly 3.7% chance to $p \in [0, 0.5]$ once $X = 200$ has been observed. This statement of course hinges on accepting \textbf{Uniform}(0, 1) as the prior pdf for $p$. Laplace advocated such a choice citing “insufficient reasoning” – applicable to cases where he felt there was no reason to \textit{a priori} prefer any one value for $p$ over any other value. Bayes would also advocate a similar choice, but only if he felt, \textit{a priori}, that $P(X = j) = 1/(n + 1)$ for each $j = 0, \ldots, n$ [notice that this probability is not conditional on $p$]. The difference between these two postulates is subtle but important. Laplace’s argument breaks down under reparametrization, but Bayes’ argument remain invariant. Of course the modern version of Bayesian statistics allows any arbitrary prior pdf $\pi(p)$ for $p$, for which the \textit{a posteriori} (post-data) uncertainty about $p$ given $X = x$ is simply $\pi(p|x) \propto L_x(p)\pi(p)$. The choice of the prior pdf $\pi(p)$ remains the most critical aspect of a Bayesian analysis.