1

Introductory Statistical Concepts

1.0 Preliminaries and Overview

When analyzing numerical data, subject to random uncertainty, which have been collected in some scientific or real-life context, the first "golden rule" is to study the data, for example, using dotplots, bivariate scatterplots, relative frequency histograms, and contingency tables, before applying any formal statistical technique. Complicated data sets deserve several hours, days, or even weeks of study. When studying a data set, you should realize that data are not simply numbers but rather measurements or counts of real entities (e.g., birthweights of babies, numbers of students passing a college test, a measurement of a real chemical). Therefore, any tentative conclusions should be made in the contexts of their meaning in relation to these entities, the real background of the data, and how the data were collected. The same set of numerical data might mean something entirely different in different scientific or real-life contexts.

Sometimes, upon viewing the data, you may discover a particularly distinctive feature that yields a decisive conclusion. In this case, it may not be necessary, or indeed technically feasible, to proceed to a more formal analysis. For example, when investigating the years of service of French generals during the late eighteenth century (see Wetzler, 1983, Appendix), the conclusion was reached, upon viewing a distinctive spike in the scatterplot, that a number of the generals had been rather abruptly dismissed during the French Revolution. As another example, the State of Wisconsin was advised during a court action in 1986, and based upon data for a carefully collected random sample of $n = 120$ nursing homes, that the state was not adequately reimbursing the actual costs (in dollars per patient per day) for nursing homes with costs in excess of \$45. This conclusion was validated by a distinctive blip in an otherwise linear bivariate scatterplot. The State of Wisconsin conceded the case, largely because the state recognized that data based upon a representative sample (a random sample of 120 nursing homes from 600 nursing homes in Wisconsin) had been collected. George and Wecker (1985) also emphasize the importance of using good statistics in legal cases.

For the first of these analyses, a computer package was not used, since formal summary statistics such as the sample mean and variance would not be particularly relevant. For the second analysis, any attempt to fit a regression model without first carefully considering the data could have led to many hours of fruitless analysis. Similarly, you should try to avoid any "black box" data analytic technique that cannot be combined with an interaction between your thought processes and the data. It is particularly

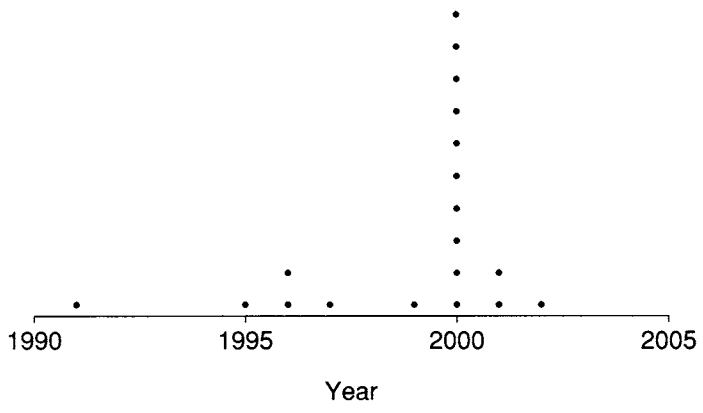

Figure 1.0.1. A dotplot with a spike.

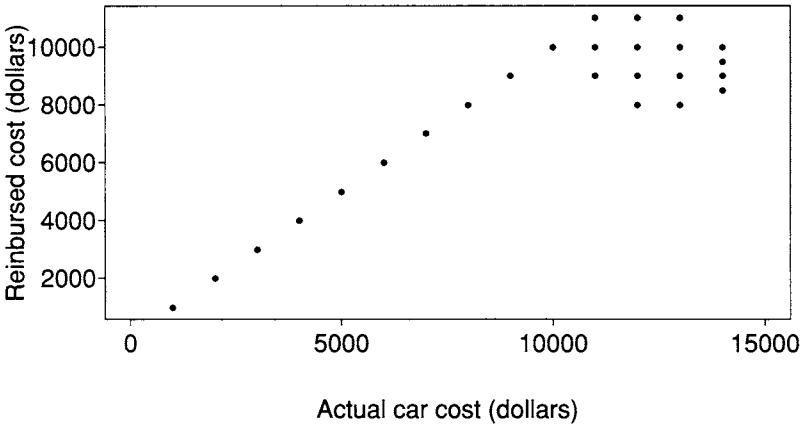

Figure 1.0.2. A bivariate scatterplot with a blip.

important to consider the dotplots and scatterplots. Two further data plots, with a spike and a blip, respectively, are described in Figures 1.0.1 and 1.0.2.

When viewing the data, we should pay careful attention to any outlying observations (see Figure 1.0.3). Outliers are discussed in greater detail by Barnett and Lewis (1978). For example, an outlier can enhance the apparent correlation between two variables that may not otherwise be obviously correlated. Carefully consider the origins and meaning of each outlier and make a careful intuitive decision as to whether or not to include it in the sample. Don't automatically reject outliers, using a "black box" technique,

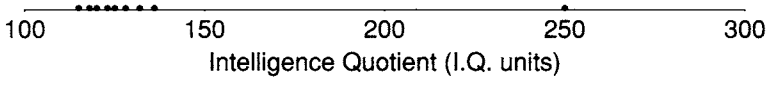

Figure 1.0.3. A dotplot with an outlier.

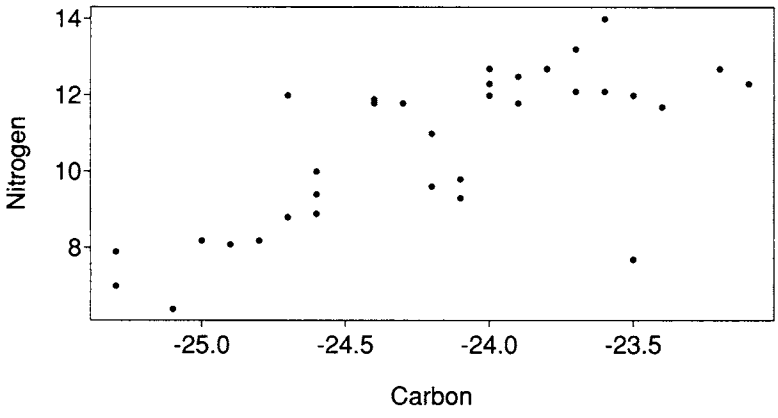

Figure 1.0.4. A scatterplot of the readings on $n = 34$ skeletons.

since they may be quite informative, particularly if they are part of a random sample. Outliers can of course strongly influence any formal analysis, so it is essential to be aware of them. We also regard it as important not to "impute" values for missing data, since the modeling process for the whole data set can then become confused with the imputation process, and the imputed values can exaggerate the information content of the data. Likelihood and Bayesian methods will be able to readily handle missing data problems (just integrate the sampling distribution with respect to the missing observations), without any need to impute the data.

A typical archaeologists' diagram for recording the (transformed) nitrogen and carbon content of skeletons compresses the carbon axis, creating a tendency for the skeletons to be divided into groups, according to nitrogen content alone. In cases where there are two groups, the group of skeletons with the higher nitrogen content is often taken to be Mesolithic, and the group with lower nitrogen content, Neolithic. In Figure 1.0.4, however, we report the entire scatterplot of the readings on $n = 34$ skeletons, at the Lepenski Vir site in the Danube valley, but with the carbon scale substantially broadened when compared with the archaeologists' procedures. Our scatterplot suggests division into several groups rather than two groups. Indeed, McLachlan's package for multivariate mixtures (McLachlan and Basford, 1988) indicates at least six groups. The strong case for at least four or five groups can be confirmed by matching the groups with gender. Further discussions of these data are provided by Bonsall, Lennon, and McSweeney (1997).

For many data sets, it is also of interest either (a) to draw inferences about unknown parameters of interest, for example, the density of a fluid, the recovery rate for patients receiving a particular treatment, or population means, or proportions, or (b) to be able to predict future observations, given the current and previous observations, for example, economic forecasting, the forecasting of the paths of hurricanes, or the prediction of the probability of failure of an engineering design. It is then useful to formalize the random variability or uncertainty in the data, using a mathematical probability model, that is, by taking the numerical observations to be realizations of random variables whose joint

distribution comprises the mathematical or "sampling" model. A second "golden rule" is to realize that "true models" are available only in limited situations. In many cases, a variety of different models can be taken to plausibly represent a data set with a finite number n of observations. Which model to use depends partly on statistical technique, but also on the meaning and usefulness of the model in relation to the actual context of the data.

Given a particular sampling model, a key question is, How should the applied statistician use the data to draw inferences about any unknown parameters appearing in the model? In this text, we adhere to the principle, Given the truth of the sampling model, all information in the data is summarized by the likelihood function. The beautiful concept of likelihood links all major philosophies of statistics and provides a cornerstone of the Bayesian paradigm. Its properties and applications are developed in detail throughout this chapter.

It is possible to draw objectively acceptable conclusions from data, when appropriate randomization is performed at the design stage, ideally with further replications of the experiment, to detect unlucky randomizations. For uncontrolled data, an appropriate model can be more difficult to find, and any conclusions are subjective and subject to the effects of "lurking" or "confounding" variables (see Section 1.2 (H)). In general, the conclusions are subject to "shades of subjectivity," depending upon the way the data are collected. For example, Brown et al. (1997) experienced considerable practical difficulties in collecting a random sample while surveying primary-care patients for drug or alcohol abuse. This was mainly the case because the interviewers were under considerable pressure to complete their interviews within time periods agreed upon with the clinics. The conclusions therefore needed to be qualified accordingly. There are also frequently problems with the selective reporting of significant results (see Dawid and Dickey, 1977). Furthermore, the sample size should be chosen with care at the design stage (Donner, 1984).

Both Fisherian and Bayesian statistics depend heavily upon the concept of "probability." What is probability? For a statistical experiment $\mathcal{E}$, with sample space S, mathematicians will tell you that a probability distribution $p(\cdot)$ is a real-valued function defined on all events (strictly speaking, events are constrained to be "measurable subsets") contained in S and satisfying the Kolmogorov axioms (see Exercise 1.1.l). However, philosophically speaking, there are three main types of probabilities:

(A) *Classical probability:* This is defined by an "m over k" rule and is appropriate whenever $S = \{e_1, e_2, \ldots, e_k\}$ possesses k outcomes that are judged to be "equally likely," and when an event A consists of m of these k outcomes. When the equally likely assumption is made objectively, such as when the outcome that occurs has been chosen at random from the k outcomes, or the equally likely assumption has been tested by replicating the experiment numerous times under identical conditions, then the probability $p(A)$ of the event A, defined by $p(A) = m/k$, can be referred to as an "objective classical probability." When the equally likely assumption is made subjectively (e.g., in the absence of evidence to distinguish that any particular outcome is more likely than any other), then $p(A) = m/k$ can be referred to as a "subjective classical probability." See

also Exercises 1.1.a and 1.1.b, which tell us that population proportions can be identified with classical probabilities when individuals are chosen at random from the population.

(B) *Frequency probability:* This will be defined by equation (1.1.1). The frequency probability of an event is the long-run proportion of times the event occurs in a large number of replications of the experiment. Objective classical probability provides an example of frequency probability. Therefore, since the objective classical probability that a roulette wheel will give a black number is 9/19, this can also be used to predict the long-run performance of the wheel.

(C) *Subjective probability:* This measures an individual's uncertainty in an event and may vary from individual to individual. You may calibrate your subjective probabilities by judging whether events A are equally likely to events for an objective auxiliary experiment, for example, the spinning pointer of Exercise 1.1.k. In principle, you should assign probabilities to all events $A \subseteq S$ and ensure that your probabilities satisfy the Kolmogorov axioms. An individual who always tries to represent his uncertainty by a subjective probability distribution is referred to as a "Bayesian."

In general, a number between zero and unity can be regarded as a probability only if all other events in the sample space are envisioned, probabilities are assigned to every event, and the laws of probability, as defined by the Kolmogorov axioms, are satisfied by the entire collection of probabilities (referred to as a "probability distribution"). Many "probabilities" quoted in science and the media do not satisfy these conditions. For a sample space with either finitely many outcomes or outcomes that can be arranged in an infinite sequence, it is sufficient to check that the values assigned to the individual outcomes sum to unity.

Consider situations where you possess some information regarding an unknown parameter θ, for values of θ lying in a parametric space Θ. Then a big question is whether or not you can represent this information by a subjective probability distribution on Θ. Some Bayesians say, You should always represent your information by a subjective probability distribution on Θ, since there are some very simple axioms that tell us that if you don't act like this, then you are irrational, incoherent, and moreover, a sure loser! We do not concur with this type of "normative approach," largely because we are unaware of an axiom system that is simple enough, when compared with the Kolmogorov axioms, to justify this viewpoint. Moreover, some information, such as medical knowledge or evidence in a court case, may be too diverse or eclectic to be representable by probabilities. (These views are open to discussion. For a more traditionally Bayesian approach, see Bernardo and Smith, 1994, section 2.3. Many Bayesians believe that the uncertainty in any event is representable by a probability. These aspects are pursued in Exercise 1.1.k and were previously debated by Leonard, 1980, and discussants. In our current chapter, we also debate the likelihood principle. See Section 1.5 (C), Exercises 1.5.c–1.5.f.)

Other topics discussed in the current chapter include Akaike's and Schwarz's information criteria, AIC and BIC, for deciding between different choices of sampling

models. These subtract a penalty per parameter from the log-likelihood function (see equations 1.1.5 and 1.1.6). Information criteria are best justified and compared by computer simulation of sets of observations from particular choices of their true sampling model (see Sections 1.2 (F) and (G)). However, it is also important to consider all possible diagnostics, for example, residual analyses for regression models, when comparing models (see Exercise 1.5.l) and also to consider the real-life or scientific reasonability of the candidates.

Any formal statistical procedure, whether for inference about parameters, prediction of future observations, or choice of sampling model, should possess desirable long-run frequency properties (e.g., good mean squared error (MSE) for estimation of parameters, accurate frequency coverage for approximate confidence intervals, high long-run probability of choosing a reasonable model). In situations where these cannot be developed theoretically, computer simulations can produce accurate and meaningful results. Graduate students and research specialists are encouraged to create novel statistical procedures, but then always to check their new ideas by using frequency simulations.

In Section 1.3, procedures are described for obtaining approximate confidence intervals that closely relate to the multivariate normal likelihood approximation (1.3.11) and that refer to the concept of transforming the parameters to achieve possibly better approximate normality. It is particularly important for the research worker to numerically check any theoretical suggestions when using theoretical approximations, since the numerical work may produce some surprises or suggest adjustment terms to the approximations. For example, an "approximately normally distributed random variable X" might not yield values for $p(X < -1.96)$ or $p(X > 1.96)$ that are particularly close to 0.025, as required for an exact result. A variety of practical justifications of the approximations employed in the text are included (e.g., Sections 1.2 (C), 1.4 (A) and (B)). Some key properties of the multivariate normal distribution are developed in Exercise 1.1.n.

A multivariate normal approximation (1.3.11) and related parameter transformations will provide a central theme to a variety of Bayesian ideas developed later in the text, such as the construction of "prior distributions" for several parameters, computational procedures using importance sampling, Laplacian approximations, and rejection sampling. It is more important for the reader to understand the multivariate normal approximation and related approximate confidence intervals than to research the complicated asymptotic theory of maximum likelihood estimators. For any particular model, it is better to check the validity of (1.3.11) computationally and for finite sample sizes.

The works of Sir Ronald Fisher provide excellent background to this chapter. See, for example, Fisher (1925, 1935, 1959) and Bennett (1971–4). Fisher always mixed his techniques with practical common sense.

1.1 Sampling Models and Likelihoods

Numerical data often arise as a result of some statistical experiment $\mathcal{E}$, that is, an occurrence with a random or uncertain outcome. Suppose that on a single repetition of $\mathcal{E}$, you observe n numerical observations $y_1, \ldots, y_n$. Let the sample space S denote

the set of all possible realizations of the column vector $\mathbf{y} = (y_1, \ldots, y_n)^T$. Then S is a subset of n-dimensional Euclidean space R^n, and the vector $\mathbf{y}$ consists of the n observations, arranged in a column.

You might be prepared to make the quite strong assumption that $\mathbf{y} = (y_1, \ldots, y_n)^T$ is a numerical realization of a random vector $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ (i.e., the column vector of the random variables $Y_1, \ldots, Y_n$), which possesses some probability distribution P, defined on events in S. For example, if $\mathcal{E}$ can be repeated a large number of times under identical conditions (i.e., *replicated*), then $P = P(\cdot)$ can be defined in terms of the frequency notion of probability. That is, for any event A contained in S,

$$P(A) = \text{prob}(\mathbf{Y} \in A) = \lim_{m \to \infty} r_m(A), \tag{1.1.1}$$

whenever the limit on the right-hand side exists, where the relative frequency $r_m(A)$ denotes the proportion of times that $\mathbf{y} \in A$, during the first m replications of $\mathcal{E}$. However, if $\mathcal{E}$ can be performed only once, then the assumption that $\mathbf{y}$ is a numerical realization of $\mathbf{Y}$ cannot always be made objectively. In some situations where you use random sampling from a population or in other situations where outcomes of the experiment can be regarded as equally likely, it will still be possible to define P in an objective fashion. However, in many cases, part of the modeling process, that is, the specification of P, will need to be performed subjectively and by reference to the scientific or social background of the data. In many cases, P will be incompletely known, even after a variety of modeling assumptions, and it is therefore frequently necessary to infer reasonable choices of P, based upon the vector $\mathbf{y}$ of observations, for a single repetition of $\mathcal{E}$.

For simplicity, assume that $\mathbf{Y}$ is either a continuous random vector with density $p(\mathbf{y})$, for $\mathbf{y} \in S$, or a discrete random vector with probability mass function $p(\mathbf{y})$, for $\mathbf{y} \in S$. Then, following the tenets of *parametric statistical inference*, you might wish to make an assumption of the form

$$p(\mathbf{y}) = f(\mathbf{y} \mid \boldsymbol{\theta}) \qquad \left(\mathbf{y} \in S,\ \boldsymbol{\theta} \in \Theta \subseteq R^k\right), \tag{1.1.2}$$

where $f(\mathbf{y} \mid \boldsymbol{\theta})$ is specified as a function of both $\mathbf{y} \in S$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^T \in \Theta$. Here $\boldsymbol{\theta}$ is some vector of unknown parameters, and Θ is the parameter space. If $n \geq k$, you can now make inferences about a k-dimensional vector $\boldsymbol{\theta}$ of unknown parameters, rather than an entire function $p(\cdot)$.

Box (1980) distinguishes between this *inference* problem and the problem of statistically *modeling* the choice of functional form for f. Modeling involves both creating an appropriate choice for f in relation to the scientific background and checking the reasonability of this choice against the data. Modeling requires substantial inductive thought, while inference requires deduction, that is, the calculation of mathematical conclusions, given that the functional form of the model is assumed true. This blend of inductive and deductive thought is part of the *inductive synthesis* (Aitken, 1944, p. 3).

Following Birnbaum's (1962) philosophy of "the irrelevance of observations not actually observed" (e.g., why use procedures involving significance probabilities, minimum variance criteria, and confidence statements, which average across the sample space?) and Edwards's famous 1972 treatise on likelihood, it is a reasonable and widely

held contention, though without a watertight scientific proof, that the *likelihood function* summarizes all the information about $\boldsymbol{\theta}$ contained in the data, when the functional form of f is assumed true, and you condition on the numerical realization of $\mathbf{y} = (y_1, \ldots, y_n)^T$ actually observed. The likelihood function is a function of $\boldsymbol{\theta}$ and satisfies

$$l(\boldsymbol{\theta} \mid \mathbf{y}) = f(\mathbf{y} \mid \boldsymbol{\theta}) \qquad (\boldsymbol{\theta} \in \Theta), \tag{1.1.3}$$

where $\mathbf{y}$ is the observed $\mathbf{Y}$. Hence (1.1.3) provides a complete solution, at least in principle, for the inference problem. For example, if $k = 1$ or 2, you can just consider a sketch of the likelihood function and draw all your conclusions from this graphical procedure, together with your background experience. Any vector $\hat{\boldsymbol{\theta}}$ maximizing (1.1.3), as a function of $\boldsymbol{\theta} \in \Theta$, with $\mathbf{y}$ fixed, provides a *maximum likelihood estimate* of $\boldsymbol{\theta}$. In intuitive terms, this gives the realization of $\boldsymbol{\theta}$ most likely to have given rise to the current data set, an important finite sample property. Different choices or hypotheses, say, $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, for $\boldsymbol{\theta}$ may be compared by choosing the hypothesis that maximizes (1.1.3), with $\mathbf{y}$ fixed. However, areas under the likelihood function are not immediately relevant; they should in particular not be interpreted as probabilities.

Note that the expected log-likelihood

$$E\big[\log l(\boldsymbol{\theta} \mid \mathbf{Y})\big] = \int_S f(\mathbf{y} \mid \boldsymbol{\theta}) \log f(\mathbf{y} \mid \boldsymbol{\theta})\, d\mathbf{y}$$

is the negative of the *entropy* associated with the sampling density $f(\mathbf{y} \mid \boldsymbol{\theta})$. Under broad regularity conditions, $n^{-1} \log l(\hat{\boldsymbol{\theta}} \mid \mathbf{Y})$ will converge, to this expectation, with probability one, as n tends to infinity. This introduces entropy as an information criterion in statistics and tells us that $\hat{\boldsymbol{\theta}}$ is associated with minimizing the disorder or lack of information about the sampling model. The modeling process can be assisted by a closely related criterion. Suppose that you use your knowledge or intuition to constrain f to belong to some family $\mathcal{F}$ of meaningful functional forms, but where the dimension k may vary among members of the family. Then choose $f \in \mathcal{F}$ to maximize

$$\text{GIC} = \text{General Information Criterion} = \log l\big(\hat{\boldsymbol{\theta}} \mid \mathbf{y}\big) - \frac{\alpha k}{2}. \tag{1.1.4}$$

Here $\log l(\hat{\boldsymbol{\theta}} \mid \mathbf{y})$ denotes the supremum of the log-likelihood function, and $\frac{1}{2}\alpha$ provides a *penalty per parameter* in the model. The choices $\alpha = 2$ (Akaike, 1978) and $\alpha = \log(n/2\pi)$, (Schwarz, 1978) have been suggested, with some theoretical justification (e.g., Stone, 1977, 1979), providing

$$\text{AIC} = \text{Akaike's Information Criterion} = \log l\big(\hat{\boldsymbol{\theta}} \mid \mathbf{y}\big) - k \tag{1.1.5}$$

and

$$\text{BIC} = \text{Bayesian Information Criterion} = \log l\big(\hat{\boldsymbol{\theta}} \mid \mathbf{y}\big) - \frac{k}{2} \log \frac{n}{2\pi}. \tag{1.1.6}$$

These criteria help us to find a concise model, with just a few parameters. It is also important to consider the meaning of the model in a scientific context. We should avoid

complicated models, or model choice based upon small sample sizes. L. J. Savage (see, Lindley, 1983) felt that "a model should be as big as an elephant," but this can be contrasted with the late Toby Mitchell's philosophy, "the greater the amount of information, the less you know," that is, a complicated model might be difficult to interpret. An overall predictive check of the modeling assumptions proposed by Box (1980, 1983) gives only a partial answer, and it more generally seems to be very difficult to check convincingly any particular model without a particular set of alternatives in mind by reference to the current data alone. In other words, applied scientific judgment is invariably needed. Of course, even if a model is obviously incorrect, it might still help us to extract some conclusions from the data, since a model with k parameters helps us to reduce the dimensionality of the problem from n to k and to make some conditional conclusions. This process can be repeated with different tentative models. We may hence use the model to "telescope" low-dimensional pictures of the data. In other words, some models are objective, others are subjective, and the modeling process helps us to perceive the data.

Three examples are now described to illustrate the algebraic manipulations needed when constructing likelihoods.

Worked Example 1A: *Maximum Likelihood for the Geometric Distribution*

Let $Y_1, Y_2, \ldots, Y_n$ denote a random sample from a geometric distribution, with the probability mass function (p.m.f.)

$$p(Y_i = y_i \mid \theta) = \theta(1-\theta)^{y_i - 1} \qquad (y_i = 1, 2, \ldots). \tag{1.1.7}$$

(a) Find the likelihood of θ.

(b) The maximum likelihood estimate $\hat{\theta}$ of θ maximizes the probability of obtaining the observations actually observed. Find $\hat{\theta}$.

(c) The invariance property of maximum likelihood estimates tells that for any function $\eta = g(\theta)$ of θ, $\hat{\eta} = g(\hat{\theta})$ is the maximum likelihood estimate of $g(\theta)$. Find the maximum likelihood estimate of $\eta = \theta(1-\theta) = p(Y_1 = 1)$.

Model Answer 1A:

(a) The joint p.m.f. of $Y_1, \ldots, Y_n$ is

$$\begin{aligned} &p(Y_1 = y_1, \ldots, Y_n = y_n \mid \theta) \\ &= \prod_{i=1}^{n} p(y_i \mid \theta) = \prod_{i=1}^{n} \theta(1-\theta)^{y_i - 1} = \theta^n \prod_{i=1}^{n} (1-\theta)^{y_i - 1} \\ &= \theta^n (1-\theta)^{\sum_{i=1}^{n}(y_i - 1)} = \theta^n (1-\theta)^{\sum_{i=1}^{n} y_i - n} \\ &= \theta^n (1-\theta)^{n(\bar{y}-1)}, \end{aligned}$$

where $\bar{y} = n^{-1}\sum_{i=1}^{n} y_i$. If $y_1, \ldots, y_n$ are the numerical realizations of $Y_1, Y_2, \ldots, Y_n$, this p.m.f. also provides the likelihood

$$l(\theta \mid \mathbf{y}) = l(\theta \mid y_1, \ldots, y_n) = \theta^n (1-\theta)^{n(\bar{y}-1)} \qquad (0 < \theta < 1),$$

which, as a function of θ, is a beta curve.

(b) It is often technically easier to maximize the likelihood by maximizing the log-likelihood. In our example,

$$\log l(\theta \mid \mathbf{y}) = n \log \theta + n(\bar{y}-1)\log(1-\theta),$$

$$\frac{\partial \log l(\theta \mid \mathbf{y})}{\partial \theta} = \frac{n}{\theta} - \frac{n(\bar{y}-1)}{1-\theta},$$

implying

$$\frac{\partial^2 \log l(\theta \mid \mathbf{y})}{\partial \theta^2} = -\frac{n}{\theta^2} - \frac{n(\bar{y}-1)}{(1-\theta)^2}.$$

Setting the first derivative equal to zero, we find that $\hat{\theta}$ satisfies

$$\frac{n}{\hat{\theta}} = \frac{n(\bar{y}-1)}{1-\hat{\theta}}$$

and

$$1 - \hat{\theta} = (\bar{y}-1)\hat{\theta},$$

so that $\hat{\theta} = 1/\bar{y}$. This single stationary point provides a global maximum, since the second derivative of the log-likelihood is negative, for all values of θ. This implies that the likelihood curve is convex.

(c) $\hat{\eta} = \hat{\theta}(1-\hat{\theta}) = \bar{y}^{-2}(\bar{y}-1)$, since, by the invariance property of maximum likelihood estimates, $\hat{\theta} = 1/\bar{y}$.

Worked Example 1B: *Likelihood Methods for the Poisson Distribution*

Let $Y_1, Y_2, \ldots, Y_n$ denote a random sample from a Poisson distribution with mean μ and p.m.f.

$$p(y) = \frac{e^{-\mu}\mu^y}{y!} \qquad (y = 0, 1, 2, \ldots; 0 < \mu < \infty).$$

Note that $U = Y_1 + Y_2 + \cdots + Y_n$ possesses a Poisson distribution with mean $n\mu$.

(a) Find the likelihood $l(\mu \mid \mathbf{y})$ of μ, given that $Y_1 = y_1, \ldots, Y_n = y_n$.

(b) Show that $l(\mu \mid \mathbf{y}) = h(\mathbf{y})l^*(\mu \mid \mathbf{y})$, where $h(\mathbf{y})$ does not depend on μ, and the likelihood kernel $l^*(\mu \mid \mathbf{y})$ depends only upon $y_1, y_2, \ldots, y_n$ via the sample mean $\bar{y}$, when n is specified.