Part I

Probability Functions, Probability Density Functions, and Their Cumulative Counterparts

Chapter 1

Discrete Probability and Cumulative Probability Functions

> While being shown a house to buy, Garp and his wife Helen witness a single-engine plane, presumably suffering catastrophic mechanical failure, plowing right into the side of the house. Garp takes this as a good sign – "The odds of another plane hitting this house are astronomical!" – and agrees right then and there to buy the house. (John Irving, *The World According to Garp*).

1.1 Overview

The first random variable typically encountered by students of basic statistics is known as a *discrete* random variable, after which they proceed to study *continuous* random variables. Discrete random variables do not always receive as much attention as continuous random variables receive, but in a nonparametric framework, the importance of their study should not be understated. Whether the discrete random variable is the number of times a single-engine plane crashes into a home or whether option "a", "b", or "c" was selected by a respondent on a questionnaire, it plays a fundamental role in statistical analysis.

A discrete random variable is one that can take on a *countable* number of values. They come in many different flavours and go by a variety of names including *nominal* (*unordered*) and *ordinal* (*ordered*) categorical variables. Examples would include the number of heads in three tosses of a coin where the random variable takes on the values $\{0, 1, 2, 3\}$, or an individual's employment status being classified as either "employed" or "unemployed" (i.e., an *unordered* categorical variable), or a response to a survey question recorded as one of "a", "b", or "c" where "a" indicates "most preferred" and

"c" "least preferred" (i.e., an *ordered* categorical variable). Their defining features are that their support[1] is *discrete*, repeated values in a random sample are to be expected, and *counting* the number of sample realizations for a particular outcome is a sensible thing to do.

Although the probability function for a discrete support random variable plays a key role in statistical inference, in applied settings this function is generally unknown and must be estimated. There are three approaches we might entertain when estimating the unknown probability function for a discrete random variable:

i. Presume a parametric family (e.g., binomial) and estimate under this presumption.
ii. Use the (nonsmooth) sample proportions.
iii. Use a kernel-smoothed approach.

The first two are standard fare and are routinely taught in introductory courses on data analysis. The third, however, is likely far less familiar. One drawback with the first approach is that if the parametric family we have assumed is not compatible with the underlying data generating process (DGP), then the resulting estimates can be statistically *biased* and *inconsistent*. One drawback with the second approach is that, even though it is *unbiased* and *consistent*, there may be very few realizations of a particular outcome in the sample at hand, and hence the sample proportion for such an outcome will be highly variable.[2] The third approach introduces some finite sample bias by smoothing the sample proportions in a particular manner,[3] but this smoothing also reduces finite sample variance. The estimator that uses kernel smoothing is *asymptotically unbiased* and *consistent*, and may therefore exhibit better finite sample performance than either of its peers.

One of the benefits of beginning the study of nonparametric methods with kernel-smoothed estimators of probability functions is that, at least for the unordered case, there is no need for the type of approximation that is required when studying the kernel-smoothed estimators of density functions, which we will do in Chapter 2. We obtain simple and exact expressions for quantities such as the bias and variance of the estimator, its *summed* mean square error, and optimal smoothing parameters, among others. And for a special ordered case, we are introduced to an approximation technique that is widely used when studying kernel-smoothed estimators of density functions (this appears as an exercise). Another benefit is that when we migrate to the mixed-data case (i.e., datasets containing a mix of continuous

[1] By *support* we simply mean the *sample space* or set of all possible outcomes, i.e., it is the set of all outcomes whose probability (or probability density that we study in Chapter 2) is strictly positive.

[2] See Simonoff (1996) who proposes the use of discrete support kernel functions for smoothing sparse contingency tables.

[3] Essentially it *shrinks* the sample proportions in the direction of the discrete uniform distribution.

and discrete support random variables) the powerful potential uncovered by smoothing discrete support random variables in the manner outlined below will be revealed.

1.2 Parametric Probability Function Estimation

Suppose that we were interested in modeling a univariate probability function for some discrete random variable X. Furthermore, without loss of generality,[4] assume that $X \in \mathcal{D} = \{0, 1, \ldots, c-1\}$ where c is the number of outcomes taken on by X, and assume that $\{X_1, X_2, \ldots, X_n\}$ represents n independent random draws from the probability distribution $p(x)$. We denote the probability function $p(x) = \Pr(X = x)$, $x \in \mathcal{D}$, $0 \leq p(x) \leq 1$, $\sum_{x \in \mathcal{D}} p(x) = 1$ (the last two are necessary conditions for *proper* probabilities). In general, $p(x)$ is unknown and must be estimated.

Suppose that we took a parametric approach towards modeling the unknown probability function $p(x)$. The parametric approach would presume a parametric distribution for the unknown $p(x)$.[5] By way of illustration, we might presume that the data were generated from the binomial distribution given by

$$p(x; \pi) = \binom{n}{x} \pi^x (1 - \pi)^{n-x},$$

where n is the number of trials, π the probability of a *success* on each trial, and $\binom{n}{x} = n!/((n-x)!x!)$ with $x! = x \times (x-1) \times (x-2) \times \cdots \times 1$ and $0! \equiv 1$. We shall make use of this parametric model in the following illustrative example.

Example 1.1. Boy-Girl Ratio in Families (Adapted from Berry and Lindgren (1990), page 563).

> Occasionally, we hear parents remark something along the lines of "we have three boys and wanted a girl so thought we would try again", which begs the question of whether the sex of successive children in a family is akin to a coin toss, i.e., whether it behaves like a sequence of independent *Bernoulli* trials with the probability π of having a boy and $1 - \pi$ of having a girl. If so, the number of boys in a family of given size is binomially distributed, and we can compute the probability of obtaining x boys in a family of 8 children under this presumption where $x \in \mathcal{D} = \{0, 1, 2, \ldots, 8\}$. In a random sample of $n = 1,000$ families having eight children, there were 4,040 boys so our estimate of π is $\hat{\pi} = 4,040/8,000 =$

[4] The generality is that here we assume X is integer-valued, but it could just as easily be the characters "a" and "b".

[5] Common distributions for discrete random variables include the hypergeometric, Poisson, binomial, and negative binomial, by way of example.

Table 1.1: Boy-girl ratio in families. The null probability is $p_0(x) = \binom{8}{x}0.505^x(1-0.505)^{8-x}$ and the expected frequency is $e_x = 1,000 \times p_0(x)$.

x	Null Probability	Expected Frequency	Observed Frequency
0	0.0036	3.6	10
1	0.0294	29.4	34
2	0.1050	105.0	111
3	0.2143	214.3	215
4	0.2733	273.3	239
5	0.2231	223.1	227
6	0.1138	113.8	115
7	0.0332	33.2	34
8	0.0042	4.2	15

0.505. We compute the expected number of families with $x = 0, 1, \ldots, 8$ boys (i.e., $1,000 \times p_0(x)$ where $p_0(x)$ is the *null* probability if the sex of successive children in a family behave like independent Bernoulli trials). Our null is therefore H_0: $p(x) = \binom{8}{x}0.505^x(1-0.505)^{8-x}$ for all $x \in \mathcal{D} = \{0, 1, \ldots, 8\}$ versus the alternative that it is some other discrete distribution. The results presented in Table 1.1 summarize the observed frequencies as well as the probabilities and frequencies under the null that the number of children of a given sex is binomially distributed.

We can use the χ^2 goodness-of-fit procedure[6] to test the null that the data were generated by the binomial distribution. To measure how close the observed frequencies are to the expected frequencies, we calculate the statistic $\chi^2_\nu = \sum_{j=1}^{c}(o_j - e_j)^2/e_j$ where o_j and e_j denote observed and expected frequencies under the null, respectively. For our data, the statistic is $\chi^2 = 44.24$. The degrees of freedom here is $\nu = 7$ and equals the number of outcomes ($c = 9$) minus 1 minus the number of estimated parameters (we estimated one parameter, $\hat{\pi} = 0.505$). The critical value at the 5% level of significance is $\chi^2_{1-0.05,7} = 14.07$. The P-value is $1.92e-07$, which is extremely strong evidence against the null, and we would therefore reject the null that the data were generated by a binomial distribution at all conventional levels of significance.

Table 1.1 and Figure 1.1 reveal that the binomial distribution

[6] Tests for goodness-of-fit are used to determine whether a set of data is consistent with a proposed model.

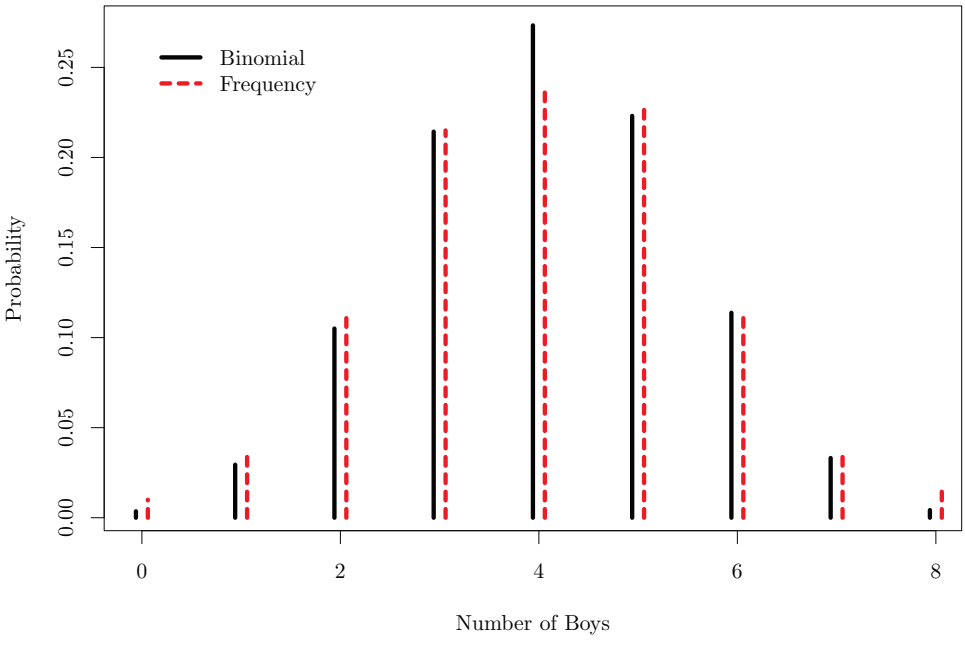

Figure 1.1: Parametric (binomial) versus nonsmooth nonparametric (sample proportion) probability estimates for the number of boys in families with eight children.

predicts more families with equal numbers of boys and girls and fewer families having all boys or all girls than is supported by the data.[7] The dice are stacked against you if you think that "trying again" behaves just like an independent coin toss. The results from the goodness-of-fit test conclusively reject this parametric model for the unknown probability distribution $p(x)$.

This example highlights the dilemma faced by practitioners who wish to model unknown discrete probability distributions. We can always *presume* a functional form for the underlying parametric model and estimate probabilities under this presumption. However, if we entertain the possibility that the parametric model might be misspecified, we would naturally test for correct specification of the presumed parametric model (e.g., test for correct specification of the binomial distribution as we did above). If the parametric model is rejected (as was the case above), then we return to where we began, having ruled out perhaps one of a number of potential parametric distributions. Furthermore, repeatedly testing alternative parametric specifications opens the *pre-test* can of worms, a fact that is often conveniently ignored by practitioners.

[7]Figure 1.1 compares the binomial probabilities with the *frequency* estimator of the probabilities defined in Section 1.3 below.

Against this backdrop, we might instead consider a nonparametric approach, and we shall consider two popular methods. The first is the familiar *frequency* estimator that was used in the illustrative example above (i.e., the *sample proportion*, which computes the *relative frequency* of occurrence and is a *nonparametric nonsmooth* approach). The second is a kernel[8] estimator that smooths a discrete support random variable in a particular manner (i.e., a *nonparametric kernel-smoothed* approach). We now turn our attention to the nonsmooth frequency estimator of the unknown probability function $p(x)$.

1.3 Nonsmooth Probability Function Estimation

All readers will no doubt be aware of an extremely popular nonparametric estimator of unknown probabilities, namely the *sample proportion* $p_n(x)$ defined below, which we refer to as the *nonsmooth* or *frequency* estimator. Students of introductory statistics know that this estimator is unbiased (i.e., $\mathrm{E}\, p_n(x) = p(x)$) and has variance $p(x)(1 - p(x))/n$. It will be instructive to derive these results because familiarity with this proof concept will lend transparency to the proof for the kernel-smoothed approach that we consider afterwards.

Let $X \in \mathcal{D} = \{0, 1, \ldots, c-1\}$ be a discrete random variable having finite support. Suppose that we have a sample of n independent random draws from the probability distribution $p(x)$, denoted $\{X_1, X_2, \ldots, X_n\}$. The univariate frequency estimator of $p(x)$ is the familiar sample proportion given by

$$\begin{aligned} p_n(x) &= \frac{\#X_i \text{ equal to } x}{n} \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i = x), \end{aligned}$$

where "$\#X_i$ equal to x" is simply the number of sample realizations equal to any particular outcome x and where $\mathbf{1}(\cdot)$ is an *indicator function* defined by

$$\mathbf{1}(X_i = x) = \begin{cases} 1 & \text{if } X_i = x \\ 0 & \text{otherwise.} \end{cases}$$

This indicator function is for *counting* and is limited to conducting a binary operation, *equal* or *not equal*. Hence the expression $n^{-1}\sum_{i=1}^{n} \mathbf{1}(X_i = x)$ simply considers each member of the sample of n observations $\{X_1, X_2, \ldots, X_n\}$, assigns to each the value 1 if it equals the particular outcome x and 0 otherwise, adds up all of the 1s and divides by the number of observations n. In the end, this is simply the sample proportion of observations equal to x.

[8] The term *kernel* simply refers to the use of weight functions having particular properties.

Recall that the *expected value* of a discrete random variable is obtained by multiplying each element of the outcome space $\mathcal{D}$ by its probability of occurrence and taking the sum thereof, while the expected value of some *function* of a discrete random variable is obtained by multiplying the function evaluated at each element of the outcome space by its probability of occurrence and taking the sum thereof. The expected value of $p_n(x)$ is therefore given by

$$\begin{aligned}
\mathrm{E}\, p_n(x) &= \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\, \mathbf{1}(X_i = x) \\
&= \mathrm{E}\, \mathbf{1}(X_1 = x) \\
&= \sum_{t\in\mathcal{D}} \mathbf{1}(t = x)p(t) \\
&= \mathbf{1}(x = x)p(x) + \sum_{t\in\mathcal{D}, t\neq x} \mathbf{1}(t = x)p(t) \\
&= 1 \times p(x) + \sum_{t\in\mathcal{D}, t\neq x} 0 \times p(t) \\
&= p(x),
\end{aligned}$$

where the second line follows from the identical distribution assumption (i.e., under identical distributions $\mathrm{E}\, \mathbf{1}(X_1 = x) = \mathrm{E}\, \mathbf{1}(X_2 = x) = \cdots = \mathrm{E}\, \mathbf{1}(X_n = x)$, so $\sum_{i=1}^{n} \mathrm{E}\, \mathbf{1}(X_i = x) = \sum_{i=1}^{n} \mathrm{E}\, \mathbf{1}(X_1 = x) = n \times \mathrm{E}\, \mathbf{1}(X_1 = x)$), and the third line follows from the definition of the expected value of a function of a discrete random variable described above. Moving from the third to the sixth line, note that $t = x$ for only one value of $t \in \mathcal{D}$ (e.g., suppose $x = 2$, although x could be any outcome that we might consider). Hence $\mathbf{1}(t = x)$ equals 1 for $t = x$ and $\mathbf{1}(t = x)p(t) = 1 \times p(x) = p(x)$ for $t = x$, while $\mathbf{1}(t = x)p(t) = 0 \times p(t) = 0$ for the remaining outcomes in $\mathcal{D}$ (i.e., all $t \in \mathcal{D}$ for which $t \neq x$). Since $\mathrm{E}\, p_n(x) = p(x)$, this estimator is clearly *unbiased* (i.e., $\mathrm{Bias}\, p_n(x) = \mathrm{E}\, p_n(x) - p(x) = 0$).

The variance of $p_n(x)$ is given by

$$\begin{aligned}
\mathrm{Var}\, p_n(x) &= \mathrm{E}\left((p_n(x) - \mathrm{E}\, p_n(x))^2\right) \\
&= \mathrm{E}\left(\left(\frac{1}{n}\sum_{i=1}^{n} (\mathbf{1}(X_i = x) - \mathrm{E}\, \mathbf{1}(X_i = x))\right)^2\right) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n} \mathrm{E}\, \eta_i^2 + \sum_{i}\sum_{j,i\neq j} \mathrm{E}\, \eta_i\eta_j\right) \\
&= \frac{1}{n}\,\mathrm{E}\,(\mathbf{1}(X_1 = x) - \mathrm{E}\, \mathbf{1}(X_1 = x))^2 \\
&= \frac{1}{n}\left(\mathrm{E}\, \mathbf{1}^2(X_1 = x) - (\mathrm{E}\, \mathbf{1}(X_1 = x))^2\right)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n}\left(\mathrm{E}\,\mathbf{1}(X_1 = x) - (\mathrm{E}\,\mathbf{1}(X_1 = x))^2\right) \\
&= \frac{1}{n}\left(p(x) - p(x)^2\right) \\
&= \frac{p(x)(1-p(x))}{n},
\end{aligned}
$$

where $\eta_i = \mathbf{1}(X_i = x) - \mathrm{E}\,\mathbf{1}(X_i = x)$, $\mathbf{1}^2(\cdot) = \mathbf{1}(\cdot)$, $\mathrm{E}\,\eta_i\eta_j = 0$ for $i \neq j$ since we assumed i.i.d. draws (*independent and identically distributed*), and $\mathrm{E}\,\mathbf{1}(X_1 = x) = p(x)$ from the previous derivation. To go from line three to line four note that the second term in brackets is 0 since, for i.i.d. draws, $\mathrm{E}\,\eta_i\eta_j = 0$ when $i \neq j$, and note that $\sum_{i=1}^{n} \mathrm{E}\,\eta_i^2 = n\,\mathrm{E}\,\eta_1^2 = n\,\mathrm{E}\,(\mathbf{1}(X_1 = x) - \mathrm{E}\,\mathbf{1}(X_1 = x))^2$ since $\mathrm{E}\,\eta_1^2$ is a constant. To go from line four to line five recall that $\mathrm{E}\left((\hat{\theta} - \mathrm{E}\,\hat{\theta})^2\right)$ $= \mathrm{E}\left(\hat{\theta}^2 - 2\hat{\theta}\,\mathrm{E}\,\hat{\theta} + (\mathrm{E}\,\hat{\theta})^2\right) = \mathrm{E}\,\hat{\theta}^2 - 2\,\mathrm{E}\,\hat{\theta}\,\mathrm{E}\,\hat{\theta} + (\mathrm{E}\,\hat{\theta})^2 = \mathrm{E}\,\hat{\theta}^2 - 2(\mathrm{E}\,\hat{\theta})^2 + (\mathrm{E}\,\hat{\theta})^2$ $= \mathrm{E}\,\hat{\theta}^2 - (\mathrm{E}\,\hat{\theta})^2$ for any estimator $\hat{\theta}$.

The mean square error (MSE) criterion is perhaps the most important criterion used to evaluate the performance of an estimator $\hat{\theta}$ of some population characteristic θ. The MSE reflects the *bias*, *precision* (i.e., variance), and overall *accuracy* in statistical estimation as a function of the sample size, and is defined as $\mathrm{E}\left((\hat{\theta} - \theta)^2\right)$. Recalling that the MSE of an estimator can be expressed as its variance plus the square of its bias, the MSE of $p_n(x)$ is given by

$$
\begin{aligned}
\mathrm{MSE}\,p_n(x) &= \mathrm{Var}\,p_n(x) + (\mathrm{Bias}\,p_n(x))^2 \\
&= \frac{p(x)(1-p(x))}{n} + 0^2 \\
&= \frac{p(x)(1-p(x))}{n}.
\end{aligned}
$$

This is of *large order of magnitude* $O(n^{-1})$ and small order $o(1)$. Hence its *root MSE* (i.e., $\sqrt{\mathrm{MSE}}$) is of $O\left(n^{-1/2}\right)$ and $o(1)$, which is the familiar rate of convergence typically associated with *correctly specified* parametric models. In other words, it is *root-n-consistent* (note that an incorrectly specified parametric model has a bias term that never vanishes, hence such estimators are *inconsistent*).[9]

This tells us that this estimator is of large order in probability $O_p(n^{-1/2})$ and small order in probability $o_p(1)$, i.e.,

$$
p_n(x) - p(x) = O_p\left(n^{-1/2}\right) = o_p(1)
$$

(see Appendix A for an overview of orders of magnitude and probability).

[9] MSE is measured in units of X *squared*, while $\sqrt{\mathrm{MSE}}$ is measured in the same units as X - either can be reported. To say that an estimator has MSE of $O(n^{-1})$ simply means that the MSE is proportional to $1/n$ (MSE is non-stochastic).