1
Introduction

Everyone is interested in records, weather records, sports records, crime statistics, and so on. Record values are kept for almost every conceivable phenomenon. What was the coldest day last year (or ever), which city has the lowest crime rate, what was the shortest time recorded to complete a marathon, who holds the record in eating the most number of hot dogs in the shortest period, what was the highest stock value thus far? The list could go on and on; there is even a book that lists all kinds of records broken during a given year—the well-known *Guinness Book of World Records*! Naturally, if there is a subject concerning statistical values that interests the majority of people in the world, it has to be of interest to statisticians. However, how does one relate record values to statistical theory? The easiest way to explain this is with some examples. To begin with, consider a sports event: Not only do we want to know who holds the record for running 100 meters in the Olympics, but we also want to *predict* the next record-breaking time. Similarly, we want to determine if Miami, Florida, will still have the highest auto theft rate next year, or will Los Angeles still be the most polluted city in the US next year. Or we would like to predict the next highest closing price of a particular stock. In all of these examples, we want to use past data to predict the future. And prediction of the future using past data requires statistical theory.

Besides arising naturally in our day-to-day activities, observing record values also has a place in destructive stress testing and industrial quality control experiments. In these experiments it is often of interest to estimate a guarantee value or a population quantile. Generally, we would do this by observing the entire sample and then using the appropriate order statistic to estimate the guarantee value or the quantile of interest. Instead, we can observe the sample sequentially and record only successive minimum or maximum values. Then, measurements are made only at "record-setting" items, and the total number of measurements made is considerably smaller than n, the total sample size. It turns out that we can still estimate the guarantee value from these record-setting measurements. This strategy is extremely useful when items are available for testing, after they are manufactured,

but before they are shipped out. Let us say that we have a shipment of wooden beams and we want to make an inference about the breaking strength of these beams. We take a sample of fifty beams to test their breaking strength. In classical sampling, we would destroy all fifty beams. But in the setting of record-breaking data, here is what we would do: We stress the first beam until it breaks and record the breaking stress. The next beam is then stressed only up to the level that broke the first one. If it does not break, we move on to the third beam and stress it up to the value that broke the first beam. If the second beam breaks, its breaking stress is recorded; then we stress the third beam only up to the value that broke the second one. As is obvious, the data will consist of lower and lower breaking stress values. Moreover, the total number of beams broken will surely be less than fifty. Also as mentioned earlier, we will still be able to estimate the required quantile or guarantee value based on the statistical theory of successive minima. Of course, besides estimating the guarantee value, one may want to predict a future record, estimate underlying parameters, or estimate the underlying probability distribution function of the variable being measured. These and other problems have given rise to a plethora of papers and books on record-breaking data.

However, although record values have been around forever, "record-breaking data" as it is called, is relatively new to the field of statistics, owing its birth to Chandler in 1952. Chandler (1952) studied the stochastic behavior of random record values arising from the "classical record model," that is, the record model where the underlying sample from which records are observed is considered to consist of independent identically distributed observations from a continuous probability distribution. Among the many properties that Chandler established for the random record sequence, perhaps the most important and somewhat surprising one was that the expected value of the waiting time between records has infinite expectation. Chandler's work was followed by that of Dwass (1960) and Renyi (1962), who established limit theorems for some of the sequences associated with record-breaking data. Dwass (1964) studied the frequency of records indexed by i, $an \leq i \leq bn$ and showed that this frequency is asymptotically a Poisson count with mean $\ln(b/a)$. Afterward, the subject of record values caught the attention of several mathematicians and statisticians, and work on it increased tremendously. There have been numerous articles on moments of records, characterizations, inference from records, and the like. One only has to survey the recent statistical literature to note the fairly large volume of work that is still being carried out in this field. Also, statisticians have started moving away from the classical model. There are a number of situations where due to improvements in technology or techniques, the underlying population may have a trend in it. Hence the classical record-breaking model will not provide an adequate explanation for these data. In fact, as soon as

the basic "independent, identically distributed" assumption for the record model of Chandler is extended to better reflect reality, the problem becomes much harder. A perfect example here is the field of sports. Improved training techniques, diet, health care, and so on, all lead to better performances. Thus the simple record model cannot explain sports records due perhaps to the changing underlying population. Yang (1975), Ballerini and Resnick (1985), and Smith (1988) are just a few of the authors who have moved away from the simple model and have studied models that allow for a changing population.

Although the literature on record values is not enormous compared to other subject areas in statistics, today there are over 300 papers and several books published on record-breaking data. With such a volume of work on records and record-breaking data, it is imperative that related results be brought together in one place. That has been the purpose of most of the books published on record-breaking data and that is also the purpose of this book. Most of the earlier literature on this topic has focused on the stochastic behavior of records, prediction of future record values, and characterization problems. Inference, both parametric and nonparametric, followed later. The manuscripts on record-breaking data have also followed the same trend. With the exception of the book by Arnold et al. (1998), all other books have focused on the stochastic behavior of records, characterizations, and prediction. Arnold et al. (1998) presented a comprehensive review of most of the results on records, including a chapter devoted to the results on inference from record-breaking data. However, that chapter is somewhat brief. The authors focused mainly on estimation of parameters from record-breaking data and not on the general problem of parametric and nonparametric inference from such data. Gulati and Padgett (1994d) gave a brief survey on estimation from such data up until that time.

The purpose of this present monograph is then to fill the gap mentioned above. We focus on cataloging the results on nonparametric inference from record-breaking data. The general problem of parametric and nonparametric inference from record-breaking data has its birth in two articles by Samaniego and Whitaker (1986, 1988). In the first paper (Samaniego and Whitaker, 1986), they develop and study the properties of the maximum likelihood estimator of the mean of an underlying exponential distribution. In their 1988 paper, however, they use record-breaking data to develop a nonparametric maximum likelihood estimator of the underlying distribution function. Under repeated sampling, the nonparametric maximum likelihood estimator is shown to be strongly consistent and asymptotically normal. Their estimator has since been used to develop and study properties of smooth nonparametric function estimates by Gulati and Padgett (1992, 1994, 1995), among others. Besides function estimation from record-breaking

data, there are also results on distribution-free tests and nonparametric prediction from such data, as well as a paper on nonparametric Bayesian estimation from record-breaking data (Tiwari and Zalkikar, 1991). All of these results are catalogued here.

In view of the purpose of the monograph, the layout of the book follows. First, the problem of record-breaking data is defined and the notation introduced. This is done in Chapter 2. We also present a summary of the stochastic results on record-breaking data in Chapter 2. Because of the existence of several manuscripts on stochastic results from record-breaking data, and especially in light of the recent monograph by Arnold et al. (1998), the presentation on stochastic results is somewhat brief. In Chapter 3, we discuss some of the major results on parametric inference from such data. Expressions for the estimates of the parameters for various distributions can be found in the book by Arnold et al. (1998). Hence they are not repeated here. Work along the lines of Samaniego and Whitaker (1986), however, has not been discussed by Arnold et al. (1998) and therefore is presented in Chapter 3. The main emphasis of this book, however, begins in Chapter 4. There we tabulate and discuss all the known work on nonparametric inference from such data, starting with the distribution-free tests of Foster and Stuart (1954), leading up to Samaniego and Whitaker's work. Later chapters present some details of the work done in nonparametric function estimation and other results in more recent years. Finally, we consider models that incorporate trend and give a brief outline of some of the work done there.

2
Preliminaries and Early Work

How does one describe record-breaking data in a statistical framework? There are several models for such data, and the classical record model is described here. This model arises, for example, in industrial quality control experiments and destructive stress testing, where one records successive minimum values. As mentioned in Chapter 1, in such experiments one is often interested in estimating a guarantee value or a population quantile. In the classical record model this is done by observing the data sequentially and recording only successive minimum values (since the quantile of interest is normally a lower quantile). Thus, one measures only "record-setting" items and in general, the number of measurements made is considerably smaller than the total sample size. This "measurement savings" is important when the measurement process is costly, time consuming, or destructive.

Consider again the wooden beam example. Suppose a building code prohibits the use of a particular type of beam unless it has probability at least 0.95 of surviving some severe stress, x (see Glick, 1978). In other words, the fifth percentile $x_{0.05}$ should satisfy $x_{0.05} \geq x$. Since it is always better to underestimate the percentile than overestimate it, one considers the smallest failure stress observed in laboratory testing. It is safe to assume that for a large sample, this point will lie below the distribution's fifth percentile. In fact from Glick (1978), the breaking stress of the weakest item in a sample of size 90 lies below the fifth percentile with a probability of 0.99; that is, this minimum value will be the 0.99 tolerance limit for the fifth percentile of the distribution. So we may take a random sample of 90 beams, with our goal being the measurement of the breaking stress of the weakest beam. We want to destroy only a few of the beams and so record sampling is one way to measure the weakest beam. As mentioned in Chapter 1, the breaking stress value of the first beam is our first record value. Thereafter, successive beams are stressed only up to the value at which the previous breakage occurred, with smaller and smaller breaking stress values being recorded until the sample has been exhausted. Note, of

course, that the last breaking stress value recorded will be the breaking stress of the weakest item and the estimate of the required tolerance limit. Moreover, on average, we will destroy only about 5 beams (Glick, 1978), leaving the remainder intact for shipment.

The verbal description above is now quantified in a more exact mathematical framework. Then some stochastic properties and statistical characterizations are briefly summarized in the remainder of this chapter.

2.1 Notation and Terminology

The notation and basic statistical framework for the record values (successive minima) is now introduced. Let $Y_1, Y_2, \ldots$ be a random sample from a continuous cumulative distribution function (c.d.f.) F with density function f. Then, since only successive minimum values are recorded, the observed data consist of the sequence $X_1, K_1, X_2, K_2, \ldots, X_r, K_r$, where $X_1 = Y_1$, X_i, $i = 2, 3, \ldots, r$, is the ith new minimum, and K_i is the number of trials following the observation of X_i to obtain a new record (or to exhaust all available observations in the case of $i = r$, K_r). The sampling schemes for generating these data are:

1) Data are obtained via the *inverse sampling scheme*, where items are presented sequentially and sampling is terminated when the rth record is observed. In this case, the total number of items sampled N_r is a random variable and K_r is defined to be 1 for convenience.

2) Records are obtained under the *random sampling scheme*, that is, a random sample, $Y_1, Y_2, \ldots, Y_n$, from c.d.f. F is examined sequentially and successive minimum values are recorded. For this sampling scheme the number of records R_n obtained is a random variable and, given a value of r, $\sum_{i=1}^{r} K_i = n - 1$.

To understand the terminology better, we look at another example. Consider the process of measuring the thickness of a manufactured item using a micrometer. In order to measure the minimum thickness of $n = 100$ items one first measures an item at random. The thickness of this item then is $X_1 = Y_1$. The gap in the micrometer created by the first item serves as a standard in judging subsequent items, and a new measurement is made only if a subsequent item fits inside this gap. Hence, if the second measurement is made at the sixth trial, then $X_2 = Y_6$, $K_1 = 5$, and $N_2 = 6$ (note that N_1 is always equal to one). Now the gap created by this sixth item serves as a standard for judging subsequent items. Once again, by using this method, the number of actual measurements

made will be substantially less than 100, and yet it will serve equally well in determining the minimum thickness.

Regardless of the sampling scheme, we define the following sequences: $\{X_i, 1 \le i \le r\}$ is the *record value sequence*, $\{N_i, 1 \le i \le r\}$ is defined to be the *record time sequence* (note that $N_1 = 1$ by default), and finally, $\{K_i, 1 \le i \le r\}$ is the *interrecord time sequence*. With the above notation and terminology, we have what is called the *Classical* Record Model.

2.2 Stochastic Behavior

The stochastic behavior of the classical record model was first studied by Chandler (1952) who showed that the record times (N_is) and the interrecord times (K_is) both had infinite expectation, although the mode of the K_is was one. Chandler also gave an expression for the joint distribution of $X_1, X_2, \ldots, X_r$, and obtained tables for the percentile points for X_i for $i = 1, 2, \ldots, 9$, for the normal and the rectangular distributions.

The fact that N_2 has an infinite expectation discouraged many statisticians from working on record-breaking data for a while (see Galambos, 1978). Development began again with Dwass (1960) and Renyi (1962) who gave "strong law of large numbers"- and "central limit theorem"-type results for the R_ns and the N_is. Dwass (1964) also showed the frequency of the record highs among the observations indexed by i, $an \le i \le bn$, where n is the sample size and $0 < a < b$ is asymptotically a Poisson count with mean $\ln(b/a)$. Since Dwass, a number of statisticians have investigated the behavior of record values. (See Neuts, 1967, Resnick, 1973(a,b,c), and Shorrock, 1972(a,b), 1973, for some of the articles on the subject.) Glick (1978) gave an informal summary of results to that date. Review articles have also been written by Galambos (1978), Nagaraja (1988), and Nevzorov (1987). In addition, the topic of records and record-breaking data has been discussed in a number of books. Besides the very thorough and comprehensive work by Arnold et al. (1998), work on records has been reviewed in the books by Galambos (1987), Resnick (1987), and Ahsanullah (1995). What follows next is a brief summary of the stochastic results about the number of records R_n, record values the X_is, the interrecord times K_is, and the record times N_is. For details of these results, the reader is referred to the book by Arnold et al. (1998).

Record Values

a) The joint distribution of $X_1, X_2, \ldots, X_r$ is given by the probability density function (Chandler, 1952, and Glick, 1978)

$$g(x_1, x_2, \ldots, x_r) = f(x_r) \prod_{i=1}^{r-1} \frac{f(x_i)}{1-F(x_i)}.$$

b) Define $H(y) = -\ln(1 - F(y))$. Then we have the large sample results that (Resnick, 1973a)

i) $[H(x_r) - r]/\sqrt{r}$ is asymptotically normal with mean 0 and variance 1, and

ii) $H(x_r)/\sqrt{r}$ converges to one with probability one.

In fact, Resnick (1973a) characterized the three types of limit distributions to which the record values X_r can converge as $r \to \infty$. Depending on the underlying distribution F, a record value sequence satisfies exactly one of the following convergences in distribution as $r \to \infty$.

i) $$P\left\{\frac{X_{N_r} - G^{-1}(r)}{G^{-1}(r + \sqrt{r}) - G^{-1}(r)} < x\right\} \to \Phi(x),$$

ii) $$P\left\{\frac{X_{N_r}}{G^{-1}(r)} < x\right\} \to \begin{cases} 0, & x < 0 \\ \Phi(\alpha \ln x)), & x \geq 0 \end{cases}$$

and

iii) $$P\left\{\frac{X_{N_r} - \bar{x}}{\bar{x} - G^{-1}(r)} < x\right\} \to \begin{cases} \Phi(-\alpha \ln(-x)), & x < 0 \\ 1, & x \geq 0, \end{cases}$$

where $\Phi(x)$ denotes the standard normal distribution function.

Frequency of Records

Results for the statistical behavior of the number of records R_n in a random sample of size n are listed next:

a) $E(R_n) = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}$ (Glick, 1978);

b) $\mathrm{Var}(R_n) = \sum_{i=1}^{n} \frac{1}{i} - \sum_{i=1}^{n} \frac{1}{i^2}$ (Glick, 1978).

c) $[\ln(R_n) - n]/n^{1/2}$ is asymptotically normal with mean zero and variance one (Resnick, 1973c).

d) $R_n/\ln(n)$ converges to one with probability one (or almost surely, a.s.) as the sample size $n \to \infty$ (Renyi, 1962). That is, as n increases, about $\ln(n)$ records will be found in a random sample of n values.

e) The frequency of record highs among the observations indexed by i, $an \le i \le bn$ $(0 < a < b)$ is asymptotically a Poisson count with mean $\ln(b/a)$ (Dwass, 1964).

Record Times

The record times N_i also have interesting stochastic properties, as described next.

a) The value of N_i does not depend on the underlying distribution function (Chandler, 1952).

b) $[\ln(N_r) - r]/\sqrt{r}$ is asymptotically normal with mean 0 and variance 1 (Renyi, 1962, Resnick, 1973c).

c) $N_r/\ln(r)$ converges to one with probability one as the number of records $r \to \infty$ (Renyi, 1962, Dwass, 1960, and Galambos, 1978).

d) The distribution of the ratio N_r/N_{r+1} is asymptotically uniform over the unit interval (Tata, 1969).

e) The successive ratios N_r/N_{r+1}, N_{r+1}/N_{r+2}, . . . are asymptotically independent uniform variates (Shorrock, 1972b and Resnick, 1973c).

Waiting Time Between Records

For the interrecord time sequence, perhaps the most surprising result is the infinite expected value, the first property below. The asymptotic distributional behavior of K_r is similar to that of N_r.

a) $E(K_i) = \infty$ for all i, although the mode of the K_is is 1 (Chandler, 1952).

b) $[\ln(K_r) - r]/\sqrt{r}$ is asymptotically normal with mean 0 and variance 1 (Neuts, 1967).

c) $K_r/\ln(r)$ converges to one with probability one as $r \to \infty$ (Neuts, 1967).

In addition to the brief summary presented above of some of the main results on records, it can be shown that the record time sequence, the record-value sequence, and the sequence of the number of records each forms a Markov chain. There are also results on the moments of the record value sequence (both parametric and nonparametric) and on characterizations. Once again, the reader is referred to the book by Arnold et al. (1998) for a detailed discussion on all of these results.