7

SAMPLING AND DESCRIPTIVE STATISTICS

The distinction between Probability and Statistics is somewhat fuzzy, but largely has to do with the perspective of what is known versus what is to be determined. One may think of Probability as the study of models for (random) experiments when the model is fully known. When the model is not fully known and one tries to infer about the unknown aspects of the model based on observed outcomes of the experiment, this is where Statistics enters the picture. In this chapter we will be interested in problems where we assume we know the outputs of random variables, and wish to use that information to say what we can about their (unknown) distributions.

Suppose, for instance, we sample from a large population and record a numerical fact associated with each selection. This may be recording the heights of people, recording the arsenic content of water samples, recording the diameters of randomly selected trees, or anything else that may be thought of as repeated, random measurements. Sampling an individual from a population in this case may be viewed as a random experiment. If the sampling were done at random with replacement with each selection independent of any other, we could view the resulting numerical measurements as i.i.d. random variables $X_1, X_2, \ldots, X_n$. A more common situation is sampling without replacement, but we have previously seen (See Section 2.3) that when the sample size is small relative to the size of the population, the two sampling methods are not dramatically different. In this case we have the results of n samples from a distribution, but we don't actually know the distribution itself. How might we use the samples to attempt to predict such things as expected value and variance?

7.1 THE EMPIRICAL DISTRIBUTION

A natural quantity we can create from the observed data, regardless of the underlying distribution that generated it, is a discrete distribution that puts equal probability on each observed point. This distribution is known as the empirical distribution. Some values of X_i can of course be repeated, so the empirical distribution is formally defined as follows.

DEFINITION 7.1.1. *Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables. The "empirical distribution" based on these is the discrete distribution with probability mass function given by*

$$f(t) = \frac{1}{n}\#\{X_i = t\}.$$

We can now study the empirical distribution using the tools of probability. Doing so does not make any additional assumptions about the underlying distribution, and inferences about it based on the empirical distribution are traditionally referred to as "descriptive statistics". In later chapters, we will see that making additional assumptions lets us make "better" inferences, provided the additional assumptions are valid.

It is important to realize that the empirical distribution is itself a random quantity, as each sample realisation will produce a different discrete distribution. We intuitively expect it to carry information about the underlying distribution, especially as the sample size n grows. For example, the expectation computed from the empirical distribution should be closely related to the true underlying expectation, probabilities of events computed from the empirical distribution should be related to the true probabilities of those events, and so on. In the remainder of this chapter, we will make this intuition more precise and describe some tools to investigate the properties of the empirical distribution.

7.2 DESCRIPTIVE STATISTICS

7.2.1 *Sample Mean*

Given a sample of observations, we define the sample mean to be the familiar definition of average.

DEFINITION 7.2.1. *Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables. The "sample mean" of these is*

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

It is easy to see that $\overline{X}$ is the expected value of a random variable whose distribution is the empirical distribution based on $X_1, X_2, \ldots, X_n$ (see Exercise 7.2.4). Suppose the X_j random variables have a finite expected value μ. The sample mean $\overline{X}$ is not the same as this expected value. In particular μ is a fixed constant while $\overline{X}$ is a random variable. From the statistical perspective, μ is usually assumed to be an unknown quantity while $\overline{X}$ is something that may be computed from the results of the sample $X_1, X_2, \ldots, X_n$. How well does $\overline{X}$ work as an estimate of μ? The next theorem begins to answer this question.

THEOREM 7.2.2. *Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample of random variables whose distribution has finite expected value μ and finite variance σ^2. Let $\overline{X}$ represent the sample mean. Then*

$$E[\overline{X}] = \mu \quad \text{and} \quad SD[\overline{X}] = \frac{\sigma}{\sqrt{n}}.$$

Proof -

$$\begin{aligned} E[\overline{X}] &= E[\frac{X_1 + X_2 + \cdots + X_n}{n}] \\ &= \frac{E[X_1] + E[X_2] + \cdots + E[X_n]}{n} \\ &= \frac{n\mu}{n} = \mu \end{aligned}$$

To calculate the standard deviation, we consider the variance and use Theorem 4.2.6 and Exercise 6.1.12 to obtain

$$\begin{aligned} Var[\overline{X}] &= Var[\frac{X_1 + X_2 + \cdots + X_n}{n}] \\ &= \frac{Var[X_1] + Var[X_2] + \cdots + Var[X_n]}{n^2} \\ &= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \end{aligned}$$

Taking square roots then shows $SD[\overline{X}] = \frac{\sigma}{\sqrt{n}}$. ■

The fact that $E[\overline{X}] = \mu$ means that, on average, the quantity $\overline{X}$ is accurately describing the unknown mean μ. In the language of statistics $\overline{X}$ is said to be an "unbiased estimator" of the quantity μ. Note also that $SD[\overline{X}] \to 0$ as $n \to \infty$ meaning that the larger the sample size, the more accurately $\overline{X}$ reflects its average of μ. In other words, if there is an unknown distribution from which it is possible to sample, averaging a large sample should produce a value close to the expected value of the distribution. In technical terms, this means that the sample mean is a "consistent estimator" of the population mean μ.

7.2.2 *Sample Variance*

Given a sample of observations, we define the sample variance below.

DEFINITION 7.2.3. *Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables. The "sample variance" of these is*

$$S^2 = \frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n-1}.$$

Note that this definition is not universal; it is common to define sample variance with n (instead of $n-1$) in the denominator, in which case the definition matches the variance of the empirical distribution of $X_1, X_2, \ldots, X_n$ (Exercise 7.2.4). The definition given here produces a quantity that is unbiased for the underlying population variance, a fact that follows from the next theorem.

THEOREM 7.2.4. *Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample of random vairables whose distribution has finite expected value μ and finite variance σ^2. Then S^2 is an unbiased estimator of σ^2, i.e.*

$$E[S^2] = \sigma^2.$$

Proof - First note that

$$E[\overline{X}^2] = Var[\overline{X}] + (E[\overline{X}])^2 = \frac{\sigma^2}{n} + \mu^2$$

whereas

$$E[X_j^2] = Var[X_j] + E[X_j]^2 = \sigma^2 + \mu^2.$$

Now consider the quantity $(n-1)S^2$.

$$\begin{aligned} E[(n-1)S^2] &= E[(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2] \\ &= E[X_1^2 + X_2^2 + \cdots + X_n^2] - 2E[(X_1 + X_2 + \cdots + X_n)\overline{X}] \\ &\quad + E[\overline{X}^2 + \overline{X}^2 + \cdots + \overline{X}^2] \end{aligned}$$

But $X_1 + X_2 + \cdots + X_n = n\overline{X}$, so

$$\begin{aligned} E[(n-1)S^2] &= E[X_1^2 + X_2^2 + \cdots + X_n^2] - 2nE[\overline{X}^2] + nE[\overline{X}^2] \\ &= E[X_1^2 + X_2^2 + \cdots + X_n^2] - nE[\overline{X}^2] \\ &= n(\sigma^2 + \mu^2) - n(\frac{\sigma^2}{n} + \mu^2) \\ &= (n-1)\sigma^2 \end{aligned}$$

Dividing by $n-1$ gives the desired result, $E[S^2] = \sigma^2$. ■

A more important property (than unbiasedness) is that S^2 and its variant with n in the denominator are both "consistent" for σ^2, just as $\overline{X}$ was for μ, in the sense that $Var[S^2] \to 0$ as $n \to \infty$ under some mild conditions.

7.2.3 *Sample proportion*

Expectation and variance are commonly used summaries of a random variable, but they do not characterize its distribution completely. In general, the distribution of a random variable X is fully known if we can compute $P(X \in A)$ for any event A. In particular, it is enough to know probabilities of the type $P(X \leq t)$, which is precisely the cumulative distribution function of X evaluated at t.

Given a sample of i.i.d. observations $X_1, X_2, \ldots, X_n$ from a common distribution defined by a random variable X, the probability $P(X \in A)$ of any event A has the natural sample analog $P(Y \in A)$, where Y

is a random variable following the empirical distribution based on $X_1, X_2, \ldots, X_n$. To understand this quantity, recall that Y essentially takes values $X_1, X_2, \ldots, X_n$ with probability $1/n$ each, and so we have

$$P(Y \in A) = \sum_{X_i \in A} \frac{1}{n} = \frac{\#\{X_i \in A\}}{n}$$

In other words, $P(Y \in A)$ is simply the proportion of sample observations for which the event A happened. Not surprisingly, $P(Y \in A)$ is a good estimator of $P(X \in A)$ in the following sense.

THEOREM 7.2.5. *Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample of random variables with the same distribution as a random variable X, and suppose that we are interested in the value $p = P(X \in A)$ for an event A. Let*

$$\hat{p} = \frac{\#\{X_i \in A\}}{n}.$$

Then, $E(\hat{p}) = P(X \in A)$ and $Var(\hat{p}) \to 0$ as $n \to \infty$.

Proof - Let

$$Y = \#\{X_i \in A\} = \sum_{i=1}^{n} Z_i,$$

where for $1 \leq i \leq n$,

$$Z_i = \begin{cases} 1 & \text{if } X_i \in A \\ 0 & \text{otherwise} \end{cases}$$

It is easy to see $P(Z_i = 1) = P(X_i \in A) = p$. Further Z_i's are independent because X_i's are independent (See Theorem 3.3.6 and Exercise 7.2.1). Thus, Y has the Binomial distribution with parameters n and p, with expectation np and variance $np(1-p)$. It immediately follows that

$$E(\hat{p}) = E(Y/n) = p \text{ and } \mathrm{Var}(\hat{p}) = p(1-p)/n$$

which has the limiting value 0 as $n \to \infty$. ■

This result is a special case of the more general "law of large numbers" we will encounter in Section 8.2. It is important because it gives formal credence to our intuition that the probability of an event measures the limiting relative frequency of that event over repeated trials of an experiment.

DEFINITION 7.2.6. *In terms of our notation above, the analog of the cumulative distribution function of X is the cumulative distribution function of Y, which is traditionally denoted by*

$$\hat{F}_n(t) = P(Y \leq t) = \frac{\#\{X_i \leq t\}}{n}$$

and known as the "empirical cumulative distribution function" or ECDF of $X_1, X_2, \ldots, X_n$.

EXERCISES

Ex. 7.2.1. Verify that the proofs of Theorem 3.3.5 and Theorem 3.3.6 hold for continuous random variables.

Ex. 7.2.2. Let X and Y be two continuous random variables having the same distribution. Let $f : \mathbb{R} \to \mathbb{R}$ be a piecewise continuous function. Then show that $f(X)$ and $f(Y)$ have the same distribution.

Ex. 7.2.3. Verify Exercise 7.2.2 for discrete random variables.

Ex. 7.2.4. Let P be the empirical distribution defined by sample observations $X_1, X_2, \ldots, X_n$. In other words, P is the discrete distribution with probability mass function given in Definition 7.1.1. Let Y be a random variable with distribution P.

(a) Show that $E(Y) = \overline{X}$.

(b) Show that $Var(Y) = \frac{n-1}{n} S^2$.

Ex. 7.2.5. Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with finite expectation μ, finite variance σ^2, and finite $\gamma = E(X_1 - \mu)^4$. Compute $Var(S^2)$ in terms of μ, σ^2, and γ and show that $Var(S^2) \to 0$ as $n \to \infty$.

Ex. 7.2.6. Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with finite expectation μ and finite variance σ^2. let $S = \sqrt{S^2}$, the non-negative root of the sample variance. The quantity S is called the "sample standard deviation". Although $E[S^2] = \sigma^2$, it is not true that $E[S] = \sigma$. In other words, S is not an unbiased estimator for σ. Follow the steps below to see why.

(a) Let Z be a random variable with finite mean and finite variance. Prove that $E[Z^2] \geq E[Z]^2$ and give an example to show that equality may not hold. (Hint: Consider how these quantities relate to the variance of Z).

(b) Use (a) to explain why $E[S] \leq \sigma$ and give an example to show that equality may not hold.

7.3 SIMULATION

The preceding discussion gives several mathematical statements about random samples, but it is difficult to develop any intuition about what these statements mean unless we look at actual data. Data is of course abundant in our world, and we will look at some real life data sets later in this book. However, the problem with real data is that we do not usually know for certain the random variable that generated it. To hone our intuition, it is therefore useful to be able to generate random samples from a distribution we specify. The process of doing so using a computer program is known as "simulation".

Simulation is not an easy task, because computers are by nature not random. Simulation is in fact not a random process at all; it is a completely deterministic process that tries to mimic randomness. We will not go into how simulation is done, but simply use R to obtain simulated random samples.

R supports simulation from many distributions, including all the ones we have encountered. The general pattern of usage is that each distribution has a corresponding function that is called with the sample size an argument, and further arguments specifying parameters. The function returns the simulated observations as a vector. For example, 30 Binomial$(100, 0.75)$ samples can be generated by

```
> rbinom(30, size = 100, prob = 0.75)

 [1] 74 84 87 75 69 71 80 75 79 68 72 75 78 75 76 78 82 70 74 76 74 77 70 73 76
[26] 70 70 76 72 77
```

We usually want to do more than just print simulated data, so we typically store the result in a variable and make further calculations with it; for example, compute the sample mean, or the sample proportion of cases where a particular event happens.

```
> x <- rbinom(30, size = 100, prob = 0.75)
> mean(x)

[1] 73.63333

> sum(x >= 75) / length(x)

[1] 0.4333333
```

R has a useful function called `replicate` that allows us to repeat such an experiment several times.

```
> replicate(15, {
+     x <- rbinom(30, size = 100, prob = 0.75)
+     mean(x)
+ })

 [1] 73.23333 75.53333 74.50000 75.46667 75.36667 75.63333 73.53333 74.66667
 [9] 75.43333 73.96667 74.40000 75.16667 74.40000 74.16667 75.30000
```

```
> replicate(15, {
+     x <- rbinom(30, size = 100, prob = 0.75)
+     sum(x >= 75) / length(x)
+ })
```

```
 [1] 0.5000000 0.4333333 0.8666667 0.5333333 0.6000000 0.5000000 0.5666667
 [8] 0.5333333 0.6333333 0.4000000 0.5333333 0.5333333 0.5666667 0.5333333
[15] 0.4666667
```

This gives us an idea of the variability of the sample mean and sample proportion computed from a sample of size 30. We know of course that the sample mean has expectation $100 \times 0.75 = 75$, and we can compute the expected value of the proportion using R as follows.

```
> 1 - pbinom(74, size = 100, prob = 0.75)
```

```
[1] 0.5534708
```

So the correponding estimates are close to the expected values, but with some variability. We expect the variability to go down if the sample size increases, say, from 30 to 3000.

```
> replicate(15, {
+     x <- rbinom(3000, size = 100, prob = 0.75)
+     mean(x)
+ })
```

```
 [1] 75.00300 75.11233 74.95167 74.99033 75.06167 74.96633 74.86000 74.94633
 [9] 75.08333 74.92700 75.03167 75.02633 75.05000 74.95467 75.03167
```

```
> replicate(15, {
+     x <- rbinom(3000, size = 100, prob = 0.75)
+     sum(x >= 75) / length(x)
+ })
```

```
 [1] 0.5706667 0.5780000 0.5433333 0.5440000 0.5863333 0.5426667 0.5496667
 [8] 0.5440000 0.5516667 0.5486667 0.5423333 0.5480000 0.5526667 0.5403333
[15] 0.5573333
```

Indeed we see that the estimates are much closer to their expected values now.

We can of course replicate this process for other events of interest, and indeed for many other distributions. We will see in the next section how we can simulate observations following the normal distribution using the funtion `rnorm`, and the exponential distribution using the funtion `rexp`. It is also interesting to think about how one can simulate observations from a given distribution when a function to do so is not already available. The following examples explore some simple approaches.

EXAMPLE 7.3.1. When trying to formulate a method to simulate random variables from a new distribution, it is customary to assume that we already have a method to generate random variables from Uniform$(0, 1)$. Let us see this can be used to generate random observations from a Poisson(λ) distribution using its probability mass function.

Let X denote an observation from the Poisson(λ) distribution, and $U \sim$ Uniform$(0, 1)$. Denote $p_i = P(X = i)$. The basic idea is as follows:

$$\begin{aligned} & & p_0 &= P(U \le p_0) \\ P(U \le p_0 + p_1) = p_0 + p_1 & \Rightarrow & p_1 &= P(p_0 < U < p_0 + p_1) \\ P(U \le p_0 + p_1 + p_2) = p_0 + p_1 + p_2 & \Rightarrow & p_2 &= P(p_0 + p_1 < U < p_0 + p_1 + p_2) \end{aligned}$$

and so on. Thus, if we set Y to be 0 if $U \le p_0$, and k if U satisfies $\sum_{i=0}^{k-1} p_i < U < \sum_{i=0}^{k} p_i$, then Y has the same distribution as X.

To use this idea to generate 50 observations from Poisson(5), we can use the following code in R, noting that $\sum_{i=0}^{k} p_i = P(X \leq k)$.

```
> replicate(50,
+         {
+             U <- runif(1)
+             Y <- 0
+             while (U > ppois(Y, lambda = 5)) Y <- Y + 1
+             Y
+         })
```

```
 [1]  4  8  3  4  7  3  7  7  5  3  4  5  1  2  8  6  3  4  2  8  5  7  2  4  4
[26]  3  4  8  5  6  8  3  7  9  5  5  5  7  8  4  5  3  3  8  2  8  2  7  8 14
```

Of course, there is nothing in this procedure that is specific to the Poisson distribution. By replacing the call to `ppois()` suitably, the same process can be used to simulate random observations from any discrete distribution supported on the non-negative integers. ■

EXAMPLE 7.3.2. The process described in the previous example cannot be used for continuous random variables. In such cases, Lemma 5.3.7 oftens proves useful. The first part of the lemma states that if $U \sim$ Uniform $(0, 1)$, and F_X is the distribution function of a continuous random variable X, then $Y = F_X^{-1}(U)$ has the same distribution as X. This can be used to generate observations from X provided we can compute F_X^{-1}.

Consider the case where we want X to have the Exp(1) distribution. Then, $F_X(x) = 1 - e^{-x}$ for $x > 0$. Solving for $F_X(x) = u$, we have

$$\begin{aligned} 1 - e^{-x} &= u \\ \Rightarrow e^{-x} &= 1 - u \\ \Rightarrow x &= -\log(1 - u), \end{aligned}$$

that is, $F_X^{-1}(u) = -log(1 - u)$. Thus, we can simulate 50 observations from the Exp(1) distribution using the following R code.

```
> -log(1 - runif(50))
```

```
 [1] 0.17983033 0.59899225 0.39765691 0.46661641 1.83186881 0.75753630
 [7] 0.15224550 3.01323320 0.02324019 2.62589324 0.50319325 0.06495110
[13] 1.73626921 0.79253356 0.46701605 1.31246443 1.94788764 0.32681347
[19] 0.96975851 0.52949759 0.74217408 0.85115821 0.04679527 0.35540345
[25] 0.25261271 0.91725848 0.54630522 1.53183895 0.52956653 1.02305166
[31] 1.65161608 1.30340256 0.27096431 1.05641695 0.58749136 0.19851994
[37] 0.04194768 1.43645222 0.70200050 1.09493028 0.40181847 1.76807864
[43] 3.24628447 0.65443582 0.08138553 1.23594540 0.28568794 1.90748439
[49] 0.27814493 0.54204644
```

multiple values at once, and the fact that the expression for $F_X^{-1}(u)$ can be easily vectorized. We can multiply the resulting observations by $1/\lambda$ to simulate observations from the Exp(λ) distribution. ■

EXAMPLE 7.3.3. (TODO: Simulate Bivariate Normal) Suppose we want to simulate observations (X, Y) from a bivariate normal distribution. To start with, let us assume that both mean parameters are 0, both variance parameters are 1, and the correlation coefficient ρ (which is also the covariance) is specified.

This problem is somewhat tricky, because the definition of bivariate normal does not directly provide a way to simulate it. All we know is that any linear combination $aX + bY$ has a univariate normal distribution. ■

EXERCISES

Ex. 7.3.1. (a) Show that both the sample mean and the sample variance of a sample obtained from the Poisson(λ) distribution will be unbiased estimators of λ.

(b) Which of these estimators is better? To answer this question, simulate random observations from the Poisson(λ) distribution for various values of λ using the R function *rpois*. Explore the behaviour of the two estimates by varying λ as well as the sample size.

Ex. 7.3.2. Exercise 2.3.7 described the technique called "capture-recapture" which bilogists use to estimate the size of the population of a species that cannot be directly counted. Suppose the unknown population size is N, and fifty members of the species are selected and given an identifying mark. Sometime later a sample of size twenty is taken from the population, and it is found to contain X of the twenty previously marked. Equating the proportion of marked members in the second sample and the population, we have $\frac{X}{20} = \frac{50}{N}$, giving an estimate of $\hat{N} = \frac{1000}{X}$.

Recall that X has a hypergeometric distribution that involves N as a parameter. It is not easy to compute $E[\hat{N}]$ and $Var[\hat{N}]$; however, Hypergeometric random variables can be simulated in R using the function `rhyper`. For each $N = 50$, 100, 200, 300, 400, and 500, use this function to simulate 1000 values of $\hat{N}$ and use them to estimate $E[\hat{N}]$ and $Var[\hat{N}]$. Plot these estimates as a function of N.

Ex. 7.3.3. Suppose p is the unknown probability of an event A, and we estimate p by the sample proportion $\hat{p}$ based on an i.i.d. sample of size n.

(a) Write $Var[\hat{p}]$ and $SD[\hat{p}]$ as functions of n and p.

(b) Using the relations derived above, determine the sample size n, as a function of p, that is required to acheive $SD(\hat{p}) = 0.01$. How does this required value of n vary with p?

(c) Design and implement the following simulation study to verify this behaviour. For $p = 0.01$, 0.1, 0.25, 0.5, 0.75, 0.9, and 0.99,

(i) Simulate 1000 values of $\hat{p}$ with $n = 500$.

(ii) Simulate 1000 values of $\hat{p}$ with n chosen according to the formula derived above.

In each case, you can think of the 1000 values as i.i.d. samples from the distribution of $\hat{p}$, and use the sample standard deviation as an estimate of $SD[\hat{p}]$. Plot the estimated values of $SD(\hat{p})$ against p for both choices of n. Your plot should look similar to Figure 7.1.

(d) (FIXME: Open-ended question) Do you think the standard deviation $SD[\hat{p}]$ is a good way to measure how well $\hat{p}$ measures p? If not, what alternatives can you think of?

Ex. 7.3.4. TODO: Give several other distributions as specific examples and specific events. Mention corresponding R functions.

7.4 PLOTS

As we will see in later chapters, making more assumptions about the underlying distribution of X allows us to give concrete answers to many important questions. This is indeed a standard and effective approach to doing statistics, but in following that approach there is a danger of forgetting that assumptions have been made, which we should guard against by doing our best to convince ourselves beforehand that the assumptions we are making are reasonable.

Doing this is more of an art than a science, and usually takes the form of staring at plots obtained from the sample observations, with the hope of answering the question: "does this plot look like what I would have expected it to look like had my assumptions been valid?" Remember that the sample $X_1, X_2, \ldots, X_n$ is a random sample, so any plot derived from it is also a "random plot". Unlike simple quantities such as sample mean and sample variance, it is not clear what to "expect" such plots to look like, and the only way to really hone our instincts to spot anomalies is through experience. In this section, we introduce some commonly used plots and use simulated data to give examples of how such plots might look like when the usual assumptions we make are valid or invalid.

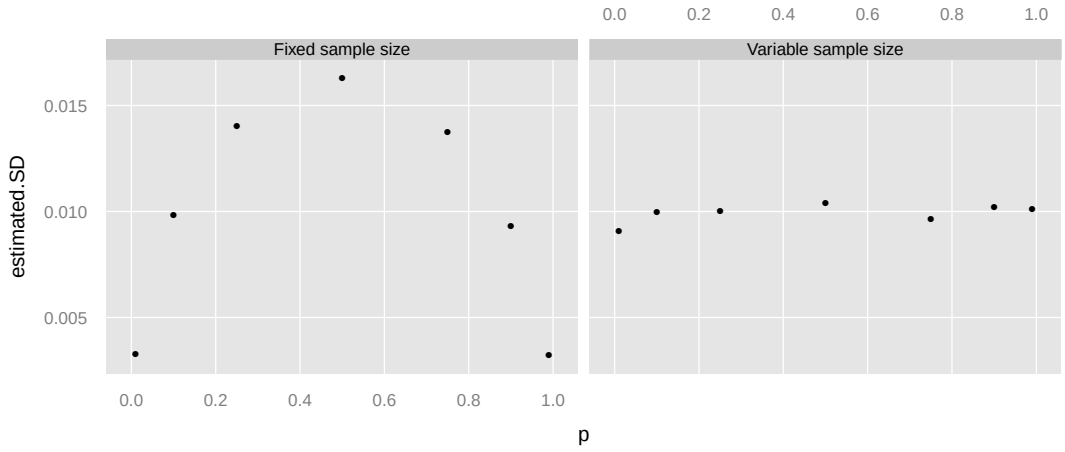

Figure 7.1: Estimated standard deviation in estimating a probability using sample proportion as a function of the probability being estimated. See exercise 7.3.3.

7.4.1 *Empirical Distribution Plot for Discrete Distributions*

The typical assumption made about a random sample is that the underlying random variable belongs to a *family* of distributions rather than a very specific one. For example, we may assume that the random variable has a Poisson(λ) distribution for some $\lambda > 0$, without placing any further restriction on λ, or a Binomial(n, p) distribution for some $0 < p < 1$. Such families are known as parametric families.

When the data $X_1, X_2, \ldots, X_n$ are from a discrete distribution, the simplest representation of the data is its empirical distribution, which is essentially a table of the frequencies of each value that appeared. For example, if we simulate 1000 samples from a Poisson distribution with mean 3, its frequency table may look like

```
> x <- rpois(1000, lambda = 3)
> table(x)

x
  0   1   2   3   4   5   6   7   8   9
 56 154 206 236 153 111  48  21  13   2

> prop.table(table(x))

x
    0     1     2     3     4     5     6     7     8     9
0.056 0.154 0.206 0.236 0.153 0.111 0.048 0.021 0.013 0.002
```

The simplest graphical representation of such a table is through a plot similar to Figure 7.2, which represents a larger Poisson sample with mean 30, resulting in many more distinct values. Although in theory all non-negative integers have positive probability of occurring, the probabilites are too small to be relevant beyond a certain range. This plot does not have a standard name, although it may be considered a variant of the Cleveland Dot Plot. We will refer to it as the *Empirical Distribution Plot* from now on.

We can make similar plots for samples from Binomial or any other distribution. Unfortunately, looking at this plot does not necessarily tell us whether the underlying distribution is Poisson, in part because the shape of the Poisson distribution varies with the λ parameter. A little later, We will discuss a modification of the empirical distribution plot, known as a rootogram, that helps make this kind of comparison a little easier.

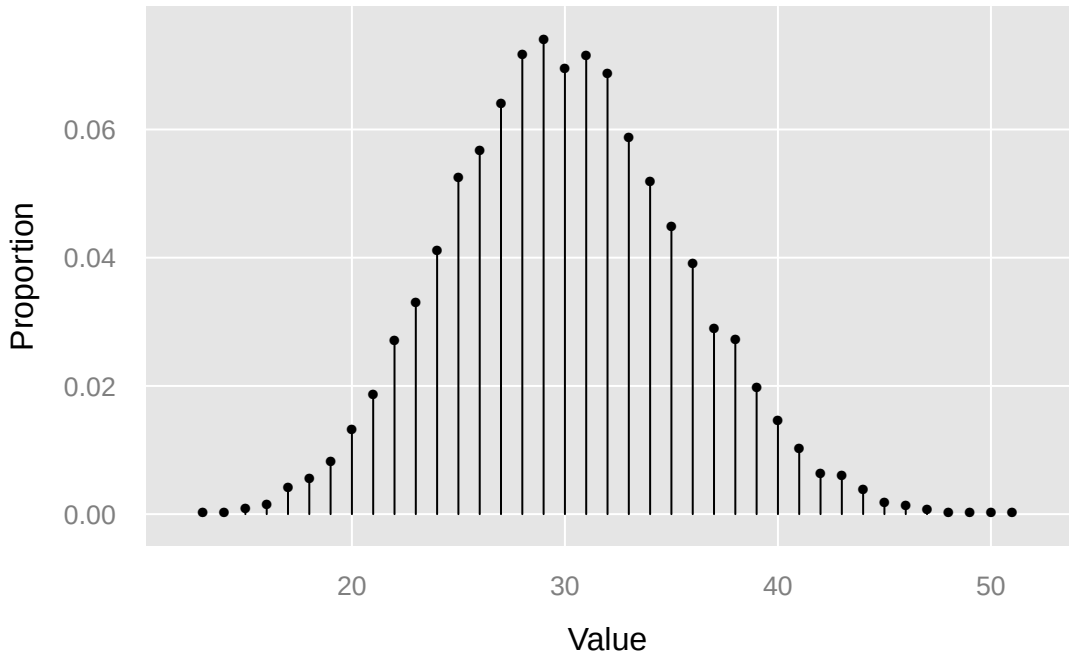

Figure 7.2: Empirical frequency distribution of 10000 random samples from the Poisson(30) distribution.

7.4.2 *Histograms for Continuous Distributions*

In the case of continuous distributions, we similarly want to make assumptions about a random sample being from a parametric family of distributions. For example, we may assume that the random variable has a Normal(μ, σ^2) distribution without placing any further restriction on the parameters μ or σ^2 (except of course that $\sigma^2 > 0$), or that it has an Exponential(λ) distribution with any value of the parameter $\lambda > 0$. Such families are known as parametric families. For both these examples, the *shape* of the distribution does not depend on the parameters, and this makes various diagnostic plots more useful.

The empirical distribution plot above is not useful for data from a continuous distribution, because by the very nature of continuous distributions, all the data points will be distinct with probability 1, and the value of the empirical distribution function will be exactly $1/n$ at these points.

The plot that is most commonly used instead to study distributions is the histogram. It is similar to the empirical distribution plot, except that it does not retain all the information contained in the empirical distribution, and instead divides the range of the data into arbitrary bins and counts the frequencies of data points falling into each bin. More precisely, the histogram estimates the probability density function of the underlying random variable by estimating the density in each bin as a quantity such that the probability of each bin is proportional to the number of observations in that bin. By choosing the bins judiciously, for example by having more of them as sample size increases, the histogram strikes a balance that ensures that the histogram "converges" to the true underlying density as $n \to \infty$.

Figure 7.3 gives examples of histograms where data are simulated from the normal and exponential distributions for varying sample sizes. Five replications are shown for each sample size. We can see that for large sample sizes, the shape of the histograms are recognizably similar to the shapes of the corresponding theoretical distributions seen in Figure 5.1 and Figure 5.2 in Chapter 5. Moreover, the shape is consistent over the five replications. This is not true, however, for small sample sizes. Remember that the histograms are based on the observed data, and are therefore random objects themselves. As we saw with numerical properties like the mean, estimates have higher variability when the sample size is small, and get less variable as sample size increases. The same holds for graphical estimates, although making this statement precise is more difficult.

7.4.3 *Hanging Rootograms for Comparing with Theoretical Distributions*

Graphical displays of data are almost always used for some kind of comparison. Sometimes these are implicit comparisons, say, asking how many peaks does a density have, or is it symmetric? More often, they are used to compare samples from two subpopulations, say, the distribution of height in males and females. Sometimes, as discussed above, they are used to compare an observed sample to a hypothesized distribution.

In the case of the empirical distribution plot, a simple modification is to add the probability mass function of the theoretical distribution. This, although a reasonable modification, is not optimal. Research into human perception of graphical displays indicates that the human eye is more adept at detecting departures from straight lines than from curves. Taking this insight into account, John Tukey suggested "hanging" the vertical lines in an empirical distribution plot (which are after all nothing but sample proportions) from their expected values under the hypothesized distribution. He further suggested a transformation of what is plotted: instead of the sample proportions and the correponding expected probabilities, he suggested plotting their square roots, thus leading to the name *hanging rootogram* for the resulting plot. The reason for making this transformation is as follows. Recall that for a proportion $\hat{p}$ obtained from a sample of size n,

$$Var[\hat{p}] = \frac{p(1-p)}{n} \approx \frac{p}{n}$$

provided p is close to 0. In Chapter 9, we will encounter the Central Limit Theorem and the Delta Method, which can be used to show that as the sample size n grows large, $Var[\sqrt{\hat{p}}] \approx c/n$ for a constant c. This means that unlike $\hat{p} - p$, the variance of $\sqrt{\hat{p}} - \sqrt{p}$ will be approximately independent of p. Figure 7.4 gives examples of hanging rootograms.

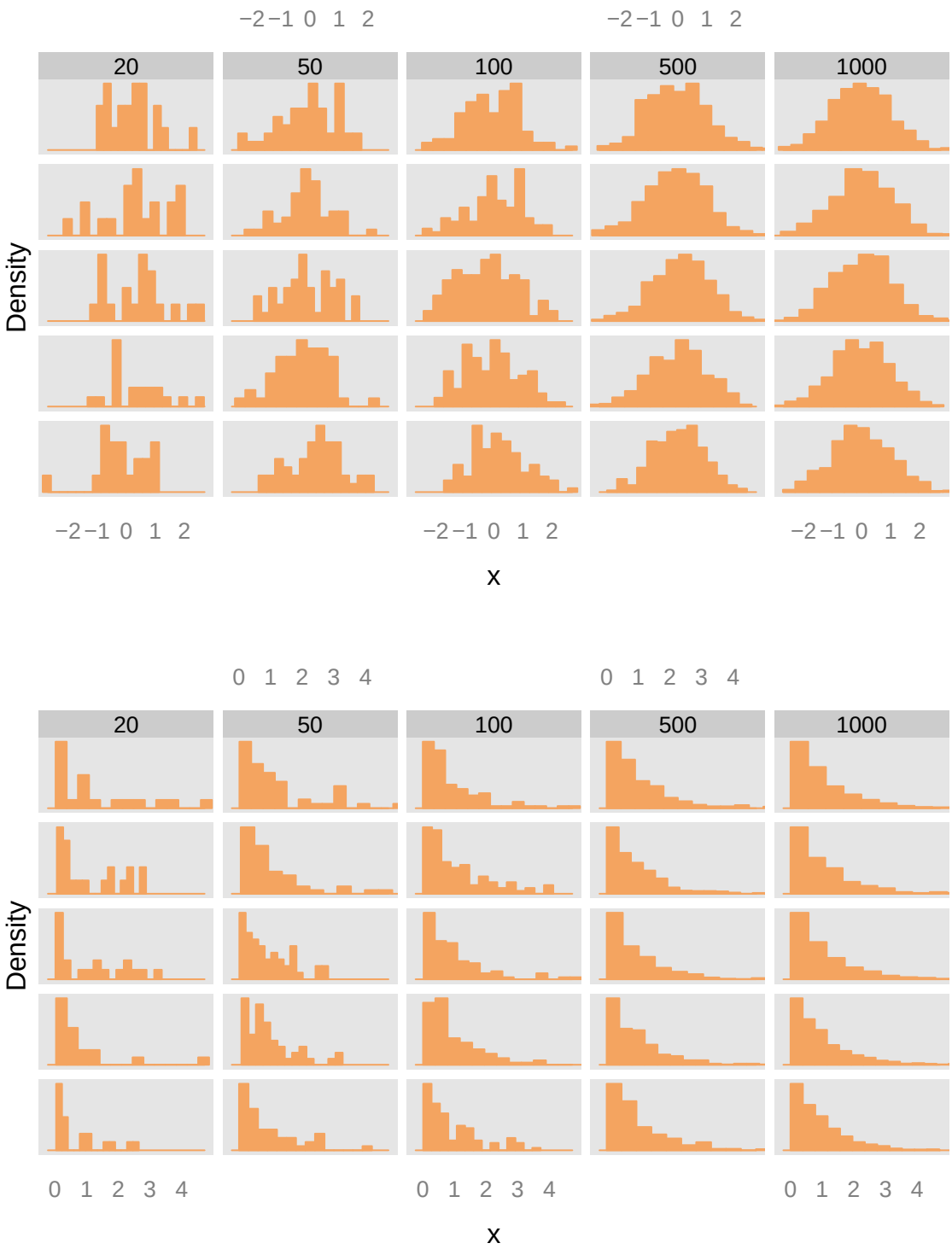

Figure 7.3: Histograms of random samples from the Normal$(0, 1)$ (top) and Exponential(1) (bottom) distributions. Columns represent increasing sample sizes, and rows are independent repetitions of the experiment.

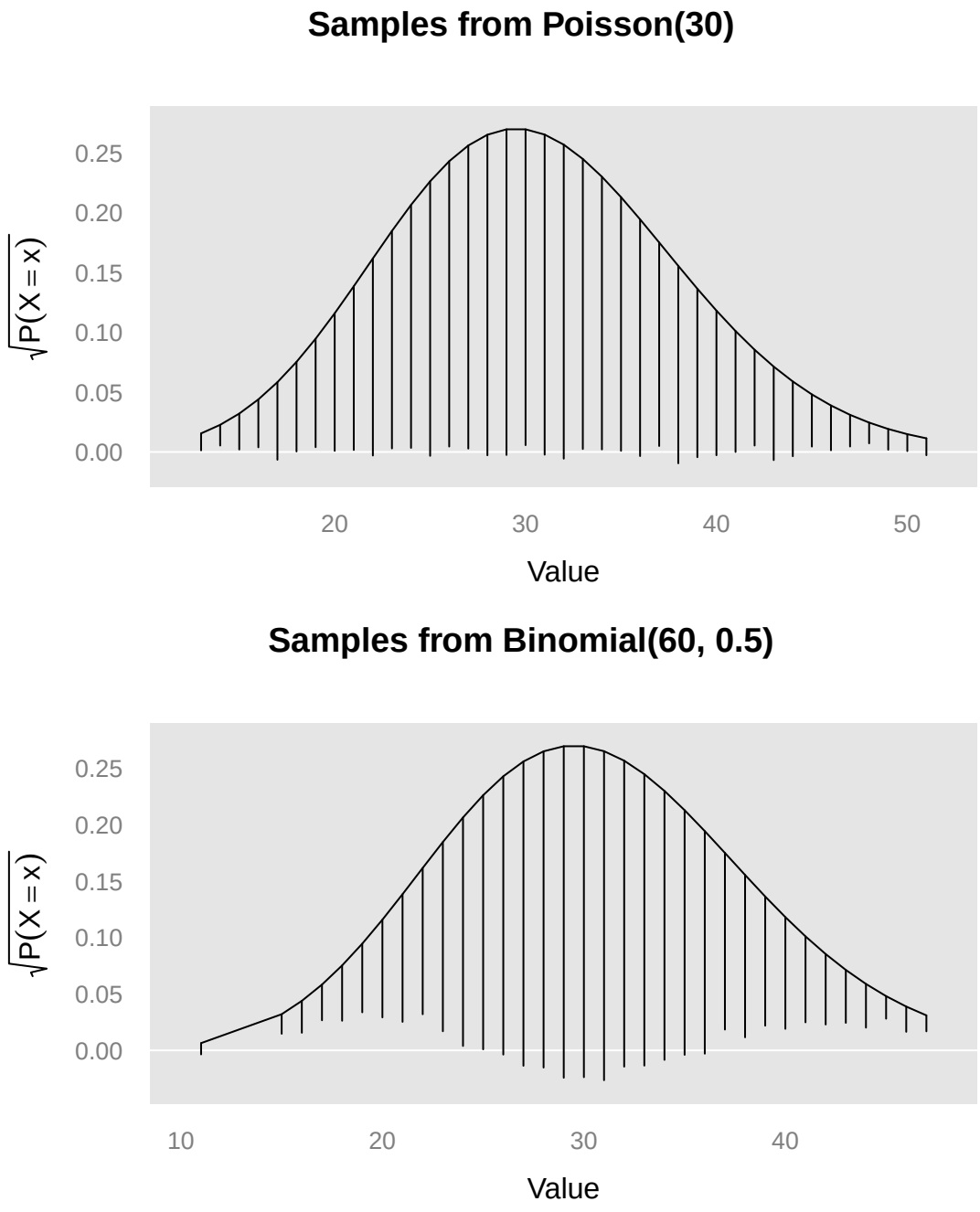

Figure 7.4: Hanging rootogram of 10000 random samples compared with the Poisson(30) distribution. In the top plot, the samples are also from Poisson(30), whereas in the bottom plot the samples are from the Binomial$(100, 0.3)$ distribution, which has the same mean but different variance. Note the similarities with Figure 2.2

7.4.4 *Q-Q Plots for Continuous Distributions*

Just as histograms were binned versions of the empirical distribution plot, we can plot binned versions of hanging rootograms for data from a continuous distribution as well. It is more common however, to look at quantile-quantile plots (QQ plots), which do not bin the data, but instead plot what is essentially a transformation of the empirical CDF.

Recall that the ECDF of observations $X_1, X_2, \ldots, X_n$ is given by

$$\hat{F}_n(t) = P(Y \leq t) = \frac{\#\{X_i \leq t\}}{n}$$

The top plot in Figure 7.5 is a conventional ECDF plot of 200 observations simulated from a Normal$(1, 0.5^2)$ distribution. The bottom plot has the sorted data values on the y-axis and 200 equally spaced numbers from 0 to 1. A little thought tells us that this plot is essentially the same as the ECDF plot, with the x- and y-axes switched, and using points instead of lines. Naturally, we expect that for reasonably large sample sizes, the ECDF plot obtained from a random sample will be close to the true cumulative distribution function of the underlying distribution. If we know the shape of the distribution we expect the data to be from, we can compare it with the shape seen in the plot.

Although this is a fine idea in principle, it is difficult in practice to detect small differences between the observed shape and the theorized or expected shape. Here, we are helped again by the insight that the human eye finds it easier to detect deviations from a straight line than from curves. By keeping the sorted data values unchanged, but transforming the equally spaced probability values to the corresponding quantile values of the theorized distribution, we obtain a plot that we expect to be linear. . Quantiles are defined as follows: For a given CDF F, the quantile corresponding to a probability value $p \in [0, 1]$ is a value x such that $F(x) = p$. Such an x may not exist for all p and F, and the definition of quantile needs to be modified to take this into account. However, for most standard continuous distributions used in Q-Q plots, the above definition is adequate. Such a plot with Normal$(0, 1)$ quantiles is shown in Figure 7.6.

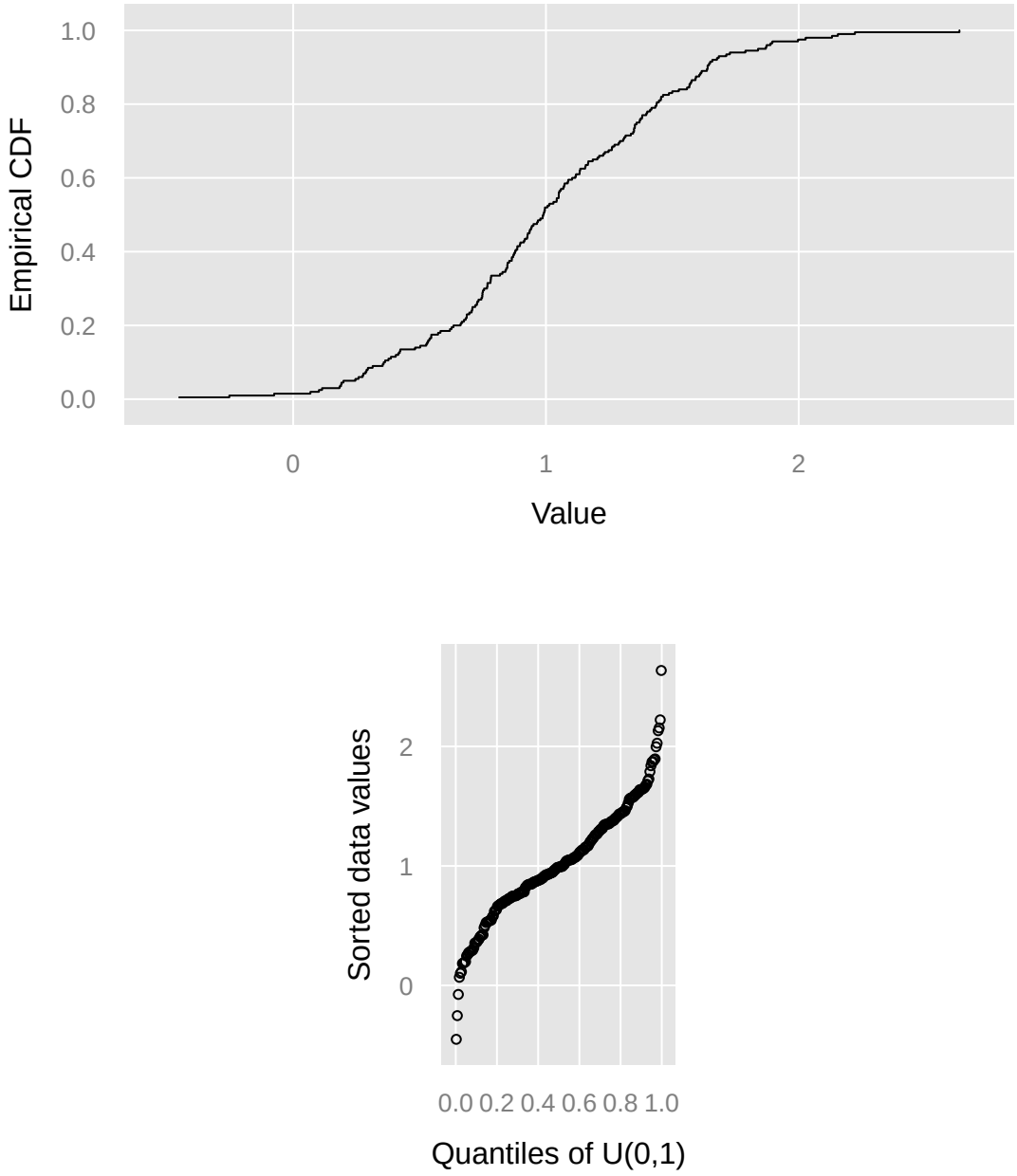

Figure 7.5: Conventional ECDF plot (top) and its "inverted" version (bottom), with x- and y-axes switched, and points instead of lines.

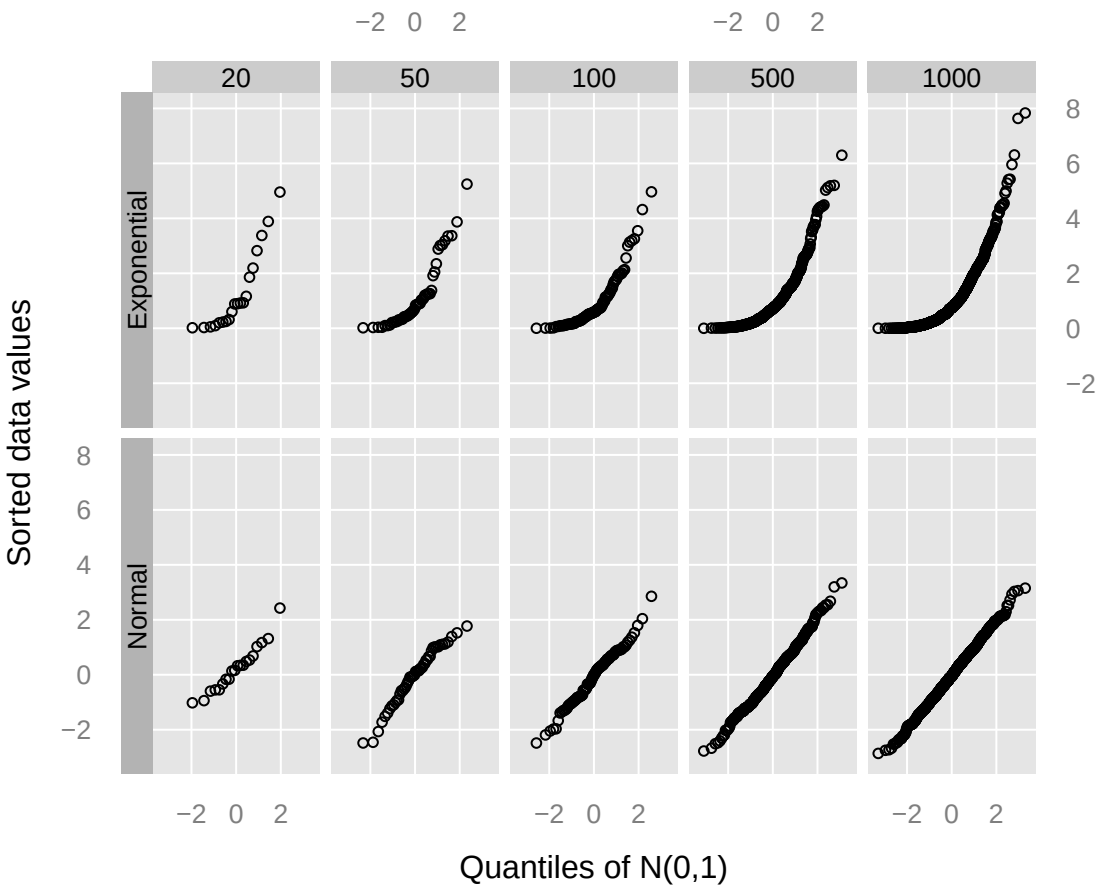

Figure 7.6: Normal Q-Q plots of data generated from Normal and Exponential distributions, with varying sample size. The Q-Q plots are more or less linear for Normal data, but exhibit curvature indicative of a relatively heavy right tail for exponential data. Not surprisingly, the difference becomes easier to see as the sample size increases.