2

Frequentist Inference

Before the computer age there was the calculator age, and before "big data" there were small data sets, often a few hundred numbers or fewer, laboriously collected by individual scientists working under restrictive experimental constraints. Precious data calls for maximally efficient statistical analysis. A remarkably effective theory, feasible for execution on mechanical desk calculators, was developed beginning in 1900 by Pearson, Fisher, Neyman, Hotelling, and others, and grew to dominate twentieth-century statistical practice. The theory, now referred to as *classical*, relied almost entirely on frequentist inferential ideas. This chapter sketches a quick and simplified picture of frequentist inference, particularly as employed in classical applications.

We begin with another example from Dr. Myers' nephrology laboratory: 211 kidney patients have had their *glomerular filtration rates* measured, with the results shown in Figure 2.1; `gfr` is an important indicator of kidney function, with low values suggesting trouble. (It is a key component of `tot` in Figure 1.1.) The mean and standard error (1.1)–(1.2) are $\bar{x} = 54.25$ and $\widehat{\text{se}} = 0.95$, typically reported as

$$54.25 \pm 0.95; \tag{2.1}$$

± 0.95 denotes a frequentist inference for the accuracy of the estimate $\bar{x} = 54.25$, and suggests that we shouldn't take the ".25" very seriously, even the "4" being open to doubt. Where the inference comes from and what exactly it means remains to be said.

Statistical inference usually begins with the assumption that some probability model has produced the observed data $\boldsymbol{x}$, in our case the vector of $n = 211$ `gfr` measurements $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$. Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ indicate n independent draws from a probability distribution F, written

$$F \to \boldsymbol{X}, \tag{2.2}$$

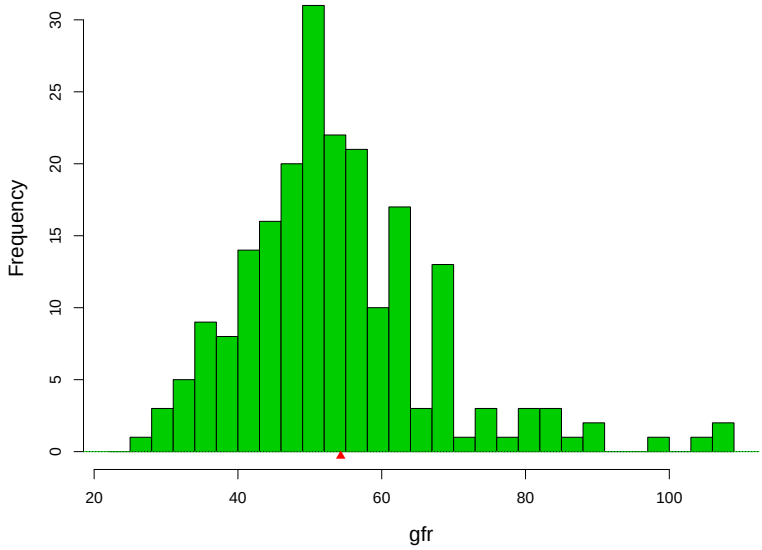

Figure 2.1 Glomerular filtration rates for 211 kidney patients; mean 54.25, standard error .95.

F being the underlying distribution of possible `gfr` scores here. A realization $\boldsymbol{X} = \boldsymbol{x}$ of (2.2) has been observed, and the statistician wishes to *infer* some property of the unknown distribution F.

Suppose the desired property is the *expectation* of a single random draw X from F, denoted

$$\theta = E_F\{X\} \tag{2.3}$$

(which also equals the expectation of the average $\bar{X} = \sum X_i/n$ of random vector (2.2)[1]). The obvious estimate of θ is $\hat{\theta} = \bar{x}$, the sample average. If n were enormous, say 10^{10}, we would expect $\hat{\theta}$ to nearly equal θ, but otherwise there is room for error. How much error is the inferential question.

The estimate $\hat{\theta}$ is calculated from $\boldsymbol{x}$ according to some known algorithm, say

$$\hat{\theta} = t(\boldsymbol{x}), \tag{2.4}$$

$t(\boldsymbol{x})$ in our example being the averaging function $\bar{x} = \sum x_i/n$; $\hat{\theta}$ is a

[1] The fact that $E_F\{\bar{X}\}$ equals $E_F\{X\}$ is a crucial, though easily proved, probabilistic result.

realization of

$$\hat{\Theta} = t(\boldsymbol{X}), \tag{2.5}$$

the output of $t(\cdot)$ applied to a theoretical sample $\boldsymbol{X}$ from F (2.2). We have chosen $t(\boldsymbol{X})$, we hope, to make $\hat{\Theta}$ a good estimator of θ, the desired property of F.

We can now give a first definition of frequentist inference: *the accuracy of an observed estimate* $\hat{\theta} = t(\boldsymbol{x})$ *is the probabilistic accuracy of* $\hat{\Theta} = t(\boldsymbol{X})$ *as an estimator of* θ. This may seem more a tautology than a definition, but it contains a powerful idea: $\hat{\theta}$ is just a single number but $\hat{\Theta}$ takes on a range of values whose spread can define measures of accuracy.

Bias and variance are familiar examples of frequentist inference. Define μ to be the expectation of $\hat{\Theta} = t(\boldsymbol{X})$ under model (2.2),

$$\mu = E_F\{\hat{\Theta}\}. \tag{2.6}$$

Then the bias and variance attributed to estimate $\hat{\theta}$ of parameter θ are

$$\text{bias} = \mu - \theta \quad \text{and} \quad \text{var} = E_F\left\{(\hat{\Theta} - \mu)^2\right\}. \tag{2.7}$$

Again, what keeps this from tautology is the attribution to the single number $\hat{\theta}$ of the probabilistic properties of $\hat{\Theta}$ following from model (2.2). If all of this seems too obvious to worry about, the Bayesian criticisms of Chapter 3 may come as a shock.

Frequentism is often defined with respect to "an infinite sequence of future trials." We imagine hypothetical data sets $\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, \boldsymbol{X}^{(3)}, \ldots$ generated by the same mechanism as $\boldsymbol{x}$ providing corresponding values $\hat{\Theta}^{(1)}$, $\hat{\Theta}^{(2)}$, $\hat{\Theta}^{(3)}, \ldots$ as in (2.5). The frequentist principle is then to attribute for $\hat{\theta}$ the accuracy properties of the ensemble of $\hat{\Theta}$ values.[2] If the $\hat{\Theta}$s have empirical variance of, say, 0.04, then $\hat{\theta}$ is claimed to have standard error $0.2 = \sqrt{0.04}$, etc. This amounts to a more picturesque restatement of the previous definition.

2.1 Frequentism in Practice

Our working definition of frequentism is that *the probabilistic properties of a procedure of interest are derived and then applied verbatim to the procedure's output for the observed data*. This has an obvious defect: it requires calculating the properties of estimators $\hat{\Theta} = t(\boldsymbol{X})$ obtained from

[2] In essence, frequentists ask themselves "What would I see if I reran the same situation again (and again and again. . .)?"

the true distribution F, even though F is unknown. Practical frequentism uses a collection of more or less ingenious devices to circumvent the defect.

1. The plug-in principle. A simple formula relates the standard error of $\bar{X} = \sum X_i/n$ to $\mathrm{var}_F(X)$, the variance of a single X drawn from F,

$$\mathrm{se}\left(\bar{X}\right) = \left[\mathrm{var}_F(X)/n\right]^{1/2}. \tag{2.8}$$

But having observed $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ we can estimate $\mathrm{var}_F(X)$ without bias by

$$\widehat{\mathrm{var}}_F = \sum (x_i - \bar{x})^2 / (n-1). \tag{2.9}$$

Plugging formula (2.9) into (2.8) gives $\widehat{\mathrm{se}}$ (1.2), the usual estimate for the standard error of an average $\bar{x}$. In other words, the frequentist accuracy estimate for $\bar{x}$ is itself estimated from the observed data.[3]

2. Taylor-series approximations. Statistics $\hat{\theta} = t(\boldsymbol{x})$ more complicated than $\bar{x}$ can often be related back to the plug-in formula by local linear approximations, sometimes known as the "delta method."[†] For example, $\hat{\theta} = \bar{x}^2$ has $d\hat{\theta}/d\bar{x} = 2\bar{x}$. Thinking of $2\bar{x}$ as a constant gives †1

$$\mathrm{se}\left(\bar{x}^2\right) \doteq 2\,|\bar{x}|\,\widehat{\mathrm{se}}, \tag{2.10}$$

with $\widehat{\mathrm{se}}$ as in (1.2). Large sample calculations, as sample size n goes to infinity, validate the delta method which, fortunately, often performs well in small samples.

3. Parametric families and maximum likelihood theory. Theoretical expressions for the standard error of a maximum likelihood estimate (MLE) are discussed in Chapters 4 and 5, in the context of parametric families of distributions. These combine Fisherian theory, Taylor-series approximations, and the plug-in principle in an easy-to-apply package.

4. Simulation and the bootstrap. Modern computation has opened up the possibility of numerically implementing the "infinite sequence of future trials" definition, except for the infinite part. An estimate $\hat{F}$ of F, perhaps the MLE, is found, and values $\hat{\Theta}^{(k)} = t(\boldsymbol{X}^{(k)})$ simulated from $\hat{F}$ for $k = 1, 2, \ldots, B$, say $B = 1000$. The empirical standard deviation of the $\hat{\Theta}$s is then the frequentist estimate of standard error for $\hat{\theta} = t(\boldsymbol{x})$, and similarly with other measures of accuracy.

This is a good description of the bootstrap, Chapter 10. (Notice that

[3] The most familiar example is the observed proportion p of heads in n flips of a coin having true probability π: the actual standard error is $[\pi(1-\pi)/n]^{1/2}$ but we can only report the plug-in estimate $[p(1-p)/n]^{1/2}$.

Table 2.1 *Three estimates of location for the* `gfr` *data, and their estimated standard errors; last two standard errors using the bootstrap,* $B = 1000$.

	Estimate	Standard error
mean	54.25	.95
25% Winsorized mean	52.61	.78
median	52.24	.87

here the plugging-in, of $\hat{F}$ for F, comes *first* rather than at the end of the process.) The classical methods 1–3 above are restricted to estimates $\hat{\theta} = t(\boldsymbol{x})$ that are smoothly defined functions of various sample means. Simulation calculations remove this restriction. Table 2.1 shows three "location" estimates for the `gfr` data, the mean, the 25% Winsorized mean,[4] and the median, along with their standard errors, the last two computed by the bootstrap. A happy feature of computer-age statistical inference is the tremendous expansion of useful and usable statistics $t(\boldsymbol{x})$ in the statistician's working toolbox, the `lowess` algorithm in Figures 1.2 and 1.3 providing a nice example.

5. Pivotal statistics. A pivotal statistic $\hat{\theta} = t(\boldsymbol{x})$ is one whose distribution does *not* depend upon the underlying probability distribution F. In such a case the theoretical distribution of $\hat{\Theta} = t(\boldsymbol{X})$ applies exactly to $\hat{\theta}$, removing the need for devices 1–4 above. The classic example concerns Student's two-sample t-test.

In a two-sample problem the statistician observes two sets of numbers,

$$\boldsymbol{x}_1 = (x_{11}, x_{12}, \ldots, x_{1n_1}) \quad \boldsymbol{x}_2 = (x_{21}, x_{22}, \ldots, x_{2n_2}), \tag{2.11}$$

and wishes to test the *null hypothesis* that they come from the same distribution (as opposed to, say, the second set tending toward larger values than the first). It is assumed that the distribution F_1 for $\boldsymbol{x}_1$ is *normal*, or *Gaussian*,

$$X_{1i} \overset{\text{ind}}{\sim} \mathcal{N}(\mu_1, \sigma^2), \qquad i = 1, 2, \ldots, n_1, \tag{2.12}$$

the notation indicating n_1 independent draws from a normal distribution[5]

[4] All observations below the 25th percentile of the 211 observations are moved up to that point, similarly those above the 75th percentile are moved down, and finally the mean is taken.

[5] Each draw having probability density $(2\pi\sigma^2)^{-1/2} \exp\{-0.5 \cdot (x - \mu_1)^2/\sigma^2\}$.

with expectation μ_1 and variance σ^2. Likewise

$$X_{2i} \overset{\text{ind}}{\sim} \mathcal{N}(\mu_2, \sigma^2) \qquad i = 1, 2, \ldots, n_2. \tag{2.13}$$

We wish to test the null hypothesis

$$H_0 : \mu_1 = \mu_2. \tag{2.14}$$

The obvious test statistic $\hat{\theta} = \bar{x}_2 - \bar{x}_1$, the difference of the means, has distribution

$$\hat{\theta} \sim \mathcal{N}\left(0, \sigma^2\left(\tfrac{1}{n_1} + \tfrac{1}{n_2}\right)\right) \tag{2.15}$$

under H_0. We could plug in the unbiased estimate of σ^2,

$$\hat{\sigma}^2 = \left[\sum_1^{n_1}(x_{1i} - \bar{x}_1)^2 + \sum_1^{n_2}(x_{2i} - \bar{x}_2)^2\right] \Big/ (n_1 + n_2 - 2), \tag{2.16}$$

but Student provided a more elegant solution: instead of $\hat{\theta}$, we test H_0 using the two-sample t-statistic

$$t = \frac{\bar{x}_2 - \bar{x}_1}{\widehat{\text{sd}}}, \qquad \text{where } \widehat{\text{sd}} = \hat{\sigma}\left(\tfrac{1}{n_1} + \tfrac{1}{n_2}\right)^{1/2}. \tag{2.17}$$

Under H_0, t is pivotal, having the same distribution (Student's t distribution with $n_1 + n_2 - 2$ degrees of freedom), no matter what the value of the "nuisance parameter" σ.

For $n_1 + n_2 - 2 = 70$, as in the leukemia example (1.5)–(1.6), Student's distribution gives

$$\Pr_{H_0}\{-1.99 \le t \le 1.99\} = 0.95. \tag{2.18}$$

The hypothesis test that rejects H_0 if $|t|$ exceeds 1.99 has probability exactly 0.05 of mistaken rejection. Similarly,

$$\bar{x}_2 - \bar{x}_1 \pm 1.99 \cdot \widehat{\text{sd}} \tag{2.19}$$

is an exact 0.95 *confidence interval* for the difference $\mu_2 - \mu_1$, covering the true value in 95% of repetitions of probability model (2.12)–(2.13).[6]

[6] Occasionally, one sees frequentism defined in careerist terms, e.g., "A statistician who always rejects null hypotheses at the 95% level will over time make only 5% errors of the first kind." This is not a comforting criterion for the statistician's clients, who are interested in their own situations, not everyone else's. Here we are only assuming hypothetical repetitions of the specific problem at hand.

What might be called the *strong definition of frequentism* insists on exact frequentist correctness under experimental repetitions. Pivotality, unfortunately, is unavailable in most statistical situations. Our looser definition of frequentism, supplemented by devices such as those above,[7] presents a more realistic picture of actual frequentist practice.

2.2 Frequentist Optimality

The popularity of frequentist methods reflects their relatively modest mathematical modeling assumptions: only a probability model F (more exactly a family of probabilities, Chapter 3) and an algorithm of choice $t(\boldsymbol{x})$. This flexibility is also a defect in that the principle of frequentist correctness doesn't help with the choice of algorithm. Should we use the sample mean to estimate the location of the `gfr` distribution? Maybe the 25% Winsorized mean would be better, as Table 2.1 suggests.

The years 1920–1935 saw the development of two key results on *frequentist optimality*, that is, finding the *best* choice of $t(\boldsymbol{x})$ given model F. The first of these was Fisher's theory of maximum likelihood estimation and the Fisher information bound: in parametric probability models of the type discussed in Chapter 4, the MLE is the optimum estimate in terms of minimum (asymptotic) standard error.

In the same spirit, the Neyman–Pearson lemma provides an optimum hypothesis-testing algorithm. This is perhaps the most elegant of frequentist constructions. In its simplest formulation, the NP lemma assumes we are trying to decide between two possible probability density functions for the observed data $\boldsymbol{x}$, a null hypothesis density $f_0(\boldsymbol{x})$ and an alternative density $f_1(\boldsymbol{x})$. A testing rule $t(\boldsymbol{x})$ says which choice, 0 or 1, we will make having observed data $\boldsymbol{x}$. Any such rule has two associated frequentist error probabilities: choosing f_1 when actually f_0 generated $\boldsymbol{x}$, and vice versa,

$$\begin{aligned} \alpha &= \Pr_{f_0}\{t(\boldsymbol{x}) = 1\}, \\ \beta &= \Pr_{f_1}\{t(\boldsymbol{x}) = 0\}. \end{aligned} \tag{2.20}$$

Let $L(\boldsymbol{x})$ be the *likelihood ratio*,

$$L(\boldsymbol{x}) = f_1(\boldsymbol{x})/f_0(\boldsymbol{x}) \tag{2.21}$$

[7] The list of devices is not complete. Asymptotic calculations play a major role, as do more elaborate combinations of pivotality and the plug-in principle; see the discussion of approximate bootstrap confidence intervals in Chapter 11.

and define the testing rule $t_c(\boldsymbol{x})$ by

$$t_c(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \log L(\boldsymbol{x}) \geq c \\ 0 & \text{if } \log L(\boldsymbol{x}) < c. \end{cases} \tag{2.22}$$

There is one such rule for each choice of the cutoff c. The Neyman–Pearson lemma says that only rules of form (2.22) can be optimum; for any other rule $t(\boldsymbol{x})$ there will be a rule $t_c(\boldsymbol{x})$ having smaller errors of both kinds,[8]

$$\alpha_c < \alpha \quad \text{and} \quad \beta_c < \beta. \tag{2.23}$$

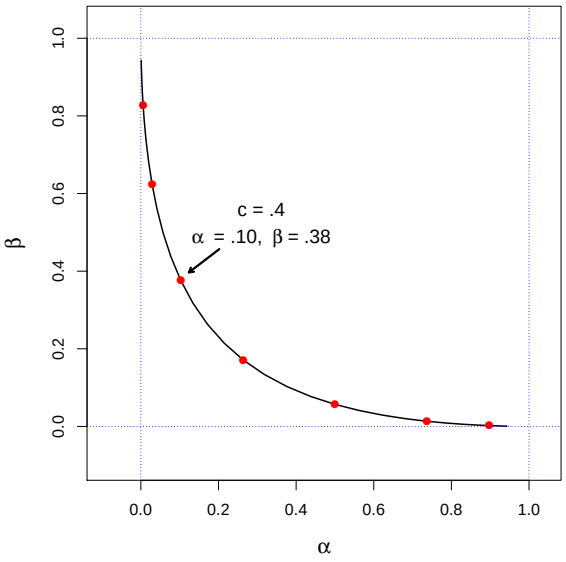

Figure 2.2 Neyman–Pearson alpha–beta curve for $f_0 \sim \mathcal{N}(0, 1)$, $f_1 \sim \mathcal{N}(.5, 1)$, and sample size $n = 10$. Red dots correspond to cutoffs $c = .8, .6, .4, \ldots, -.4$.

Figure 2.2 graphs (α_c, β_c) as a function of the cutoff c, for the case where $\boldsymbol{x} = (x_1, x_2, \ldots, x_{10})$ is obtained by independent sampling from a normal distribution, $\mathcal{N}(0, 1)$ for f_0 versus $\mathcal{N}(0.5, 1)$ for f_1. The NP lemma says that any rule not of form (2.22) must have its (α, β) point lying above the curve.

[8] Here we are ignoring some minor definitional difficulties that can occur if f_0 and f_1 are discrete.

Frequentist optimality theory, both for estimation and for testing, anchored statistical practice in the twentieth century. The larger data sets and more complicated inferential questions of the current era have strained the capabilities of that theory. Computer-age statistical inference, as we will see, often displays an unsettling ad hoc character. Perhaps some contemporary Fishers and Neymans will provide us with a more capacious optimality theory equal to the challenges of current practice, but for now that is only a hope.

Frequentism cannot claim to be a seamless philosophy of statistical inference. Paradoxes and contradictions abound within its borders, as will be shown in the next chapter. That being said, frequentist methods have a natural appeal to working scientists, an impressive history of successful application, and, as our list of five "devices" suggests, the capacity to encourage clever methodology. The story that follows is not one of abandonment of frequentist thinking, but rather a broadening of connections with other methods.

2.3 Notes and Details

The name "frequentism" seems to have been suggested by Neyman as a statistical analogue of Richard von Mises' frequentist theory of probability, the connection being made explicit in his 1977 paper, "Frequentist probability and frequentist statistics." "Behaviorism" might have been a more descriptive name[9] since the theory revolves around the long-run behavior of statistics $t(\boldsymbol{x})$, but in any case "frequentism" has stuck, replacing the older (sometimes disparaging) term "objectivism." Neyman's attempt at a complete frequentist theory of statistical inference, "inductive behavior," is not much quoted today, but can claim to be an important influence on Wald's development of decision theory.

R. A. Fisher's work on maximum likelihood estimation is featured in Chapter 4. Fisher, arguably the founder of frequentist optimality theory, was not a pure frequentist himself, as discussed in Chapter 4 and Efron (1998), "R. A. Fisher in the 21st Century." (Now that we are well into the twenty-first century, the author's talents as a prognosticator can be frequentistically evaluated.)

†1 [p. 15] *Delta method.* The delta method uses a first-order Taylor series to approximate the variance of a function $s(\hat{\theta})$ of a statistic $\hat{\theta}$. Suppose $\hat{\theta}$ has mean/variance (θ, σ^2), and consider the approximation $s(\hat{\theta}) \approx s(\theta) +$

[9] That name is already spoken for in the psychology literature.

$s'(\theta)(\hat{\theta} - \theta)$. Hence $\text{var}\{s(\hat{\theta})\} \approx |s'(\theta)|^2\sigma^2$. We typically plug-in $\hat{\theta}$ for θ, and use an estimate for σ^2.