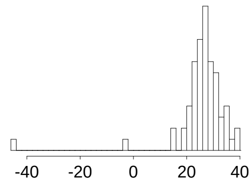

Figure 8.3 *Histogram of Simon Newcomb's measurements for estimating the speed of light, from Stigler (1977). The data represent the amount of time required for light to travel a distance of 7442 meters and are recorded as deviations from* 24,800 *nanoseconds.*

Example: comparing data to replications from a fitted normal distribution

The most fundamental way to check model fit is to display replicated datasets and compare them to the actual data. Here we illustrate with a simple case, from a famous historical dataset that did not fit the normal distribution. The goal of this example is to demonstrate how the lack of fit can be seen using predictive replications.

Figure 8.3 shows the data, a set of measurements taken by Simon Newcomb in 1882 as part of an experiment to estimate the speed of light. We (inappropriately) fit a normal distribution to these data, which in the regression context can be done by fitting a linear regression with no predictors:

R code
```
light <- lm (y ~ 1)
```

The next step is to simulate 1000 replications from the parameters in the fitted model (in this case, simply the constant term β_0 and the residual standard deviation σ):

R code
```
n.sims <- 1000
sim.light <- sim (light, n.sims)
```

We can then use these simulations to create 1000 fake datasets of 66 observations each:

R code
```
n <- length (y)
y.rep <- array (NA, c(n.sims, n))
for (s in 1:n.sims){
  y.rep[s,] <- rnorm (n, sim.light$beta[s], sim.light$sigma[s])
}
```

Visual comparison of actual and replicated datasets. Figure 8.4 shows a plot of 20 of the replicated datasets, produced as follows:

R code
```
par (mfrow=c(5,4))
for (s in 1:20){
  hist (y.rep[s,])
}
```

The systematic differences between data and replications are clear. In more complicated problems, more effort may be needed to effectively display the data and replications for useful comparisons, but the same general idea holds.

Checking model fit using a numerical data summary. Data displays can suggest more focused test statistics with which to check model fit, as we illustrate in Section 24.2. Here we demonstrate a simple example with the speed-of-light measurements. The graphical check in Figures 8.3 and 8.4 shows that the data have some extremely low values that do not appear in the replications. We can formalize this check by defining a *test statistic*, $T(y)$, equal to the minimum value of the data, and then calculating $T(y^{\text{rep}})$ for each of the replicated datasets: