4.9: The Law of Large Numbers

We just looked at the results of one fictitious IQ experiment with a sample size of \(N=100\). The results were somewhat encouraging: the true population mean is 100, and the sample mean of 98.5 is a pretty reasonable approximation to it. In many scientific studies that level of precision is perfectly acceptable, but in other situations you need to be a lot more precise. If we want our sample statistics to be much closer to the population parameters, what can we do about it?

The obvious answer is to collect more data. Suppose that we ran a much larger experiment, this time measuring the IQ's of 10,000 people. We can simulate the results of this experiment using R, using the rnorm() function, which generates random numbers sampled from a normal distribution. For an experiment with a sample size of n = 10000, and a population with mean = 100 and sd = 15, R produces our fake IQ data using these commands:

```
IQ <- rnorm(n=10000, mean=100, sd=15) #generate IQ scores
IQ <- round(IQ) # make round numbers
```

Cool, we just generated 10,000 fake IQ scores. Where did they go? Well, they went into the variable IQ on my computer. You can do the same on your computer too by copying the above code. 10,000 numbers is too many numbers to look at. We can look at the first 100 like this:

```
IQ <- rnorm(n=10000, mean=100, sd=15)
IQ <- round(IQ)
print(IQ[1:100])
```

run | restart | restart & run all

```
  [1]  97  98 101 114 110 105  84  95  96 103  86 118  99  93  64 101
117 104
 [19] 106  73  81  98 100 111 103 100  91 115 107  98 107  76  70 107
104  86
 [37] 120  91 103 129  92  98 105 108  96  87  94  97 102  80  98  76
131 107
 [55] 104 114  90 109 104  86 124  73 131 114 104  83  99  91  83 105
107 107
 [73] 125  74 112  87  76 103 105  88  97  86  99  90 117 121  86 109
132  89
 [91]  97 132  76 131  98 111 118  98  94  98
```

We can compute the mean IQ using the command mean(IQ) and the standard deviation using the command sd(IQ), and draw a histogram using hist(). The histogram of this much larger sample is shown in Figure 4.8.4*c*. Even a moment's inspections makes clear that the larger sample is a much better approximation to the true population distribution than the smaller one. This is reflected in the sample statistics: the mean IQ for the larger sample turns out to be 99.9, and the standard deviation is 15.1. These values are now very close to the true population.s

I feel a bit silly saying this, but the thing I want you to take away from this is that large samples generally give you better information. I feel silly saying it because it's so bloody obvious that it shouldn't need to be said. In fact, it's such an obvious point that when Jacob Bernoulli – one of the founders of probability theory – formalized this idea back in 1713, he was kind of a jerk about it. Here's how he described the fact that we all share this intuition:

> *For even the most stupid of men, by some instinct of nature, by himself and without any instruction (which is a remarkable thing), is convinced that the more observations have been made, the less danger there is of wandering from one's goal (see Stigler, 1986, p65).*

Okay, so the passage comes across as a bit condescending (not to mention sexist), but his main point is correct: it really does feel obvious that more data will give you better answers. The question is, why is this so? Not surprisingly, this intuition that we all share turns out to be correct, and statisticians refer to it as the law of large numbers. The law of large numbers is a mathematical law that

applies to many different sample statistics, but the simplest way to think about it is as a law about averages. The sample mean is the most obvious example of a statistic that relies on averaging (because that's what the mean is… an average), so let's look at that. When applied to the sample mean, what the law of large numbers states is that as the sample gets larger, the sample mean tends to get closer to the true population mean. Or, to say it a little bit more precisely, as the sample size "approaches" infinity (written as \(N \rightarrow \infty\)) the sample mean approaches the population mean (\(\bar{X} \rightarrow \mu\)).

I don't intend to subject you to a proof that the law of large numbers is true, but it's one of the most important tools for statistical theory. The law of large numbers is the thing we can use to justify our belief that collecting more and more data will eventually lead us to the truth. For any particular data set, the sample statistics that we calculate from it will be wrong, but the law of large numbers tells us that if we keep collecting more data those sample statistics will tend to get closer and closer to the true population parameters.
