

Chapter 4

One Sample Tests

In this chapter, we will follow up with some more concrete examples based upon the concepts introduced in the last chapter. We will learn how to determine whether a descriptive statistic (such as the median, mean, standard deviation, etc.) is or is not consistent with a predicted value. We will also learn how to determine the range of values over which we can expect a statistic to vary, that is, how to determine the sampling distribution. In fact, these are really two sides of the same coin.

In order to use Monte Carlo and bootstrapping methods, we are going to need to know how to probe our sampling distributions for some information. The three most common types of information we extract from a sampling distribution are 1) the standard deviation (which is the “standard error of the mean” under traditional methods), 2) 95% confidence intervals, and 3) a probability of some particular value arising from the conditions that generated the sampling distribution (a “significance”). These three are tightly coupled, and can be thought of as different ways of expressing the same basic information.

Going to the movies

Consider the data shown in Figure 4.1, which shows a histogram of how many people (out of a sample of 100) have seen x movies in the past week. (These data – moviewatch.txt - are available on the webpage; you are encouraged to download them and work through the examples in this chapter.) The data are highly skewed, because the vast majority of people (70%) have either seen 0 or 1 movie in the past week (and 7 people in the sample must be movie critics, having seen 7 or more movies per week).

Let’s say we wanted to find out if people typically watched more than one movie per week. We could tackle this problem a couple different ways. First, let’s take a bad, but very easy approach: let us test whether mean number of movies watched was larger than 1. The mean of our moviewatch data is 1.59, and traditional statistics using Central Limit Theorem (CLT) can easily tell us how likely it is that this mean comes from a sampling distribution that is truly centered around 1.

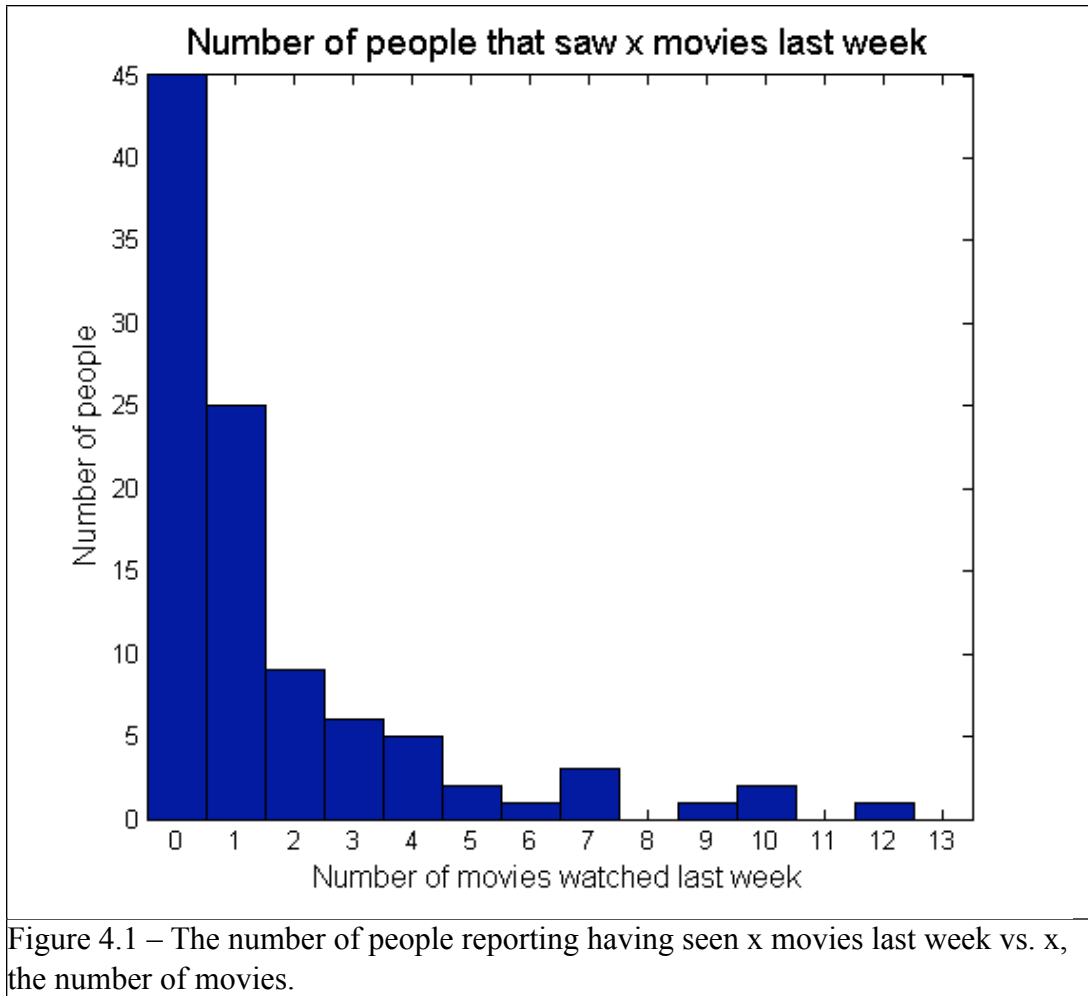


Figure 4.1 – The number of people reporting having seen x movies last week vs. x , the number of movies.

With the MATLAB Statistics Toolbox, we can simply do a one-sample t-test of the hypothesis that our measured mean is “significantly” greater than 1 without the need to understand the principle of generating sampling distribution under CLT. We type:

```
>> [h,p,ci,stats] = ttest(moviewatch, 1, .05, 'right')
```

and then look at the output:

```
h =
    1

p =
    0.0092

ci =
    1.1812
    Inf

stats =
    tstat: 2.3962
     df: 99
     sd: 2.4622
```

If you remember what a t-test is about, this should be fairly clear even if you are new to MATLAB. If you are rusty on the t-test, however, what the above command is saying is “test to see if the mean of ‘moviewatch’ is greater than a mean of 1.0”. Don’t worry about the details of the t-test; in later chapters, once we have become comfortable with the concept of sampling distributions, we will revisit a few of the popular traditional statistical tests.

What the output is telling us is that, if the true mean were 1 movie per week, and the data were distributed normally, and we were willing to accept the mean as a good measure of central tendency for these data, then there is about 1% chance (the p-value of 0.0092) that we would have seen a mean as large or larger than the one we actually obtained.

A more do-it-yourself approach, but one still reliant on the above assumptions, is following. First, we compute the standard deviation of the data, and then use it to compute the expected standard error of the sampling distribution of the mean using Central Limit Theorem:

```
>> mymean = mean(moviewatch)      % the mean
>> myn = length(moviewatch)      % compute number of samples
>> mysd = std(moviewatch)        % the standard deviation
>> myse = mysd./sqrt(myn)        % the standard error by CLT
```

Now we can picture what the sampling distribution of the mean should look like – we just need to draw a Gaussian distribution whose mean is our measured mean (1.590), and whose standard deviation is the standard error we just computed (0.246). We also know that around 95% of the distribution should fall between the mean and ± 2 standard errors, which is about 1.098 and 2.082. This gives us a way to check our drawing.

```
>> xvals = linspace(0,3);        % make and x-axis
>> distofmeans = normpdf(x, mymean, myse); % normal dist.
>> figure; plot(xvals, distofmeans) % plot it

% and draw a dashed line at x = 1 for reference
>> line([1 1], [0 max(distofmeans)], 'LineStyle',
':')
```

The result is shown in Figure 4.2, and should look very much like what you get when you enter the above commands. Notice that this analysis gives us qualitatively the same result as the traditional t-test: it looks fairly unlikely that our measured mean, 1.59, and a mean of 1 belong to the same distribution. To be more quantitative about this, we could compute the area of our sampling distribution less than a mean of 1:

```
>> normcdf(1, mymean, myse)

ans =
```

0.0083

And this gives us about a 1% chance of seeing a mean as small or smaller than 1 given that the true mean is equal to 1.59, our measured mean. Notice that we've asked the mirror image question from the traditional t-test – “could a value of 1 come from a distribution centered on 1.59?” vs. “could a value of 1.59 come from a distribution centered on 1?” – but it amounts to the same thing and we get the same answer when we assume the same standard error about these numbers.¹

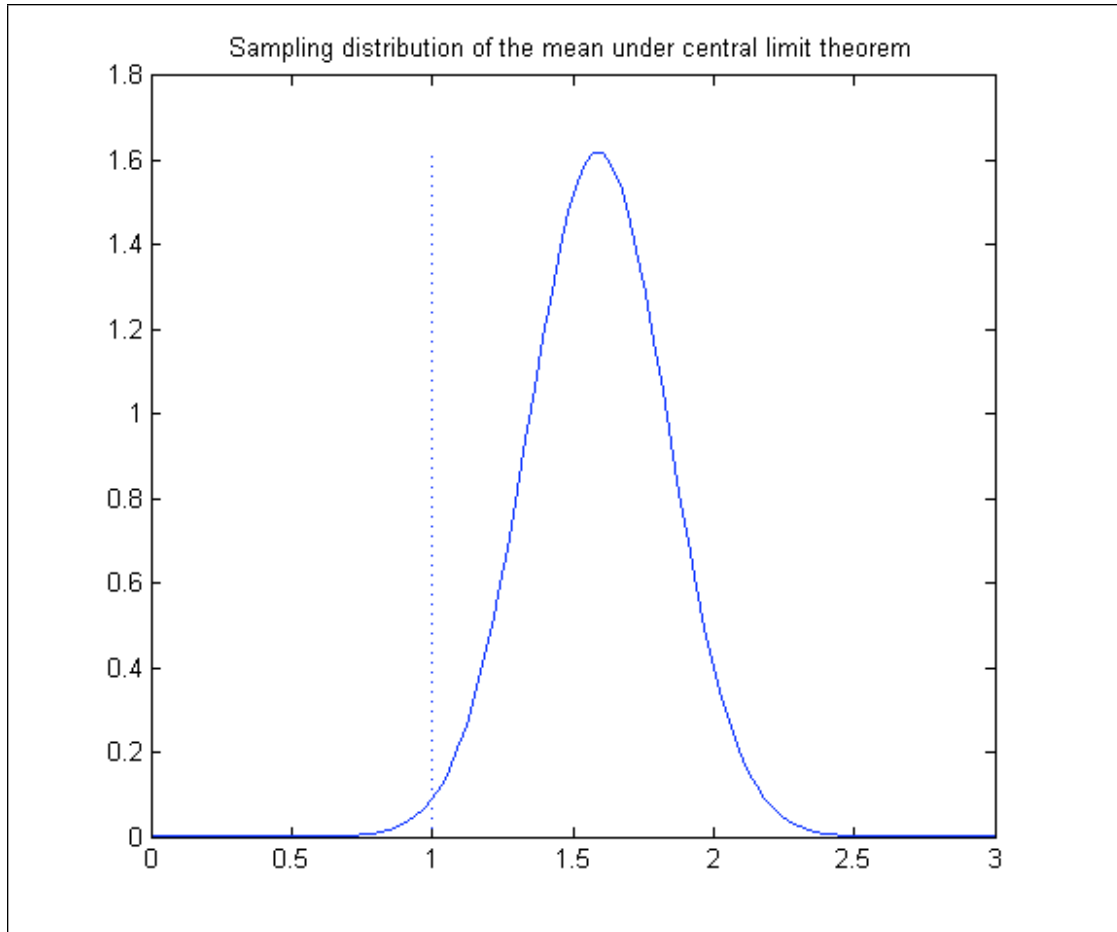


Figure 4.2 – The sampling distribution of the mean for the number of movies per week by Central Limit Theorem. The dashed line shows that an average of 1 movie per week is highly unlikely.

Alternatively, we can report our mean value with its 95% confidence interval. To compute the confidence interval under Central Limit Theorem we can use inverse normal probability density function `norminv()` or use the “ci” value from `ttest()` function.

¹ The small discrepancy comes from the fact that the t-test uses Gosset’s (i.e. Student’s) t distribution, rather than the standard normal distribution, which is technically correct when estimating the population variance from a sample variance. The difference is negligible for large ($n > 30$) sample sizes.

```
>> ci = norminv([.025, .975], mymean, myse)

ci =

    1.1074    2.0726
```

In English, this function call says “Give me the 2.5 and 97.5 percentiles of a normal distribution whose mean is the same as mean of my distribution and whose standard deviation is the same as standard error of my distribution”. Note that the 95% confidence interval does not include 1, and thus agrees with the previous analyses.

When using `ttest()` to get the intervals, we have to use the two tailed option:

```
>> [h,p,ci,stats]=ttest(moviewatch,1,.05);

>> ci =
    1.1014
    2.0786
```

Again, the same answer.

Thus ends our examination of these data using traditional methods. Now let us go back to Figure 4.1. It should be clear from a look at the original data that population distribution is almost certainly not normally distributed. This raises two questions. First, is our use of Central Limit Theorem still valid? That is, can we still use the formula $se = sd/\sqrt{n}$ to estimate the width of the sampling distribution? Second, and perhaps more importantly, is the mean in this case a good indicator of how many movies a typical person sees in a week? With a skewed data distribution such as this, the median would probably serve us better. There is, however, no equivalent of the CLT for the median, so we need to get the sampling distribution some other way. We have two choices: Monte Carlo simulation or the Bootstrap. If we had some theoretical reason to believe that the frequency of movie-going happened according to a particular law, we would know what distribution to use. We do not happen to know this, however, so we can either go “distribution fishing” to see if we can catch a distribution that matches our data, or we can simply use the Bootstrap. So, let's open up the MATLAB editor and write a loop very similar those used in previous chapters:

```
n = length(moviewatch);
nrep = 1000;           % number of bootstrapped resamples
bootmeds = zeros(nrep,1);
bootmeans = zeros(nrep,1);
for i = 1:nrep
    onebootsample = randsample(moviewatch, n, true);
    bootmeds(i) = median(onebootsample);
    bootmeans(i) = mean(onebootsample);
end
```

This bootstraps the median as well as the mean for comparison. Now let's look at them both (you can either put this in your program or type it at the command line):

```
figure;
subplot(2,1,1);
hist(bootmeds);
xlim([0, 3]);
title('bootstrap sample medians');
ylabel('frequency');
subplot(2, 1, 2);
hist(bootmeans);
xlim([0, 3]);
title('bootstrap sample means');
ylabel('frequency');
```

The resulting distributions are shown in Figure 4.3. Let us first examine the bottom part – sampling distribution of the mean obtained not from CLT and assumption of normality, but using bootstrap resampling from the original distribution. Although the original moviewatch distribution was highly skewed and far from normal, the sampling distribution of the mean looks quite normal, and very similar to the probability density function in Figure 2. To see if we get the same answer from the bootstrap analysis as we got from the traditional analysis, we can simply count number of simulations (out of 1000) that gave mean larger than 1:

```
>> sum(bootmeans>1)
ans =
    995
```

In other words, 995 out of 1000 simulations yielded mean above 1 and only 5 out of 1000 simulations yielded mean equal to or smaller than 1, suggesting that it is highly unlikely that the true mean would be equal to 1. Your exact value may be a little different because the bootstrap sampling is a random process, but the conclusion will be the same. To obtain 95% confidence interval from the bootstrapped sampling distribution, we will use `quantile()` function, which is the analog of `norminv()` function for empirical distributions:

```
>> quantile(bootmeans, [.025, .975])
ans =
    .1300    2.1000
```

Again, your exact numbers may be somewhat different, but should be very similar to the values obtained using traditional methods.

So, to answer our first question, we can conclude that the Central Limit Theorem is very robust and provides us with adequate sampling distribution of the mean even when

the original population distribution is not normally distributed, as long as the sample size is reasonably large.

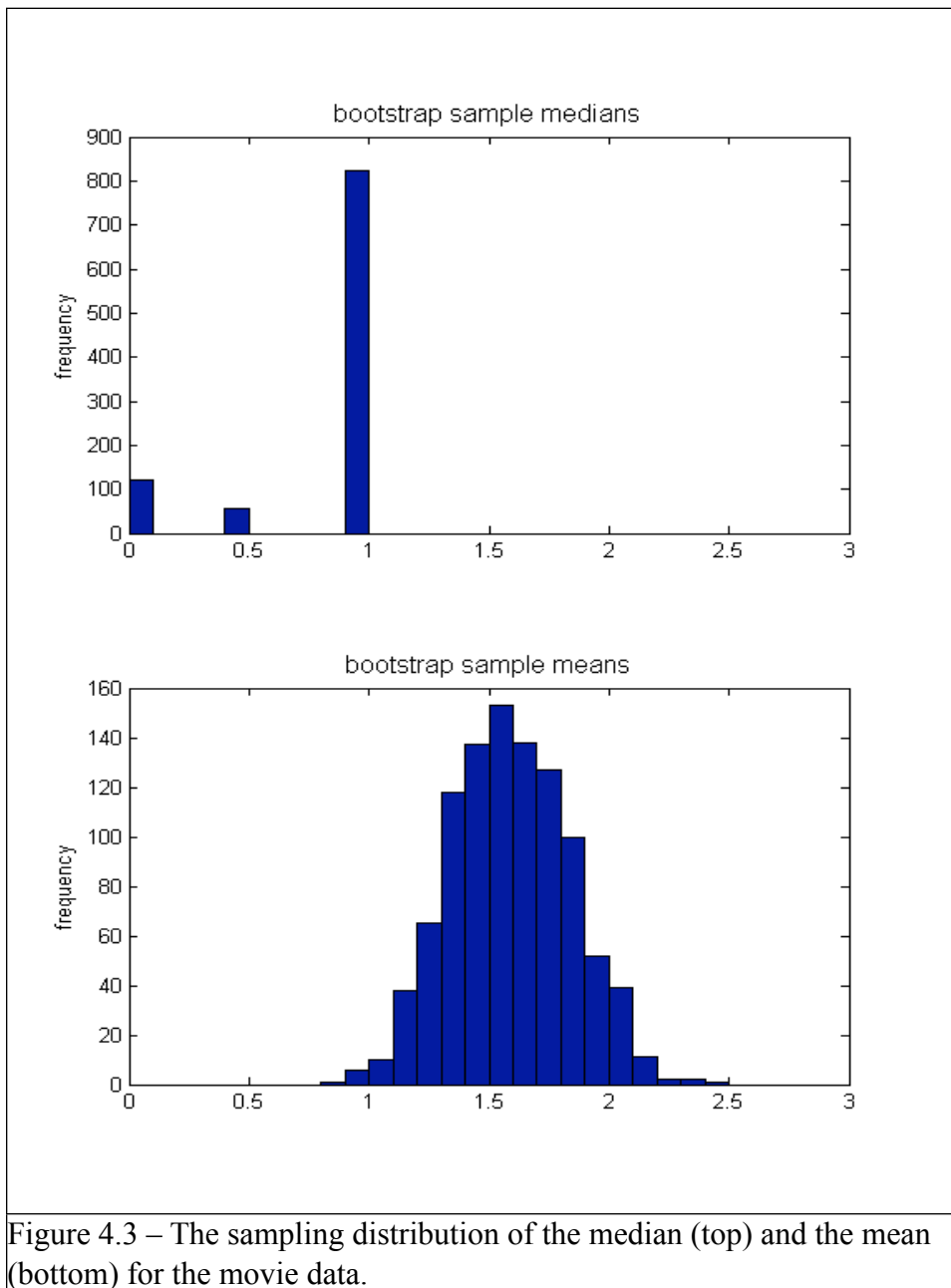


Figure 4.3 – The sampling distribution of the median (top) and the mean (bottom) for the movie data.

Now, should we be happy about the answer and conclude that in general, people watch more than one movie a week? Really? Even when 70% of the people actually watch zero or one movie only? As mentioned earlier, the median is worth considering as a measure of central tendency when a distribution is highly skewed. The sampling distribution of the median is shown in the upper panel of Fig. 4.3 and - wow – the distribution of the median is very different from that of the mean. All of the medians

correspond to 1 movie per week or fewer, with about 10% at 0, and a small percentage at 0.5 (corresponding to cases in which exactly 50 people saw no movies, and the other 50 saw at least one movie). More importantly, when we analyzed the mean using traditional methods (above), we concluded that a typical person saw more than one movie per week, because typical was to be represented by ‘mean’, the only measure of central tendency that CLT can address. Here, using a more reasonable analysis, *we come to exactly the opposite conclusion*; we expect the typical person to see one or fewer movies per week.

In summary, this example illustrates that the Bootstrap is not simply a different route to the same conclusions as would be reached by traditional methods. Its power lies in its ability to handle problems for which the traditional methods fail to provide an answer and/or lure us into a misleading analysis.

Are response times Gaussian?

As a second example, consider the data shown in Figure 4.4 (which we will also be using in the homework exercises and can be downloaded from the website). These are reaction times (in second) from a simple detection task where participant had to indicate which side of the screen an “x” symbol appeared. The distribution looks somewhat skewed. Moreover, the mean (dashed line) is somewhat larger than the median (dotted line), which is further evidence of skew.

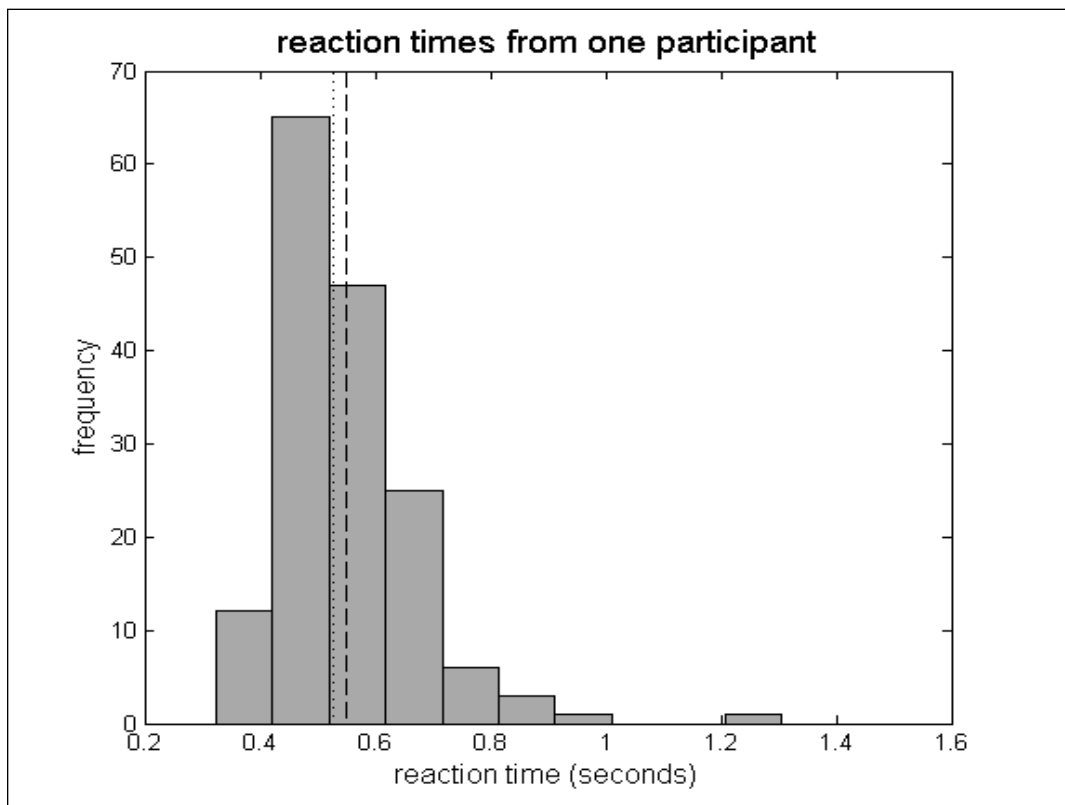


Figure 4.4 – A sample of 160 reaction times from one participant. Visual inspection indicates a potential positive skew.

We would like to know whether the skew we see in the data is significant or whether it might just represent a sample from a normal distribution that was skewed by the vagaries of chance. To do this, we can compute the skew for a lot of bootstrapped replications of these data, and see how much of this bootstrapped sampling distribution of the skew falls on the positive side of zero, which is the skew of a normal distribution by definition (were the skew in the other direction, we would instead be interested in how much of the distribution fell on the negative side of zero). Can we compute a sampling distribution of a skew? Of course we can. Although the concept of “sampling distribution” is most closely associated with the mean, we can compute a sampling distribution for any summary statistic that we can compute from our data. In fact, this is one of the main benefits of departing from the traditional methods; we are free from what has been called “the tyranny of the mean.”

The skew of a distribution is defined as

$$\frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s} \right)^3$$

This formula might look intimidating at first, but it should make sense if you think about it. First, note that it is just the sum of something divided by the number of addends (elements being summed). In other words, it is an *average*, just like the mean and standard deviation (the mean being the average value of the data, and the standard deviation being the average unsigned distance of the data points from the average value). But what is the skew an average of? First, notice that, in the numerator, the difference between each data point and the mean is computed (exactly as in the formula for variance or standard deviation). This difference, however, is cubed instead of squared, and this has two effects: it makes big differences *really* big, and it preserves the sign of differences. Thus, in a skewed distribution, the values on the extended tail become really big when cubed, and will thus dominate the sum, resulting in a positive value when the extended tail to the right of the mean, or a negative value when the extended tail is to the left. Second, notice that we take each difference and divide by the standard deviation. All this means is that we are converting the differences into units of standard deviation, just like in z-score or a t-value. This makes the skew impervious to the overall standard deviation of the distribution. Thus, computing the skew on the above RT distribution would give the same value whether the data were in seconds or in milliseconds, even though the standard deviation of the later would be numerically much larger (120 msec vs. 0.12 sec for our example distribution). This is as it should be, since skew is a measure of shape, not of dispersion (like the standard deviation), or location (like the mean).

At any rate, we can now easily get the sampling distribution of the skew via bootstrapping:

```
n = length(data);
nrep = 10000;
bootskew = zeros(nrep,1);

for i=1:nrep
    onebootsample = randsample(data, n, true);
    bootmean = mean(onebootsample);    %mean of resample
    bootstd = std(onebootsample);     %std of resample
    bootskew(i) = sum(((onebootsample - bootmean)/bootstd).^3)./n;
end

hist(bootskew);
```

And the resulting sampling distribution is shown in Figure 4.5. Note that the sampling distribution of skew is certainly not normally distributed and we could not obtain it in any other way than by bootstrap simulation. Can we still make inferences about the significance of the skewness when the sampling distribution is not normally distributed? Absolutely!

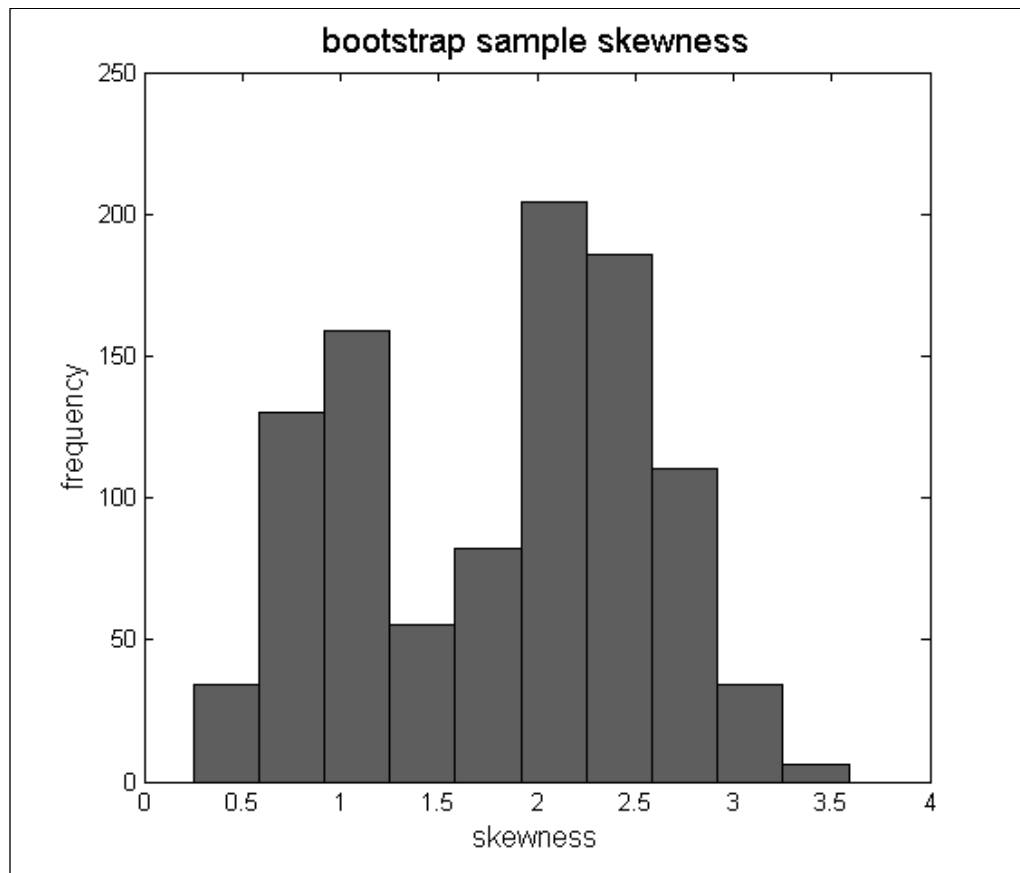


Figure 4.5 – A sampling distribution of skews obtained from 1000 bootstrap replications of the reaction time data shown in Fig. 4.4

As the entire domain (the extent along the x axis) of this distribution is positive, we can be very confident that the true skew of the population from which the original sample was drawn was greater than zero. To be even more precise, we can calculate values of skew that enclose 95% of the distribution – 95% confidence intervals:

```
>> quantile(bootskew, [.025 .975])
```

```
ans =
```

```
0.5485    3.0189
```

So it seems almost certain – as certain as we can be in statistics! – that our RT distribution is in fact skewed. You probably already learned in introductory psychology class that positive skew is very typical for individual reaction time data. As mean is highly susceptible to extreme values, and reaction times often contain extreme high values (participant was yawning, sneezing, or checking a cell phone instead of focusing on the task), it is much more appropriate to use median to characterize speed of participants responses.

Summary

In this chapter, we have learned how to take a single sample of data, compute some summary statistic and its sampling distribution on that sample, and decide whether the value of that summary statistic is in accord with some theory that specifies what the value of that summary statistic should be. For example, is the mean IQ of a sampled group different than the population IQ? Is the median amount of cell divisions of a sampled group greater than the amount of cell division known to occur normally? Is the mean weight gain of experimental animals less than the known mean weight gain of animals over the same time period?

We learned that the bootstrap method is an alternative to traditional statistics that can provide the same answers as traditional statistics when asked the same questions. Additionally, it can provide answers that traditional statistic cannot. Importantly, understanding the logic behind generating sampling distribution using Monte Carlo or bootstrap methods can provide a deeper understanding of the principles that gave rise to the formulas used in traditional statistics and that have become occluded by “blackboxiness” of the traditional statistics for most of us.

Why to use traditional statistics at all? If assumptions of CLT are met AND the mean is of interest, traditional statistics provides a valid answer. Additionally, the values taken from the traditional sampling distribution of the mean (p-values, confidence intervals etc) computed by anyone, anywhere will always be precisely the same, which is nice. Of course, if the assumptions are not met, however, the results from traditional statistics will be precisely wrong.

In the next chapter, we will learn how to draw conclusion about the differences between a summary statistics of two samples of data. For example, is the median weight

loss of an experimental diet group different than that of a control group maintaining their normal diet? Does the variability of weight loss of an experimental group exposed to the same diet increase relative to the control group? (If the variability increases by a substantial amount, then a decrease in the median weight could be offset by the fact that many patients are actually gaining weight, doing them potential harm.)

