BOOTSTRAPPING FOR LEARNING STATISTICS

Tim Hesterberg
Google, Seattle, WA, USA
timhesterberg@gmail.com

Collecting data, producing plots such as histograms and scatter plots, and calculating numerical statistics such as means, medians, and regression coefficients are relatively concrete operations. In contrast, ideas related to the random variability of those statistics—sampling distributions, standard error, confidence intervals, central limit theorems, hypothesis tests, P-values, and statistical significance—are relatively abstract, and more difficult for students to understand. Bootstrap methods and permutation tests take those concrete tools, that students are used to using with data, and apply them to sampling distributions. This promotes understanding. We demonstrate using two examples—one involving linear regression, the other comparing two sample means. We finish by discussing why the bootstrap works, and what to watch out for.

LINEAR REGRESSION EXAMPLE—BUSHMEAT

The goal of this section is to demonstrate the basic mechanics of the bootstrap and show how to use the bootstrap to get visual and numerical measures of variation.

The consumption of "bushmeat," the meat of wild animals, threatens the survival of some wild animals in Africa. This pressure might be reduced if alternative supplies of protein were available. Brashares et al. (2004) studied the relationship between fish supply and demand for bushmeat in Ghana. Part of their data is shown in the following table and Figure 1, data from 30 years of local fish supply and biomass of 41 species in nature preserves. They found a relationship between fish supply and the hunting of bushmeat, measured by the change in biomass; we investigate that relationship below. They also corroborated that relation with other data, such as the supply of bushmeat in local markets.

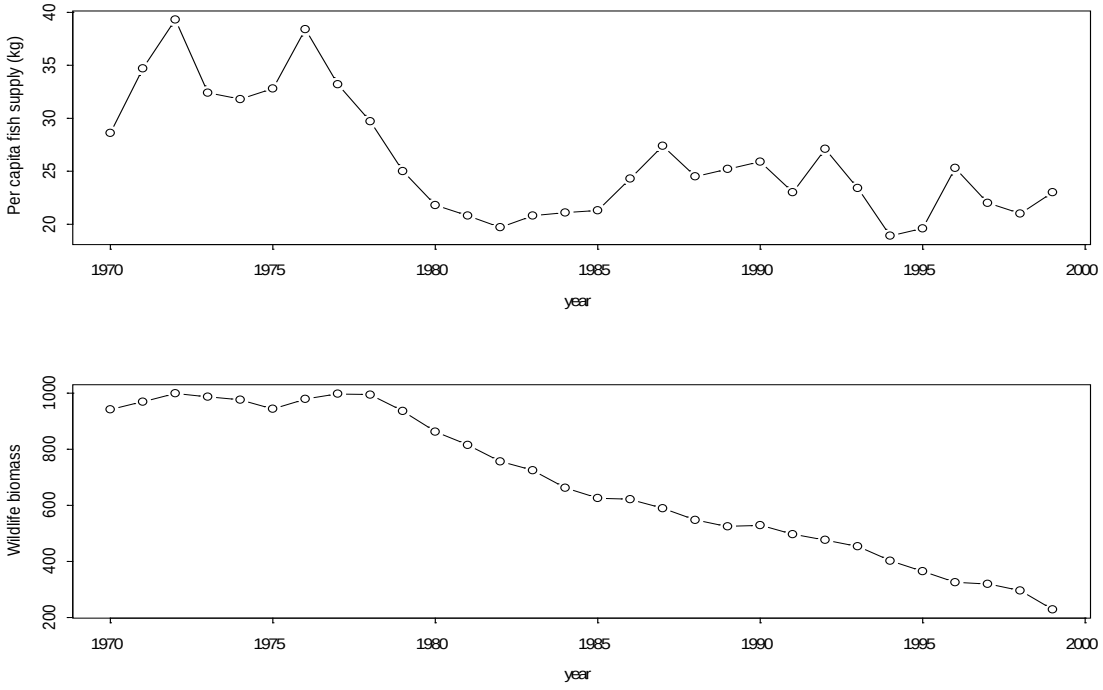

Fig. 1. *Bushmeat and Fish:* 30 years of data of local supply of fish per capita and biomass of 41 species in natural parks in Ghana

In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

Table 1: Bushmeat: Local Supply of Fish per Capita (kg), and Biomass of 41 Species in Nature Preserves

Year	Fish	Biomass	Year	Fish	Biomass	Year	Fish	Biomass
1970	28.6	942.54	1980	21.8	862.85	1990	25.9	529.41
1971	34.7	969.77	1981	20.8	815.67	1991	23.0	497.37
1972	39.3	999.45	1982	19.7	756.58	1992	27.1	476.86
1973	32.4	987.13	1983	20.8	725.27	1993	23.4	453.80
1974	31.8	976.31	1984	21.1	662.65	1994	18.9	402.70
1975	32.8	944.07	1985	21.3	625.97	1995	19.6	365.25
1976	38.4	979.37	1986	24.3	621.69	1996	25.3	326.02
1977	33.2	997.86	1987	27.4	589.83	1997	22.0	320.12
1978	29.7	994.85	1988	24.5	548.05	1998	21.0	296.49
1979	25.0	936.36	1989	25.2	524.88	1999	23.0	228.72

There is a general decline in biomass over the study period, with a steeper decline in years with less fish. Figure 2 shows a positive relationship between fish and change in biomass. The correlation is 0.67, and regression slope is 0.63, suggesting that an increase of 1 kg of fish per capita per year results in a gain (or less of a loss) of about two-thirds of one percent of biomass. The *y* intercept is -21.1, giving a prediction of a 21% decline in biomass were the fish supply to disappear. On the other hand, the *x* intercept of the regression line is 33.3, suggesting that 33.3 kg of fish per capita would suffice to forestall further wildlife declines.

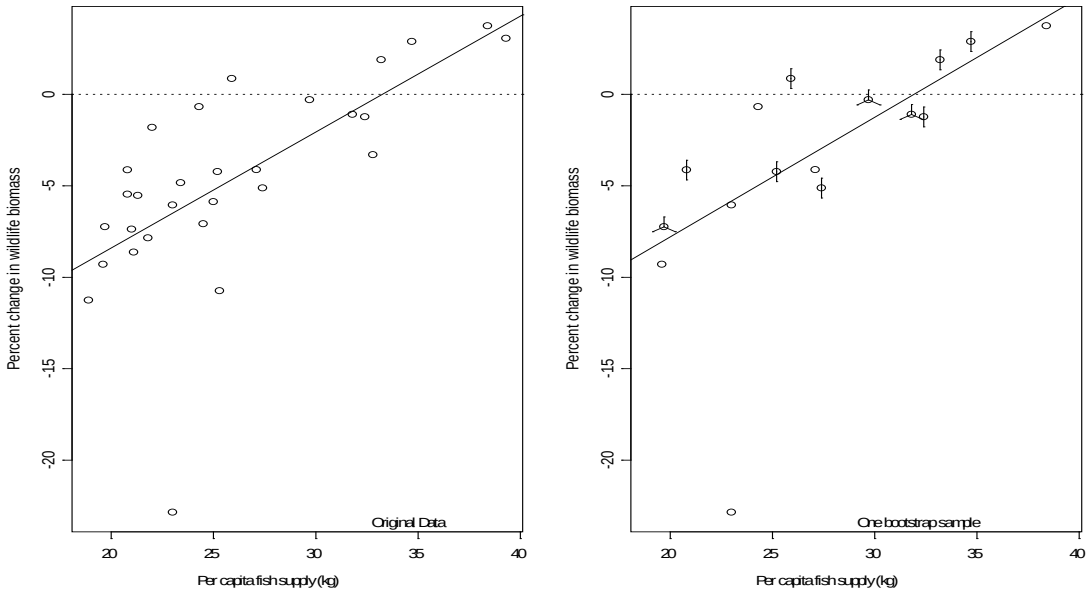

Figure 2. *Change in Biomass versus Fish Supply:* scatterplot of change in biomass versus fish supply for 29 years, with a linear regression line superposed. The left panel shows the original data, and the right panel depicts one bootstrap sample, as a "sunflower" plot, where the number of petals is the number of times an observation is repeated.

However, those estimates are obtained from a limited amount of data. How accurate are they? We use the bootstrap to estimate this. We begin by taking one *bootstrap sample*—a random sample with replacement, of the same size, from the original data. Here we pick 29 random years from the original 29 years (omitting 1971 because the change in biomass between 1970 and 1971 is unknown): 1995, 1982, 1989, 1990, 1973, 1990, 1982, 1987, 1973, 1974, 1978, 1974, 1978, 1987, 1983, 1982, 1977, 1978, 1991, 1983, 1971, 1992, 1976, 1977, 1999, 1986, 1989, 1971, and 1974. The right panel of Figure 2 shows this bootstrap sample. Because we are sampling with replacement, some of the original observations are omitted whereas others appear more than once. We compute the same statistics for this sample that we calculated for the original sample. For this

bootstrap sample, the correlation is 0.68, slope is 0.65, *y* intercept is –20.9, and the *x* intercept is 31.9. We repeat the process tens or thousands of times—drawing thousands of random bootstrap samples and computing the statistics of interest. We use the variability in these *bootstrap statistics* to estimate the variability in the original statistics.

Figure 3 shows two views of the bootstrap output. The left panel is a graphical bootstrap: regression lines for 25 bootstrap samples. We see how the regression lines vary. The farther to either side we look, the more the *y* value of the lines varies. This helps students see how extrapolation provides less accurate answers.

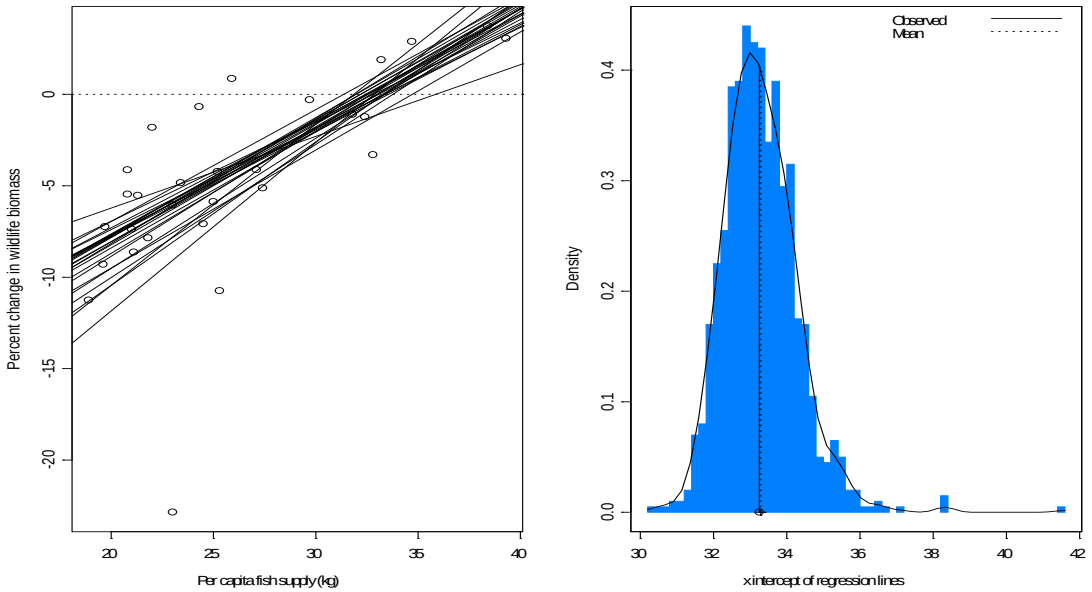

Figure 3. *Bootstrap Regression Lines and x Intercept:* (*L*) Regression lines calculated from 40 bootstrap samples of the bushmeat data; (*R*) Histogram and density curve for *x*-intercepts of regression lines from 1000 bootstrap samples (These are the estimated values of fish supply that would result in zero loss of biomass.)

The regression suggests that increasing the fish supply would reduce bushmeat harvest, and the *x*-intercept suggests that 33.3kg of fish would stop the loss of wildlife. We can use the bootstrap to get an idea how accurate that number is. We'll use more bootstrap samples for better accuracy; we create 1000 bootstrap regression lines, record where each line intercepts the *x* axis, and plot a histogram of those *x* intercepts, to obtain the right panel of Figure 3. The original value of 33.3 falls in the middle of this distribution. The middle 95% range is from 31.6 to 35.5, giving a rough idea of the reliability of the estimate. We are reasonably confident (95%) that the supply of fish needed to forestall loss of biomass lies in that interval, assuming that historical data is representative of the future (other factors such as global warming and population growth could change matters). The interval (31.6, 35.5) = (33.3 – 1.7, 33.3 + 2.2) stretches farther to the right, which tells a pessimistic tale—it takes more fish to gain confidence on the positive side than to lose confidence on the negative side.

To summarize, we draw random samples from the data with replacement (bootstrap samples), and compute the statistic(s) of interest (like the *x* intercept). The variation of these bootstrap statistics gives an idea of the accuracy of the original statistic(s). The range of the middle 95% of the bootstrap statistics gives a 95% confidence interval for the true unknown parameter; this is known as a *bootstrap percentile interval*. While there are more accurate bootstrap intervals, this one is good for helping intro stats students understand the idea of a confidence interval.

The bootstrap SE for the prediction at any value of x is the standard deviation of the regression lines at that x. For comparison, the classical estimate of the standard error is $s\sqrt{1/n+(x-\bar{x})^2/\sum_i (x_i-\bar{x})^2}$. Not only is the picture more intuitive for students, it is also more accurate – the classical formula assumes homoscedasticity, but here the residuals are more variable

when *x* is smaller. This makes sense in the context of the application; when *x* is large, there is less hunting, and less variation in hunting pressure.

TWO-SAMPLE MEANS EXAMPLE – TV ADVERTISEMENTS

We look next at data collected by a student for a statistics project, comparing time spent on commercials in basic and extended (extra-cost) cable channels. We'll start by bootstrapping the mean of one sample, then the difference in means of two samples, and finally test whether the difference is statistically significant using a permutation test. Here are the data, from apstats.4t.com.

Table 2: Number of Minutes of Commercials during Random Half-Hour Periods from 7A.M. to 11 P.M.

Basic	7.0	10.0	10.6	10.2	8.6	7.6	8.2	10.4	11.0	8.5
Extended	3.4	7.8	9.4	4.7	5.4	7.6	5.0	8.0	7.8	9.6

The means of the basic and extended channel commercial times are 9.21 and 6.87, respectively, a difference of 2.34 minutes. How accurate are these numbers? There is not much data—the poor student could only stand to watch 10 hours of random TV!

The average for the basic channel is 9.21 minutes. To assess the accuracy of this number by bootstrapping, we draw bootstrap samples of the same size as the original data (10 observations with replacement from the original 10 basic-channel numbers), and compute the mean for each bootstrap sample. I would normally show a histogram, but omit that here to save space; it appears approximately normal, centered at 9.21, with a standard deviation of 0.42 (the sample standard deviation of the bootstrap means).

What we have just done, in a perfectly natural way, is to calculate the *bootstrap standard error*. The bootstrap standard error (SE) is the standard deviation of the bootstrap distribution. A SE, by definition, is an estimate of the standard deviation of a statistic. Students who are taught that the formula for the standard error of a mean is $s/\sqrt{n}$ may never truly understand that a SE is the standard deviation of the distribution of a statistic, and how that is distinct from the distribution of data. The bootstrap reinforces this idea by calculating the SE directly from a distribution for the statistic.

We follow the same process for the extended channel; this histogram is also approximately normal and centered at the original sample mean; the bootstrap SE is 0.63. For comparison, the formula SEs are 0.44 and 0.67; the bootstrap standard errors are a bit shorter, but otherwise the bootstrap and formula standard errors are quite similar. This similarity lets students check their work.

Now consider the problem we are really interested in—comparing the two samples. To bootstrap the difference in means, we draw bootstrap samples from each original sample independently, compute the difference in means, and repeat say 1000 times. The 1000 differences in means comprise the bootstrap distribution for the difference in means. This distribution is shown in the left panel of Figure 4. The distribution is centered at the original difference in means, 2.34 minutes, and is approximately normal with a standard deviation of 0.76 (the bootstrap SE). I believe that the combination of the picture and standard deviation is more intuitive for students than is the formula SE $\sqrt{s_1^2/n_1 + s_2^2/n_2}$. The distribution is approximately normal, illustrating the Central Limit Theorem.

Before turning to permutation tests, we share two thoughts:

- It is interesting for students to see histograms and normal quantile plots of bootstrap distributions from a series of samples with the same skewness but different sample sizes, to see how well the CLT works with different sample sizes.
- It is good for students to see bootstrap distributions from data with different sample sizes but similar sample standard deviations *s*, to see how the standard error changes with *n*.

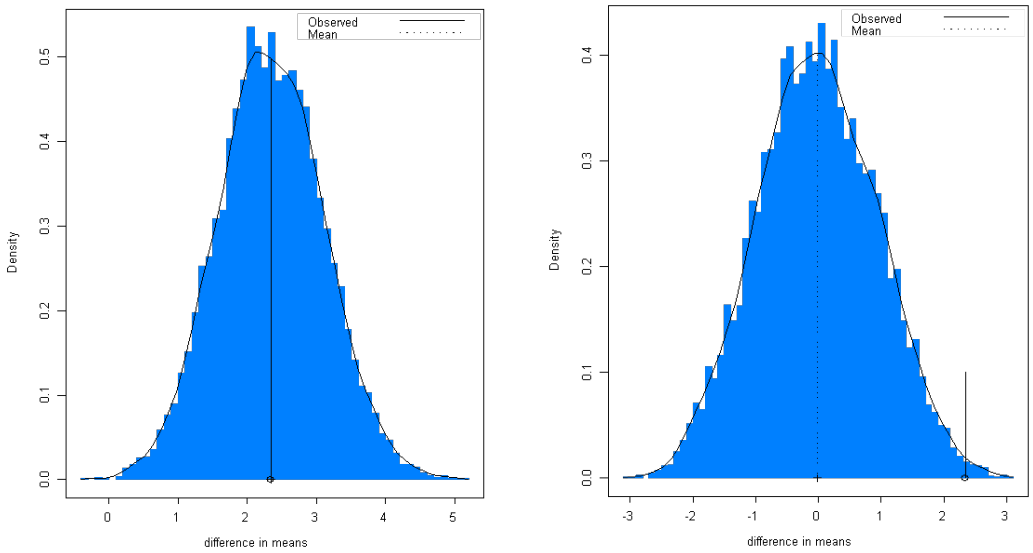

Figure 4. *Sampling Distributions for Difference in Advertisements:* left: histogram of the bootstrap distribution for the difference in mean advertisement time, in basic versus extended TV channels. Right: same, for the permutation distribution.

Permutation Tests for P-Values

So far the data appear to support the student's working hypothesis, that basic cable TV channels have more advertising than extended channels. But is the difference statistically significant? Or could it have easily occurred just by chance? The bootstrap percentile interval, the middle 95% of the bootstrap distribution, excludes zero, suggesting that more than chance is involved. But a formal test would be better.

Statistical significance is measured using a *P*-value—the probability of observing a statistic this large or larger, if the null hypothesis is true. In other words, assuming there is no real difference between basic and extended channels, how often would a difference of 2.34 minutes or more occur just by chance?

To answer this question we must resample in a way that is consistent with the null hypothesis. Suppose there is no real difference between the two populations, that there is really just one population. Then we may pool all observations to form an estimate of the combined population and draw samples from that. A *permutation sample* is created by drawing n_1 observations *without* replacement from the pooled data to label as one sample, leaving the remaining n_2 observations as the second sample. We calculate the statistic of interest, e.g. the difference in means of the two samples. We repeat this many times—1000 or more. The one-sided *P*-value is 0.0054, the fraction of times the random statistic exceeds the original statistic, the area to the right of 2.54 in the right panel of Figure 4. In this case we use a one-sided test, because the student's hypothesis was that basic channels would have more commercials. For a two-sided test we multiply the one-sided *P*-value by two.

One pedagogical advantage of this procedure is that it gives students a visual picture of the *P*-value. Another advantage is that it works directly with the statistic of interest, the difference in sample means, rather than forcing students to transform that statistic of interest into a *t*-statistic so they can use a *t*-table to calculate a *P*-value. A third is that it provides students a way to check their work; when permutation distributions are approximately symmetric, the *t*-test *P*-value should be close to the permutation-test *P*-value. In this case they are close; the *t*-test *P*-value is 0.00498.

What if the permutation and *t*-test *P*-values differ? Then the *t*-test is wrong, sometimes dramatically so, e.g. by a factor of 4 in a consulting project (Hesterberg 2002). This occurs when the permutation distribution is skewed, due to skewness in the data, and with unequal sample sizes so the skewness does not cancel out. The permutation test is the gold standard, accurate even for very small samples. Indeed, Fisher originally justified the *t*-test as an approximation to the permutation test, in the pre-computer era when permutation distributions were difficult to compute.

WHY THE BOOTSTRAP WORKS

A sampling distribution is the distribution of a statistic when drawing random samples from a population. It would be nice to estimate the sampling distribution by doing exactly that—drawing samples from the population and computing the statistic, and repeating many times. The problem is that the population is unknown, or it would be too expensive to draw repeated samples from that population.

The bootstrap substitutes an estimate for the population for the population. We draw samples from that estimate, compute the statistic, and repeat many times.

In the usual *nonparametric bootstrap*, the estimate is the empirical distribution, i.e. we draw samples from the data. In upper-level courses we may use a *parametric bootstrap* instead, in which we estimate parameters from the original data, then draw from parametric distributions with those parameters.

The fact that the bootstrap mimics real life offers a variety of pedagogical benefits. It reinforces the role that random sampling plays in statistics. It lets us investigate what would happen with different sampling methods, e.g. stratified sampling. We can do "what-if" analyses, e.g. to see how sample size matters we can draw bootstrap samples of different sizes (with replacement from the same original sample).

When the Bootstrap Fails; When it is Better

The bootstrap fails when the empirical distribution is a poor estimate of the population. This is particularly true in small samples. For a single mean, the bootstrap percentile interval is like a *t*-interval, but using z_α instead of t_α, calculating *s* with a divisor of *n* instead of *n*-1, and making a skewness adjustment based on skewness estimated from the sample. That does not work well in small samples.

People tend to think of bootstrapping for very small samples, and rely on classical methods for medium and larger samples. That is turned around. The bootstrap relies on the data to tell the shape of the population. In very small samples the data cannot do that reliably, and the bootstrap is too variable. Conversely, in medium samples, people do not realize how poor classical methods are. For example, it requires $n = 5000$ before the 95% one-sample *t*-interval is reasonably accurate, if the population has the skewness of an exponential distribution (Hesterberg 2008).

CONCLUSION

We presented two examples that demonstrate the pedagogical value of resampling, for understanding sampling distributions, standard errors, P-values, confidence intervals, the Central Limit Theorem, prediction accuracy and extrapolation.

For further examples and discussion of resampling for teaching, see Hesterberg et al. (2005) for introductory statistics, Chihara and Hesterberg (2011) for mathematical statistics, and other articles at http://www.timhesterberg.net/bootstrap.

REFERENCES

Brashares, J. S., Arcese, P., Sam, M. K., Coppolillo, P. B., Sinclair, A. R. E., & Balmford, A. (2004). Bushmeat hunting, wildlife declines, and fish supply in West Africa. *Science, 306*(5699), 1180-1183.

Chihara, L.. & Hesterberg, T. (2011) *Mathematical Statistics with Resampling and R*. Wiley.

Hesterberg, T. (2002). Performance evaluation using fast permutation tests. In B. Gavish (Ed.), *Proceedings of the Tenth International Conference on Telecommunication Systems Management* (pp. 465–74). Monterey, CA: American Telecommunication Systems Management Association.

Hesterberg, T. (2008). It's time to retire the "n >= 30" rule. *Proceedings of the American Statistical Association, Statistical Computing Section*.

Hesterberg, T., Moore, D. S., Monaghan, S., Clipson, A., & Epstein, R. (2005). *Bootstrap methods and permutation tests* (2nd ed). W. H. Freeman.