

Hints on Exercise 4.5

I would suggest first reading through this whole thing without reference to the text. Then, laying this aside, go back and make sure you can work the exercise in the text. The notation in the text and in the solution was not consistent. I am using a more consistent notation in what follows, but it's almost the same as the text.

Let $S = \{y_1, \dots, y_n\}$ be a random sample from a population with mean μ , variance σ^2 , and distribution function P . Let \hat{P} be the empirical distribution function. Let \bar{y} be the sample mean for S .

The basic idea of resampling is that now we have a discrete uniform distribution. It has the probability mass function,

$$\begin{aligned}\hat{p}(x) &= \frac{1}{n} \quad \text{for } x = y_1, \dots, y_n \\ &= 0 \quad \text{otherwise.}\end{aligned}$$

The mean of this distribution is

$$\mu_{\hat{P}} = \bar{y},$$

and the variance is

$$\sigma_{\hat{P}}^2 = \frac{1}{n} \sum_i (y_i - \bar{y})^2.$$

Let $S^* = \{y_1^*, \dots, y_n^*\}$ be a random sample taken with replacement from S . Let \bar{y}^* be the sample mean for S^* .

For the questions in this exercise, we must first recognize there is some sloppy notation. No distinction is made between a random variable and its realization. We usually make this distinction by use of upper and lower case letters. Here, the author (me) has carelessly not made this distinction. (Unfortunately, almost no one uses precise notation in this context.) When we speak of an expected value of something, we usually want that something to be a random variable.

In these problems if S is a random sample and we want to talk about expectations of statistics computed from it, to use careful notation, we should denote its elements by Y_i , instead of y_i .

Likewise, the bootstrap samples S_j^* : if we want to talk about expectations of statistics computed from them, we should denote their elements by Y_i^{*j} .

Expectations relating to the empirical distribution are conditional distributions (conditional on the given sample). We denote these expectation operators as $E_{\hat{P}}(\cdot)$. We could also denote them as $E_S(\cdot)$.

Throughout this exercise, another thing to remember is if a random variable X has distribution D with mean μ_D and variance σ_D^2 , and we have a random sample of size n , using an expansion like that above, we can work out

$$\begin{aligned}E_D(\bar{X}^2) &= (E_D(\bar{X}))^2 + V_D(\bar{X}) \\ &= \mu_D^2 + \sigma_D^2/n,\end{aligned}$$

and

$$E_D \left(\sum_i (X_i - \bar{X})^2 \right) = (n-1)\sigma_D^2.$$

Part (a)

Show that

$$E_{\hat{P}}(\bar{y}^*) = \bar{y}.$$

The solution on page 380 is correct, except that “ x ” should be “ y ”, and “ \bar{x}_b ” should be “ \bar{y}^* ”. (The notation used in the solution is a common notation used in discussing the bootstrap. I used that notation a long time ago.)

At this point you might also want to work out the conditional variance (that is, the variance conditional on the original sample S),

$$\begin{aligned} V_{\hat{P}}(\bar{Y}^*) &= V_{\hat{P}}(Y_i^*)/n \\ &= \sigma_{\hat{P}}^2/n \\ &= \frac{1}{n^2} \sum_i (y_i - \bar{y})^2, \end{aligned}$$

and the expectation of the square,

$$\begin{aligned} E_{\hat{P}}((\bar{Y}^*)^2) &= \left(E_{\hat{P}}(\bar{Y}^*) \right)^2 + V_{\hat{P}}(\bar{Y}^*) \\ &= \bar{y}^2 + \frac{1}{n^2} \sum_i (y_i - \bar{y})^2, \\ &= \frac{n-1}{n} \bar{y}^2 + \frac{1}{n^2} \sum_i y_i^2. \end{aligned}$$

These expressions could be used in Part (c).

Part (b)

Show that

$$E_P(\bar{y}^*) = \mu.$$

One way of thinking of this (and of the question in Part (d)) is as an “overall” expectation of a conditional expectation:

$$\begin{aligned} E_P(\bar{Y}^*) &= E_{P\hat{P}}(\bar{Y}^*) \\ &= E_P \left(E_{\hat{P}}(\bar{Y}^*) \right) \\ &= E_P(\bar{Y}) \\ &= \mu. \end{aligned}$$

(Note what may appear to be an anomaly here: $E_P(\cdot) = E_{P\hat{P}}(\cdot)$.)

Another way of thinking about Parts (a) and (b) is as an expectation with respect to the original distribution at each point in the sum that forms \bar{Y}^* :

$$\begin{aligned} E_P(\bar{Y}^*) &= E_P\left(\frac{1}{n}\sum_i Y_i^*\right) \\ &= \frac{1}{n}\sum_i E_P(Y_i^*) \\ &= \mu. \end{aligned}$$

Note that in Parts (a) and (b) above there was no replication of the bootstrap sampling. Now, suppose that we take m samples S_j^* . Each of these m samples is just like the one in Parts (a) and (b), and furthermore, the samples are independent of each other. For each of these samples, we compute \bar{y}^{*j} , and compute

$$V = \frac{1}{m-1}\sum_j \left(\bar{y}^{*j} - \bar{\bar{y}}^*\right)^2, \quad (1)$$

where $\bar{\bar{y}}^* = \frac{1}{m}\sum_{j=1}^m \bar{y}^{*j}$. Notice that this is the same as the overall mean, $\frac{1}{mn}\sum_{j=1}^m \sum_{i=1}^n y_i^{*j}$. Hence, equation (1) is like the “mean squares among groups” in analysis of variance. *The basic decomposition of sums of squares in AOV prevades statistics.* You should immediately work out the “total sum of squares” and the “within sum of squares” for this problem.

Remember that \hat{P} is the ECDF; that is, sampling from \hat{P} is the same as sampling from a discrete uniform distribution over the original sample.

Treat each sample S_j as a random sample of size n from this this distribution. (In this context, it is correct, but redundant, to say “ S_j as a simple random sample with replacement”. The fact that the number of mass points, n , and the size of the sample S_j is n is irrelevant for purposes of the exercise.) You should just approach the problem in this general form, and work out things relating to $E_{\hat{P}}$ in terms of the parameters $\mu_{\hat{P}}$ and $\sigma_{\hat{P}}^2$. Then, for things relating to E_P , the parameters of the ECDF become random variables, as in the previous questions.

The only slight complication in this is that we take multiple samples, and compute statistics from each sample separately. This does not cause any additional problem, however, because the samples are independent. Thus if $j \neq k$, then \bar{Y}^{*j} is independent of \bar{Y}^{*k} .

Part (c)

Derive $E_{\hat{P}}(V)$.

First, expand the terms:

$$E_{\hat{P}}(V) = E_{\hat{P}}\left(\frac{1}{m-1}\sum_j \left(\bar{Y}^{*j} - \bar{\bar{Y}}^{*j}\right)^2\right)$$

$$= \frac{1}{m-1} \left(\sum_j E_{\hat{P}}((\bar{Y}^{*j})^2) - m E_{\hat{P}}((\overline{\bar{Y}^{*j}})^2) \right) \quad (2)$$

At this point, we could expand this further and continue. Remember that the expansion of something like $(\sum_i x_i)^2$ involves something like

$$\sum_i x_i^2 + \sum_{i \neq j} x_i x_j,$$

and that if $i \neq j$, then x_i and x_j are independent realizations. This allows you to take expectations of various quantities. For example, we could write $(\overline{\bar{Y}^{*j}})^2$ as

$$\frac{1}{m^2} \sum_j (\bar{Y}^{*j})^2 + \frac{1}{m^2} \sum_{j \neq k} \bar{Y}^{*j} \bar{Y}^{*k},$$

and then work out the individual expectations. So we would have for equation (2)

$$E_{\hat{P}}((\overline{\bar{Y}^{*j}})^2) - \frac{1}{m} \left(E_{\hat{P}}(\bar{Y}^{*j}) \right)^2.$$

We have already worked out both of these.

We could also get this by substituting directly into equation (2). You don't have to expand everything!

Think about what the original expression is in terms of the random variables \bar{Y}^{*j} . These random variables have a variance of σ_P^2/n , so if we take a sample of them, and compute an expression as in equation (1), what is its expected value?