Simulations in statistical inference

Recall the confidence interval for the population mean:

$\bar{x} \pm z\,\mathsf{SE}(\bar{x})$

The idea was that $\bar{x}$ follows approximately the normal curve.

What if we are interested in an estimator $\hat{\theta}$ for some parameter θ and the normal approximation is not valid for the estimator $\hat{\theta}$? What if there is no formula for SE($\hat{\theta}$)?

In such situations, simulations can often be used to estimate these quantities quite well. In fact, simulations may result in better estimates even in cases where the normal approximation is applicable!

The Monte Carlo Method

What is the average height of all people living in the United States?

This is difficult to determine exactly but can easily be estimated quite well:

Sample $n = 100$ (say) people at random. Then use the average height of these n people as an estimate of the average height of all people in the US.

This is an example of the general problem where we are interested in a unknown **parameter** θ of a population.

We estimate θ with a **statistic (estimator)** $\hat{\theta}$ which is based on a sample of n observations $X_1, \ldots, X_n$ drawn at random from the population:

$\hat{\theta} =$ average of the sample $= \frac{1}{n} \sum_{i=1}^{n} X_i$

The Monte Carlo Method

$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i$ tends to be close to the uncomputable population mean θ, even for moderate sample sizes such as $n = 100$.

This example is a special case of the **Monte Carlo Method** or **Simulation**:

- We approximate a fixed quantity θ by the average of independent random variables that have expected value θ.
- By the law of large numbers, the approximation error can be made arbitrarily small by using a large enough sample size.

The Monte Carlo Method

The Monte Carlo Method can also be used for more involved quantities. For example, we can use it to compute the standard error (SE) of a statistic $\hat{\theta}$.

Recall that the standard error tells roughly how far off the statistic will be from its expected value. The precise definition is

$$\mathrm{SE}(\hat{\theta}) = \sqrt{\mathrm{E}(\hat{\theta} - E(\hat{\theta}))^2}.$$

- Get many (say 1,000) samples of 100 observations each.
- Compute $\hat{\theta}$ for each sample, resulting in 1,000 estimates $\hat{\theta}_1, \ldots, \hat{\theta}_{1000}$.
- Compute the standard deviation of these 1,000 estimates:

  $$s(\hat{\theta}_1, \ldots, \hat{\theta}_{1000}) = \sqrt{\frac{1}{999} \sum_{i=1}^{1000} (\hat{\theta}_i - \mathrm{ave}(\hat{\theta}_i))^2} \quad \bar{\hat{\theta}}$$

Note that this is not an average of independent random variables. But it can be shown that the law of large numbers still applies and Monte Carlo works: $s(\hat{\theta}_1, \ldots, \hat{\theta}_{1000}) \approx \mathrm{SE}(\hat{\theta})$.

We can use Monte Carlo only if we can draw many samples of size 100!

The Bootstrap principle

We have an estimate $\hat{\theta}$ for a parameter θ and want to know how accurate $\hat{\theta}$ is:
we would like to find SE$(\hat{\theta})$ or give a confidence interval for θ.

The bootstrap can do this in quite general settings.

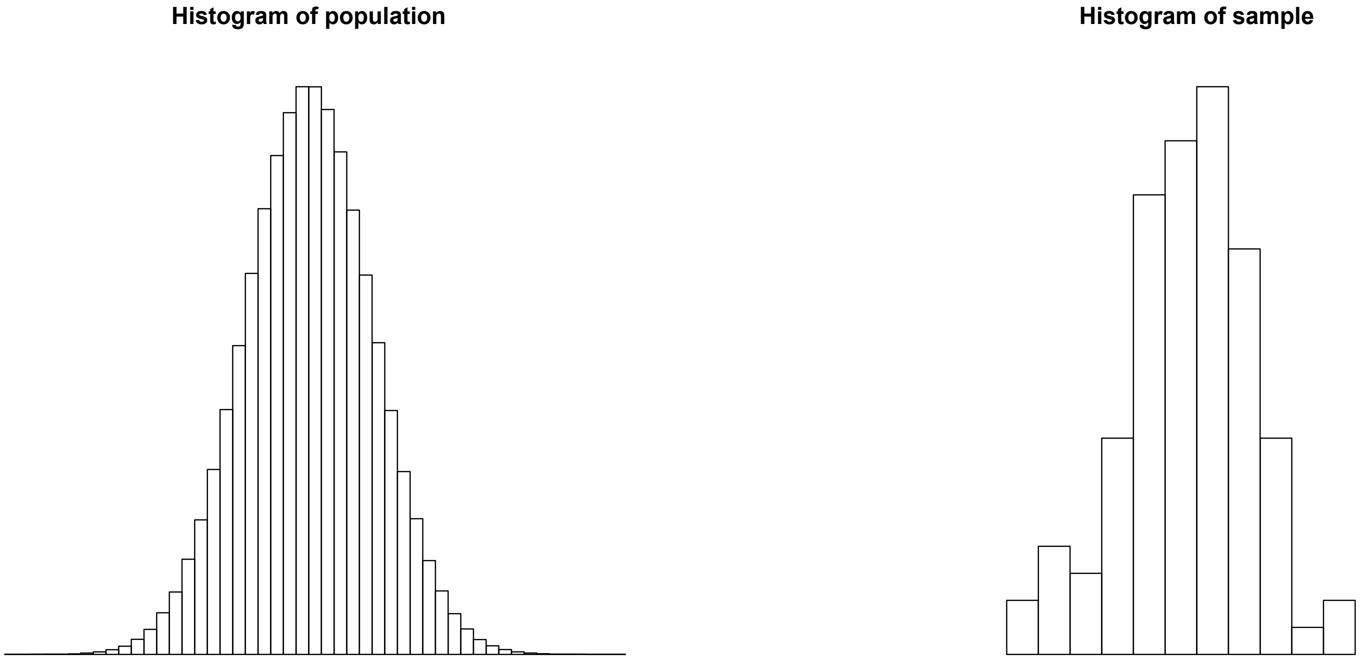
Example: $\theta =$ average height of all people in the US.
θ is unknown but can be estimated by the average height $\hat{\theta}$ of 100 randomly selected people.

This illustrates the **plug-in principle**:

We can't compute the population mean because we can't access the whole population.
So we 'plug in' the sample in place of the population and compute the mean of the sample instead.

The Bootstrap principle

The rationale for the plug-in principle is that the sample mean $\hat{\theta}$ will be close to the population mean θ because the sample histogram is close to the population histogram.



The bootstrap principle

The bootstrap uses the plug-in principle and the Monte Carlo Method to approximate quantities such as SE$(\hat{\theta})$.

Here is the reasoning behind the bootstrap:

Suppose we can draw as many samples from the population as we wish. Then we can approximate SE$(\hat{\theta})$ with Monte Carlo:

- Draw a sample $X_1, \ldots, X_n$ and use it to compute $\hat{\theta}$.
- Repeat B times (say B=1,000) to get $\hat{\theta}_1, \ldots, \hat{\theta}_B$.
- The standard deviation of these B estimates is close to SE$(\hat{\theta})$ if B is large, by the law of large numbers.

However, we have only one sample $X_1, \ldots, X_n$ and we can't simulate more because the population is not accessible.

The bootstrap uses the plug-in principle to get around this: It simulates from the sample instead of from the population.

① Draw a sample from the population.

② Draw bootstrap samples from the sample.

The bootstrap principle

③ Compute the estimator for each of the samples. ④ Use copies of estimators to approximate the value of interest.

The bootstrap pretends that the sample histogram is the population histogram and then uses Monte Carlo to simulate the quantity of interest.

Simulating a bootstrap sample $X_1^*, \ldots, X_n^*$ means that we draw n times with replacement from $X_1, \ldots, X_n$.

The bootstrap consists of two steps:

- Draw B bootstrap samples and compute $\hat{\theta}^*$ for each bootstrap sample:

$$\begin{array}{l} X_1^{*1}, \ldots, X_n^{*1} \ \rightarrow \ \hat{\theta}_1^* \\ \qquad \vdots \\ X_1^{*B}, \ldots, X_n^{*B} \ \rightarrow \ \hat{\theta}_B^* \end{array}$$

- Use $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$ to approximate the quantity of interest.
  For example, we approximate SE$(\hat{\theta})$ by the standard deviation of $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$.

More about the bootstrap

The **nonparametric bootstrap** simulates a bootstrap sample $X_1^*, \ldots, X_n^*$ by drawing with replacement from $X_1, \ldots, X_n$.

Sometimes a parametric model is appropriate for the data, e.g. a normal distribution with unknown mean and standard deviation. Then one may be better off with the **parametric bootstrap**, which simulates the bootstrap samples from this model, using estimates for the unknown parameters.

So far, the bootstrap samples were drawn independently. If there is dependence in the data (time series), then this needs to be incorporated, e.g. with the **block bootstrap**.

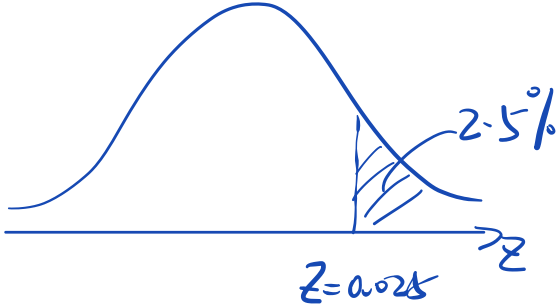
Bootstrap confidence intervals

If the sampling distribution of $\hat{\theta}$ is approximately normal, then

$$\left[\hat{\theta} - z_{\alpha/2}\,\mathrm{SE}(\hat{\theta}), \hat{\theta} + z_{\alpha/2}\,\mathrm{SE}(\hat{\theta})\right]$$

$\alpha=5\%$

is an approximate $(1-\alpha)$-confidence interval for θ.



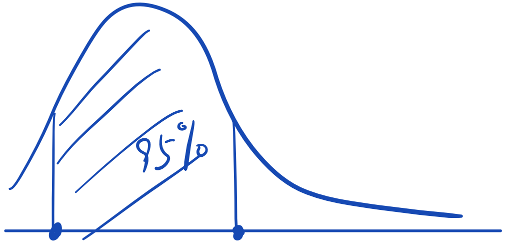
SE$(\hat{\theta})$ can be estimated by the bootstrap.

If $\hat{\theta}$ is far from normal, then we have to use the bootstrap to estimate the whole sampling distribution of $\hat{\theta}$, not just SE$(\hat{\theta})$.

The sampling distribution of $\hat{\theta}$ can be approximated by that of $\hat{\theta}^*$, which in turn can be approximated by the histogram of $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_B$.

This gives the **bootstrap percentile interval**

$$\left[\hat{\theta}^*_{(\alpha/2)}, \hat{\theta}^*_{(1-\alpha/2)}\right]$$



where $\hat{\theta}^*_{(\alpha/2)}$ is the $\alpha/2$ percentile of the $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_B$.

Bootstrap confidence intervals

An alternative to bootstrapping the distribution of $\hat{\theta}$ is to do so for $\hat{\theta} - \theta$.

The hope is that this approach is less sensitive to θ and therefore produces a more accurate confidence interval.

This results in the **bootstrap pivotal interval**

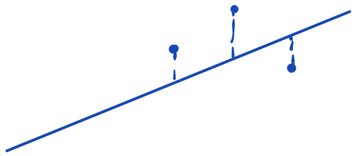
$$\left[2\hat{\theta} - \hat{\theta}^*_{(1-\alpha/2)}, 2\hat{\theta} - \hat{\theta}^*_{(\alpha/2)}\right]$$

Bootstrapping for regression

We have data $(X_1, Y_1), \ldots, (X_n, Y_n)$ from the simple linear regression model

$$Y_i \;=\; a + bX_i + e_i$$

From the data we can compute estimates $\hat{a}, \hat{b}$. How can we use the bootstrap to get standard errors and confidence intervals?

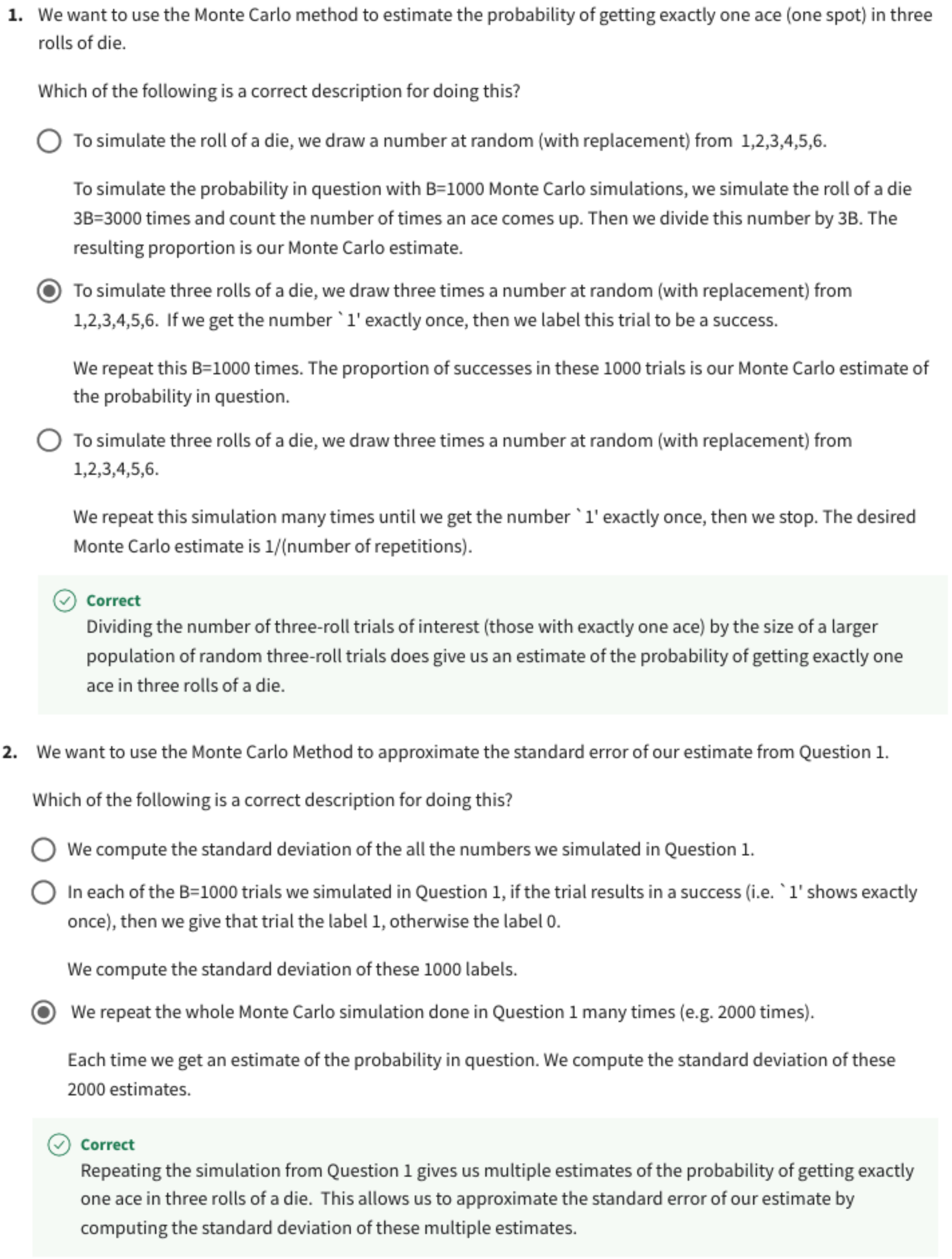


- Compute the residuals $\hat{e}_i = Y_i - \hat{a} - \hat{b}X_i$
- Resample from those residuals to get $e_1^*, \ldots, e_n^*$
- Compute the bootstrapped responses $Y_i^* = \hat{a} + \hat{b}X_i + e_i^*$

This gives a bootstrap sample $(X_1, Y_1^*), \ldots, (X_n, Y_n^*)$, from which we can estimate the parameters $\hat{a}^*$ and $\hat{b}^*$ in the usual way.

least squares

Quiz



1. We want to use the Monte Carlo method to estimate the probability of getting exactly one ace (one spot) in three rolls of die.

Which of the following is a correct description for doing this?

○ To simulate the roll of a die, we draw a number at random (with replacement) from 1,2,3,4,5,6.

To simulate the probability in question with B=1000 Monte Carlo simulations, we simulate the roll of a die 3B=3000 times and count the number of times an ace comes up. Then we divide this number by 3B. The resulting proportion is our Monte Carlo estimate.

● To simulate three rolls of a die, we draw three times a number at random (with replacement) from 1,2,3,4,5,6. If we get the number `1' exactly once, then we label this trial to be a success.

We repeat this B=1000 times. The proportion of successes in these 1000 trials is our Monte Carlo estimate of the probability in question.

○ To simulate three rolls of a die, we draw three times a number at random (with replacement) from 1,2,3,4,5,6.

We repeat this simulation many times until we get the number `1' exactly once, then we stop. The desired Monte Carlo estimate is 1/(number of repetitions).

✓ **Correct**
Dividing the number of three-roll trials of interest (those with exactly one ace) by the size of a larger population of random three-roll trials does give us an estimate of the probability of getting exactly one ace in three rolls of a die.

2. We want to use the Monte Carlo Method to approximate the standard error of our estimate from Question 1.

Which of the following is a correct description for doing this?

○ We compute the standard deviation of the all the numbers we simulated in Question 1.

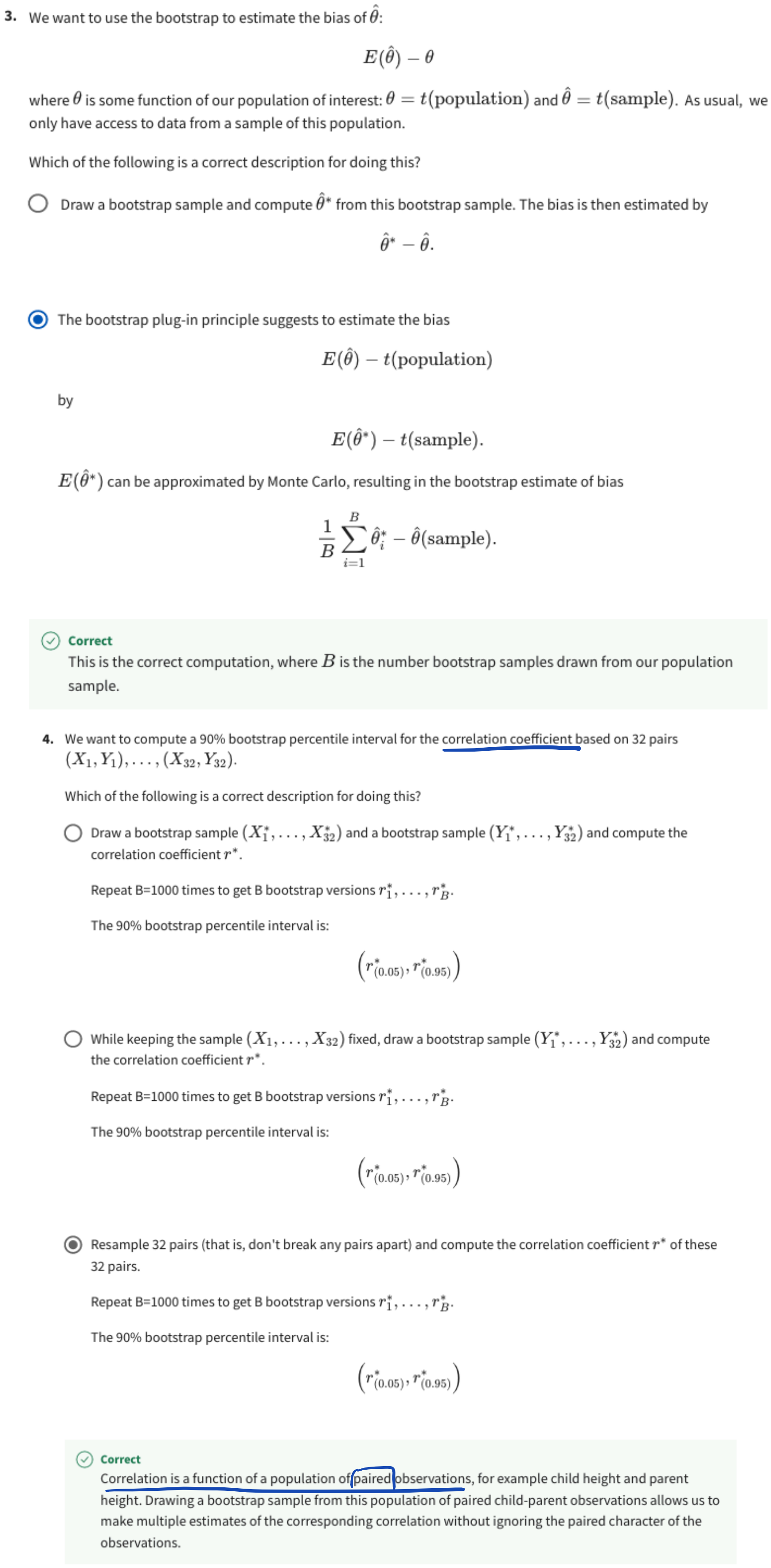
○ In each of the B=1000 trials we simulated in Question 1, if the trial results in a success (i.e. `1' shows exactly once), then we give that trial the label 1, otherwise the label 0.

We compute the standard deviation of these 1000 labels.

● We repeat the whole Monte Carlo simulation done in Question 1 many times (e.g. 2000 times).

Each time we get an estimate of the probability in question. We compute the standard deviation of these 2000 estimates.

✓ **Correct**
Repeating the simulation from Question 1 gives us multiple estimates of the probability of getting exactly one ace in three rolls of a die. This allows us to approximate the standard error of our estimate by computing the standard deviation of these multiple estimates.



3. We want to use the bootstrap to estimate the bias of $\hat{\theta}$:

$$E(\hat{\theta}) - \theta$$

where θ is some function of our population of interest: $\theta = t(\text{population})$ and $\hat{\theta} = t(\text{sample})$. As usual, we only have access to data from a sample of this population.

Which of the following is a correct description for doing this?

○ Draw a bootstrap sample and compute $\hat{\theta}^*$ from this bootstrap sample. The bias is then estimated by

$$\hat{\theta}^* - \hat{\theta}.$$

● The bootstrap plug-in principle suggests to estimate the bias

$$E(\hat{\theta}) - t(\text{population})$$

by

$$E(\hat{\theta}^*) - t(\text{sample}).$$

$E(\hat{\theta}^*)$ can be approximated by Monte Carlo, resulting in the bootstrap estimate of bias

$$\frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_i^* - \hat{\theta}(\text{sample}).$$

✓ **Correct**

This is the correct computation, where B is the number bootstrap samples drawn from our population sample.

4. We want to compute a 90% bootstrap percentile interval for the correlation coefficient based on 32 pairs $(X_1, Y_1), \dots, (X_{32}, Y_{32})$.

Which of the following is a correct description for doing this?

○ Draw a bootstrap sample $(X_1^*, \dots, X_{32}^*)$ and a bootstrap sample $(Y_1^*, \dots, Y_{32}^*)$ and compute the correlation coefficient r^*.

Repeat B=1000 times to get B bootstrap versions $r_1^*, \dots, r_B^*$.

The 90% bootstrap percentile interval is:

$$\left(r^*_{(0.05)}, r^*_{(0.95)}\right)$$

○ While keeping the sample $(X_1, \dots, X_{32})$ fixed, draw a bootstrap sample $(Y_1^*, \dots, Y_{32}^*)$ and compute the correlation coefficient r^*.

Repeat B=1000 times to get B bootstrap versions $r_1^*, \dots, r_B^*$.

The 90% bootstrap percentile interval is:

$$\left(r^*_{(0.05)}, r^*_{(0.95)}\right)$$

● Resample 32 pairs (that is, don't break any pairs apart) and compute the correlation coefficient r^* of these 32 pairs.

Repeat B=1000 times to get B bootstrap versions $r_1^*, \dots, r_B^*$.

The 90% bootstrap percentile interval is:

$$\left(r^*_{(0.05)}, r^*_{(0.95)}\right)$$

✓ **Correct**

Correlation is a function of a population of paired observations, for example child height and parent height. Drawing a bootstrap sample from this population of paired child-parent observations allows us to make multiple estimates of the corresponding correlation without ignoring the paired character of the observations.