STAT0032: Exercise Sheet #3

The exercises in this sheet focus on confidence intervals. As before (see Sheet #2), questions marked with the indication "(COMPUTER IMPLEMENTATION)" require programming, with R being the suggested language.

1. The following emphasises the message that, in a confidence interval, the randomness is at the boundaries of the interval, not in the parameter. Let X be a random variable whose distribution is parameterised by some θ. If $L(X)$ is a transformation of a random variable such that $P(L(X) \leq \theta) = 1 - \alpha_1$, and $U(X)$ is another transformation such that $P(U(X) \geq \theta) = 1 - \alpha_2$, and $L(x) \leq U(x)$ for all x, show that $P(L(X) \leq \theta \leq U(X)) = 1 - \alpha_1 - \alpha_2$.

2. Simon Newcomb performed experiments to measure the speed of light in 1878. Three days of measurements are provided in the Moodle page, file NEWCOMB.DAT. The unit of measurement is km/s. Using any method and software package of your choice to find confidence intervals for the speed of light. Perform some diagnostics of your assumptions.

3. (Adapted from Wasserman, Chapter 6) The following is a somewhat artificial example, but also helps to clarify the interpretation of confidence intervals. Say we want to estimate some property θ of the world which, as in Newcomb's study, is corrupted by measurement error. In this case, each data point is given as $Y^{(i)} = \theta + X^{(i)}$, where our data is $Y^{(i)}$, that is, $X^{(i)}$ is not observed. X is discrete, with probabilities $P(X = 1) = P(X = -1) = 1/2$.

   Now, suppose our sample size is $n = 2$, that is, our data is just some i.i.d. $\{Y^{(1)}, Y^{(2)}\}$. Let us define the following confidence interval:

   $$C = \begin{cases} Y^{(1)} - 1, & \text{if } Y^{(1)} = Y^{(2)}; \\ \{Y^{(1)} + Y^{(2)}\}/2, & \text{if } Y^{(1)} \neq Y^{(2)}. \end{cases}$$

   (a) What is the coverage of C?

   (b) Now suppose we observe the data $Y^{(1)} = 15$ and $Y^{(2)} = 17$. What is C? What can you say about θ? Do you see some problem with your confidence interval?

4. In class we saw one method for finding confidence intervals: finding **pivots** such as $Q(\mathbf{X}, \mu) \equiv (\hat{\mu}(\mathbf{X}) - \mu)/\hat{\sigma}(\mathbf{X})$, where we explicitly represented empirical average $\hat{\mu}$ and empirical standard deviation $\hat{\sigma}$ as functions of the data $\mathbf{X}$. A pivot is not a statistic as it depends on the unknown parameter, but it has a distribution which does not ($N(0, 1)$ in a typical example).

   Another way of building a confidence interval is by **inverting a test statistic**. For instance, suppose we have data $X^{(1)}, \dots, X^{(n)} \sim N(\mu, \sigma^2)$ from a known σ^2 but unknown μ. Consider some test for $H_0 : \mu = \mu_0$ at a level α. Describe how the machinery behind this test can be converted into a $1 - \alpha$ confidence interval for μ.

5. Now let us see how confidence intervals are complementary to hypothesis testing. Remember that the idea of confidence interval is quantifying the uncertainty we get given our assumptions, but this is not the same as validating our assumptions. Describe (or simulate with a program) which kind of confidence interval we will get for the problem of estimating the mean of a Normal of known variance when we assume the mean is positive, but when in reality it is not. To be more precise, suppose our sample size is $n = 10000$, our true data generating process is $N(-1, 1)$ and our estimator is $\hat{\mu} = \max(0, \bar{X})$, and the confidence interval is the usual one for the mean with known variance (even if this interval includes negative numbers, it will still asymptotically have the right coverage for $\mu > 0$). What are your conclusions?

6. (Computer implementation) Now let us see how the pipeline of "test, then infer" would work. Generate "random problems": sample parameters μ for a series of problems from a uniform distribution in $[-1, 1]$ (in R this would correspond to function runif). Then from each of these cases, sample 1000 points from a $N(\mu, 1)$ generating several synthetic datasets of sample size $n = 1000$ each. For each problem, do a hypothesis test $H_0 : \mu < 0$ against $H_1 : \mu \geq 0$. Only for those which pass the test at a level 0.05, calculate the 95% confidence interval as done in the previous question. What was the fraction of intervals which captured the true mean, and what do you conclude from this?

7. This is an example adapted from Efron and Tibshirani (1993, An Introduction to Bootstrap), also described by Wasserman. The data (available in Moodle, file placebo.dat) records the response of eight subjects to a placebo, an old treatment and a new treatment. In order to show that the efficacy of the new treatment is at least as good as the old, the regulatory agency Food and Drug Administration (FDA) requires a "proof of bioequivalence". Namely, for Z defined as the difference between the old treatment and the placebo, and for Y defined as the difference between the new treatment and the old, bioequivalence happens if

$$\theta \equiv \frac{E[Y]}{E[Z]} \leq 0.20.$$

An estimate of θ can be obtained by a **plug-in** strategy: get an estimate of $E[Y]$, an estimate of $E[Z]$ and take the ratio. That is, define

$$\hat{\theta} \equiv \frac{\bar{Y}}{\bar{Z}}.$$

Describe how you would use the bootstrap to find a confidence interval for θ, and why a simple CLT approximation is not straightforward.

(Computer implementation) Write a program to calculate the Normal-based and the bootstrap pivotal confidence interval for this parameter.

8. Another type of bootstrap interval is the **bootstrap percentile interval**, and it is as simple as it gets. If $\theta^\star_\alpha$ is the empirical α quantile of the bootstrap sample, an α interval is defined as

$$(\theta^\star_{\alpha/2}, \theta^\star_{1-\alpha/2}).$$

Provide a (formal or informal) justification for this interval, and with that comment on how you expect it to behave compared to the other two methods we discussed in the previous item.

(Computer implementation) Complement the program you wrote to also include the percentile interval.

9. (Adapted from Wasserman, Chapter 8) Let $X^{(1)}, \dots, X^{(n)} \sim Uniform(0, \theta)$. One estimator of θ (as a matter of fact, what we call a **maximum likelihood estimator**, as discussed in Chapter 3) is $\hat{\theta} \equiv \max\{X^{(1)}, \dots, X^{(n)}\}$. The density of such a distribution is

$$p(x) = \begin{cases} 1/\theta, & \text{for } 0 \leq x \leq \theta \\ 0, & \text{otherwise.} \end{cases}$$

As now we have a fully specified probability model (as opposed to the previous question), we might want to exploit the bootstrap in a different way. In **parametric bootstrap**, we generate bootstrap datasets from a model with parameters given by estimates generated from our original datasets. This is in contrast to what we saw in class, where we detailed what is also known as the **nonparametric bootstrap**, as it does not assume a particular shape for the distribution of the data.

For instance, we can plug-in the maximum likelihood estimator $\hat{\theta}$ to generate B bootstrap samples $X^{(1)\star}, \dots, X^{(n)\star} \sim \text{Uniform}(0, \hat{\theta})$. For each synthetic sample, we can now generate the corresponding $\theta^\star_{n,b}$.

(a) Find the distribution of $\hat{\theta}$ analytically.

(Computer implementation) Compare the true distribution of $\hat{\theta}$ to the histograms from the parametric and the nonparametric bootstrap. For instance, set $n = 50$ and $\theta = 1$. Plot histograms of corresponding bootstrap distributions of $\hat{\theta}$.

(b) The nonparametric bootstrap does poorly here. Show that $P(\theta^\star_{n,b} = \hat{\theta}) = 0$ for the parametric bootstrap, but $P(\theta^\star_{n,b} = \hat{\theta}) \approx 0.632$ in the nonparametric case. (For that, first you need to think what is the probability of a particular data point being included in a bootstrap sample. Then use a known result from calculus concerning e^x to find $\lim_{n\to\infty}(1 - (1/n))^n$.)

10. **The studentised pivotal interval**. We can take bootstrapping to the next level. Define the following pivot:

$$Z_n \equiv \frac{T_n - \theta}{\hat{se}_{boot}},$$

where T_n is a particular statistic computed from our data, θ is the parameter of interest, and $\hat{se}_{boot}$ is the standard deviation ("standard error") of T_n based on the bootstrap. Now define,

$$Z^\star_{n,b} \equiv \frac{T^\star_{n,b} - T_n}{\hat{se}^\star_b},$$

where $\hat{se}^\star_b$ is an estimate of the standard error of $T^\star_{n,b}$, **NOT** T_n. The empirical distribution formed from the bootstrap samples $Z^\star_{n,1}, \dots, Z^\star_{n,B}$ can be used to approximate the distribution of Z_n. So if $z^\star_\alpha$ is the α quantile of the empirical distribution of $Z^\star_{n,1}, \dots, Z^\star_{n,n}$, then $P(Z_n \leq z^\star_\alpha) \approx \alpha$. Moreover,

$$C_n = \left(T_n - z^\star_{1-\alpha/2}\hat{se}_{boot}, T_n - z^\star_{\alpha/2}\hat{se}_{boot}\right)$$

will be an (asymptotic) confidence interval for θ with approximate coverage α. With actual data, this type of interval is in general a better approximation to an exact confidence interval than the usual pivotal interval.

a. How would you compute C_n? What is the price you pay for that?

b. (COMPUTER IMPLEMENTATION) Implement this method, apply it to our previous drug experiment test.