STATG006: Solutions to Exercise Sheet #3

The exercises in this sheet focus on confidence intervals. As before, we provide solutions, sometimes detailed, sometimes a sketch that should point you to the complete solution.

1. There are many possibilities out there, but studies of public opinion are among the most common. I expect the forums to contain some concrete examples.

2. One way is to exploit the following about probabilities of events: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. In this case, we define our events A and B as "$L(X) \leq \theta$" and "$U(X) \geq \theta$". The only step now that requires a bit thought is to show that $P(A \cup B) = 1$ in this case.

3. An interval based on means plus or minus three standard deviations will have an approximate coverage of 99%. QQ plots should reveal at least a couple of measurements that would not fit a Gaussian model. Regardless of Gaussianity, a comparison between days should also reveal something about possible outliers. Should we remove them? Although in this exercise we are not given a complete picture of the measurement process, we may be able to provide some justification for that.

4. For (a), check that the probability of the event $Y^{(1)} = Y^{(2)}$ is $1/2$, and therefore $P(Y^{(1)} \neq Y^{(2)}) = 1/2$. Calculate the probability of $P(C \text{ contains } \theta \mid Y^{(1)} = Y^{(2)})$ and $P(C \text{ contains } \theta \mid Y^{(1)} \neq Y^{(2)})$. Then use the Law of Total Probability to get $P(C \text{ contains } \theta) = 3/4$.

   For (b), it should be clear that $\theta = 16$. Also, by definition $C = \{16\}$ (a single point). Yet, our claim is that this is a 0.75 confidence interval. *Our claim is correct.* As we discussed in class, the probabilistic statement is that *over the data distribution* we capture the true parameter with that probability. It says nothing about any particular interval, although through other external means (like technological advances that remove measurement error, or the special arrangement of the problem here) we may be able to validate whether a particular interval does capture the intended target.

5. One typical rejection region for the test would be for samples $\mathbf{x}$ such that $|\bar{x} - \mu_0| > z_{1-\alpha/2}\sigma/\sqrt{n}$, where $z_{1-\alpha/2}$ is the corresponding quantile of a standard Normal. The probability statement concerning $\bar{X}$ is:

$$P\left(\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu_0 \leq \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \mid \mu = \mu_0\right) = 1 - \alpha$$

But the above is true for *every* μ_0. Therefore, the following is true:

$$P\left(\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

It follows that the interval $[\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}]$ will give us a confidence interval with coverage $1 - \alpha$.

6. After doing the usual calculations (e.g., using the interval from the previous item with our different choice of $\hat{\mu}$) this interval will shrink to zero and with very high probability be centered at zero. That means an interval which is very narrow and yet has coverage of zero! This is not a failure of the interval construction. It is shown exactly what it should show: which values of the parameters are compatible with the data, across the distribution of the data given by the model assumptions. Narrow confidence intervals are not the same thing as high precision answers. Everything is predicted on the model assumptions, which might not be true. This example might have looked very artificial and extreme, but in reality we are getting intervals not for "true" parameters but for approximations based on a approximate model. The hope is that the approximations are good enough, but this leaves room for hypothesis testing to at least assess some of the assumptions to some extent.

7. The actual coverage goes back to 0.95 after all. However, in general we might not be that lucky, in the sense that here the assumptions tested and the confidence interval calculated have a direct relationship – i.e. a single hypothesis and a single confidence interval. With multiple hypothesis tests, there are more chances of mistakes if proper control of Type I error is not done, which will propagate to the coverage of the proposed intervals. In that case, the coverage will not be as the advertised, although we should expect it to be closer to the right coverage by first testing our assumptions.

8. It is not obvious from here that this ratio of averages would converge to a Gaussian distribution. Moreover, our sample size is small anyway. The application of bootstrapping is as in the slides. See EX3.R.

9. On a qualitative, informal basis, we can think of explanations for the validity of this interval. In a semi-formal way, the following is adapted from section 8.5.2. from Wasserman. The main idea is to transform the distribution of $\theta^\star$ into something manageable. If $\theta^\star_{n,b}$ is the b-th bootstrap estimate of θ, let $U^\star_b \equiv m(\theta^\star_{n,b})$. This transformation is defined[1] such that it is monotone, and $U = m(\hat{\theta})$ is such that $U \sim N(\phi, c^2)$, where $\phi \equiv m(\theta)$.

   Let $u^\star_\beta$ be the β sample quantile of the bootstrap $U^\star_b$s. Monotonicity will preserve quantiles, so that $u^\star_{\alpha/2} = m(\theta^\star_{\alpha/2})$. By the fact that $U \sim (\phi, c^2)$, the $\alpha/2$ quantile of

[1] In reality, a transformation with exactly these properties will rarely exist, but there may exist approximate ones. That's a shortcoming of this method.

this distribution is $\phi + z_{\alpha/2}c$, where $z_{\alpha/2}$ is the corresponding quantile of a standard Gaussian. Therefore, $u^\star_{\alpha/2} = \phi + z_{\alpha/2}c$ and $u^\star_{1-\alpha/2} = \phi + z_{(1-\alpha)/2}c$. This implies,

$$\begin{aligned} P(\theta^\star_{\alpha/2} \leq \theta \leq \theta^\star_{1-\alpha/2}) &= P(m(\theta^\star_{\alpha/2}) \leq m(\theta) \leq m(\theta^\star_{\alpha/2})) \\ &= P(u^\star_{\alpha/2} \leq \phi \leq u^\star_{1-\alpha/2}) \\ &= P(U + cz_{\alpha/2} \leq \phi \leq U + cz_{1-\alpha/2}) \\ &= P(z_{\alpha/2} \leq \tfrac{U-\phi}{c} \leq z_{1-\alpha/2}) \\ &= 1 - \alpha. \end{aligned}$$

10. For part(a), the reasoning is similar to what we did in the previous exercise sheet for the minimum of a sample. The maximum is relatively simpler and I will not provide the full derivation here, but just start from the same idea: what is the equivalent event to "the maximum of the sample is less than or equal to some x"? That is, if $X_{\max} \equiv \max(\{X^{(1)}, \dots, X^{(n)}\})$, how would you write the event $X_{\max} \leq x$ in a way to exploit the known distribution of $\{X^{(1)}, \dots, X^{(n)}\}$? See EX3.R for hints concerning part (b). The main calculus result you need to use for showing the requested statement is $\lim_{n\to\infty}(1 + x/n)^n = e^x$. The relation between that and $P(\theta^\star_{n,b} = \hat{\theta})$ is the following: the event equivalent to "$\theta^\star_{n,b} = \hat{\theta}$" is "the maximum of the sample is in the bootstrap sample". Even simpler, as the data is i.i.d., this is equivalent to "a particular data point $X^{(i)}$ is in the bootstrap sample". Now you just have to show that the probability of that event is $1 - (1 - (1/n))^n$.

11. For (a), the problem is that we need $\hat{se}^\star_b$. Surely if we need the bootstrap to approximate the standard error of T_n, then there is no easy way of exactly calculating the standard error of $T^\star_{n,b}$. With no shame, we can *bootstrap the bootstrap*: with each bootstrap sample b, we use bootstrap internally to estimate $\hat{se}^\star_b$ (that is a lot of nested estimation taking place here), which we can plug in the formula for $Z^\star_{n,b}$. The price we pay is two nested bootstraps. So if a particular bootstrap sample requires B iterations, two nested bootstraps require B^2. If we need a large B (the choice of B is another story we have not discussed, but just imagine "setting it large enough so that multiple runs of the bootstrap seem to agree"[2]), then this can be costly. In are in your own for (b), but feel free to ask about it.

[2]Yes, we can statistically test that too!