5A Confidence intervals

Class problems

5A1 (Soda fountain) A soda fountain dispenses soft drinks in cups. The volume of a drink (in milliliters) is approximately normally distributed with mean μ and a known standard deviation of $\sigma = 3$, caused by random variations in the dispensing process. In nine cups the following volumes were observed: 304, 298, 301, 302, 301, 300, 305, 300, 306.

(a) Find a confidence interval for μ at 95% confidence level.

(b) Find a confidence interval for μ at 99% confidence level.

(c) When performing an experiment like in (a) — dispensing nine cups and then calculating a confidence interval — what is the probability that the calculated interval will (i) contain μ, (ii) be completely below μ, (iii) be completely above μ?

(d) How many cups do we need to take in order to obtain a 95% confidence interval that has length at most 1 milliliter (0.5 ml on each side of point estimate)?

(e) How many cups for a 95% confidence interval of length 0.1 milliliters?

Solution.

(a) The sample mean is $m(\vec{x}) \approx 301.89$. Because σ is known, and the individual volumes were assumed to be normal, our confidence interval is simply

$$m(\vec{x}) \pm z\frac{\sigma}{\sqrt{n}},$$

where $n = 9$, $\sigma = 3$ and $z > 0$ is a number such that a standard normal distributed Z has $\mathrm{P}(|Z| \leq z) = 0.95$. Equivalently, we require $\mathrm{P}(|Z| > z) = 0.05$. Because the standard normal distribution is symmetric around zero, the tail probabilities $\mathrm{P}(Z < -z)$ and $\mathrm{P}(Z > z)$ are equal, so one tail must be $\mathrm{P}(Z > z) = 0.05/2 = 0.025$.

Thus we need to find z such that the cumulative distribution function has value $F_Z(z) = \Phi(z) = 1 - 0.025 = 0.975$. From tables, or by the R command `z <- qnorm(0.975)` we obtain $z \approx 1.96$. (Note that there are other ways of performing the same calculation.) Our confidence interval is

$$m(\vec{x}) \pm z\frac{\sigma}{\sqrt{n}} \;=\; 301.89 \pm 1.96\frac{3}{\sqrt{9}} \;=\; 301.89 \pm 1.96.$$

(b) Same as before, but now we require $z > 0$ to be such that each tail has probability $\mathrm{P}(Z < -z) = \mathrm{P}(Z > z) = 0.01/2 = 0.005$. Using tables or the R command `z <- qnorm(0.995)` we obtain $z \approx 2.58$. The confidence interval at 99% level is

$$m(\vec{x}) \pm z\frac{\sigma}{\sqrt{n}} \;=\; 301.89 \pm 2.58\frac{3}{\sqrt{9}} \;=\; 301.89 \pm 2.58.$$

(c) (i) 95% (ii) 2.5% (iii) 2.5%

(d) If a 95% confidence interval is calculated from n observations, using our assumptions (normally distributed observations with known $\sigma = 3$), and the same procedure as in (a), but with some other value of n, we will have a confidence interval of length

$$2z\frac{\sigma}{\sqrt{n}} = \frac{2 \cdot 1.96 \cdot 3}{\sqrt{n}}.$$

Setting this equal to 1 (milliliter), and solving for n, we obtain

$$n = (2 \cdot 1.96 \cdot 3)^2 \approx 138.3.$$

Since we want the interval to be 0.01 or shorter, we need to pour 139 cups.

(e) Approximately $(2 \cdot 1.96 \cdot 3/0.1)^2 \approx 13830$ cups.

5A2 (Opinion poll) An opinion poll reported in July 2016 by Helsingin Sanomat, 89% of Finns think that president Niinistö has performed his duties well or extremely well. The poll was conducted by recording the opinions of 1002 Finns of ages 15–79 years, and the margin of error was reported as approximately 3 percentage points (in both directions). Let us assume that the margin of error was calculated by using the conservative interval for the binary model (see e.g. Lecture 4B).

(a) From the reported numbers, deduce what confidence level was used.

(b) How many Finns should have been recorded, in order to obtain a margin of error of 1 percentage point (in both directions), at the same confidence level?

Solution.

(a) The conservative confidence interval for the binary model is

$$\hat{p} \pm z \cdot \frac{0.5}{\sqrt{n}},$$

where $\hat{p}$ is the relative frequency of ones in the data set, and the number z is such that

$$\mathrm{P}(|Z| \le z) = \beta,$$

where β is the confidence level. The margin of error (= half-length of confidence interval) is

$$z \cdot \frac{0.5}{\sqrt{n}} = z \cdot \frac{0.5}{\sqrt{1002}}.$$

Since the margin of error was reported as 0.03, we can solve

$$z \;=\; 0.03 \cdot \frac{\sqrt{1002}}{0.5} \;\approx\; 1.90.$$

This corresponds to the confidence level

$$\beta = \mathrm{P}(|Z| \le 1.90) \;=\; \Phi(1.90) - \Phi(-1.90) \;\approx\; 94.3\%.$$

The calculation says that the confidence level was approximately 94.3%. We may guess that they actually used the common 95% level.

(If they actually used 95% and thus $z = 1.96$, and the procedure we assumed, then their margin of error was $1.96 \cdot 0.5/\sqrt{1002} \;\approx\; 3.10\%$, which may have been reported as "approximately 3%".)

(b) To have margin of error 0.01 at 95% confidence level, using the conservative binary interval, we need

$$0.01 \;=\; 1.96 \cdot \frac{0.5}{\sqrt{n}},$$

so

$$n \;=\; \left(1.96 \cdot \frac{0.5}{0.01}\right)^2 \;=\; 9604.$$

We need to record the opinions of about 9600 Finns.

Home problems

5A3 (Multiparty opinion poll) From a large population, a random sample of $n = 100$ persons were asked which of the four parties A,B,C,D they support. The numbers of the supporters were 70, 28, 2 and 0.

(a) For each party separately, consider the *binary* question (supporting party X or not), and calculate a 95% confidence interval for the proportion of party-X supporters in the population. Give the results as intervals like $[0.500, 0.600]$, with both endpoints expressed in 3 decimals (or 1 decimal in percent form).

(b) Repeat the previous, now using the conservative length of confidence intervals (see e.g. Lecture 4B).

(c) Do the calculated intervals seem reasonable and meaningful? If not, explain in what way, and what you think might be the reason. Note. Here a small amount of thinking is enough. We do not require a complete solution. Solutions for such situations exist in the literature, but they are somewhat more complicated.

(d) If the true proportion of some party X in the population is nonzero, can the sample proportion be zero? Apply common sense, not calculations.

(e) If the true proportion of some party X in the population is zero, can the sample proportion be nonzero? Apply common sense, not calculations.

Grading.
1/2 points for (a),(b) each.
1/3 points for (c),(d),(e) each.
Total rounded up. In (d) it is enough to have just some consideration. For example, observing that the confidence intervals span negative values.

Solution.

(a) For party A, from the sample we have the point estimate $\hat{p} = 70/100 = 0.700$, and the approximate margin of error (at 95% confidence level) is $1.96 \cdot \sqrt{\hat{p}(1-\hat{p})/n} \approx 0.090$. From this we have the confidence interval

$$[0.700 - 0.090, 0.700 + 0.090] = [0.610, 0.790].$$

For party B, we obtain MOE=0.088 and interval $[0.192, 0.368]$.

For party C, we obtain MOE=0.027 and interval $[-0.007, 0.047]$.

For party D, we obtain MOE=0.000 and interval $[0.000, 0.000]$.

(b) The point estimates $\hat{p}$ are as before, but now for all parties we use the same "conservative" margin of error $1.96 \cdot 0.5/\sqrt{n} = 0.098$. So the conservative confidence intervals are:

- A: $[0.602, 0.798]$

- B: $[0.182, 0.378]$
- C: $[-0.078, 0.118]$
- D: $[-0.098, 0.098]$

In all four cases, the interval is wider than in (a), as it should because we are using the conservative (maximum) interval length.

(c) The confidence intervals of parties A and B seem reasonable; they are wide, but this is because of the smallish sample size ($n = 100$).

The intervals for parties C and D are somewhat strange. The true proportion of party supporters (or anything) in the population cannot be negative, so it seems odd that the intervals span some negative values. The main reason for this is that we used the normal distribution to approximate the binomial distribution; the normal distribution can easily go to the negative side. Near the ends of a binomial distribution, the normal approximation is not very good.

One thing is easy to do. We could simply cut away any negative portion of the confidence intervals. This cannot really harm, because the true proportion is surely nonnegative.

But there are other problems. In (a), we calculated the confidence interval for party D as a single point 0.000. This seems intuitively too narrow (and indeed it is). On the other hand, in (b) we calculated a very wide interval for party D.

In general, the confidence intervals for a binary proportion are quite bad near the ends (mainly because of the normal approximation). In the literature, you can find more accurate formulas for such cases, for example in Brown, Cai ja DasGupta (2001): Interval estimation for a binomial proportion, *Statistical Science* 16:101–133. This is outside the course, but it is good to know such formulas exist, in case you need them. — Another, and perhaps easier solution is provided by Bayesian credible intervals.

(d) Yes. This can easily happen if the proportion of those supporters is small in the population and/or the random sample is small. (So from seeing zero in the sample, we cannot really infer that the true proportion is zero.)

(e) No. (If we saw even one supporter for party C in the sample, we know with absolute certainty that there is at least that one supporter in the population, so the population proportion is not zero.)

5A4 In a physical experiment, particle decays are observed at random intervals. The intervals are assumed independent and each of them exponentially distributed with unknown mean μ (so the rate parameter of the distribution is $\lambda = 1/\mu$). 30 consecutive intervals were observed, and their mean was 12.09 seconds and standard deviation 11.47 seconds.

(a) Calculate a confidence interval for μ at 90% confidence level. Use the general method for estimating an unknown mean (Lecture 4B).

(b) The individual intervals were not normally distributed. Do you think the general method is still reasonable? Why / why not? Explain in words.

(c) Would the same method work well if we had $n = 3$ observations? Why / why not?

Grading. 1 p for (a), 0.5 p for each of (b) and (c).

Solution.

(a) For 90% confidence level, we need a number $z > 0$ such that $\Phi(z) = 0.95$, so that the right tail has probability $P(Z > z) = 1 - \Phi(z) = 0.05$, and (by symmetry) the left tail has the same probability $P(Z < -z) = 0.05$. (Then the two tails have 0.10 probability in total.)

So we find from tables, or by the R command `qnorm(0.95)` that $z \approx 1.64$, and calculate the confidence interval as

$$m(\vec{x}) \pm z \cdot \frac{\operatorname{sd} \vec{x}}{\sqrt{n}} \;=\; 12.09 \pm 1.64 \times \frac{11.47}{\sqrt{30}} \;\approx\; 12.09 \pm 3.43.$$

(b) Although the individual intervals were exponentially distributed (which is far from normal), the sum of 30 independent such intervals is reasonably close to normally distributed, so we may approximate that $m(\vec{x})$ has normal distribution.

Because the sample is not very big ($n = 30$), and the standard deviation was estimated from the sample, it might be a good idea to use the t distribution instead of the normal. (In the t distribution with 29 degrees of freedom, the 0.95 point is at 1.70 instead of 1.64, so the difference is not that big.) — If we want to get more involved, we could try to apply the knowledge that the individual obsevations are exponential, and look for special, more efficient methods for this situation.

(c) Probably not. The sum of three exponentially distributed random variables is not very close to normal distribution, and also $n = 3$ is so small that we cannot assume that $\operatorname{sd}(\vec{x})$ is very close to the standard deviation in of the generating (exponential) distribution. We would need some more powerful methods to work with this small non-normal data.

In such a situation, one could search the literature for confidence intervals specially tailored to the *exponential* distribution, since we were assuming that is the shape of the generating distribution. — Another solution would be to apply Bayesian inference, which can easily be applied even with small data. The calculated interval may be wide, but at least it will be valid, and from small data one cannot really assume very precise knowledge of the parameters.