Practice questions

1. A medical journal publishes analyses about the efficacy θ of treatments that can be proven to exceed some threshold at the 95% confidence level.

   (a) If a journal issue publishes 30 analyses what is the probability that at least one of them does not meet the threshold? Assume that each analysis barely meets the 95% confidence level requirement (it does not exceed it) and that the analyses are mutually independent.

   Solution: The probability of all 30 analyses meeting the threshold is $\mathrm{P}(ALL) = 0.95^{30}$, so the probability that at least one analysis one does not meet the threshold equals to $\mathrm{P}(ALL^c) = 1 - \mathrm{P}(ALL) = 1 - 0.95^{30} \approx 0.785$.

   (b) If the journal editors want to be 95% confident that all 30 analyses are correct, how should they modify the requirement for each treatment?

   Solution: If p is the confidence level of a single analysis then $\mathrm{P}(ALL) = p^{30}$. For the journal to achieve $\mathrm{P}(ALL) = 95\%$ it needs to impose a confidence level of $p = \sqrt[30]{95\%} \approx 99.8\%$.

2. The measurements of ten random athlete heights in centimeters are

$$152, 163, 188, 201, 192, 176, 194, 166, 215, 184.$$

   (a) Assuming the heights are independent normal random variables with known standard deviation $\sigma = 20$, give a 95% confidence interval for the mean height.

   Solution: The (symmetric) 95% confidence interval for the mean height of normal samples with known standard deviation σ is $[\overline{X} - z\frac{\sigma}{\sqrt{n}}, \overline{X} + z\frac{\sigma}{\sqrt{n}}]$, where z is chosen so that $\mathrm{P}(-z \le \mathrm{Normal}(0,1) \le z) = 95\%$. The latter condition holds for $z = 1.96$. We calculate

$$\overline{X} = \frac{152 + 163 + 188 + 201 + 192 + 176 + 194 + 166 + 215 + 184}{10} = 183.1$$

   to obtain the confidence interval $[170.7, 195.5]$.

   (b) How many samples do you need for a 95% confidence interval of width 5cm?

   Solution: The width of the confidence interval is $\frac{2\sigma}{\sqrt{n}} \cdot z$, so to ensure width at most w, n should be $(2\sigma z/w)^2$. When $w = 5$ and $z = 1.96$, $(2\sigma z/w)^2$, $(2\sigma z/w)^2 = 245.8624$, so $n = 246$ samples are sufficient.

3. A large company conducts a job satisfaction survey among its 6250 employees. Out of 250 employees that are sampled (with repetition), 142 are satisfied with their jobs.

   (a) Calculate a 99% confidence interval for the number of employees that are satisfied with their job.

   Solution: We first come up with a confidence interval for the fraction of satistfied employees p given $n = 250$ Indicator(p) (also known as Bernoulli(p)) independent samples. The sample mean is

$$\overline{X} = \frac{142}{250} = 0.568.$$

The "traditional" formula based on the normal approximation of a sum of indicator samples gives the 99% confidence interval $[A - zB, A + zB]$ for p, where $z = 2.576$ is the 99 percentile two-sided threshold for the Normal(0, 1) random variable, and

$$A = \frac{\overline{X} + z^2/2n}{1 + z^2/n} \approx 0.567, \qquad B = \frac{\sqrt{\overline{X}(1-\overline{X})/n + z^2/4n^2}}{1 + z^2/n} \approx 0.031$$

The 99%-confidence interval for p is $[A - zB, A + zB] \approx [0.487, 0.646]$. The 99%-confidence interval for the number $6250p$ of satisfied employees is $[3041, 4036]$.
The "simplified" formula gives the 99%-confidence interval

$$\left[\overline{X} - z\sqrt{\overline{X}(1-\overline{X})/n}, \overline{X} + z\sqrt{\overline{X}(1-\overline{X})/n}\right] \approx [0.487, 0.648]$$

for p and the corresponding interval $[3045, 4054]$ for $6250p$.

(b) Find a confidence interval of width 100 for the number of satisfied employees and estimate the confidence level for it.

Solution: If we use the "simplified" formula, the width is $w = N \cdot 2z\sqrt{\overline{X}(1-\overline{X})/n}$, where $N = 6250$ is the total number of employees. If we set $w = 100$, for $\overline{X} = 0.568$ and $n = 250$ we get

$$z = \frac{w\sqrt{n}}{2N\sqrt{\overline{X}(1-\overline{X})}} \approx 0.255$$

which gives a confidence level of $\mathrm{P}(-z \le \mathrm{Normal}(0,1) \le z) \approx 0.201$, or only about 20% for the interval $[N\overline{X} - 50, N\overline{X} + 50] = [3540, 3640]$.

4. Let MAX be the sample maximum for 10 independent samples of a Uniform$(0, \theta)$ random variable.

(a) What is the CDF of the random variable MAX/θ?

Solution:

$$\begin{aligned}\mathrm{P}(MAX/\theta \le z) &= \mathrm{P}(MAX \le \theta z) \\ &= \mathrm{P}((X_1 \le \theta z) \cap \cdots \cap (X_{10} \le \theta z)) \\ &= \mathrm{P}(X_1 \le \theta z) \cdots \mathrm{P}(X_{10} \le \theta z) \\ &= z^{10}\end{aligned}$$

for $0 \le z \le 1$.

(b) Find values z_- and z_+ for which $\mathrm{P}(MAX/\theta < z_-) = \mathrm{P}(MAX/\theta > z_+) = 2.5\%$.

Solution: From part (a), $z_- = \sqrt[10]{2.5\%} \approx 0.692$ and $z_+ = \sqrt[10]{97.5\%} \approx 0.997$.

(c) Find a 95% confidence interval for θ based on your answers in part (b).

Solution: We can rewrite the event $z_- \le MAX/\theta \le z_+$ as $MAX/z_+ \le \theta \le MAX/z_-$, so by part (b) the desired 95%-confidence interval is

$$[MAX/z_+, MAX/z_-] \approx [1.003MAX, 1.446MAX].$$

(d) **(Optional)** There are multiple 95% confidence intervals for θ. What is the narrowest one that you can find?

Solution: We want to minimize the length $(1/z_- - 1/z_+)MAX$ of the confidence interval under the constraint $\mathrm{P}(MAX < z_-) + \mathrm{P}(MAX > z_+) = 5\%$, namely $z_-^{10} + (1 - z_+^{10}) = 5\%$ for $0 \leq z_- \leq z_+ \leq 1$. If we make the substitution $z_- = (z_+^{10} - 0.95)^{1/10}$ and plot the function $f(z_+) = 1/(z_+^{10} - 0.95)^{1/10} - 1/z_+$ we can observe that it is decreasing on $[0.95^{1/10}, 1]$ so the optimal choice is $z_+ = 1$ and $z_- = 0.05^{1/10} \approx 0.741$. (You should be able to prove this using calculus.) The narrowest 95% interval is therefore $[MAX, 1.349 MAX]$.