CS 70 Discrete Mathematics and Probability Theory
Summer 2021 DIS 07B

1 Confidence Interval Introduction

We observe a random variable X which has mean μ and standard deviation $\sigma \in (0,\infty)$. Assume that the mean μ is unknown, but σ is known.

We would like to give a 95% confidence interval for the unknown mean μ. In other words, we want to give a random interval (a,b) (it is random because it depends on the random observation X) such that the probability that μ lies in (a,b) is at least 95%.

We will use a confidence interval of the form $(X-\varepsilon, X+\varepsilon)$, where $\varepsilon > 0$ is the width of the confidence interval. When ε is smaller, it means that the confidence interval is narrower, i.e., we are giving a more *precise* estimate of μ.

(a) Using Chebyshev's Inequality, calculate an upper bound on $\mathbb{P}[|X-\mu| \geq \varepsilon]$.

(b) Explain why $\mathbb{P}(|X-\mu| < \varepsilon)$ is the same as $\mathbb{P}[\mu \in (X-\varepsilon, X+\varepsilon)]$.

(c) Using the previous two parts, choose the width of the confidence interval ε to be large enough so that $\mathbb{P}[\mu \in (X-\varepsilon, X+\varepsilon)]$ is guaranteed to exceed 95%. [Note: Your confidence interval is allowed to depend on X, which is observed, and σ, which is known. Your confidence interval is not allowed to depend on μ, which is unknown.]

(d) The previous three parts dealt with the case when you observe one sample X. Now, let n be a positive integer and let $X_1, \ldots, X_n$ be i.i.d. samples, each with mean μ and standard deviation $\sigma \in (0,\infty)$. As before, assume that μ is unknown but σ is known.

Here, a good estimator for μ is the *sample mean* $\bar{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$. Calculate the mean and variance of $\bar{X}$.

(e) We will now use a confidence interval of the form $(\bar{X}-\varepsilon, \bar{X}+\varepsilon)$ where $\varepsilon > 0$ again represents the width of the confidence interval. Imitate the steps of (a) through (c) to choose the width ε to be large enough so that $\mathbb{P}[\mu \in (\bar{X}-\varepsilon, \bar{X}+\varepsilon)]$ is guaranteed to exceed 95%.

To check your answer, your confidence interval should be *smaller* when n is larger. Intuitively, if you collect more samples, then you should be able to give a more *precise* estimate of μ.

Solution:

(a) Since $\mathbb{E}[X] = \mu$ and $\operatorname{Var} X = \sigma^2$, then by Chebyshev's Inequality,

$$\mathbb{P}\{|X-\mu| \geq \varepsilon\} \leq \frac{\operatorname{Var} X}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2}.$$

(b) Note that $|X-\mu|<\varepsilon$ if and only if $-\varepsilon<X-\mu<\varepsilon$, if and only if $\mu-\varepsilon<X<\mu+\varepsilon$. However, the first inequality says that $\mu<X+\varepsilon$ and the second inequality says that $\mu>X-\varepsilon$, that is, $X-\varepsilon<\mu<X+\varepsilon$, which is the same thing as saying $\mu\in(X-\varepsilon,X+\varepsilon)$. So, the events $\{|X-\mu|<\varepsilon\}$ and $\{\mu\in(X-\varepsilon,X+\varepsilon)\}$ are identical.

(c) We want $\mathbb{P}[\mu\in(X-\varepsilon,X+\varepsilon)]\geq 0.95$, which is equivalent to

$$\mathbb{P}[|X-\mu|\geq\varepsilon]=1-\mathbb{P}[|X-\mu|<\varepsilon]=1-\mathbb{P}[\mu\in(X-\varepsilon,X+\varepsilon)]\leq 0.05.$$

However, we have the bound $\mathbb{P}[|X-\mu|\geq\varepsilon]\leq\sigma^2/\varepsilon^2$, so we just need to choose ε big enough so that $\sigma^2/\varepsilon^2\leq 0.05$. To do this, we want $\varepsilon^2\geq 20\sigma^2$, or $\varepsilon\geq\sqrt{20}\sigma\approx 4.47\sigma$. Our confidence interval is therefore $(X-4.47\sigma,X+4.47\sigma)$.

(d) For the mean, use linearity of expectation. We have

$$\mathbb{E}[\bar{X}]=\mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n}X_i\Big]=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i]=\frac{1}{n}\sum_{i=1}^{n}\mu=\frac{1}{n}\cdot n\mu=\mu.$$

For the variance, recall two facts. One is that for a constant c, the scaling of the variance is $\operatorname{Var}(cX)=c^2\operatorname{Var}X$. The second fact is that $X_1,\ldots,X_n$ are independent, so they are pairwise uncorrelated, that is, for any distinct $i,j\in\{1,\ldots,n\}$, $\operatorname{cov}(X_i,X_j)=0$; this implies that $\operatorname{Var}(\sum_{i=1}^{n}X_i)=\sum_{i=1}^{n}\operatorname{Var}X_i$. Using these facts,

$$\operatorname{Var}\bar{X}=\operatorname{Var}\Big(\frac{1}{n}\sum_{i=1}^{n}X_i\Big)=\frac{1}{n^2}\operatorname{Var}\Big(\sum_{i=1}^{n}X_i\Big)=\frac{1}{n^2}\sum_{i=1}^{n}\operatorname{Var}X_i=\frac{1}{n^2}\sum_{i=1}^{n}\sigma^2=\frac{1}{n^2}\cdot n\sigma^2=\frac{\sigma^2}{n}.$$

(e) By Chebyshev's Inequality,

$$\mathbb{P}\{|\bar{X}-\mu|\geq\varepsilon\}\leq\frac{\operatorname{Var}\bar{X}}{\varepsilon^2}=\frac{\sigma^2}{n\varepsilon^2}.$$

We want $\sigma^2/(n\varepsilon^2)\leq 0.05$, and to do this, we choose $\varepsilon^2\geq 20\sigma^2/n$, or $\varepsilon\geq\sqrt{20}\sigma/\sqrt{n}$.

2 Poisson Confidence Interval

For n a positive integer, you collect $X_1, \ldots, X_n$ i.i.d. samples drawn from a Poisson distribution (with unknown mean λ). However, you have a bound on the mean: from a confidential source, you know that $\lambda \leq 2$. For $0 < \delta < 1$, find a $1-\delta$ confidence interval for λ using Chebyshev's Inequality.

Solution: Our estimator for λ is the sample mean $\frac{1}{n}\sum_{i=1}^{n} X_i$. We apply Chebyshev's Inequality for $\varepsilon > 0$:

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \lambda\right| > \varepsilon\right) \leq \frac{\operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)}{\varepsilon^2} = \frac{\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right)}{n^2\varepsilon^2} = \frac{\sum_{i=1}^{n}\operatorname{Var} X_i}{n^2\varepsilon^2} = \frac{\operatorname{Var} X_1}{n\varepsilon^2} = \frac{\lambda}{n\varepsilon^2}$$
$$\leq \frac{2}{n\varepsilon^2}.$$

We want the probability of error to be at most δ, so we set

$$\frac{2}{n\varepsilon^2} \leq \delta \implies \varepsilon \geq \sqrt{\frac{2}{n\delta}}.$$

Our $1-\delta$ confidence interval for λ is

$$\left[\frac{1}{n}\sum_{i=1}^{n} X_i - \sqrt{2/(n\delta)}, \frac{1}{n}\sum_{i=1}^{n} X_i + \sqrt{2/(n\delta)}\right].$$

3 Vegas

On the planet Vegas, everyone carries a coin. Many people are honest and carry a fair coin (heads on one side and tails on the other), but a fraction p of them cheat and carry a trick coin with heads on both sides. You want to estimate p with the following experiment: you pick a random sample of n people and ask each one to flip their coin. Assume that each person is independently likely to carry a fair or a trick coin.

(a) Let X be the proportion of people whose coin flip results in heads. Find $\mathbb{E}[X]$.

(b) Given the results of your experiment, how should you estimate p? (*Hint:* Construct an unbiased estimator for p using part (a))

(c) How many people do you need to ask to be 95% sure that your answer is off by at most 0.05?

(d) Suppose n is large. Construct an approximate 98% confidence interval for p.

Solution:

(a) Let X_i be the indicator that the ith person's coin flips heads. Then $X = \frac{1}{n}\sum_{i=1}^{n} X_i$. Apply linearity.

$$\mathbb{E}[X] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \mathbb{E}[X_i]$$

By total probability,

$$\mathbb{E}[X_i] = p\cdot 1 + (1-p)\cdot\frac{1}{2} = \frac{1}{2}(p+1).$$

(b) We want to construct an estimate $\hat{p}$ such that $\mathbb{E}[\hat{p}] = p$. Then, if we have a large enough sample, we'd expect to get a good estimate of p. In other words, we measure X, the fraction of people whose coin flips heads. How can we use this observation to construct $\hat{p}$? From (a), $\mathbb{E}[X] = \frac{1}{2}(p+1)$. By applying (reverse) linearity to isolate p, we find that

$$p = 2\,\mathbb{E}[X] - 1 = \mathbb{E}[2X-1].$$

Thus, our estimator $\hat{p}$ should be $2X-1$.

(c) We want to find n such that $P[|\hat{p}-p| \le 0.05] > 0.95$. Another way to state this is that we want

$$P[|\hat{p}-p| > 0.05] \le 0.05.$$

Notice that $\mathbb{E}[\hat{p}] = p$ by construction, so we can immediately apply Chebyshev's inequality on $\hat{p}$. What we get is:

$$\mathbb{P}[|\hat{p}-p| > 0.05] \le \mathbb{P}[\hat{p}-p| \ge 0.05] \le \frac{\mathrm{Var}[\hat{p}]}{0.05^2}$$

If $\frac{\text{Var}[\hat{p}]}{0.05^2} \leq 0.05$, then we have $\mathbb{P}[|\hat{p}-p| > 0.05] \leq 0.05$ as desired. So, we want n such that $\text{Var}[\hat{p}] \leq 0.05^3$.

$$\text{Var}[\hat{p}] = \text{Var}[2X-1] = 4\,\text{Var}[X] = \frac{4}{n^2}\,\text{Var}\left[\sum_{i=1}^{n} X_i\right] = \frac{4}{n}\,\text{Var}[X_1].$$

But X_i is an indicator (Bernoulli variable), so its variance is bounded by $\frac{1}{4}$ (note that $p(1-p)$ is maximized at $p=\frac{1}{2}$ to yield a value of $\frac{1}{4}$). Therefore we have

$$\text{Var}[\hat{p}] \leq \frac{4}{n}\frac{1}{4} = \frac{1}{n}.$$

So, we choose n such that $\frac{1}{n} \leq 0.05^3$, so $n \geq \frac{1}{0.05^3} = 8000$.

(d) If n is large, the distribution of $S_n = \sum_{i=1}^{n} X_i$ is approximately normal. We must select ε such that

$$\mathbb{P}[p \in (\hat{p}-\varepsilon, \hat{p}+\varepsilon)] = \mathbb{P}[|\hat{p}-p| < \varepsilon] = 0.98.$$

Since $\hat{p}-p = \frac{1}{n}S_n - p$, $\hat{p}-p$ is also approximately normal. Let $\sigma^2 = \text{Var}[X_1] = p(1-p)$. Then $\hat{p}-p \xrightarrow{d} \text{Normal}\left(0, \frac{4\sigma^2}{n}\right)$. For n large, $\hat{p}-p \approx \frac{2\sigma}{\sqrt{n}}Z$ where $Z \sim \text{Normal}(0,1)$.

$$\begin{aligned}
\mathbb{P}[|\hat{p}-p| < \varepsilon] &= \mathbb{P}\left[-\varepsilon < \hat{p}-p < \varepsilon\right] \\
&\approx \mathbb{P}\left[-\varepsilon < \frac{2\sigma}{\sqrt{n}}Z < \varepsilon\right] \\
&= \mathbb{P}\left[-\frac{\varepsilon\sqrt{n}}{2\sigma} < Z < \frac{\varepsilon\sqrt{n}}{2\sigma}\right] \\
&= \Phi\left(\frac{\varepsilon\sqrt{n}}{2\sigma}\right) - \Phi\left(-\frac{\varepsilon\sqrt{n}}{2\sigma}\right) \\
&= 2\Phi\left(\frac{\varepsilon\sqrt{n}}{2\sigma}\right) - 1 \qquad \text{by symmetry of } Z
\end{aligned}$$

Since we want $\mathbb{P}[|\hat{p}-p| < \varepsilon] = 0.98$,

$$\begin{aligned}
2\Phi\left(\frac{\varepsilon\sqrt{n}}{2\sigma}\right) - 1 &= 0.98 \\
\frac{\varepsilon\sqrt{n}}{2\sigma} &= \Phi^{-1}\left(\frac{0.98+1}{2}\right) \\
\varepsilon &= \frac{2\sigma}{\sqrt{n}}\Phi^{-1}(0.99) \\
\varepsilon &\leq \frac{1}{\sqrt{n}}\Phi^{-1}(0.99),
\end{aligned}$$

where in the last line we use the fact that $\sigma^2 \leq \frac{1}{4}$. Our confidence interval is

$$\left[\hat{p} - \frac{1}{\sqrt{n}}\Phi^{-1}(0.99), \hat{p} + \frac{1}{\sqrt{n}}\Phi^{-1}(0.99)\right].$$