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 $\mathbb{P}[|X - \mu| \geq \varepsilon]$.

$$P(|X-\mu| \geq \varepsilon) \leq \frac{Var(X)}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2}$$

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

$|X-\mu| < \varepsilon \iff -\varepsilon < X-\mu < \varepsilon \iff \mu - \varepsilon < X < \mu + \varepsilon$

First inequality says $\mu < X+\varepsilon$, second says $\mu > X-\varepsilon$

so $\mu \in (X-\varepsilon, X+\varepsilon)$

(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.]

want to choose ε s.t. $P(\mu \in (X-\varepsilon, X+\varepsilon)) \geq 0.95$

i.e. $P(|X-\mu| < \varepsilon) \geq 0.95$

$\iff P(|X-\mu| \geq \varepsilon) \leq 0.05$

$P(|X-\mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}$ by Chebychev's inequality

choose ε big enough s.t. $\frac{\sigma^2}{\varepsilon^2} \leq 0.05$

$\varepsilon^2 \geq 20\sigma^2$

$\varepsilon \geq \sqrt{20}\,\sigma \approx 4.47\sigma$

our confidence interval is $(X-4.47\sigma,\ X+4.47\sigma)$

(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 $\bar{X}$.

$E[\bar{X}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$

$Var(\bar{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}$

(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.

want to find ε s.t. $P(\mu \in (\bar{X} - \varepsilon, \bar{X} + \varepsilon)) \geq 0.95$

i.e. $P(|\bar{X} - \mu| \geq \varepsilon) \leq 0.05$

$P(|\bar{X} - \mu| \geq \varepsilon) \leq \dfrac{Var(\bar{X})}{\varepsilon^2} = \dfrac{\sigma^2}{n\varepsilon^2}$ by Chebychev's Inequality

want $\dfrac{\sigma^2}{n\varepsilon^2} \leq 0.05$

$$\varepsilon^2 \geq \frac{20\sigma^2}{n}$$

$$\varepsilon \geq \frac{\sqrt{20}\sigma}{\sqrt{n}}$$

$$\left(\bar{X} - \frac{\sqrt{20}\sigma}{\sqrt{n}}, \bar{X} + \frac{\sqrt{20}\sigma}{\sqrt{n}}\right)$$

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.

$1-\delta$ confidence interval means we want probability of error (prob. thing we're trying to estimate falls outside our confidence interval) to be at most δ

---

estimator for λ is $\frac{1}{n}\sum_{i=1}^{n} X_i$

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

$$= \frac{Var\left(\sum_{i=1}^{n} X_i\right)}{n^2\varepsilon^2}$$

$$= \frac{\sum_{i=1}^{n} Var(X_i)}{n^2\varepsilon^2}$$

$$= \frac{\cancel{n}\, Var(X_i)}{n^{\cancel{2}}\varepsilon^2}$$

$$= \frac{\lambda}{n\varepsilon^2}$$

$$\leq \frac{2}{n\varepsilon^2}$$

set $\frac{2}{n\varepsilon^2} \leq \delta$ (since we want error $\leq \delta$)

$$\varepsilon \geq \sqrt{\frac{2}{n\delta}}$$

$1-\delta$ C.I. is $\left[\frac{1}{n}\sum_{i=1}^{n} X_i - \sqrt{\frac{2}{n\delta}}, \ \frac{1}{n}\sum_{i=1}^{n} X_i + \sqrt{\frac{2}{n\delta}}\right]$

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 $\mathbb{E}[X]$.

let $X_i = \begin{cases} 1 & \text{if ith person's coin flips heads} \\ 0 & \text{o.w.} \end{cases}$

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

$= \frac{1}{n}(n\,E[X_i])$

$= E[X_i]$

$E[X_i] = P(X_i = 1)$

$= P(\text{ith person heads} \mid \text{ith person fair})\,P(\text{ith person fair})$

$+ P(\text{ith person heads} \mid \text{ith person trick})\,P(\text{ith person trick})$

$= \frac{1}{2}(1-p) + 1 \cdot p$

$= \frac{1}{2}(p+1)$

$E[X] = \frac{1}{2}(p+1)$

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

want to construct an estimate $\hat{p}$ s.t. $E[\hat{p}] = p$

since $E[X] = \frac{1}{2}(p+1)$

$$p = 2E[X] - 1$$
$$= E[2X-1]$$

let $\hat{p} = 2X - 1$

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

find n s.t. $P(|\hat{p} - p| \leq 0.05) \geq 0.95$

i.e. $P(|\hat{p} - p| > 0.05) \leq 0.05$

since $E[\hat{p}] = p$

$$P(|\hat{p} - p| > 0.05) \leq P(|\hat{p} - p| \geq 0.05) \leq \frac{Var(\hat{p})}{0.05^2}$$

want $\frac{Var(\hat{p})}{0.05^2} \leq 0.05 \Rightarrow$ want n s.t.

$Var(\hat{p}) \le 0.05^3$ $\quad Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i) = nVar(X_1)$

$$Var(\hat{p}) = Var(2\bar{X}-1) = 4Var(\bar{X}) = \frac{4}{n^2} Var(\sum_{i=1}^{n} X_i)$$

$$= \frac{4}{n} Var(X_1)$$

$Var(X_1) = p(1-p)$ which is maximized at $p = \frac{1}{2}$

giving a variance of $\frac{1}{4}$

$\left(\frac{d}{dp} p(1-p) = 0 \iff 1-2p = 0 \iff p = \frac{1}{2} \right)$

so $Var(X_1) \le \frac{1}{4} \Rightarrow Var(\hat{p}) \le \frac{4}{n}\left(\frac{1}{4}\right) = \frac{1}{n}$

choose n s.t. $\frac{1}{n} \le 0.05^3$

$$\boxed{n \ge \frac{1}{0.05^3} = 8000}$$

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

when n large, $\frac{1}{n}S_n \to N(\mu, \frac{\sigma^2}{n})$ by CLT

where $S_n = \sum_{i=1}^{n} X_i$, $\mu = E[X_i] = \frac{1}{2}(p+1)$, $\sigma^2 = Var(X_i)$
$= p(1-p)$

need to select ε s.t. $P(p \in (\hat{p}-\varepsilon, \hat{p}+\varepsilon)) \geq 0.98$

i.e. $P(|\hat{p}-p| < \varepsilon) \geq 0.98$

Note that $\hat{p} = 2(\frac{1}{n}S_n) - 1$

$\hat{p} - p = \frac{2}{n}S_n - 1 - p$ so $\hat{p} - p \approx N(0, \frac{4\sigma^2}{n})$

$\to \hat{p} - p \approx \frac{2\sigma}{\sqrt{n}} Z$ for $Z \sim N(0,1)$

*recall that $X \sim N(\mu, \sigma^2) \to aX + b \sim N(a\mu + b, a^2\sigma^2)$

$$P(|\hat{p}-p| < \varepsilon) = P(-\varepsilon < \hat{p} - p < \varepsilon)$$
$$= P\left(-\frac{\varepsilon\sqrt{n}}{2\sigma} < Z < \frac{\varepsilon\sqrt{n}}{2\sigma}\right)$$
$$= \Phi\left(\frac{\varepsilon\sqrt{n}}{2\sigma}\right) - \Phi\left(-\frac{\varepsilon\sqrt{n}}{2\sigma}\right)$$

$$= 2\Phi\left(\frac{\varepsilon\sqrt{n}}{2\sigma}\right) - 1$$

Set $2\Phi\left(\frac{\varepsilon\sqrt{n}}{2\sigma}\right) - 1 = 0.98$

$$\frac{\varepsilon\sqrt{n}}{2\sigma} = \Phi^{-1}\left(\frac{1.98}{2}\right)$$

$$\varepsilon = \frac{2\sigma}{\sqrt{n}}\Phi^{-1}(0.99)$$

$$\varepsilon \leq \frac{1}{\sqrt{n}}\Phi^{-1}(0.99) \quad \text{since } \sigma^2 \leq \frac{1}{4}$$

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