Discussion 7B

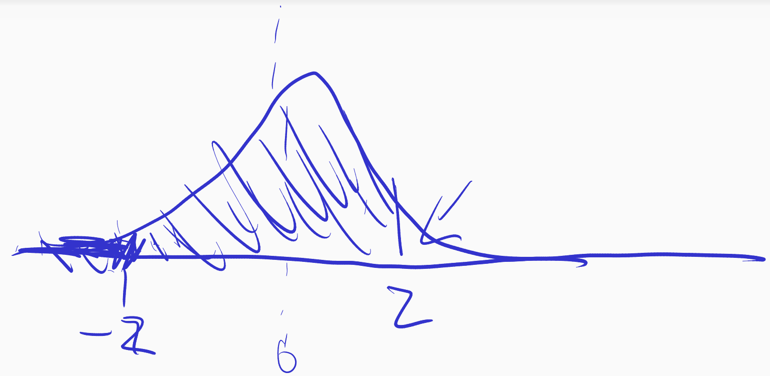
Tarang Srivastava - CS70 Summer 2020

Mini Review

$\Phi(\cdot)$ is the cdf of the standard normal random variable.

leave your answers

$$\phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx$$



$$\phi(z) - (1 - \phi(z))$$

$$2\phi(z) - 1$$

Provide a confidence level that the true parameter μ is with a certain range of the estimated parameter:

$$\boxed{P(|\hat{\mu} - \mu| \leq \epsilon) \geq 1 - \delta}$$

We can think of ϵ as the error in our estimate, and $1 - \delta$ as our confidence level.

Sample mean $\bar{X}$

"error"

$$\underbrace{P(|X - \mu| \leq \varepsilon)}_{\text{probability that } X \text{ is within } \mu \text{ by } \varepsilon} \geq 1 - \delta$$

0.95

0.05

95%

Stats

Lecture Review

Some More Review

Question 1

Question 1

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.]

Question 1

(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 $\mathbb{P}\{\mu \in (X-\varepsilon, X+\varepsilon)\}$.

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) 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.

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

b) $P(|X-\mu| < \varepsilon) = P(\mu \in (X-\varepsilon, X+\varepsilon))$

$|X-\mu| < \varepsilon$

$X-\mu < \varepsilon$ $\qquad$ $-(X-\mu) < \varepsilon$

$X-\varepsilon < \mu$ $\qquad$ $X-\mu > -\varepsilon$

$\qquad$ $X+\varepsilon > \mu$

$X-\varepsilon < \mu < X+\varepsilon \Longrightarrow \mu \in (X-\varepsilon, X+\varepsilon)$

Question 1

(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.]

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.

(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 $(X - 4.47\sigma, X + 4.47\sigma)$.

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

$\boxed{P(\mu \in \underbrace{(X-\varepsilon, X+\varepsilon)}_{\text{part (b)}}) \geq \underset{1-\delta}{95\%}}$

$(X - 4.47\sigma, X + 4.47\sigma)$

$P(|X-\mu| < \varepsilon) \geq 95\%$

$+1 \qquad +1$

complement

Solve for epsilon $1 - 0.95 \geq 1 - P(|X-\mu| < \varepsilon) \geq$

$\underset{\delta}{0.05} \geq \frac{\sigma^2}{\varepsilon^2} > P(|X-\mu| \geq \varepsilon)$

Question 2

Question 2

$X \sim \text{Poisson}(\lambda)$

$E[X] = \lambda$ $\quad \boxed{Var(X) = \lambda}$

You collect n samples (n is a positive integer) $X_1, \dots, X_n$, which are i.i.d and known to be drawn from a Poisson distribution (with unknown mean). However, you have a bound on the mean: from a confidential source, you know that $\lambda \leq 2$. Find a $1-\delta$ confidence interval ($\delta \in (0,1)$) for λ using Chebyshev's Inequality. (Hint: a good estimator for λ is the *sample mean* $\bar{X} := n^{-1}\sum_{i=1}^{n} X_i$)

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

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

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

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

$$\boxed{P(|\hat{\mu} - \mu| \geq \varepsilon) \leq \delta} \quad 0.05$$

$$-\lambda| \geq \varepsilon) < \frac{Var(\bar{X})}{\varepsilon^2} \qquad Var(X+Y) = Var \quad Var(Y) \text{ independ}$$

$$P(|\bar{X} - \lambda| \geq \varepsilon) \leq \frac{Var(\bar{X})}{\varepsilon^2} = \frac{Var(\frac{1}{n}\sum_{i=1}^{n} x_i)}{\varepsilon^2} = \frac{Var(\sum_{i=1}^{n} X_i)}{n^2\varepsilon^2}$$

$Var(X+Y) = Var(X) + Var(Y)$ independent

$$= \frac{\sum_{i=1}^{n} Var(X_i)}{n^2\varepsilon^3}$$

$$\frac{\lambda}{n\varepsilon^2} < \delta \qquad \mu \in (X - \varepsilon, X + \varepsilon)$$

$$\frac{\lambda}{n\varepsilon^2} < \boxed{\frac{2}{n\varepsilon^2} < \delta}$$

solve for epsilon

it won't

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

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

$$\left(\bar{X} - \sqrt{\frac{2}{n\delta}}, \bar{X} + \sqrt{\frac{2}{n\delta}}\right)$$

Quesiton 2

Question 3

Question 3

We would like to test the hypothesis claiming that a coin is fair, i.e. $P(H) = P(T) = 0.5$. To do this, we flip the coin $n = 100$ times. Let Y be the number of heads in $n = 100$ flips of the coin. We decide to reject the hypothesis if we observe that the number of heads is less than $50 - c$ or larger than $50 + c$. However, we would like to avoid rejecting the hypothesis if it is true; we want to keep the probability of doing so less than 0.05. Please determine c. (*Hints: use the central limit theorem to estimate the probability of rejecting the hypothesis given it is actually true. Table is provided in the appendix.*)

X_i represent i th flip, Bernoulli

$E[X_i] = \frac{1}{2}$

CLT $Y =$

$Y = \sum_{i=1}^{100} X_i$

$Var(X_i) = \frac{1}{4}$

$E[X_i^2] - E[X]^2$

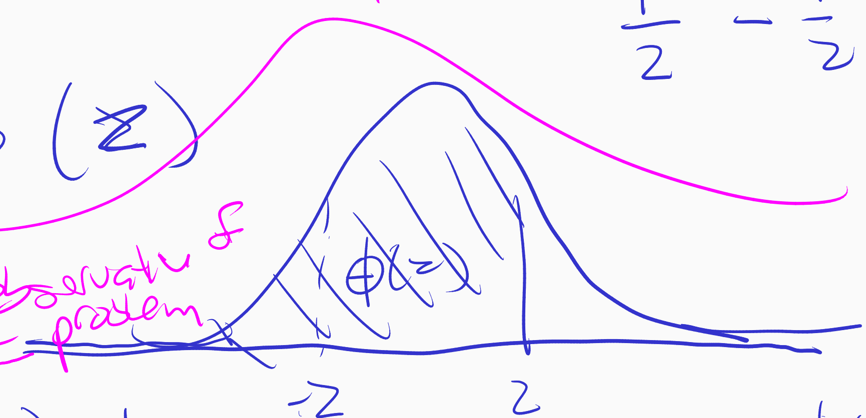
$E[X_i]$

$\frac{1}{2} - \frac{1}{2}^2 = \frac{1}{4}$

$$P\left(\frac{Y - n\frac{1}{2}}{\sqrt{n\frac{1}{4}}} \le z\right) \approx \Phi(z)$$

CLT

$$P\left(\frac{Y - 50}{\cancel{\frac{10}{2}}\ 5} \le z\right) \approx \Phi(z)$$



observation of the problem

$$P(|Y - 50| < c)$$

$$P\left(\frac{|Y-50|}{5} < \frac{c}{5}\right) \approx 2\Phi\left(\frac{c}{5}\right) - 1$$

$$\Phi(z) - (1 - \Phi(z)) = 2\Phi(z) - 1$$

$$2\Phi\left(\frac{c}{5}\right) - 1 = 0.95$$

$$\Phi\left(\frac{c}{5}\right) = 0.975$$

Let X_i be the random variable denoting the result of the i-th flip:

$$X_i = \begin{cases} 1 & \text{if the } i\text{-th flip is head,} \\ 0 & \text{if the } i\text{-th flip is tail.} \end{cases}$$

Then we have $Y = \sum_{i=1}^{n} X_i$. If the hypothesis is true, then $\mu = \mathbb{E}[X_i] = \frac{1}{2}$ and $\sigma^2 = \mathrm{Var}(X_i) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$. By central limit theorem, we know that

$$P\left(\frac{Y - n\mu}{\sqrt{n\sigma^2}} \le z\right) \approx \Phi(z)$$

$$P\left(\frac{Y - 100 \cdot \frac{1}{2}}{\sqrt{100 \cdot \frac{1}{4}}} \le z\right) \approx \Phi(z)$$

$$P\left(\frac{Y - 50}{5} \le z\right) \approx \Phi(z)$$

where

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \mathrm{d}x.$$

We will reject the hypothesis when $|Y - 50| > c$. We also want $P(|Y - 50| > c) < 0.05$, or equivalently $P(|Y - 50| \le c) > 0.95$. We have

$$P(|Y - 50| \le c) = P\left(\frac{|Y - 50|}{5} \le \frac{c}{5}\right) \approx 2\Phi\left(\frac{c}{5}\right) - 1.$$

The reason this is $\approx 2\Phi(\frac{c}{5}) - 1$ is because the probability we are looking for is the probability that Y is within $\frac{c}{5}$ standard deviations of the mean. By an area argument, we can see that this is $\Phi(\frac{c}{5}) - (1 - \Phi(\frac{c}{5})) = 2\Phi(\frac{c}{5}) - 1$. Let $2\Phi(\frac{c}{5}) - 1 = 0.95$, so $\Phi(\frac{c}{5}) = 0.975$ or $\frac{c}{5} = 1.96$. That is $c = 9.8$ flips. So we see that if we observe more that $50 + 10 = 60$ or less than $50 - 10 = 40$ heads, we can reject the hypothesis.

We would like to test the hypothesis claiming that a coin is fair, i.e. $P(H) = P(T) = 0.5$. To do this, we flip the coin $n = 100$ times. Let Y be the number of heads in $n = 100$ flips of the coin. We decide to reject the hypothesis if we observe that the number of heads is less than $50 - c$ or larger than $50 + c$. However, we would like to avoid rejecting the hypothesis if it is true; we want to keep the probability of doing so less than 0.05. Please determine c. (*Hints: use the central limit theorem to estimate the probability of rejecting the hypothesis given it is actually true. Table is provided in the appendix.*)

Question 3