Chapter 8 Sampling Distributions of Estimators

Yurun (Ellen Ying)

Oct 2, 2022

8.1 The Sampling Distribution of a Statistic

Problem 1

Suppose that a random sample $X_1, \ldots, X_n$ is to be taken from the uniform distribution on the interval $[0, \theta]$ and that θ is unknown. How large a random sample must be taken in order that

$$\Pr(|\max\{X_1, \ldots, X_n\} - \theta| \leq 0.1\theta) \geq 0.95,$$

for all possible θ?

Answer: Let random variable $T = \max\{X_1, \ldots, X_n\}$. We can write its c.d.f. as

$$\Pr(T \leq t) = (\frac{t}{\theta})^n,$$

where $t \in [0, \theta]$. This is because if the maximum among n observations is less than or equal to t, all off them are less than or equal to t. The fact that all of them are independent of each other gives the above expression.
From the given condition, we have

$$\Pr(|T - \theta| \leq 0.1\theta) = \Pr(0.9\theta \leq T \leq 1.1\theta) = 1 - \Pr(T \leq 0.9\theta).$$

The last equality holds because $T \leq 1.1\theta$ with a probability 1. So that we have

$$1 - \Pr(T \leq 0.9\theta) = 1 - 0.9^n \geq 0.95.$$

Solving for n yields $n \geq 29$.

Problem 3

Suppose that a random sample is to be taken from the normal distribution with unknown mean θ and standard deviation 2.
a. How large a random sample must be taken in order that $\mathrm{E}_\theta(|\overline{X}_n - \theta|^2) \leq 0.1$ for every possible value θ?
b. How large a random sample must be taken in order that $\mathrm{E}_\theta(|\overline{X}_n - \theta|) \leq 0.1$ for every possible value of θ?

Answer.a: Since X_i are taken from a normal distribution independently, we can have

$$\mathrm{E}[\overline{X}_n] = \theta \text{ and } \mathrm{Var}[\overline{X}_n] = \mathrm{E}[\overline{X}_n^2] - \mathrm{E}[\overline{X}_n]^2 = \frac{\sigma^2}{n}.$$

Therefore, we have

$$\begin{aligned}
\mathrm{E}_\theta(|\overline{X}_n - \theta|^2) &= \mathrm{E}[\overline{X}_n^2] - 2\theta\mathrm{E}[\overline{X}_n] + \theta^2 \\
&= \frac{\sigma^2}{n} + \theta^2 - 2\theta^2 + \theta^2 \\
&= \frac{4}{n} \leq 0.1
\end{aligned}$$

Solve for n and we have $n \geq 40$.

Answer.b: Let $Y = \overline{X}_n - \theta$. Since $\overline{X}_n$ is a normal distribution with mean θ and variance $\frac{\sigma^2}{n}$, Y is a normal distribution with mean 0 and the same variance. We can calculate the mean of $|Y|$ as follows:

$$\begin{aligned}
\mathrm{E}[|Y|] &= \int_{-\infty}^{-\infty} |y| f(y) dy \\
&= \int_{-\infty}^{0} |y| f(y) dy + \int_{0}^{-\infty} |y| f(y) dy \\
&= -\int_{-\infty}^{0} y f(y) dy + \int_{0}^{-\infty} y f(y) dy) \\
&= -\frac{1}{\sqrt{2\pi\sigma^2/n}} \left(\int_{-\infty}^{0} y \exp[-\frac{y^2}{2\sigma^2/n}] dy + \int_{0}^{-\infty} y \exp[-\frac{y^2}{2\sigma^2/n}] dy) \right) \\
&= \frac{1}{\sqrt{2\pi\sigma^2/n}} \left(\frac{\sigma^2}{n} \exp[-\frac{y^2}{2\sigma^2/n}] \Big|_{-\infty}^{0} - \frac{\sigma^2}{n} \exp[-\frac{y^2}{2\sigma^2/n}] \Big|_{0}^{-\infty} \right) \\
&= \sqrt{\frac{2\sigma^2}{n\pi}}.
\end{aligned}$$

Therefore, we have

$$\mathrm{E}_\theta(|\overline{X}_n - \theta|) = \sqrt{\frac{8}{n\pi}} \leq 0.1.$$

Solve for n and we have $n \geq \frac{800}{\pi} \approx 255$.

Problem 5

Suppose that a random sample is to be taken from the Bernoulli distribution with unknown parameter p. Suppose also that it is believed that the value of p is in the neighborhood of 0.2. How large a random sample must be taken in order that $\Pr(|\overline{X}_n - p| \leq 0.1) \geq 0.75$ when $p = 0.2$?

Answer: We can express the probability we need to calculate as follows

$$\Pr(-0.1 + p \leq \overline{X}_n \leq 0.1 + p) = \Pr(0.1 \leq \overline{X}_n \leq 0.3) = \Pr(\overline{X}_n \leq 0.3) - \Pr(\overline{X}_n \leq 0.1).$$

This calls for us to calculate the c.d.f. of the sampling distribution of $\overline{X}_n$. Since $\overline{X}_n = \frac{Y}{n}$ where $Y = \sum_i^n X_i \sim \mathrm{Bern}(n, p)$, we have

$$\begin{aligned}
F_{\overline{X}_n}(x) &= \Pr(\overline{X}_n \leq x) \\
&= \Pr(Y \leq nx) \\
&= \sum_{0}^{nx} \binom{n}{nx} p^{nx} (1-p)^{n-nx}.
\end{aligned}$$

Substitute back to the original express yields a equation that needs to be solved

$$F_{\overline{X}_n}(0.3) - F_{\overline{X}_n}(0.1) \geq 0.75.$$

Using R to solve the equation yields $n \geq 17$.
Appendix: R code for reference

```r
# write a function to calculate the c.d.f. of X_n
Pr <- function(n, x, p = 0.2) {
  nx <- seq(0, n*x, by = 1)
  Pr <- sum(gamma(n+1)/(gamma(nx+1)*gamma(n-nx+1))*p^(nx)*(1-p)^(n - nx))
  return(Pr)
}

# iterate through different values of n
N <- seq(10, 100, by = 1) # generate a sequence of n
P <- c() # placeholder for the desired probability
for (n in N) {
  p <- Pr(n = n, x = 0.3, p = 0.2) - Pr(n = n, x = 0.1, p = 0.2) # calculate the probability
  P <- c(P, p) # append the new value to the vector
}

min(N[which(P >= 0.75)])
```

8.2 The Chi-Squared Distributions

Problem 1

Suppose that we will sample 20 chunks of cheese in Example 8.2.3. Let $T = \sum_{i=1}^{20}(X_i - \mu)^2/20$, where X_i is the concentration of lactic acid in the ith chunck. Assume that $\sigma^2 = 0.09$. What number c satisfies $\Pr(T \le c) = 0.9$?

Answer: Let $Z_i = \frac{X_i}{\sigma}$, and we can write T as

$$T = \sum_{i=1}^{20}(X_i - \mu)^2/20 = \frac{\sigma^2}{20}\sum_{i=1}^{20} Z_i^2.$$

Therefore, $W = \frac{20T}{\sigma^2}$ is a Chi-squared distribution. Now we have

$$\Pr(T \le c) = \Pr(W \le \frac{20c}{\sigma^2}) = 0.9.$$

Thus, $\frac{20c}{\sigma^2} = 28.41198$. Solving for c yields $c \approx 0.1279$

Problem 5

Suppose that a point (X, Y, Z) is to be chosen at random in three-dimensional space, where X, Y, and Z are independent random variables and each has the standard normal distribution. What is the probability that the distance from the origin to the point will be less than 1 unit?

Answer: The distance can be expressed as $d = \sqrt{X^2 + Y^2 + Z^2}$, and thus the square $d^2 = X^2 + Y^2 + Z^2$ is a Chi-squared distribution with 3 degrees of freedom. The probability of d being less than 1 unit is

$$\Pr(d < 1) = \Pr(d^2 < 1) \approx 0.1987$$

8.4 The t Distributions

Problem 1

Suppose that X has the t distribution with m degrees of freedom $(m > 2)$. Show that $\mathrm{Var}(X) = m/(m-2)$.

Answer: Since $\mathrm{E}[X] = 0$, $\mathrm{Var}[X] = \mathrm{E}[X^2]$. Using the p.d.f. of a t distribution, we have

$$\mathrm{E}[X^2] = \frac{\Gamma(\frac{m+1}{2})}{(m\pi)^{1/2}\Gamma(\frac{m}{2})} \int_{-\infty}^{+\infty} x^2(1+\frac{x^2}{m})^{-\frac{m+1}{2}} dx.$$

Since the integrand is an even function, the above equation can be written into

$$\mathrm{E}[X^2] = 2\frac{\Gamma(\frac{m+1}{2})}{(m\pi)^{1/2}\Gamma(\frac{m}{2})} \int_{0}^{+\infty} x^2(1+\frac{x^2}{m})^{-\frac{m+1}{2}} dx.$$

Change the variable x to

$$y = \frac{\frac{x^2}{m}}{1+\frac{x^2}{m}},$$

we have

$$\begin{aligned}
\mathrm{E}[X^2] &= 2\frac{\Gamma(\frac{m+1}{2})}{(m\pi)^{1/2}\Gamma(\frac{m}{2})} \int_{0}^{+\infty} \frac{my}{1-y}(\frac{1}{1-y})^{-\frac{m+1}{2}} \frac{1}{2}(\frac{my}{1-y})^{-\frac{1}{2}} \frac{m}{(1-y)^2} dy \\
&= 2\frac{\Gamma(\frac{m+1}{2})}{(m\pi)^{1/2}\Gamma(\frac{m}{2})} \frac{m^{\frac{3}{2}}}{2} \int_{0}^{+\infty} y^{\frac{1}{2}}(1-y)^{\frac{m-4}{2}} dy \\
&= 2\frac{\Gamma(\frac{m+1}{2})}{(m\pi)^{1/2}\Gamma(\frac{m}{2})} \frac{m^{\frac{3}{2}}}{2} \frac{\Gamma(\frac{3}{2})\Gamma(\frac{m-2}{2})}{\Gamma(\frac{m+1}{2})} \\
&= \frac{m}{m-2}.
\end{aligned}$$

The second to last step holds based on the p.d.f. of a beta function with parameters $\alpha = \frac{3}{2}$ and $\beta = \frac{m-2}{2}$.

8.5 Confidence Intervals

Problem 1

Suppose that $X_1, \dots, X_n$ form a random sample from the normal distribution with unknown mean μ and known variance σ^2. Let Φ stand for the c.d.f. of the standard normal distribution, the c.d.f. of the standard normal distribution, and let Φ^{-1} be its inverse. Show that the following interval is a coefficient γ confidence interval for μ if X_n is the observed average of the data values:

$$(\overline{X}_n - \Phi^{-1}\left(\frac{1+\gamma}{2}\right)\frac{\sigma}{\sqrt{n}},\ \overline{X}_n + \Phi^{-1}\left(\frac{1+\gamma}{2}\right)\frac{\sigma}{\sqrt{n}})$$

Answer: Let $Z = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$, which is a standard normal distribution. Let c be a real number satisfying $\Pr(-c < Z < c) = \gamma$, we then have $c = \Phi^{-1}(\frac{\gamma+1}{2})$. Substituting Z in the probability expression gives

$$\begin{aligned}
\Pr(-c < Z < c) &= \Pr(-c < \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} < c) \\
&= \Pr(\overline{X}_n - \frac{c\sigma}{\sqrt{n}} < \mu < \overline{X}_n + \frac{c\sigma}{\sqrt{n}}) \\
&= \Pr(\overline{X}_n - \Phi^{-1}\left(\frac{1+\gamma}{2}\right)\frac{\sigma}{\sqrt{n}} < \mu < \overline{X}_n + \Phi^{-1}\left(\frac{1+\gamma}{2}\right)\frac{\sigma}{\sqrt{n}}),
\end{aligned}$$

which means the given expression is the confidence interval for μ.

Problem 3

Suppose that $X_1, \dots, X_n$ form a random sample from the normal distribution with unknown mean μ and unknown variance σ, and let the random variable L denote the length of the shortest confidence interval for mu that can be constructed from the observed values in the sample. Find the value of $\mathrm{E}(L^2)$.

Answer: From the given conditions, we know that

$$L^2 = 4\left[T_{n-1}^{-1}\left(\frac{r+1}{2}\right)\right]^2 \frac{\sigma'^2}{n}.$$

Therefore,

$$\mathrm{E}[L^2] = \mathrm{E}\left[4\left[T_{n-1}^{-1}\left(\frac{r+1}{2}\right)\right]^2 \frac{\sigma^2}{n(n-1)}Y\right],$$

where

$$Y = \frac{\sum_{i=1}^{n}(X_i - \overline{X}_n)2}{\sigma^2},$$

which is a Chi-squared distribution with $n-1$ degrees of freedom. Since its expectation is $n-1$, we have

$$\mathrm{E}[L^2] = 4\left[T_{n-1}^{-1}\left(\frac{r+1}{2}\right)\right]^2 \frac{\sigma^2}{n}.$$

8.6 Unbiased Estimator

Problem 1

Let $X_1, \dots, X_n$ be a random sample from the Poisson distribution with mean θ.
a. Express the $Var_\theta(X_i)$ as a function $\sigma^2 = g(\theta)$.
b. Find the M.L.E. of $g(\theta)$ and show that it is unbiased.

Answer: a. Since X is from a Poisson distribution, $Var_\theta(X_i) = \theta$.
b. The M.L.E. of $g(\theta)$ is $\hat{\theta} = \overline{X}_n$. We know that $\mathrm{E}_\theta[\overline{X}_n] = n\frac{\theta}{n} = \theta = Var_\theta(X_i)$, so the M.L.E. $\hat{\theta}$ is an unbiased estimator.

Problem 3

Suppose that X is a random variable whose distribution is completely unknown, but it is known that all the moments $\mathrm{E}(X^k)$, for $k = 1, 2, \dots$, are finite. Suppose also that $X_1, \dots, X_n$ form a random sample from this distribution.
a. Show that for $k = 1, 2, \dots$, the kth sample moment $(1/n)\sum_{i=1}^n X_i^k$ is an unbiased estimator of $\mathrm{E}(X^k)$.
b. Find an unbiased estimator of $[\mathrm{E}(X)]^2$.

Answer: a. Let $\delta(\mathbf{X}) = (1/n)\sum_{i=1}^n X_i^k$. We have the expectation of this statistic

$$\mathrm{E}_\theta[\delta(\mathbf{X})] = \frac{1}{n}\sum_{i=1}^n \mathrm{E}[X_i^k] = \mathrm{E}[X^k].$$

The last equality holds because for all $k = 1, 2, \dots$, $\mathrm{E}(X^k)$ are finite.