Math 362: Mathematical Statistics II

Le Chen
le.chen@emory.edu

Emory University
Atlanta, GA

Last updated on April 13, 2021

2021 Spring

Chapter 5. Estimation

Chapter 5. Estimation

Rationale. Point estimate doesn't provide precision information.

By using the variance of the estimator, one can construct an interval such that with a high probability that interval will contain the unknown parameter.

- The interval is called **confidence interval**.
- The high probability is **confidence level**.

E.g. 1. A random sample of size 4, $(Y_1 = 6.5,\ Y_2 = 9.2,\ Y_3 = 9.9,\ Y_4 = 12.4)$, from a normal population:

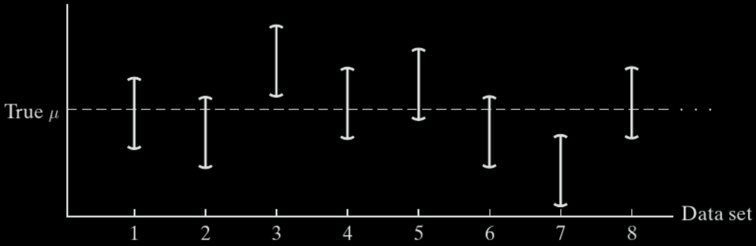
$$f_Y(y;\mu) = \frac{1}{\sqrt{2\pi}\,0.8} e^{-\frac{1}{2}\left(\frac{y-\mu}{0.8}\right)^2}.$$

Both MLE and MME give $\mu_e = \bar{y} = \frac{1}{4}(6.5 + 9.2 + 9.9 + 12.4) = 9.5$.
The estimator $\widehat{\mu} = \overline{Y}$ follows normal distribution.

Construct 95%-confidence interval for μ ...

"The parameter is an unknown constant and no probability statement concerning its value may be made."

–Jerzy Neyman, original developer of confidence intervals.



Possible 95% confidence intervals for μ

In general, for a normal population with σ known, the $100(1-\alpha)\%$ **confidence interval** for μ is

$$\left(\bar{y} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

Comment: There are many variations

1. One-sided interval such as

$$\left(\bar{y} - z_{\alpha}\frac{\sigma}{\sqrt{n}}, \bar{y}\right) \quad \text{or} \quad \left(\bar{y}, \bar{y} + z_{\alpha}\frac{\sigma}{\sqrt{n}}\right)$$

2. σ is unknown and sample size is small: z-score $\rightarrow$ t-score
3. σ is unknown and sample size is large: z-score by CLT
4. Non-Gaussian population but sample size is large: z-score by CLT

Theorem. Let k be the number of successes in n independent trials, where n is large and $p = \mathbb{P}(\textit{success})$ is unknown. An approximate $100(1-\alpha)\%$ confidence interval for p is the set of numbers

$$\left(\frac{k}{n} - z_{\alpha/2}\sqrt{\frac{(k/n)(1-k/n)}{n}},\ \frac{k}{n} + z_{\alpha/2}\sqrt{\frac{(k/n)(1-k/n)}{n}}\right).$$

Proof: It follows the following facts:

- $X \sim$binomial(n,p) iff $X = Y_1 + \cdots + Y_n$, while Y_i are i.i.d. Bernoulli(p):

$$\mathbb{E}[Y_i] = p \quad \text{and} \quad \mathrm{Var}(Y_i) = p(1-p).$$

- **Central Limit Theorem**: Let $W_1, W_2, \cdots, W_n$ be an sequence of i.i.d. random variables, whose distribution has mean μ and variance σ^2, then

$$\frac{\sum_{i=1}^n W_i - n\mu}{\sqrt{n\sigma^2}} \quad \textit{approximately} \text{ follows } \quad N(0,1), \quad \text{when } n \text{ is large.}$$

- When the sample size n is large, by the central limit theorem,

$$\frac{\sum_{i=1}^{n} Y_i - np}{\sqrt{np(1-p)}} \quad \overset{\text{ap.}}{\sim} \quad N(0,1)$$

$$\|$$

$$\frac{X - np}{\sqrt{np(1-p)}} = \frac{\frac{X}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} \approx \frac{\frac{X}{n} - p}{\sqrt{\frac{p_e(1-p_e)}{n}}}$$

- Since $p_e = \frac{k}{n}$, we see that

$$\mathbb{P}\left(-z_{\alpha/2} \leq \frac{\frac{X}{n} - p}{\sqrt{\frac{\frac{k}{n}\left(1-\frac{k}{n}\right)}{n}}} \leq z_{\alpha/2}\right) \approx 1 - \alpha$$

i.e., the $100(1-\alpha)\%$ confidence interval for p is

$$\left(\frac{k}{n} - z_{\alpha/2}\sqrt{\frac{(k/n)(1-k/n)}{n}},\ \frac{k}{n} + z_{\alpha/2}\sqrt{\frac{(k/n)(1-k/n)}{n}}\right).$$

$\square$

E.g. 1. Use *median test* to check the randomness of a random generator.

Suppose $y_1, \cdots, y_n$ denote measurements presumed to have come from a continuous pdf $f_Y(y)$. Let k denote the number of y_i's that are less than the median of $f_Y(y)$. If the sample is random, we would expect the difference between $\frac{k}{n}$ and $\frac{1}{2}$ to be small. More specifically, a 95% confidence interval based on k should contain the value 0.5.

Let $f_Y(y) = e^{-y}$. The median is $m = 0.69315$.

```r
#! /usr/bin/Rscript
main <- function() {
  args <- commandArgs(trailingOnly = TRUE)
  n <- 100 # Number of random samples.
  r <- as.numeric(args[1]) # Rate of the exponential
  # Check if the rate argument is given.
  if (is.na(r)) return("Please provide the rate and try again.")

  # Now start computing ...
  f <- function (y) pexp(y, rate = r)-0.5
  m <- uniroot(f, lower = 0, upper = 100, tol = 1e-9)$root
  print(paste("For rate ", r, "exponential distribution,",
              "the median is equal to ", round(m,3)))
  data <- rexp(n,r) # Generate n random samples
  data <- round(data,3) # Round to 3 digits after decimal
  data <- matrix(data, nrow = 10,ncol = 10) # Turn the data to a matrix
  prmatrix(data) # Show data on terminal
  k <- sum(data > m) # Count how many entries is bigger than m
  lowerbd = k/n - 1.96 * sqrt((k/n)*(1-k/n)/n);
  upperbd = k/n + 1.96 *sqrt((k/n)*(1-k/n)/n);
  print(paste("The 95% confidence interval is (",
              round(lowerbd,3), ",",
              round(upperbd,3), ")"))
}
main()
```

Try commandline ...

```
Math362:./Example-5-3-2.R 1
[1] "For rate  1 exponential distribution, the median is equal to  0.693"
       [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9] [,10]
 [1,] 1.324 1.211 0.561 0.640 2.816 2.348 0.788 2.243 1.759 0.103
 [2,] 0.476 2.288 0.106 0.079 0.636 1.941 0.801 3.838 0.612 0.030
 [3,] 1.085 0.305 0.354 1.013 0.687 1.656 1.043 0.389 1.476 2.158
 [4,] 1.267 1.031 0.917 0.681 0.912 0.236 0.054 0.862 0.065 0.402
 [5,] 0.957 1.003 1.665 1.137 0.378 1.182 0.659 1.923 1.127 0.364
 [6,] 0.307 0.127 0.203 0.394 1.392 2.378 4.192 0.365 3.227 0.337
 [7,] 0.707 0.049 0.391 1.967 1.220 2.605 0.887 1.749 1.479 1.526
 [8,] 0.662 0.141 0.318 0.523 0.646 1.202 0.442 0.174 1.178 0.177
 [9,] 0.397 0.493 0.214 0.522 2.024 4.109 1.268 1.041 0.948 0.382
[10,] 2.260 0.292 0.437 0.962 0.224 4.221 0.594 0.218 0.601 0.941
[1] "The 95% confidence interval is ( 0.422 , 0.618 )"
Math362:./Example-5-3-2.R 10
[1] "For rate  10 exponential distribution, the median is equal to  0.069"
       [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9] [,10]
 [1,] 0.199 0.069 0.013 0.025 0.000 0.107 0.068 0.116 0.066 0.146
 [2,] 0.027 0.076 0.044 0.458 0.052 0.127 0.100 0.100 0.014 0.061
 [3,] 0.014 0.078 0.044 0.072 0.028 0.141 0.038 0.022 0.037 0.093
 [4,] 0.042 0.015 0.250 0.132 0.292 0.072 0.105 0.244 0.046 0.054
 [5,] 0.134 0.074 0.182 0.057 0.021 0.038 0.095 0.196 0.004 0.048
 [6,] 0.016 0.021 0.163 0.030 0.139 0.063 0.054 0.006 0.023 0.051
 [7,] 0.227 0.055 0.091 0.121 0.066 0.114 0.004 0.021 0.035 0.211
 [8,] 0.113 0.083 0.129 0.338 0.160 0.008 0.014 0.167 0.050 0.127
 [9,] 0.053 0.073 0.054 0.098 0.004 0.036 0.274 0.276 0.004 0.159
[10,] 0.045 0.469 0.152 0.003 0.129 0.017 0.084 0.072 0.162 0.007
[1] "The 95% confidence interval is ( 0.392 , 0.588 )"
Math362:
```

Instead of the C.I. $\left(\frac{k}{n} - z_{\alpha/2}\sqrt{\frac{(k/n)(1-k/n)}{n}},\ \frac{k}{n} + z_{\alpha/2}\sqrt{\frac{(k/n)(1-k/n)}{n}}\right)$.

One can simply specify the mean $\dfrac{k}{n}$ and

the **margin of error**: $$d := z_{\alpha/2}\sqrt{\frac{(k/n)(1-k/n)}{n}}.$$

$$\max_{p\in(0,1)} p(1-p) = p(1-p)\Big|_{p=1/2} = 1/4 \quad\Longrightarrow\quad d \le \frac{z_{\alpha/2}}{2\sqrt{n}} =: d_m.$$

Comment:

1. When p is close to $1/2$, $d \approx \frac{z_{\alpha/2}}{2\sqrt{n}}$, which is equivalent to $\sigma_p \approx \frac{1}{2\sqrt{n}}$.

   E.g., $n = 1000$, $k/n = 0.48$, and $\alpha = 5\%$, then

$$d = 1.96\sqrt{\frac{0.48 \times 0.52}{1000}} = 0.0309\underline{7} \quad \text{and} \quad d_m = \frac{1.96}{2\sqrt{1000}} = 0.0309\underline{9}$$

$$\sigma_p = \sqrt{\frac{0.48 \times 0.52}{1000}} = 0.015\underline{7}9873 \quad \text{and} \quad \sigma_p \approx \frac{1}{2\sqrt{1000}} = 0.015\underline{8}1139.$$

2. When p is away from $1/2$, the discrepancy between d and d_m becomes big....

E.g. Running for presidency. Max and Sirius obtained 480 and 520 votes, respectively. What is probability that Max will win?

What if the sample size is $n = 5000$, and Max obtained 2400 votes.

Choosing sample sizes

$$d \leq z_{\alpha/2}\sqrt{p(1-p)/n} \quad \Longleftrightarrow \quad n \geq \frac{z_{\alpha/2}^2 p(1-p)}{d^2} \qquad \text{(When } p \text{ is known)}$$

$$d \leq \frac{z_{\alpha/2}}{2\sqrt{n}} \quad \Longleftrightarrow \quad n \geq \frac{z_{\alpha/2}^2}{4d^2} \qquad \text{(When } p \text{ is unknown)}$$

E.g. Anti-smoking campaign. Need to find an 95% C.I. with a margin of error equal to 1%. Determine the sample size?

Answer: $n \geq \frac{1.96^2}{4 \times 0.01^2} = 9640.$

E.g.' In order to reduce the sample size, a small sample is used to determine p. One finds that $p \approx 0.22$. Determine the sample size again.

Answer: $n \geq \frac{1.96^2 \times 0.22 \times 0.78}{\times 0.01^2} = 6592.2.$