Math 170S

Lecture Notes Section 7.5 *†

Confidence interval for percentiles

Instructor: Swee Hong Chan

NOTE: Materials that appear in the textbook but do not appear in the lecture notes might still be tested. Please send me an email if you find typos.

*Version date: Tuesday 3rd November, 2020, 19:59.

†This notes is based on Hanbaek Lyu's and Liza Rebrova's notes from the previous quarter, and I would like to thank them for their generosity. "*Nanos gigantum humeris insidentes* (I am but a dwarf standing on the shoulders of giants)".

1 Example: Median

The friendly instructor (FI) wants to know the median for the grade of Midterm. He selected the grade of 32 students randomly:

72	75	76	78	80
81	82	83	85	85
85	86	86	87	88
89	91	92	93	94
95	95	96	96	98
100	100	100	100	100
100	100			

Help the FI create a confidence interval for the median of Midterm.

2 Settings: Median

- **Object:** Y is an **unknown continuous** random variable.
- **Input:**
  - Random samples $Y_1, \ldots, Y_n$ for Y, ordered from smallest to largest.
- **Output:** Two real numbers a and b (with $a < b$) and α that allows us to say

  "The **median** m is contained in the interval $[a, b]$ with confidence (approximately) 1-α."

  The values a and b will be chosen from $Y_1, \ldots, Y_n$.

The interval $[a, b]$ is the **confidence interval** for m, and the **confidence constant** is $1 - \alpha$.

3 Answer I for Example: Median

Let a be the seventh lowest grade and b be the eighth lowest grade:

$$a = Y_7 = 82; \qquad b = Y_8 = 83.$$

We now compute the probability that the median m is between a and b,

$$P(Y_7 < m < Y_8).$$

For the event $Y_7 < m < Y_8$ to happen, we need to have **exactly** 7 students score below m, and exactly $n - 7$ students score above m.

This is **binomial random variable** on $n = 32$ people with $x = 7$ success and with success probability $p = P(X < m) = \frac{1}{2}$, so we have (BT)

$$\begin{aligned} P(Y_7 < m < Y_8) &= \binom{n}{x} (p)^x (1-p)^{n-x} \\ &= \binom{32}{7} \left(\frac{1}{2}\right)^7 \left(\frac{1}{2}\right)^{25} \approx 0.00078. \end{aligned}$$

The confidence constant is equal to

$$1 - \alpha = P(Y_7 < m < Y_8) = 0.00078,$$

and the confidence interval is

$$[a, b] = [Y_7, Y_8] = [82, 83].$$

Hence we conclude that the median m is contained in the interval $\big[82, 83\big]$ with confidence 0.00078.

4 Median: Form 1

Theorem 1. *Suppose that*

$$a = Y_x; \qquad b = Y_{x+1},$$

for some $x \in \{1, \ldots, n-1\}$. *Then the confidence constant* $1 - \alpha$ *is equal to*

$$1 - \alpha = \binom{n}{x} \frac{1}{2^n}. \qquad \square$$

5 Second Answer to Example: Median

Let a be the seventh lowest grade and b be the tenth lowest grade:

$$a = Y_7 = 82; \qquad b = Y_{10} = 85.$$

We want to compute the probability that the median m is between a and b,

$$P(Y_7 \ < \ m \ < Y_{10}).$$

The event $Y_7 < m < Y_{10}$ consists of the union of these three (mutually exclusive) cases:

- Exactly 7 students score below m, and $n - 7$ score above m, which has probability

$$\binom{32}{7} \frac{1}{2^{32}}.$$

- Exactly 8 students score below m, and $n - 8$ score above m, which has probability

$$\binom{32}{8} \frac{1}{2^{32}}.$$

- Exactly 9 students score below m, and $n - 9$ score above m, which has probability

$$\binom{32}{9} \frac{1}{2^{32}}.$$

Combining these three events, we have

$$\begin{aligned} & P(Y_7 \ < \ m \ < Y_{10}) \\ = & \binom{32}{7} \frac{1}{2^{32}} + \binom{32}{8} \frac{1}{2^{32}} + \binom{32}{9} \frac{1}{2^{32}} \\ \approx & \ 0.0098. \end{aligned}$$

Then the confidence constant is equal to

$$1 - \alpha \ = \ P(Y_7 \ < \ m \ < Y_{10}) \ \approx \ 0.0098,$$

and the confidence interval is

$$[a, b] = [Y_7, Y_{10}] = [82, 85].$$

Hence we conclude that the median m is contained in the interval $\big[82, 85\big]$ with confidence 0.0098.

6 Median: Form 2

Theorem 2. *Suppose that*

$$a = Y_i; \qquad b = Y_j$$

with $i < j$. *Then the confidence constant is equal to*

$$1 - \alpha = \frac{1}{2^n} \sum_{x=i}^{j-1} \binom{n}{x}. \qquad \square$$

Remark 3. This theorem should be used **only when** $j - i$ **is small**.

7 Third answer: Example median

Let a be the 12th lowest grade and b be the 25-th lowest grade:

$$a = Y_{12} = 86; \qquad b = Y_{25} = 98.$$

It follows from Theorem 2 that

$$P(Y_{12} \ < \ m \ < Y_{25}) \ \approx \ \sum_{x=12}^{24} \binom{32}{x} \frac{1}{2^{32}}.$$

It is impractical to compute such a large sum. Instead, we appreciate the sum with normal random variables (see Section 5.7)

$$\sum_{x=i}^{j-1} \binom{n}{x} \frac{1}{2^n} \ = \ \Phi\left(\frac{(j-0.5)-(\frac{n}{2})}{\sqrt{\frac{n}{4}}}\right) - \Phi\left(\frac{(i-0.5)-(\frac{n}{2})}{\sqrt{\frac{n}{4}}}\right),$$

where Φ is the cdf of the standard normal distribution.

Plugging in n,x,i, and j, we get

$$\begin{aligned}
&P(Y_{12} \ < \ m \ < Y_{25}) \\
&\approx \Phi\left(\frac{(25-0.5)-(\frac{32}{2})}{\sqrt{\frac{32}{4}}}\right) \ - \ \Phi\left(\frac{(12-0.5)-(\frac{32}{2})}{\sqrt{\frac{32}{4}}}\right) \\
&\approx \Phi(3.01)-\Phi(-1.59) \ = \ (0.9987)-(1-0.9441) \\
&= 0.9428.
\end{aligned}$$

Then the confidence constant is equal to

$$1-\alpha \ = \ P(Y_{12} \ < \ m \ < Y_{25}) \ = \ 0.9428,$$

and the confidence interval is

$$[a,b] = [Y_{12}, Y_{25}] = [86, 98].$$

Hence we conclude the median m is contained in the interval $\big[86, 98\big]$ with confidence 0.9428.

8 Median: Form 3

Theorem 4. *Suppose that*

$$a = Y_i; \qquad b = Y_j$$

with $i < j$ *such that* $j - i$ ***is large****. Then the confidence constant is* ***approximately***

$$1-\alpha \approx \Phi\left(\frac{(j-0.5)-(\frac{n}{2})}{\sqrt{\frac{n}{4}}}\right) - \Phi\left(\frac{(i-0.5)-(\frac{n}{2})}{\sqrt{\frac{n}{4}}}\right).$$

□

9 Confidence interval for percentiles: Settings

- **Object:** Y is an **unknown continuous** RV.
- **Input:**
  - Random samples $Y_1, \ldots, Y_n$ for Y, ordered from smallest to largest.
- **Output:** Two real numbers a and b (with $a < b$) and α that allows us to say

  "The **p-th percentile** π_p is contained in the interval $[a, b]$ with confidence 1-α."

  The values a and b will be chosen from $Y_1, \ldots, Y_n$.

The interval $[a, b]$ is the **confidence interval** for π_p, and the **confidence constant** is $1 - \alpha$.

10 Percentiles: Theorem

Theorem 5. *Suppose that*

$$a = Y_i; \qquad b = Y_j$$

with $i < j$. Then the confidence constant is equal to

$$1 - \alpha \;=\; \sum_{x=i}^{j-1} \binom{n}{x} p^x (1-p)^{n-x}.$$

Also suppose that $j - i$ ***is large****. Then the confidence constant is* ***approximately***

$$1 - \alpha \;\approx\; \Phi\left(\frac{(j-0.5) - np}{\sqrt{np(1-p)}}\right) \;-\; \Phi\left(\frac{(i-0.5) - np}{\sqrt{np(1-p)}}\right).$$

□

11 Example: Percentiles

Consider the Midterm-grade example from before. Help the FI create a confidence interval for the first quartile of Midterm.

12 Answer for Example: Percentiles

Let a be the 7-th lowest grade and b be the 12-th lowest grade:

$$a = Y_7 = 82; \qquad b = Y_{12} = 86; \qquad p = \frac{1}{4}.$$

We have from Theorem 5 that

$$\begin{aligned} 1-\alpha &\approx \Phi\left(\frac{(12-0.5)-(\frac{32}{4})}{\sqrt{(32)(\frac{1}{4})(\frac{3}{4})}}\right) - \Phi\left(\frac{(7-0.5)-(\frac{32}{4})}{\sqrt{(32)(\frac{1}{4})(\frac{3}{4})}}\right) \\ &\approx \Phi(1.43) - \Phi(-0.61) \\ &= (0.9236) - (1-0.7291) \\ &= 0.6527. \end{aligned}$$

The confidence interval is equal to

$$[a, b] = [Y_7, Y_{12}] = [82, 86].$$

Hence we conclude that

"The first quartile is contained in the interval $\big[82, 86\big]$ with confidence 0.6527."

Remark 6. Note that in this section we d **not** assume any prior knowledge on the unknown random variables (other than it is continuous). This is in contrast to materials from other sections (e.g., normal random variables in Section 7.1 and Bernoulli random variables in Section 7.3). This is why the methods here are sometimes called *distribution-free methods*.