Introduction to Probability

Dimitri P. Bertsekas and John N. Tsitsiklis
2nd edition
2008
ISBN: 9781886529236

SECTION 5.1. Some Useful Inequalities
Problem 1

Problem 1. A statistician wants to estimate the mean height h (in meters) of a population, based on n independent samples $X_i, ..., X_n$, chosen uniformly from the entire population. He uses the sample mean $M_n = (X_1 + ... + X_n)/n$ as the estimate of h, and a rough guess of 1.0 meters for the standard deviation of the samples X_i.
(a) How large should n be so that the standard deviation of M_n is at most 1 centimeter?
(b) How large should n be so that Chebyshev's inequality guarantees that the estimate is within 5 centimeters from h, with probability at least 0.99?
(c) The statistician realizes that all persons in the population have heights between 1.4 and 2.0 meters, and revises the standard deviation figure that he uses based on the bound of Example 5.3. How should the values of n obtained in parts (a) and (b) be revised?

Solution:

(a)

$$
\begin{aligned}
var[M_n] &= var\left[\frac{(X_1 + ... + X_n)}{n}\right] && \text{definition } X_n \\
&= \frac{1}{n^2} var[X_1 + ... + X_n] \\
&= \frac{1}{n^2}\left[var[X_1] + ... + var[X_n]\right] \\
&= \frac{1}{n^2}\left[var[X_i] + ... + var[X_i]\right] \\
&= \frac{1}{n^2}\left[n * var[X_i]\right] \\
&= \frac{n * var[X_i]}{n^2} \\
var[M_n] &= \frac{var[X_i]}{n}
\end{aligned}
$$

$$
\begin{aligned}
sd[M_n] &= \sqrt{var[M_n]} \\
&= \sqrt{\frac{var[X_i]}{n}} \\
&= \frac{\sqrt{var[X_i]}}{\sqrt{n}}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{\sqrt{10.000cm^2}}{\sqrt{n}} \\
&= \frac{100cm}{\sqrt{n}}
\end{aligned}
$$

$$
\begin{aligned}
\frac{100cm}{\sqrt{n}} &< 1cm \\
\frac{100cm}{1cm} &< \sqrt{n} \\
100 &< \sqrt{n} \\
\sqrt{n} &> 100 \\
(\sqrt{n})^2 &> (100)^2 \\
n &> 10.000
\end{aligned}
$$

Free Material:

[3, section 19.3, page 798]
`http://crosslinks.mit.edu/topic/variance/`

Books:

[1, section 2.4, page 81]
[2, section 4.5, page 132]

(b)
Although the problem asks to use Chebyshev, we can start using Markov, just to see their difference. "Markov Inequality" states that:

$$P(M_n \geq a) \leq \frac{E[M_n]}{a}$$

$$\begin{aligned}
E[M_n] &= E\left[\frac{(X_1 + ... + X_n)}{n}\right] \\
&= \frac{1}{n}E[X_1 + ... + X_n] \\
&= \frac{1}{n}\left[E[X_1] + ... + E[X_n]\right] \\
&= \frac{1}{n}\left[E[X_i] + ... + E[X_i]\right] \\
&= \frac{1}{n}\left[n * E[X_i]\right] \\
&= \frac{n * E[X_i]}{n} \\
&= E[X_i]
\end{aligned}$$

$$P(M_n \geq a) \leq \frac{E[X_i]}{a}$$

Now it is clear that we cannot derivate anything using just the "Markov Inequality", because $E[M_n]$ does not depends on n. So let us try using "Chebyshev Inequality".

Expected Value:

Free Material:

[3, section 18.4, page 751]
`http://crosslinks.mit.edu/topic/expected-value/`

Books:

[1, section 2.4, page 81]
[2, section 4.4, page 128]

Markov Inequality:

Free Material:

[3, section 19.1, page 789]
`http://crosslinks.mit.edu/topic/markovs-theorem/`

Books:

[1, section 5.1, page 265]
[2, section 8.2, page 388]

"Chebyshev Inequality" is just a simple derivation from "Markov Inequality", but will help us in this problem.

$$\begin{aligned}
P(|M_n - E[M_n]| \geq a) &\leq \frac{var[M_n]}{a^2} \\
&\leq \frac{10000cm^2}{n} * \frac{1}{a^2}
\end{aligned}$$

$$\begin{aligned}
P(|M_n - E[M_n]| \geq 5cm) &\leq \frac{10000cm^2}{n} * \frac{1}{(5cm)^2} \\
&\leq \frac{10000cm^2}{n} * \frac{1}{25cm^2} \\
&\leq \frac{10000cm^2}{25cm^2} * \frac{1}{n} \\
&\leq 400 * \frac{1}{n} \\
&\leq \frac{400}{n}
\end{aligned}$$

But the problem asks the probability of being withing 5cms and not outside. So we need to "invert" the probability. And we want this probability to be at least 0.99.

$$
\begin{aligned}
P(|M_n - E[M_n]| \leq 5cm) &= 1 - P(|M_n - E[M_n]| \geq 5cm) \\
0.99 &<= P(|M_n - E[M_n]| \leq 5cm) \\
0.99 &<= 1 - P(|M_n - E[M_n]| \geq 5cm) \\
0.99 &\leq 1 - \frac{400}{n} \\
0.99 - 1 &\leq [1 - \frac{400}{n}] - 1 \\
-0.01 &\leq -\frac{400}{n} \\
0.01 &\geq \frac{400}{n} \\
n &\geq \frac{400}{0.01} \\
n &\geq 40.000
\end{aligned}
$$

Free Material:

[3, section 19.2, page 792]
`http://crosslinks.mit.edu/topic/chebyshevs-inequality/`

Books:

[1, section 5.1, page 267]
[2, section 8.2, page 389]

(c)
Given that all sampled heights are bounded, we can use the "variance definition" to find a bound to the variance. If this bound is smaller than the initial variance of $1m^2$ maybe we can have a better approximation to the probability. We start with the "variance definition".

Gamma (γ), in the following equations, has no special meaning, it is real number.

$$E[(X - \gamma)^2] = E[X^2] - 2\gamma E[X] + \gamma^2$$

We know that this equation is minimized when $\gamma = E[x]$. Wich means that when γ is different than $E[x]$ the value of the equation is greater.

$$var[X] = E[(X - E[X])^2] \leq E[(X - \gamma)^2]$$

If we choose γ as $\frac{a+b}{2}$ with a and b as the lower and higher bounds.

$$
\begin{aligned}
var[X] &= E[(X - E[X])^2] \leq E\left[\left(X - \frac{a+b}{2}\right)^2\right] \\
E\left[\left(X - \frac{a+b}{2}\right)^2\right] &= E\left[X^2 - 2X\left(\frac{a+b}{2}\right) + \left(\frac{a+b}{2}\right)^2\right] \\
&= E\left[X^2 - 2X\frac{a}{2} - 2X\frac{b}{2} + \left(\frac{a+b}{2}\right)^2\right] \\
&= E\left[X^2 - Xa - Xb + \left(\frac{a+b}{2}\right)^2\right] \\
&= E\left[X^2 - Xa - Xb + \left(\frac{a^2+2ab+b^2}{4}\right)\right] \\
&= E\left[X^2 - Xa - Xb + \frac{2ab}{4} + \left(\frac{a^2+b^2}{4}\right)\right. \\
&= E\left[X^2 - Xa - Xb + \frac{ab}{2} + \left(\frac{a^2+b^2}{4}\right)\right] \\
&= E\left[X^2 - Xa - Xb + \frac{ab}{2} + (\frac{ab}{2}) + \left(\frac{a^2+b^2}{4}\right) - (\frac{ab}{2})\right]
\end{aligned}
$$

$$
\begin{aligned}
&= E\left[X^2 - Xa - Xb + \frac{2ab}{2} + \left(\frac{a^2+b^2}{4}\right) - (\frac{2ab}{4})\right] \\
&= E\left[X^2 - Xa - Xb + ab + \left(\frac{a^2+b^2-2ab}{4}\right)\right] \\
&= E\left[X^2 - Xa - Xb + ab\right] + \left(\frac{a^2+b^2-2ab}{4}\right) \\
&= E[(X-a)(X-b)] + \frac{(a-b)^2}{4}
\end{aligned}
$$

The interesting part now is that a and b are the bounds of the variable. So $X - b$ is always negative, because X must be smaller than b. Wich makes

$$(X - a)(X - b) < 0$$

.

And the expected value will be negative, which will always decrease $\frac{(a-b)^2}{4}$. Giving us.

$$= E[(X-a)(X-b)] + \frac{(a-b)^2}{4} \leq \frac{(a-b)^2}{4}$$

Which gives us the expectacular result that:

$$var[x] \leq \frac{(a-b)^2}{4}$$

Going back to the problem, the statistician realized that the bound is 1.4 and 2.0. So we have that the maximum possible variance is:

$$\begin{aligned} var[X] \leq \frac{(a-b)^2}{4} &= \frac{(1.4m - 2.0m)^2}{4} \\ &= \frac{(0.6m)^2}{4} \\ &= \frac{0.36m^2}{4} \\ &= 0.09m^2 \\ &= 900cm^2 \\ sd[x] &= \sqrt{var[X]} \\ &= \sqrt{900cm^2} \\ &= 30cm \end{aligned}$$

Doing parts (a) and (b) again we have.
(new a)

$$\begin{aligned} \frac{30cm}{\sqrt{n}} &< 1cm \\ \frac{30cm}{1cm} &< \sqrt{n} \\ 30 &< \sqrt{n} \\ (30)^2 &< n \\ n &> 900 \end{aligned}$$

(new b)

$$\begin{aligned} 0.01 &\geq \frac{900cm^2}{25cm^2} * \frac{1}{n} \\ 0.01 &\geq 36 * \frac{1}{n} \\ n &\geq \frac{36}{0.01} \\ n &\geq 3.600 \end{aligned}$$

References

[1] Dimitri P. Bertsekas and John N. Tsitsiklis, Introduction to Probability, 2nd edition, 2008, ISBN: 9781886529236, http://www.athenasc.com/probbook.html

[2] Sheldon Ross A First Course in Probability, 5th edition, 2012, ISBN: 9780137463145

[3] Mathematics for Computer Science, https://courses.csail.mit.edu/6.042/spring18/