Median, Quantiles, Variance

MATH/STAT 394: Probability I
Summer 2021 A Term

Introduction to Probability
D. Anderson, T.Seppäläinen, B. Valkó

§ 3.3, 3.4

Aaron Osgood-Zimmerman

Department of Statistics

Logistics

- midterm solutions will be up this evening
- I'm about half way through grading them. done tomorrow at the latest
- Feedback:
  - pace of the course is generally viewed as fast
  - use of class time is generally well received
  - homeworks feel difficult compared to class examples
  - more practice problems requested
- Some of this is forced by the condensed summer term
- Some of this is by design
- I've included more practice problems in the first part of today's lecture
- I'll post some practice problem resources later today
- I'll update the course calendar in case you would like to read ahead
- curves

Practice solution

Practice

The annual maximum one-day rainfall can be modeled by a r.v. X with p.d.f.

$$f(x) = \begin{cases} \frac{2}{\pi}\frac{1}{x^2+1} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

What is the expectation of X?

Solution

- First, you can check that f is a valid p.d.f. (see quiz lecture 15)
- Idea: we would like to compute $\int_{-\infty}^{+\infty} xf(x)dx = \frac{2}{\pi}\int_0^{+\infty} \frac{x}{x^2+1}dx$
- As $x \to +\infty$, $\frac{x}{x^2+1} \sim \frac{1}{x}$ and $\int_1^{+\infty} \frac{1}{x}dx = +\infty$ so we are going to show that the expectation is infinite in this case
- For $x \geq 1$, $\frac{x}{x^2+1} \geq \frac{1}{2x}$, so for any $b > 1$,

$$\int_0^b xf(x)dx \geq \int_1^b xf(x)dx \geq \frac{1}{\pi}\int_1^b \frac{1}{x}dx = \frac{1}{\pi}\log b$$

- Therefore

$$\mathbb{E}[X] = \int_0^{+\infty} xf(x)dx \geq \lim_{b\to+\infty} \log b = +\infty, \text{the expectation is infinite!}$$

Recap

Expectation of a function of a r.v.

- if X is discrete with p.m.f. p, $\mathbb{E}[g(X)] = \sum_{k \in \mathcal{X}} g(k)p(k)$
- if X is continuous with p.d.f. f $\mathbb{E}[g(X)] = \int_{-\infty}^{+\infty} g(x)f(x)dx$

Properties of Expectation

- For 1_A the indicator r.v. of an event A,

$$\mathbb{E}[1_A] = \mathbb{P}(A)$$

- **Linearity of the expectation**: for any r.v. X, Y and $a, b \in \mathbb{R}$

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b, \qquad \mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

Outline

Median, quantiles

Variance

Median

Motivation

- The expectation often gives a good summary of a r.v.
- Yet, if the r.v. has some abnormally large values, the expectation may be a bad indicator of where the center of the distribution lies
- Another indicator is often used: the median that tells us where to split the distribution of X to have equal mass on the left and right sides of the median

Definition

*The **median** of a continuous r.v. X is a value m s.t.*

$$\mathbb{P}(X \geq m) = \mathbb{P}(X \leq m) = 1/2$$

*More generally, the **median** of a r.v. X is any value m such that*

$$\mathbb{P}(X \geq m) \geq 1/2 \qquad \mathbb{P}(X \leq m) \geq 1/2$$

Note:

- Fundamentally, is a value in $\mathbb{R}$ that splits the dist. of X in two equal parts
- Generally the median is not unique, see next example.
- In the second definition, we want to take into account the possibility that m has a non-zero probability.
- Namely the median of a uniform dist. on $\{-1, 0, 1\}$ is 0 according to the second definition.
- Without the "$\geq 1/2$" instead of "$= 1/2$" the median would not exist in this case.

Median

Example

Let X be uniformly distributed on $\{-100, 1, 2, 3, \ldots 9\}$. So X has a prob. dist.

$$\mathbb{P}(X = -100) = 1/10, \qquad \mathbb{P}(X = k) = 1/10 \quad \text{for } k \in \{1, \ldots 9\}$$

What are the expectation and the median of X?

Solution

- $\mathbb{E}[X] = -100 \cdot 1/10 + (1 + 2 + \ldots + 9) \cdot 1/10 = -5.5$
- On the other hand,

$$\mathbb{P}(X \leq 4.5) = p(-100) + p(1) + p(2) + p(3) + p(4) = 1/2$$
$$P(X \geq 4.5) = p(5) + \ldots + p(9) = 1/2$$

- So 4.5 is **a** median for X
- Note that any $m \in [4, 5]$ is a median for X, we usually take the mid-point of the interval
- Note that 4.5 illustrates much better the fact that 90% of the dist. is in $\{1, \ldots, 9\}$
- Whereas the mean is dominated by -100 (which represents only one value amoing the 10 possible)

Quantiles

Motivation

- What else could characterize our r.v.?
- Typically we would like to know if some observation of our r.v. is **rare** or not
- Namely we would like to have access to a value x, such that if $X \geq x$ then the probability of this observation is small
- This is formalized with the definitions of **quantiles**

Quantiles

Definition

*Given $0 \leq p \leq 1$ (e.g. $p = 90/100$), the **pth quantile** of a continuous r.v. X is any value x_p such that*

$$\mathbb{P}(X \leq x_p) = p \qquad \mathbb{P}(X \geq x_p) = 1 - p$$

*More generally the **p^{th} quantile** of a r.v. X is any value x_p such that*

$$\mathbb{P}(X \leq x_p) \geq p \qquad \mathbb{P}(X \geq x_p) \geq 1 - p$$

Notes

- Note that for $p = 1/2$ we retrieve the median!
- Here for $p = 90/100$, the p^{th} quantile tells us that there is less than 10% of chance to observe a value greater than x_p

Outline

Variance

Motivation

- The expectation summarizes the r.v. to a single point
- Generally the distribution should gather around the mean, but by how much?
- The variance informs us about the *dispersion* of the r.v. around the mean

Variance

Definition

*The **variance** of a r.v. X with mean μ is defined as*

$$\mathrm{Var}(X) = \mathbb{E}\left[(X-\mu)^2\right]$$

It is often denoted σ_X^2

*The square root of the variance σ is called the **standard deviation**.*

In terms of p.m.f. or p.d.f. we have that

$$\mathrm{Var}(X) = \sum_{k\in\mathcal{X}} (x-\mu)^2 p(k) \qquad \text{for a discrete r.v. with p.m.f. } p$$

$$\mathrm{Var}(X) = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)dx \qquad \text{for a continuous r.v. with p.d.f. } f$$

Note:

- Note that the variance is simply defined through the expectation of a function of the r.v.
- Numerous key characteristics of a r.v. are defined that way, namely the expectation is our main tool
- Note that, as for the expectation, the variance may be finite, infinite or undefined

Variance

Example
Consider two investment strategies.

1. One, denoted X, yields a profit of 1\$ or -1\$ with equal prob. 1/2,
2. Another one, denoted Y yields a profit of 100\$ or -100\$ with equal prob. 1/2

What is the mean and the variance of each investment?

Solution

- Clearly $\mathbb{E}[X] = \mathbb{E}[Y] = 0$
- On the other hand,

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = (1-0)^2 \cdot \frac{1}{2} + (-1-0)^2 \cdot \frac{1}{2} = 1$$

$$\mathrm{Var}(Y) = \mathbb{E}[(Y - E[Y])^2] = (100-0)^2 \cdot \frac{1}{2} + (-100-0)^2 \cdot \frac{1}{2} = 10000$$

- Though X and Y have same mean, Y varies much more, this is reflected in the variance.

Variance

Exercise

Consider $X \sim \text{Ber}(p)$, what is $\text{Var}(X)$?

Solution

- Recall that $\mathbb{E}[X] = p$
- So
$$\text{Var}[X] = \mathbb{E}[(X-p)^2] = (1-p)^2 \cdot p + (0-p)^2 \cdot (1-p) = p(1-p)$$
- Note that if $X = 1_A$ for some event A, then
$$\text{Var}(X) = \mathbb{P}(A)\mathbb{P}(A^c)$$

Variance

Lemma
The variance of a r.v. X can also be expressed as

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

Proof

$$\begin{aligned}
\mathrm{Var}(X) = \mathbb{E}[(X-\mu)^2] &= \mathbb{E}[X^2 - 2X\mu + \mu^2] \\
&= \mathbb{E}[X^2] - 2\mathbb{E}[X]\mu + \mu^2 \qquad \text{(linearity of the expectation)} \\
&= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \qquad \text{(because } \mu = \mathbb{E}[X]\text{)}
\end{aligned}$$

Practice next lecture

Practice

Consider that the time to wait for your bus is modeled by $X \sim \mathrm{Exp}(\lambda)$.

1. What is the 90/100th quantile for $\lambda = 1$?
2. Let's say that the 90/100th quantile is 20min (λ is unknown). Give an upper bound on the prob. that you wait more 30min