Probability and Mathematical Statistics in Data Science

Lecture 12: 5.1: Definition Section 5.2: Functions of Random Variables

Expectation

- Random variables have distributions, so they have centers and spreads.
- The expected value (mean value or expectation) of a random variable describes its theoretical long-run average value.
- We typically use μ or $E(X)$ to denote the mean of a random variable X.

Notes about E(X)

- Same units as X
- Not necessarily an attainable value (for example, in 2020, there was an average of 1.93 children under 18 per family in the United States)
- Expectation is a long-run average value of X
- Center of mass or center of gravity (balancing point) for the distribution

Example

- Consider a university having 15, 000 students and let X = the number of courses for which a randomly selected student is registered. The pmf is as follows:

x	1	2	3	4	5	6	7
$p(x)$	.01	.03	.13	.25	.39	.17	.02
Number registered	150	450	1950	3750	5850	2550	300

Example

- Since each of 150 students is taking one course, these 150 contribute 150 courses to the total. Similarly, 450 students contribute 2(450) courses, and so on. The population average value of *X* is then:

$$\frac{1(150) + 2(450) + 3(1950) + \cdots + 7(300)}{15{,}000} = 4.57$$

- Since 150/15,000 = 0.01 = p(1), 450/15,000 = 0.03 = p(2), and so on, we can calculate the average as follows:

$$1*p(1) + 2*p(2) + \ldots\ldots\ldots\ldots\ldots + 7*p(7)$$

Expected Value: Definition

- Suppose X is a discrete random variable whose probability model is given by

Value of X	x_1	x_2	……..	x_k
Probability	p_1	p_2	……..	p_k

- The expected value of X is given by

$$\mathrm{E(X)} = \mu_X = \sum_{x \in D} x \cdot p(x) = x_1 p_1 + x_2 p_2 + \cdots x_k p_k$$

Difference Between Expectation and Measure of Center

What's the difference between $E(X)$ (expectation/population mean) and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (sample mean)?

Actually, they live in different worlds.

- Expected value of a RV X is the *true* mean in the ideal world, while mean of a sample $x_1, x_2, x_3, \ldots$ is the *observed* mean in the empirical world.

- .In other words, expectation is a concept in probability, and sample mean is a concept in statistics. In inference, we use $\bar{x}$ to infer $E(X)$

Expected Value is the Population Mean

Example

- Population values: 1, 1, 2, 2, 3, 3, 4, 4
- E(X) = Mean = (1+1+2+2+3+3+4+4)/8 = 20/8 = 2.5
- E(X) = 1*(0.25) + 2*(0.25) + 3*(0.25) + 4*(0.25) = 2.5

Example: Categorical Data: Smoke: Yes(1), No(0)

Population values: 1, 0, 0, 1, 1, 1, 1, 0,0, 1

E(X) = Mean = (1+0+0+1+1+1+1+0+0+1)/10 = 6/10 = 0.6

E(X) = 1*(0.6) + 0*(0.4) = 0.6

Expected Value of a random variable

- The ***Expectation*** or ***Expected Value*** of a random variable X is defined to be the sum of all the products $(\boldsymbol{x} \times \boldsymbol{f}(\boldsymbol{x}))$ over all possible values x of the random variable X: that is, the expectation of a random variable is a ***weighted average*** of all the possible values that the random variable can take, weighted by the probability of each value.

- $\boldsymbol{E(X) = \sum x \cdot P(X = x) = \sum x \cdot f(x)}$

- **For example**, toss a coin 3 times, let X = # of heads. Write down $f(x)$, and then use the formula to compute the expectation of X.

Hypertension Example

Hypertension Many new drugs have been introduced in the past several decades to bring hypertension under control—that is, to reduce high blood pressure to normotensive levels. Suppose a physician agrees to use a new antihypertensive drug on a trial basis on the first four untreated hypertensives she encounters in her practice, before deciding whether to adopt the drug for routine use. Let X = the number of patients of four who are brought under control. Then X is a discrete random variable, which takes on the values 0, 1, 2, 3, 4.

$Pr(X = r)$	.008	.076	.265	.411	.240
r	0	1	2	3	4

$$E(X) = 0(.008) + 1(.076) + 2(.265) + 3(.411) + 4(.240) = 2.80$$

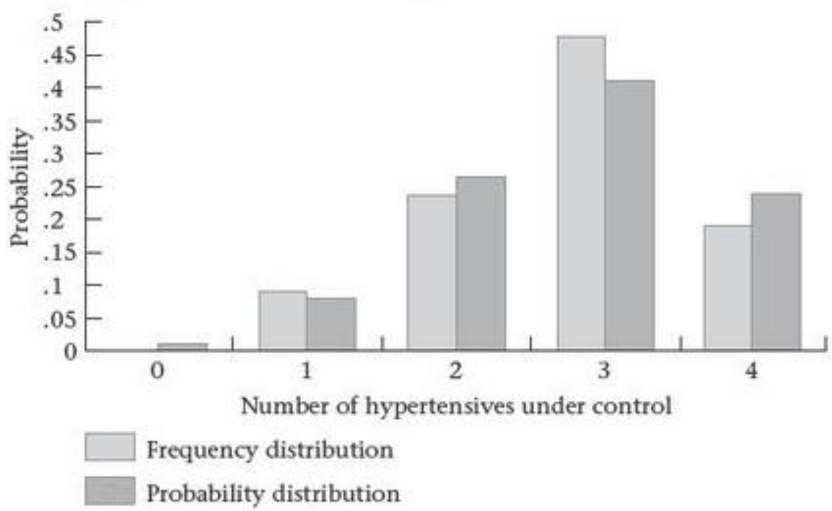
Thus on average about 2.8 hypertensives would be expected to be brought under control for every 4 who are treated.

Hypertension Example

Comparison of the sample-frequency distribution and the theoretical-probability distribution for the hypertension-control example

Number of hypertensives under control = r	Probability distribution $Pr(X = r)$	Frequency distribution
0	.008	.000 = 0/100
1	.076	.090 = 9/100
2	.265	.240 = 24/100
3	.411	.480 = 48/100
4	.240	.190 = 19/100

Comparison of the frequency and probability distribution for the hypertension-control example



$$\bar{x} = [0(0) + 1(9) + 2(24) + 3(48) + 4(19)]/100 = 2.77$$

hypertensives controlled per 4-patient clinical practice, while $\mu = 2.80$.

$$\bar{x} = 0(0/100) + 1(9/100) + 2(24/100) + 3(48/100) + 4(19/100)$$

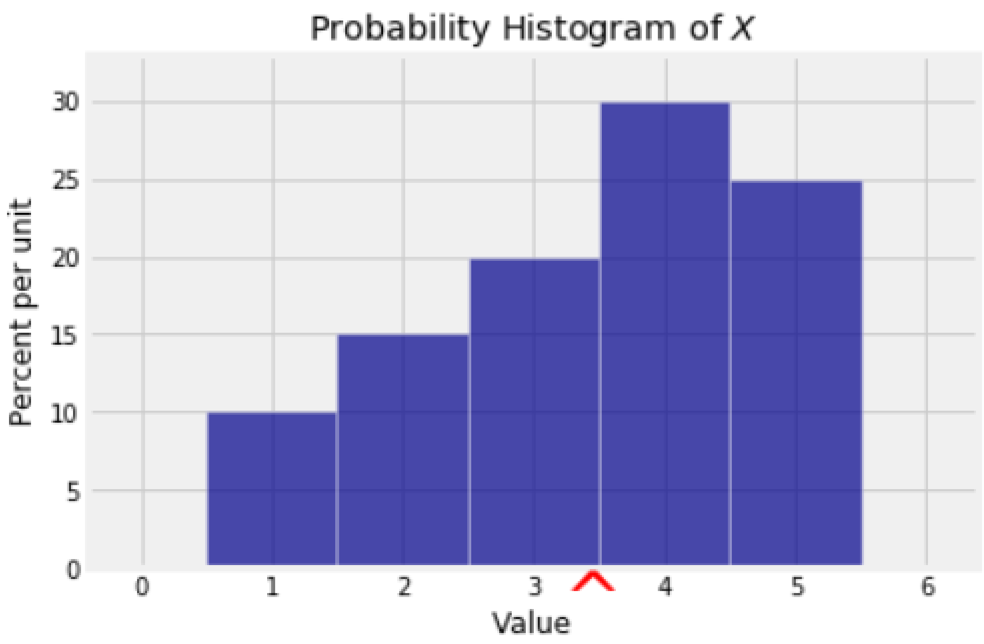
Expectation: Long run average

For example, suppose X has the following distribution table.

k	1	2	3	4	5
$P(X=k)$	0.1	0.15	0.2	0.3	0.25

Then

$$E(X) \ = \ 1(0.1) + 2(0.15) + 3(0.2) + 4(0.3) + 5(0.25) \ = \ 3.45$$



The Expected Value of a Function

- The cost of a certain vehicle diagnostic test depends on the number of cylinders *X* in the vehicle's engine. Suppose the cost function is given by $h(X) = 20 + 3X + .5X^2$

- Since *X* is a random variable, so is $Y = h(X)$. The pmf of *X* and derived pmf of *Y* are as follows:

x	4	6	8
$p(x)$	.5	.3	.2

$\Rightarrow$

y	40	56	76
$p(y)$	.5	.3	.2

The Expected Value of a Function

With D^* denoting possible values of Y,

$$E(Y) = E[h(X)] = \sum_{D^*} y \cdot p(y)$$
$$= (40)(.5) + (56)(.3) + (76)(.2)$$
$$= h(4) \cdot (.5) + h(6) \cdot (.3) + h(8) \cdot (.2)$$
$$= \sum_{D} h(x) \cdot p(x)$$

- It was not necessary to determine the pmf of *Y* to obtain *E*(*Y*); instead, the desired expected value is a weighted average of the possible *h*(*x*) (rather than *x*) values.

The Expected Value of a Function

- A bookstore purchases ten copies of a book at \$60 each to sell at \$12, and any unsold copies after three months can be redeemed for \$2. If the number of copies sold is X, what is the profit of the bookstore?

- The profit is $h(X) = 12X + 2(10 – X) – 60 = 10X − 40$.
- What is the expectation of h(X)?

- If Random Variable X has range D and pmf p(x), then the expected value of function h(X) is given by

$$E(h(X)) = \sum_{x \in D} (h(x) \cdot p(x))$$

The Expected Value of a Function

- The Linear Function Case

- In the case of linear function, we have a much more convenient formula

$$E(aX + b) = a \cdot E(X) + b$$

Bookstore Example:

$$h(X) = 10X - 40$$

$$E(h(X)) = E(10X - 40) = 10E(X) - 40$$