STAT115: Introduction to Biostatistics

University of Otago

Ōtākou Whakaihu Waka

Lecture 8: Random Variables

Outline

- Data summaries: sample mean and standard deviation
- Summaries are limited
  - To go further we needed statistical models
    - Use probability to describe the variation in the data
- Had an introduction to probability
- Today we will introduce idea of a random variable
  - Useful in helping us use probability to describe data

Example: bovine leptospirosis

- An inspector visits cattle & dairy farms for signs of bovine leptospirosis
- If they visit three farms, the sample space has eight possible outcomes
  - LLL, LLC, LCL, LCC, CLL, CLC, CCL, CCC
    - L: evidence of leptospira at farm
    - C: farm is clear
  - Each outcome has an associated probability
- If the inspector visits 30 farms, there are 1 073 741 824 possible outcomes
- The way the problem is expressed makes it difficult to answer questions:
  - How many farms would we expect to have evidence of leptospira?
  - How likely is it that 24 or more farms will have evidence of leptospira?
- We need a better way of writing/expressing things

Random variable

- A random variable assigns a numerical value to each outcome in sample space
- For our purposes, we can use a simpler definition:
  - A random variable is a (random) process with a numerical outcome
- Common to represent a random variable with capital letter
  - e.g. X or Y or Z
- The possible values are given with lowercase letters
  - e.g. x, y, z

Random variables: leptospirosis example

- Y represents the number of farms with evidence of leptospira
- Visit three farms
  - Four possible values: $y_1 = 0$, $y_2 = 1$, $y_3 = 2$, $y_4 = 3$
- Visit 30 farms
  - 31 possible values: $y_1 = 0$, $y_2 = 1$, $\ldots$, $y_{31} = 30$.
- We may use i (or j) as an index of possible values
  - e.g. $i = 2$ is the second possible value; $y_i = y_2 = 1$
- We use the k to represent the number of possible values
  - $k = 4$ if we visit three farms
  - $k = 31$ if we visit 30 farms

Probability distribution

- A random variable has an associated probability distribution
- For the leptospirosis example

i	1	2	3	4	Total
y_i	0	1	2	3	
$\Pr(Y = y_i)$	0.25	0.15	0.4	0.2	1

- $\Pr(Y = y_i)$: the probability that (the random variable) Y takes the value y_i
  - e.g. for $i = 3$: $\Pr(Y = 2) = 0.4$, the probability that Y takes the value 2

Probability distribution: example

- Suppose we open an online store that sells two products
- A given online visitor may:
  - With probability 0.4 buy nothing: we receive \$0
  - With probability 0.3 buy item A: we receive \$20
  - With probability 0.2 buy item B: we receive \$35
  - With probability 0.1 buy item A and B: we receive \$50
- If Y represents the money we receive from an online visitor

i	1	2	3	4	Total
y_i	0	20	35	50	
$\Pr(Y = y_i)$	0.4	0.3	0.2	0.1	1

Using probability distributions

- With these definitions we can start to ask useful questions
  - How likely is it that 2 or more farms will have evidence of leptospira?
  - How likely is it that we will receive $20 or below from an online visitor?

Using probability distributions

- With these definitions we can start to ask useful questions
  - How likely is it that 2 or more farms will have evidence of leptospira?
  - How likely is it that we will receive \$20 or below from an online visitor?
- We use results from last week to answer those questions
- Using the online store as an example
  - Think of the y values as events: $y_1 = 0$, $y_2 = 20$, $y_3 = 35$, $y_4 = 50$
  - The events are mutually exclusive
  - $\Pr(Y \leq 20) = \Pr(Y = 0 \text{ or } Y = 20) = \Pr(Y = 0) + \Pr(Y = 20) = 0.4 + 0.3 = 0.7$

Expectation

- We can't yet answer the other question from earlier
  - How many farms would we expect to have evidence of leptospira? or
  - How much money do we expect to receive from an online visitor?
- We want to find $E[Y]$, the expected value of the random variable Y
  - The expected value is the same as the mean and is often represented by μ
- To find this, we weight each possible value by its corresponding probability

$$E[Y] = \sum_{i=1}^{k} y_i \Pr(Y = y_i)$$

- k is the number of possible values (in both our examples $k = 4$)
  - $E[Y] = y_1 \Pr(Y = y_1) + y_2 \Pr(Y = y_2) + y_3 \Pr(Y = y_3) + y_4 \Pr(Y = y_4)$

Expectation: leptospirosis example

- How many farms would we expect to have evidence of leptospira?

i	1	2	3	4	Total
y_i	0	1	2	3	
$\Pr(Y = y_i)$	0.25	0.15	0.4	0.2	1

$$E[Y] = \underbrace{0 \times 0.25}_{0} + \underbrace{1 \times 0.15}_{0.15} + \underbrace{2 \times 0.4}_{0.8} + \underbrace{3 \times 0.2}_{0.6}$$

$$= 1.55$$

- The expected (mean) number of farms with evidence of leptospira infection is 1.55

Expectation: online store

- How much money do we expect to receive from an online visitor?

i	1	2	3	4	Total
y_i	0	20	35	50	
$\Pr(Y = y_i)$	0.4	0.3	0.2	0.1	1

$$\begin{aligned} E[Y] &= \underbrace{0 \times 0.4}_{0} + \underbrace{20 \times 0.3}_{6} + \underbrace{35 \times 0.2}_{7} + \underbrace{50 \times 0.1}_{5} \\ &= 18 \end{aligned}$$

- We expect to receive \$18 from an online visitor

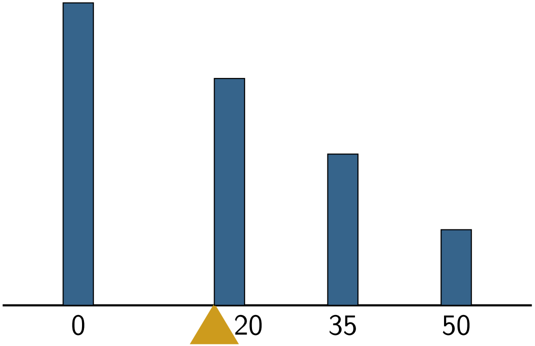
Expection: intuition

i	1	2	3	4	Total
y_i	0	20	35	50	
$\Pr(Y = y_i)$	0.4	0.3	0.2	0.1	1

- If we saw 100 online visitors
  - We would expect 40 of them to spend nothing: receive \$0
  - We would expect 30 of them to spend \$20: receive \$600
  - We would expect 20 of them to spend \$35: receive \$700
  - We would expect 10 of them to spend \$50: receive \$500
- We would expect to receive \$1800 per 100 visitors = \$18 per visitor
- Multiplying y_i by $\Pr(Y = y_i)$ is taking a ‘direct route’ to this answer

Expectation: intuition

- Another way we can look at expectation is by thinking of the probability distribution as a old-fashioned scale
- The expected value balances the probability distribution (gold triangle)



Variance

- We could also ask questions that relate to variability
  - How much would we expect income from our store to vary from one day to the next?
- For small problems (like those we have been looking at)
  - Probably preferable to base this off the probability distribution
- For larger problems (which we are moving toward)
  - We need a measure of variability
  - Typically use variance / standard deviation

Variance

- The variance of the random variable Y is $\text{Var}(Y)$
  - Find the average of squared deviations from the mean
  - Weight the squared deviations by their probability

$$\text{Var}(Y) = \sum_{i=1}^{k} (y_i - E[Y])^2 \Pr(Y = y_i)$$

- For $k = 4$
  - $\text{Var}(Y) = (y_1 - E[Y])^2 \Pr(Y = y_1) + (y_2 - E[Y])^2 \Pr(Y = y_2) + (y_3 - E[Y])^2 \Pr(Y = y_3) + (y_4 - E[Y])^2 \Pr(Y = y_4)$

Variance: leptospirosis example

- What is the variance in the number of farms that have evidence of leptospira?
  - We know $E[Y] = 1.55$

i	1	2	3	4	Total
y_i	0	1	2	3	
$\Pr(Y = y_i)$	0.25	0.15	0.4	0.2	1

$$\text{Var}(Y) = \underbrace{(0-1.55)^2 \times 0.25}_{2.4025 \times 0.25} + \underbrace{(1-1.55)^2 \times 0.15}_{0.3025 \times 0.15} + \underbrace{(2-1.55)^2 \times 0.4}_{0.2025 \times 0.4} + \underbrace{(3-1.55)^2 \times 0.2}_{2.1025 \times 0.2}$$

$$= 1.1475$$

Standard deviation

- The standard deviation is the square root of variance
  - $\mathsf{sd}(Y) = \sqrt{\mathsf{Var}(Y)}$
- For the leptospirosis example
  - $\mathsf{sd}(Y) = \sqrt{1.1475} = 1.07$
- The standard deviation is (approximately) the average deviation from the mean
- Often the variance will be represented by σ^2
  - The standard deviation as σ

Example: online visitors

- What is the variance in the amount we receive from an online visitor?
  - We know $E[Y] = 18$

i	1	2	3	4	Total
y_i	0	20	35	50	
$\Pr(Y = y_i)$	0.4	0.3	0.2	0.1	1

$$\text{Var}(Y) = \underbrace{(0-18)^2 \times 0.4}_{324 \times 0.4 \quad +} + \underbrace{(20-18)^2 \times 0.3}_{4 \times 0.3 \quad +} + \underbrace{(35-18)^2 \times 0.2}_{289 \times 0.2 \quad +} + \underbrace{(50-18)^2 \times 0.1}_{1024 \times 0.1}$$

$$= 291$$

$$\text{sd}(Y) = 17.1$$

We've seen this before

- We saw expectation (mean), standard deviation, and variance in Week 1
    - Sample mean, sample variance, sample standard deviation
    - These are summaries of a particular data set (a sample)
- Today we've found these quantities for a distribution
    - Summaries of a random variable
    - Tells us something about what realizations from the distribution should look like

Summary

- Introduced random variables
- Probability distribution of random variable
- Saw several summaries of random variables
  - Mean
  - Variance
  - Standard deviation