

Statistics 571

Statistical Methods for Bioscience I

Section I: Bret Larget
Section II: Jun Zhu
Department of Statistics
University of Wisconsin–Madison

Fall 2005

Course Information

- Read the entire syllabus carefully. Complete the survey sheet.
- Before the lectures, download the lecture notes and hand-outs from the course website. You are advised to attend the lecture section you are enrolled in, even though the two lecture sections are similar.
- Late homework without Instructor’s prior permission will be penalized.
- You are advised to attend the discussion section that you are enrolled in. If you wish to attend a different section that is closed, you need to obtain TA’s permission to attend that section.
- Block the dates and time for the exams **NOW**.
- Read “Introduction to R” on the course website.
- Policy on Auditors: Attend at least 80% of the lectures.
- Where to get help beyond the lectures?
 - Reading materials.
 - Course website, discussion sections, office hours, etc.

Introduction to Statistics

What is statistics?

- Statistics is a branch of scientific inquiry that helps to determine cause and association, and to make predictions.
- It provides methods to organize and summarize data from sample (i.e. any subset of a population).
- It also provides methods to use information in the data to draw various conclusions about a population (i.e. all individuals or objects of a particular type).

Population vs. sample

- A plate vs. a spoonful of chef's special tonight.
- A book vs. a few pages of the book.
- All vs. 100 plants in a field.

Introduction to Statistics

Probability vs. statistics

- Probability is the mathematics of chance and randomness. In probability, properties of a population are assumed known and questions about a sample are posed and answered. Thus the approach here is deductive.
- In statistics, properties of a sample are available and conclusions about a population is drawn based on the sample. Thus the approach here is inductive.

Three main topics

- Descriptive statistics: display and summarize data in a sample.
- Probability: Given a population, study the uncertainty associated with a sample taken from the population.
- Statistics: Given a sample, learn methods to draw conclusions about a population, while taking into account of uncertainties in the sample.

Descriptive Statistics

Example: height of seedlings

Thirteen (13) red pine seedlings were sampled from a nursery in Wisconsin. The heights of these seedlings were (in cm):

42 23 43 34 49 56 31 47 61 54 46 34 26

- Graphical methods describe data by visual/graphical techniques.
 - Stem-and-leaf plot, dot plot
 - Histogram
- Numerical methods extract summarizing numbers that characterize the data set and reveal main features.
 - Measures of location/center:
 - * Sample mean
 - * Sample median
 - * Sample quantiles, box plot
 - Measures of spread:
 - * Sample range
 - * Interquartile range (IQR)
 - * Sample variance, standard deviation

Descriptive Statistics

Stem-and-leaf plot

- Select leading digits for stem values and trailing digits for leaves.
- List all possible stems:

For the height of seedlings, 42 23 43 34 49 56 31 47 61 54 46 34 26

2 |
3 |
4 |
5 |
6 |

- Record leaf for each observation (or obs), besides the corresponding stem value:

2 | 36
3 | 144
4 | 23679
5 | 46
6 | 1

- Indicate the units for stems and leaves.

stem = 10 cm; leaf = 1 cm

Descriptive Statistics

Remarks

- R output:

The decimal point is 1 digit(s) to the right of the |

```
2 | 36
3 | 144
4 | 23679
5 | 46
6 | 1
```

- An alternative is a dot plot.
- Stem-and-leaf plots and dot plots have information about the shape, center, spread of the data distribution, as well as outliers and all the observations.

Descriptive Statistics

Histogram

- Divide data into non-overlapping classes.
- Decide the number of obs (i.e. frequencies) in each class (i.e. tally).
- Draw rectangles with height = frequencies and base = class intervals.
- For the height of seedlings,

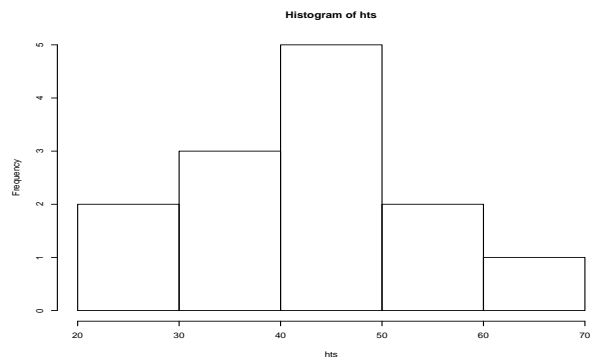
class	frequencies
19.5-29.5	2
29.5-39.5	3
39.5-49.5	5
49.5-59.5	2
59.5-69.5	1

Descriptive Statistics

Remarks

- Histogram is a pictorial representation of the data frequency distribution.
- Note the boundary values for the class intervals.
- Histograms have information about the shape, center, spread of the data distribution.
- Comparison of graphical methods:

methods	advantages	disadvantages
stem-and-leaf	quick and easy; all data values	awkward with large/disparate data
dot plot	compact	same as stem-and-leaf
histogram	general shape	art with deciding classes



Descriptive Statistics

Sample mean

- Given a data set of y_1, y_2, \dots, y_n from a population, sample mean provides a measure of location/center of the data set.
- To compute the sample mean:
 - add all the values $\sum_{i=1}^n y_i = y_1 + y_2 + \dots + y_n$
 - divide by the number of observations n
 - formula for sample mean is

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$$

- For the height of seedlings,

$$y_1 = 42, y_2 = 23, y_3 = 43, \dots, y_{13} = 26$$

and thus

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n} = \frac{546}{13} = 42.$$

- Note that \bar{y} is the balance point of the histogram (or dot plot).
- Sometimes $\sum_{i=1}^n y_i$ is abbreviated as $\sum y_i$.

Descriptive Statistics

Sample median

- To compute the sample median:
 - arrange the data in a list of ascending order
 - take the middle value in the list
 - formula for sample median is

$$\tilde{y} = \begin{cases} \left(\frac{n+1}{2}\right)^{th} \text{ value} & ; n \text{ is odd} \\ \text{avg of } \left(\frac{n}{2}\right)^{th} \text{ and } \left(\frac{n+1}{2}\right)^{th} \text{ value} & ; n \text{ is even} \end{cases}$$

- For the height of seedlings, $\tilde{y} = 43$.

2 | 36
3 | 144
4 | 23679
5 | 46
6 | 1

- If n is even, then the sample median is the average of the middle two values. For example, suppose the ordered data values are 2, 4, 5, 7, 8, 14 and the sample median is $\frac{5+7}{2} = 6$.

Descriptive Statistics

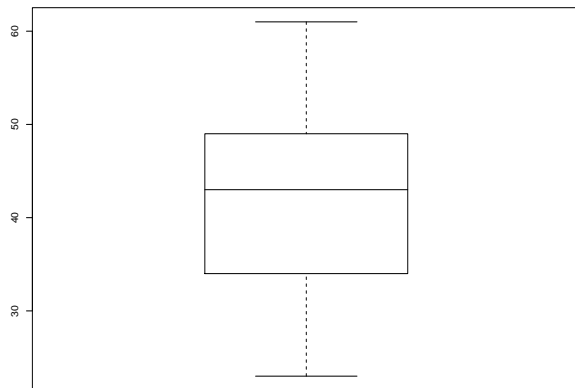
Sample quantiles

- Sample median is a special example of a sample quantile (or percentile).
- Denote the p^{th} sample quantile as $y_{[p]}$ with $0 \leq p \leq 1$.
- To compute the p^{th} sample quantile:
 - arrange data in a list of ascending order
 - compute $n \times p$
 - If $n \times p$ is an integer, then $y_{[p]}$ is the average of $(n \times p)^{th}$ and $(n \times p + 1)^{th}$ data values in the list.
 - If $n \times p$ is not an integer, then round up to $[n \times p]$ and use the $[n \times p]^{th}$ data value in the list.
- For the height of seedlings, suppose $p = 0.20$. Then $y_{[0.20]}$ is the 0.20^{th} sample quantile, or the 20^{th} percentile. Since $n \times p = 13 \times 0.20 = 2.6$, round up to $[n \times p] = 3$ and $y_{[0.20]}$ is the third data value in the list, which is 31.
- If $n \times p$ is an integer, say 2 (for a different data set with $n = 10$ and $p = 0.20$), then take the average of the second and the third data values.

Descriptive Statistics

Box plot

- $y_{[0.50]}$ is the sample median; $y_{[0.25]}$ is called the first quartile; and $y_{[0.75]}$ is called the third quartile.
- For the height of seedlings, $y_{[0.25]} = 34$ (the 4th obs) and $y_{[0.75]} = 49$ (the 10th obs).
- A box plot displays several quantiles simultaneously.



Descriptive Statistics

Sample range

- Sample range is the difference between the largest obs and the smallest obs, which is a measure of spread/variability of the data set.
- For the height of seedlings, the sample range is $61 - 23 = 38$ (see box plot).

Interquartile range (IQR)

- Interquartile range (IQR) is the difference between the third quartile and the first quartile (i.e. $y_{[0.75]} - y_{[0.25]}$).
- For the height of seedlings, the IQR is $49 - 34 = 15$ (see box plot).

Descriptive Statistics

Sample variance

- Sample variance is denoted by s^2 with

$$s^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - 1}$$

- Sample variance measures the average squared deviation.
- For the height of seedlings, $\bar{y} = 42$, $y_1 = 42$, $y_2 = 23$, $y_3 = 43$, \dots , $y_{13} = 26$,

$$s^2 = \frac{(42 - 42)^2 + (23 - 42)^2 + (43 - 42)^2 + \dots + (26 - 42)^2}{12} = 138.17$$

- Why dividing by $n - 1$ but not n ?
- For hand calculation, use working formulas

$$s^2 = \frac{1}{n - 1} \left[\sum_{i=1}^n y_i^2 - \frac{(\sum_{i=1}^n y_i)^2}{n} \right]$$

or

$$s^2 = \frac{1}{n - 1} \left[\sum_{i=1}^n y_i^2 - n(\bar{y})^2 \right]$$

Descriptive Statistics

Sample standard deviation

- Sample standard deviation (SD) is the square root of sample variance

$$s = \sqrt{s^2}$$

- For the height of seedlings, sample standard deviation is

$$s = \sqrt{138.17} = 11.75$$

- Sample standard deviation is a typical deviation, as $\pm 1s$ captures about 2/3 of bell-shaped data.
- Coefficient of variation (CV) is the ratio of standard deviation and mean

$$c = \frac{s}{\bar{y}}$$

which is a measure of relative spread of data.

- For the height of seedlings, sample coefficient of variation is

$$c = \frac{11.75}{42} = 0.28.$$

Descriptive Statistics

A quick summary

- Sample mean and standard deviation
- Sample median and IQR
- Suppose data values are

2 4 6 7 8 10 12

Then

$$\bar{y} = 7, s = 3.42, \tilde{y} = 7, \text{IQR} = 6$$

- Suppose data values are

2 4 6 7 8 10 102

Then

$$\bar{y} = 19.9, s = 36.32, \tilde{y} = 7, \text{IQR} = 6$$

Descriptive Statistics

Key R commands

```
> # enter data
> hts = c(42, 23, 43, 34, 49, 56, 31, 47, 61, 54, 46, 34, 26)
> hts
[1] 42 23 43 34 49 56 31 47 61 54 46 34 26

> # sample size
> length(hts)
[1] 13

> # stem-and-leaf plot
> stem(hts)

The decimal point is 1 digit(s) to the right of the |

 2 | 36
 3 | 144
 4 | 23679
 5 | 46
 6 | 1

> # histogram plot
> hist(hts)

> # sample mean
> mean(hts)
[1] 42

> # sample variance
> var(hts)
[1] 138.1667

> # sample standard deviation
> sd(hts)
[1] 11.75443

> # coefficient of variation
> sd(hts)/mean(hts)
```

```
[1] 0.2798674
```

```
> # more summary stats  
> summary(hts)
```

```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   
    23     34     43     42     49     61
```

```
> # IQR  
> IQR(hts)
```

```
[1] 15
```

```
> # box plot  
> boxplot(hts)
```

Descriptive Statistics

Example 1: weight of soil

The data represent the actual weight of 15 2-lb. bags of soil used for a lab experiment.

```
2.36 2.27 2.42 2.13 2.19 2.33 2.54 2.21 2.06 2.36  
2.51 2.45 2.12 2.32 2.29
```

The decimal point is 1 digit(s) to the left of the |

```
20 | 6  
21 | 239  
22 | 179  
23 | 2366  
24 | 25  
25 | 14
```

Descriptive Statistics

Example 1: weight of soil

- sample mean $\bar{y} = 2.30$
- sample median $\tilde{y} = 2.32$ (i.e. $y_{[0.50]} = 2.32$)
- sample variance $s^2 = 0.0205$
- sample standard deviation $s = 0.14$
- first quartile $y_{[0.25]} = 2.19$
- third quartile $y_{[0.75]} = 2.42$
- IQR = $2.42 - 2.19 = 0.23$

Descriptive Statistics

Example 2: weight of soil

13		5
14		
15		
16		1
17		9
18		2
19		
20		7
21		0
22		35
23		
24		6
25		1
26		04
27		9
28		
29		
30		2
31		8

Descriptive Statistics

Example 2: weight of soil

- sample mean $\bar{y} = 2.30$
- sample median $\tilde{y} = 2.25$ (i.e. $y_{[0.50]} = 2.25$)
- sample variance $s^2 = 0.27$
- sample standard deviation $s = 0.52$
- first quartile $y_{[0.25]} = 1.82$
- third quartile $y_{[0.75]} = 2.64$
- IQR = $2.64 - 1.82 = 0.82$
- Compared to example 1, the mean is the same but the spread is larger.

Descriptive Statistics

Example 3: weight of soil

19 | 3
20 | 5
21 | 8
22 | 47
23 | 344789
24 | 1124

- sample mean $\bar{y} = 2.30$
- sample median $\tilde{y} = 2.34$ (i.e. $y_{[0.50]} = 2.34$)
- sample variance $s^2 = 0.021$
- sample standard deviation $s = 0.15$
- first quartile $y_{[0.25]} = 2.24$
- third quartile $y_{[0.75]} = 2.41$
- IQR = $2.41 - 2.24 = 0.17$
- Compared to example 1, the mean and spread are similar, but the data distribution is skewed.

Descriptive Statistics

Example 4: weight of soil

20 | 69
21 | 35779
22 | 9
23 |
24 | 01358
25 | 15

- sample mean $\bar{y} = 2.30$
- sample median $\tilde{y} = 2.29$ (i.e. $y_{[0.50]} = 2.29$)
- sample variance $s^2 = 0.029$
- sample standard deviation $s = 0.17$
- first quartile $y_{[0.25]} = 2.15$
- third quartile $y_{[0.75]} = 2.45$
- IQR = $2.45 - 2.15 = 0.30$
- Compared to example 1, the mean and spread are similar, but the data distribution is bimodal.
- Examples 1–4 suggest the need to look at data.

Descriptive Statistics

Basic concepts

- A *population* is a whole set/group of individuals/objects which we are interested in studying. It can be thought of as the complete set of possible observations (finite or infinite).
- A *sample* is a (small) part of the population which we actually observe in order to learn about the population. Oftentimes we try to draw samples that are representative of the population.
- For example,
 - Weight of soil: what is the actual average weight of a 2-lb. bag?
 - Quality of goods: defective rates?
 - Voting populations: what percentage favors each candidate?
- *Data* are observations in the sample we use to learn about the population.
- Statistical problems involve the collection, description, analysis, and interpretation of data.

Descriptive Statistics

Types of data

- There are two broad classes of data: quantitative (i.e. numerical) and qualitative (i.e. categorical) data.
- For *quantitative data*, each observation has a number associated with it.
- For example, weight, milk yield, or # of cows on a farm.
- Quantitative data can be either continuous or discrete.
- In the example, weight and milk yield are continuous data and # of cows on a farm are discrete data.
- For *qualitative data*, each observation can be put into a category, which is either nominal or ordered.
- For example, 15 cows are assigned to 3 types of beds (A, B, C) or 3 different diet types (I,II,III):

bed types	# of cows	diet types	# of cows
A:Hay	10	I: high in VC	5
B:Cement	6	II: low in VC	5
C:Others	4	III: control	5

Descriptive Statistics

Key R commands

```
# example 1
> wts1 = c(2.36, 2.27, 2.42, 2.13, 2.19, 2.33, 2.54, 2.21, 2.06, 2.36,
+ 2.51, 2.45, 2.12, 2.32, 2.29)
> wts1
[1] 2.36 2.27 2.42 2.13 2.19 2.33 2.54 2.21 2.06 2.36 2.51 2.45 2.12 2.32 2.29
> stem(wts1)

The decimal point is 1 digit(s) to the left of the |

20 | 6
21 | 239
22 | 179
23 | 2366
24 | 25
25 | 14

> var(wts1)
[1] 0.02049714
> sd(wts1)
[1] 0.1431682
> summary(wts1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 2.060  2.200  2.320  2.304  2.390  2.540
>
> # save results
> wts1.mean = mean(wts1)
> wts1.mean
[1] 2.304
> wts1.sd = sd(wts1)
> wts1.sd
[1] 0.1431682
> wts1.cv = wts1.sd/wts1.mean
> wts1.cv
[1] 0.06213899
>
> # example 2
> wts2 = c(1.35, 1.61, 1.79, 1.82, 2.07, 2.10, 2.23, 2.25, 2.46, 2.51,
```

```

+ 2.60, 2.64, 2.79, 3.02, 3.18)
> stem(wts2)

The decimal point is at the |

1 | 4
1 | 688
2 | 1123
2 | 55668
3 | 02

> stem(wts2, scale=4)

The decimal point is 1 digit(s) to the left of the |

13 | 5
14 |
15 |
16 | 1
17 | 9
18 | 2
19 |
20 | 7
21 | 0
22 | 35
23 |
24 | 6
25 | 1
26 | 04
27 | 9
28 |
29 |
30 | 2
31 | 8

> var(wts2)
[1] 0.2697981
> sd(wts2)
[1] 0.5194209
> summary(wts2)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 1.350  1.945   2.250   2.295  2.620   3.180

```

```

>
> # example 3
> wts3 = c(1.93, 2.05, 2.18, 2.24, 2.27, 2.33, 2.34, 2.34, 2.37, 2.38,
+ 2.39, 2.41, 2.41, 2.42, 2.44)
> stem(wts3)

The decimal point is 1 digit(s) to the left of the |

18 | 3
20 | 58
22 | 47344789
24 | 1124

> stem(wts3, scale=2)

The decimal point is 1 digit(s) to the left of the |

19 | 3
20 | 5
21 | 8
22 | 47
23 | 344789
24 | 1124

> var(wts3)
[1] 0.02142857
> sd(wts3)
[1] 0.146385
> summary(wts3)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 1.930  2.255   2.340   2.300  2.400   2.440
>
> # example 4
> wts4 = c(2.06, 2.09, 2.13, 2.15, 2.17, 2.17, 2.19, 2.29, 2.40, 2.41,
+ 2.43, 2.45, 2.48, 2.51, 2.55)
> stem(wts4)

The decimal point is 1 digit(s) to the left of the |

20 | 69
21 | 35779
22 | 9

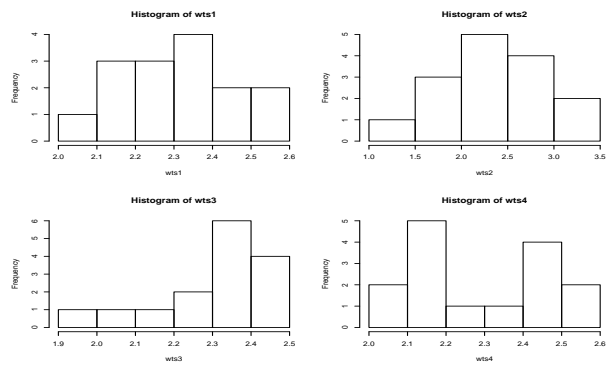
```

```

23 |
24 | 01358
25 | 15

> var(wts4)
[1] 0.02854095
> sd(wts4)
[1] 0.1689407
> summary(wts4)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.060  2.160  2.290  2.299  2.440  2.550
>
> par(mfrow=c(2,2))
> hist(wts1); hist(wts2); hist(wts3); hist(wts4)

```



Probability Model

Notion of chance

- Recall that in probability, properties of the population are assumed known and questions about a sample are posed and answered.
- Even though probability is a branch of mathematics that describes chance (or randomness), it is a language for statistics.
- *Long-run relative frequency* such as past flood records used to predict the chance of flood tomorrow.
- *Subjective notion* such as expert opinions used to state the reliability of a new GPS unit.
- *Classical notion* that is based on assumed symmetry in situations. For example, if a bag of 50 M&M chocolate contains 5 blue ones and 1 piece is picked at “random”, then the probability of it being blue is $\frac{5}{50} = 0.1 = 10\%$.
- Probability theory models/quantifies/describes any of these notions of chance.
- Our focus is on the classical notion of probability.

Probability Model

Elements in a probability model

- An *experiment* is an action/process that generates data. For example, recording the weight of a bag of soil, counting the number of cows, rolling a die, tossing a coin, etc. It usually has more than one possible outcomes and is theoretically repeatable.
- The individual possible outcomes of an experiment are called *elementary outcomes*.
- The entire group of elementary outcomes is called the *sample space* and is denoted as \mathcal{S} . For example, roll a die once and record the result. The sample space is $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$
- A collection of elementary outcomes is called an *event* or a subset of \mathcal{S} . For example, roll a die once and record the result. Possible events are:

$$A = \text{"the result is even"} = \{2, 4, 6\}$$

$$B = \text{"the result is not 4 nor 5"} = \{1, 2, 3, 6\}$$

$$C = \text{"the result is 3"} = \{3\}$$

$$D = \text{"the result is 3 or even"}$$

$$E = \{2, 6\}, F = \{1, 3, 5\}, \mathcal{S} = \{1, 2, 3, 4, 5, 6\}$$

$$\emptyset = \{ \} \text{ is an empty set.}$$

Probability Model

Operations of elements

- The event " U or V " consists of the elementary outcomes in U , in V , or in both. Also written as $U \cup V$ and read as " U union V ". For example,
 A or $C = \{2, 4, 6\}$ or $\{3\} = \{2, 3, 4, 6\} = D$
 A or $B = \{1, 2, 3, 4, 6\}$
- The event " U and V " consists of the elementary outcomes in both U and V . Also written as $U \cap V$ and read as " U intersect V ". For example,
 A and $B = \{2, 4, 6\}$ and $\{1, 2, 3, 6\} = \{2, 6\} = E$
 B and $C = \{3\} = C$
 A and $F = \emptyset$
- The event "not U " consists of all elementary outcomes in \mathcal{S} that are not in U . Also written as \bar{U} and read as "the complement of U ". For example,
not $A = \text{not } \{2, 4, 6\} = \{1, 3, 5\} = F$

Probability Model

Operations of elements

- Note the following:
 $B \cup E = B$ and $B \cap E = E$, as E is contained in B

$$D \cup F = \mathcal{S}$$

$A \cap C = \emptyset$, as A and C have nothing in common.

- Two events are *mutually exclusive*, if they do not have any elementary outcomes in common. For example, $A = \{2, 4, 6\}$ and $C = \{3\}$ are mutually exclusive

$D = \{2, 3, 4, 6\}$ and $F = \{1, 3, 5\}$ are not mutually exclusive.

Probability Model

Probability model

Probabilities are defined in terms of a model. A *probability model* consists of a probability assignment for each of the events in \mathcal{S} . There are two main issues:

- I. How to assign probability?
- II. Rules an assignment must follow.

Basic rules

A probability assignment must follow three basic rules to ensure that the assignment is consistent with our intuitive notion of chance and that the math is valid.

- (i). For any event U , $0 \leq P(U) \leq 1$
- (ii). $P(\mathcal{S}) = 1$
- (iii). If U and V are mutually exclusive, then

$$P(U \text{ or } V) = P(U) + P(V)$$

which is known as “the addition rule”.

Probability Model

Example: fair die rolled once

Roll a die once and record the result. Recall that $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$. If the die is fair, then each side has probability $1/6$. That is,

$$P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = \frac{1}{6}.$$

The assignment follows the three basic rules. Furthermore, we can compute the probability of an event.

$$\begin{aligned} P(E) &= P(\{2, 6\}) \\ &= P(2 \text{ or } 6) \\ &= P(2) + P(6) = 1/6 + 1/6 = 1/3 \end{aligned}$$

$$\begin{aligned} P(D) &= P(\{2, 3, 4, 6\}) \\ &= P(2 \text{ or } 3 \text{ or } 4 \text{ or } 6) \\ &= P(2) + P(3) + P(4) + P(6) = 2/3 \end{aligned}$$

$$\begin{aligned} P(C \text{ or } E) &= P(\{3\} \text{ or } \{2, 6\}) \\ &= P(C) + P(E) \\ &= 1/6 + 1/3 = 1/2 \end{aligned}$$

$$\begin{aligned} P(B \text{ or } D) &= P(\{1, 2, 3, 6\} \text{ or } \{2, 3, 4, 6\}) \\ &\neq P(B) + P(D) \\ &= 2/3 + 2/3 = 4/3. \end{aligned}$$

Probability Model

Derived rules

- In the example of rolling a fair die once, we cannot apply (iii) to compute $P(B \text{ or } D)$ because B and D are not mutually exclusive.
- One additional rule derived from (i)–(iii):

(iv). For any two events U, V ,

$$P(U \text{ or } V) = P(U) + P(V) - P(U \text{ and } V)$$

which is consistent with (iii).

- Thus in the example,

$$\begin{aligned} P(B \text{ or } D) &= P(B) + P(D) - P(B \text{ and } D) \\ &= P(\{1, 2, 3, 6\}) + P(\{2, 3, 4, 6\}) - P(\{2, 3, 6\}) \\ &= 2/3 + 2/3 - 1/2 = 5/6. \end{aligned}$$

- For any three events U, V, W :

$$\begin{aligned} P(U \cup V \cup W) &= P(U) + P(V) + P(W) \\ &\quad - P(U \cap V) - P(U \cap W) - P(U \cap W) \\ &\quad + P(U \cap V \cap W) \end{aligned}$$

Probability Model

Derived rules

- An event U and the event “not U ” are always mutually exclusive. By (iii),

$$P(U) + P(\text{not } U) = P(\mathcal{S}) = 1$$

- Thus the derived rule:

(v). For any event U ,

$$P(\text{not } U) = 1 - P(U)$$

- For example,

$$\begin{aligned} P(\text{not } E) &= P(\text{not } \{2, 6\}) \\ &= P(\{1, 3, 4, 5\}) \\ &= 2/3 \end{aligned}$$

or simply by (v),

$$\begin{aligned} P(\text{not } E) &= 1 - P(E) \\ &= 1 - P(\{2, 6\}) \\ &= 1 - 1/3 = 2/3 \end{aligned}$$

Probability Model

Conditional probability

- Again roll a fair die and we know that $P(2) = 1/6$. Suppose we were told that the result is an even number, what is the probability that it is a 2?

$$P(2 \text{ given it is an even number}) = 1/3$$

and is not $1/6$.

- Additional information can alter the probability of an event.
- In general, for two events U, V , denote the *conditional probability of U given V* as $P(U|V)$, which is the probability of the event U given (or knowing) that the event V has occurred. The exact definition is:

$$P(U|V) = \frac{P(U \text{ and } V)}{P(V)},$$

provided that $P(V) \neq 0$.

- In the example above,

$$P(2|\text{even}) = \frac{P(2 \text{ and even})}{P(\text{even})} = \frac{P(2)}{P(\{2, 4, 6\})} = \frac{1/6}{3/6} = 1/3.$$

Probability Model

Independence

- There are 6 cards in a box, each with a number ranging from 1 to 3 and a color either red or green.

card	#1	#2	#3	#4	#5	#6
feature	1R	2R	3R	1G	2G	3G

Given that a card randomly drawn from the box is red colored, what is probability that the card has a number 2? Note that $P(2|\text{red}) = 1/3$ and $P(2) = 1/3$.

- The extra knowledge that the card is red colored does not alter the probability, which leads to the notion of independence.
- For two events U, V , U and V are *independent* if

$$P(U|V) = P(U)$$

Probability Model

Independence

- Suppose U and V are independent, then

$$P(U) = P(U|V) = \frac{P(U \text{ and } V)}{P(V)}$$

- Thus if U and V are independent, then

$$P(U \text{ and } V) = P(U) \times P(V)$$

which is known as “the multiplication rule”.

- The multiplication rule can be used to check if two events are independent.
- For example, in the example of rolling a fair die once, $D = \{2, 3, 4, 6\}$ and $E = \{2, 6\}$ are not independent, because

$$P(D \text{ and } E) = 1/3, P(D) = 2/3, P(E) = 1/3$$

and $P(D \text{ and } E) \neq P(D) \times P(E)$.

- But $A = \{2, 4, 6\}$ and $B = \{1, 2, 3, 6\}$ are independent, because $P(A \text{ and } B) = P(A) \times P(B)$, even though A and B are not mutually exclusive.

Probability Model

Independence

- The multiplication rule can also be used to find $P(U \text{ and } V)$. If U and V are independent, then $P(U \text{ and } V) = P(U) \times P(V)$. In general,

$$P(U \text{ and } V) = P(U|V) \times P(V) = P(U) \times P(V|U)$$

A quick summary

- Probability: notion of chance, probability model.
- Elements of probability model: experiment, elementary outcomes, sample space, events, operations of events (or, and, not), mutually exclusive events.
- Probability model: Three basic rules (i)–(iii), derived rules (iv) and (v), conditional probability, and independence.

Random Variables

Definition

- A random variable (r.v.) is a variable that depends on the outcome of a chance situation.
- A r.v. is often denoted by capital letters (e.g. Y).
- More rigorously, given a sample space \mathcal{S} of an experiment, a r.v. is a function (or rule) that assigns a *number* to each *elementary outcome* in the sample space \mathcal{S} .

Random Variables

Example: coin tossed three times

Toss a coin *independently* three times and record the number of times that the coin landed on heads. Let $Y = \#$ of heads. Then the sample space is:

1st toss	2nd toss	3rd toss	Y
H	H	H	3
H	H	T	2
H	T	H	2
H	T	T	1
T	H	H	2
T	H	T	1
T	T	H	1
T	T	T	0

Random Variables

Discrete r.v. and probability distribution

- Like for data, r.v.'s can be discrete or continuous, but cannot be categorical.
- A r.v. is *discrete* if there are either a finite number of possible values (e.g. y_1, \dots, y_n) or at most there is one for every integer (e.g. y_1, y_2, \dots).
- The *probability distribution of a discrete r.v.* is described by the probability of each possible value of the r.v.

Random Variables

Example: coin tossed three times

Continue with the example of tossing a coin three times. Suppose the coin is fair

$$P(H) = P(T) = 1/2$$

Then

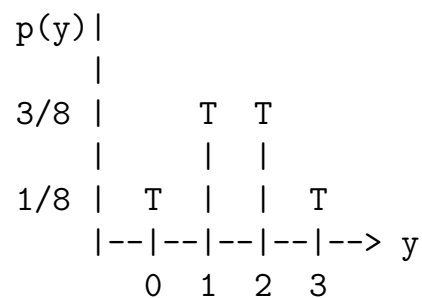
$$\begin{aligned} P(Y = 3) &= P(\text{HHH}) \\ &= P(H) \times P(H) \times P(H) \\ &= 1/2 \times 1/2 \times 1/2 = 1/8 \\ P(Y = 2) &= P(\text{HHT or HTH or THH}) \\ &= P(\text{HHT}) + P(\text{HTH}) + P(\text{THH}) \\ &= P(H)P(H)P(T) + P(H)P(T)P(H) \\ &\quad + P(T)P(H)P(H) \\ &= 1/8 + 1/8 + 1/8 = 3/8 \end{aligned}$$

Similarly $P(Y = 1) = 3/8, P(Y = 0) = 1/8$.

Random Variables

Example: coin tossed three times

- For notational convenience, define $p(y) = P(Y = y)$ (e.g. $p(2) = P(Y = 2)$).
- A line graph shows the probability distribution of Y .



- A frequency table also shows the probability distribution of Y .

y	$p(y)$
0	1/8
1	3/8
2	3/8
3	1/8

Random Variables

Summary measures

- The probability distribution of a discrete r.v. Y gives complete information about Y and hence complete information about the population.
- It is often helpful to have some numerical summary measures such as the center/location or spread/variability of the population (as we did with sample data).

	population (r.v.)	sample (observed data)
mean	μ_Y	\bar{y}
variance	σ_Y^2	s^2
standard deviation	σ_Y	s

Random Variables

Example: fair die game

A game involves rolling a fair die. If the result is 1, 2, or 3, then the player receives nothing \$0. If the result is 4, then the player receives \$9. If the result is 5 or 6, then the player receives \$15. Let the r.v. $Y =$ winning \$. Then the probability distribution of Y is

$$p(0) = 1/2, p(9) = 1/6, p(15) = 1/3.$$

Repeat the game 600 times. What will be the player's average winnings per game?

Roughly, 300 times the player will receive \$0; 100 times \$9; and 200 times \$15. Thus the total winning is about

$$30 \times 0 + 100 \times 9 + 200 \times 15 = 3900$$

and so on average the player will expect $3900/600 = 6.5$ dollars per game.

Random Variables

Expectation

- *Expectation* of a r.v. Y is the population mean of the probability distribution of Y .
- Expectation is denoted as $E(Y)$ or μ_Y .
- Expectation can be thought of as a typical value.
- For a discrete r.v. Y , the formula for expectation of Y is

$$E(Y) = \sum y \times p(y)$$

summing over all possible values y of the r.v. Y .

- In the fair die game,

$$E(Y) = 0 \times 1/2 + 9 \times 1/6 + 15 \times 1/3 = 6.50.$$

Random Variables

Properties of expectation

For a r.v. Y ,

- If $Z = Y + c$ where c is fixed constant, then $E(Z) = E(Y) + c$.
- If $Z = 2Y$, then $E(Z) = 2E(Y)$.
- If $Z = kY$ where k is fixed constant, then $E(Z) = kE(Y)$.
- For example, if the entrance fee for the fair die game is \$7, then the expected winning is -\$0.50. Let $Z = Y - 7$. Since $E(Y) = 6.5$, $E(Z) = E(Y) - 7 = 6.5 - 7 = -0.50$

Random Variables

Variance

- *Variance* of a r.v. Y measures the population spread/variability of the probability distribution of Y .
- Variance is denoted as $Var(Y)$ or σ_Y^2 .
- Variance can be thought of as the amount Y deviates from the expectation μ_Y .
- For a discrete r.v. Y , the formula for variance of Y is

$$Var(Y) = E(Y - \mu_Y)^2 = \sum (y - \mu_Y)^2 \times p(y)$$

summing over all possible values y of the r.v. Y .

- In the fair die game,

$$\begin{aligned} Var(Y) &= (0 - 6.5)^2 \times 1/2 + (9 - 6.5)^2 \times 1/6 + (15 - 6.5)^2 \times 1/3 \\ &= 46.25. \end{aligned}$$

- Standard deviation is the square root of variance

$$\sigma_Y = \sqrt{Var(Y)}$$

or simply σ .

- In the fair die game, $\sigma_Y = \sqrt{46.25} = 6.8$.

Random Variables

Properties of variance

For a r.v. Y ,

- $Var(Y) \geq 0$. If $Var(Y) = 0$, then Y is a constant.
- If $Z = Y + c$ where c is fixed constant, then $Var(Z) = Var(Y)$.
- If $Z = 2Y$, then $Var(Z) = 4Var(Y)$.
- If $Z = kY$ where k is fixed constant, then $Var(Z) = k^2Var(Y)$.
- For example, if the entrance fee for the fair die game is \$7, then the variance is still 46.25.

Random Variables

Independence

- Two r.v.'s X and Y are *independent* if knowledge of the value of one r.v. has no effect on the probability about the other r.v.
- For discrete r.v.'s, X and Y are independent if

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

for *any* x and y .

- For example, the size and the weight of a cow are probably not independent.

Random Variables

More on two r.v.'s

For r.v.'s X and Y ,

- $X + Y$ is a r.v.
- $E(X + Y) = E(X) + E(Y)$
- $E(X - Y) = E(X) - E(Y)$
- $Var(X + Y) = Var(X) + Var(Y)$, if X and Y are independent.
- $Var(X - Y) = Var(X) + Var(Y)$, if X and Y are independent.
- For example X = milk yield from cow #1 and Y = milk yield from cow #2.