Module 3 summary: Probability: The Language of Uncertainty

Contingency Tables

- To explore the relationship between two categorical variables for the object, we create contingency tables, also known as cross-tabulations.
- Contingency tables present one categorical variable as the rows and the other categorical variable as the columns.
- These tables are used to calculate the conditional probabilities or percentages, which makes it easier for us to explore potential associations between variables.

2

Contingency Tables

- Consider the relationship between the cut of a diamond and its clarity.

```
> tally( ~ cut + clarity,margins=TRUE,data=Diamonds)

     clarity
cut          I1   SI1  SI2  VS1   VS2 VVS1 VVS2   IF Total
Fair        210   408  466  170   261   17   69    9  1610
Good         96  1560 1081  648   978  186  286   71  4906
Very Good    84  3240 2100 1775  2591  789 1235  268 12082
Premium     205  3575 2949 1989  3357  616  870  230 13791
Ideal       146  4282 2598 3589  5071 2047 2606 1212 21551
Total       741 13065 9194 8171 12258 3655 5066 1790 53940
```

3

Contingency Tables

- Can calculate the conditional column percentages using the following code:
- Use round() to reduce the decimal points, otherwise R uses 6 decimal points.

```
> round(tally( ~ cut | clarity,margins=TRUE, format = "proportion", data=Diamonds),3)

           clarity
cut            I1   SI1   SI2   VS1   VS2  VVS1  VVS2    IF
  Fair      0.283 0.031 0.051 0.021 0.021 0.005 0.014 0.005
  Good      0.130 0.119 0.118 0.079 0.080 0.051 0.056 0.040
  Very Good 0.113 0.248 0.228 0.217 0.211 0.216 0.244 0.150
  Premium   0.277 0.274 0.321 0.243 0.274 0.169 0.172 0.128
  Ideal     0.197 0.328 0.283 0.439 0.414 0.560 0.514 0.677
  Total     1.000 1.000 1.000 1.000 1.000 1.000 1.000 1.000
```

4

Interpretation of Contingency table

- These are conditional column probabilities.
- If we add all the probabilities for a column together, they will equal 1, e.g. sum column 1=(0.283, 0.130, 0.113, 0.227, 0.197) = 1.00.
- Comparing the worst clarity I1 with the best, IF (Flawless). Probability of IF diamonds having Ideal cut is 0,667 vs I1 of 0.197.

Clustered Bar Charts using contingency table

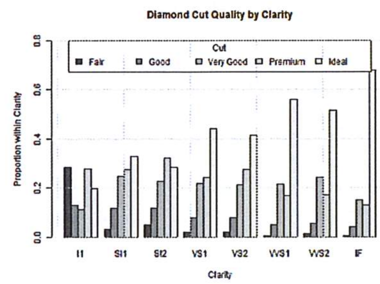
- Clustered bar charts are a great way to visualise two qualitative variables.
- First, must create contingency table in an object called table.
- This will make it easy for us to create the clustered bar chart.
- > table<-tally(~ cut | clarity, format = "proportion", data=Diamonds)

Clustered Bar Charts using contingency table

```
> barplot(table, main = "Diamond Cut Quality by Clarity",ylab="Proportion within Clarity",ylim=c(0,.8),legend=rownames(table),beside=TRUE, args.legend=c(x = "top",horiz=TRUE,title="Cut"),xlab="Clarity")
> grid()
```

- Notice how grid() was added after the plot was produced in R. Grid lines can help the viewer quickly read off and compare values on the plot axes.

When the clarity of a diamond increases, the quality of the cut also tends to increase.



Probability:

- **Probability** is defined as the proportion of times a random **event** occurs in a very large number of **trials**. Its value is between 0 and 1.
- The probability of an event = ***f/n***, where *f* is the frequency or number of times an event occurs and *n* is the total sample size.
- As the sample size *n* increases, the sample will begin to approximate the true population probability.

9

Table 10 of the Australian Bureau of Statistics 2011 -2012 National Health Survey

FRUIT INTAKE

Fruit Intake : Table

Table 1. Self-reported Fruit Intake–2011–12, Persons '000

	Age group (years)							
	18–24	25–34	35–44	45–54	55–64	65–74	75 +	Total
Males								
Usual daily intake of fruit								
< 1 serve	232	463	463	365	262	129	73	2036
1 serve	397	563	626	479	345	243	143	2691
2 serves	254	379	348	380	369	257	174	2182
3 serves or r	188	202	215	265	269	195	164	1498
Total	1120	1606	1551	1489	1265	824	651	8406
Females								
Usual daily intake of fruit								
< 1 serve	159	264	280	251	191	83	64	1332
1 serve	417	602	599	456	307	209	165	2755
2 serves	318	503	446	503	463	323	248	2802
3 serves or r	136	231	261	324	343	242	214	1747
Total	1070	1600	1587	1534	1301	857	688	8636
Persons								
Usual daily intake of fruit								
< 1 serve	431	727	743	616	463	212	137	3368
1 serve	813	1164	1124	934	652	452	306	5445
2 serves	672	883	795	863	853	560	420	4984
3 serves or r	323	433	476	589	603	433	378	3244
Total	2189	3206	3138	3023	2566	1681	1239	17042

10

Rules

$$Pr(2 \text{ serves}) = 4984/17042 = .292$$

$$Pr(<1 \text{ serve}) + Pr(1 \text{ serve}) + Pr(2 \text{ serves}) + Pr(3 \text{ serves or more}) = 1$$

$$\frac{3368+5445+4984+3244}{17042} = \frac{17042}{17042} = 1$$

11

Rules

- Two events are **mutually exclusive** if, when one event occurs, the other cannot and vice versa.
- Mutually exclusive sets have no **intersection**: *Pr*(A ∩ B) = 0. We use ∩ to denote an intersection.
- Example:The levels of fruit consumption are mutually exclusive. A person cannot occupy more than one category at a particular time.

12

Rules Intersection, Union and Complement

$$Pr(1 \text{ serve} \cap 2 \text{ serves}) = 0$$

$$Pr(< 1 \text{ serve } \cap \text{Male}) = \frac{2471}{17042} = .145$$

$$Pr(1 \text{ serve} \cup < 1 \text{ serve}) = \frac{3368 + 5445}{17042} = .517$$

$$Pr(\overline{< 1 \text{ serve}}) = Pr(1 \text{ serve}) + Pr(2 \text{ serves}) + Pr(3 \text{ serves or more})$$
$$= \frac{5445 + 4984 + 3244}{17042} = .802$$

or...

$$Pr(\overline{< 1 \text{ serve}}) = 1 - Pr(< 1 \text{ serve}) = 1 - \frac{3368}{17042} = .802$$

13

Conditional Probability

The probability that an event, B, will occur given that another event, A has already occurred.

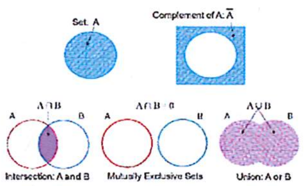
$$Pr(B|A) = \frac{Pr(A \cap B)}{Pr(A)}$$

Using an example..

$$Pr(< 1 \text{ serve} \mid \text{Male}) = \frac{Pr(\text{Male} \cap < 1 \text{ serve})}{Pr(\text{Male})} = \frac{2471/17042}{8842/17042} = .280$$

15

Rules



14

Conditional probability

Using conditional probability to check independence.

The two events A and B are independent if and only if *Pr*(A|B) = *Pr*(A) or *Pr*(B|A) = *Pr*(B).

Use this rule to reconfirm that gender and fruit consumption are dependent:

$$Pr(< 1 \text{ serve} \mid \text{Male}) = .280$$

$$Pr(< 1 \text{ serve}) = \frac{3368}{17042} = .198$$

16

Permutations

- There are six candidates. You need to vote for the top three (order matters).
- How many possible ways can you assign your votes, 1st, 2nd and 3rd preference? This is an example of a permutation problem, 3 possible ways:

Veronica Paskett	Milagros Depaolo	Loraine Muntz	Thuy Silverberg	Myriam Hakes	Maude Dimery
1st	2nd	3rd	-	-	-
-	1st	2nd	3rd	-	-
-	2nd	1st	-	3rd	-

Permutations

Calculation of number of possible permutation of k observations out of n observations;

$$P(n,k) = \frac{n!}{(n-k)!}$$

The n! is known as the factorial of a number. For example 6! = 6 x 5 x 4 x 3 x 2 x 1 = 720.

Solving the voting problem:

$$P(6,3) = \frac{6!}{(6-3)!} = \frac{6!}{3!} = \frac{720}{6} = 120$$

Using R:

```
> factorial(6)/factorial(6-3)
[1] 120
```


Combinations

- In combination, we need to know how many possible combinations of selecting four out of ten friends exist (with combinations the order does not matter), 3 possibilities are:

Leah	Rosalie	Marlena	Tarra	Graham	Gilberto	Marcos	Gladis	Otha	Jeremiah
Ticket	-	Ticket	-	Ticket	-	-	-	-	Ticket
-	-	-	-	-	-	Ticket	Ticket	Ticket	Ticket
-	Ticket	-	Ticket	-	Ticket	-	-	-	Ticket

Combinations

$$C(n,k) = \frac{n!}{(n-k)!k!}$$

This is known as the choose formula or the binomial coefficient (we will revisit this in Module 4). Solving, we find:

$$C(n,k) = \frac{n!}{(n-k)!k!} = \frac{10!}{(10-4)!4!} = \frac{10!}{6!4!} = \frac{3628800}{17280} = 210$$

Using R:

```
> choose(10,4)
[1] 210
```