Chapter 13 Independence

1 Probability and applications

1.1 Joint probabilities

Data sets with two categorical variables

Bivariate categorical data sets are usually summarised with a contingency table.

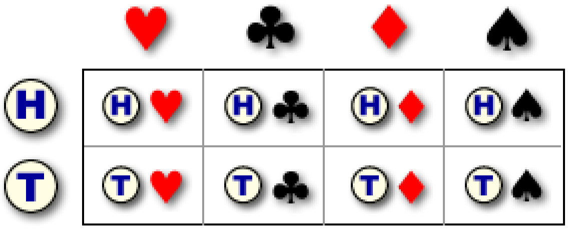
For example, a study examined 686 tourists and classified each by educational level and by whether they were 'information seekers' (who requested destination-specific literature from travel agents) or 'non-seekers':

	Information seeker?		
Education	**Yes**	**No**	**Total**
Some high school	13	27	40
High school degree	64	118	182
Some college	100	123	223
College degree	59	69	128
Graduate degree	67	46	113
Total	303	383	686

Joint probabilities

Bivariate categorical data can be modelled as a random sample from an underlying population of pairs of categorical values. The population proportion for each pair (x, y) is denoted by pxy and is called the joint probability for (x, y).

In games of chance, we can often work out the joint probabilities. For example, if a gambler draws a card from a shuffled deck and also tosses a coin, there are eight possible combinations,



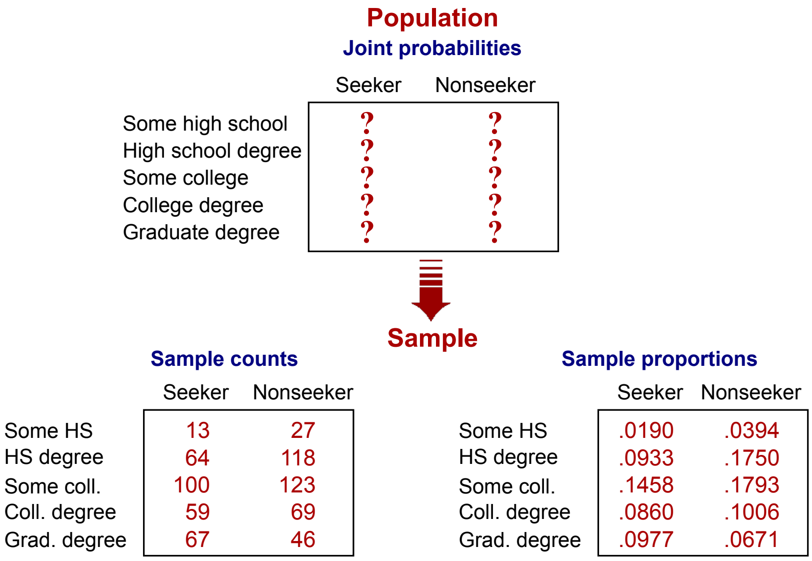
	♥	♣	♦	♠
H	H ♥	H ♣	H ♦	H ♠
T	T ♥	T ♣	T ♦	T ♠

Since these are equally likely,

$$p_{head,heart} = p_{head,club} = ... = p_{tail,spade} = \frac{1}{8} = 0.125$$

Interest in the model

In practice, we usually only have a random sample (summarised by a contingency table) and do not know the underlying joint probabilities. The sample proportions however provide **estimates**.



Population

Joint probabilities

	Seeker	Nonseeker
Some high school	?	?
High school degree	?	?
Some college	?	?
College degree	?	?
Graduate degree	?	?

Sample

Sample counts

	Seeker	Nonseeker
Some HS	13	27
HS degree	64	118
Some coll.	100	123
Coll. degree	59	69
Grad. degree	67	46

Sample proportions

	Seeker	Nonseeker
Some HS	.0190	.0394
HS degree	.0933	.1750
Some coll.	.1458	.1793
Coll. degree	.0860	.1006
Grad. degree	.0977	.0671

1.2 Marginal probabilities

Probabilities for a single variable

A model for two categorical variables is characterised by the joint probabilities p_{xy}.

The **marginal probability**, p_x, for a variable X is the proportion of (x, y) pairs in the population with $X = x$. This can be found by adding all joint probabilities for pairs with this x-value.