Lecture 04. Statistics and probability theory.

Question on the smokers in some population. The joint distribution of the pair of the categorical variables (Smoker, Gender)

	Gender		Total
Smoker	Male	Female	
Yes	1630	1684	
No	5550	8232	
Total			

	Gender		Total
Smoker	Male	Female	
Yes	1630	1684	3314
No	5550	8232	13782
Total	7180	9916	17096

The marginal distributions

	Gender		Total
Smoker	Male	Female	
Yes	1630	1684	
No	5550	8232	
Total			

	Gender		Total
Smoker	Male	Female	
Yes			3314
No			13782
Total	7180	9916	17096

The joint distribution

	Gender		Total
Smoker	Male	Female	
Yes	0.095	0.098	0.193
No	0.324	0.481	0.805
Total	0.419	0.579	1*

* - due to rounding errors, this result may differ slightly.

The conditional distributions

	Gender		Total
Smoker	Male	Female	
Yes	0.2270	0.1698	
No	0.7730	0.8302	
Total	1	1	

	Gender		Total
Smoker	Male	Female	
Yes	0.4919	0.5081	1
No	0.4027	0.5973	1
Total			

Independence of the distributions

Smoker	Gender		Total
	Male	Female	
Yes			3314
No			13782
Total	7180	9916	17096

$a + b = c + d = S$

Smoker	Gender		Total
	Male	Female	
Yes			a
No			b
Total	c	d	S

Smoker	Gender		Total
	Male	Female	
Yes	$a \cdot \frac{c}{S}$	$a \cdot \frac{d}{S}$	a
No	$b \cdot \frac{c}{S}$	$b \cdot \frac{d}{S}$	b
Total	c	d	S

Smoker	Gender		Total
	Male	Female	
Yes			$\frac{a}{S}$
No			$\frac{b}{S}$
Total	$\frac{c}{S}$	$\frac{d}{S}$	1

Smoker	Gender		Total
	Male	Female	
Yes	$\frac{a}{S} \cdot \frac{c}{S}$	$\frac{a}{S} \cdot \frac{d}{S}$	$\frac{a}{S}$
No	$\frac{b}{S} \cdot \frac{c}{S}$	$\frac{b}{S} \cdot \frac{d}{S}$	$\frac{b}{S}$
Total	$\frac{c}{S}$	$\frac{d}{S}$	1

Generalization

Variable 2	Variable 1			Total
		...		
		...		
$\vdots$		...		
		...		
Total		...		

Probabilistic model $(\mathbf{S}, \mathbf{P})$ (S is the sample space, P is the probability assigned to the events.)

A random phenomenon has outcomes that we cannot predict but that nonetheless have a regular distribution in very many repetitions. The probability of an event is the proportion of times the event occurs in many repeated trials of a random phenomenon.

Sample space S:

The sample space S of a random phenomenon is the set of all possible outcomes, $S = \{\omega : \omega \text{ is an outcome}\}$.

Example.

Toss a coin. There are only two possible outcomes

ω_1="tail", ω_2="head"

and the sample space is $S = \{\omega_1, \omega_2\} = \{head, tail\}$ or, more briefly, $S = \{H, T\}$. It is often convenient to quantify by putting H = 1, T = 0 then $S = \{0, 1\}$.

Example.
We are tossing a coin n times and record outcomes. There are 2^n possible outcomes.
A coin is tossed repeatedly n times. The joint outcome may be recorded as a sequence of H's and T's, where H = "head," T = "tail." Since there are 2 outcomes for each trial, there are 2^n possible joint outcomes: $\omega = (a_1, a_2, a_3, \ldots, a_n)$, where a_i =H or T, $i = 1, 2, 3, \ldots, n$.

Example. We are rolling a die and record the count of spots on the up-face. There are six possible outcomes

$$\begin{array}{lll} \omega_1 = 1, & \omega_2 = 2, & \omega_3 = 3, \\ \omega_4 = 4, & \omega_5 = 5, & \omega_6 = 6 \end{array}$$

and the sample space is $S = \{1, 2, 3, 4, 5, 6\}$

Example. We are first tossing a coin and next we are rolling a die, and record the results. There are twelve possible joint outcomes

$$\begin{array}{lll} \omega_1 = (H, 1), & \omega_2 = (H, 2), & \omega_3 = (H, 3), \\ \omega_4 = (H, 4), & \omega_5 = (H, 5), & \omega_6 = (H, 6), \\ \omega_7 = (T, 1), & \omega_8 = (T, 2), & \omega_9 = (T, 3), \\ \omega_{10} = (T, 4), & \omega_{11} = (T, 5), & \omega_{12} = (T, 6), \end{array}$$

and the sample space is $S = \{H, T\} \times \{1, 2, 3, 4, 5, 6\}$

Event

An event is a set of outcomes of a random phenomenon. That is, any subset A of the sample space S is an event.

There are two particular events:

the empty set $\emptyset$ - the event that never occurs,
the whole sample set S - the event that always occurs.

Example. We are tossing a coin two times. The event A involving at least one head observed has the form $A = \{(H, H), (H, T), (T, H)\}$.

Example. The one-element event $\{\omega\}$ means that the outcome is exactly ω.

Probability P

In a probability model, **events have probabilities**. What properties must any assignment of probabilities to events have? Here are some basic facts about any probability model. These facts follow from the idea of probability as "the long run proportion of repetitions on which an event occurs."

Fundamental rules of probability

1. Any probability is a number between 0 and 1: $0 \leq P(A) \leq 1$, for every event A.

2. If S is the sample space in a probability model, then $P(S) = 1$. The event $\emptyset$ has probability 0: $P(\emptyset) = 0$.

3. If two events have no outcomes in common (are disjoint), the probability that one or the other occurs is the sum of their individual probabilities:

for every events $A, B \subset S$, if $A \cap B = \emptyset$, then

$$P(A \cup B) = P(A) + P(B).$$

This is the addition rule for disjoint events.

4. The probability that an event does not occur is 1 minus the probability that the event does occur:

$$P(A^c) = 1 - P(A)$$

for every event $A \subset S$, A^c is a complement of A.

Assigning probabilities: finite number of outcomes: $S = \{\omega_1, \omega_2, \ldots, \omega_n\}$

The individual outcomes of a random phenomenon are always disjoint ($\{\omega_i\} \cap \{\omega_j\}$, if $\omega_i \neq \omega_j$). So the addition rule provides a way to assign probabilities to events with more than one outcome: start with probabilities for individual

outcomes:

$$P(\{\omega_1\}) = p_1,\ P(\{\omega_2\}) = p_2,\ \ldots,\ P(\{\omega_n\}) = p_n$$
$$p_1,\ p_2, \ldots, p_n \geq 0,\ p_1 + p_2 + \cdots + p_n = 1.$$

and add to get probabilities for events: for every event $A = \{\omega_{i_1}, \omega_{i_1}, \ldots, \omega_{i_1}\}$ we define

$$P(A) = \sum_{\omega \in A} P(\{\omega\}) = P(\{\omega_{i_1}\}) + P(\{\omega_{i_2}\}) + \cdots + P(\{\omega_{i_n}\}) = p_{i_1} + p_{i_2} + \cdots + p_{i_n}.$$

This idea works well when there are only a finite (fixed and limited) number of outcomes.

A particular case: **equally likely outcomes.** If a random phenomenon has k possible outcomes, all equally likely, then each individual outcome has probability $1/k$. The probability of any event A is

$$P(A) = \frac{\text{count of outcomes in } A}{\text{count of outcomes in } S} = \frac{\text{count of outcomes in } A}{k}$$

Example. We are tossing a coin n times and record outcomes. A coin is tossed repeatedly n times. The joint outcome may be recorded as a sequence of H's and T's, where H = "head," T = "tail." There are 2^n possible joint outcomes. If all of these are assumed to be equally likely so that each particular joint outcome has probability $\frac{1}{2^n}$:

$$\mathrm{P}(\{a_1, a_2, a_3, \ldots, a_n)\}) = \frac{1}{2^n},$$

where a_i =H or T, $i = 1, 2, 3, \ldots, n$.

Example. We are rolling a symmetric die. The possible outcomes are $1, 2, 3, \ldots, 6$. Since the die is symmetric, the outcomes are equally likely, we have

$$\mathrm{P}(\{1\}) = \mathrm{P}(\{2\}) = \mathrm{P}(\{3\}) = \cdots = \mathrm{P}(\{6\}) = \frac{1}{6}$$

and for every event $A \subset S = \{1, 2, 3, 4, 5, 6\}$ we have

$$\mathrm{P}(A) = \frac{|A|}{|S|} = \frac{|A|}{6},$$

where $|A|$ is a count of outcomes in A.

Example. We are rolling two dice, the first is white and the other is red. Describe the probabilistic space, assign probability to the events.

$$\text{outcomes}: \ \omega = (k_1, k_2), \ k_1, k_2 = 1, 2, 3, 4, 5, 6,$$

where k_1 and k_2 are spots on the white and red die, respectively;

the probability space:

$$S = \{\omega : \ \omega = (k_1, k_2), \ k_1, k_2 = 1, 2, 3, 4, 5, 6\}$$
$$|S| = 6 \cdot 6 = 36$$

Since outcomes are equally likely, the probability assigned to any event A is equal to

$$\mathrm{P}(A) = \frac{|A|}{|S|} = \frac{|A|}{36}.$$

Example. We are rolling two dice, of the same colour (they are indistinguishable). Describe the probabilistic space, assign probability to events.

outcomes:

$$\omega = (k_1, k_2), \ k_1 \leq k_2, \ k_1, k_2 = 1, 2, 3, 4, 5, 6$$

where k_1 and k_2 are spots on dice.

the probability space:

we divide the outcomes into two disjoint groups:
the first group S_1 consists of outcomes (k_1, k_2) with $1 \leq k_1 < k_2 \leq 6$. We record such an outcome in the following cases

1. k_1 spots ocurred on the first die and k_2 spots on the other
or
2. k_2 spots ocurred on the first die and k_1 spots on the other

the second group S_2 consists of outcomes (k, k) with $1 \leq k \leq 6$. We record such an outcome in the case when the same number of spots occured on the both diece.
The numbers of elements of the groups S_1 and S_2 are equal to

$$|S_1| = 1 \cdot 5 + 1 \cdot 4 + 1 \cdot 3 + \cdots + 1 \cdot 1 = 15, \ |S_2| = 6.$$

We see that outcomes of each group S_1 and S_2 are equally likely

$$P(\{(1,2)\}) = P(\{(1,3)\}) = \cdots = P(\{(k_1,k_2)\}) = \cdots = P(\{(5,6)\}) = p_1,$$
$$1 \le k_1 < k_2 \le 6$$

$$P(\{(1,1)\}) = P(\{(2,2)\}) = \cdots = P(\{(6,6)\}) = p_2$$

Moreover outcomes of the group S_1 occur twice as often as outcomes of the group S_2, therefore

$$p_1 = 2 \cdot p_2$$

Finally, adding the probabilities all of single outcomes we get

$$1 = |S_1| \cdot p_1 + |S_2| \cdot p_2 = 15p_1 + 6p_2 = 15 \cdot 2p_2 + 6p_2 = 36p_2$$

Hence

$$p_2 = 1/36,\ p_1 = 2 \cdot p_2 = 1/18.$$

Infinite discrete (countable) sample space

Example. (Spaces of different sizes)

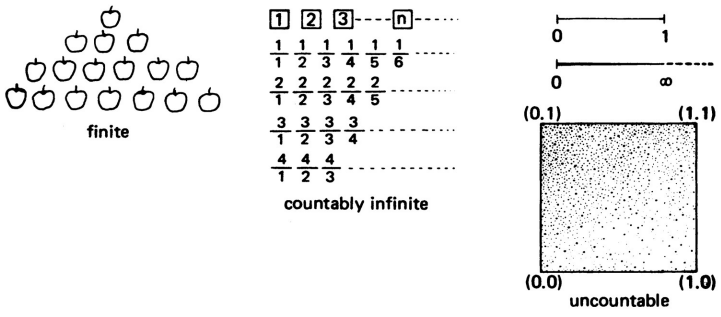


Figure 12

In this case a random phenomenon has infinite number of possible outcomes labeled with natural numbers: $\omega_1, \omega_2, \omega_3, \ldots$. Then the sample space $S = \{\omega_1, \omega_2, \omega_3, \ldots\}$. To describe the probablity we first have to determine the probabilities of single outcomes

$$\mathrm{P}(\{\omega_1\}) = p_1,\ \mathrm{P}(\{\omega_2\}) = p_2,\ \mathrm{P}(\{\omega_3\}) = p_3,\ \ldots,$$
$$p_1, p_2, p_3, \ldots, p_n, \cdots \ge 0$$
$$p_1 + p_2 + p_3 + \cdots + p_n + \cdots = \sum_{k=1}^{\infty} p_k = 1.$$

For any event $A \subseteq S$ we define the probability $\mathrm{P}(A)$ as follows

$$\mathrm{P}(A) = \sum_{\omega \in A} \mathrm{P}(\{\omega\}).$$

Example. We are tossing a symmetric coin until the head occurrs. In such a case outcomes are as follows: $\omega_1 = (H)$, $\omega_2 = (T, H)$, $\omega_3 = (T, T, H)$,..., $\omega_n = (\underbrace{T, T, \ldots, T}_{n-1}, H)$, ...
The sample space: $S = \{\omega_1, \omega_2, \ldots, \omega_n, \ldots\}$.
Events are arbitrary subsets $A \subseteq S$.

Assigning the probabilities: we take $\mathrm{P}(\{\omega_n\}) = \frac{1}{2^n}$. As a consequence, the probability of any event A is defined as

$$\mathrm{P}(A) = \sum_{\omega_k \in A} \mathrm{P}(\{\omega_k\}) = \sum_{\omega_k \in A} \frac{1}{2^k} \text{ for } A \subseteq S$$

Exercise 1. . Distribution of blood types. All human blood can be "ABO-typed" as one of O, A, B, or AB, but the distribution of the types varies a bit among groups of people. Here is the distribution of blood types for a randomly chosen person in the United States:

Blood type	A	B	AB	O
U.S. probability	0.40	0.11	0.04	?

(a) What is the probability of type O blood in the United States?

(b) Maria has type B blood. She can safely receive blood transfusions from people with blood types O and B. What is the probability that a randomly chosen American can donate blood to Maria?

Answer.

Probabilistic model:
possible outcomes: A, B, AB, 0, sample space $S = \{A, B, AB, 0\}$

(a) Assignig probabilities:
1. $\mathrm{P}(\{A\}) = 0.40$, $\mathrm{P}(\{B\}) = 0.11$, $\mathrm{P}(\{AB\}) = 0.04$, $\mathrm{P}(\{0\}) = 1 - (0.40 + 0.11 + 0.04) = 0.45$

(b) $\mathrm{P}(\{B, 0\}) = 0.11 + 0.45 = 0.55$