Probability

Probability	a number between 0 and 1 that indicates how likely it is that a specific event or set of events will occur.
Simple experiment	some well-defined act or process that leads to a single well-defined outcome. For example, a coin toss will yield either a heads or a tails; a birth will yield either a boy or a girl. (NOTE: Statisticians do NOT use the term "Experiment" in the same way a Social psychologist or a chemist would).
Sample space	the set of all possible distinct outcomes of an experiment. For example, if you toss a coin once, the possible outcomes are H or T; toss it twice, and the possible outcomes are HH, HT, TH, and TT.
Sample point, or elementary event	any member of the sample space. One possible result of a single trial of the experiment. e.g., getting a Heads when tossing a coin; getting the Ace of Hearts when pulling a card from a deck.
Event, or event class	some subset of the outcomes of an experiment; any set of elementary events. e.g. getting a "heart" when you pull a card from a deck is achieved by 13 different elementary events.
Mutually exclusive outcomes	Any set of events that cannot occur simultaneously. For example, for the variable GENDER, a person cannot be both male and female. Conversely, for ETHNICITY, an individual could claim both European and Asian ethnic heritages.
Independent events	events that have nothing to do with each other; the occurrence of one event in no way affects the occurrence of the other. For example, the result of one coin toss does not affect the possible value of the next.

NOTE: Mutually exclusive and independent are not one and the same!!! If someone is Male, we know they are not female; male and female are mutually exclusive events. But, if one coin toss comes up heads, we know nothing about the value of the next coin toss.

PROBABILITY AXIOMS:

1. $0 \leq P(E_i) \leq 1$, where E_i = Event i (NOTE: A is often used instead of E_i).
2. P(S) = 1, where S = the Sample Space
3. When all sample points are equally likely,

$$P(E_i) = \frac{Number\ of\ elementary\ events\ in\ E_i}{Total\ number\ of\ possible\ events}$$

For example, if E_1 = getting a 1 when you roll a fair die, $P(E_1) = 1/6$; if E_1 = getting an even number when you roll a fair die, $P(E_1) = 3/6$.

4. Total number of sample points = $n_1 * n_2 * n_3 * ... * n_k$, where n_i = number of possible outcomes for the ith variable. All outcomes need not be equally likely. Hence, if you toss a coin three times, there are 8 possible outcomes.

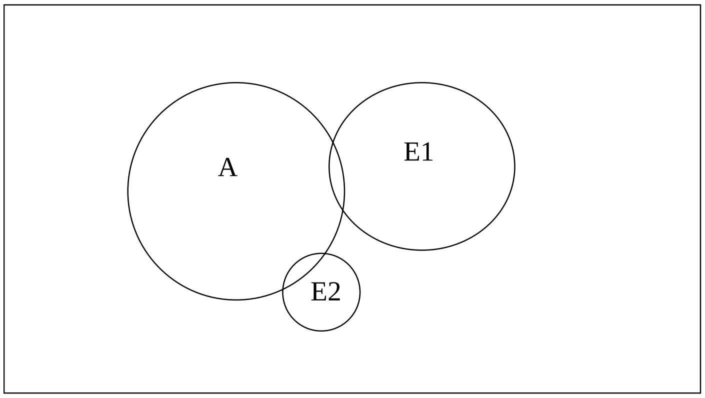
Summary of probability rules: Let A and B be two events of interest in a particular experiment. (The rules will make more sense once you see the examples.)

Rule Name/ Explanation	General Rule	Rule for mutually exclusive events	Rule for independence
Complements - Prob that A does not occur	$P(\overline{A}) = 1 - P(A)$		
Conditional probability - Prob that A occurs given that B has or will occur [“Probability of A given B”]	$P(A\|B) = \dfrac{P(A \cap B)}{P(B)}$	$P(A\|B) = 0$	$P(A\|B) = P(A)$
Joint Probability - Prob that both A & B occur in one replication of the experiment; the prob of the intersection of A & B [“Probability of A and B”]	$P(A \cap B) =$ $P(B)P(A\|B) =$ $P(A)P(B\|A)$	$P(A \cap B) = 0$	$P(A \cap B) =$ $P(A)P(B)$
Additive probability, aka Probability of a union - Prob that either A or B or both occur in one replication of the experiment [“Probability of A or B”]	$P(A \cup B) =$ $P(A) + P(B) -$ $P(A \cap B)$	$P(A \cup B) =$ $P(A) + P(B)$	$P(A \cup B) =$ $P(A) + P(B) -$ $P(A)P(B)$
Marginal probability	$P(A) =$ $\sum P(A \cap E_i) =$ $\sum P(E_i)P(A\|E_i)$		
Bayes' Rule (Another formula for conditional probability)	$P(E_i\|A) =$ $\dfrac{P(E_i)P(A\|E_i)}{\sum_j P(E_j)P(A\|E_j)}$		

Probability – Intuitive/Substantive Examples

The above formulas are actually pretty easy; it is knowing when to use them that is the trick. It is especially easy to mix up conditional probability, joint probability, and additive probability. Here are some examples that might give you an intuitive feel for the different types of probabilities. (Any numbers I give are just made up; also, drawings are not to scale.)

1. 5% of the population dies from heart attacks every year. 1% of all those aged 20-25 die from heart attacks every year. 10% of those aged 80 and above die from heart attacks every year.

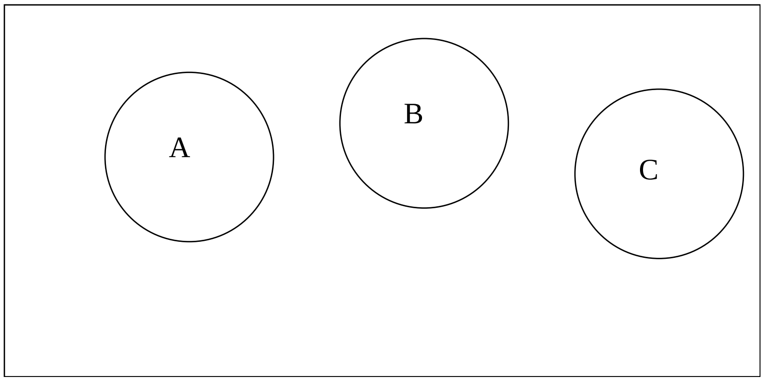


The first statement gives the marginal probability of dying from a heart attack, i.e. if A = dying from heart attack, then P(A) = .05. Everything not within A represents people who do not die or who die from other causes. However, as the next two statements show, the probability of dying from a heart attack varies by age; that is, these statements give the conditional probability of your dying from a heart attack given your age. Hence, if E1 = aged 20-25 and E2 = aged 80 and above, then P(A | E1) = .01 and P(A | E2) = .10. Also, E1 and E2 are mutually exclusive events; you can't be both ages 20-25 and 80+.

The areas where the circles overlap reflect their joint probabilities, e.g. the probability of dying from a heart attack and being age 80+, or $P(A \cap E2)$. If these were drawn perfectly, 1% of E1 would overlap with A and 10% of E2 would overlap with A.

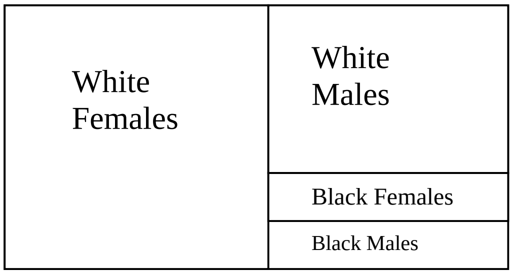
NOTE: If I know nothing about a person and I am asked to predict the probability of their dying in the next year from a heart attack, my best guess is 5%. But, if I know their age, a likely better guess is given by the conditional probability of death given age. That is what much of statistics is about: using information about people to better explain or better predict what happens to them.

2. 5% of the population dies from heart attacks every year, 5% dies from cancer, and 5% dies from other causes.



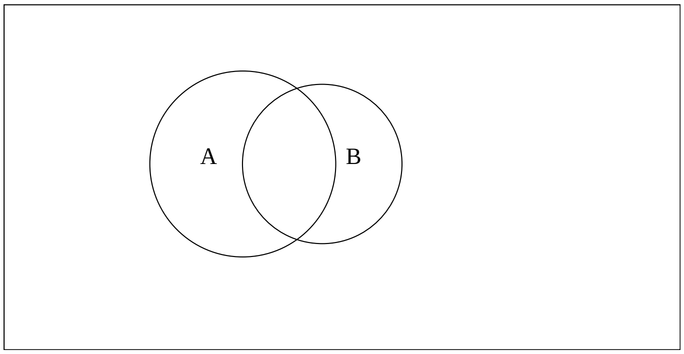
Again, you are given marginal probabilities. These happen to be mutually exclusive events, so add them all up and you get P(Dying in a year) = 15%. Everything not contained in the three circles reflects the 85% of the population that does not die.

3. Of those who voted in a recent election, 50% were white females, 35% were white males, 9% were black females, and 6% were black males.



Here, you are given several joint probabilities, e.g. the probability of being both white and female = P(white $\cap$ female) = .50. From the information given, you could easily determine the marginal probabilities for race and gender, e.g. P(White) = P(White $\cap$ Female) + P(White $\cap$ Male) = .85. Similarly, P(Black) = .15, P(Female) = .59, P(Male) = .41.

4. At a boys school, 10% of the students were on the football team and 8% were on the track team. Half of the boys who played football were also on the track team. Altogether, 13% of the boys were on at least one of the two teams, i.e. they were on the football team or the track team or both.



You have two marginal probabilities here: P(Football) = .10 and P(Track) = .08. There is one conditional probability: P(Track Team | Football team) = .50. And, there is one additive probability: P(Football $\cup$ Track) = .13.

Why isn't this last number .18? Because track and football are not mutually exclusive events; some boys were in both sports, as is shown by the overlap in the two circles. Since 10% of the boys are in football and half of those are also in track, the probability of being on both the track and football teams is .05, i.e. P(Football $\cap$ Track) = P(Football) * P(Track Team | Football team) = .10 * .50 = .05. Hence, to find the probability of being on football or track, you take P(Football) + P(Track) – P(Football $\cap$ Track) = .10 + .08 - .05 = .13.

Probability – Math Examples

1. In a family of 11 children, what is the probability that there will be more boys than girls?

SOLUTION. The easiest way to solve this is via the complements rule. Let event A = Having more boys than girls. $\overline{A}$ is therefore more girls than boys. Each of these events are equally likely, so P(A) = .50.

Note that the problem would be a little more difficult if there were 10 children, since you would then have to figure out the probability that there were an equal number of boys and girls. We'll see later how to do that.

2. You are playing *Let's Make a Deal* with Monte Hall. You are offered your choice of door #1, door #2, or door #3. Monte tells you that goats are behind two of the doors; but, behind the other door is a new car. You choose door #1. Monte, who knows what is behind each door, then opens door #3, revealing a goat. He then offers you the choice of either keeping your own door, #1, or else switching to Monte's remaining door, #2. Should you switch?

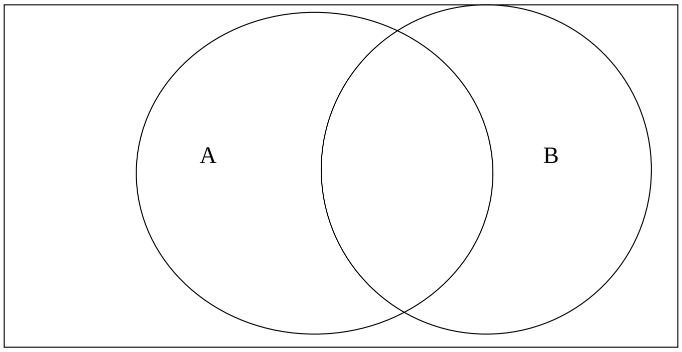
SOLUTION. The easiest way to see this is by using the complements rule. Let A = switch, $\overline{A}$ = doesn't switch. Note that resolving not to switch is the same as not having the option to switch. When you first pick, you have a 1/3 chance of winning; ergo, if you don't switch, your probability of winning stays at 1/3. Hence, you have a 2/3 chance of winning if you switch and only 1/3 if you don't switch. So, switch.

Hardly anybody believes this simple proof though, at least not right away! Indeed, this question embroiled the country in controversy in 1991. Marilyn vos Savant, allegedly the world's smartest person, received 1000s of letters when she addressed this question in her Parade Magazine column. More recently, the problem was included in the best-selling book, The Curious Incident of the Dog in the Night-Time, about an autistic child who tries to solve the murder mystery of the neighbor's dog. For further discussion of this burning issue, see the Appendix on The Monty Hall Controversy.

3. A survey shows that 44% of a magazine's readers are Protestants, 55% are Democrats, and 30% are Protestant Democrats. That is, if event A = Protestant, and event B = Democrat, then P(A) = .44, P(B) = .55, $P(A \cap B) = .30$. Answer the following:

a. What proportion of Democrats are Protestants?

SOLUTION. This is a question about conditional probability. Given that a reader is a Democrat, what is the probability that s/he is also a Protestant? This diagram may help you to visualize it a bit better.



44% of the sample space falls within A, 55% falls within B, and 30% falls in the intersection of A and B. Formulaically, the solution is

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{.30}{.55} = .55$$

Put into words, a little over half (55%) of the magazines readers are Democrats, and of those a little over half (55%) are also Protestants.

b. What proportion of Protestants are Democrats?

SOLUTION. Again, conditional probability. Of those readers that are Protestant, what percentage are also Democrats?

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{.30}{.44} = .68$$

A little less than half (44%) of the readers are Protestant, and of those more than 2/3 (68%) are Democrats. That is, if a reader is Protestant, there is better than a 2:1 chance they are also Democrat.

c. What proportion of the population is either Democrat or Protestant, or both?

SOLUTION. This is a question about the probability of a union. Note that

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = .44 + .55 - .30 = .69$$

Note that we can’t just add up the % Protestant and the % Democrat, because then the Protestant Democrats would get counted twice (once as Protestants and then again as Democrats).

4. Prove that Race and Gender are independent in the following table:

Race \ Gender	Male	Female	Total
White	35	35	70
Nonwhite	15	15	30
Total	50	50	100

SOLUTION. A = White, B = Male, P(A) = .70, P(B) = .50, P(A $\cap$ B) = .35. A and B are independent because

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{.35}{.50} = .70 = P(A),$$

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{.35}{.70} = .50 = P(B)$$

5. A new, less expensive method has been developed for testing for the AIDS virus. Fifty percent of the people who test positive actually have AIDS. Of those who test negative, 5% have AIDS. Twenty percent of the population tests positive.

a. What percentage of the population will receive false positive scores - that is, the test will say they have AIDS when they really don't? [i.e. what % will be unnecessarily scared?]

b. What percentage of the population will receive false negative scores - that is, the test will say they don't have AIDS when they really do? [i.e. what % will have a false sense of security]

c. What percentage of the population has AIDS? (Don't worry, these are fictitious numbers.)

d. What is the probability that someone who has AIDS will test positive?

SOLUTION. Let us use the following terms:

A = has Aids, $\overline{A}$ = does not have AIDS
E_1 = positive test, E_2 = negative test

We are told:

$P(E_1)$ = P(Positive test) = .20, implying
$P(E_2)$ = P(Negative test) = .80

$P(A \mid E_1)$ = P(having Aids if you test positive) = .50, implying
$P(\overline{A} \mid E_1)$ = P(not having Aids if you test positive) = .50

$P(A \mid E_2)$ = P(having aids if you test negative) = .05, implying
$P(\overline{A} \mid E_2)$ = P(not having aids if you test negative) = .95.

a. We are asked to find what percentage of the population does not have AIDS and receives a positive score, that is, $P(E_1 \cap \overline{A})$. Use joint probability.

$$P(\overline{A} \cap E_1) = P(E_1)P(\overline{A} \mid E_1) = .20 * .50 = .10$$

That is, half of the 20% of the population that tests positive will needlessly worry that they have AIDS.

b. We are asked to find what percentage of the population has AIDS and receives a negative score, that is, $P(A \cap E_2)$. Again use joint probability.

$$P(A \cap E_2) = P(E_2)P(A \mid E_2) = .80 * .05 = .04$$

That is, 5% of the 80% who test negative, or 4% altogether, will have AIDS but still test negative.

c. We are asked to find P(A), i.e. the probability of having AIDS. Use the marginal probability theorem:

$$P(A) = \sum P(E_i)P(A \mid E_i) = (.20 * .50) + (.80 * .05) = .14$$

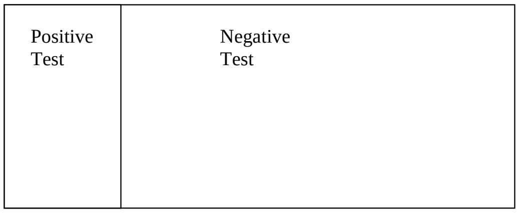
That is, 14% of the population has AIDS.

d. We are asked to find P(E_1 | A), i.e. what percentage of the people who have AIDS will be identified by this test? Use Bayes theorem.

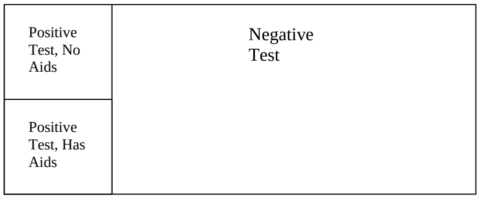
$$P(E_1|A) = \frac{P(E_1)P(A|E_1)}{\sum_j P(E_j)P(A|E_j)} = \frac{.20*.50}{(.20*.50)+(.80*.05)} = \frac{.10}{.14} = .714$$

So, this test will identify a little over 71% of those who have AIDS.

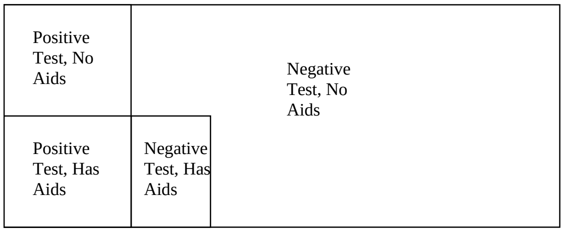
ALTERNATIVE APPROACH. If you find a problem like this confusing, you might try drawing a diagram and see if that helps. So, begin by noting that 20% test positive (which also means that 80% test negative):



Further, we are told that half of those who test positive have AIDS – so divide the left-hand portion into two equal parts:



Now, the right-hand side consists of those who tested negative. We are told that 5% of those who tested negative actually have AIDS, which means that the other 95% of that group does not have AIDS. So, the diagram looks roughly like this:



Each of the four boxes represents a joint probability. It should be fairly easy to figure out the probability of each of those combinations and to then solve the original problems from there.

6. A researcher is doing a study of gender discrimination in the American labor force. She has come up with a 3-part classification of occupations (Occupation 1, Occupation 2, and Occupation 3) and a 2-part classification for wages ("good" and "bad"). She finds that, by gender, the distribution of occupation and wages is as follows:

	Women			Men		
Pay/Occ	Occ 1	Occ 2	Occ 3	Occ 1	Occ 2	Occ 3
Good Pay	20%	7%	10%	7%	10%	60%
Bad Pay	50%	8%	5%	8%	5%	10%

From the table, it is immediately apparent that 37% of all women receive good pay, compared to 77% of the men. At the same time, it is also very clear that the types of occupations are very different for men and women. For women, 70% are in occupation 1, which pays poorly, while 70% of men are in occupation 3, which pays very well. Therefore, the researcher wants to know whether differences in the types of occupations held by men and women account for the wage differential between them. How can she address this question?

SOLUTION. This problem is best addressed by asking a "what if" sort of question: Suppose women were distributed across occupations the same way men were, but within each occupation had the same wage structure that they do now. If differences in types of occupations alone account for the wage discrepancies, then this approach should control for those differences and wage differentials should disappear.

We will use the following terms:

Event A = Receives Good pay, $\overline{A}$ = Bad pay, E_i = Employed in occupation i.

Given these definitions, this problem requires that we combine the occupational distribution for men $(P(E_i))^M$ with the conditional probabilities that a woman receives good wages given the occupation she is in $(P(A \mid E_i))^F$

For men $P(E_1)^M = .15$, $P(E_2)^M = .15$, $P(E_3)^M = .70$

For women $P(A \mid E_1)^W = 2/7$, $P(A \mid E_2)^W = 7/15$, $P(A \mid E_3)^W = 10/15$

Using the marginal probability theorem, we get

$$P(A) = \sum P(E_i)^M \, P(A \mid E_i)^W = (.15 * \tfrac{2}{7}) + (.15 * \tfrac{7}{15}) + (.70 * \tfrac{10}{15}) = .58$$

Hence, if women had the same occupational distribution while continuing to make the same salaries within occupations that they do now, 58% of women would make good wages. This is much more than the 37% of women who currently make good wages, but still well short of the male figure of 77%. Differences in occupational structure account for much of the difference between men and women, but not all.

NOTE: This sort of “what if” question comes up all the time. For example, in demography it is often difficult to compare death rates across populations, because one population might be relatively old (and hence has a lot of people at high risk of dying) while another is relatively young. Therefore, a common approach is to standardize using the age composition of one population and the age-specific death rates of the other. For example, Mexico and the United States have very similar crude death rates, i.e. about the same proportion of people die in both countries each year. This is highly deceptive because high fertility in recent years has caused Mexico's population to be much younger than is the United States'.