Methods of Mathematical Statistics

Notes by Tim Brown and Guoqi Qian

Module 1: Probability

Contents

1 Overview **2**

1.1 Housekeeping . . . 2

1.2 The Big Picture . . . 2

1.3 Methods of Mathematical Statistics Outline . . . 4

2 Probability Basics — 1.1 **4**

2.1 Intuitive Setup . . . 4

2.2 Method 1 — Relative Frequency . . . 5

2.3 Method 2 — Symmetry . . . 6

2.4 Method 3 — Subjective Belief . . . 7

2.5 Definitions . . . 7

2.6 Operations on Events . . . 8

2.7 Rules of probability . . . 11

2.8 Consequences . . . 11

3 Methods of Enumeration — 1.2 **14**

3.1 How to count? . . . 14

3.2 Example: Sampling without Replacement . . . 16

4 Conditional Prob.— 1.3 **18**

4.1 Bayesian and Frequentist Statistics . . . 18

4.2 Definition . . . 18

4.3 Example: Conditional Probability Definition . . . 18

4.4 Multiplication Rule . . . 20

4.5 Example: System Failure . . . 20

5 Bayes' Theorem — 1.5 **20**

5.1 Law of Total Probability . . . 20

5.2 Bayes' Theorem . . . 21

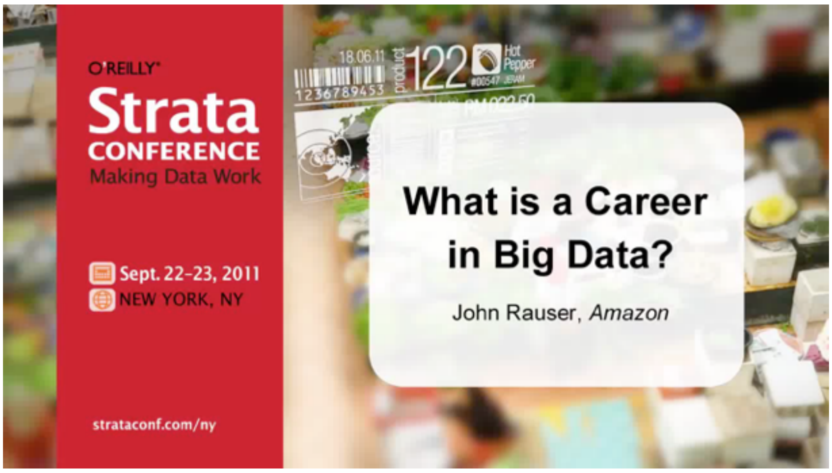
5.3 Bayes Example: Sampling w'out Replacement . . . 23

6 Independence — 1.4 **25**

6.1 Definition . . . 25

6.2 Example: Rocket Failure . . . 25

6.3 Example: Pairwise Independence — Soft Drinks . . . 27



O'REILLY

Strata

CONFERENCE

Making Data Work

Sept. 22-23, 2011

NEW YORK, NY

strataconf.com/ny

What is a Career in Big Data?

John Rauser, *Amazon*

Figure 1: https://www.youtube.com/watch?v=0tuEEnL61HM

1 Overview

1.1 Housekeeping

General Housekeeping

Textbook: Probability and Statistical Inference, Hogg, Tanis and Zimmerman, 9th Edn — most of the book = fast pace + double workload!

References: to textbook at the end of section titles — lectures not always the same order or content as textbook.

Office hours: MON and THU from 10:00 to 12:00. By appointment: send me email.

Enrolment: Anyone not enrolled in subject? Sign up if you have not yet. Workshops and labs start this week.

Assignments etc: On the learning management system including dates. Assignment 1 is due on March 17.

First week Survey: Please let me know your background.

Midterm Exam: Week 7, between April 16 and April 19.

1.2 The Big Picture

What is a Career in Data Science?

Figure 1 is the start of the youtube video referenced in the figure title.

The Big Picture

Figure 2 shows where our subject fits in Data Science.

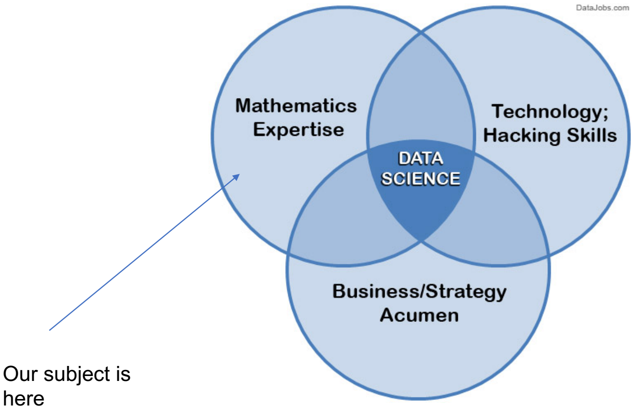


Figure 2: Where does our subject fit?

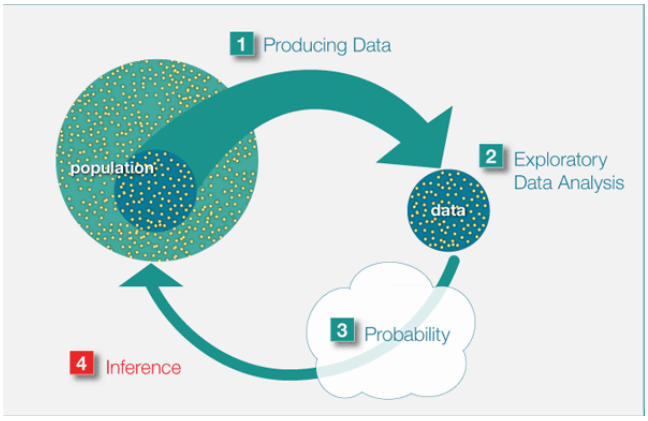


Figure 3: Our Subject Mostly 3 and 4

The Big Picture

Figure 3 gives more on the phases in Data Science problems and where our subject fits.

4 Phases in Data Science

1. Producing Data: collecting it as it arises, or conducting an experiment or random sample from a population.

2. Exploratory Analysis: organising the data (computer science), producing summaries (mean, standard deviation etc).

3. Probability: understanding the chance mechanisms that generated the data: random variables and their properties are important.

4. Inference: using probability to draw inferences about the data: find estimates of population quantities, understand the uncertainty in the estimates, decide on hypotheses and predict.

1.3 Methods of Mathematical Statistics Outline

Probability — first half of our subject

Module 1: the basics for understanding random samples — important background: sets, permutations and combinations, functions (1:1 and onto).

Module 2: discrete distributions and random variables — important background: series and the exponential function, differentiation.

Module 3: continuous distributions and random variables — important background: integration.

Module 4: distributions for two random variables, correlation — important background: two variable calculus.

Module 5: transformations of random variables and limits — important background: limits, multivariable calculus.

Statistical Inference — second half of our subject

Probability is important because the underlying idea of statistical inference is: *data* = *model* + *residual*, with both the model and the residuals assumed to obey laws of probability.

Module 6: data description and point estimation — important background: Modules 1 to 5.

Module 7: quantifying uncertainty through interval estimation — important background: Modules 1 to 5.

Module 8: tests of hypotheses — important background: Modules 1 to 5.

Module 9: special cases including regression, analysis of variance and contingency tables — important background: Modules 1 to 5.

2 Probability Basics — 1.1

2.1 Intuitive Setup

Probability in English

Everything varies,

So uncertainty can't be removed,

But probability tries to quantify it.

Probability ranges from 0 to 1 corresponding to impossibility and certainty.

Probability is assigned to *events*, which are collections of *outcomes*.

How is probability assigned?

3 Methods described in next slides.

2.2 Method 1 — Relative Frequency

1. Relative Frequency

Suppose: a tennis player made 1000 first serves in last year's matches. Among these serves, 750 were successful.

Question: what is the probability that his next first serve will be successful?

Assuming: nothing has changed, a reasonable answer is 0.75 or 75%.

What are the implications of this assessment?

Prediction: If the tennis player makes 40 new serves, about 30 would be expected to be successful.

1. Prediction – A simulation in R

This course uses R, which is very common in data science for sophisticated statistics. Here is an R simulation of 40 first serves:

```r
# Simulate 40 first serves, save the results in y
# Show the results
(y <- sample(x = c("S", "F"), size = 40, replace = T,
    prob = c(0.75, 0.25)))

##  [1] "S" "S" "F" "S" "S" "S" "S" "F" "F" "S" "S"
## [12] "S" "F" "S" "S" "F" "S" "F" "S" "S" "S" "S"
## [23] "S" "S" "S" "S" "S" "S" "S" "S" "S" "S" "S"
## [34] "F" "S" "S" "S" "F" "S" "S"

table(y)

## y
##  F  S
##  8 32
```

1. Prediction – Another simulation

```r
# Another sample Not the same total or order
(y <- sample(x = c("S", "F"), size = 40, replace = T,
    prob = c(0.75, 0.25)))

##  [1] "S" "S" "S" "S" "F" "S" "F" "S" "S" "S" "S"
## [12] "S" "S" "S" "S" "S" "S" "S" "S" "S" "S" "S"
## [23] "S" "S" "S" "S" "S" "F" "S" "S" "S" "S" "S"
## [34] "S" "F" "S" "S" "S" "S" "F"

table(y)

## y
##  F  S
##  5 35
```

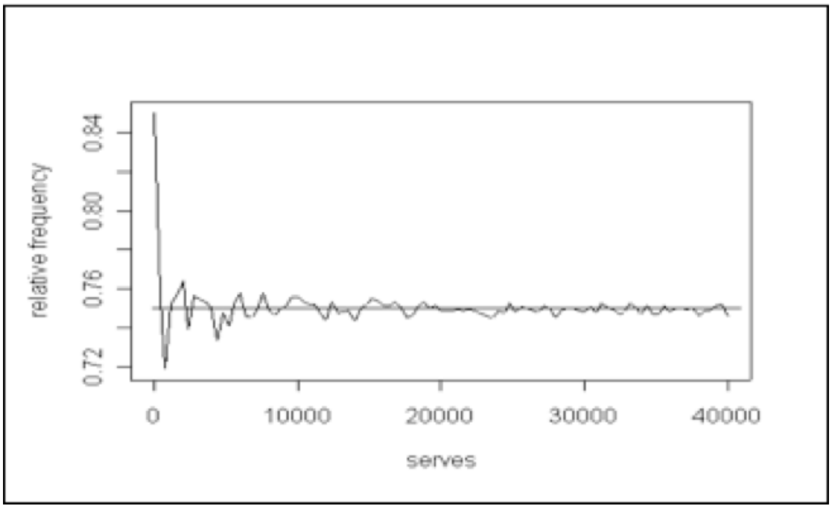


Figure 4: 40,000 first serves simulated — relative frequency approaches 0.75

1. Simulating with more and more first serves

1. Relative frequency morals

Relative Frequencies stabilise around the probability in a large number of repeated trials.

Law of Large Numbers: remember Tobias Mayer, the first data scientist.

2.3 Method 2 — Symmetry

2. Symmetry or Equally Likely Outcomes

Die is rolled — it is uniform in manufacture, we'd expect the labels 1 to 6 to all have equal chance of coming up.

Two Coins: If we throw two coins at the same time, there are four possible outcomes:

1. Heads first, Tails second: HT,
2. TH,
3. HH,
4. TT.

And: they might be assumed to be equally likely in the absence of evidence to the contrary.

Probabilities are the ratio of the number of outcomes in an event to the total number of outcomes,

So counting is needed — discussed in Section 1.2.

2.4 Method 3 — Subjective Belief

3. Subjective Belief

I think the sun will come up tomorrow (based on experience and understanding of physical models for the movement of the sun), so I decide to assign probability one to this.

My investment outlook is negative so I think it is more likely than not that the ASX has gone down since 1 pm today.

How much would I bet on this?

Leads to quantitative assignment of probabilities if this is a "fair" game using expectations based on probabilities — will discuss later in the course.

2.5 Definitions

Definitions

Experiment: a controlled process for producing an outcome of interest.

Observation: no control — common in business.

Trial: performing an experiment once.

Outcome: an observed result of an experiment.

Random experiment: an experiment in which the outcome cannot be predicted with certainty.

Similarly for a random observation.

Definition of Sample Space & Examples

The Sample space, or Outcome space, denoted by S, is the set of all possible outcomes of an experiment.

Examples Write down the outcome space

1. In the examples for assigning probabilities by symmetry: **Sol:** roll a die: $S = \{1, 2, 3, 4, 5, 6\}$ toss 2 coins: $S = \{(H, T), (T, H), (H, H), (T, T)\}$.
2. Wait until there is a head: **Sol:** $S = \{H, (T, H), (T, T, H), \ldots\}$
3. The lifetime of a light bulb: **Sol:** $S = [0, \infty)$ (0 is possible!)
4. Observe the longitude, latitude and the depth of the epicentre of an earthquake: **Sol:** $S = [0, 360) \times [0, 360) \times [0, 6371)$ where long'de and lat'de are in degrees and depth in km.

Definition of Events

Event: A subset, A, of the outcomes in the sample space, S, is called an event. The Venn diagram in Figure 5 illustrates this.

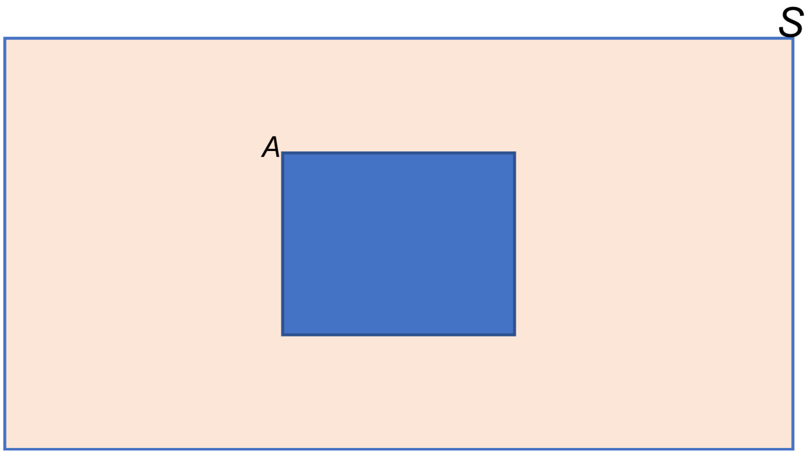


Figure 5: A coloured in blue

Example

There are 4 white marbles and 2 black marbles in a bag. Two are drawn at random.

- What are the outcomes in the Sample Space? **Sol:** Depends on whether order is recorded, whether labelled and whether replaced. Assume ordered, numbered but not replaced. Label Whites 1 to 4 and Blacks 5 and 6. Then sample space is $S = \{(i,j) : i,j = 1,2,\ldots,6, i \neq j\}$. There are 30 outcomes in total.

- What are the events – called elementary events – that correspond to each outcome? **Sol:** $A_{(i,j)} = \{(i,j)\}, i,j = 1,2,\ldots,6, i \neq j$ — each elementary event is a set with one ordered pair in it.

Example Ctd.

- What is the event, B, that the second draw is a black? **Sol:** $B = \{(i,j) : i = 1,2,\ldots,6, j = 5,6, i \neq j\}$ — there are 10 outcomes in this event.

- C, that the first draw is a white? **Sol:** $C = \{(i,j) : i = 1,2,\ldots,4, j = 1,2,\ldots,6, i \neq j\}$ — there are 20 outcomes in this event.

- What assumptions would you make to assign probabilities? **Sol:** Equally likely outcomes in the absence of information to the contrary — assumes bag is well shaken before each draw and person who draws is blindfolded.

2.6 Operations on Events

Union of Events

Union: If A, B are events, then $A \cup B$ is the event with all of the outcomes of both events, i.e., in *either A or B* or *both*. Figure 6 illustrates this.

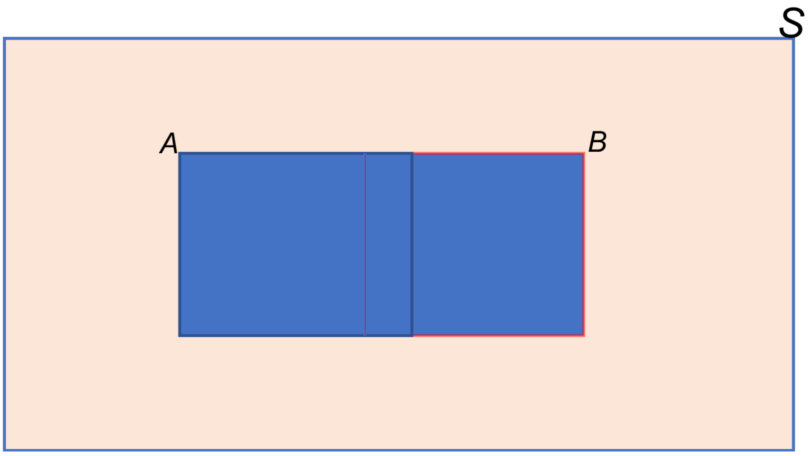


Figure 6: $A \cup B$ coloured in blue

Intersection of Events

Intersection: If A, B are events, then $A \cap B$ is the event with outcomes that are in *both* sets. Figure 7 illustrates this.

Complement of An Event

Complement: If A is an event, then A^c is the event with all outcomes which are *not* in A. Figure 8 illustrates this.

Disjoint Events

Disjoint, or *mutually exclusive*, events have no outcomes in common. Figure 9 illustrates this.

Laws for Events

Commutative Union: $A \cup B = B \cup A$

Commutative Intersection: $A \cap B = B \cap A$

Associative Union: $A \cup (B \cup C) = (A \cup B) \cup C$

Associative Intersection: $A \cap (B \cap C) = (A \cap B) \cap C$

Distributive Union: $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

Distributive Intersection: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

De Morgan's Laws 1: $(A \cap B)^c = A^c \cup B^c$

De Morgan's Laws 2: $(A \cup B)^c = A^c \cap B^c$

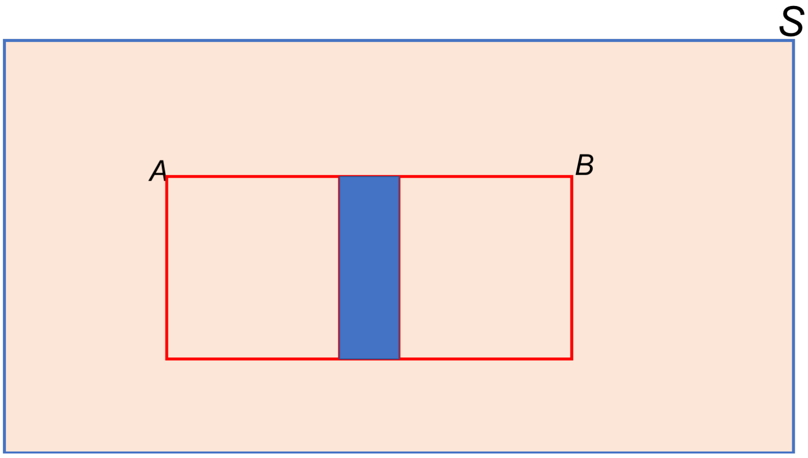


Figure 7: $A \cap B$ coloured in blue

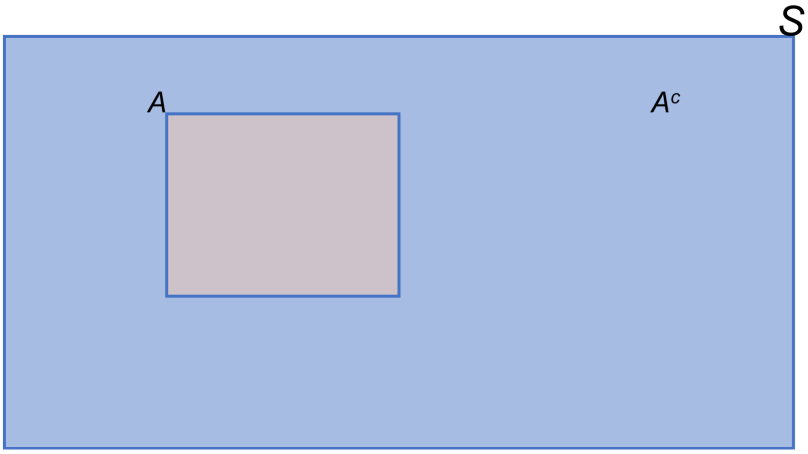


Figure 8: A^c coloured in blue

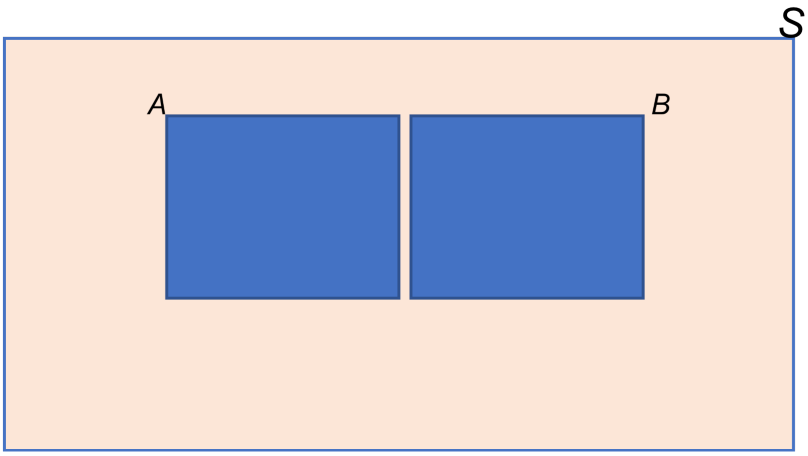


Figure 9: A and B disjoint

2.7 Rules of probability

Rules of Probability

Probability, P, must obey some basic rules or axioms:

1. For any event A, $P(A) \geq 0$.
2. The probability of the sample space is 1, $P(S) = 1$.
3. For disjoint events $A, B, C \ldots$, $P(A \cup B \cup C \cup \ldots) = P(A) + P(B) + P(C) + \ldots$.

Two disjoint or mutually exclusive events

2.8 Consequences

Consequences of the rules of probability

For any events A, B, C:

$$\text{A.} \qquad \boxed{P(A^c) = 1 - P(A)}$$

because by rules 2 & 3, and the facts that $A \cap A^c = \emptyset$ & $A \cup A^c = S$, it follows that

$$\begin{aligned} 1 &= P(S) \\ &= P(A \cup A^c) \\ &= P(A) + P(A^c) \end{aligned}$$

giving the required identity on rearranging.

$$\text{B.} \qquad \boxed{P(\emptyset) = 0}$$

from A. and rule 2 since $S^c = \emptyset$.

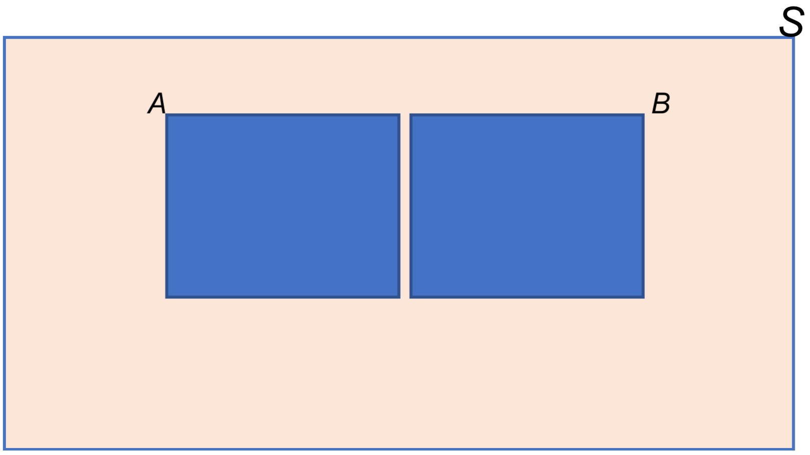


Figure 10: $P(A \cup B) = P(A) + P(B)$

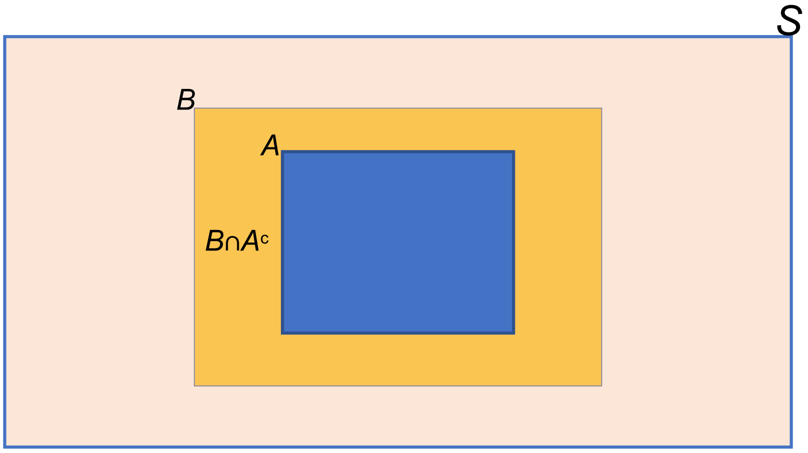


Figure 11: $A \subseteq B$

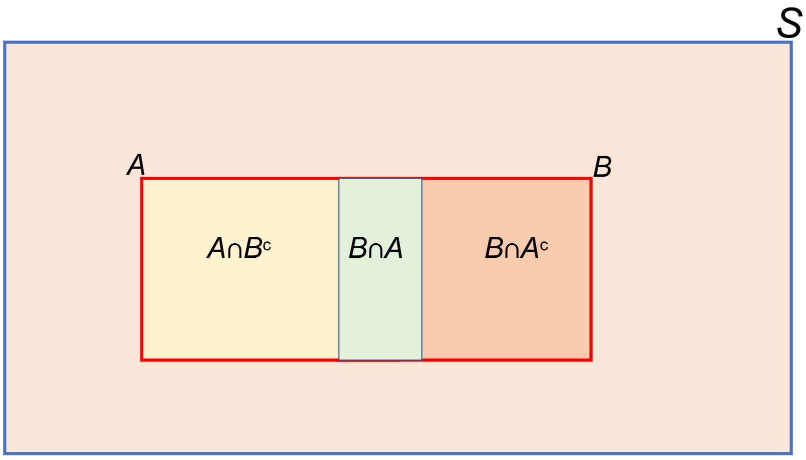


Figure 12: $A \cup B$ divided into $A \cap B^c$, $A \cap B$, $A^c \cap B$

A is a subset of B

Consequences of the rules of probability 2

$$\text{C.} \qquad \boxed{\text{If } A \subseteq B, \text{ then } P(A) \leq P(B)}$$

because by rule 3, and the facts that $B \cap (A \cap B^c) = \emptyset$ & $B \cup (A \cap B^c) = B$ (see Figure 11), it follows that

$$\begin{aligned} P(B) &= P(A \cup (B \cap A^c)) \\ &= P(A) + P(B \cap A^c), \end{aligned}$$

and, by rule 1 $P(B \cap A^c) \geq 0$, giving the required inequality.

$$\text{D.} \qquad \boxed{P(A) \leq 1}$$

from C. and rule 2, since $A \subseteq S$.

A union B in general

Consequences of the rules of probability 3

$$\text{E.} \qquad \boxed{P(A \cup B) = P(A) + P(B) - P(A \cap B)}$$

because by rule 3, and the facts that $(A \cap B^c) \cap (A \cap B) \cap (A^c \cap B) = \emptyset$ & $(B \cap A^c) \cup (A \cap B) \cup (A^c \cap B) = A \cup B$ (see Figure 12), it follows from rule 3 that

$$\begin{aligned} P(A \cup B) &= P(A \cap B^c) + P(A \cap B) + P(A^c \cap B) \\ &= P(A) + P(A^c \cap B), \\ &= P(A) + P(A \cap B) + P(A^c \cap B) - P(A \cap B) \\ &= P(A) + P(B) - P(A \cap B) \end{aligned}$$

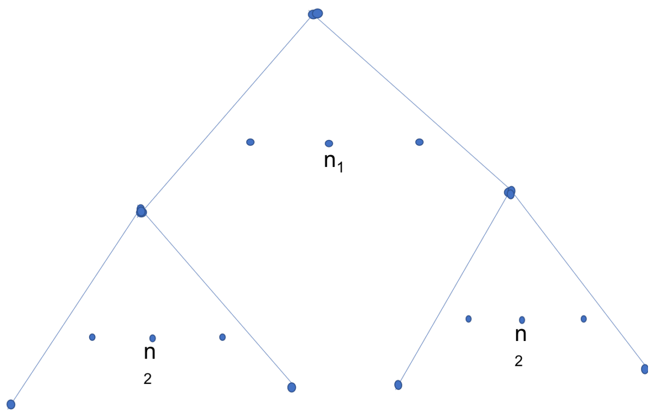


Figure 13: Tree with n_1 nodes at first level, each with n_2 nodes at second level

the steps following from rule $(A \cap B^c) \cap (A \cap B) = \emptyset$ & $(A \cap B^c) \cup (A \cap B) = A$ and similarly for B.

Consequences of the rules of probability 4

$$F. \qquad \begin{aligned} P(A \cup B \cup C) = P(A) &+ P(B) + P(C) \\ &- P(A \cap B) - P(B \cap C) - P(A \cap C) \\ &+ P(A \cap B \cap C) \end{aligned}$$

which follows from applying E. to $A \& B \cup C$, $\quad B \& C \quad$ and $A \cap B \& A \cap C$ (also applying the distributive intersection rule for events).

3 Methods of Enumeration — 1.2

3.1 How to count?

Two basic rules for counting: Multiplication principle

If one choice has n_1 possibilities and another has n_2 possibilities, then a choice of both has $n_1 \times n_2$ possibilities (see Figure 13).

Two basic rules for counting: Addition principle

If one event has n_1 outcomes and a disjoint one has n_2 outcomes, then the union has $n_1 + n_2$ outcomes (see Figure 14).

Permutations

The number of *permutations* or *orderings* or *vectors without replacement* of length r from a set of size n is

$$\begin{aligned} {}_nP_r &= n(n-1)\dots(n-r+1) \\ &= \frac{n(n-1)\dots 1}{(n-r)(n-r-1)\dots 1} \\ &= \frac{n!}{(n-r)!}. \end{aligned}$$

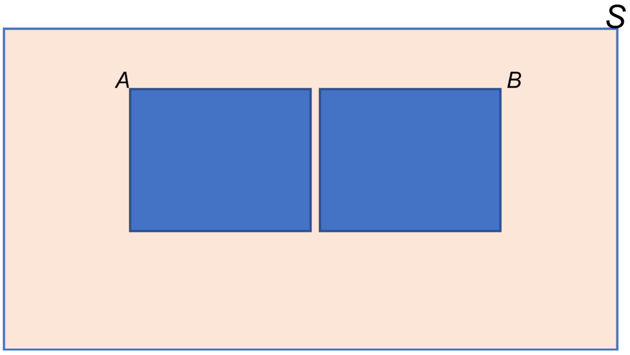


Figure 14: number$(A \cup B)$ = number(A) + number(B)

because any ordering $(i_1, i_2, \dots i_r)$ has n choices for i_1, $(n-1)$ for $i_2 \dots (n-r+1)$ for the rth. So the multiplication principle applied r times gives the first line. The third line relates the answer to factorials denoted !. Note that the choices for i_2 are different depending on which i_1 has been chosen - they are all choices except i_1. However, the multiplication principle still applies because there are always $n-1$ choices for i_2.

Combinations

The number of *combinations* or *samples without replacement* or *subsets* of size r from a set of size n is

$$\begin{aligned} {}_nC_r &= \frac{(n-1)\dots(n-r+1)}{r(r-1)\dots 1} \\ &= \frac{n(n-1)\dots 1}{r(r-1)\dots 1 \times (n-r)(n-r-1)\dots 1} \\ &= \frac{n!}{r!(n-r)!} \\ &= \binom{n}{r}. \end{aligned}$$

The last equality is the definition of the bracket and this number ${}_nC_r$ is often called a *Binomial Coefficient*.

Combinations — reasons

- If the subset was ordered then it would be a ordering of size r from the set of size n, and there are ${}_nP_r$ of these.
- An ordering can be chosen by first choosing a subset size r from the set of size n and then ordering that subset of size r.
- There are ${}_rP_r = r!$ such orderings.
- Hence, by the multiplication principle:

$$_nP_r =_n C_r \times r!$$

which gives the result on rearrangement.

3.2 Example: Sampling without Replacement

Example: Sampling without Replacement

Suppose that 10000 lines of code have 30 lines which could be improved (for example by clarification of purpose, simplification or documentation). If a random sample of 10 lines of code is taken, what is the chance that:
(a) 1 or
(b) 0
lines of the code sampled could be improved?

Example Solution

- The sample space has all the samples of size 10 taken without replacement is $\binom{10,000}{10}$.
- The 1 line of code in the sample that could be improved must be chosen from the 30 improvable lines. There are $\binom{30}{1} = 30$ such choices.
- The other 9 lines of code in the sample must be chosen from the $10,000 - 30 = 9,970$ lines of code that cannot be improved. There are $\binom{9,970}{9}$ ways to do that.
- The multiplication principle says that the total number of samples with one improvable line is $\binom{30}{1} \times \binom{9,970}{9}$.
- Random sampling without replacement means that all the samples are equally likely so the answer is

$$\frac{\binom{30}{1} \times \binom{9,970}{9}}{\binom{10,000}{10}}$$

Example Solution Ctd 1

There are very big numbers in this calculation but modern symbolic packages compute them and the ratio exactly. The next two slides show the calculations for (a) and (b) using the package Mathematica. They show that the probability of one line in the sample needing improvement is about 0.03.

Example Solution Ctd 2

This is the Mathematica command to find the number of sample points:

```
Binomial[10000, 10]
```

This is the output of it:

```
2, 743, 355, 077, 591, 282, 538, 231, 819, 720, 749, 000
```

This calculates the probability that one line of code in the sample could be improved:

```
Binomial[30, 1] * Binomial[10000 − 30, 9]/Binomial[10000, 10]
```

Here is the output:

$$\frac{2143751464028247883152007617}{73351740042547661450048655635}$$

This asks to evaluate the fraction as a decimal: $\boldsymbol{N}\left[\frac{\mathbf{2143751464028247883152007617}}{\mathbf{73351740042547661450048655635}}\right]$

Here is the result:

```
0.0292256
```

Example Solution Ctd 3

Now for the probability that there are no lines of code that could be improved, Input and Output:

Binomial[30, 0] ∗ Binomial[10000 − 30, 10]/Binomial[10000, 10]

$$\frac{1016852777770732245908435612997}{1047882000607823735000695080500}$$

$$\boldsymbol{N}\left[\frac{\mathbf{1016852777770732245908435612997}}{\mathbf{1047882000607823735000695080500}}\right]$$

0.970389

Example Solution Ctd 4

Same calculations in R — note answers agree with Mathematica to the same number of figures:

```r
# prob1 is the probability of 1 line
(den <- choose(10000, 10))

## [1] 2.743355e+33

num1 <- choose(30, 1) * choose(10000 - 30, 9)
(prob1 <- num1/den)

## [1] 0.02922564

# prob0 is the probability of 0 lines
num0 <- choose(30, 0) * choose(10000 - 30, 10)
(prob0 <- num0/den)

## [1] 0.9703886
```

Ex. Sampling W'out Replacement — Comments, Questions

- There is a chance of 0.9995 that there is at most one line of code that could be improved.
- Do we really have to work out those big numbers? Approximations?
- What happens with sampling with replacement? Comparison?
- In practice, we'd observe the number of lines of code that could be improved in a sample, and want to infer from that how good the code is in total
- To do this we could use the probabilities calculated to make an inference about the plausibility of 30 in total
- For example, if we observed 5 lines of code in our sample of 10 that could be improved, then 30 in total is implausible

4 Conditional Prob.— 1.3

4.1 Bayesian and Frequentist Statistics

Conditional Probability

- Continuing the last example, given an observed number of 5 lines of code in our sample, the probability that the total number of lines of code among the 10,000 is 30 or 50 or 100 ... would help us decide the quality of the code?
- This is the approach of Bayesian statistics — use the data to make a conclusion about the true number of lines of code that could be improved.

Conditional Probability

- The frequentist approach, on the other hand, would be to start with the *Assumption* that there were 30 or 50 or 100 lines of code that could be improved, and then to compute the probability of the observed 5 lines of improvable code in the sample
- This probablity could then be plotted against the total number of lines, and the maximum picked out — this is called the *maximum likelihood estimate.*
- Conditional probabilities are the heart of the Bayesian approach, but are also very important for frequentists.

4.2 Definition

Conditional Probability
Definition. The conditional probability of an event A, given that event B has occurred, is defined by

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

provided that $P(B) > 0$. Effectively, we replace the whole sample space by the event B and so the probability of A now only refers to the part of A in common with B, i.e., $A \cap B$. Further, the probabilities need to be scaled up by the reciprocal of $P(B)$ to preserve the probability of the whole space is 1, i.e., $P(B|B) = 1$.

4.3 Example: Conditional Probability Definition

Example : Conditional Prob. Definition
Two cards are drawn at random in order without replacement from a deck of cards. (a) What is the conditional probability of getting an ace on the second draw given that the first draw is an ace? (b) What is the conditional probability of getting an ace on the first draw given that the second draw is an ace?

Example Solution

The sample space $S = \{(i,j) : i,j \in \{1,\ldots,52\}, i \neq j\}$, and equally likely outcomes are assumed. Let A_1, A_2 be the events that the first or second draw is an ace. Using the multiplication principle and letting n stand for number,

$$\begin{aligned} P(A_1) &= \frac{n(A_1)}{n(S)}, \text{ so} \\ P(A_1) &= \frac{4 \times 51}{52 \times 51} \\ P(A_2) &= \frac{n(A_2)}{n(S)}, \text{ so} \\ P(A_2) &= \frac{4 \times 3 + 48 \times 4}{52 \times 51} \\ P(A_1 \cap A_2) &= \frac{n(A_1 \cap A_2)}{n(S)}, \quad \text{so} \quad P(A_1 \cap A_2) = \frac{4 \times 3}{52 \times 51} \end{aligned}$$

Example Solution Ctd

Answering (a)

$$P(A_2|A_1) = \frac{P(A_1 \cap A_2)}{P(A_1)}$$

which gives

$$P(A_2|A_1) = \frac{4 \times 3}{4 \times 51} = \frac{3}{51}$$

and

$$P(A_1|A_2) = \frac{P(A_1 \cap A_2)}{P(A_2)},$$

so

$$P(A_1|A_2) = \frac{4 \times 3}{4 \times 51} = \frac{3}{51}$$

Conditional Probability Example: Comment

- The answer to (a) is not perhaps surprising since there are 51 cards left after the first choice of which 3 are aces given that the first drawer was an ace.
- But why should (b) have the same answer?
- A thought experiment leads to the idea of symmetry of events, in this case called *exchangeability.* (Think about recording the order of the draws, but then randomly shuffling the order — how could this change the probabilities?)

Rules of Conditional Probability

Conditional probability satisfies the same rules as ordinary probability. For unions, the conditioning event needs to be the same. For example, if A_1 and A_2 are disjoint events and $P(B) > 0$,

$$P(A_1 \cup A_2|B) = P(A_1|B) + P(A_2|B).$$

4.4 Multiplication Rule

Multiplication Rule

The definition of conditional probability can be re-expressed as:

$$P(A \cap B) = P(A|B) \times P(B)$$

or equally

$$\boxed{P(A \cap B) = P(B|A) \times P(A).} \tag{1}$$

In this form, it enables us to work out the probability of two events occuring together, if we know – or make assumptions – about the probability of one event and the conditional probability of the other event given the first. It can be natural to make assumptions about conditional probabilities for events that occur sequentially in time, as the next example shows.

4.5 Example: System Failure

Example: System Failure

A device has two components. It will operate if at least one of the two components is operating. The probability that one component will fail when both are working in a one-year period is 0.01. However, when one fails, the probability of the other failing is 0.03 in that one-year period due to added strain. The two components cannot fail at the same time. What is the probability that the device fails during a one-year period?

Solution — System Failure

Let A_1 (A_2) be the event that the first (respectively second) component fails first in the given year and B be the event that both components fail in the year. Then, because $B = (B \cap A_1) \cup (B \cap A_2)$ and the events $B \cap A_1$, $B \cap A_2$ are disjoint,

$$P(B) = P(B \cap A_1) + P(B \cap A_2).$$

Using the Multiplication Rule, Equation 1, for Conditional Probabilities on each component

$$\begin{aligned} P(B) &= P(B|A_1) \times P(A_1) + P(B|A_2) \times P(A_2) \\ &= 0.03 \times 0.01 + 0.03 \times 0.01 \\ &= 0.0006 \end{aligned}$$

5 Bayes' Theorem — 1.5

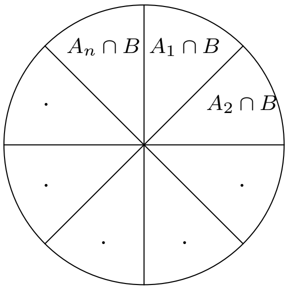
5.1 Law of Total Probability

The Law of Total Probability

Assumption 1 Suppose $A_1, A_2, \ldots, A_n$ ($n = 1, 2, \ldots$) are disjoint events each with non-zero probability.

Assumption 2 Suppose the event B is the union of $B \cap A_1, B \cap A_2, \ldots, B \cap A_n$ –i.e., the event B is partitioned into disjoint bits by the events $A_1, A_2, \ldots, A_n$.

Venn Diagram for B:



The Law of Total Probability

Assumption 1 Suppose $A_1, A_2, \ldots, A_n$ $(n = 1, 2, \ldots)$ are disjoint events each with non-zero probability.

Assumption 2 Suppose the event B is the union of $B \cap A_1, B \cap A_2 \ldots B \cap A_n$ –ie the event B is partitioned into disjoint bits by the events $A_1, A_2 \ldots A_n$.

Rule (c) for Probability:

$$P(B) = P(B \cap A_1) + P(B \cap A_2) + \ldots + P(B \cap A_n)$$

Multiplication Rule 1 gives *the Law of Total Probability*:

$$\boxed{P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \ldots + P(B|A_n)P(A_n)} \qquad (2)$$

5.2 Bayes' Theorem

The Law of Total Probability to Bayes Theorem

System Failure Ex. *Assumptions* were the probabilities $P(B|A_i), P(A_i), i = 1$ or 2, and could then find $P(B)$

Sampling Ex. *Assumptions* were the total number of lines of code and the total number that could be improved.

So could calculate probabilities for the number of lines of code *in a random sample* that could be improved.

But to *infer* from an *observed* number in the sample to the **total** number of lines of improvable code, the *reverse* conditional probabilities are needed.

Enter the Reverend Thomas Bayes who had a remarkable posthumous essay published in 1764 containing (a specific case of) *Bayes Theorem.*

Figure 15: 1702 – 1761

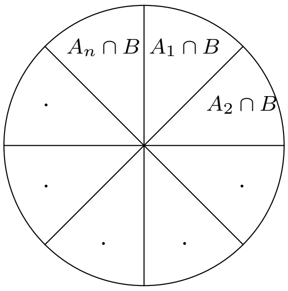
The Reverend Thomas Bayes,1702 – 1761

Figure 15 shows Bayes. The Theorem enables us to draw conclusions about the truth from data.

Bayes Theorem — Assumptions as Before

Assumption 1 Suppose $A_1, A_2, \ldots, A_n$ $(n = 1, 2, \ldots)$ are disjoint events each with non-zero probability.

Assumption 2 Suppose the event B is the union of $B \cap A_1, B \cap A_2, \ldots, B \cap A_n$ –i.e., the event B is partitioned into disjoint bits by the events $A_1, A_2, \ldots, A_n$.



Bayes Theorem

Assumption 1 Suppose $A_1, A_2, \ldots, A_n$ $(n = 1, 2, \ldots)$ are disjoint events each with non-zero probability.

Assumption 2 Suppose the event B is the union of $B \cap A_1, B \cap A_2, \ldots, B \cap A_n$ –i.e., the event B is partitioned into disjoint bits by the events $A_1, A_2, \ldots, A_n$.

Multiplication Rule (Equation (1)) gives for any i in $1, \ldots, n$

$$P(B \cap A_i) = P(A_i|B)P(B)$$

Law of Total Probability used on $P(B)$ then gives *Bayes Theorem*:

$$\boxed{P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \ldots + P(B|A_n)P(A_n)}}. \quad (3)$$

Understanding and Using Bayes Theorem

Denominator: is a weighted average of the probabilities $P(B|A_j)$ with weights $P(A_j), j = 1, \ldots, n$.

Numerator: is the ith term in the weighted average.

Reverse conditional probability: is the ratio of this term to the weighted average.

Use in statistics: B is usually the observed data and $A_1, \ldots, A_n$ are the possible true states or underlying hypotheses.

Illustrated: in the next example.

5.3 Bayes Example: Sampling w'out Replacement

Bayes Example: Sampling w'out Replacement

Suppose that the standard for acceptable code is that at most 10% can be improved. Past experience suggests that 90% of blocks of 1000 lines of code meet the standard, with 10% failing to meet the standard. A random sample of size 20 is taken from 1000 lines of code and, on inspection, it is observed that 5 of the 20 lines of code can be improved. Does it seem plausible that the standard has been met?

Bayes Ex. — Sampl. w'out Repl't: Solution

Definitions Let A_i be the event that, in total, i ($i = 0, 1, \ldots, 1000$) lines of code can be improved and B be the *event* that 5 lines of code could be improved in the random sample of 20 lines.

Standard? For the standard to be met, one of the events $A_0, \ldots, A_{100}$ must be true.

Sampling Ex. means we can work out the probabilities $P(B|A_i)$ because, *given* the total number of lines of code that can be improved,

the combinations calculations gave us:

$$P(B|A_i) = \begin{cases} \frac{\binom{i}{5}\binom{1000-i}{15}}{\binom{1000}{20}}, & i = 5, \ldots, 985 \\ 0, & i = 0, \ldots, 4, 986, \ldots, 1000 \end{cases} \quad (4)$$

Bayes Ex. — Solution Ctd

What about $P(A_i)$ $(i = 0, 1, \ldots, 1000)$?

We haven't been told in detail, so we need to make some assumptions.

We have been told that 90% meet the standard.

So, since the event that the code meets the standard is the union of the disjoint events $A_0, A_1, \ldots, A_{100}$,

Rule (c) gives:

$$0.9 = P(A_0) + P(A_1) + \ldots + P(A_{100}) \tag{5}$$

Simplest assumption is that the events $A_0, A_1, \ldots, A_{100}$ are equally likely.

So we assume

$$P(A_0) = P(A_1) = \ldots = P(A_{100}) = 0.9/101 = 0.00891 \tag{6}$$

Bayes Ex. — Solution Ctd 2

Simlarly we assume

$$P(A_{101}) = P(A_{102}) = \ldots = P(A_{1000}) = 0.1/900 = 0.00011 \tag{7}$$

Bayes Theorem (3) gives

$$P(A_0|B) = P(A_1|B) = \ldots = P(A_4|B) = 0 \tag{8}$$

since $P(B|A_i) = 0, i = 0, 1, 2, 3, 4$ (the number of lines in total that can be improved must be at least the number in the sample that can be improved!),

and

$$P(A_5|B) = \frac{P(B|A_5)P(A_5)}{P(B|A_0)P(A_0)\ldots + P(B|A_{1000})P(A_{1000})} \tag{9}$$

Bayes Ex. — Solution Ctd 3

Combining (4) to (7) in (8) gives

$$P(A_5|B) = \frac{\frac{\binom{5}{5}\binom{995}{15}}{\binom{1000}{20}} \times \frac{0.9}{101}}{5 \times 0 + \frac{\binom{5}{5}\binom{995}{15}}{\binom{1000}{20}} \times \frac{0.9}{101} + \ldots + \frac{\binom{985}{5}\binom{15}{15}}{\binom{1000}{20}} \times \frac{0.1}{900} + 15 \times 0} \tag{10}$$

The answer to this is small because only fluke could give us all five lines of code in our sample of 20 if there are only 5 in the 1000 that could be improved.

But notice that (4) simplifies: for $5 \leq i \leq 985$

$$\frac{\binom{i}{5}\binom{1000-i}{15}}{\binom{1000}{20}} = \binom{20}{5} \times \frac{i}{1000} \times \cdots \times \frac{i-4}{996} \times \frac{1000-i}{995} \times \cdots \times \frac{1000-i-14}{981}. \tag{11}$$

Sampling with or without replacement

So if i is small relative to 1000

$$\frac{\binom{i}{5}\binom{1000-i}{15}}{\binom{1000}{20}} \approx \binom{20}{5} \times \left(\frac{i}{1000}\right)^5 \times \left(1 - \frac{i}{1000}\right)^{15}. \tag{12}$$

Which is the probability that we'll see later applies when there is sampling with replacement.

The reason is that whatever happens during the first few selections of lines of code in the sample does not change the probabilities much.

This leads to the idea of *independent events.*

6 Independence — 1.4

6.1 Definition

Independent Events

Events are *independent* if the occurrence of any combination of them (or not) does not change the probabilities of the others.

So if A and B are independent, then $P(A|B) = P(A)$, $P(B|A) = P(B)$, $P(B|A^c) = P(B)$ and $P(A|B^c) = P(A)$.

The multiplication rule shows that, if A and B are independent, then

$$P(A \cap B) = P(A)P(B). \tag{13}$$

A proof shows that (13) is enough to ensure independence of two events

For n events to be independent, we need that for any choices of the events $A_i, A_j, \ldots, A_z$

$$P(A_i \cap A_j \cap \cdots \cap A_z) = P(A_i)P(A_j)\cdots P(A_z). \tag{14}$$

6.2 Example: Rocket Failure

Example: Rocket Failure

A rocket has a built-in redundancy system. In this system, if component 1 fails, it is bypassed and component 2 is used. If component 2 fails, it is bypassed and component 3 is used. Suppose that the probability of failure of any one of these components is 0.15 and assume that the failures of these components are mutually independent events. What is the probability that the rocket does not fail?

Solution — Rocket Failure

Definition Let A_i denote the event that component $i, i = 1, 2, 3$ fails.

Because the system fails precisely if all components fail, the probability that the system does *not* fail is given by

$$P((A_1 \cap A_2 \cap A_3)^c) = 1 - P(A_1 \cap A_2 \cap A_3).$$

Independence gives the right hand side as

$$1 - P(A_1)P(A_2)P(A_3).$$

So the probability that the rocket does not fail is $1 - 0.15^3 = 0.9966$.

Comments — Rocket Failure

The desired probability is the probability that at least one of the components does not fail.

This is the probability of the union of three events.

Using the rule in Consequence F involves

- the individual probabilities,
- the probabilities of the intersections two at a time and
- the probability of the intersection of all three.

This is a lot more complicated than our calculation above.

Comments — Rocket Failure

Even though each component has a significant chance of 15% of failing, the system as a whole is unlikely to fail because of the design with 3 components in parallel.

The calculation relies heavily on the assumption of independence.

If instead there was a chance that all three components could be part of a faulty batch, the conditional probability of the second component failing could be much higher than the 15% — the multiplication rule could then make the probability of all three failing much higher than 0.15^3.

This is similar to the system failure example when the failure of one component led to a higher probability that the other one would fail.

6.3 Example: Pairwise Independence — Soft Drinks

Example: Soft drinks

Two companies producing soft drinks, P and C, compete for shelf space in 3 supermarket chains, B, D and X. At a particular time, competition is so fierce that the possibilities are equally likely that each company has more shelf space in each supermarket chain.

1. **What is an appropriate sample space?**
2. Consider the events P_B,P_D, P_X that P has more shelf space in supermarket chains, B, D, X (respectively). Are these events independent?
3. What about the events BD, DX, BX that chains
   (a) B and D,
   (b) D and X,
   (c) B and X (respectively)

   allocate more shelf space for the *same* soft drink?

Solution — Soft drinks-1

An appropriate sample space has sample points which are ordered triples recording the soft drink that has more shelf space in supermarket chains B, D and X in this order. Therefore,

$$\begin{aligned} S = \{&(P,P,P),(P,P,C),(P,C,P),(P,C,C),\\ &(C,P,P),(C,P,C),(C,C,P),(C,C,C)\}. \end{aligned}$$

For example, the sample point (P,C,P) corresponds to the outcome that P has more shelf space in supermarket B, C in D and P in X. Notice that there are $2 \times 2 \times 2 = 8$ outcomes in the sample space.

Example: Soft drinks

Two companies producing soft drinks, P and C, compete for shelf space in 3 supermarket chains, B, D and X. At one time, competition is so fierce that the possibilities are equally likely for which company has more shelf space in which supermarket chain.

1. What is an appropriate sample space?
2. **Consider the events P_B,P_D, P_X that P has more shelf space in supermarket chains, B, D, X (respectively). Are these events independent?**
3. What about the events BD, DX, BX that chains
   (a) B and D,
   (b) D and X,
   (c) B and X (respectively)

   allocate more shelf space for the *same* soft drink?

Solution — Soft drinks-2

The event
$$P_B = \{(P,P,P),(P,P,C),(P,C,P),(P,C,C)\}$$
so
$$P(P_B) = \frac{4}{8} = \frac{1}{2}.$$

Write out the outcomes in P_D and P_X to convince yourself that these events also have probability one half.

The event
$$P_B \cap P_D = \{(P,P,P),(P,P,C)\}$$
so
$$P(P_B \cap P_D) = \frac{2}{8} = \frac{1}{2} \times \frac{1}{2} = P(P_B) \times P(P_D)$$

Solution — Soft drinks-2 ctd

Hence, P_B and P_D are independent. Check for yourself that the pairs P_D and P_X, as well as P_B and P_X, are also independent.

Finally,
$$P_B \cap P_D \cap P_X = \{(P,P,P)\}$$
so
$$\begin{aligned} P(P_B \cap P_D \cap P_X) &= \frac{1}{8} \\ &= \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \\ &= P(P_B) \times P(P_D) \times P(P_X) \end{aligned}$$

Hence, the events P_B, P_D, P_X are independent.

Example: Soft drinks

Two companies producing soft drinks, P and C, compete for shelf space in 3 supermarket chains, B, D and X. At one time, competition is so fierce that the possibilities are equally likely for which company has more shelf space in which supermarket chain.

1. What is an appropriate sample space?
2. Consider the events P_B,P_D, P_X that P has more shelf space in supermarket chains, B, D, X (respectively). Are these events independent?
3. **What about the events BD, DX, BX that chains**

   (a) B and D,

   (b) D and X,

   (c) B and X (respectively)

   allocate more shelf space for the *same* soft drink?

Solution — Soft drinks-3

The event
$$BD = \{(P,P,P),(P,P,C),(C,C,P),(C,C,C)\}$$
because either B and D supermarkets allocate more shelf space for P — in which case X can either have allocated more shelf space for P or C — or they allocate more shelf space for C and X supermarket is again unconstrained.

so
$$P(BD) = \frac{4}{8} = \frac{1}{2}.$$

Solution — Soft drinks-3

Write out the outcomes in DX and BX to convince yourself that these events also have probability one half.

The event
$$BD \cap BX = \{(P,P,P),(C,C,C)\}$$

so
$$P(BD \cap BX) = \frac{2}{8} = \frac{1}{2} \times \frac{1}{2} = P(BD) \times P(BX)$$

Solution — Soft drinks-3 ctd

Hence, BD and BX are independent. Check for yourself that the pairs BD and DX, as well as DX and BX, are also independent.

Finally,
$$BD \cap DX \cap BX = \{(P,P,P),(C,C,C)\} = \quad BD \cap BX \quad !!$$

so
$$\begin{aligned} P(BD \cap DX \cap BX) &= \frac{1}{4} \\ &\neq P(BD) \times P(DX) \times P(BX) \end{aligned}$$

Hence, the events BD, DX, BX are *not* independent.

Knowing BD and BX occurred tells us that DX also occurred.