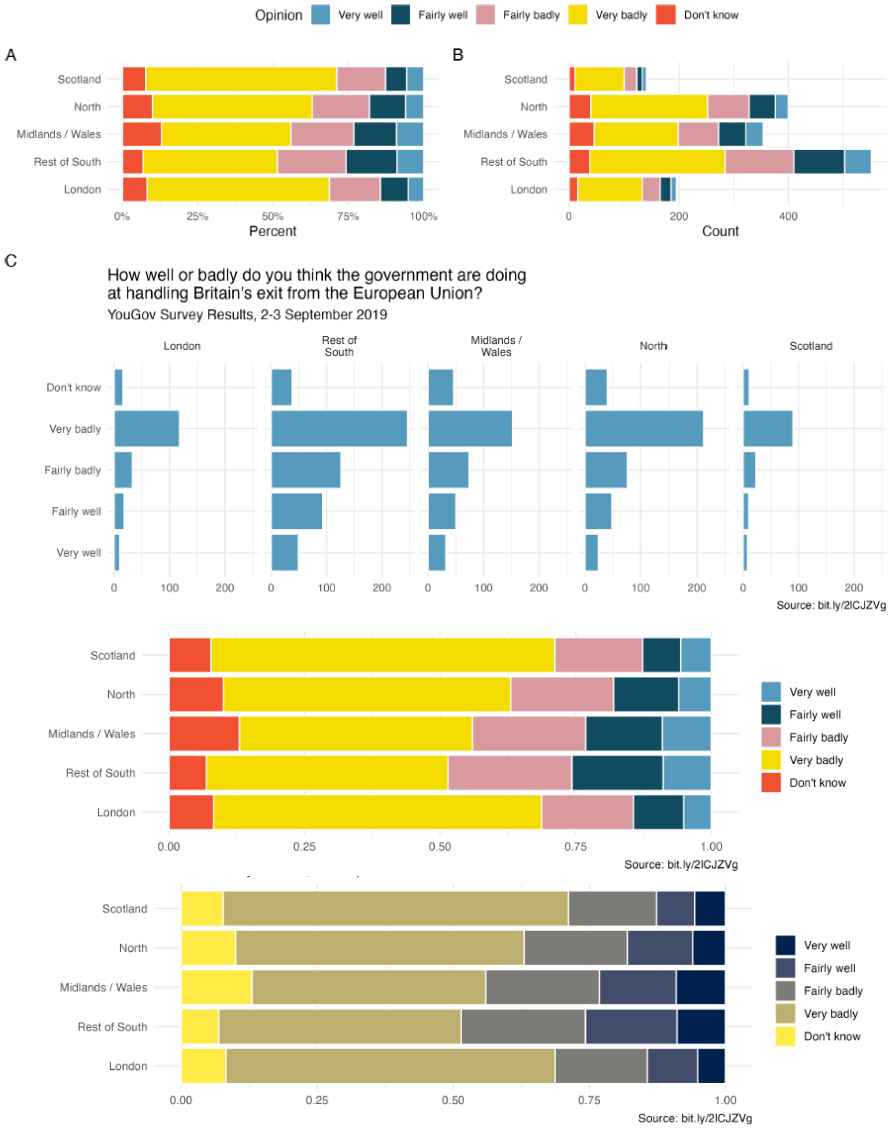
Notes and Exercises Chapter 6 – Applications and Probability
Math 217
Saidi

Chapter 6

Elements of good visualization:

- Color
- Order
- Labels

Example 1: Below, you will see different ways to represent the same results of a survey question: ***How well or badly do you think the government are doing at handling Britain's exit from the European Union? (YouGov Survey Results 9/2019)***



Notice color, length, and labels are used differently to provide information to the viewer.

* Note: A "cividis" palette is generated by optimizing the "viridis" colormap and is optimal for viewing by those with or without color vision deficiency (CVD), a different visual perception of colors that affects 8.5% of the human population

Extra Topic: Intro to Probability

Pr{E} is the likelihood of an event occurring, and the numerical value is always between 0 and 1(inclusive).

A ***chance operation*** is one whose outcome is not deterministic – it must be defined such that each time the chance operation is performed, the event ***E*** either occurs or does not occur

$$\Pr\{E\} = \frac{\#of\ times\ E\ occurs}{\#of\ times\ chance\ operation\ is\ repeated}$$

Law of large numbers states that as the number of experiments increases, the actual ratio of outcomes will converge on the theoretical, or expected, ratio of outcomes.

As an example:

Let E: Heads

Then $\Pr(\{E\} = 0.5 \leftrightarrow \frac{\#of\ heads}{\#of\ tosses}$ (for a long series of tosses of a fair coin)

Probability Tree Diagrams

Toss three coins and represent it with a probability tree diagram
Sample Space:

Calculate the following events:

Pr(3 heads) $\Pr(HHH) = \frac{1}{8} = 0.125$ (same prob. as for Pr(TTT))

Pr(2 heads, 1 tail) $\Pr(HHT) = \frac{3}{8} = 0.375$ (there are 3 ways in the sample space to get 2 heads and 1 tail)

To summarize the probability of getting x heads with 3 coin tosses:

x	0	1	2	3
$P(X = x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

Example 2: Nitric Oxide for Newborns with Hypoxic Respiratory Failure

Hypoxic respiratory failure is a serious condition that affects some newborns. If a newborn has this condition, it is often necessary to use extracorporeal membrane oxygenation (ECMO) to save the life of the child. However, ECMO is an invasive procedure that involves inserting a tube into a vein or artery near the heart, so physicians hope to avoid the need for it. Another treatment for hypoxic respiratory failure is to have the newborn inhale nitric oxide (NO). To test the effectiveness of this treatment, newborns suffering hypoxic respiratory failure were assigned at random to either be given

nitric oxide or a control group.

In the **treatment group (NO) 45.6% of the newborns had a negative outcome,** meaning that either they needed ECMO or that they died. In the **control group (ECMO), 63.6% of the newborns had a negative outcome.**

Create a probability tree diagram and answer the question: what is the probability of getting a negative outcome?

False Positives and False Negatives

The ***sensitivity*** of a test indicates how likely the test will detect a positive outcome. The ***specificity*** of the test correctly indicates that a disease is absent if the person really does not have the disease.

Example 3 This is Conditional Probability

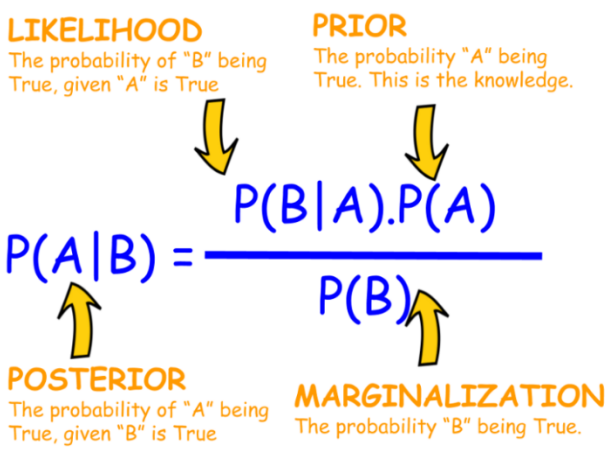
A test detects a positive outcome with 95% accuracy (sensitivity). There is a 90% chance of correctly indicating the disease is absent if the person really does not have the disease (specificity).

Suppose 8% of the population is known to have the disease. Use a probability tree diagram to solve the following.

a. What is the probability that a randomly chosen person will test positive?
b. What is the probability that, given a person tested positive, they actually have the disease? (This is called positive predictive probability)

Bayes' Theorem

Calculate the conditional probability that the first event occurred, given that we only know that the second event occurred.



We start with a hypothesis and a degree of belief in that hypothesis. That means, based on domain expertise or prior knowledge, we assign a non-zero probability to that hypothesis.

Then, we gather data and update our initial beliefs. If the data support the hypothesis, then the probability goes up, if it does not match, then probability goes down.

https://towardsdatascience.com/bayes-rule-with-a-simple-and-practical-example-2bce3d0f4ad0

We have to know 3 pieces of information.

1. The **sensitivity** of the test – the probability that the test will indicate a true positive outcome
2. The **specificity** of the test – the probability that the test will indicate a true negative outcome
3. The **"prior"** known probability about the population regarding the test in question. This means that if we choose a random person from the general population without any testing, what is the probability that the person would test positive?

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

Example 4: In Canada, about 0.35% of women over 40 will develop breast cancer in any given year. A common screening test for cancer is the mammogram, but this test is not perfect. In about 11% of patients with breast cancer, the test gives a false negative: it indicates a woman does not have breast cancer when she does have breast cancer. Similarly, the test gives a false positive in 7% of patients who do not have breast cancer: it indicates these patients have breast cancer when they actually do not.

If we tested a random woman over 40 for breast cancer using a mammogram and the test came back positive {that is, the test suggested the patient has cancer} what is the probability that the patient actually has breast cancer?
Create a probability tree diagram to answer this.

The Binomial Distribution and Formula

The binomial random variable is a chance operation defined by the independent-trials model, which are a series of **independent trials** and the probability of success for each trial is equivalent. The notation for a "success" is p.

For a binomial random variable Y, the probability that n trials result in k successes (and $n-k$ failures):

$$\Pr\{k \text{ successes}\} = \Pr(Y = k) = C_{n,k} * p^k(1-p)^{n-k}$$

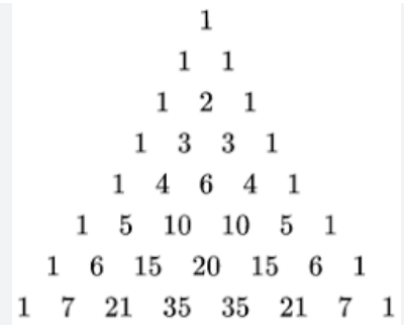
$C_{n,k}$ is called the binomial coefficient, calculated by: $C_{n,k} = \frac{n!}{k!(n-k)!}$

$n! = n(n-1)(n-2) * \ldots * 2 * 1$, and it is called ***n-factorial***. (Note: 0! = 1)

In Table 3.6.2, binomial probabilities for $n = 5$ are shown.

Table 3.6.2 Binomial probabilities for $n = 5$

Number of Successes j	Number of Failures $n-j$	Probability
0	5	$1p^0(1-p)^5$
1	4	$5p^1(1-p)^4$
2	3	$10p^2(1-p)^3$
3	2	$10p^3(1-p)^2$
4	1	$5p^4(1-p)^1$
5	0	$1p^5(1-p)^0$



Example 5 Blood Type
In the U.S., 85% of the population has Rh positive blood. Randomly select 6 people and count the number with Rh positive blood.

a. Identify n and p

b. Create a table to show the probability distribution.

c. Calculate the probability that **at least 1 person is Rh positive**
(that is, 1 or 2 or 3 or 4 or 5 or 6 people are Rh positive)

The Binomial Distribution Using R

Example 6: Suppose there are 12 multiple choice questions in an English class quiz. Each question has 5 possible answers, and only one of them is correct. **Find the probability of having 4 or less correct answers if a student attempts to answer every question at random.**

Solution

Since only one out of the 5 possible answers is correct, the probability of answering a question correctly by random is 1/5 = 0.2. We can find the probability of having exactly 4 correct answers by random attempts:

R code:

```
dbinom(4, size = 12, prob = 0.2)
```

To find the probability of having 4 or less correct answers by random attempts, we apply the function dbinom with x = 0, …,4)

```
dbinom(0, size = 12, prob = 0.2) +
dbinom(1, size = 12, prob = 0.2) +
dbinom(2, size = 12, prob = 0.2) +
dbinom(3, size = 12, prob = 0.2) +
dbinom(4, size = 12, prob = 0.2)
```

Alternatively, use the cumulative probability function - pbinom:

```
pbinom(4, size = 12, prob = 0.2)
```

Answer
The probability of four or less questions answered correctly by random in a 12-question multiple choice quiz is 92.7%.

5-Hand Poker Probabilities

Check out the website to see how combinations work in poker probabilities (https://en.wikipedia.org/wiki/Poker_probability)

The Geometric and Negative Binomial Distributions (from OpenIntroStat – Diaz)

The geometric distribution describes the probability of observing the first success on the nth trial.

Geometric Distribution

If the probability of a success in one trial is p and the probability of a failure is $1-p$, then the probability of finding the first success in the n^{th} trial is given by

$$(1-p)^{n-1}p \qquad (3.30)$$

The mean (i.e. expected value), variance, and standard deviation of this wait time are given by

$$\mu = \frac{1}{p} \qquad \sigma^2 = \frac{1-p}{p^2} \qquad \sigma = \sqrt{\frac{1-p}{p^2}} \qquad (3.31)$$

The **negative binomial distribution** is more general: it describes the probability of observing the **kth success on the nth trial.**

Negative binomial distribution

The negative binomial distribution describes the probability of observing the kth success on the nth trial:

$$\text{P(the kth success on the nth trial)} = \binom{n-1}{k-1} p^k (1-p)^{n-k}$$

where p is the probability an individual trial is a success. All trials are assumed to be independent.

Example 7: Each day a high school football coach tells his star kicker, Brian, that he can go home after he successfully kicks four 35-yard field goals. Suppose we say each kick has a probability p of being successful. If p is small {e.g. close to 0.1 {would we expect Brian to need many attempts before he successfully kicks his fourth field goal?

Solution

We are waiting for the fourth success (k = 4). If the probability of a success (p) is small, then the number of attempts (n) will probably be large. This means that Brian is more likely to need many attempts before he gets k = 4 successes. To put this another way, the probability of n being small is low. To identify a negative binomial case, we check 4 conditions. The first three are common to the binomial distribution.

TIP: Is it negative binomial? Four conditions to check.

(1) The trials are independent.
(2) Each trial outcome can be classified as a success or failure.
(3) The probability of a success (p) is the same for each trial.
(4) The last trial must be a success.

TIP: Binomial versus negative binomial

In the binomial case, we typically have a fixed number of trials and instead consider the number of successes. In the negative binomial case, we examine how many trials it takes to observe a fixed number of successes and require that the last observation be a success

Example 8: On 70% of days, a hospital admits at least one heart attack patient. On 30% of the days, no heart attack patients are admitted. Identify each case below as **a binomial or negative binomial case**, and **compute the probability.**

(a) What is the probability the hospital will admit a heart attack patient on exactly three days this week?

(b) What is the probability the second day with a heart attack patient will be the fourth day of the week?

(c) What is the probability the fifth day of next month will be the first day with a heart attack patient?

Solution: In each part, p = 0:7.
(a) The number of days is fixed, so this is binomial. The parameters are k = 3 and n = 7: 0.097.
(b) The last \success" (admitting a heart attack patient) is fixed to the last day, so we should apply the negative binomial distribution. The parameters are k = 2, n = 4: 0.132.
(c) This problem is negative binomial with k = 1 and n = 5: 0.006. Note that the negative binomial case when k = 1 is the same as using the geometric distribution.

The Poisson Distribution

The **Poisson distribution** is often useful for estimating the number of events in a large population over a unit of time. For instance, consider each of the following events:

- having a heart attack,
- getting married, and
- getting struck by lightning.

The Poisson distribution helps us describe the number of such events that will occur in a short unit of time for a fixed population if the individuals within the population are independent.

Poisson distribution

Suppose we are watching for events and the number of observed events follows a Poisson distribution with rate λ. Then

$$P(\text{observe } k \text{ events}) = \frac{\lambda^k e^{-\lambda}}{k!}$$

where k may take a value 0, 1, 2, and so on (the specific number of successes)

The mean and standard deviation of this distribution are λ and $\sqrt{\lambda}$ respectively.

Is it Poisson?

A random variable may follow a Poisson distribution if we are looking for the number of events, the population that generates such events is large, and the events occur independently of each other. The mean number of events should be known for a given population with consistent probabilities.

Homework Probability

1. If a woman takes an early pregnancy test, she will either test positive, meaning that the test says she is pregnant, or test negative, meaning that the test says she is not pregnant. Suppose that if a woman really is pregnant, there is a 98% chance that she will test positive. Also, suppose that if a woman really is *not* pregnant, there is a 99% chance that she will test negative.
   (a) Suppose that 1,000 women take early pregnancy tests and that 100 of them really are pregnant. What is the probability that a randomly chosen woman from this group will test positive?

   (b) Suppose that 1,000 women take early pregnancy tests and that 50 of them really are pregnant. What is the probability that a randomly chosen woman from this group will test positive?

2. (a) Consider the setting of problem #1, part (a). Suppose that a woman tests positive. What is the probability that she really is pregnant?

   (b) Consider the setting of Exercise 3.2.6, part (b). Suppose that a woman tests positive. What is the probability that she really is pregnant?

3. In the United States, 44% of the population has type A blood. Consider taking a sample of size 4. Let Y denote the number of persons in the sample with type A blood. Find
   (a) $\Pr\{Y = 0\}$
   (b) $\Pr\{Y = 1\}$
   (c) $\Pr\{Y = 2\}$
   (d) $\Pr\{0 \leq Y \leq 2\}$
   (e) $\Pr\{0 < Y \leq 2\}$

4. Neuroblastoma is a rare, serious, but treatable disease. A urine test, the VMA test, has been developed that gives a positive diagnosis in about 70% of cases of neuroblastoma. It has been proposed that this test be used for large-scale screening of children. Assume that 300,000 children are to be tested, of whom 8 have the disease. We are interested in whether or not the test detects the disease in the 8 children who have the disease. Find the probability that
   (a) all eight cases will be detected.
   (b) only one case will be missed.
   (c) two or more cases will be missed. [*Hint:* Use parts (a) and (b) to answer part (c).]

5. Further apply the concepts you've learned in this part in R with computational labs that walk you through a data analysis case study. 6.3 R labs (Intro to data - Flight delays https://www.openintro.org/go?id=ims-r-lab-intro-to-data)