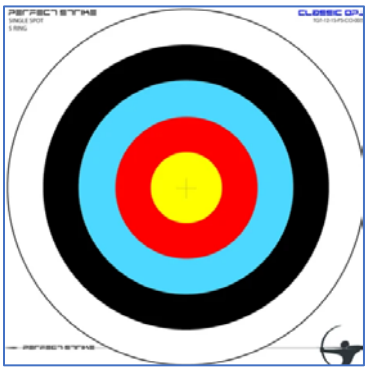
BENG 100. Introduction to Probability for Bioengineering

Class handouts

Throwing a dart: an example of randomness

I am throwing a dart to a squared target of the dimension 100 cm × 100 cm. Inside this squared target is a circle with diameter = 100 cm. The outcome of whether my first throw ends inside the circle is random.



There is randomness about where the dart will land on the target.

Computer simulation challenge

Optional materials for course discussion. The following questions are typically considered graduate level materials.

Can you write a program to simulate this dart throwing experiment? Can you write a program to simulate a total of 100,000 times of the dart-throwing experiment? Please assume that in a hypothetical computer language, the function “rand(a,b)” generates a uniformly distributed random real number in the range of (a,b), where a, b are real numbers.

- Tip: If you have not learned any programing language, providing the gist of how you will write such a program will suffice.
- Answer: consider the following pseudo code:
    - While n < 100,000, assign x_n = rand (0, 100), assign y_n = rand (0, 100), n ++.
- This pseudo code generates 100,000 pairs of x,y coordinates at random.

The π challenge

Optional materials for course discussion. The following questions are typically considered graduate level materials.

Suppose that human beings have not discovered the mathematical constant π. However, the human race has invented computers and you have written your computer program that can simulate the dart throwing experiments. Can you write a computer program to estimate the area of the circle with 100 cm in diameter WITHOU using the mathematical constant π?

- Tip: please use the assumption that in a hypothetical computer language, the function "rand(a,b)" generates a random real number in the range of (a,b), where a, b are real numbers.
- Behind the scenes: If you can, CONGRATULATIONS! You have written a Monte Carlo computer algorithm. Monte Carlo computer algorithms utilize randomness to solve problems. Monte Carlo computer algorithms are a powerful class of modern computational algorithms.
- Answer: We will make an assumption that the chance of a dart (a pair of x,y generated in the answer to the previous simulation challenge) landing within the circle is proportional the ratio of the area of the circle (2R=100 cm) and the area of the square (width = 100 cm). This assumption is satisfied if x,y are generated from uniform distribution (we will discuss Uniform distribution later in this course). The function rand(a,b) generates a random number from a uniform distribution in (a,b).

  Then, a "thrown dart" (x_n, y_n) lands within the circle is equivalent to $(x\_n - 50)^2 + (y\_n - 50)^2 < 50^2$, that is the distance of (x_n, y_n) to the center of the cycle (50,50) is smaller than the radius (50 cm).

  Next, the ratio of the area of the circle to the area of the square can be estimated by the proportion of the "thrown darts" inside the circle among all the "thrown darts" in the square:

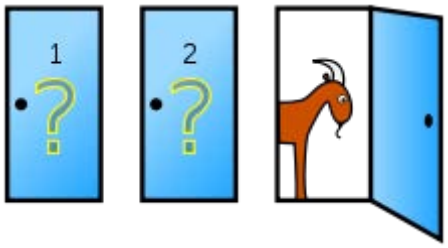
  - In-circle = 0;
  - While n < 100,000,
    - assign x_n = rand (0, 100);
    - assign y_n = rand (0, 100);
    - if $((x\_n - 50)^2 + (y\_n - 50)^2 < 50^2)$, then In-circle ++;
    - n ++.
  - Ratio = In-circle / 100,000;

The Ratio above is our estimated π because πR^2 is the area of the circle and R^2 is the area of the square.

The Monty Hall problem

The Monty Hall problem is a brain teaser, in the form of a probability puzzle, loosely based on the American television game show *Let's Make a Deal* and named after its original host, Monty Hall (Wikipedia).

Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?



In search of a new car, the player picks a door, say 1. The game host then opens one of the other doors, say 3, to reveal a goat and offers to let the player switch from door 1 to door 2.

The Monty Hall problem (Continued)

Below is a simpler formulation of the Monty Hall problem, assuming the guest of the show picked Door 1 to start the game, as well as a solution to this simplified formulation.

* (The Monte Hall Problem [1]) You are in a game show, and the host gives you the choice of three doors. Behind one door is a car and behind the others are goats. You pick a door, say door 1. The host who knows what is behind the doors opens a different door and reveals a goat (the host can always open such a door because there is only one door behind which is a car). The host then asks you: "Do you want to switch?" The question is, is it to your advantage to switch your choice?

Solution:

Yes, if you switch, your chance of winning the car is $\frac{2}{3}$. Let W be the event that you win the car if you switch. Let C_i be the event that the car is behind door i, for $i = 1, 2, 3$. Then $P(C_i) = \frac{1}{3} \quad i = 1, 2, 3$.
Then,

$$\begin{aligned} P(W) &= \sum_{i=1}^{3} P(W|C_i)P(C_i) \\ &= P(W|C_1)P(C_1) + P(W|C_2)P(C_2) + P(W|C_3)P(C_3) \\ &= 0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} \\ &= \frac{2}{3} \end{aligned}$$

Proof of chain rule of conditional probability

The chain rule of conditional probably (chapter 1.4.0) can be proved as follows:

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

Applications of the hypergeometric distribution

- **Color of cards**. A deck of cards contains 20 cards: 6 red cards and 14 black cards. 5 cards are drawn randomly *without replacement*. What is the probability that exactly 4 red cards are drawn?

The probability of choosing exactly 4 red cards is:
P(4 red cards) = # samples with 4 red cards and 1 black card / # of possible 4 card samples

Recall that the PMF of a hypergeometric random variable X(b,r,k):

$$P(X = x) = \frac{\binom{b}{x}\binom{r}{k-x}}{\binom{b+r}{k}}$$

where b=6, r=14, k=5, x=4.

$$P(X = 4) = 0.0135$$

- **Patients in a clinic.** A clinic has 101 cancer and 95 non-cancer patients. The clinical records of 10 patients are drawn at random from this clinic. What is the probability exactly 7 of the drawn records are from cancer patients? What is the probability that exactly 7 of the drawn records are from non-cancer patients?

Recall that the PMF of a hypergeometric random variable X(b,r,k):

$$P(X = x) = \frac{\binom{b}{x}\binom{r}{k-x}}{\binom{b+r}{k}}$$

where b=101, r=95, k=10, x=7. P(X=7)=0.130.

  - What is the probability that exactly 7 of the drawn records are from non-cancer patients?

Recall that the PMF of a hypergeometric random variable X(b,r,k):

$$P(X = x) = \frac{\binom{b}{x}\binom{r}{k-x}}{\binom{b+r}{k}}$$

where b=95, r=101, k=10, x=7. Alternatively, b=101, r=95, k=10, x=3.

Genotype vs. phenotype

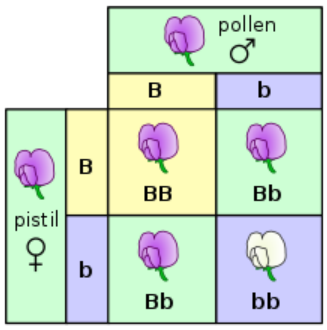
The term Genotype is explained by Wikipedia as: “A **Genotype** is an organism’s complete set of heritable genes, or genes that can be passed down from parents to offspring. These genes help encode the characteristics that are physically expressed (phenotype) in an organism, such as hair color, height, etc. The term was coined by the Danish botanist, plant physiologist and geneticist Wilhelm Johannsen in 1903.

The genotype is one of three factors that determine phenotype, along with inherited epigenetic factors and non-inherited environmental factors. Not all organisms with the same genotype look or act the same way because appearance and behavior are modified by environmental and growing conditions. Likewise, not all organisms that look alike necessarily have the same genotype.

One's genotype differs subtly from one's genomic flash card sequence, because it refers to how an individual *differs* or is specialized within a group of individuals or a species. So, typically, one refers to an individual's genotype with regard to a particular gene of interest and the combination of alleles the individual carries (see homozygous, heterozygous). Genotypes are often denoted with letters, for example *Bb*, where *B* stands for one allele and *b* for another.

Somatic mutations which are acquired rather than inherited, such as those in cancers, are not part of the individual's genotype. Hence, scientists and physicians sometimes talk about the genotype of a particular cancer, that is, of the disease as distinct from the diseased.

An example of a characteristic determined by a genotype is the petal color in a pea plant. The collection of all genetic possibilities for a single trait are called alleles; two alleles for petal color are purple and white.”



“Here the relation between genotype and phenotype is illustrated, using a Punnett square, for the character of petal colour in a pea plant. The letters B and b represent alleles for colour and the pictures show the resultant flowers.”

“Any given gene will usually cause an observable change in an organism, known as the phenotype. The terms genotype and phenotype are distinct for at least two reasons:

- To distinguish the source of an observer's knowledge (one can know about genotype by observing DNA; one can know about phenotype by observing outward appearance of an organism).

- Genotype and phenotype are not always directly correlated. Some genes only express a given phenotype in certain environmental conditions. Conversely, some phenotypes could be the result of multiple genotypes. The genotype is commonly mixed up with the phenotype which describes the end result of both the genetic and the environmental factors giving the observed expression (e.g. blue eyes, hair color, or various hereditary diseases).

A simple example to illustrate genotype as distinct from phenotype is the flower colour in pea plants (see Gregor Mendel). There are three available genotypes, PP (homozygous dominant), Pp (heterozygous), and pp (homozygous recessive). All three have different genotypes but the first two have the same phenotype (purple) as distinct from the third (white).

A more technical example to illustrate genotype is the single-nucleotide polymorphism or SNP. **A SNP occurs when corresponding sequences of DNA from different individuals differ at one DNA base**, for example where the sequence AAGC**C**TA changes to AAGC**T**TA. This contains two alleles: C and T. SNPs typically have three genotypes, denoted generically AA Aa and aa. (Note that instead of using AA Aa and aa, the above figure used "BB, Bb, bb" to denote the three genotypes). In the example above, the three genotypes would be CC, CT and TT."

- **Question**: Now that let us assume that for specific SNP inside a particular gene, there are two alleles, C and A. A genomics study included a total of 1,005 people. Among them, 999 carried the C allele, and the rest 6 people carried the A allele. (In such a case, the A allele is often referred to as the minor allele.) Three out of the 1,005 people have colon cancer. Assuming that this specific SNP of this particular gene (the genotype) is independent of the colon cancer (the phenotype), what is the probability that exactly 2 colon-cancer patients in this study carry the A allele? What is the probability that 2 or more colon-cancer patients in this study carry the A allele?

Answer to the probability that exactly 2 colon-cancer patients in this study carry the A allele.

Recall that the PMF of a hypergeometric random variable X(b,r,k):

$$P(X = x) = \frac{\binom{b}{x}\binom{r}{k-x}}{\binom{b+r}{k}}$$

where b=6, r=999, k=3, x=2. Alternatively, b=999, r=6, k=3, x=1.

Answer to the probability that 2 or more colon-cancer patients in this study carry the A allele.

$$P(X = 2\ or\ 3) = P(X = 2) + P(X = 3) = \frac{\binom{6}{2}\binom{999}{1}}{\binom{1005}{3}} + \frac{\binom{6}{3}\binom{999}{0}}{\binom{1005}{3}}$$

The instructor's cheat sheet to hypothesis testing

To perform a hypothesis test, we will need the following steps:

1. Forming two competing hypotheses, called **the null (H0) and the alternative hypothesis (H1)**.
    a. Tip: The keyword is "competing". If we consider H0 and H1 as two sets, they must be disjoint.
        i. For example, please consider whether these following pairs of hypotheses can be regarded as competing hypotheses:
            1. A table surface is flat vs. a table surface is not flat.
            2. $u=0$ vs. $u>0$.
            3. $u=0$ vs. $u\geq 0$.
            4. $u=0$ vs. $u\neq 0$.
    b. Tip: put the simpler hypothesis as H0. Either of the two hypotheses can be regarded as H0, however, the procedure for testing the two hypotheses will be easier if the simpler hypothesis is designated as H0.
        i. For example, which way of formatting H0 and H1 is better:
            1. H0: $u=0$; H1: $u\neq 0$
            2. H0: $u\neq 0$; H1: $u=0$
2. Generating or getting data. Our general idea is to use the data generated from experiments to test the hypothesis, that is to argue which of the two competing hypotheses is more likely to be supported by data.
    a. Tip: the key word is "argue". We will see that the entire testing is a process of forming an argument.
    b. Tip: the basis of this argument is the data.
        i. If there are already data generated by experiments, no need to do anything.
        ii. If there are no data generated yet, do the experiments to generate the data.
        iii. The experiments must be relevant to the hypotheses, thus can be used for an argument about the hypotheses (we will revisit this point in the discussion of Test Statistic).
3. Summarizing the data into a **Test Statistic**. The data can be a large set of numbers, which can hardly be directly used for making any argument. Thus, we must summarize the data into a single number to form our argument. This single number is called a Statistic.
    a. Tip: the first keyword is a single number. For example, suppose my 3 experiments generated 3 data points, i.e. x1, x2, x3, which of the following is a Statistic:
        i. x1+x2+x3
        ii. x1-x2-x3
        iii. x1-x2
        iv. x2
    b. Tip: the second concept is to form an argument. To be able to form an argument, the magnitude of this Statistic must reflect the degree of support to one of the two hypotheses. When the magnitude of this Statistic reflects the degree of support to

one of the two hypotheses, we call this Statistic the Test Statistic. Thus, we can make an argument based on how large of small the Test Statistic is.

i. For example, suppose we are testing two brands of an experimental reagent, Brand X, a classical brand, and Brand Y, a new brand. for which produced a greater yield of DNA from a DNA extraction experiment. We performed the experiment with Brand X 3 times, with the DNA yield of x1, x2, x3. We performed the experiment with Brand B 4 times, with the DNA yield of y1, y2, y3, y4. What hypotheses can we formulate?
    1. Answer: H0: the average yield of Brand X equals the average yield of Brand Y. H1: the average yield of the new brand (Y) is larger than the average yield of the classical Brand (X).

ii. Which of the following statistic (denoted as t) is a good Test Statistic? There can be more than 1 correct answer:
    1. t= x1+x2+x3
    2. t= x1+x2+x3+y1+y2+y3+y4
    3. t= (x1+x2+x3+y1+y2+y3+y4)/7
    4. t= (x1+x2+x3)-(y1+y2+y3+y4)
    5. t= [(x1+x2+x3)/3-(y1+y2+y3+y4)/4]/σ, where σ is a constant
    6. t= [(x1+x2+x3)/3]/[(y1+y2+y3+y4)/4], assuming y1+y2+y3+y4>0

iii. Tip: there are more than one Test Statistic for testing a pair of competing hypotheses.

c. At this point, based on the magnitude of our Test Statistic (t), we can already subjectively judge which hypothesis more believable. (Please recall at the beginning of this course, we mentioned that probability can be interpreted as subjective belief.)

d. Our last question is how to quantify our subjective belief? This question leads to the introduction of the p-value.

4. Calculating **p-value**. The p-value can be thought as the probability of seeing the currently observed value of the Test Statistic (t) or *more extreme* values of this Test Statistic (T) if we were to repeat the experiments in future for many times and the null hypothesis is true:
    a. The above statement can be written as: P-value = $P(T \geq t \mid H0)$
    b. Note that we have unconsciously introduced a random variable, called the Test Statistic (T). Also note that t is the observed value of this RV based on the currently finished experiments.
    c. To calculate the p-value, we note that p-value = $1 - F_{T|H0}(t)$, where $F_{T|H0}(t)$ is the CDF of T when H0 is true. Please do not be scared by T|H0. T|H0 is just a random variable (distribution). We call this distribution the null distribution or the distribution of the Test Statistic under the null hypothesis.
    d. Finally, as long as we can obtain the CDF of the null distribution, we can calculate the p-value.
    e. How to derive the CDF of the null distribution? There are two ways.

i. First, people can try to derive the mathematical form of the CDF under some assumptions on the distribution of the original experiment. Each of these previously derived and documented CDFs is usually titled with a name, such as the T distribution. Coupled with the name of the CDF is statistical test, such as the T test.
ii. Second, people can use computer simulation to produce random outcomes under H0. In the Brand X vs. Brand Y example, one way to simulate random outcomes is to keep the yields but randomly swap the labels (x, y). Summarizing these simulated data points can produce an approximation to the null distribution.

5. Making a decision based on p-value. The decision is either "Reject H0" or not reject. We use p-values to quantitatively assess how much belief we have for H1, i.e. against H0. The smaller the p-value, the great belief we have against the H0. People often choose an arbitrary cutoff to p-value, such as 0.05, to make a binary judgement. For example, people often say: "Since the p-value is smaller than 0.05, we reject the null hypothesis" or "We cannot reject the null hypothesis because the p-value is greater than 0.05."
6. (Optional) Making a decision based on the **acceptance region**. This is an alternative (slightly old fashioned) way for making a decision. Since the magnitude of the observed value (t) of our Test Statistic directly relates to the degree of evidence for (or against) H0, we can use a threshold (δ) on the Test Statistic for making the decision. For example, if the larger t is the greater the evidence is against H0, then we can decide to:
    a. Reject H0, if $t > \delta$.
    b. Not to reject H0, if $t \leq \delta$.

    In this case, $(-\infty, \delta)$ is the **Acceptance Region**. The statistical test is completely defined when the Test Statistic (T) and the acceptance region are given. For a completely defined test, we can obtain the following probabilities:

    a. P(Reject H0|H0). The action (decision) of "Reject H0|H0" is called the **Type I error**, also referred to as the false positive.
    b. P(Do not reject H0|H1). The action (decision) of "Not to reject H0|H1" is called **the Type II error**, also referred to also false negative.
    c. In the above example, $P(\text{Reject H0}|\text{H0}) = P(t > \delta|\text{H0})$, $P(\text{Do not reject H0}|\text{H1}) = P(t \leq \delta|\text{H1})$.
    d. Since the CDF of T|H0 is often given, at least in those "named tests" such as the T test, $P(t > \delta|\text{H0})$ can be computed by $P(t > \delta|\text{H0}) = 1\text{-}F_{T|H0}(\delta)$. This probability is called the **significance level**, denoted as α.
    e. As long as the CDF of T|H0, denoted as $F_{T|H0}(\delta)$, is known, we can calculate α based on the acceptance region $(-\infty,\delta)$ or calculate δ based on the significance level (α). This is because $\alpha = P(t > \delta|\text{H0}) = 1\text{-}F_{T|H0}(\delta)$.