Statistical Data Analysis

Prof. Dr. Nizamettin AYDIN

naydin@yildiz.edu.tr

http://www3.yildiz.edu.tr/~naydin

Probability

Probability as a Measure of Uncertainty

- Plots and summary statistics are used to learn about the distribution of variables and to investigate their relationships.
  - However, we always remain uncertain about the true distributions and relationships in the population since we almost never have access to all of its members.
  - Furthermore, our findings based on the observed sample can change if different samples from the population were obtained.
- Therefore, when we generalize our findings from a sample to the whole population, we should explicitly specify the extent of our uncertainty.
  - We use probability as a measure of uncertainty.

Some Commonly Used Genetic Terms

- Gene
  - a segment of double-stranded DNA, which itself is made of a sequence of four different nucleotides:
    - adenine (A), guanine (G), thymine (T), or cytosine (C).
- Single Nucleotide Polymorphisms (SNPs)
  - Genetic variation is caused by changes in the DNA sequence of a gene.
  - SNPs are the most common type of genetic variation.
  - SNPs occur when a single nucleotide is replaced by another one.
    - An example of a SNP would be replacing “G” in the sequence {TAGCAAT} by “T” to create {TATCAAT}.
- Alleles
  - alternate forms of a gene
  - responsible for variation in phenotypes.
    - Phenotypes, in general, are observable traits, such as eye color, disease status, and blood pressure, due to genetic factors and/or environmental factors
  - In the above example, the alleles could be denoted as T and G.
    - We denote the genes with bold face letters (e.g., **A**) and the two different alleles as capital and small letters (e.g., *A* and *a*).

Some Commonly Used Genetic Terms

- Genotype
  - Genetic materials are stored on chromosomes.
  - Human somatic cells have two copies of each chromosome
    - one inherited from each parent; hence, they are called diploid.
  - Each pair of similar chromosomes are called homologous chromosomes.
  - The genotype (i.e., genetic makeup) of an individual for the bi-allelic gene **A** can take one of the three possible forms:
    - AA, aa, or Aa.
- Homozygous vs. heterozygous
  - The first two genotypes, AA and aa, are called homozygous,
    - which means the same version of the allele was inherited from both parents.
      - That is, both homologous chromosomes have the same allele.
  - The last genotype, Aa, is called heterozygous,
    - which means different alleles were inherited.

Some Commonly Used Genetic Terms

- Phenotype
  - the set of observable characteristics of an individual resulting from the interaction of its genotype with the environment
- Recessive vs. dominant
  - The presence of a specific allele does not always result in its corresponding trait (a characteristic such as eye color).
  - Some alleles are recessive,
    - producing their trait only when both homologous chromosomes carry that specific variant.
  - On the other hand, some alleles are dominant,
    - producing their traits when they appear on at least one of the homologous chromosomes.
      - {For example, suppose that the allele a for gene A is responsible for a specific disease.
      - Furthermore, assume that a is a recessive allele.
      - Then, only a person with genotype aa will be affected by the disease.
      - Individuals with genotype AA or Aa will not have the disease.}

Random phenomena and their sample space

- A phenomenon is called random if its outcome (value) cannot be determined with certainty before it occurs.
  - For example, coin tossing and genotypes are random phenomena.
- The collection of all possible outcomes S is called the sample space.

Coin tossing :	$S = \{H, T\}$,
Die rolling :	$S = \{1, 2, 3, 4, 5, 6\}$,
Bi-allelic gene :	$S = \{A, a\}$,
Genotype :	$S = \{AA, Aa, aa\}$.

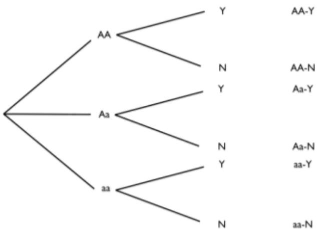
Random phenomena and their sample space

- The sample space might include an infinite number of possible outcomes.
  - For example, the value of blood pressure is random since it cannot be determined with certainty before measuring it.
    - The corresponding sample space for blood pressure values is (theoretically) the set of positive real numbers, which is infinite.
- For a complex random phenomenon that is a combination of two or more other random phenomena, it might be easier to view the sample space with tree diagrams.

Random phenomena and their sample space

- For example, suppose that we suspect that gene **A** is related to a specific disease, but genetic variation alone does not determine the disease status.
  - Rather, it affects the risk of the disease.
  - Further, we suspect that smoking (an environmental factor) is also related to the disease.
- In this case, the random phenomenon we are interested in is the combination of genotype and smoking status
- All possible combinations (i.e., sample space) are identified using the following tree diagram.

Random phenomena and their sample space



- Genotypes (*AA*, *Aa*, and *aa*) are represented by the first set of branches
- Smoking status (Y for smokers and N for nonsmokers) is represented by the second set of branches
- Following each branch from root to tip, we obtain the following sample space;
  - $S = \{AA\text{-}Y, AA\text{-}N, Aa\text{-}Y, Aa\text{-}N, aa\text{-}Y, aa\text{-}N\}$.
    - For example, *Aa - Y* represents the outcome of having heterozygous genotype and smoking.

Probability Measure

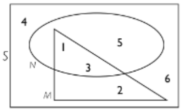
- To each possible outcome in the sample space, we assign a probability P,
  - which represents how certain we are about the occurrence of the corresponding outcome.
    - For an outcome o, we denote the probability as $P(o)$,
      - where $0 \leq P(o) \leq 1$.
- The total probability of all outcomes in the sample space is always 1.
  - Coin tossing : $P(H) + P(T) = 1$
  - Die rolling : $P(1) + P(2) + P(3) + P(4) + P(5) + P(6) = 1$
- Therefore, if the outcomes are equally probable,
  - the probability of each outcome is $1/n_S$,
    - where n_S is the number of possible outcomes.

Random events

- An event is a subset of the sample space S.
  - A possible event for die rolling is
    - $E = \{1,3,5\}$.
      - This is the event of rolling an odd number.
  - For the genotype example,
    - $E = \{AA, aa\}$
      - This is the event that a person is homozygous.
- An event occurs when any outcome within that event occurs.
- We denote the probability of event E as $P(E)$.
- The probability of an event is the sum of the probabilities for all individual outcomes included in that event.

Random events – Example 1

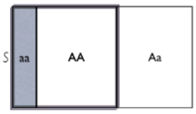
- Consider the die rolling example presented in the form of a Venn diagram below.



- All the possible outcomes are contained inside the sample space *S*, which is represented by the rectangle.
- We define two events.
  - The event *M* (shown as a triangle) occurs when the outcome is less than 4.
  - The event *N* (shown as an oval) occurs when the outcome is an odd number.
- In this example, *P(M)* = 1/2 and *P(N)* = 1/2

Random events – Example 2

- As a running example, we consider a bi-allelic gene **A** with two alleles *A* and *a*.
- We assume that allele *a* is recessive and causes a specific disease.
  - Then only people with the genotype *aa* have the disease.
    - A schematic representation for a bi-allelic gene with a recessive allele *a* that causes a specific disease.



- The *shaded area* shows the disease event (*D*).
- The *unshaded area* shows the no-disease event (*ND*).
- The *area with shaded border lines* shows the homozygous event (*HM*).
- The *remaining part* of the sample space, which includes the outcome *Aa* only, corresponds to the heterozygous event

Random events - Example

- We can define four events as follows:
  - The homozygous event : *HM* = {*AA*, *aa*};
  - The heterozygous event : *HT* = {*Aa*};
  - The no-disease event : *ND* = {*AA*,*Aa*};
  - The disease event : *D* = {*aa*}:
- Assume that the probabilities for different genotypes are
  - *P*(*AA*) = 0.49, *P*(*Aa*) = 0.42, and *P*(*aa*) = 0.09.
- Then,
  - *P*(*HM*) = 0.49 + 0.09 = 0.58*;*
  - *P*(*HT*) = 0.42*;*
  - *P*(*ND*) = 0.49 + 0.42 = 0.91*;*
  - *P*(*D*) = 0.09.

Complement

- For any event *E*, we define its complement, E^c, as the set of all outcomes that are in the sample space *S* but not in *E*.
  - For the gene-disease example, the complement of the homozygous event *HM* = {*AA*, *aa*} is the heterozygous event {*Aa*};
    - we show this as $HM^c = HT$.
  - Likewise, the complement of the disease event, *D* = {*aa*}, is the no-disease event, *ND* = {*AA*, *Aa*};
    - we show this as $D^c = ND$.
- The probability of the complement event is
  - 1 minus the probability of the event:

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

Complement - example

- For the event that the outcome is an odd number, we have
  - $P(N^c) = 1 - P(N) = 1 - (1/2) = 1/2$
    - equal to the probability that the outcome is an even number.
- In the gene disease example, the probability of the complement of the homozygous event is
  - $P(HM^c) = 1 - P(HM) = 1 - 0.58 = 0.42.$
    - equal to the probability of the heterozygous event *P*(*HT*) = 0.42.
- Likewise, the probability of the complement of the disease event is
  - $P(D^c) = 1 - P(D) = 1 - 0.09 = 0.91$
    - equal to the probability of the no-disease event, *P(ND)* = 0.91.

Complement

- The odds of an event shows how much more certain we are that the event occurs than we are that it does not occur.
- For event *E*, we calculate the odds as follows: $\frac{P(E)}{P(E^c)} = \frac{P(E)}{1 - P(E)}$
- For the gene-disease example, the odds for *ND* (i.e., not having the disease) are

$$\frac{P(ND)}{P(ND^c)} = \frac{P(ND)}{1 - P(ND)} = \frac{0.91}{1 - 0.91} = 10.11$$

- Therefore, it is almost 10 times more likely that a person is not affected by the disease than it is for having the disease.
  - In this case, we say that the odds for not having the disease are 10 to 1.

Union

- For two events E_1 and E_2 in a sample space S, we define their union $E_1 \cup E_2$ as the set of all outcomes that are at least in one of the events.
- The union $E_1 \cup E_2$ is an event by itself, and it occurs when either E_1 or E_2 (or both) occurs.
  - For example, the union of the heterozygous event, *HT*, and the disease event, *D*, is
    - $\{Aa\} \cup \{aa\} = \{Aa, aa\}$.
- When possible, we can identify the outcomes in the union of the two events and find the probability by adding the probabilities of those outcomes.

Union

- For the die rolling example (slide 13)

$$P(M \cup N) = P(\{1, 2, 3, 5\}) = \frac{4}{6} = \frac{2}{3}$$

- Note that in general this is not equal to the sum of the probabilities of the two events:

$$P(M \cup N) \neq \tfrac{1}{2} + \tfrac{1}{2}$$

- Only under a specific condition, we can write the probability of the union of two events as the sum of their probabilities.
- For the union of the heterozygous event, *HT* , and the disease event, *D*,

$$P(HT \cup D) = P(\{Aa, aa\}) = 0.42 + 0.09 = 0.51$$

- In this special case, the probability of the union of the two events is equal to the sum of their individual probabilities.

Intersection

- For two events E_1 and E_2 in a sample space S, we define their intersection $E_1 \cap E_2$ as the set of outcomes that are in both events.
- The intersection $E_1 \cap E_2$ is an event by itself, and it occurs when both E_1 and E_2 occur.
  - For example, the intersection of the heterozygous event and the no-disease event is $HM \cap ND = \{AA\}$.
- The intersection of M and N in the dye rolling example (slide 13) is

$$M \cap N = \{1,3\}$$

  - In this case, the intersection of the two events includes outcomes that are less than 4 and odd.
- The intersection of the heterozygous event and the no-disease event is $HM \cap ND = \{AA\}$.

Intersection - Example

- For the die rolling example (slide 13)

$$P(M \cap N) = P(\{1, 3\}) = \frac{2}{6} = \frac{1}{3}$$

- For the gene-disease example (slide 14)

$$P(HM \cap ND) = P(AA) = 0.49$$

- Now consider the intersection of the heterozygous event and the disease event.
  - There is no common element between *HT* and *D*.
  - Therefore, the intersection is the empty set
    - $HT \cap D = \{\}$,
  - its probability is
    - $P(HT \cap D) = P(\emptyset) = 0$.

Joint vs. marginal probability

- We refer to the probability of the intersection of two events, $P(E_1 \cap E_2)$, as their joint probability.
- In contrast, we refer to probabilities $P(E_1)$ and $P(E_2)$ as the marginal probabilities of events E_1 and E_2.
- For any two events E_1 and E_2, we have
  - $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$.
    - That is, the probability of the union $P(E_1 \cap E_2)$ is the sum of their marginal probabilities minus their joint probability.
- {The union of the heterozygous and the no-disease events is
  - $P(HM \cup ND) = P(HM) + P(ND) - P(HM \cap ND)$
    $= 0.58 + 0.91 - 0.49 = 1$}

Disjoint events

- Two events are called disjoint or mutually exclusive if they never occur together:
  - if we know that one of them has occurred, we can conclude that the other event has not.
- Disjoint events have no elements (outcomes) in common, and their intersection is the empty set.
- {For the above example (slide 14), if a person is heterozygous, we know that he does not have the disease
  - so the two events HT and ND are disjoint.}

Disjoint events

- For two disjoint events E_1 and E_2, the probability of their intersection (i.e., their joint probability) is zero:
  - $P(E_1 \cap E_2) = P(\varphi) = 0$
- Therefore, the probability of the union of the two disjoint events is simply the sum of their marginal probabilities:
  - $P(E_1 \cup E_2) = P(E_1) + P(E_2)$
- In general, if we have multiple disjoint events, E_1, E_2, ..., E_n, then the probability of their union is the sum of the marginal probabilities:
  - $P(E_1 \cup E_2 \cup ... \cup E_n) = P(E_1) + P(E_2) + ... + P(E_n)$

Disjoint events - Example

- The probability of the union of the heterozygous and disease events is
  - $P(HT \cup D) = 0.42 + 0.09 = 0.51.$
- Likewise, when we roll a die, the events {1, 2}, {4}, and{5, 6}are disjoint.
- The occurrence of one event prevents the occurrence of the others.
- Therefore, the probability of their union is
  - $P(\{1,2\} \cup \{4\} \cup \{5,6\}) = 1/3 + 1/6 + 1/3 = 5/6$
- Now consider the three events{1, 2, 3},{4}, and{5, 6}.
  - These events are disjoint, and their union is the sample space *S*.

Partition

- When two or more events are disjoint and their union is the sample space *S*,
  - we say that the events form a partition of the sample space.
- Two complementary events *E* and E^c always form a partition of the sample space
  - since they are disjoint and their union is the sample space.

Conditional Probability

- Very often, we need to discuss possible changes in the probability of one event based on our knowledge regarding the occurrence of another event.
- The conditional probability, denoted $P(E_1|E_2)$, is
  - the probability of event E_1 given that another event E_2 has occurred.
- The conditional probability of event *E*1 given event E_2 can be calculated as follows: (assuming $P(E_2) \neq 0$)

$$P(E_1|E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)}$$

  - This is the joint probability of the two events divided by the marginal probability of the event on which we are conditioning .

Conditional Probability - Example

- Consider the die rolling example (slide 13).
- The intersection of the two events is
  - $M \cap N = \{1, 3\}$

  with probability
  - $P(E_1 \cap E_2) = 2/6 = 1/3.$
- Therefore, the conditional probability of an outcome less than 4, given that the outcome is an odd number, is

$$P(M|N) = \frac{P(M \cap N)}{P(M)} = \frac{1/3}{1/2} = \frac{2}{3}$$

Conditional Probability - Example

- Consider the gene-disease example (slide 14).
- Suppose we know that a person is homozygous and are interested in the probability that this person has the disease, $P(D|HM)$.
- The probability of the intersection of *D* and *HM* is
  - $P(D \cap HM) = P(\{aa\}) = 0.09$
- Therefore, the conditional probability of having the disease knowing that the genotype is homozygous can be obtained as follows:

$$P(D|HM) = \frac{P(D \cap HM)}{P(HM)} = \frac{0.09}{0.58} = 0.16$$

- In this case, the probability of the disease has increased from $P(D) = 0.09$ to $P(D|HM) = 0.16$.

Conditional Probability - Example

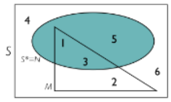
- Now let us find the conditional probability of not having the disease knowing that the person has a homozygous genotype: $P(ND|HM)$.
- The joint probability of ND and HM is
  - $P(ND \cap HM) = P(\{AA\}) = 0.49$.
- The conditional probability is therefore

$$P(ND|HM) = \frac{P(ND \cap HM)}{P(HM)} = \frac{0.49}{0.58} = 0.84$$

- The information that the person is homozygous decreases the probability of no disease from its 0.91 to 0.84.
- Note that the two events ND and D are complementary, and the conditional probability of ND given HM is
  - $P(ND|HM) = 1 - P(D|HM) = 1 - 0.16 = 0.84$.

Conditional Probability

- In general, all the probability rules we discussed so far apply to conditional probabilities.



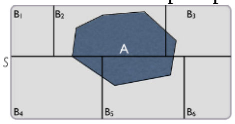
- Conditioning on an event only reduces the sample space (e.g., from the large rectangle to the shaded oval in in the figure).
- Within this shrunken sample space, all probability rules are valid.
- For example,

$$P(E_1^c|E_2) = 1 - P(E_1|E_2),$$

$$P(E_1 \cup E_2|E_3) = P(E_1|E_3) + P(E_2|E_3) - P(E_1 \cap E_2|E_3)$$

The law of total probability

- By rearranging the equation for conditional probabilities, we have
  - $P(E_1 \cap E_2) = P(E_1|E_2)P(E_2)$.
- Now suppose that a set of K events B_1, B_2, ... , B_K forms a partition of the sample space.



- Using the above equation, we have
  - $P(A) = P(A|B_1)P(B_1) + \cdots + P(A|B_K)P(B_K)$
- This is known as the law of total probability

The law of total probability

- The law of total probability can be written as

$$P(A) = \sum_{k=1}^{K} P(A|B_k)P(B_k)$$

where B_1, B_2, ... , B_K form a partition of the sample space, and A is an event in the sample space.

- For die rolling example, consider the three events
  - $B_1 = \{1, 2\}$, $B_2 = \{3,4\}$, and $B_3 = \{5, 6\}$,
    - whose probabilities are $P(B_1) = P(B_2) = P(B_3) = 1/3$.
- These events form a partition of the sample space.
- The conditional probabilities of M (outcome less than four) given either of these three events are
  - $P(M|B_1) = 1$, $P(M|B_2) = 1/2$, $P(M|B_3) = 0$.

The law of total probability

- If we know that the event $B_1 = \{1, 2\}$ has occurred, we know for sure that the outcome is less than 4.
- Given $B_2 = \{3, 4\}$, the possible outcomes are now 3 and 4.
- One of two possible outcomes corresponds to the event M, that is, the conditional probability of M given B_2 is $1/2$.
- If we know that the event $B_3 = \{5, 6\}$ has occurred,
  - then the probability that the number is less than 4 is zero: $P(M|B3) = 0$.
- Using the law of total probability, we have

$$P(M) = P(M|B_1)P(B_1) + P(M|B_2)P(B_2) + P(M|B_3)P(B_3)$$

$$= 1 \times \frac{1}{3} + \frac{1}{2} \times \frac{1}{3} + 0 \times \frac{1}{3} = \frac{1}{2},$$

which is the same as the probability we found directly based on the outcomes included in M.

Independent events

- Two events E_1 and E_2 are independent if our knowledge of the occurrence of one event does not change the probability of occurrence of the other event.
  - $P(E_1|E_2) = P(E_1)$
  - $P(E_2|E_1) = P(E_2)$
- For example, if a disease is not genetic, knowing a person has a specific genotype (e.g., AA) does not change the probability of having that disease.

Independent events

- When two events E_1 and E_2 are independent, the probability that E_1 and E_2 occur simultaneously, i.e., their joint probability, is the product of their marginal probabilities:
  - $P(E_1 \cap E_2) = P(E_1) \times P(E_2)$
- Therefore, the probability of the union of two independent events is as follows:
  - $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1) \times P(E_2)$
- In general, if events E_1, E_2 , ... , E_n are independent
  - $P(E_1 \cap E_2 \cap ... \cap E_2) = P(E_1) \times P(E_2) \times ... \times P(E_n)$

Independent events - Example

- If we toss two fair coins simultaneously, then the probability of observing heads on both coins is
  - $P(H_1 \cap H_2) = 1/2 \times 1/2 = 1/4.$
- The probability of the union of two independent events as follows:
  - $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1) \times P(E_2)$
- For the above coin tossing example, the probability that at least one of the two coins is heads is
  - $P(H_1 \cup H_2) = 1/2 + 1/2 - 1/2 \times 1/2$
    $= 1 - 1/4 = 3/4 = 0.75$

Disjoint vs Independent events

- Events are disjoined (mutually exclusive) if the occurrence of one event excludes the occurrence of the other(s).
  - They cannot happen at the same time.
    - For example: when tossing a coin, the result can either be H or T but cannot be both.
    - Therefore
      - $P(H \cap T) = 0$
      - $P(H \cup T) = P(H) + P(T)$
      - $P(H \mid T) = 0$
      - $P(H \mid T^c) = P(H) / \{1 - P(T)\}$

Disjoint vs Independent events

- Events are independent if the occurrence of one event does not influence (and is not influenced by) the occurrence of the other(s).
  - They can happen at the same time.
    - For example, when tossing two coins, the result can be H_1H_2 , H_1T_2 , T_1H_2 , or T_1T_2.
    - Considering probability of coming H_1H_2:
      - $P(H_1 \cap H_2) = P(H_1)\,P(H_2)$
      - $P(H_1 \cup H_2) = P(H_1) + P(H_2) - P(H_1)\,P(H_2)$
      - $P(H_1 \mid H_2) = P(H_1)$
      - $P(H_1 \mid H_2^c) = P(H_1)$
- This means that disjoint events are not independent, and independent events cannot be disjoint.

Bayes' theorem

- Sometimes, we know the conditional probability of E_1 given E_2, but we are interested in the conditional probability of E_2 given E_1.
- For example, suppose that the probability of having lung cancer is $P(C) = 0.001$ and that the probability of being a smoker is $P(SM) = 0.25$.
- Further, suppose we know that if a person has lung cancer, the probability of being a smoker increases to $P(SM|C) = 0.40$.
- We are, however, interested in the probability of developing lung cancer if a person is a smoker, $P(C|SM)$.

Bayes' theorem

- In general, for two events E_1 and E_2, the following equation shows the relationship between $P(E_2|E_1)$ and $P(E_1|E_2)$:

$$P(E_2|E_1) = \frac{P(E_1|E_2)P(E_2)}{P(E_1)}$$

- This formula is known as Bayes' theorem or Bayes' rule.
- For the above example,

$$P(C|SM) = \frac{P(SM|C)P(C)}{P(SM)} = \frac{0.4 \times 0.001}{0.25} = 0.0016$$

- Therefore, the probability of lung cancer for smokers increases from 0.001 to 0.0016.

Bayes' theorem

- Now suppose that a set of K events B_1, B_2, ..., B_K forms a partition of the sample space.
- We can write the Bayes' theorem for each of the partitioning events as follows:

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A)}$$

- Here, B_i is one of the partitioning events, and A is an event in the sample space.

Bayes' theorem

- Using the law of total probability (slide 34), we have

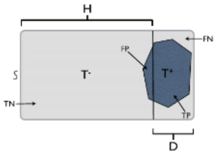
$$P(A) = \sum_{k=1}^{K} P(A|B_k)P(B_k)$$

- Therefore, we can write the general form of Bayes' theorem as

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{k=1}^{K} P(A|B_k)P(B_k)}$$

Application of Bayes' Theorem

- A Venn diagram illustrating a typical medical diagnosis test ("sweat test" to diagnose Cystic Fibrosis)



  - Here, the following abbreviations are used
    - S : sample space,
    - H : healthy,
    - D : diseased,
    - T^- : negative test result,
    - T^+ : positive test result.
  - The true positive TP : The shaded area to the right of vertical line
  - The false positive FP : The shaded area to the left of the vertical line
  - The true negative TN : The unshaded area to the left of the vertical line
  - The false negative FN : The unshaded area to the right of the vertical line

Application of Bayes' Theorem

- The sweat test is a simple procedure to detect CF by measuring the concentration of salt in a person's sweat.
  - A high level of salt above a certain cutoff indicates CF.
- The conditional probability of a positive diagnosis for CF patient, $P(T^+|D)$, is called the sensitivity of the test.
- The conditional probability of a negative result for a healthy person, $P(T^-|H)$, is called the specificity of the test.
- The probability of the CF disease for a child whose parents are both carriers is $P(D) = 0.25$.
  - Note that the gene causing CF is recessive.
- Therefore, if we denote the allele causing CF as *a* and the normal allele as *A*, only people with *aa* genotype have CF.
- People with *Aa* genotype are carriers.
  - If both parents are carriers, the chance of transmitting *a* is 0.5 for each parent

Application of Bayes' Theorem

- Assuming that chromosomes from two parents are transmitted independently, there is the probability $P(D) = 0.5×0.5 = 0.25$ that the child becomes affected (i.e., *aa* genotype).
  - Then, the probability of being healthy is
    - $P(H) = 1 − 0.25 = 0.75$.
- Assuming that the probability of false positive for the sweat test is $P(T^+|H) = 0.04$ and the probability of false negative is $P(T^-|D) = 0.07$
- Because T^+ and T^- are complementary events, we have

$$P(T^-|H) = 1 - P(T^+|H) = 1 - 0.04 = 0.96,$$
$$P(T^+|D) = 1 - P(T^-|D) = 1 - 0.07 = 0.93.$$

Application of Bayes' Theorem

- Now we can calculate the updated probability of the disease knowing that the outcome of the test is positive.
- Using the general form of Bayes' theorem, the conditional probability of the disease given a positive test result is

$$P(D|T^+) = \frac{P(T^+|D)P(D)}{P(T^+|D)P(D) + P(T^+|H)P(H)}$$
$$= \frac{0.93 \times 0.25}{0.93 \times 0.25 + 0.04 \times 0.75} = 0.89.$$

  - Therefore, the positive test result increases the probability of having the disease from $P(D) = 0.25$ to $P(D|T^+) = 0.89$.

Bayesian Statistics

- In the CF diagnosis example discussed, we assigned the probability of 0.25 to the disease event before seeing any new empirical data.
  - This probability is called the prior probability.
    - In this case, the prior probability of disease was $P(D) = 0.25$.
- After obtaining new evidence, namely positive test results, we updated the probability of the disease from $P(D)$ to $P(D|T^{+})$.
  - We call this updated probability the posterior probability.
    - In this case, the posterior probability of the disease was $P(D|T^{+}) = 0.89$
- Therefore, based on the test result, we become more certain that the child is affected by the disease.

Interpretation of Probability as the Relative Frequency

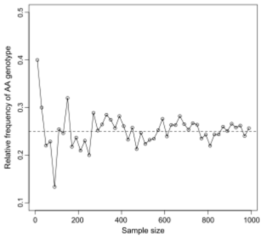
- The random phenomena we have been discussing so far can be observed repeatedly.
  - A coin can be tossed or a die can be rolled many times.
  - We can observe the genotypes of many people.
- These repeated experiments or observations are called trials.
- For such random phenomena, the probability of an event can be interpreted in terms of the relative frequency.
- The above view of probability is the basis of Frequentist Statistics

Interpretation of Probability as the Relative Frequency

- As an example, suppose that the probability of genotype *AA* is $P(AA) = 1/4$.
  - This probability could be interpreted as 1 out of 4 people in the population have genotype *AA*.
- Suppose that we take a simple random sample of size *n* from the population.
  - If the genotype *AA* is observed n_{AA} times in the sample, the relative frequency of *AA* in the sample is n_{AA}/n.
- If our probability assumption is true (i.e., $P(AA) = 1/4$), this sample relative frequency would be approximately 1/4.
  - In this case, as our sample size *n* increases, the sample relative frequency becomes closer to the probability of 1/4;
    - that is, it reaches the probability $P(AA) = 1/4$.

Interpretation of Probability as the Relative Frequency

- Simulation study of the relative frequency of *AA* genotype for different sample size values.



  - The plot shows how the sample relative frequency of *AA* genotype approaches the probability *P(AA)* = 1/4 as the sample size increases.

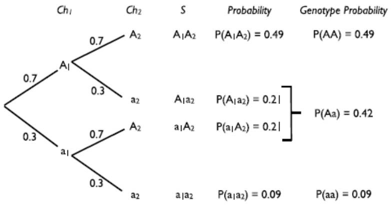
Interpretation of Probability as the Relative Frequency

- Note that the above interpretation of probability requires two important assumptions.
  - We assume that the probability of events does not change from one trial to another.
    - For example, the probability of *AA* must remain 1/4.
      - If the population changes as we are sampling people (e.g., genotype *AA* becomes more prevalent), then the sample relative frequency will not converge to 1/4.
  - We also assume that the outcome of one trial does not affect the outcome of another trial.

Using Tree Diagrams to Obtain Joint Probabilities

- Previously, we used tree diagrams to find the sample space for the combination of two random phenomena.
- Tree diagrams can also be used for calculating their joint probabilities.
- As an example, assume that the alleles on the homologous chromosomes are independent
  - i.e., the allele inherited from the mother has no influence on the allele inherited from the father.
- Also assume that for a biallelic gene **A**, the allele probabilities are $P(A) = 0.7$ and $P(a) = 0.3$.
- Then to find the genotype probabilities, we can use the tree diagram (shown in next slide).

Using Tree Diagrams to Obtain Joint Probabilities



- The first set of branches represents possible alleles for one chromosome (Ch_1), and the second set represents possible alleles for the other chromosome (Ch_2).
- Since these events are independent, knowing the allele on the first chromosome has no influence on the probability of the allele on the second chromosome.

Using Tree Diagrams to Obtain Joint Probabilities

- The sample space is obtained by following a branch from root to tip:
  - $S = \{A_1A_2, A_1a_2, a_1A_2, a_1a_2\}$
- Since these events are independent, their joint probabilities are obtained by multiplying their marginal probabilities:
  - $P(A_1A_2) = 0.7 \times 0.7 = 0.49$
- Likewise, the probability of having a on the first chromosome and allele A on the second chromosome is
  - $P(a_1A_2) = 0.3 \times 0.7 = 0.21$
- Following similar approach, we can find the probability of each possible combination of two chromosomes.
  - These probabilities are given in the column after the sample space in the figure (previous slide).

Using Tree Diagrams to Obtain Joint Probabilities

- The labeling of the chromosomes is arbitrary.
- Therefore, we can drop the indices for A_1A_2 and a_1a_2 and write them as genotypes AA and aa, respectively.
- The genotype Aa can be considered as an event that includes two outcomes,
  - A_1a_2 and a_1A_2.
- Therefore, $P(Aa) = 0.21+0.21 = 0.42$
  - This probability is shown in the last column in the figure (slide 53).

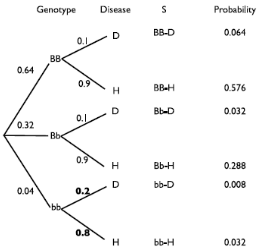
Using Tree Diagrams to Obtain Joint Probabilities

- The above example can be generalized.
- Assume that the probability of observing the A allele is $P(A) = p$ and the probability of observing the a allele is $P(a) = q$.
- Then the genotype probabilities are
  - Homozygous AA: $P(A_1A_2) = p \times p = p^2$,
  - Heterozygous Aa: $P(A_1a_2 \cup a_1A_2) = p \times q + q \times p = 2pq$,
  - Homozygous aa: $P(a_1a_2) = q \times q = q^2$.
- Suppose, for example, that the allele probabilities for gene **B** are $P(B) = 0.8$ and $P(b) = 0.2$ and that the alleles on homologous chromosomes are independent (i.e., they are transmitted from parents independently).
- Then the genotype probabilities are
  - $P(BB) = 0.8 = 0.64$,
  - $P(bb) = 0.2 = 0.04$,
  - $P(Bb) = 2 \times 0.8 \times 0.2 = 0.32$.

Using Tree Diagrams to Obtain Joint Probabilities

- Tree diagrams can also be used to find probabilities when the outcomes are not independent.
- Suppose that gene **B** in previous example is related to a specific disease, but it is not the only factor to determine the disease status.
- In particular, the probability of having the disease is 0.2 for the bb genotype, whereas this probability is 0.1 for the other two genotypes, BB and Bb.
- Therefore, the probability of the disease depends on the genotype.

Using Tree Diagrams to Obtain Joint Probabilities



- The first set of branches represents the genotype, and the second set represents the disease status.
- The probabilities on the first set of branches are for different genotypes: $P(BB) = 0.64$, $P(Bb) = 0.32$, and $P(bb) = 0.04$.
- The probabilities on the second set of branches are conditional probabilities for the disease status given the genotype: $P(D|BB) = 0.1$, $P(D|Bb) = 0.1$, and $P(D|bb) = 0.2$.
- Since the healthy (H) and disease (D) events are complementary, the remaining conditional probabilities are $P(H|BB) = 1 - 0.1 = 0.9$, $P(H|Bb) = 1 - 0.1 = 0.9$, and P(H|bb) = 1 − 0.2 = 0.8.

Using Tree Diagrams to Obtain Joint Probabilities

- Unlike the tree for independent events, the probabilities on the second set of branches depend on the outcomes on the first set of branches.
- As before, we follow the branches from the root to tip and obtain the sample space:
  - $S = \{BB - D, BB - H, Bb - D, Bb - H, bb - D, bb - H\}$.
- To find their probabilities, which are in fact the joint probabilities of genotype and disease status, we multiply the probabilities on the corresponding branches.
- For example, the probability of $Bb - D$ is the product of the conditional probability $P(D|Bb)$ and the marginal probability $P(Bb)$:
  - $P(Bb - D) = P(Bb)P(D|Bb) = 0.32 \times 0.1 = 0.032$.