Statistical Data Analysis

Assist. Prof. Dr. Zeyneb KURT

(Slides have been prepared by
Prof. Dr. Nizamettin AYDIN,
updated by Zeyneb KURT)

zeyneb@yildiz.edu.tr
http://avesis.yildiz.edu.tr/zeyneb/

Probability (cont’d)

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

- As an die rolling example, we define two events.
  - The event *M* occurs when the outcome is less than 4.
  - The event *N* occurs when the outcome is an odd number.
- In this example, *P(M)* = 1/2 and *P(N)* = 1/2

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

- 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*,

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

- 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 homozygous event and the no-disease event is $HM \cap ND = \{AA\}$.
- The intersection of M and N in the dye rolling example is

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

  - In this case, the intersection of the two events includes outcomes that are less than 4 and odd.

Intersection - Example

- For the die rolling example

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

- For the gene-disease example

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

- 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, respectively.
- 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 $P(E_1 \cup E_2)$ is the sum of their marginal probabilities minus their joint probability.
- The union of the homozygous 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 example, if a person is heterozygous, we know that he does not have the disease
  - so the two events HT and D 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, ..., En, 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 $E1$ 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.
- 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

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

Conditional Probability - Example

- Consider the gene-disease example.
- 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

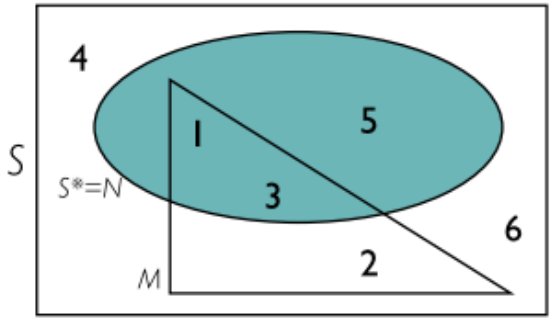
- Now let us find the conditional probability of not having the disease knowing that the person has a homozygous genotype: $P(ND|HM)$.
- The joint probability of *ND* and *HM* is
  - $P(ND \cap HM) = P(\{AA\}) = 0.49$.
- The conditional probability is therefore

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

- 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.



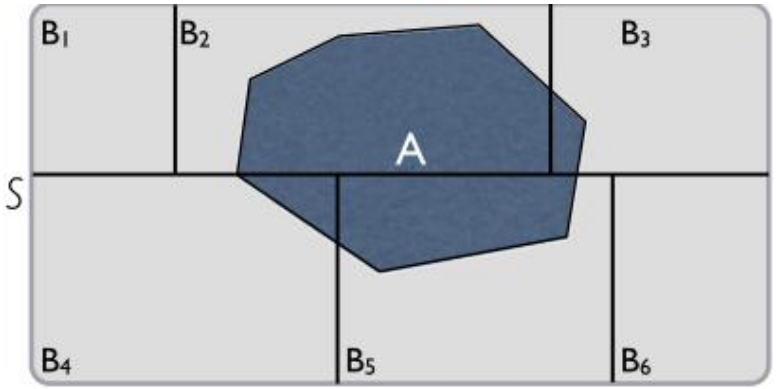
  - Conditioning on an event only reduces the sample space (e.g., from the large rectangle to the shaded oval in in the figure).
- Within this shrunken sample space, all probability rules are valid.
- For example,

$$P\left(E_1^c \middle| E_2\right) = 1 - P(E_1|E_2),$$

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

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
  - $P(E_1 \cap E_2 \cap ... \cap E_n) = P(E_1) \times P(E_2) \times ... \times P(E_n)$

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
      - $P(H \cap T) = 0$
      - $P(H \cup T) = P(H) + P(T)$
      - $P(H \mid T) = 0$
      - $P(H \mid T^c) = P(H) / \{1 - P(T)\} = 1$

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, we have

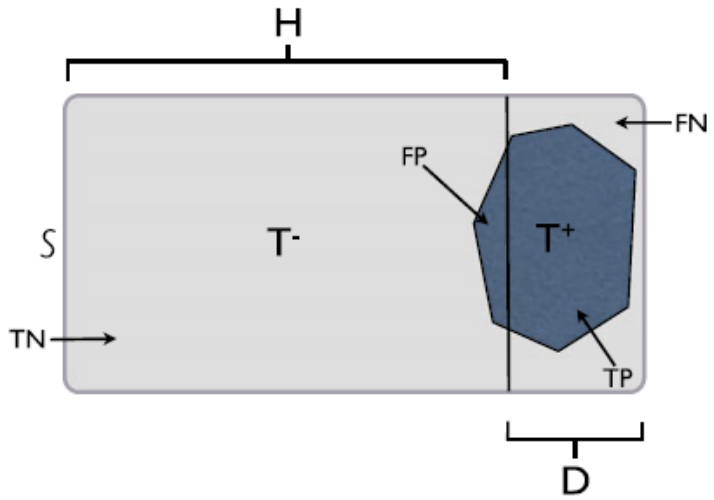
$$P(A) = \sum_{k=1}^{K} P(A|B_k)P(B_k)$$

- 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 $P(D) = 0.5 \times 0.5 = 0.25$ 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

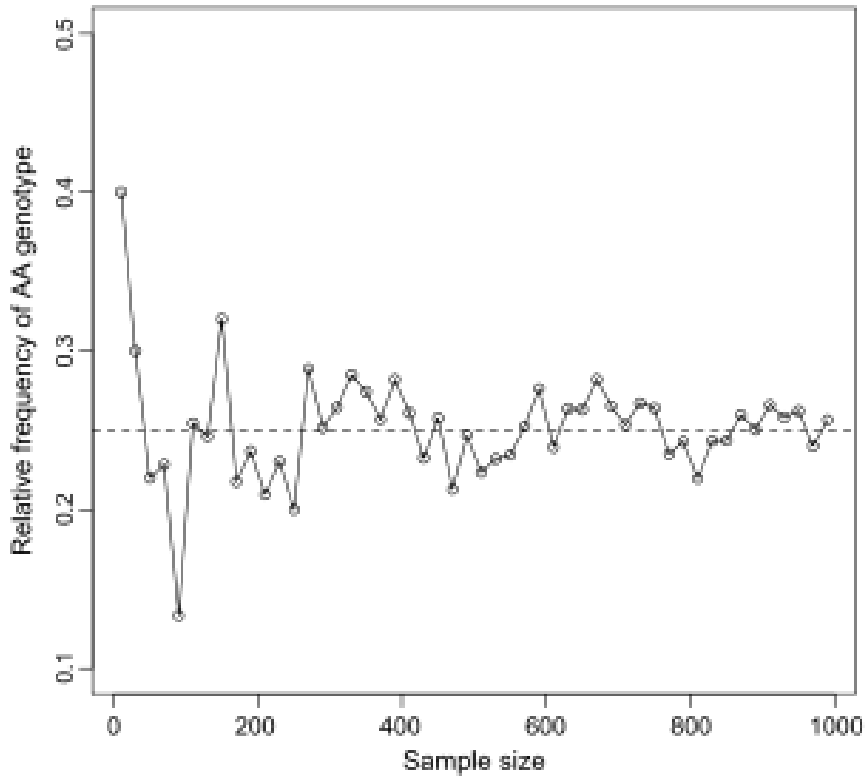
- The random phenomena we have been discussing so far can be observed repeatedly.
  - A coin can be tossed or a die can be rolled many times.
  - We can observe the genotypes of many people.
- These repeated experiments or observations are called trials.
- For such random phenomena, the probability of 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.

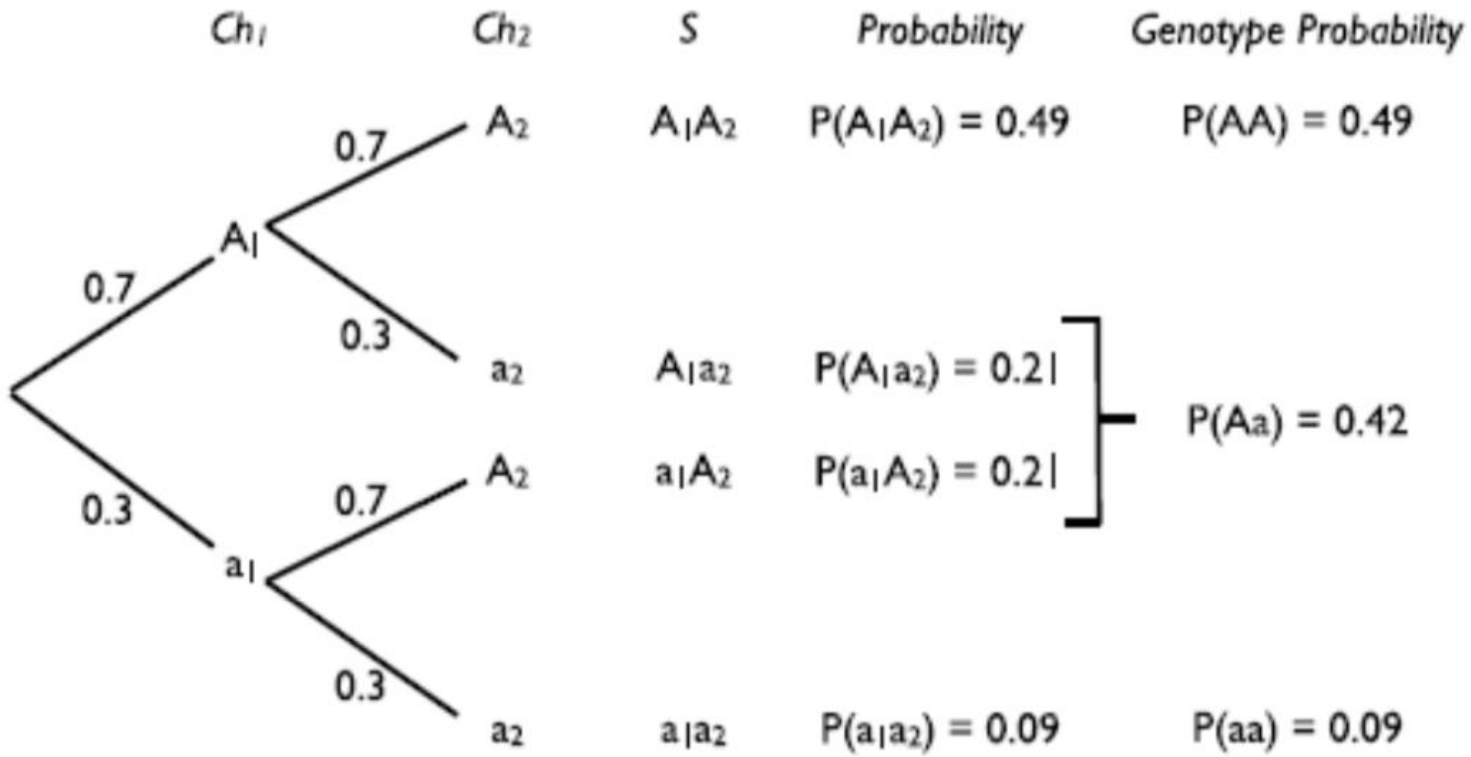
Interpretation of Probability as the Relative Frequency

- 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.

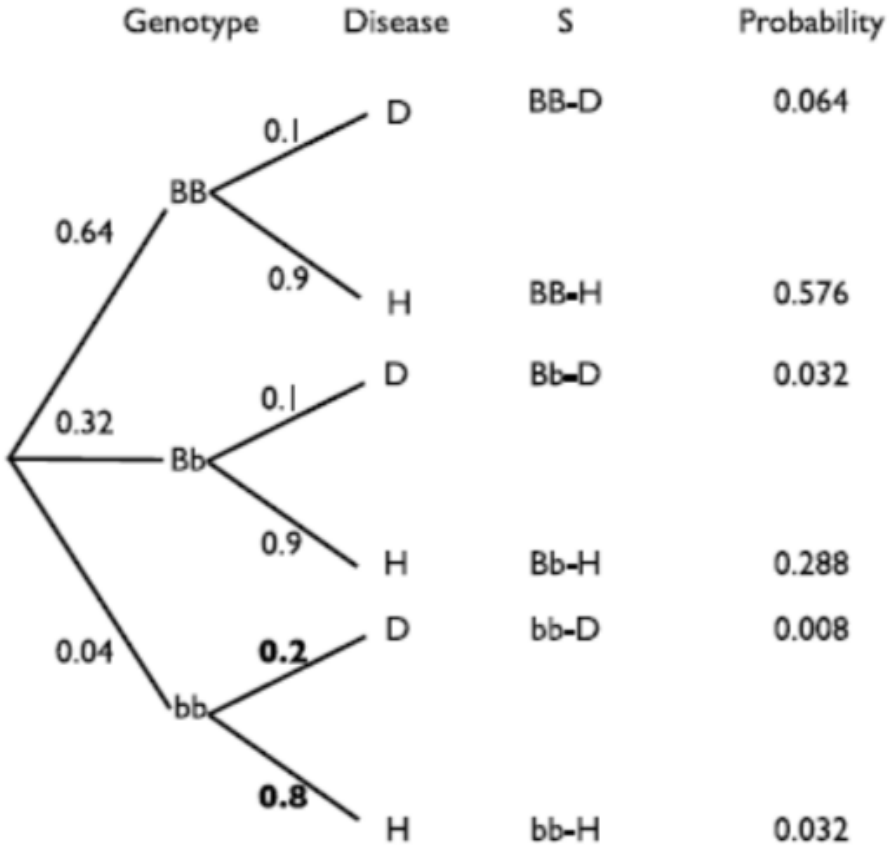
Using Tree Diagrams to Obtain Joint Probabilities

- 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
  - $P(BB) = 0.8\text{^}2 = 0.64$,
  - $P(bb) = 0.2\text{^}2 = 0.04$,
  - $P(Bb) = 2\times0.8\times0.2 = 0.32$.

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 $P(Bb)$:
  - $P(Bb - D) = P(Bb)P(D|Bb) = 0.32 \times 0.1 = 0.032$.

Random Variables
and
Probability Distributions

Random variables

- We are interested in calculating the probabilities associated with both quantitative and qualitative events.
- For example,
  - we can determine the probability that a machinist selected at random from the workers in a large automotive plant would suffer an accident during an 8-hour shift.
  - We can also find the probability that a machinist selected at random would work more than 80 hours without suffering an accident.
- These qualitative and quantitative events can be classified as events (or outcomes) associated with qualitative and quantitative variables.

Qualitative Random variables

- For example,
  - in the automotive plant accident study, the randomly selected machinist's accident report would consist of checking one of the following:
    - No Accident, Minor Accident, or Major Accident.
  - Thus, the data on 100 machinists in the study would be observations on a qualitative variable because the possible responses are the different categories of accident and are not different in any measurable, numerical amount.
- Because we cannot predict with certainty what type of accident a particular machinist will suffer, the variable is classified as a qualitative random variable.

Qualitative Random variables

- Other examples of qualitative random variables that are commonly measured are
  - political party affiliation,
  - socioeconomic status,
  - the species of insect discovered on an apple leaf,
  - the brand preferences of customers.
  - …
- There are a finite (and typically quite small) number of possible outcomes associated with any qualitative variable.

Quantitative Random variables

- Many times the events of interest in an experiment are quantitative outcomes associated with a quantitative random variable, since the possible responses vary in numerical magnitude.
  - For example, in the automotive plant accident study, the number of consecutive 8-hour shifts between accidents for a randomly selected machinist is an observation on a quantitative random variable.
  - Events of interest, such as the number of 8-hour shifts between accidents for a randomly selected machinist, are observations on a quantitative random variable.

Quantitative Random variables

- Other examples of quantitative random variables are:
  - the change in earnings per share of a stock over the next quarter,
  - the length of time a patient is in remission after a cancer treatment,
  - the yield per acre of a new variety of wheat,
  - the number of persons voting for the incumbent in an upcoming election.
  - …

Random variables

- Formally, a random variable X assigns a numerical value to each possible outcome (and event) of a random phenomenon.
- For instance, we can define X based on possible genotypes of a bi-allelic gene **A** as follows:

$$X = \begin{cases} 0 & \text{for genotype } AA, \\ 1 & \text{for genotype } Aa, \\ 2 & \text{for genotype } aa. \end{cases}$$

- In this case, the random variable assigns 0 to the outcome AA, 1 to the outcome Aa, and 2 to the outcome aa.

Random variables

- The way we specify random variables based on a specific random phenomenon is not unique.
- Alternatively, we can define a random variable Y as:

$$Y = \begin{cases} 0 & \text{for genotypes } AA \text{ and } aa, \\ 1 & \text{for genotype } Aa. \end{cases}$$

- In this case, Y assigns 0 to the homozygous event and assigns 1 to the heterozygous event.

Random variables

- When the underlying outcomes are numerical, the values the random variable assigns to each outcome can be the same as the outcome itself.
  - For the die Rolling example, we can define a random variable Z to be equal to $1, 2, \ldots, 6$ for outcomes $1, 2, \ldots, 6$, respectively.
  - Alternatively, we can define a random variable W and set W to 1 when the outcome is an odd number and to 2 when the outcome is an even number.
- The set of values that a random variable can assume is called its range.
  - For the above examples, the range of X is $\{0, 1, 2\}$, and the range of Z is $\{1, 2, \ldots, 6\}$.

Random variables

- After we define a random variable, we can find the probabilities for its possible values based on the probabilities for its underlying random phenomenon.
- This way, instead of talking about the probabilities for different outcomes and events,
  - we can talk about the probability of different values for a random variable.
- {For example,
  - suppose $P(AA) = 0.49$, $P(Aa) = 0.42$, and $P(aa) = 0.09$.
  - Then, we can say that $P(X = 0) = 0.49$,
    - i.e., X is equal to 0 with probability of 0.49.}
  - Note that the total probability for the random variable is still 1.

Random variables

- The probability distribution of a random variable specifies its possible values (i.e., its range) and their corresponding probabilities.
  - For the random variable X defined based on genotypes, the probability distribution can be simply specified as follows:

$$P(X = x) = \begin{cases} 0.49 & \text{for } x = 0, \\ 0.42 & \text{for } x = 1, \\ 0.09 & \text{for } x = 2. \end{cases}$$

    - Here, x denotes a specific value (i.e., 0, 1, or 2) of the random variable.

Discrete vs. continuous random variables

- We divide the random variables into two major groups:
  - discrete and continuous.
- When observations on a quantitative random variable can assume only a countable number of values, the variable is called a discrete random variable.
  - These variables can be categorical (nominal or ordinal), such as genotype, or counts, such as the number of patients visiting an emergency room per day

Discrete vs. continuous random variables

- When observations on a quantitative random variable can assume any one of the uncountable number of values in a line interval, the variable is called a continuous random variable.
  - Typical continuous random variables are temperature, pressure, height, weight, and distance.
- The distinction between discrete and continuous random variables is pertinent when we are seeking the probabilities associated with specific values of a random variable.

Probability distribution

- The probability distribution of a random variable provides the required information to find the probability of its possible values.
- We need to know the probability of observing a particular sample outcome in order to make an inference about the population from which the sample was drawn.
- To do this, we need to know the probability associated with each value of the variable.
- Viewed as relative frequencies, these probabilities generate a distribution of theoretical relative frequencies called the probability distribution of the variable.

Probability distribution

- Probability distributions differ for discrete and continuous random variables.
  - For discrete random variables, we will compute the probability of specific individual values occurring.
  - For continuous random variables, the probability of an interval of values is the event of interest.
- The probability distributions discussed here are characterized by one or more parameters.
- The parameters of probability distributions we assume for random variables are usually unknown.

Probability distribution

- Typically, we use Greek alphabets such as μ and σ to denote these parameters and distinguish them from known values.
  - We usually use μ to denote the mean of a random variable and use σ^2 to denote its variance.
- For a population of size N, the mean and variance are calculated as follows:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} \qquad \sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

Discrete probability distributions

- The probability distribution for a discrete random variable displays the probability $P(y)$ associated with each value of y.
  - This display can be presented as a table, a graph, or a formula.
- The probability distribution of a discrete random variable is fully defined by the probability mass function (pmf).
  - This is a function that specifies the probability of each possible value within range of random variable.

Discrete probability distributions

- For the genotype example, the pmf of the random variable X is

$$P(X = x) = \begin{cases} 0.49 & \text{for } x = 0, \\ 0.42 & \text{for } x = 1, \\ 0.09 & \text{for } x = 2. \end{cases}$$

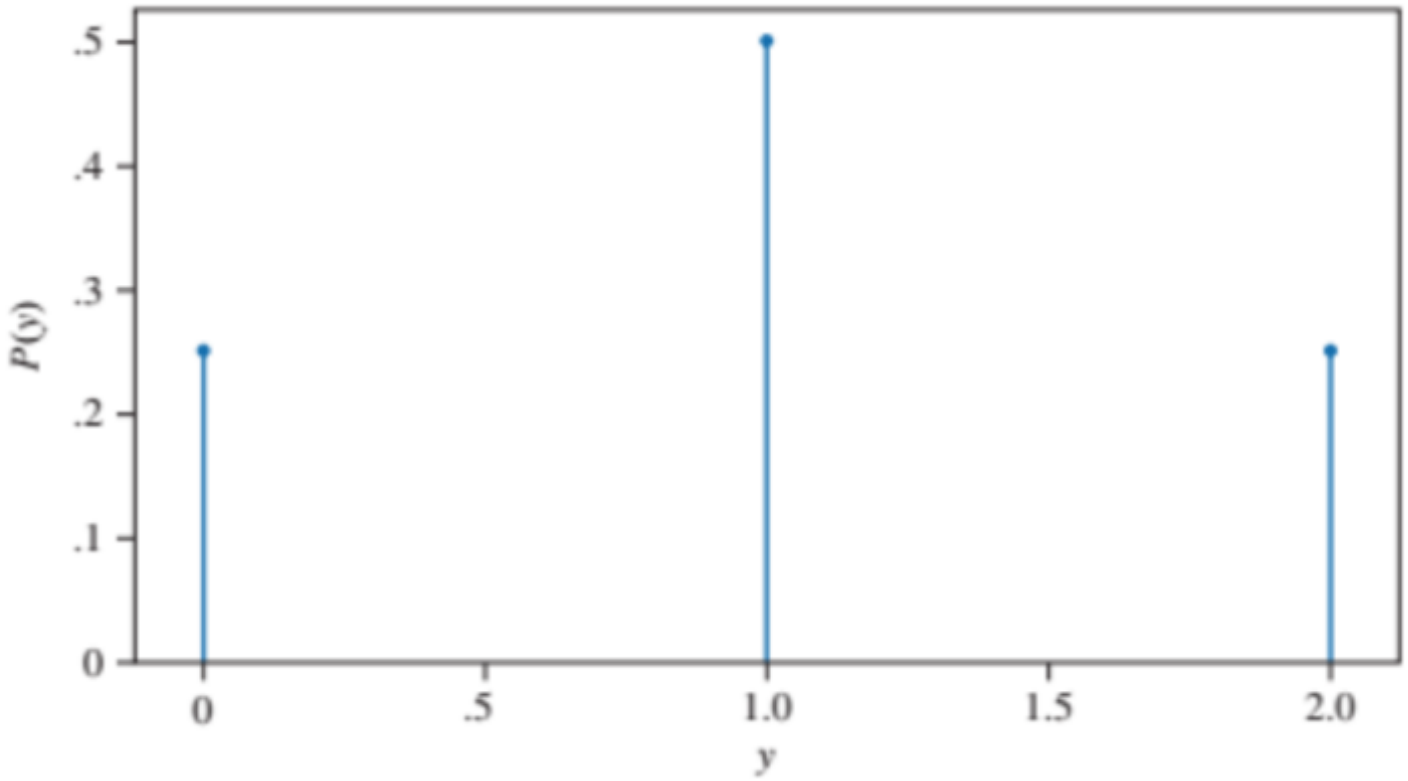
- As another example, suppose Y is a random variable that is equal to 1 when a newborn baby has low birthweight, and is equal to 0 otherwise.
  - We say Y is a *binary* random variable.
- Further, assume that the probability of having a low birthweight for babies is 0.3.
  - Then the pmf for the random variable Y is

$$P(Y = y) = \begin{cases} 0.7 & \text{for } y = 0, \\ 0.3 & \text{for } y = 1. \end{cases}$$

Discrete probability distributions

- Example:
  - Probability distribution for the number of heads when two coins are tossed

y	$P(y)$
0	.25
1	.50
2	.25



Properties of Discrete Random Variables

- The probability distribution for the discrete random variable given in previous slide illustrates three important properties of discrete random variables.
  - The probability associated with every value of y lies between 0 and 1.
  - The sum of the probabilities for all values of y is equal to 1.
  - The probabilities for a discrete random variable are additive.
    - Hence, the probability that $y = 1$ or 2 is equal to $P(1) + P(2)$

Bernoulli Distribution

- Binary random variables are abundant in scientific studies.
  - Examples include disease status (healthy and diseased), gender (male and female), survival status (dead, survived), and a gene with two possible alleles (A and a).
- The binary random variable X with possible values 0 and 1 has a Bernoulli distribution with parameter θ,
  - where, $P(X = 1) = \theta$ and $P(X = 0) = 1 - \theta$.
- We denote this as $X \sim$ Bernoulli (θ), where $0 \leq \theta \leq 1$.
  - Here θ is unknown parameter.
- If θ were known, we could fully specify the probability mass function:

$$P(X = x) = \begin{cases} 1 - \theta & \text{for } x = 0 \\ \theta & \text{for } x = 1 \end{cases}$$

Bernoulli Distribution

- For example, let X be a random variable representing the five-year survival status of breast cancer patient,
  - where $X = 1$ if the patient survived and $X = 0$ otherwise.
- Suppose that the probability of survival is $\theta = 0.8$: $P(X = 1) = 0.8$
- Therefore, the probability of not surviving is

$$P(X = 0) = 1 - \theta = 0.2$$

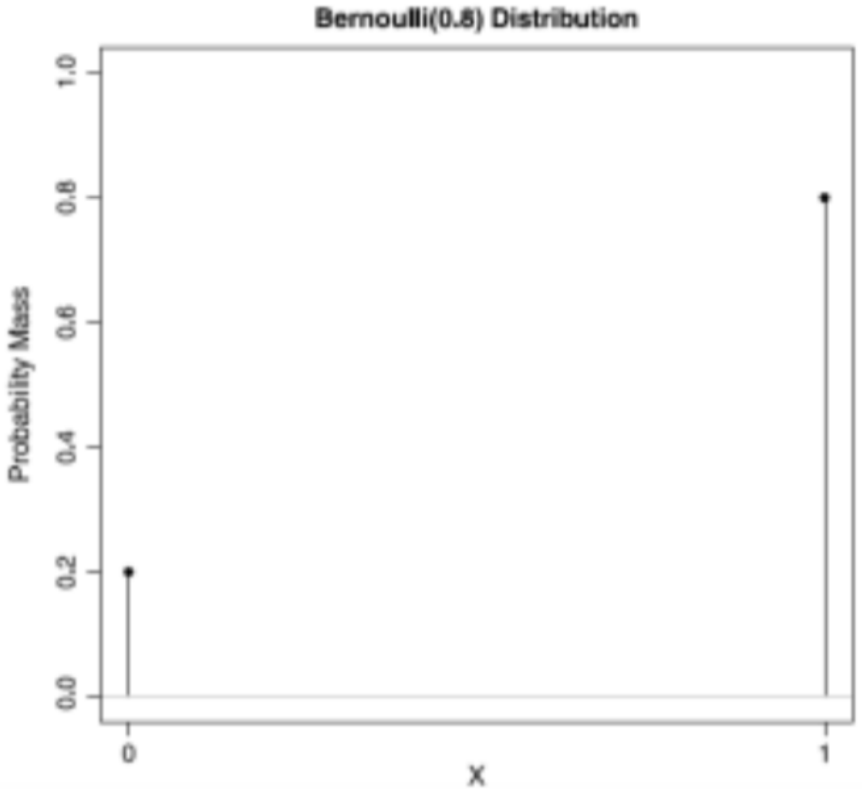
- Then X has a Bernoulli distribution with parameter $\theta = 0.8$, and we denote this as

$$X \sim \text{Bernoulli}(0.8).$$

- The pmf for this distribution is

$$P(X = x) = \begin{cases} 0.2 \; for \; x = 0 \\ 0.8 \; for \; x = 1 \end{cases}$$

- Plot of the pmf for Bernoulli(0.8) distribution



Bernoulli Distribution

- The mean of a binary random variable, X, with Bernoulli(θ) distribution is θ .
  - We show this as $\mu = \theta$.
    - In this case, the mean can be interpreted as the proportion of the population who have the outcome of interest.
- The variance of a random variable with Bernoulli(θ) distribution is

$$\sigma^2 = \theta(1 - \theta) = \mu(1 - \mu)$$

- The standard deviation is obtained by taking the square root of variance

$$\sigma = \sqrt{\theta(1 - \theta)} = \sqrt{\mu(1 - \mu)}$$

Bernoulli Distribution

- In the above example, $\mu = 0.8$.
  - 80% of patients survive.
- The variance of the random variable is

  $$\sigma^2 = 0.8 \times 0.2 = 0.16,$$

- Its standard deviation is $\sigma = 0.4$.
- This reflects the extent of variability in survival status from one person to another.
  - For this example, the amount of variation is rather small.
    - Therefore, we expect to see many survivals ($X = 1$) with occasional death ($X = 0$).

Bernoulli Distribution

- For comparison, suppose that the probability of survival for bladder cancer is $\theta = 0.6$.
- Then, the variance becomes

$$\sigma^2 = 0.6 \times (1 - 0.6) = 0.24.$$

- This reflects a higher variability in the survival status for bladder cancer patients compared to that of breast cancer patients.