STAT 230

PROBABILITY

PROFESSOR N. MOHAMMAD • WINTER 2014 • UNIVERSITY OF WATERLOO

Last Revision: March 4, 2015

Table of Contents

Future Modifications

1 Introduction

Definition 1.1. Probability is the study of randomnesses and uncertainty.

- Variability in population, processes or phenomena

Definition 1.2. Classic Interpretatation makes assumptions about the physical world to deduce probability.

$$\frac{\text{\# ways an event can occur}}{\text{Total \# of outcomes}}$$

Definition 1.3. Relative-frequency interpretation is when the probability of specific outcome is defined as the proportion of times it occurs over the long run. May only be used if the experiment may be repeated.

Definition 1.4. Frequency is just the amount of times an event occurs while **relative frequency** is a fraction of the amount of times an event occurs over the total amount of possible outcomes.

Definition 1.5. Personal-probability interpretation is the degree to which a given individual believes the event wil happen.

Quote. Coherent means that personal probability of one event does not contradict personal probability of another.

Example 1.1. If the probability of finding a parking space is 0.2, the probability of not finding one should be 0.8.

1.1 Probability Models

Definition 1.6. Experiment is any action, phenomenon, or process whose outcome is subject to uncertainty.

Definition 1.7. Trial is a single repetition of an experiment.

Definition 1.8. Sample space, denoted by S, is the set of possible distinct outcomes. In a single trial, only one outcome may occur. The sample space may be either **discrete** or **continuous**.

Definition 1.9. An **event** is any subset of outcomes contained in the sample space S.

- **Simple** - one outcome
- **Compound** - more than one outcome

Probability notation

$P(\text{event})$ is used to denote the probability of an event occurring. The **compliment** is the probability of an event not happening and is denoted as $P(A^c)$ or $P(\overline{A})$

Definition 1.10. The odds in **favour** of an event is the odds of an event occurring compared to its compliment.

$$\frac{P(A)}{P(A^c)} = \frac{P(A)}{1 - P(A)}$$

The odds against is the reciprocal of odds in favour.

Theorem 1.1.

$$\sum_{i=0}^{k} P(A_i) = 1$$

The set $P(A_i), i = 1, 2, \ldots$ is the probability distribution on S.

Theorem 1.2. If A and B are two events wirh $A \subseteq B$, then $P(A) \leq P(B)$

Definition 1.11. Two events are **mutually exclusive** or **disjoint** if they cannot happen simultaneously.

$$A \cap B = \emptyset$$

1.2 Counting Techniques

Counting Principle • Permutations • Combinations

Definition 1.12. Addition rule: When there are m ways to perform A, and n ways to perform B, there are $m + n$ ways to perform A **OR** B.

Definition 1.13. Product rule: Where there are p ways to perform A, and q ways to perform B, there are $m \times n$ ways to perform A **AND** B.

Definition 1.14. Uniform Probability Model:

$$P(A) = \frac{\text{Outcomes in } A}{\text{Outcomes in } S}$$

Definition 1.15. Permutation is an arrangement of elements in an ordered list.

Note. The amount of ways to arrange n items (all of them have to be used) is $n!$

$$P_{k,n} = \frac{n!}{(n-k)!}$$

Definition 1.16. Any unordered sequence of k objects taken from a set of n distinct objects is called a **combination**.

$$C_{k,n} = \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Example 1.2. Binomial Theorem:

$$(1+x)^n = \binom{n}{0} + \binom{n}{1}x + \binom{n}{2}x^2 + \cdots + \binom{n}{n}x^n$$

2 Chapter 4

Definition 2.1. The **union** of two events A and B, $A \cup B$ is the set containing all outcomes in either A or B.

Definition 2.2. The **intersection** of two events A and B, $A \cup B$ is the set containing all outcomes that are in both A and B.

Definition 2.3. The **empty set**, $\emptyset$ is the set containing no outcomes.

Definition 2.4. The probability of **either** event happening is the sum of their individual probabilities:

$$P(A \cup B) = P(A) + P(B)$$

only if A and B are mutually exclusive.

Note. For non-mutually exclusive events, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - [P(A \cap B) + P(A \cap C) + P(B \cup C)] + P(A \cap B \cap C)$$

Definition 2.5. Two events are **independent** if and only if

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

Definition 2.6. The conditional probability of A given B is defined by

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

Note. If two events A and B are independent, then

$$P(A) = P(A|B)$$

Quote. Show all steps to combinations and permutation. (formulas)

Definition 2.7. A set of events is **exhaustive** when at least one of the events must occur.

Disease Problems

Definition 2.8. Terminology:

- **False positive** is when test indicates positive and is wrong (true status is negative). $(T \mid D^c)$
- **False negative** is when a test indicates a negative status and is wrong (true status is positive). $(T^c \mid D)$
- **Sensitivity/ True positive** is when a test indicates a positive status and is correct. $(T \mid D)$
- textbfSpecificity/ True negative is when a test indicates a negative status and is wrong (true status is negative) $(T^c \mid D^c)$

Example 2.1. A cheap blood test for HIV has the following characteristics:

- False positive rate is 0.5%
- False negative rate is 2%
- Around 0.04% of Canadian males are infected with HIV

Determine the probability that if a male tests positive for HIV, he actually has HIV.

Solution. Let D represent has disease and T represent test positive. We are looking for $P(D \mid T)$. This is equivalent to

$$P(D \mid T) = \frac{P(DT)}{P(T)}$$

Given information:

$$P(D) = 0.0004, P(D^c) = 0.9996$$
$$P(T^c \mid D) = 0.02, P(T \mid D) = 0.98$$
$$P(T \mid D^c) = 0.005, P(T^c \mid D^c) = 0.995$$

Now for the calculations

$$P(DT) = P(T \mid D) \cdot P(D) = 0.98 \cdot 0.0004 = 0.000392$$

$$P(T) = P(DT) + P(D^cT) = P(DT) + P(T \mid D^c) \cdot P(D^c) = 0.000392 + 0.005 \cdot 0.9996 = 0.00539$$
$$P(D \mid T) = \frac{0.000392}{0.00539} = 0.0727$$

Theorem 2.1. Bayes's Theorem states: Let $A_1, A_2, \ldots, A_k$ be mutually exclusive with prior probabilities $P(A_i)\, i = 1, 2, \ldots, k$. Then for any other event B where $P(B) > 0$, the positerior probability of A_j given that B has occurred is

$$P(A_j|B) = \frac{P(A_J \cap B)}{P(B)}$$

3 Distributions

Definition 3.1. A **variable** is any characteristic whose value may change from one object to another in the population. They are denoted by lowercase letters.

Types of Variables

Definition 3.2. **Numeric variable** is a quantitative variable.

- **Continuous variable**: Can take values consisting of an entire interval ith infinite number of real values. (eg: time)
- **Discrete variable**: Can only take finite number of real values (even if limits may approach infinity). (eg number of people)

Definition 3.3. **Categorical variables** have values that describe a quality or characeristic of a data unit.

- **Ordinal variables** take on values that can be ordered/ranked. (grades, clothing size)
- **Nominal variables** take on value that are not able to be sequentialized. (gender)

Definition 3.4. **Univariate Data** only involve a single variable. Main purpose of these is for central tendency analysis.

Definition 3.5. **Bivariate data** involves two variables, and deals with causes or relationships. Analysis of variables simultaneously for correlation.

Definition 3.6. **Multivariate data** involve more than two variables for more in depth analysis.

Definition 3.7. A **random variable** is a function whose domain is the sample space and whose range is the set of possible values of the variable. It maps each outcome in the sample space with an outcome based on what the r.v. represents.

Example 3.1. Let X denote number of heads in two coin flips. $P(X=0)=\frac{1}{4}$.

Sample space: $\{TT, HT, TH, HH\}$

$$X = \begin{cases} 0: & TT \\ 1: & HT, TH \\ 2: & HH \end{cases}$$

Definition 3.8. The **probability function** of a random variable, X is a function

$$f(x) = P(X=x) \qquad \forall x \in \mathbb{S}$$

Properties include:

$$f(x) \geq 0 \qquad \forall x \in \mathbb{S}$$
$$\sum_{x \in \mathbb{S}} f(x) = 1$$

Definition 3.9. A **probability distribution** of X is a description of the probabilities associated with all the possible values of X. It shows how the total probably of 1 is distributed among the various values of X.

Definition 3.10. The **cumulative distribution function** is defined by

$$F(x) = P(X \leq x) = \sum_{i}^{x} f(i)$$

For a number x, $F(x)$ is the probability that the observed value will be at most x.

Method 3.1. To calculate the cdf, first $P(x)$ must be determined for all values of x. Then add each $f(x)$ to all the previous ones to obtain $P(x)$. For example if the range was $\{0, 1, 2\}$

$$F(1) = P(X \leq 1) = P(1) + (0)$$

$$F(x) = \begin{cases} \# : & x < 0 \\ \# : & 0 \leq x < 1 \\ \# : & 1 \leq x < 2 \end{cases}$$

Note. For any number a and b with $a \leq b$,

$$P(a \leq X \leq b) = F(b) - F(a-)$$

$a-$ represents the largest X value that is strictly less than a.

Note. In addition: $f(x) = F(x) - F(x-1)$

3.1 Uniform Distribution

Definition 3.11. Uniform distribution is where each outcome has equal probability. If the range was from a to b,

$$P(X = x) = \begin{cases} \frac{1}{b-a+1} = \frac{1}{n} & : x \in [a, b] \\ 0 & : x \notin [a, b] \end{cases}$$

Examples of this distribution include rolling a fair die, flipping a fair coin.

3.2 Binomial Distribution

Definition 3.12. Binomial distributions are based on the probability experiments for which the results of each trial can be either a success or faillure. Requirements include:

- Experiment is repeated for fixed number of trials
- Only two possible outcomes for each trial
- Trials are independent

- Probability of success is fixed
- If sampling is without replacement, only use binomial distribution if the sample size is at most 5% of the population size.

Formula

- p denotes the probability of success in a trial.
- $q = 1 - p$ denotes the probability of failure in a trial.
- X is the range of successes in n tries (where $x = 0, \cdots, n$)

$$P(X = x) = b(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x},\ 0 \leq x \leq n$$

We write $x \sim$ Binomial(n, p).

On chapter 5c

4 cpt 7, pg 35

Note. Mean is denoted as μ.

Definition 4.1. Variance is a measure of the spread of the recorded values on a variable. It is a measure of dispersion.

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$