Probability and Statistics by DeGroot Notes

Nathan Ueda

March 4, 2024

Contents

1 Introduction to Probability

1.1 The History of Probability

The use of probability to measure uncertainty and variability dates back hundredes of years.

1.2 Interpretations of Probability

Probability Interpretations

- Frequency: If an experiment is carried out many times, the frequency with which a particular outcome occurred would define its probability.
- Classical: If an outcome of some experiment must be one of n different, equally likely outcomes, the probability of each outcome is $\frac{1}{n}$.
- Subjective: An entity assigns probabilities to each possible outcome.

Probability theory does not depend on intepretation.

1.3 Experiments and Events

Probability allows us to quantify how likely an outcome is to occur.

Experiments: Any process in which the possible outcomes can be identified ahead of time.
Events: A well defined set of possible outcomes of the experiment (such as rolling an even number on a fair dice).

Although there is controversy in regard to the proper meaning and interpration of some of the probabilities that are assigned to the outcomes of many experiements, once these probabilities are assigned, there is complete agreemenet upon the mathematical theory of probability.

Almost all work in the mathematical theory of probability is related to:

- Methods for determining probabilities of certain events from given probabilities for each possible outcome in an experiment.
- Methods for revising probabilities of events when additional relevant information is obtained.

1.4 Set Theory

Sample Space: The collection of all possible outcomes of an experiment.
Empty Set: Subset of S containing no elements, denoted $\emptyset$, representing any events that cannot occur.
Complement: For some set A, its complement, denoted A^c, is the set containing all elements of S not in A.
Union: For n sets $A_1, \ldots, A_n$, their union, denoted $A_1 \cup \ldots \cup A_n$ or $\bigcup_{i=1}^{n} A_i$, is defined as the set containing all outcomes that belong to at least one of these n sets.
Intersection: For n sets $A_1, \ldots, A_n$, their intersection, denoted $A_1 \cap \ldots \cap A_n$ or $\bigcap_{i=1}^{n} A_i$, is defined as the set containing the elements common to all these n sets.
Disjoint/Mutually Exclusive: Two sets A and B are disjoint/mutually exclusive if they have no outcomes in common, that is, if $A \cap B = \emptyset$, representing that both A and B cannot occur.

1.5 The Definition of Probability

Axioms of Probability:

1. For every event A, $P(A) \geq 0$.
2. $P(S) = 1$.
3. $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

Basic Theorems:

1. $P(\emptyset) = 0$.
2. For every finite sequence of n disjoint events, $A_1, \ldots, A_n$, $P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$.
3. For every event A, $P(A^c) = 1 - P(A)$.
4. If $A \subset B$, then $P(A) \leq P(B)$.
5. For every event A, $0 \leq P(A) \leq 1$.
6. For every two events A and B, $P(A \cap B^c) = P(A) - P(A \cap B)$.
7. For every two events A and B, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
8. Bonferroni Inequality: For all events $A_1, \ldots, A_n$, $P(\bigcap_{i=1}^{n} A_i) \geq 1 - \sum_{i=1}^{n} P(A_i^c)$.

1.6 Finite Sample Spaces

Simple Sample Space

- Has a finite number (n) of possible outcomes.
- Each outcome has an equal probability ($\frac{1}{n}$).
- If an event A has m outcomes, then $P(A) = \frac{m}{n}$.

1.7 Counting Methods

Multiplication Rule: An experiment with k parts where the ith part has n_i possible outcomes (regardless of which specific outcomes have occurred in the other parts) has a sample space $S = n_1 n_2 \ldots n_k$.

Permutations ($P_{n,k}$):

- Number of ways to arrange a set (order matters).
- Sampling considering n different items and making k choices from them.
  - Sampling with replacement: n^k.
  - Sampling without replacement: $n(n-1)\ldots(n-k+1)$.
    - n options for first choice, $n-1$ options for second choice, $n-k+1$ options for kth choice.

- The number of permutations of n different items is $P_{n,n} = n!$.
- The number of permutations of n different items making k choices ($0 \leq k \leq n$) is

$$P_{n,k} = n(n-1)\dots(n-k+1)$$

$$P_{n,k} = n(n-1)\dots(n-k+1)\left(\frac{1}{1}\right)$$

$$P_{n,k} = n(n-1)\dots(n-k+1)\left(\frac{(n-k)(n-k-1)\dots 1}{(n-k)(n-k-1)\dots 1}\right)$$

$$P_{n,k} = \frac{n(n-1)\dots(n-k+1)(n-k)(n-k-1)\dots 1}{(n-k)(n-k-1)\dots 1}$$

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

1.8 Combinatorial Methods

Combinations ($C_{n,k}$):

- Number of subsets (order does not matter).
- Permutations may be thought of as combinations of size k chosen out of n, multiplied by the number of ways to arrange the size k subsets, $k!$. More formally, this says

$$P_{n,k} = C_{n,k}k!$$

- Combinations (binomial coefficient) are the number of distinct subsets of size k that can be chosen from a set of size n (this is the same formula as for permutations, except we are dividing out the number of ways we can rearrange the subsets, $k!$, since order does not matter):

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

- Combinations without replacement: $\binom{n+k-1}{k}$.

1.9 Multinomial Coefficients

The total number of different ways of dividing n elements into k groups is

$$\binom{n}{n_1, n_2, \dots, n_k} = \binom{n}{n_1}\binom{n-n_1}{n_2}\binom{n-n_1-n_2}{n_3}\cdots\binom{n-n_1-\cdots-n_{k-2}}{n_{k-1}} = \frac{n!}{n_1!n_2!\cdots n_k!}$$

1.10 The Probability of a Union of Events

- Union of two events A_1 and A_2:

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$$

- Union of three events A_1, A_2, and A_3:

$$P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3) - [P(A_1 \cap A_2) + P(A_1 \cap A_3) + P(A_2 \cap A_3)]$$
$$+P(A_1 \cap A_2 \cap A_3)$$

- Union of n events $A_1, \ldots, A_n$:

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) -$$
$$\sum_{i<j<k<l} P(A_i \cap A_j \cap A_k \cap A_l) + \cdots + (-1)^{n+1} P(A_1 \cap A_2 \cap \cdots \cap A_n)$$

2 Conditional Probability

2.1 The Definition of Conditional Probability

- Conditional probability is the updating of probabilities when certain events are observed.
- The updated probability of event A after we learn that event B has occured is the conditional probability of A given B, denoted $P(A|B)$.
- When we go from $P(A)$ to $P(A|B)$, we say we are conditioning on B.
- For $P(B) > 0$

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

.

- Intuitively, this is saying the probability of A occurring, given B has occurred is equal to the outcomes where both A and B occurred (which makes sense since we know B has occurred and we want to find the probability of A also occurring) divided by the probability of B occurring (which makes sense since we know B occurred, we can renormalize the sample space to only contain outcomes where B occurred).

Multiplication Rule for Conditional Probabilities

- For 2 events A, B
  - If $P(B) > 0$

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

  - If $P(A) > 0$

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

- For n events $A_1, \ldots, A_n$ such that $P(A_1 \cap \ldots \cap A_{n-1}) > 0$

$$P(A_1 \cap \ldots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \ldots P(A_n|A_1 \cap \ldots A_{n-1})$$

- For n events $A_1, \ldots, A_n, B$ such that $P(B) > 0$ and $P(A_1 \cap \ldots \cap A_{n-1}) > 0$

$$P(A_1 \cap \ldots \cap A_n|B) = P(A_1|B)P(A_2|A_1 \cap B)P(A_3|A_1 \cap A_2 \cap B) \ldots P(A_n|A_1 \cap \ldots A_{n-1} \cap B)$$

Law of Total Probability

- Tells us that to get the unconditional probability of A, we can divide the sample space into disjoint slices B_j, find the conditional probability of A within each of these slices, then take a weighted sum of the conditional probabilities, where the weights are the probabilities $P(B_j)$.
- Often used in tandum with Bayes' Rule.
- Relates conditional probability to unconditional probability .
- Partition: Let S denote the sample space and consider k events $B_1, \ldots, B_k$ in S such that $B_1, \ldots, B_k$ are disjoint and $\bigcup_{i=1}^{k} B_i = S$. Then events $B_1, \ldots, B_k$ form a partition in S. In other words, only one of these events can occur and combined they fill the entire sample space.
- Suppose events $B_1, \ldots, B_k$ form a partition of the space S and $P(B_j) > 0$ for $j = 1, \ldots, k$. Then
$$P(A) = \sum_{j=1}^{k} P(B_j \cap A) = \sum_{j=1}^{k} P(B_j)P(A|B_j)$$
.
- Conditional LOTP
$$P(A|C) = \sum_{j=1}^{k} P(B_j|C)P(A|B_j \cap C)$$
.

2.2 Independent Events

Independence of 2 Events:

- Two events are independent if learning that one occurred does not change the probability of the other event.
- Two events A and B with positive probabilities are independent if
$$P(A \cap B) = P(A)P(B)$$
- Similarly, two events A and B with positive probabilities are independent if
$$P(A|B) = P(A) \text{ and } P(B|A) = P(B)$$
- If two events A and B are independent, then A and B^C are independent

Independence of k Events:

- For k events, if knowing what happened with any particular subset of the events gives us no information about what happened with the events not in the subset, the events are independent.
- k events are (mutually) independent if
  - Any pair $P(A_i \cap A_j) = P(A_i)P(A_j)$ for $i \neq j$

- Any triplet $P(A_i \cap A_j \cap A_k) = P(A_i)P(A_j)P(A_k)$ for $i \neq j \neq k$
- ...
- The n-tuplet $P(A_1 \cap \ldots \cap A_n) = P(A_1) \ldots P(A_n)$

When deciding whether or not to model events as independent, try to answer the following question: "If I were to learn that some of these events occurred, would I change the probabilities of any of the others?"

Conditional Independence:

- Two events A and B are conditionally independent given E if $P(A \cap B|E) = P(A|E)P(B|E)$.
- It is defined as independence but with respect to the conditional probabilities.

2.3 Bayes' Theorem

- Bayes' Theorem relates $P(A|B)$ to $P(B|A)$.
- This is important as often it is easier to solve either $P(A|B)$ or $P(B|A)$.
- Suppose that we are interested in which of several disjoint events $A_1, \ldots, A_k$ will occur and that we will get to observe some other event B. If $P(B|A_i)$ is available for each i, then Bayes' theorem is a useful formula for computing the conditional probabilities of the A_i events given B, that is, $P(A_i|B)$ for each i.

Bayes' Theorem for 2 Events

- $$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$
- It is often common to use LOTP in the denominator

$$P(A|B) = \frac{P(B|A)P(A)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$$

Bayes' Theorem for n Events

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{n} P(B|A_j)P(A_j)}$$

Prior vs. Posterior Probabilities

- Prior probabilities represent probabilities *before* new evidence is introduced.
  - $P(A)$
- Posteror probabilities represent probabilities *after* the new evidence in taken into consideration.
  - $P(A|B)$

Conditional Version of Bayes' Theorem

$$P(A_i|B \cap C) = \frac{P(B|A_i \cap C)P(A_i|C)}{\sum_{j=1}^{n} P(B|A_j \cap C)P(A_j|C)}$$

2.4 The Gambler's Ruin Problem

High Level Problem Overview: Consider two gamblers with finite resources who repeatedly play the same game against each other. Using the tools of conditional probability, we can calculate the probability that each of the gamblers will eventually lose all of his money to the opponent. As this game ends only when one of the gamblers has no money left, this problem is fittingly called, the Gambler's Ruin problem.

Problem Statement

- Two gamblers, A and B, make \$1 bets each round.
- A has a probability p of winning, B has probability $1 - p$ of winning.
- A starts with k dollars, B starts with $k - i$ dollars, for a combined fortune of k.
- The game continues until one of the gamblers has 0 dollars left.
- The goal is to determine the probability that A will reach k dollars before 0.

3 Random Variables and Distributions

3.1 Random Variables and Discrete Distributions

Random Variables

- A function that maps each outcome in the sample space to the real line.
- Two main types of random variables: discrete and continuous random variables.
- Main tool used for modeling unknown quantities in statistical analysis.

Distributions

- A distribution specifies the probabilities of all events associated with a random variable (maps outcomes to probabilities).

Discrete Distributions

- Distributions that assign positive probability to at most countably many different values.
- Discrete distributions can be characterized by its probability function (p.f.) aka probability mass function (p.m.f).

Discrete Random Variables

- A random variable with a discrete distribution.
- Is a discrete r.v. if the r.v. can take only a finite number k of different values $x_1, \dots x_k$ or, at most, an infinite sequence of different values $x_1, x_2, \dots$

Probability Function

- A function that describes the probabilities of different values of a r.v.

- For a discrete r.v., the p.f. of X is defined as the function f such that for every real number x,
$$f(x) = P(X = x)$$
- If x is not a possible value of X, then $f(x) = P(X = x) = 0$
- The support of X is all values x such that $f(x) = P(X = x) > 0$.
- To be a valid p.f.
  - The probabilities of all possible outcomes of X sum to 1. More formally, if the sequence $x_1, x_2, \ldots$ includes all the possible values of X, then $\sum_{i=1}^{\infty} f(x_i) = 1$.
  - All possible values of X have a probability greater than 0 and all other values have a probability of 0.
- A p.f. is only for discrete distributions.

Example

- In an experiment in which a fair coin is tossed 2 times, we can define a random variable X to represent the number of heads in each outcome.
- The possible outcomes in the sample space are $S = \{TT, HT, TH, HH\}$.
- $X(TT) = 0,\ X(TH) = 1,\ X(HT) = 1,\ X(HH) = 2.$
- Therefore, the p.f. of X is the function f given by
  - $f(0) = P(X = 0) = 1/4$
  - $f(1) = P(X = 1) = 1/2$
  - $f(2) = P(X = 2) = 1/4$

Common Named Distributions

- Some random variables has distributions that occur so frequently they are given names.
- These common named distributions represent a family of distributions (for example, there are a family of Bernoulli distributions, not just one). What allows us to determine the specific distribution we have (the specific Bernoulli distribution we have), are parameters (for Bernoulli, this is the value for p)

Bernoulli Distributions

- A distribution of the probabilities of a single experiment asking a yes-no question.
- A r.v. X that takes on only two values 1 and 0 (success and failure).
- Has one parameter, p.
- Written $X \sim \text{Bern}(p)$.
- p.f. of X
$$P(X = x) = f(x) = \begin{cases} p & \text{for } x = 1 \\ 1 - p & \text{for } x = 0 \\ 0 & \text{otherwise} \end{cases}$$

Discrete Uniform Distributions

- A distribution where each outcome for the r.v. has an equally likely probability.
- A r.v. X where, for 2 integers a, b where $a \leq b$, the value of X is equally likely to be each of the integers $a, \ldots, b$.
- A uniform distribution on a set of k integers has probability of $1/k$ on each integer.
- Has two parameters, a, b.
- Written $X \sim \text{DUnif}(a, b)$.
- p.f. of X

$$P(X = x) = f(x) = \begin{cases} \frac{1}{b-a+1} & \text{for } x = a, \ldots, b \\ 0 & \text{otherwise} \end{cases}$$

Binomial Distributions

- A distribution that is based upon a r.v. denoting the number of successes in n independent Bernoulli trials (an experiment that can end only in success or failure).
- Suppose that n independent Bernoulli trials are performed, each with success probability p. If we let X be the number of success, then the distribution of X is called the Binomial distribution with parameters n and p.
- Has two parameters, n, p.
- Written $X \sim \text{Bin}(n, p)$.
- p.f. of X

$$P(X = x) = f(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{for } x = 0, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$

  - The probability of obtaining exactly x successes and $n-x$ failures is $p^x(1-p)^{n-x}$. Since there are $\binom{n}{x}$ ways of ordering this, its follows that $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$.
- The Bernoulli distribution is just a special case of the Binomial distribution where $n = 1$, that is, $\text{Bin}(1, p)$.

3.2 Continuous Distributions

Continuous Distributions

- Distributions that have a range of values that are uncountably infinite.
- Continuous distributions can be characterized by its probability density function (p.d.f.)
- Assign probability 0 to individual values, that is, $P(X = x) = 0$ for every number x.

Continuous Random Variables

- A random variable with a continuous distribution.

Probability Density Function

- A nonnegative function f for a r.v. X such that the integral of f over each interval $[a, b]$ gives the probability that X is in the interval, that is, $P(a \leq X \leq b) = \int_a^b f(x)\,dx$.
- Similarly, $P(X \geq a) = \int_a^\infty f(x)\,dx$ and $P(X \leq b) = \int_{-\infty}^b f(x)\,dx$.
- Density is not probability (density can be greater than 1). The p.d.f. of X, $f(x)$, itself is not the probability that X is near x, instead, the integral of f over values near x gives the probability that X is near x, and the integral is never greater than 1.
- Since the endpoints have probability 0, for an interval (a, b), $P(a < X < b) = P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b)$
- The support of X is all values x such that $f(x) = P(X = x) > 0$.
- To be a valid p.d.f.
  - The p.d.f. of X must integrate to 1, that is, $\int_{-\infty}^\infty f(x)\,dx = 1$.
  - For all x, $f(x) \geq 0$.
- To get a desired probability, integrate the p.d.f. over the appropriate range.
- In general, we can think of $f(x)\,dx$ as the probability of X being in an infinitesimally small interval containing x, of length dx,

Uniform Distributions

- A Uniform r.v. on the interval (a, b) is a completely random number, which speicifies that the p.d.f. is constant over the interval.
- Has two parameters, a, b.
- Written $X \sim \text{Unif}(a, b)$.
- p.d.f. of X

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

Normalizing Constants

- Often, we will have a p.d.f without a specified value for c, the normalizing constant.
- This happens especially often when we find sampling distributions of summaries of observed data where we can determine the p.d.f of a r.v. except for the constant factor.
- The constant is unique.
- Using the fact that the p.d.f must integrate to 1, we can solve for c.
- Example:

$$\int_0^4 cx\,dx = 1$$

$$c\int_0^4 x\,dx = 1$$

$$c\frac{x^2}{2}|_0^4 = 1$$
$$c\left(\frac{4^2}{2} - 0\right) = 1$$
$$c\left(\frac{16}{2}\right) = 1$$
$$8c = 1$$
$$c = \frac{1}{8}$$

3.3 The Cumulative Distribution Function

Cumulative Distribution Function

- A function of a r.v. X, denoted F, that gives the probability that the r.v. is less than or equal to x
$$F(x) = P(X \leq x) \text{ for } -\infty < x < \infty$$
- While a p.f is defined only for discrete r.v.s and a p.d.f is defined for only continuous r.v.s, a c.d.f. is defined for both discrete and continuous r.v.'s.
- The c.d.f. is another way of characterizing the distribution of a r.v.
- The value of F at every point x must be a number $F(x)$ in the interval $[0, 1]$ because $F(x)$ is the probability of the event $X \leq x$.
- For every value x
$$P(X > x) = 1 - F(x)$$
- For all values x_1 and x_2 such that $x_1 < x_2$
$$P(x_1 < X \leq x_2) = F(x_2) - F(x_1)$$
- For each value x
$$P(X < x) = F(x^-)$$
- For every value x
$$P(X = x) = F(x) - F(x^-)$$
  - For every value x, $P(X = x)$ is equal to the amount of the jump that occurs in F at the point x. If F is continuous at that point x, that is, if there is no jump in F at x ($F(x^-) = F(x^+) = F(x)$), then $P(X = x) = 0$.

CDF Properties

- Nondecreasing: If $x_1 \leq x_2$, then $F(x_1) \leq F(x_2)$.
- Convergence to 0 and 1 in the limits
$$\lim_{x \to -\infty} F(x) = 0$$
$$\lim_{x \to \infty} F(x) = 1$$

- Right continuous: A c.d.f. is continuous except for the possibility of having some jumps. If there are jumps, the c.d.f. is continuous from the right

$$F(x) = F(x^+)$$

at every point x.

CDF of a Discrete Distribution

- $F(x_i)$ will have a jump of magnitude $f(x_i) = P(X = x_i)$ at each possible value x_i of X.
- $F(x)$ will be constant and horizontal between every pair of successive jumps. If the interval $[a, b]$ represents this constant area between successive jumps, then $P(a < X < b) = 0$.

CDF of a Continuous Distribution

- If X is a continuous r.v., $f(x)$ is its p.d.f, and $F(x)$ is its c.d.f., then F is continuous at every x,

$$F(x) = \int_{-\infty}^{x} f(t)\, dt$$

and

$$\frac{dF(x)}{dx} = f(x)$$

at all x such that f is continuous.

- The c.d.f. of a continuous r.v. X can be obtained from the p.d.f and vice versa.
  - The p.d.f is the derivative of the c.d.f.

$$f(x) = F'(x)$$

  - The c.d.f. is an antiderivative (a function whose derivative is the original function) of the p.d.f.

Quantile Function

- A percentile is a way of expressing where an observation falls in a range of other observations. It is the value below which a percentage of data falls (divides the data into hundreths).
- If we are looking for the $100p$ (p is strictly between 0 and 1) percentile. of a r.v. X with a c.d.f. F, the inverse of F, $F^{-1}(p)$ allows us to do so.
- The p quantile of X, $F^{-1}(p)$, is the smallest value x such that $F(x) \geq p$.
- The quantile function, F^{-1}, is defined on the open interval $(0, 1)$.
- Exists for discrete and continuous distributions.
- Certain quantiles have certain names
  - Lower quartile: 25th percentile
  - Median: 50th percentile
  - Upper quartile: 75th percentile
- The quantile function is an alternative way to characterize a distribution.
- To find the quantile function $F^{-1}(p)$ when we know the c.d.f., we can set $F(x) = p$ and solve for x.

3.4 Bivariate Distributions

Joint/Bivariate Distribution Basics

- Often, in real life we are interested in several r.v.s that are related to each other. For example, if studying a random family, we may want to study the number of people in the family (which would be one r.v.) and the household income (which would be another r.v.). Joint distributions allow us to assign probabilities to these events.
- A joint distribution of two r.v.s X and Y is the collection of all probabilities of the form $P[(X,Y) \in C]$ for all sets C of pairs of real numbers such that $\{(X,Y) \in C\}$ is an event.

Discrete Joint Distribution

- Let X and Y be random variables, and consider the ordered pair (X,Y). If there are only finitely or at most countably many different possible values (x,y) for the pair (X,Y), then we say that X and Y have a discrete joint distribution.

Joint Probability Function

- The joint p.f. of discrete r.v.s X and Y is the function f, such that for every point (x,y) in the xy-plane
$$f(x,y) = P(X = x, Y = y)$$
where $f(x,y)$ is the probability of x and y occurring.

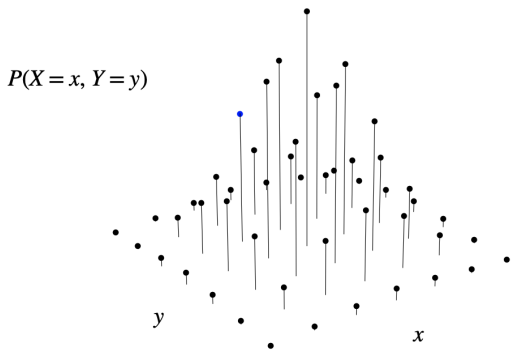

Figure 1: Joint p.f. of discrete r.v.s X and Y. The height of a vertical bar at (x,y) represents the probability $P(X = x, Y = y)$. The total summed heights of the vertical bars must be 1.

- To be a valid joint p.f.
  - The probabilities of all possible outcomes of (X,Y) sum to 1. That is
$$\sum_{\text{All } (x,y)} f(x,y) = 1$$
  - All possible values of (X,Y) have a probability greater than 0 and all other values have a probability of 0. That is, if (x,y) is not one of the possible values of (X,Y), then $f(x,y) = 0$.

- We can find the probability of the event $(X, Y) \in C$ for any set C of points (ordered pairs) in the support of (X, Y) by summing over the joint p.f. over C

$$P[(X, Y) \in C] = \sum_{(x,y)\in C} f(x, y)$$

Continuous Joint Distribution

- Two r.v.s X and Y have a continuous joint distribution if there exists a nonnegative function f (joint p.d.f.) defined over the entire xy-plane such that for every subset C of the plane,

$$P[(X, Y) \in C] = \int_C \int f(x, y)\, dxdy$$

if the integral exists.

Joint Probability Density Function

- If X and Y are continuous r.v.s with joint c.d.f $F(x, y)$, their joint p.d.f. is the derivative of the joint c.d.f. with respect to x and y (take the derivative of y treating x as a constant, then take the derivative of x treating y as a constant)

$$f(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y)$$

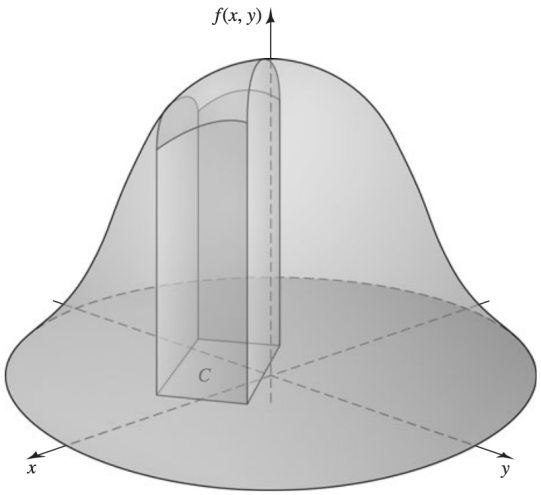

Figure 2: Joint p.d.f. of continuous r.v.s X and Y. When integrating over the region C, we are calculating the volume under the surface of the joint p.d.f and above C. The total volume under a valid joint p.d.f. is 1.

- To be a valid joint p.d.f.
  - The joint p.d.f. of X and Y must integrate to 1, that is, $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dxdy = 1$.
  - For all x, $f(x) \geq 0$.

Mixed Bivariate Distributions

- Sometimes, we must consider a mixed bivariate distribution in which we have one discrete r.v. and one continuous r.v.

Joint Probability Function/Probability Density Function

- Let X and Y be r.v.s such that X is discrete and Y is continuous. Suppose that there is a function $f(x, y)$ (join p.f./p.d.f. of X and Y) defined on the xy- plane such that, for every pair A and B of subsets of the real numbers,

$$P(X \in A, Y \in B) = \int_B \sum_{x \in A} f(x, y)\, dy$$

  if the integral exists.

Joint Cumulative Distribution Function

- The joint c.d.f. of two continuous r.v.s X and Y is define as the function F such that for all values of x and y $(-\infty < x < \infty, -\infty < y < \infty)$

$$F(x, y) = P(X \leq x, Y \leq y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f(r, s)\, dr ds$$

- The probability that an event occurs such that $X \leq x$ and $Y \leq y$.

3.5 Marginal Distributions

Marginal Distributions

- A marginal distribution is simply distribution of any 1 r.v.
- Often, we may start with a joint distribution of 2 r.v.s and want to find the distribution of just 1 of them.
- The distribution of 1 r.v. X computed from a joint distribution is called the marginal distribution of X.
- Each r.v. will have a marginal c.d.f. as well as a marginal p.d.f or p.f.
- The name marginal distribution derives from the fact that the marginal distribution are the totals that appear in the margins of tables.

Marginal Probability Function

- For discrete r.v.s X and Y that have a joint p.f. f, the marginal p.f., f_1, of X is

$$f_1(x) = P(X = x) = \sum_{\text{All } y} P(X = x, Y = y) = \sum_{\text{All } y} f(x, y)$$

  which says, to find the marginal p.f. of X, hold the 1 value of x fixed and then sum over all the outcomes of Y.

- Similarly, if f_2 is the marginal p.f. of Y

$$f_2(y) = P(Y = y) = \sum_{\text{All } x} P(X = x, Y = y) = \sum_{\text{All } x} f(x, y)$$

Marginal Probability Density Function

- For continuous r.v.s X and Y that have a joint p.d.f. f, the marginal p.d.f., f_1, of X is

$$f_1(x) = \int_{-\infty}^{\infty} f(x, y)\, dy \quad \text{for } -\infty < x < \infty$$

  which says, to find the marginal p.d.f. of X, hold the 1 value of x fixed and then integrate over all the outcomes of Y.

- Similarly, if f_2 is the marginal p.d.f. of Y

$$f_2(y) = \int_{-\infty}^{\infty} f(x, y)\, dx \quad \text{for } -\infty < y < \infty$$

Marginal p.f./p.d.f from a Discrete and Continuous Random Variable

- For discrete r.v. X and continuous r.v. Y that have a joint p.f./p.d.f f, the marginal p.f., f_1, of X is

$$f_1(x) = P(X = x) = \int_{-\infty}^{\infty} f(x, y)\, dy \quad \text{for all } x$$

- For discrete r.v. X and continuous r.v. Y that have a joint p.f./p.d.f f, the marginal p.d.f., f_2, of Y is

$$f_2(y) = \sum_x f(x, y) \quad \text{for } -\infty < x < \infty$$

Independence of Random Variables

- If we are trying to decide whether or not to model 2 r.v.s as independent, we should think about whether we would change the distribution of X after we learned the value of Y or vice versa.

- In general, the marginal distributions do not determine the joint distribution, this is the reason we wanted to study joint distributions in the first place. However, in the special case of independence, the marginal distributions are all we need in order to specify the joint distribution: we can get the joint p.f. by multiplying the marginal p.f.s.

Independence of Discrete Random Variables

- 2 discrete r.v.s X (with marginal c.d.f. F_1) and Y (with marginal c.d.f. F_2) with joint c.d.f F are independent if for all x and y

$$F(x, y) = F_1(x)F_2(y)$$

  which says that 2 discrete r.v.s are independent when the joint c.d.f. is the product of the marginal c.d.f.s of X and Y.

- If we let f be the joint p.f. of X and Y, f_1 be the marginal p.f. of X, and f_2 be the marginal p.f. of Y, this is equivalent to saying

$$f(x, y) = f_1(x)f_2(y)$$

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

  which says that 2 discrete r.v.s are independent when the joint p.f. is the product of the marginal p.f.s of X and Y.

Independence of Continuous Random Variables

- 2 continuous r.v.s X (with marginal c.d.f. F_1) and Y (with marginal c.d.f. F_2) with joint c.d.f F are independent if for all x and y

$$F(x, y) = F_1(x)F_2(y)$$

  which says that 2 continuous r.v.s are independent when the joint c.d.f. is the product of the marginal c.d.f.s of X and Y.

- If we let f be the joint p.d.f. of X and Y, f_1 be the marginal p.d.f. of X, and f_2 be the marginal p.d.f. of Y, this is equivalent to saying

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

$$f(x, y) = f_1(x)f_2(y)$$

  which says that 2 continuous r.v.s are independent when the joint p.d.f. is the product of the marginal p.d.f.s of X and Y.

3.6 Conditional Distributions

Conditional Distributions

- Recall that distributions are just collections of probabilities of events determined by r.v.s. Conditional distributions will be the probabilities of events determined by some r.v.s conditional on events determined by other r.v.s.

- The idea is that typically, there will be many r.v.s of interest in a problem. After we observe some of those r.v.s, we want to be able to adjust the probabilities associated with the r.v.s that have not yet been observed.

- The conditional distribution of 1 r.v. X given another r.v. Y will be the distribution that we would use for X after we learn the value of Y.

Conditional Probability Function

- For discrete r.v.s X and Y, the conditional p.f. of Y given $X = x$ is

$$P(Y = y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)}$$

  which is viewed as a function of y for a fixed x.

- If we let the joint p.f. be f and the marginal p.f. of X be f_1, then this is equivalent to

$$g_2(y|x) = \frac{f(x, y)}{f_1(x)}$$

  where g_2 is the conditional p.f. of Y given X.

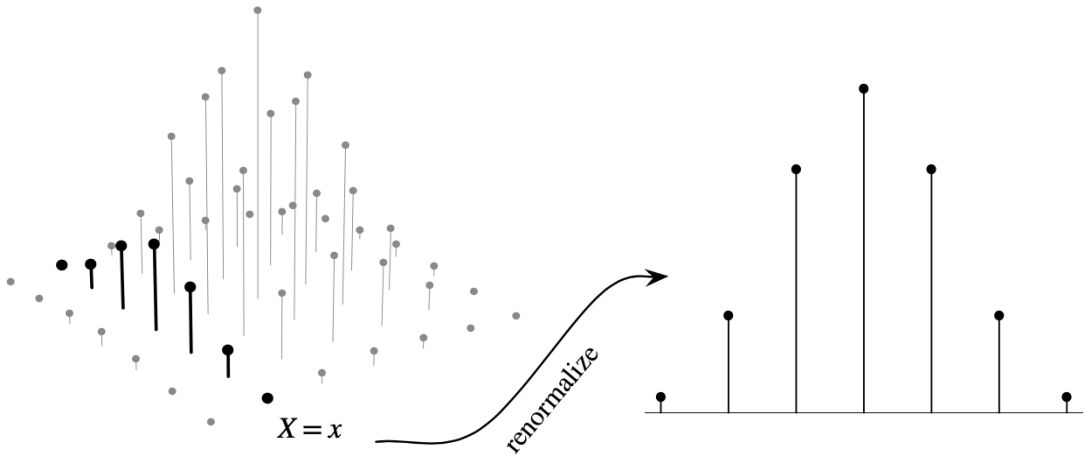

Figure 3: Conditional p.f. of Y given $X = x$. This is obtained by starting with the joint p.f (left) column compatible with $X = x$ and then renormalizing by dividing by $P(X = x)$ so the conditional p.f. sums to 1.

Conditional Probability Density Function

- For continuous r.v.s X and Y with joint p.d.f. f and X with marginal p.d.f. f_1, the conditional p.d.f. of Y given $X = x$ is

$$g_2(y|x) = \frac{f(x, y)}{f_1(x)}$$

where g_2 is the conditional p.d.f. of Y given $X = x$.

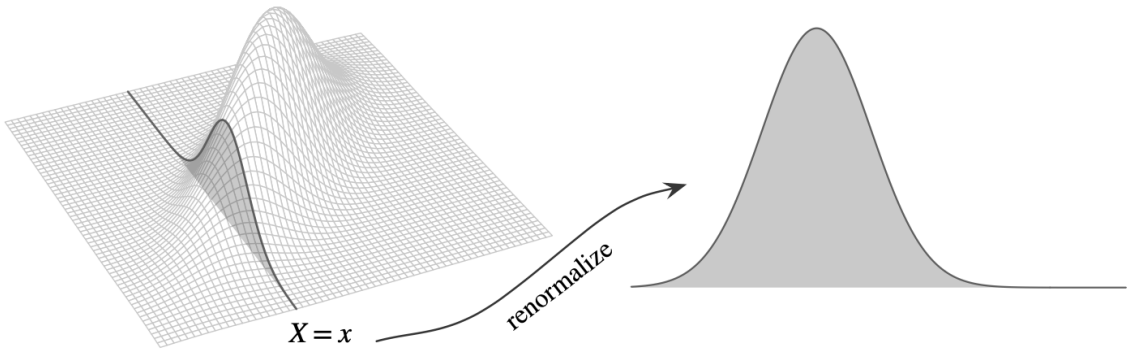

Figure 4: Conditional p.d.f. of Y given $X = x$. This is obtained by starting with the joint p.d.f (left) slice compatible with $X = x$ and then renormalizing by dividing by $P(X = x)$ so the conditional p.d.f. integrates to 1.

- It is important to keep in mind that the conditional p.d.f. of X given $Y = y$ is better thought of as the conditional p.d.f. of X given that Y is very close to y. This motivates us to remember the distinction between the conditional p.d.f. and conditioning on an event with probability 0.

Law of Total Probability with Conditional Probabilities

- If $f_2(y)$ is the marginal p.f. or p.d.f. of a r.v. Y and $g_1(x|y)$ is the conditional p.f. or p.d.f. of X given $Y = y$, then the marginal p.f. or p.d.f. of X is

$$f_1(x) = \sum_y g_1(x|y) f_2(y)$$

  if Y is discrete. If Y is continuous, the marginal p.f. or p.d.f of X is

$$f_1(x) = \int_{-\infty}^{\infty} g_1(x|y) f_2(y)\, dy$$

Bayes' Rule with Conditional Probabilities

- If $f_2(y)$ is the marginal p.f. or p.d.f. of a r.v. Y and $g_1(x|y)$ is the conditional p.f. or p.d.f. of X given $Y = y$, then the conditional p.f. or p.d.f. or Y given $X = x$ is

$$g_2(y|x) = \frac{g_1(x|y) f_2(y)}{f_1(x)}$$

  where $f_1(x)$ is obtained the above appropriate LOTP formula.

- Similarly, the conditional p.f. or p.d.f. of X given $Y = y$ is

$$g_1(x|y) = \frac{g_2(y|x) f_1(x)}{f_2(y)}$$

  where $f_2(y)$ is obtained the above appropriate LOTP formula with x and y switched and with the subscripts 1 and 2 switched.

Independence

- Suppose that X and Y are 2 r.v.s having a joint p.f., p.d.f, or p.f./p.d.f. f. Then X and Y are independent if and only if for every value of y such that $f_2(y) > 0$ and every value of x

$$g_1(x|y) = f_1(x)$$

3.7 Multivariate Distributions

Multivariate Distributions

- Multivariate distributions are simply an extension of the 2 r.v. (bivariate) distributions to n r.v. (multivariate) distributions.
- In general, the joint distribution of more than 2 r.v.s is called a multivariate distribution.
- The theory of statistical inference relies on mathematical models for observable data in which each observation is a r.v. Therefore, multivariate distributions arise naturally in the mathematical models for data.

Vector Notation

- When dealing with n r.v.s, $X_1, \ldots, X_n$, it is often more convenient to use the vector notation $\boldsymbol{X} = (X_1, \ldots, X_n)$.

- $\boldsymbol{X}$ is referred to as a *random vector*.
- Instead of speaking of the joint distrubution of the r.v.s $X_1, \ldots, X_n$ with a joint c.d.f. $F(x_1, \ldots, x_n)$, we can simply speak of the distribution of the random vector $\boldsymbol{X}$ with c.d.f. $F(\boldsymbol{x})$.
- When using vector notation, we must keep in mind that $\boldsymbol{X}$ is an n- dimensional random vector, with its c.d.f. being defined as a function on the n- dimensional space $\mathbb{R}^n$.

Multivariate Discrete Distributions

- n r.v.s $X_1, \ldots, X_n$ have a discrete joint distribution if the random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ can have only a finite number or an infinite sequence of different possible values $(x_1, \ldots, x_n)$ in $\mathbb{R}^n$.
- The joint p.f. of $X_1, \ldots, X_n$ is defined as the function f such that for every point $(x_1, \ldots, x_n) \in \mathbb{R}^n$

$$f(x_1, \ldots, x_n) = P(X_1 = x_1, \ldots, X_n = x_n)$$

$$f(\boldsymbol{x}) = P(\boldsymbol{X} = \boldsymbol{x})$$

- If $\boldsymbol{X}$ has a joint discrete distribution with p.f. f, then for every subset $C \subset \mathbb{R}^n$

$$P(\boldsymbol{X} \in C) = \sum_{x \in C} f(\boldsymbol{x})$$

Multivariate Continuous Distributions

- n r.v.s $X_1, \ldots, X_n$ have a continuous joint distribution if there is a nonnegative function (joint p.d.f. of $\boldsymbol{X}$) f defined on $\mathbb{R}^n$ such that for every subset $C \subset \mathbb{R}^n$

$$P[(X_1, \ldots, X_n) \in C] = \int \cdots \int_C f(x_1, \ldots, x_n)\, dx_1, \ldots, \, dx_n$$

$$P(\boldsymbol{X} \in C) = \int \cdots \int_C f(\boldsymbol{x})\, d\boldsymbol{x}$$

- Even if each of $X_1, \ldots, X_n$, has a continuous distribution, the vector $\boldsymbol{X}$ might not have a continuous joint distribution.

Multivariate Mixed Distributions

- If $X_1, \ldots, X_n$ are a combindation of discrete and continuous r.v.s, their joint distribution would then be represented by a function f, called the *joint p.f./p.d.f.*
- The function has the property that the probability that $\boldsymbol{X}$ lies in a subset $C \subset \mathbb{R}^n$ is calculated by summing $f(\boldsymbol{x})$ over the values of the coordinates of $\boldsymbol{x}$ that correspond to the discrete r.v.s and integrating over the coordinates that correspond to the continuous r.v.s for all points $\boldsymbol{x} \in C$.

Deriving a Marginal p.d.f. from a Multivariate Continuous Distribution

- If the joint distribution of n r.v.s $X_1, \ldots, X_n$ is known, then the marginal distributions of each single r.v. X_i can be derived from this joint distribution.

- The marginal joint p.d.f. of any k of the n random variables $X_1, \ldots, X_n$ can be found by integerating the joint p.d.f. over all possible values of the other $n-k$ variables.
- If the joint p.d.f. of $X_1, \ldots, X_n$ is f, then the marginal p.d.f. f_1 of X_1 is specified at every value x_1 by the relation

$$f_1(x_1) = \underbrace{\int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty}}_{n-1} f(x_1, \ldots, x_n)\, dx_2 \ldots\, dx_n$$

- If the joint p.d.f. of X_1, X_2, X_3, X_4 is f, then the marginal bivariate p.d.f. f_{24} of X_2 and X_4 is specified at each point (x_2, x_4) by the relation

$$f_{24}(x_2, x_4) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2, x_3, x_4)\, dx_1\, dx_3$$

Deriving a Marginal p.f. from a Multivariate Discrete Distribution

- If n r.v.s $X_1, \ldots, X_n$ have a discrete joint distribution, then the marginal joint p.f. of each subset of the n variables can be obtained from the relations similar to those for continuous distributions where the integrals are replaced by sums.

Deriving a Marginal c.d.f.

- The marginal joint c.d.f. fo any k of n r.v.s $X_1, \ldots, X_n$ can be found by computing the limiting value of the n-dimensional c.d.f. F as $x_j \to \infty$ for each of the other $n-k$ variables x_j.
- If the joint c.d.f. of $X_1, \ldots, X_n$ is F, then the marginal c.d.f. F_1 of X_1 can be obtained from the following relation

$$F_1(x_1) = P(X_1 \leq x_1) = P(X_1 \leq x_1, X_2 < \infty, \ldots, X_n < \infty)$$

$$F_1(x_1) = \lim_{x_2, \ldots, x_n \to \infty} F(x_1, x_2, \ldots, x_n)$$

- If F is the joint c.d.f. of 4 r.v.s X_1, X_2, X_3, X_4, then the marginal bivariate c.d.f. F_{24} of X_2 and X_4 is specified at every point (x_2, x_4) by the relation

$$F_{24}(x_2, x_4) = \lim_{x_1, x_3 \to \infty} F(x_1, x_2, x_3, x_4)$$

Independent Random Variables

- n r.v.s $X_1, \ldots, X_n$ are independent if, for every n sets $A_1, \ldots, A_n$ of real numbers

$$P(X_1 \in A_1, \ldots, X_n \in A_n) = P(X_1 \in A_1) \ldots P(X_n \in A_n)$$

- Let F denote the joint c.d.f. of $X_1, \ldots, X_n$ and let F_i denote the marginal univariate c.d.f. of X_i for $i = 1, \ldots, n$. The variables $X_1, \ldots, X_n$ are independent if and only if, for all points $(x_1, \ldots, x_n) \in \mathbb{R}^n$

$$F(x_1, \ldots, x_n) = F_1(x_1) \ldots F_n(x_n)$$

- If $X_1, \ldots, X_n$ have a continuous, discrete, or mixed joint distribution for which the joint p.d.f., joint p.f., or joint p.f./p.d.f. is f, and if f_i is the marginal univariate p.d.f. or p.f. of X_i for $i = 1, \ldots, n$, then $X_1, \ldots X_n$ are independent if and only if the following relation is satisfied at all points $x_1, \ldots, x_n \in \mathbb{R}^n$

$$f(x_1, \ldots, x_n) = f_1(x_1) \ldots f_n(x_n)$$

- Consider a given probability distribution of the real line that can be represented by either a p.f. or a p.d.f. f. It is said that n r.v.s $X_1, \ldots X_n$ form a *random sample* from this distribution if these r.v.s are independent and the marginal p.f. or p.d.f. of each of them is f. Such r.v.s are also said to be *independent and identically distributed (i.i.d.)*. We refer to the number n of r.v.s as the *sample size*.

- n r.v.s $X_1, \ldots, X_n$ each with marginal p.f./p.d.f. f and joint distribution g if and only if, at all points $(x_1, \ldots, x_n) \in \mathbb{R}^n$

$$g(x_1, \ldots, x_n) = f(x_1) \ldots f(x_n)$$

Conditional Multivariate Distributions

  - Suppose that n r.v.s $X_1, \ldots, X_n$ have a continuous distribution for which the joint p.d.f. is f and that f_0 denotes the marginal joint p.d.f. of the $k < n$ r.v.s $X_1, \ldots, X_k$. Then for all values of $x_1, \ldots x_k$ such that $f_0(x_1, \ldots, x_k) > 0$, the conditional p.d.f. of $(X_{k+1}, \ldots X_n)$ given that $X_1 = x_1, \ldots, X_k = x_k$ is defined as

$$g_{k+1 \ldots n}(x_{k+1}, \ldots x_n | x_1, \ldots x_k) = \frac{f(x_1, \ldots x_n)}{f_0(x_1, \ldots, x_k)}$$

  - More generally, suppose that the random vector $\boldsymbol{X} = (X_1, \ldots X_n)$ is divided into 2 subvectors $\boldsymbol{Y}$ and $\boldsymbol{Z}$ where $\boldsymbol{Y}$ is a k-dimensional random vector comprising k of the n r.v.s in $\boldsymbol{X}$, and $\boldsymbol{Z}$ is an $(n-k)$-dimensional random vector comprising of the other $n-k$ random variables in $\boldsymbol{X}$. Suppose also that the n-dimensional joint p.f., p.d.f., or p.f./p.d.f. of $(\boldsymbol{Y}, \boldsymbol{Z})$ is f and that the marginal $(n-k)$-dimensional p.f., p.d.f., or p.f./p.d.f. of $\boldsymbol{Z}$ is f_2. Then for every given point $\boldsymbol{z} \in \mathbb{R}^{n-k}$ such that $f_2(\boldsymbol{z}) > 0$, the conditional k -dimensional p.f., p.d.f., or p.f./p.d.f. g_1 of $\boldsymbol{Y}$ given $\boldsymbol{Z} = \boldsymbol{z}$ is defined as

$$g_1(\boldsymbol{y}|\boldsymbol{z}) = \frac{f(\boldsymbol{y}, \boldsymbol{z})}{f_2(\boldsymbol{z})}$$

    for $\boldsymbol{y} \in \mathbb{R}^k$.

    This can be rewritten as

$$f(\boldsymbol{y}, \boldsymbol{z}) = g_1(\boldsymbol{y}|\boldsymbol{z}) f_2(\boldsymbol{z})$$

  - To determine the conditional version given $\boldsymbol{W} = \boldsymbol{w}$ of a result just proven, simply add conditional on $\boldsymbol{W} = \boldsymbol{w}$ to every probabilistic statement in the result.

Multivariate Law of Total Probability and Bayes' Theorem

– Suppose that the random vector $\boldsymbol{X} = (X_1, \ldots X_n)$ is divided into 2 subvectors $\boldsymbol{Y}$ and $\boldsymbol{Z}$ where $\boldsymbol{Y}$ is a k-dimensional random vector comprising k of the n r.v.s in $\boldsymbol{X}$, and $\boldsymbol{Z}$ is an $(n-k)$-dimensional random vector comprising of the other $n-k$ random variables in $\boldsymbol{X}$. Suppose also that the n-dimensional joint p.f., p.d.f., or p.f./p.d.f. of $(\boldsymbol{Y}, \boldsymbol{Z})$ is f and that the marginal $(n-k)$-dimensional p.f., p.d.f., or p.f./p.d.f. of $\boldsymbol{Z}$ is f_2. If $\boldsymbol{Z}$ has a continuous joint distribution, the marginal p.d.f. of $\boldsymbol{Y}$ is

$$f_1(\boldsymbol{y}) = \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{n-k} g_1(\boldsymbol{y}|\boldsymbol{z}) f_2(\boldsymbol{z})\, d\boldsymbol{z}$$

and the conditional p.d.f. of $\boldsymbol{Z}$ given $\boldsymbol{Y} = \boldsymbol{y}$ is

$$g_2(\boldsymbol{z}|\boldsymbol{y}) = \frac{g_1(\boldsymbol{y}|\boldsymbol{z}) f_2(\boldsymbol{z})}{f_1(\boldsymbol{y})}$$

– If $\boldsymbol{Z}$ has a discrete distribution, then the multiple integreal must be replaced with a multiple summation.

Conditionally Independent Random Variables

– Let $\boldsymbol{Z}$ be a random vector with joint p.f., p.d.f., or p.f./p.d.f. $f_0(\boldsymbol{z})$. Several random variables $X_1, \ldots, X_n$ are *conditionally independent* given $\boldsymbol{Z}$ if, for all $\boldsymbol{z}$ such that $f_0(\boldsymbol{z}) > 0$, we have

$$g(\boldsymbol{x}|\boldsymbol{z}) = \prod_{i=1}^{n} g_i(x_i|\boldsymbol{z})$$