MATH 323: Probability

Anna Brandenberger

December 1, 2018

This is a transcript of the lectures given by Prof. David Stephens[1] during the fall semester of the 2018-2019 academic year (09-12 2018) for the Probability class (MATH 323). **Subjects covered** are: sample space, events, conditional probability, independence of events, Bayes' Theorem; basic combinatorial probability, random variables, discrete and continuous univariate and multivariate distributions; independence of random variables; inequalities, weak law of large numbers, central limit theorem.

[1] david.stephens@mcgill.ca, BH 1225

1 *Basics of Probability*

1.1 *Review of Set Theory*

Probability is a numerical value representing the chance of a particular event occurring given a particular set of circumstances.

Definition 1.1. A **set** S is a collection of elements $s \in S$:

- FINITE: finite number of elements
- COUNTABLE: countably infinite number of elements
- UNCOUNTABLE: uncountably infinite number of elements

Definition 1.2. The **power set** $\mathcal{P}(A)$ of A is the set of all subsets of A: $\mathcal{P}(A) = \{B \mid B \subseteq A\}$.

Set Operations

Let $A \subseteq S$:

$$A \cap \emptyset = \emptyset \qquad A \cup \emptyset = A$$
$$A \cap S = A \qquad A \cup S = S$$

Intersection $\cap$ and union $\cup$ are commutative and associative.

1. INTERSECTION: $s \in A \cap B \iff s \in A$ and $s \in B$
   extends to $A_1, A_2, \ldots, A_K$ (finite): $s \in \bigcap_{k=1}^{K} A_k \iff s \in A_k \ \forall k$
2. UNION: $s \in A \cup B \iff s \in A$ or $s \in B$
   extends to $A_1, A_2, \ldots, A_K$ (finite[2]): $s \in \bigcup_{k=1}^{K} A_k \iff \exists k$ s.t. $s \in A_k$
3. COMPLEMENT: for $A \subseteq S$, $s \in A' \iff s \in S$ but $s \notin A$
4. SET DIFFERENCE $A - B \equiv A \setminus B \equiv A \cap B'$
5. EXCLUSIVE OR (XOR): $A \oplus B \equiv A \cup B - A \cap B$

[2] technically also extends to a countably infinite number of sets.

Definition 1.3. $A_1, A_2, \ldots, A_K$ is called a **partition** of S if these sets are pairwise disjoint $A_j \cap A_k = \emptyset \ \forall j \neq k$ and exhaustive $\bigcup_{k=1}^{K} A_k = S$.

Theorem 1.1 (De Morgan's Laws)**.**

$$A' \cap B' = (A \cup B)' \qquad A \ \cup B' = (A \cap B)'$$

Proof. Let $A_1 = A \cup B$, $A_2 = A' \cap B'$. To show that $A_1' = A_2$, one must show $A_1 \cup A_2 = \emptyset$ and $A_1 \cup A_2 = S$.

$$\begin{aligned} A_1 \cap A_2 &= (A \cup B) \cap A_2 = (A \cap A_2) \cup (B \cap A_2) \\ &= (A \cap A' \cap B') \cup (B \cap A' \cap B') \\ &= \emptyset \cup \emptyset = \emptyset \end{aligned}$$

$$\begin{aligned} A_1 \cup A_2 &= A_1 \cup (A' \cap B') = (A_1 \cup A') \cap (A_1 \cup B') \\ &= (A \cup B \cup A') \cap (A \cup B \cup B') \\ &= S \cap S = S \end{aligned}$$

□

1.2 Sample Spaces and Events

Definition 1.4. The set S of possible outcomes of an experiment is the **sample space** of an experiment, and the individual elements in S are **sample points** or **outcomes**.

Definition 1.5. An **event** A is a collection of sample outcomes $A \subseteq S$. Individual sample outcomes $E_1, E_2, \ldots, E_K, \cdots \in A$ are termed **simple (elementary)** events. We say that A **occurs** if the outcome $s \in A$. S is the **certain** event; $\emptyset$ the **impossible** event.

Definition 1.6. The **probability function** $P(.)$ assigns numerical values to events: $P : \mathcal{A} \longrightarrow \mathbb{R}$, $A \longrightarrow p$ (i.e. $P(A) = p$, $A = \mathcal{P}(\mathcal{S})$).

1.3 Axioms and Consequences

Theorem 1.2. Basic probability axioms:

1. $P(A) \geq 0$
2. $P(S) = 1$
3. P is countably additive: if $A_1, A_2, \ldots$ s.t. $A_j \cap A_k = \emptyset \; \forall j \neq k$:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

Corollary 1.3. $\forall A \; P(A') = 1 - P(A)$

Proof. $S = A \cup A'$. By Theorem 1.2(3),
$P(S) = P(A) + P(A') = 1$
$\implies P'(A) = 1 - P(A)$ □

Corollary 1.4. $P(\emptyset) = 0$

Proof. Simple from Corollary 1.3. □

Corollary 1.5. $\forall A \; P(A) \leq 1$

Proof. By Theorem 1.2 and 1.3,
$1 = P(S) = P(A) + P(A') \geq P(A)$ □

Corollary 1.6. $\forall A \subseteq B$, $P(A) \leq P(B)$

Proof. Take $B = A \cup (A' \cap B)$. Then,
$P(B) = P(A) + P(A' \cap B) \geq P(A)$ □

Corollary 1.7 (General Addition Rule)**.**

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

Proof. First, $(A \cup B) = A \cup (A' \cap B)$
$\implies P(A \cup B) = P(A) + P(A' \cap B)$ [1]
Then, take $B = (A \cap B) \cup (A' \cap B)$
$\implies P(B) = P(A \cap B) + P(A' \cap B)$ [2]
So taking [1] - [2], we get:
$P(A \cup B) - P(B) = P(A) - P(A \cap B)$
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$ □

Using inductive arguments, one can construct a formula for $P\left(\bigcup_{i=1}^{n} A_i\right)$ (c.f. MATH 240) but we will mostly use:

Theorem 1.8 (Boole's Inequality)**.**

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i)$$

As a method, one can lay out probabilities in table form such as in Table 1, ensuring that $0 \leq p_\star \leq 1$ and probabilities along rows and column add up correctly.

	A	A'	Total
B	$p_{A\cap B}$	$p_{A'\cap B}$	p_B
B'	$p_{A\cap B'}$	$p_{A'\cap B'}$	$p_{B'}$
Total	p_A	$p_{A'}$	p_S

Table 1: Probabilities laid out in a table.

1.4 *Specifying probabilities*

Examples of experiments with equally likely sample outcomes include fair coins, dice and cards.

Definition 1.7 (Equally likely sample outcomes)**.** Suppose S finite, with sample outcomes $E_1, \ldots, E_N$, then $P(E_i) = \frac{1}{N}\ \forall i = 1, \ldots, N$. For an event $A \in S$, $\exists n \leq N$ s.t. $A = \bigcup_{i=1}^{n} E_{i_A}$, $i_A \in \{1, \ldots, n\}$

$$\implies P(A) = \frac{n}{N} = \frac{\#\text{ sample outcomes in } A}{\#\text{ sample outcomes in } S}$$

Definition 1.8 (Relative frequencies)**.** For a given experiment with a sample space S and interest event A, consider a finite sequence of N repeat experiments and n the number of times that A occurs. The **frequencist definition** of probability, generalizes the previous definition to $P(A) = \lim_{N\to\infty} \frac{n}{N}$, the relative frequency of appearance of A.

Definition 1.9 (Subjective assessment)**.** For a given experiment with sample space S, the probability of event A is further generalized to a numerical representation of one's own **personal degree of belief** that the actual outcome lies in A, assuming one is **internally consistent**, i.e. rational and coherent.

1.5 *Combinatorial Rules*

c.f. MATH 240 notes for examples.

1. MULTIPLICATION PRINCIPLE: a sequence of k operations i, each with n_i possible outcomes, results in $n_1 \times n_2 \times \cdots \times n_k$ possible sequences of outcomes.
2. SELECTION PRINCIPLES: selecting from a finite set can be done:
   - WITH REPLACEMENT: each selection is independent;
   - WITHOUT REPLACEMENT: set is depleted by each selection.
3. ORDERING: the order of a sequence of selections can be:
   - IMPORTANT (312 distinct from 123);
   - UNIMPORTANT (312 identical to 123).
4. DISTINGUISHABLE ITEMS: objects being selected can be:
   - DISTINGUISHABLE: individually labelled (e.g. lottery balls);
   - INDISTINGUISHABLE: labelled according to a type (e.g. colors).

Definition 1.10. A **permutation** is an ordered arrangement of distinct objects. The number of ways of ordering an n-set taking r at a time is

$$P_r^n = \frac{n!}{(n-r)!}$$

Definition 1.11. The number of ways of partitioning n distinct elements into k disjoint subsets of sizes $n_1, \dots, n_k$ is the **multinomial coefficient**

$$N = \binom{n}{n_1, \cdots, n_k} := \frac{n!}{n_1! \times n_2! \times \cdots \times n_k!}$$

a generalization of the **binomial coeff.** $\binom{n}{j} = \frac{n!}{j!(n-j)!} := \binom{n}{j, n-j}$.

Definition 1.12. The number of **combinations** C_r^n is the number of subsets of size r that can be selected from n objects

$$C_r^n = \binom{n}{r} = n! \cdot P_r^n$$

1.6 *Conditional Probability and Independence*

Definition 1.13. The **conditional probability** of A given B, where $P(B) > 0$ is

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

Given partial knowledge (B happened) i.e. knowing $s \in B \subseteq S$, one can make a more accurate probability assessment of some other event A, as $s \in A \cap B$.

Example 1.1. Given three two-sided cards, one RR, one RB and one BB, one card is selected randomly and one side displayed: it is red (R). Find the probability that the other side is red.

Answer. $S = \{R_1, R_2, B, R, B_1, B_2\}$
Card 1 (RR) selected $A = \{R_1, R_2\} \implies P(A) = 2/6$
Red side exposed: $B = \{R_1, R_2, R\} \implies P(B) = 3/6$
$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{2/6}{3/6} = 2/3$

Corollary 1.9. $P(A \mid S) = P(A) \quad P(A \mid A) = 1 \quad P(A \mid B) \leq 1$

Warning 1. Important distinction between $P(A \cap B)$ and $P(A \mid B)$:

- $P(A \cap B)$: chance of A and B occurring RELATIVE TO S
- $P(A \mid B)$: chance of A and B occurring RELATIVE TO B — non-symmetric: in general, $P(A \mid B) \neq P(B \mid A)$.

Theorem 1.10. The conditional probability function satisfies the probability axioms (Theorem 1.2).

Proof. 1. $P(A \mid B) = \frac{P(A \cap B)}{P(B)} \geq 0$

2. $P(S \mid B) = \frac{P(S \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1$

3. $P\left(\bigcup_{i=1}^{\infty}(A_i \mid B)\right) = \frac{P\left(\bigcup_{i=1}^{\infty} A_i \cap B\right)}{P(B)} = \frac{\sum_{i=1}^{\infty} P(A_i \cap B)}{P(B)} = \sum_{i=1}^{\infty} P(A_i \mid B)$ and since A_i disjoint $\forall i$, (A_i, B) also disjoint. □

Definition 1.14. Two events $A, B \in S$ are **independent** if

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

Remark 2. Notice then that $P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A)P(B)}{P(A)} = P(B)$.

Warning 3. Not the same as mutual exclusivity where $P(A \cap B) = 0$!

Warning 4. For multiple events $A_1, \ldots, A_n \in S$, we can talk about pairwise independence between any A_i, A_j with $i \neq j$, but this does not automatically generalize (c.f. Example 1.2).

Definition 1.15. Events $A_1, \ldots, A_K$ are **mutually independent** if, for all subsets $\mathcal{I}$ of $\{1, 2, \ldots, K\}$:

$$P\left(\bigcap_{k\in\mathcal{I}} A_k\right) = \prod_{k\in\mathcal{I}} P(A_k)$$

Example 1.2. Rolling two dice with independent outcomes A_1: first roll outcome is odd; A_2: second roll outcome is odd and A_3: total outcome is odd. Then $P(A_1) = P(A_2) = \frac{1}{2}$ and $P(A_1 \mid A_3) = P(A_2 \mid A_3) = \frac{1}{2}$. But $P(A_1 \cap A_2 \cap A_3) = 0$, $P(A_1 \mid A_2 \cap A_3) = 0$, etc.

Definition 1.16. Consider events $A_1, A_2, B \in S$ with $P(B) > 0$. A_1 and A_2 are **conditionally independent** given B if

$$P(A_1 \mid A_2 \cap B) = P(A_1 \mid B)$$
$$\iff P(A_1 \cap A_2 \mid B) = P(A_1 \mid B)P(A_2 \mid B)$$

Theorem 1.11 (General Multiplication Rule)**.**

$$P(A_1, A_2, \ldots, A_K) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2)\cdots \cdots P(A_K \mid A_1 \cap A_2 \cap \cdots \cap A_{K-1})$$

Proof. Trivial, recursively (*Chain Rule* for probabilities). □

We can use **probability trees** to display joint probabilities of multiples, where the junctions are the events and the branches correspond to possible sequences of choices of events. We then *multiply along a branch* to get the joint probability.

1.7 *Important Theorems*

Theorem 1.12 (Theorem of Total Probability)**.** Given a partition $B_1, \ldots, B_n$ of the sample space S into n subsets s.t. $P(B_i) > 0$, then for any event $A \in S$,

$$P(A) = \sum_{i=1}^{n} P(A \mid B_i)P(B_i)$$

Proof. *(Theorem 1.12)*

$$\begin{aligned} A &= (A \cap B_1) \cup \cdots \cup (A \cap B_n) \\ P(A) &= P(A \cap B_1) + \cdots + P(A \cap B_n) \\ &= P(A \mid B_1)P(B_1) + \cdots + P(A \mid B_n)P(B_n) \\ &= \sum_{i=1}^{n} P(A \mid B_i)P(B_i) \end{aligned}$$

Theorem 1.13 (Bayes's Theorem)**.** For any $A, B \in S$ where $P(A) > 0$ and $P(B) > 0$:

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)}$$

Furthermore, given $B_1, B_2, \ldots, B_n$ forming a partition of S:

$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{P(A)} = \frac{P(A \mid B_i)P(B_i)}{\sum_{j=1}^{n} P(A \mid B_j)P(B_j)}$$

Proof. *(Theorem 1.13).* By Definition 1.13,

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

With the stated conditions, we have

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)}$$

The second part of the theorem comes from replacing the denominator using Theorem 1.12. □

Warning 5. Although $\sum_{i=1}^{n} P(B_i \mid A) = 1$, $\sum_{i=1}^{n} P(A \mid B_i) \neq 1$ due to the use of different conditioning sets.

Remark 6. Bayes' Theorem is often used to make probability statements concerning an event B that *has not* been observed, given an event A that *has* been observed.

Remark 7 (Prosecutor's Fallacy).

$$P(\text{"Evidence"} \mid \text{"Guilt"}) \neq P(\text{"Guilt"} \mid \text{"Evidence"})$$

Remark 8. Recall that the odds on B are defined as $\frac{P(B)}{P(B')} = \frac{P(B)}{1-P(B)}$; and the conditional odds given A ($P(A) > 0$) are $\frac{P(B \mid A)}{P(B' \mid A)} = \frac{P(B \mid A}{1-P(B \mid A)}$. Thus the odds change by a factor of $\frac{P(A \mid B)}{P(A \mid B')}$ when one adds the information A.

Corollary 1.14 (Bayes' Theorem for multiple events)**.** Given events $A_1, A_2 \in S$ where $P(A_1 \cap A_2) > 0$ and given $B_1, B_2, \ldots, B_n$ forming a partition of S:

$$P(B_i \mid A_1 \cap A_2) = \frac{P(A_1 \cap A_2 \mid B_i)P(B_i)}{\sum_{k=1}^{n} P(A_1 \cap A_2 \mid B_k)P(B_k)}$$

Example 1.3 (Medical screening)**.** Given events A (positive screening test) and B (person is actually a sufferer). We are interested in $P(B \mid A)$. Suppose $P(B) = p$; $P(A \mid B) = 1 - \alpha$ (true positive rate); $P(A \mid B') = \beta$ (false positive rate) for probabilities p, α, β.

$$\begin{aligned} P(A) &= P(A \mid B)P(B) + P(A \mid B')P(B') \\ &= (1-\alpha)p + \beta(1-p) \\ P(B \mid A) &= \frac{P(A \mid B)P(B)}{P(A)} \\ &= \frac{(1-\alpha)p}{(1-\alpha)p + \beta(1-p)} \\ P(B \mid A') &= \frac{\alpha p}{\alpha p + (1-\beta)(1-p)} \end{aligned}$$

by similarly logic.

2 Random Variables and Probability Distributions

2.1 Random Variables

Definition 2.1. A **random variable** is a mapping $Y : S \to \mathbb{R},\ s \mapsto y$ from an arbitrary space S (where the experiment sample outcomes are defined) to the real line $\mathbb{R}$.

Remark 9. We use uppercase letters X,Y, Z to denote random variables and lowercase x, y, z to denote the real values taken by the random variable.

$Y(.)$ is usually a many-to-one mapping. Also, we usually suppress the dependence

Definition 2.2. A random variable is **discrete** if the image of Y is a countable set $\{y_1, y_2, \ldots, y_n, \ldots\}$ denoted $\mathcal{Y}$ for positive probabilities.

Corollary 2.1. The probabilities assigned to individual points in the set $\mathcal{Y}$ must satisfy

$$\sum_{y \in \mathcal{Y}} P(Y = y) = 1$$

Proof. folllows from the probability axioms. □

2.2 Probability Mass Function

Definition 2.3. The **probability mass function** (pmf) $p(.)$ for the random variable Y records how probability is distributed across points in $\mathbb{R}$: $p(y) = P(Y = y)$ where $p(y) > 0$ for $y \in \mathcal{Y}$. It is also denoted $f(y)$, $p_Y(y)$ or $f_Y(y)$. It must:

(a) specify probabilities, i.e. $0 \leq p(y) \leq 1$

(b) satisfy Corrolary 2.1: $\sum_{y \in \mathcal{Y}} p(y) = 1$

Remark 10. The pmf can be defined pointwise (for a discrete random variable) or in some functional form.

Remark 11. All pmfs take the form $p(y) = cg(y)$ where $g(y)$ also contains information on the set $\mathcal{Y}$ for which $P(y) > 0$.

$$\sum_{y \in \mathcal{Y}} p(y) = 1 \implies \sum_{y \in \mathcal{Y}} g(y) = \frac{1}{c}$$

2.3 Moments: expectation and variance

Definition 2.4. The **expectation** (expected value) $\mathbb{E}[Y]$ of the r.v. Y is $\mathbb{E}[Y] = \sum_{y \in \mathcal{Y}} yp(y)$, the centre of mass of the probability distribution.

Example 2.1 (Discrete Uniform Distribution)**.**
Suppose $\mathcal{Y} = \{y_1, \ldots, y_n\},\ p(y) = \frac{1}{n}$.
Then $\mathbb{E}[Y] = \frac{1}{n}\sum_{i=1}^{n} y_i = \overline{y}$
Here $p(y) = c$ is constant and the number of elements in $\mathcal{Y}$ is $\frac{1}{c}$.

Let's now deal with the expectation of a function $g(Y)$.

Definition 2.5. Let $g(.)$ real-valued function and $p(y)$ the pmf of Y:

Set of values of $g(Y)$: $\mathcal{G} = \{g_i,\ i = 1, \ldots, m :\ g_i = g(y),\ \text{some } y \text{ in } \mathcal{Y}\}$

$$\mathbb{E}[g(Y)] = \sum_{y \in \mathcal{Y}} g(y)p(y)$$

Definition 2.6. The **variance** $\mathbb{V}[Y]$ of Y is a measure of how dispersed the probability distribution is, defined as

$$\begin{aligned}\mathbb{V}[Y] &= \mathbb{E}[(Y-\mu)^2] \text{ where } \mu = \mathbb{E}[Y] \\ &= \sum_y (y-\mu)^2 p(y) \geq 0\end{aligned}$$

Definition 2.7. The **standard deviation** of Y is $\sqrt{\mathbb{V}[Y]}$.

Theorem 2.2. The expectation $\mathbb{E}[Y]$ satisfies linearity.

Remark 12. $\mathbb{V}[Y]$ can be infinite even if $\mathbb{E}[Y]$ is bounded.

Theorem 2.3. $\mathbb{E}[(Y-\mu)^2] = \mathbb{E}[Y^2] - \mu^2$

Corollary 2.4. $\mathbb{V}[aY+b] = a^2\mathbb{V}[Y]$

2.4 *Binomial distribution*

Considering an experiment with two outcomes 'success' and 'failure', define the random variable Y to be $Y = 1$ for 'success' and $Y = 0$ for 'failure', with $P(1) = p,\ P(0) = 1 = 1 - p$.

Definition 2.8. The **Bernoulli** distribution is defined with pmf

$$p_Y(y) = p^y q^{1-y} = p^y(1-p)^{1-y} \quad \forall y \in \{0,1\}, \text{ given } p \in (0,1)$$

Theorem 2.5. The expectation of the Bernoulli dist. is $\mu = \mathbb{E}[Y] = p$.

Theorem 2.6. The variance of the Bernoulli dist. is $\mathbb{V}[Y] = p(1-p)$.

We now extend the definition to a sequence of multiple experiments with two outcomes, and have the r.v. Y record the number of 'successes': $Y = y \Leftrightarrow$ "y successes in n trials". Consider a sequence with y successes and $n-y$ failures: the probability of any such sequence is $p^y q^{n-y}$. The number of possible such sequences is $C_y^n = \binom{n}{y}$.

Definition 2.9. The **Binomial** distribution is defined with pmf

$$p_y(y) \equiv P(Y=y) = \binom{n}{y} p^y q^{n-y} \quad \forall y \in \{0,1,2,\dots,n\}$$

Theorem 2.7. The expectation of the binomial dist. is $\mathbb{E}[Y] = np$.

Proof. *(Theorem 2.2)*
Let $g(Y) = ag_1(Y) + bg_2(Y)$ for separate $g_1(.)$ and $g_2(.)$.

$$\begin{aligned}&\mathbb{E}[g(Y)] = \mathbb{E}[ag_a(Y) + bg_2(Y)] \\ &= \sum_y (ag_1(y) + bg_2(y))p(y) \\ &= a\sum_y g_1(y)p(y) + b\sum_y g_2(y)p(y) \\ &= a\mathbb{E}[g_1(Y)] + b\mathbb{E}[g_2(Y)]\end{aligned}$$ □

Proof. *(Theorem 2.3)*

$$\begin{aligned}&\mathbb{E}[(Y-\mu)^2] = \mathbb{E}[Y^2 + 2Y\mu + \mu^2] \\ &= \mathbb{E}[Y^2] - 2\mu\mathbb{E}[Y] + \mu^2 \\ &= \mathbb{E}[Y^2] - 2\mu^2 + \mu^2 \\ &= \mathbb{E}[Y^2] - \mu^2\end{aligned}$$ □

Proof. *(Corollary 2.4)*

$$\begin{aligned}&\mathbb{V}[aY+b] \\ &= \mathbb{E}\left[\left(aY + b - \mathbb{E}[aY+b]\right)^2\right] \\ &= \mathbb{E}\left[\left(aY + b - (a\mathbb{E}[Y] + b)\right)^2\right] \\ &= \mathbb{E}[(aY + b - a\mu - b)^2] \\ &= \mathbb{E}[a^2(Y-\mu)^2] \\ &= a^2\mathbb{V}[Y]\end{aligned}$$ □

Proof. *(Theorem 2.5)*

$$\begin{aligned}\mathbb{E}[Y] &= \sum_{i=0}^{1} yp(y) = 0p(0) + 1p(1) \\ &= 0 \cdot q + 1 \cdot p = p\end{aligned}$$ □

Proof. *(Theorem 2.6)*

$$\begin{aligned}&\mathbb{V}[Y] = \mathbb{E}[Y^2] - \mu^2 \\ &= \sum_{i=0}^{1} y^2 p(y) - \mu^2 = 1^2 p(1) \\ &= p - p^2 = p(1-p) = pq\end{aligned}$$ □

Proof. *(Theorem 2.7)*

$$\begin{aligned}&\mathbb{E}[Y] = \sum_{y=0}^{n} y\binom{n}{y} p^y q^{n-y} \\ &= n\sum_{y=1}^{n} \frac{(n-1)!}{(y-1)!(n-y)!} p^y q^{n-y} \\ &= n\sum_{j=0}^{n-1} \frac{(n-1)!}{j!((n-1)-j)!} p^{j+1} q^{n-j-1} \\ &= np\sum_{j=0}^{n-1} \frac{(n-1)!}{j!((n-1)-j)!} p^j q^{n-1-j} \\ &= np(p+q)^{n-1} = np\end{aligned}$$ □

Theorem 2.8. The variance of the binomial dist. is $\mathbb{V}[Y] = npq$.

Proof.

$$\begin{aligned}
\mathbb{E}[Y^2] &= \mathbb{E}[Y(Y-1)] + \mathbb{E}[Y] \text{ (by linearity)} \\
\mathbb{E}[Y(Y-1)] &= \sum_{y=0}^{n} y(y-1)\binom{n}{y} p^y q^{n-y} = \sum_{y=2}^{n} y(y-1)\frac{n!}{y!(n-y)!} p^y q^{n-y} \\
&= n(n-1)\sum_{y=2}^{n} \frac{(n-2)!}{(y-2)!(n-y)!} p^y q^{n-y} \\
&= n(n-1)\sum_{j=0}^{n-2} \frac{(n-2)!}{j!(n-2-j)!} p^{j+2} q^{n-j-2} \\
&= n(n-1)p^2 \sum_{j=0}^{n-2} \frac{(n-2)!}{j!(n-2-j)!} p^j q^{n-j-2} \\
&= n(n-1)p^2(p+1)^{n-2} = n(n-1)p^2 \\
\mathbb{V}[Y] &= \mathbb{E}[Y^2] - \mu^2 = n(n-1)p^2 + np - (np)^2 \\
&= np - np^2 = npq
\end{aligned}$$

□

Remark 13. Notice how the Bernoulli distribution is a special case of the binomial distribution with $n = 1$.

2.5 Geometric Distribution

Consider independent binary trials, where one counts the number of trials carried out until the first success. Let Y be the r.v. recording the number of trials carried out up to an including the first success: $Y = y \Leftrightarrow$ "$y-1$ failures followed by 1 success".

Definition 2.10. The **geometric** distribution is defined with pmf

$$p(y) \equiv P(Y = y) = q^{y-1}p \quad \forall y \in \{1, 2, \dots\}$$

Remark 15. The geometric distribution can also be written with pmf

$$p(y) \equiv P(Y^* = y) = q^y p \quad \forall y \in \{0, 1, 2, \dots\}$$

where the r.v. Y^* then records the number of failures observed before the first success: $Y^* = Y - 1$.

Theorem 2.9. The expectation of the geometric dist. is $\mathbb{E}[Y] = \frac{1}{p}$.

Theorem 2.10. The variance of the geometric dist. is $\mathbb{V}[Y] = \frac{1-p}{p^2}$.

Definition 2.11. The **cumulative distribution function** (cdf) of a distribution is defined as $F(y) = P(Y \leq y)$.

Remark 14. Notice that

$$\sum_{y=1}^{\infty} q^{y-1}p = p\sum_{j=0}^{\infty} q^j = p\frac{1}{1-q} = 1$$

Proof. *(of Theorem 2.9)*

$$\begin{aligned}
\mathbb{E}[Y] &= \sum_{y=0}^{\infty} yq^{y-1}p = p\sum_{y=1}^{\infty} yq^{y-1} \\
&= p\frac{1}{(1-q)^2} = p\frac{1}{p^2} = \frac{1}{p}
\end{aligned}$$

□

Proof. *(of Theorem 2.10)*

$$\begin{aligned}
\mathbb{E}[Y(Y-1)] &= \sum_{y=0}^{\infty} y(y-1)p(y) \\
&= pq\sum_{y=2}^{\infty} y(y-1)q^{y-2} = pq\frac{d^2}{dq^2}\left(\sum_{y=0}^{\infty} q^y\right) \\
&= pq\frac{d^2}{dq^2}\left(\frac{1}{1-q}\right) = pq\frac{2}{(1-q)^3} = 2\frac{q}{p^2} \\
\mathbb{V}[Y] &= \mathbb{E}[Y(Y-1)] + \mathbb{E}[Y] - (\mathbb{E}[Y])^2 \\
&= 2\frac{1-p}{p^2} + \frac{1}{p} - \frac{1}{p^2} = \frac{1-p}{p^2}
\end{aligned}$$

□

Theorem 2.11. The cdf of the geometric distribution is

$$F(y) = P(Y \leq y) = 1 - (1-p)^y \quad y \in \{1, 2, \dots\}$$

Proof.

$$\begin{aligned} P(Y > y) &= \sum_{j=y+1}^{\infty} P(Y = j) = \sum_{j=y+1}^{\infty} q^{j+1} p = pq^y \sum_{j=y+1}^{\infty} q^{j-y-1} = \frac{pq^y}{1-q} \\ &= q^y \quad \forall y \in \{1, 2, \dots\} \\ P(Y \leq y) &= 1 - P(Y > y) = 1 - (1-p)^y \quad \forall y \in \{1, 2, \dots\} \end{aligned}$$

□

We can define all 'interval' probabilities such as $P(Y \leq a)$, $P(a \leq Y \leq b)$, etc. via $F(y)$, since $\forall$ distributions,

$$P(Y \in \mathcal{X}) = \sum_{y \in \mathcal{X}} p(y)$$

2.6 *Negative Binomial Distribution*

Consider independent binary trials, where one counts the number of trials carried out until the rth success. Let r.v. Y record the number of trials carried out until the rth success: $Y = y \Leftrightarrow A \cap B$: $A =$ "$r-1$ successes in the first $y-1$ trials" and $B =$ "success on the yth trial".

Definition 2.12. The **negative binomial** distribution has pmf

$$p(y) \equiv P(Y = y) = \binom{y-1}{r-1} p^r q^{y-r} \quad \forall y \in \{r, r+1, \dots\}$$

Exercise 2.2. Show for the negative binomial distribution that

$$\sum_{y=r}^{\infty} p(y) = 1$$

Remark 16. As for the geometric distribution, the negative binomial dist. can also be written with pmf

$$p(y) = P(Y^* = y) = \binom{y+r-1}{r-1} p^r q^y \quad \forall y \in \{0, 1, 2, \dots\}$$

where the r.v. Y^* records the number of failures observed up to the rth success: $Y^* = Y - r$.

Theorem 2.12. The expectation of the negative binomial distribution is $\mathbb{E}[Y] = \frac{r}{p}$.

Theorem 2.13. The variance of the negative binomial distribution is $\mathbb{V}[Y] = \frac{r(1-p)}{p^2}$.

Proofs left to the reader.

2.7 *Hypergeometric Distribution*

Consider a finite set of N items with r type I and $N-r$ type II. We select n items without replacement. Let r.v. $Y = y \Leftrightarrow$ "y type I items".

Definition 2.13. The **hypergeometric** distribution is defined with pmf

$$p(y) \equiv P(Y = y) = \frac{\binom{r}{y} \times \binom{N-r}{n-y}}{\binom{N}{n}} \quad \forall y \in \mathcal{Y}$$

$\binom{r}{y}$: number of ways of selecting y type I from the r available.
$\binom{N-r}{n-y}$: number of ways of selecting $n-y$ type II from the $N-r$ available.
$\binom{N}{n}$: total number of ways of selecting n items from N.

Remark 17. An equivalent expression for $p(y)$ is

$$p(y) = \frac{\binom{n}{y}\binom{N-n}{r-y}}{\binom{N}{r}}$$

Remark 18. When N and r are large, we can approximate the hypergeometric to a binomial distribution

$$p(y) \approx \binom{n}{y}\left(\frac{r}{N}\right)^y\left(1-\frac{r}{N}\right)^{n-y}$$

2.8 Poisson distribution

Consider the binomial experiment setting, and take the limiting case where we allow n to get larger and p to get smaller, but keeping $n \times p$ constant: i.e. take $p = \frac{\lambda}{n}$ and let n large. Then,

$$\begin{aligned} p(y) &= \binom{n}{y}p^y(1-p)^{n-y} = \frac{n!}{y!(n-y)!}\left(\frac{\lambda}{n}\right)^y\left(1-\frac{\lambda}{n}\right)^{n-y} \\ &= \frac{\lambda^y}{y!}\left(1-\frac{\lambda}{n}\right)^n \underbrace{\frac{n!}{(n-y)!n^y}\left(1-\frac{\lambda}{n}\right)^{-y}}_{\to 1 \text{ as } n\to\infty} \to \frac{\lambda^y}{y!}e^{-\lambda} \end{aligned}$$

λ: rate at which successes occur in a continuous approximation to the experiment.

An example use case would be: "If Mike makes an average number of 2 sales a day, what are the odds that he makes 3 sales tomorrow?"

Definition 2.14. The **Poisson** distribution is defined with pmf

$$p(y) = \frac{\lambda^y}{y!}e^{-\lambda} \quad \forall y \in \{0,1,2,\dots\}$$

Theorem 2.14. The expectation of the Poisson dist. is $\mathbb{E}[Y] = \lambda$.

Theorem 2.15. The variance of the Poisson dist. is $\mathbb{V}[Y] = \lambda$.

Remark 19. The Poisson distribution corresponds to a model where incidents occur independently at a constant rate λ.

Proof. *(Theorem 2.14)*

$$\begin{aligned} \mathbb{E}[Y] &= \sum_{y=0}^{\infty} yp(y) = \sum_{y=0}^{\infty} y\frac{\lambda^y}{\lambda!}e^{-\lambda} \\ &= e^{-\lambda}\sum_{y=1}^{\infty}\frac{\lambda^y}{(y-1)!} \\ &= e^{-\lambda}\lambda\sum_{y=1}^{\infty}\frac{\lambda^{y-1}}{(y-1)!} \\ &= e^{-\lambda}\lambda e^{\lambda} = \lambda \end{aligned}$$

□

Proof. *(Theorem 2.15)*

$$\begin{aligned} &\mathbb{E}[Y(Y-1)] = \sum_{y=1}^{\infty} y(y-1)\frac{\lambda^y}{y!}e^{-\lambda} \\ &= e^{-\lambda}\lambda^2\sum_{y=2}^{\infty}\frac{\lambda^{y-2}}{(y-2)!} = e^{-\lambda}\lambda^2 e^{\lambda} \\ &= \lambda^2 \\ &\mathbb{V}[Y] = \mathbb{E}[Y(Y-1] + \mathbb{E}[Y] - \mu^2 \\ &= \lambda^2 + \lambda - \lambda^2 = \lambda \end{aligned}$$

□

2.9 Moments and Moment-Generating Functions

Definition 2.15. Let $\mu = \mathbb{E}[Y]$. We define

$$\begin{aligned} \textbf{kth moment:} \quad & \mu'_k = \mathbb{E}[Y^k] = \sum_y y^k p(y) \\ \textbf{kth central moment:} \quad & \mu_k = \mathbb{E}[(Y-\mu)^k] = \sum_y (y-\mu)^k p(y) \end{aligned}$$

Definition 2.16. Generating functions $g(t)$ for the sequence $\{a_j\} \in \mathbb{R}$ is the function defined by the sum

$$g(t) = \sum_j a_j t^j$$

Example 2.3. The generating function for $\{a_j\} = \binom{n}{j} \quad \forall j = 0,1,\dots$ is

$$g(t) = (1+t)^n = \sum_{j=0}^{n}\binom{n}{j}t_j$$

Definition 2.17. Exponential generating function: $g(t) = \sum_j a_j \frac{t^j}{j!}$.

Definition 2.18. For pmf $p(y)$, the **moment generating function** (mgf) $m(t)$ is the exponential generating function for the moments of $p(y)$, requiring the sum to be finite in the neighbourhood of $t = 0$:

$$m(t) = \sum_{j=0}^{\infty} \mu'_j \frac{t^j}{j!}$$

Remark 20. It is difficult to see the use for moment generating functions right now, but it can be used to identify distributions, since two random variables with the same mgf share the same distribution (Theorem 2.18), and it can be used to calculate the moments of a random variables.

Theorem 2.16. The moment generating function can be written

$$m(t) = \mathbb{E}\left[e^{tY}\right]$$

Proof. *(Theorem 2.16)*

$$\begin{aligned} m(t) &= \sum_{j=0}^{\infty} \mu'_j \frac{t^j}{j!} = \sum_{j=0}^{\infty} \sum_{y} y^j p(y) \frac{t^j}{j!} \\ &= \sum_{y} \left(\sum_{j=0}^{\infty} \frac{(yt)^j}{j!} \right) p(y) \\ &= \sum_{y} e^{ty} p(y) \end{aligned}$$

□

Theorem 2.17. The kth derivative of $m(t)$ evaluated at $t = 0$ gives

$$m^{(k)}(0) = \frac{d^k}{dt^k} m(0) = \mu'_k$$

Proof. *(Theorem 2.17)*
Simply write out the terms. □

Theorem 2.18 (Uniqueness)**.** Suppose $m_1(t)$ and $m_2(t)$ are two mgfs for pmfs $p_1(t)$ and $p_2(t)$. Then:

$$m_1(t) = m_2(t)\ \forall |t| \leq b \text{ for some } b \iff p_1(y) = p_2(y)\ \forall y$$

Proof. *(Theorem 2.18)*
Beyond the scope of this course. □

2.10 *Probability-Generating Functions*

Definition 2.19. The **probability-generating function** (pgf) for random variable Y with $P(Y = i) = y_i$ is

$$G(t) = \mathbb{E}\left[t^Y\right] = \sum_{i=0}^{\infty} p_i t^i$$

Definition 2.20. The k**th factorial moment** $\mu_{[k]}$ is defined as

$$\mu_{[k]} = \mathbb{E}\left[Y(Y-1)\cdots(Y-k+1)\right]$$

Remark 21. Note that

$$G^{(k)}(t) = \frac{d^{(k)}(t)}{dt^k} \left(\sum_{i=0}^{\infty} p_i t^i \right) = \sum_{i=k}^{\infty} i(i-1)\cdots(i-k+1) p_i t^{i-k}$$

$$G^{(k)}(1) = \sum_{i=k}^{\infty} i(i-1)\cdots(i-k+1) p_i = \mu_{[k]}$$

Corollary 2.19. $G(t) \equiv \mathbb{E}\left[t^Y\right] = \mathbb{E}\left[\exp(Y \ln t)\right] = m(\ln t)$

2.11 Continuous Random Variables

Theorem 2.20. Properties of the cdf $F(Y)$ (c.f. Definition 2.11):

1. "starts at zero": $F(-\infty) \equiv \lim_{y \to -\infty} F(y) = 0$
2. "ends at one": $F(\infty) \equiv \lim_{y \to \infty} F(y) = 1$
3. "non-decreasing in between": $y_1 < y_2 \implies F(y_1) \leq F(y_2)$

Definition 2.21. A r.v. Y with cdf $F(y)$ is called **continuous** if $F(y)$ is continuous for all real y. i.e. $\forall y$,

$$\lim_{\varepsilon \to 0^+} F(y) = \lim_{\varepsilon \to 0^-} F(y) = F(y)$$

Warning 22. If $F(y)$ is continuous, then the probability attached to any individual point in the space is null: $P(Y = y) = 0 \; \forall y \in \mathbb{R}$.

2.12 Probability Density Function

Definition 2.22. For a continuous cdf $F(y)$, the **probability density function** (pdf) $f(y)$ is the first derivative of $F(.)$ at y when it exists:

$$f(y) = \frac{dF(y)}{dt} \implies F(y) = \int_{-\infty}^{y} f(t)dt$$

Theorem 2.21. Properties of the pdf $f(y)$:

1. non-negative: $f(y) \geq 0 \quad -\infty < y < \infty$
2. integrates to 1: $\int_{-\infty}^{\infty} f(y)dy = 1$

Proof. Follows from probability axioms - properties from Theorem 2.20. □

Warning 23. The probability density function does NOT represent probabilities; it does not integrate to 1.

We can compute the interval probability $P(a \leq y \leq b)$ as

$$P(a \leq Y \leq b) = P(Y \leq b) - P(Y \leq a) = \int_{a}^{b} f(t)dt$$

ALL THE DEFINITIONS AND THEOREMS shown for the discrete case still hold for the continuous case:

1. EXPECTATION: for continuous r.v. Y and a function $g(Y)$,

$$\mathbb{E}[Y] = \int_{-\infty}^{\infty} yf(y)dy = \int_{\mathcal{Y}} yf(y)dy = \mu$$
$$\mathbb{E}[g(Y)] = \int_{-\infty}^{\infty} g(y)f(y)dy \qquad \text{(both: only if integral is finite)}$$

2. LINEARITY OF EXPECTATION (Theorem 2.2) still holds.
3. VARIANCE: for continuous r.v. Y,

$$\mathbb{V}[Y] = \mathbb{E}[(Y - \mu)^2] = \int_{-\infty}^{\infty} (y - \mu)^2 f(y)dy$$

4. MOMENTS, CENTRAL MOMENTS AND FACTORIAL MOMENTS:

$$\mu'_k = \mathbb{E}\left[Y^k\right] = \int_{-\infty}^{\infty} y^k f(y)dy$$
$$\mu_k = \mathbb{E}\left[(Y-\mu)^k\right] = \int_{-\infty}^{\infty} (y-\mu)^k f(y)dy$$
$$\mu_{[k]} = \mathbb{E}\left[Y(Y-1)\cdots(Y-k+1)\right] = \int_{-\infty}^{\infty} y\cdots(y-k+1)f(y)dy$$

5. MOMENT GENERATING FUNCTION $m(t)$: for $|t| \leq b,\ b > 0$

$$m(t) = \mathbb{E}\left[e^{tY}\right] = \int_{-\infty}^{\infty} e^{ty} f(y)dy$$

6. FACTORIAL MOMENT GENERATING FUNCTION $G(t)$: for $1-b \leq t \leq 1+b,\ b > 0$,

$$G(t) = \mathbb{E}\left[t^Y\right] = \int_{-\infty}^{\infty} t^y f(y)dy$$

2.13 *Distributions for continuous RVs*

Definition 2.23. The **continuous uniform distribution** has a pdf that is constant on some interval $(\theta_1, \theta_2),\ \theta_1 < \theta_2$:

$$f(y) = \frac{1}{\theta_1 - \theta_2} \quad \forall\ \theta_1 \leq y \leq \theta_2 \text{ and zero otherwise.}$$

By direct calculation, we can compute that the cdf $F(y)$ is

$$F(y) = \begin{cases} 0 & y \leq \theta_1 \\ \frac{y-\theta_1}{\theta_2-\theta_1} & \theta_1 \leq y \leq \theta_2 \\ 1 & y \geq \theta_2 \end{cases}$$

and also find $\mathbb{E}\left[Y\right] = \frac{\theta_1+\theta_2}{2}$ and $\mathbb{V}\left[Y\right] = \frac{(\theta_2-\theta_1)^2}{12}$.

Definition 2.24. The **Normal** (Gaussian) distribution has pdf defined with parameters σ and μ by the pdf

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right) \quad \forall\ -\infty < y < \infty$$

It can be shown that $\mathbb{E}\left[Y\right] = \mu$ and $\mathbb{V}\left[Y\right] = \sigma^2$.
The cdf $F(y) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(t-\mu)^2\right) dt$ which cannot be computed analytically.
Instead, if Y has Normal distribution with parameters μ and σ, and we consider the r.v. $Z = \frac{Y-\mu}{\sigma}$, then $\forall z \in \mathbb{R}$,

$$\begin{aligned} F_Z(z) &= P(Z \leq z) = P\left(\sigma Z + \mu \leq \sigma z + \mu\right) = P\left(Y \leq \sigma z + \mu\right) \\ &= F_Y(\sigma z + \mu) = \int_{-\infty}^{\sigma z - \mu} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(t-\mu)^2\right) dt \\ &= \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) \text{ substituting } u = (t-\mu)/\sigma \end{aligned}$$

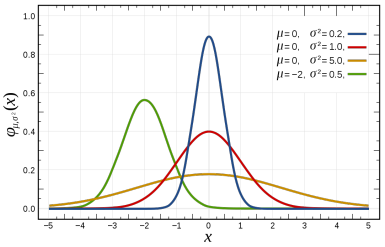

Figure 1: Normal distribution for various values of μ and σ.

Definition 2.25. The Gamma distribution with parameters $\alpha,\ \beta > 0$ has pdf

$$f(y) = \begin{cases} 0 & y < 0 \\ \frac{1}{\beta^\alpha \Gamma(\alpha)} y^{\alpha-1} e^{-y/\beta} & 0 \le y < \infty \end{cases}$$

where $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt$ which satisfies $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$, and for $n \in \mathbb{N}$, $\Gamma(n) = (n-1)!$ with $\Gamma(1) = 1$.

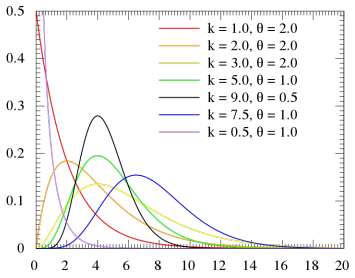

Figure 2: Gamma distribution for various values of α and β, which are respectively the shape and scale parameters.

It can be shown that $\mathbb{E}[Y] = \alpha\beta$ and $\mathbb{V}[Y] = \alpha\beta^2$.

Special cases of the Gamma Distribution are

1. chi-square distribution μ degrees of freedom: $\alpha = \mu/2,\ \beta = 2$
2. exponential distribution: $\alpha = 1$, $f(y) = \frac{1}{\beta} e^{-y/\beta} \quad 0 \le y < \infty$

The moment-generating function of the Gamma distribution is

$$m(t) = \mathbb{E}[e^{tY}] = \int_0^\infty e^{ty} \frac{1}{\beta^\alpha \Gamma(\alpha)} y^{\alpha-1} e^{-y/\beta} dy = \left(\frac{1}{1-\beta t}\right)^\alpha$$

The Gamma distribution is to the Exponential distribution what the Binomial is to the Bernoulli distribution.

e.g use case of the Gamma dist.: "How long does it take for a bucket to fill up if left out in the rain?"

The exponential distribution is known as the waiting-time distribution.

Definition 2.26. The **beta distribution** is suitable for an r.v. defined on $[0,1]$, with pdf

$$f(y) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1} = \frac{1}{B(\alpha,\beta)} y^{\alpha-1}(1-y)^{\beta-1} \quad 0 \le y \le 1$$

where $B(\alpha,\beta)$ is the **Beta** function $B(\alpha,\beta) = \int_0^1 y^{\alpha-1}(1-y)^{\beta-1} dy$.

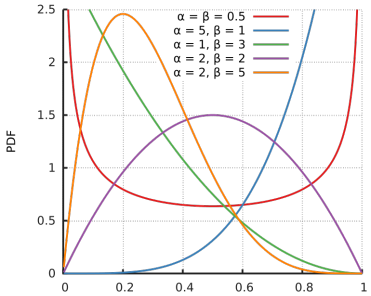

Figure 3: Beta distribution for various values of α and β.

It can be found that $\mathbb{E}[Y^k] = \frac{1}{B(\alpha,\beta)} B(\alpha+k,\beta)$ for $k = 1,2,\ldots$

Remark 24. A special case of the beta function with $\alpha = \beta = 1$ has $B(1,1) = \frac{\Gamma(1)\Gamma(1)}{\Gamma(2)} = \frac{1\cdot 1}{1} = 1$ and results in the pdf

$$f(y) = 1 \quad \forall\, 0 \le y \le 1 \quad \text{and zero otherwise}$$

the continuous uniform distribution with $\theta_1 = 0$, $\theta_2 = 1$.

3 Probability Calculation Methods

3.1 Transformations in One Dimension

Given r.v. Y defined on $\mathcal{Y}$ (the region on which Y is nontrivial), we are interested in the transformed r.v. U (defined on $\mathcal{U}$), by

$$U = h(Y) \text{ where } h(.) : \mathcal{Y} \to \mathbb{R}$$

Discrete Case

Suppose $h(.)$ maps values from (countable) $\{y_1, y_2, \dots, y_n, \dots\}$ to (countable) $\{u_1, u_2, \dots, u_n, \dots\}$ with $u_i = h(y_j)$.

We typically deal with simple transformations, e.g.

$$U = 2Y \qquad U = \exp(-Y)$$
$$U = \frac{Y}{Y+1} \qquad U = Y^2$$

- if $h(.)$ is 1-1, then identify the set $\{u_i\}_{i=1}^{\infty}$ that $\{y_i\}_{i=1}^{\infty}$ map onto, and for each i, assign $p_Y(y_i)$ to u_i to obtain $p_U(u_i)$.
- if $h(.)$ is not 1-1, we must identify the set $\mathcal{U}$ and compute for each case the probability assigned to the values in the set.

Continuous Case

We again want to see how values in the set $\mathcal{Y}$ are mapped to a set $\mathcal{U}$. Consider $\mathcal{B} \subseteq \mathcal{U}$ the set of values that points in $\mathcal{A} \subseteq \mathcal{Y}$ can be mapped to under the transformation $h(.)$:

$$\mathcal{B} = \{u : u = h(y) \,|\, y \in \mathcal{A}\} \implies P(Y \in \mathcal{A}) = P(U \in B)$$

Recall that for a continuous r.v. Y,

$$P(Y = y) = 0 \; \forall y \in \mathbb{R}$$

So to compute $F_U(.)$ or $f_U(.)$, $B \subseteq \mathcal{U}$, so fix $\mathcal{B}$. If $h(.)$ is a 1-1 transformation, then the inverse transformation $h^{-1}(.)$ can usually be computed explicitly, We know $P(U \in \mathcal{B}) = P(Y \in h^{-1}(\mathcal{B}))$, and

Theorem 3.1. $f_U(u) = f_Y(h^{-1}(u)) \left| \frac{dh^{-1}(u)}{du} \right|$

Proof. Two cases depending on $h(.)$

- 1-1 increasing:

$$F_U(u) = P(U \le u) = P(h(Y) \le u) = P(Y \le h^{-1}(u)) = F_Y(h^{-1}(u))$$
$$f_U(u) = \frac{dF_U(u)}{du} = \frac{dF_Y(h^{-1}(u)}{du} = f_Y\left(h^{-1}(u)\right) \underbrace{\frac{dh^{-1}(u)}{du}}_{\ge 0}$$

The identity $P(U \in \mathcal{B}) = P(Y \in h^{-1}(\mathcal{B})$ can be re-written as

$$\int_{\mathcal{B}} f_U(u)du = \int_{h^{-1}(\mathcal{B})} f_Y(y)dy$$

- 1-1 decreasing:

$$F_U(u) = P(U \le u) = P(h(Y) \le u) = P(Y \ge h^{-1}(u)) = 1 - F_Y(h^{-1}(u))$$
$$f_U(u) = \frac{dF_U(u)}{du} = -\frac{dF_Y(h^{-1}(u))}{du} = -f_Y(h^{-1}(u)) \underbrace{\frac{dh^{-1}(u)}{du}}_{\le 0}$$

□

Example 3.1. If $U = cY + d$ for $c > 0$ and d,

$$m_U(t) = \mathbb{E}\left[e^{tU}\right] = \mathbb{E}\left[e^{t(cY+d)}\right]$$
$$= \mathbb{E}\left[e^{(ct)Y}e^{dt}\right] = e^{dt}\mathbb{E}\left[e^{(ct)Y}\right]$$
$$\implies m_U(t) = e^{dt}m_Y(ct)$$

3.2 *Techniques for sums of random variables*

Discrete Random Variables

For discrete random variables Y_1 and Y_2 independent taking values on $\mathcal{Y}_1$ and $\mathcal{Y}_2$ respectively. Let $Y = Y_1 + Y_2$.

$$\begin{aligned} P(Y=y) &= P\left(\bigcup_{y_1\in\mathcal{Y}_1} (Y=y)\cap(Y_1=y_1)\right) \\ &= \sum_{y_1\in\mathcal{Y}_1} P\left((Y=y)\cap(Y_1=y_1)\right) \\ &= \sum_{y_1\in\mathcal{Y}_1} P\big((Y_2=y-y_1)\cap(Y_1=y_1)\big) \\ &= \sum_{y_1\in\mathcal{Y}_1} P(Y_2=y-y_1)P(Y_1=y_1) \\ &= \sum_{y_1\in\mathcal{Y}_1^*} P(Y_2=y-y_1)P(Y_1=y_1) \end{aligned}$$

The important components of this calculation are:

- the partition of the event $Y = y$ according to possible values of Y_1
- the theorem of total probability
- the independence of Y_1 and Y_2

where

$$\mathcal{Y}_1^* = \{y_1 : y_1 \in \mathcal{Y}_1 \text{ and } y - y_1 \in \mathcal{Y}_2\}$$

Theorem 3.2. If Y_1 and Y_2 are independent Poisson variables with parameters λ_1 and λ_2, then $Y = Y_1 + Y_2$ also has a Poisson distribution, with parameter $\lambda_1 + \lambda_2$.

Proof.

$$\begin{aligned} &P(Y=y) \\ &= \sum_{y_1=0}^{y-y_1} P(Y_1=y_1)P(Y_2=y-y_2) \\ &= \sum_{y_1=0}^{y-y_1} \frac{\lambda_1^{y_1}e^{-\lambda_1}}{y_1!}\frac{\lambda_2^{y-y_1}e^{-\lambda_2}}{(y-y_1)!} \\ &= e^{-(\lambda_1+\lambda_2)}\lambda_2^y \sum_{y_1=0}^{y-y_1}\left(\frac{\lambda_1}{\lambda_2}\right)^{y_1}\frac{1}{y_1!(y-y_1)!} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}\lambda_2^y}{y!}\left(1+\frac{\lambda_1}{\lambda_2}\right)^y \\ &= \frac{e^{-(\lambda_1+\lambda_2)}(\lambda_1+\lambda_2)^y}{y!} \quad y=0,1,2,\dots \end{aligned}$$

Poisson dist. with param. $\lambda_1 + \lambda_2$. □

Continuous Random Variables

If Y_1 and Y_2 are independent continuous random variables, then the sum random variable $Y = Y_1 + Y_2$ is also continuous and we aim to compute $F_Y(y) = P(Y \le y)$. By definition, with $f_Y(y)$ pdf of Y:

$$F_Y(y) = P(Y \le y) = \int_{-\infty}^{y} f_Y(t)dt \text{ where } (Y \le y) = (Y_1 + Y_2 \le y)$$

From the discrete case above, we analogize to:

Definition 3.1. Convolution formula for the cdf of $Y = Y_1 + Y_2$:

$$f_y(y) = \int_{-\infty}^{\infty} f_{Y_2}(y-y_1)f_{Y_1}(y_1)dy_1$$

where $f_Y(.)$, $f_{Y_1}(.)$ and $f_{Y_2}(.)$ are the pdf's for Y, Y_1 and Y_2 resp.

We can extend this result to more independent r.v.'s using the convolution formula recusively, e.g.

$$Y = Y_1 + Y_2 + Y_3 = (Y_1 + Y_2) + Y_3$$

Proof? .-.

4 Multivariate Distributions

We now consider probability specifications for more multiple random variables where there is not necessarily independence. We consider the vector random variable $(Y_1, Y_2, \ldots, Y_n)$.

4.1 Discrete joint distributions

Definition 4.1. For $Y_1 \ldots, Y_n$ be *discrete* random variables on $\mathbb{Z}$. The **joint (discrete bivariate for $n = 2$) probability mass function** is

$$p(y_1, \ldots, y_n) = P\left(\bigcap_{i=1}^{n}(Y_i = y_i)\right) \equiv P(Y_1 = y_1, \ldots, Y_n = y_n)$$

We can think of $p(y_1, y_2)$ as specifying probabilities in a table with Y_1 and Y_2 values in columns / rows.

We can find the joint pmf for various situations using combinatorial arguments.

Theorem 4.1. The joint pmf has the basic properties:

$$0 \leq p(y_1, \ldots, y_n) \leq 1 \quad \forall y_i \; i \in [1, n]$$

$$\sum_{i=1}^{n} \sum_{y_i=-\infty}^{\infty} p(y_1, \ldots, y_n) = \sum_{y_1=-\infty}^{\infty} \cdots \sum_{y_n=-\infty}^{\infty} p(y_1, \ldots, y_n) = 1$$

4.2 Marginal CDFs and PDFs

Definition 4.2. Given a joint pmf for Y_1 and Y_2: $P_{Y_1,Y_2}(.,.)$, the **marginal probability mass function** for Y_i is merely the pmf P_{Y_i} of Y_i, which can be found via the Theorem of Total Probability (1.12).

$$p_{Y_1}(y_1) = P(Y_1 = y_1) = \sum_{y_2=-\infty}^{\infty} p_{Y_1,Y_2}(y_1, y_2)$$

If $p_{Y_1,Y_2}(y_1, y_2)$ specfies the values in a probability table, we compute the marginal pmf for Y_1 and Y_2 by summing respectively across the rows and down the columns of the table.

		Y_2			
		1	2	3	$p_{Y_1}(.)$
Y_1	1				$p_{Y_1}(2)$
	2				
	3				
$p_{Y_2}(.)$		$p_{Y_2}(2)$			1

Definition 4.3. The **conditional mass function** for Y_2 given $Y_1 = y_1$ is

$$p_{Y_2 \mid Y_1}(y_1, y_2) = P(Y_2 = y_2 \mid Y_1 = y_1) = \frac{P_{Y_1,Y_2}(y_1, y_2)}{P_{Y_1}(y_1)}$$

Recall the fundamental result

$$p_{Y_1,Y_2}(y_1, y_2) = p_{Y_1}(y_1) p_{Y_2 \mid Y_1}(y_2 \mid y_1) \text{ when } p_{Y_1}(y_1) > 0$$

Definition 4.4. Discrete r.v.'s $Y_1, \ldots, Y_n$ are **independent** if

$$p_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n) = \prod_{i=1}^{n} p_{Y_i}(y_i)$$

$$\text{i.e. } p_{Y_i \mid Y_j}(y_i \mid y_j) = p_{Y_i} \; \forall i, j \in [1, n]$$

Definition 4.5. The **joint cumulative distribution function** cdf $F_{Y_1,Y_2}(y_1,y_2)$ and **joint probability density function** pdf $f_{Y_1,Y_2}(y_1,y_2)$

$$\begin{aligned} F_{Y_1,Y_2}(y_1,y_2) &= P\left((Y_1 \le y_1) \cap (Y_2 \le y_2)\right) \quad \forall (y_1 \in \mathbb{R},\ y_2 \in \mathbb{R}) \\ &= \int_{-\infty}^{y_1}\int_{-\infty}^{y_2} f_{Y_1,Y_2}(t_1,t_2)dt_2dt_1 \end{aligned}$$

Theorem 4.2. The joint cdf has the following properties:

$$\lim_{y_1\to-\infty}\lim_{y_2\to-\infty} F_{Y_1,Y_2}(y_1,y_2) = 0 \qquad \lim_{y_1\to\infty}\lim_{y_2\to\infty} F_{Y_1,Y_2}(y_1,y_2) = 1$$

$$F_{Y_1,Y_2}(y_1,y_2) \le F_{Y_1,Y_2}(y_1+c,y_2) \text{ and } F_{Y_1,Y_2}(y_1,y_2+c)\ \forall y_1,y_2,\ c>0$$

$$\lim_{y_1\to\infty} F_{Y_1,Y_2}(y_1,y_2) = F_{Y_2}(y_2) \qquad \lim_{y_2\to\infty} F_{Y_1,Y_2}(y_1,y_2) = F_{Y_1}(y_1)$$

Theorem 4.3. By the probability axioms, the joint pdf

- is non-negative: $f_{Y_1,Y_2}(y_1,y_2) \ge 0$
- integrates to 1: $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{Y_1,Y_2}(y_1,y_2)dy_2dy_1 = 1$

To compute $f_{Y_1,Y_2}(y_1,y_2)$ from $F_{Y_1,Y_2}(y_1,y_2)$, use

$$f_{Y_1,Y_2}(y_1,y_2) = \frac{\partial^2}{\partial y_1 \partial y_2}\left(F_{Y_1,Y_2}(y_1,y_2)\right)$$

where the partial differentiations are commutative given the second derivatives are continuous at that point.

Definition 4.6. The **marginal probability density function** pdf $f_{Y_i}(y_i)$ for each fixed y_i, $i \in \{1,2\}$ is

$$f_{Y_i}(y_i) = \int_{-\infty}^{\infty} f_{Y_1,Y_2}(y_1,y_2)dy_2$$

Definition 4.7. The **conditional probability density function** pdf is

$$f_{Y_2\,|\,Y_1}(y_2\,|\,y_1) = \frac{f_{Y_1,Y_2}(y_1,y_2)}{f_{Y_1}(y_1)}$$

The conditional pdf $f_{Y_1,Y_2}(y_1,y_2)$ is a pdf in y_2 for every fixed y_1, so it is non-negative and integrates to 1 over y_2.

As in the discrete case, we have the CHAIN RULE FACTORIZATION

$$f_{Y_1,Y_2}(y_1,y_2) = f_{Y_1}(y_1)f_{Y_2\,|\,Y_1}(y_2\,|\,y_1) = f_{Y_2}(y_2)f_{Y_1\,|\,Y_2}(y_1\,|\,y_2)$$

4.3 *Independence and Correlation*

Definition 4.8. Continuous r.v.'s $(Y_1,\ldots,Y_n)$ are **independent** if

$$f_{Y_1,\ldots,Y_n}(y_1,\ldots,y_n) = \prod_{i=1}^{n} f_{Y_i}(y_i)$$

In the bivariate case, we have

$$f_{Y_2\,|\,Y_1}(y_2\,|\,y_1) = f_{Y_2}(y_2)$$
$$f_{Y_1\,|\,Y_2}(y_1\,|\,y_2) = f_{Y_1}(y_1)$$

Remark 25. An easy way to prove that two r.v.'s are not independent is to check if $\mathcal{Y}_{12} = \mathcal{Y}_1 \times \mathcal{Y}_2$ where $\mathcal{Y}_{12}$ is the support of the joint pdf.

Extending the idea of expectation to multiple variables $Y_1, \dots, Y_n$, we get the following, respectively discrete and continuous, expressions

$$\mathbb{E}\left[g(y_1, \dots, y_n)\right] = \sum_{i=-\infty}^{\infty} \cdots \sum_{i=-\infty}^{\infty} g(y_1, \dots, y_n) f_{Y_1,\dots,Y_n}(y_1, \dots, y_n)$$

$$\mathbb{E}\left[g(y_1, \dots, y_n)\right] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(y_1, \dots, y_n) f_{Y_1,\dots,Y_n}(y_1, \dots, y_n) dy_1 \cdots dy_n$$

and linearity of expectation (Theorem 2.2) still holds.

Definition 4.9. The **covariance** $\operatorname{Cov}[Y_1, Y_2]$ between Y_1 and Y_2 with expectations μ_1 and μ_2 is

$$\operatorname{Cov}[Y_1, Y_2] = \mathbb{E}\left[(Y_1 - \mu_1)(Y_2 - \mu_2)\right] = \mathbb{E}\left[Y_1 Y_2\right] - \mu_1 \mu_2$$

For multiple variables, we consider pairwise covariance between all of them and put the results into a matrix called covariance matrix.

Definition 4.10. The **correlation** $\operatorname{Corr}[Y_1, Y_2]$ between Y_1 and Y_2 is

$$\operatorname{Corr}[Y_1, Y_2] = \frac{\operatorname{Cov}[Y_1, Y_2]}{\sqrt{\mathbb{V}\left[Y_1\right]\mathbb{V}\left[Y_2\right]}}$$

Theorem 4.4. $-1 \leq \operatorname{Corr}[Y_1, Y_2] \leq 1$

Proof. in textbook. □

Definition 4.11. Y_1, Y_2 **uncorrelated** if $\operatorname{Cov}[Y_1, Y_2] = \operatorname{Corr}[Y_1, Y_2] = 0$.

Theorem 4.5. For Y_1 and Y_2 independent,

$$\mathbb{E}\left[g_1(Y_1) g_2(Y_2)\right] = \mathbb{E}\left[g_1(Y_1)\right]\mathbb{E}\left[g_2(Y_2)\right]$$

Proof. *(Theorem 4.5)*

$$\begin{aligned}
&\mathbb{E}\left[g_1(Y_1) g_2(Y_2)\right] \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g_1(y_1) g_2(y_2) f_{Y_1,Y_2}(y_1, y_2) dy_1 dy_2 \\
&= \int_{-\infty}^{\infty} g_1(y_1) f_{Y_1}(y_1) dy_1 \int_{-\infty}^{\infty} g_2(y_2) f_{Y_2}(y_2) dy_1 \\
&= \mathbb{E}\left[g_1(Y_1)\right]\mathbb{E}\left[g_2(Y_2)\right]
\end{aligned}$$ □

Theorem 4.6. If Y_1 and Y_2 independent then they are uncorrelated.

Proof. *(Theorem 4.6)*

$$\begin{aligned}
\operatorname{Cov}[Y_1, Y_2] &= \mathbb{E}\left[Y_1 Y_2\right] - \mu_1 \mu_2 \\
&= \mathbb{E}\left[Y_1\right]\mathbb{E}\left[Y_2\right] - \mu_1 \mu_2 \\
&= 0
\end{aligned}$$ □

Warning 26. Uncorrelation doesn't imply independence!

4.4 *Transformations of random variables*

Linear Combinations

Let $Y_1, \dots, Y_n$ with $\mathbb{E}\left[Y_i\right] = \mu_i$ and $X_1, \dots, X_m$ with $\mathbb{E}\left[X_i\right] = \xi_i$, define $U_1 = \sum_{i=1}^{n} a_i Y_i$ and $U_2 = \sum_{j=1}^{m} b_j X_j$. We get the following:

Theorem 4.7.

$$\mathbb{E}\left[U_1\right] = \sum_{i=1}^{n} a_i \mathbb{E}\left[Y_i\right] = \sum_{i=1}^{n} a_i \mu_i$$

$$\mathbb{V}\left[U_1\right] = \sum_{i=1}^{n} a_i^2 \mathbb{V}\left[Y_i\right] + 2\sum_{j=2}^{n}\sum_{i=1}^{j-1} a_i a_j \operatorname{Cov}\left[Y_i, Y_j\right]$$

$$\operatorname{Cov}\left[U_1, U_2\right] = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j \operatorname{Cov}\left[Y_i, X_j\right]$$

Back to sums, consider $Y = Y_1 + Y_2$ and recall the convolution formula (Definition 3.1). An easier method than computing convolutions is to recall the moment generating function for $U = g(Y)$ (Definition 2.18) $m_U(t) = \mathbb{E}\left[e^{tU}\right] = \mathbb{E}\left[e^{tg(t)}\right]$. If Y_1 and Y_2 are independent and $Y = a_1 g_1(Y_1) + a_2 g_2(Y_2)$:

$$m_Y(t) = \mathbb{E}\left[e^{t(a_1 g_1(Y_1) + a_2 g_2(Y_2))}\right] = \mathbb{E}\left[e^{t a_1 g_1(Y_1)}\right]\mathbb{E}\left[e^{t a_2 g_2(Y_2)}\right]$$

Theorem 4.8. For $Y = Y_1 + Y_2$ where Y_1, Y_2 independent,

$$m_Y(t) = m_{Y_1}(t) m_{Y_2}(t)$$

Proof. $m_Y(t) = \mathbb{E}\left[e^{t(Y_1+Y_2)}\right] = \mathbb{E}\left[e^{tY_1}\right]\mathbb{E}\left[e^{tY_2}\right] = m_{Y_1}(t) m_{Y_2}(t)$ □

General Transformations

Approaching from a 'change of variable' perspective:

Theorem 4.9. Consider the bijective transformation

$$(Y_1, Y_2) \longrightarrow (U_1, U_2)$$

where we can write

$$Y_1 = h_1^{-1}(U_1, U_2) \text{ and } Y_2 = h_2^{-1}(U_1, U_2)$$

So for $(u_1, u_2) \in \mathbb{R}^2$,

$$f_{U_1,U_2}(u_1, u_2) = f_{Y_1,Y_2}(h_1^{-1}(u_1, u_2), h_2^{-1}(u_1, u_2))|J(u_1, u_2)|$$

$$\text{where } J(u_1, u_2) = \det \begin{bmatrix} \frac{\partial h_1^{-1}(u_1,u_2)}{\partial u_1} & \frac{\partial h_1^{-1}(u_1,u_2)}{\partial u_2} \\ \frac{\partial h_2^{-1}(u_1,u_2)}{\partial u_1} & \frac{\partial h_2^{-1}(u_1,u_2)}{\partial u_2} \end{bmatrix} \text{ is the Jacobian.}$$

Proof. c.f. textbook. □

Example 4.1 (Sum of Binomial r.v.'s)**.** For Y_1, Y_2 independent Binomial r.v.'s with parameters (n_1, p) and (n_2, p) we have for $Y = Y_1 + Y_2$

$$\begin{aligned} m_Y(t) &= m_{Y_1}(t) m_{Y_2}(t) \\ &= (pe^t + q)^{n_1}(pe^t + q)^{n_2} \\ &= (pe^t + q)^{n_1+n_2} \end{aligned}$$

Notice this is the mgf of the Binomial distribution $(n_1 + n_2, p)$.

Example 4.2 (Sum of Poisson r.v.'s)**.** Similarly to the previous example, for Y_1 and Y_2 independent Poisson r.v.'s with parameters λ_1 and λ_2 to find

$$m_Y(t) = \exp\left((\lambda_1 + \lambda_2)(e^t - 1)\right)$$

the mgf of the Poisson distribution with parameter $\lambda_1 + \lambda_2$.

Example 4.3 (Sum of Normal r.v.'s)**.** Similarly for Y_1 and Y_2 independent Normal r.v.'s with parameters (μ_1, σ_1^2) and (μ_2, σ_2^2), we get

$$m_Y(t) = \exp\left((\mu_1 + \mu_2)t + (\sigma_1^2 + \sigma_2^2)t^2/2\right)$$

the mgf of the Normal distribution with parameters $(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

5 Probability Inequalities and Theorems

5.1 Probability Inequalities

Theorem 5.1 (Markov's Inequality)**.** Suppose $g(y)$ function of a continuous r.v. Y such that $\mathbb{E}\big[|g(Y)|\big] < \infty$. For any $k > 0$ let

$$\mathcal{A}_k = \{y \in \mathbb{R} : |g(y)| \geq k\}$$

then we have

$$P\big(|g(Y)| \geq k\big) \leq \frac{\mathbb{E}\big[|g(Y)|\big]}{k}$$

Markov's Inequality holds for discrete r.v.'s as well (just replace all the integrals with sums in the proof).

It can be written equivalently as

$$P(|g(Y)| < k) \geq 1 - \frac{\mathbb{E}\big[|g(Y)|\big]}{k}$$

Proof. We want to prove that $\mathbb{E}\big[|g(Y)|\big] \geq kP\big(|g(Y)| \geq k\big)$

$$\begin{aligned}\mathbb{E}\big[|g(Y)|\big] &= \int_{-\infty}^{\infty} |g(y)| f_Y(y)dy = \int_{A_k} |g(y)| f_Y(y)dy + \int_{A_k'} |g(y)| f_Y(y)dy \\ &\geq \int_{A_k} |g(y)| f_Y(y)dy \geq \int_{A_k} k f_Y(y)dy = kP\big(|g(Y)| \geq k\big)\end{aligned}$$

□

Theorem 5.2 (Chebychev's Inequality)**.** For any distribution with finite mean μ and variance σ^2, we have that $\forall c > 0$,

$$P\left(\mu - c\sigma < Y < \mu + c\sigma\right) \geq 1 - \frac{1}{c^2}$$

Theorem 5.2 bounds how much probability can be present in the tails of the distribution.

It can be written equivalently as

$$P\big(|Y - \mu| \geq c\sigma\big) \leq \frac{1}{c^2}$$

Proof. Let $g(t) = (t-\mu)^2$ in Theorem 5.1:

$$P\big((t-\mu)^2 < k\big) \geq 1 - \frac{\mathbb{E}\big[(t-\mu)^2\big]}{k}$$

Recall $\mathbb{E}\big[(Y-\mu)^2\big] = \sigma^2$ so choose $k = c^2\sigma^2$

$$P\big((t-\mu)^2 < c^2\sigma^2\big) = P\left(\sqrt{(t-\mu)^2} < \sqrt{c^2\sigma^2}\right) = P\big(|Y-\mu| < c\sigma\big)$$

$$P\big(|Y-\mu| < c\sigma\big) \geq 1 - \frac{\mathbb{E}\big[(t-\mu)^2\big]}{c^2\sigma^2} = 1 - \frac{1}{c^2}$$

□

Theorem 5.3 (Sample mean)**.** Suppose $Y_1, \ldots, Y_n$ independent r.v.'s that have the same distribution, with expectation μ and variance σ^2. Let $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$ be the **sample mean** r.v. Then $\forall \varepsilon > 0$,

$$P\left(|\overline{Y}_n - \mu| \geq \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2}$$

Proof. Recall that for $U = \sum_{i=1}^n Y_i$ we have $\mathbb{E}[U] = n\mu$ and $\mathbb{V}\big[n\sigma^2\big]$.

$$\mathbb{E}\big[\overline{Y}_n\big] = \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^n Y_i\right] = \frac{n\mu}{n} = \mu$$

$$\mathbb{V}\big[\overline{Y}_n\big] = \frac{1}{n^2}\mathbb{V}\left[\sum_{i=1}^n Y_i\right] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

$$P\left(|\overline{Y}_n - \mu| \geq \frac{c\sigma}{\sqrt{n}}\right) \leq \frac{1}{c^2} \quad \forall c$$

Let $\varepsilon = \frac{c\sigma}{\sqrt{n}}$ such that $c = \frac{\sqrt{n}\varepsilon}{\sigma}$

$$\implies P\big(|\overline{Y}_n - \mu| \geq \varepsilon\big) \leq \frac{\sigma^2}{n\varepsilon^2}$$

where ε arbitrary. □

5.2 Convergence and Theorems

Definition 5.1. A sequence of r.v.'s $Y_1, Y_2, \ldots, Y_n, \ldots$ **converges in probability** to $c \in \mathbb{R}$ as $n \to \infty$, written as $Y_n \xrightarrow{p} c$, if $\forall \varepsilon > 0$,

$$\lim_{n\to\infty} P\left(|Y_n - c| < \varepsilon\right) = 1$$

Theorem 5.4 (Weak Law of Large Numbers)**.** Suppose $Y_1, \ldots, Y_n$ iid with expectation μ and variance σ^2. Then the sample mean $\overline{Y}_n$ as defined in Therorem 5.3 converges in probability to μ:

$$\overline{Y}_n \xrightarrow{p} \mu \quad (\text{as } n \to \infty)$$

In other words, the sample mean r.v. converges in probability to the expectation of the distribution of the Ys.

Remark 27. The n term in the denominator of Theorem 5.3 tells us that the rate of convergence is at least n.

Proof. follows directly from Chebychev (Theorem 5.2).

$$\begin{aligned} & P\left(|\overline{Y}_n - \mu| \geq \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2} \\ \Leftrightarrow\ & P\left(|\overline{Y}_n - \mu| < \varepsilon\right) \geq 1 - \frac{\sigma^2}{n\varepsilon^2}\ \forall \varepsilon > 0 \\ & \longrightarrow 1 \text{ as } n \to \infty \\ \implies & \lim_{n\to\infty} P\left(|\overline{Y}_n - \mu| < \varepsilon\right) \geq 1 \end{aligned}$$

So $\overline{Y}_n \xrightarrow{p} \mu$ □

Definition 5.2. Consider a sequence of cdfs $F_1(y), F_2(y), \ldots, F_n(y), \ldots$ such that $\lim_{n\to\infty} F_n(y) = F(y)$ at all points where $F(y)$ is continuous. Then $F(y)$ is the **limiting distribution** for the sequence.

Definition 5.3. If a sequence of r.v.'s $Y_1, Y_2, \ldots, Y_n, \ldots$ has respective cdfs $F_1(y), F_2(y), \ldots, F_n(y), \ldots$ and Y has cdf $F(y)$, this sequence **converges in distribution** to Y as $n \to \infty$, written as $Y_n \xrightarrow{d} Y$.

Theorem 5.5 (Central Limit Theorem)**.** Suppose $Y_1, \ldots, Y_n$ iid with expectation μ and variance σ^2. Consider the sample mean $\overline{Y}_n$ as defined in Therorem 5.3. The random variable $U_n = \frac{\sqrt{n}(\overline{Y}_n - \mu)}{\sigma}$ converges in distribution to a standard normal distribution $Normal(0,1)$

$$\lim_{n\to\infty} F_{U_n}(u) = \lim_{n\to\infty} P(U_n \leq u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$$

Recalling $Y_n = \frac{1}{n}\sum_{i=1}^n Y_i$

$$\begin{aligned} U_n &= \frac{\overline{Y}_n - \mu}{\sigma/\sqrt{n}} \\ &= \frac{1}{\sigma\sqrt{n}}\left(\sum_{i=1}^{n} Y_i - n\mu\right) \end{aligned}$$

In other words, irrespective of the actual distribution of $Y_1, \ldots, Y_n$, providing the conditions on expectation and variance are met, the distributions $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$ and $S_n = \sum_{i=1}^n Y_i$ can be approximated using a Normal distribution.

$$\begin{aligned} \overline{Y}_n &\approx Normal(\mu, \sigma^2/n) \\ S_n &\approx Normal(n\mu, n\sigma^2) \end{aligned}$$

Proof. Let $Z_i = \frac{Y_i - \mu}{\sigma}$ for $i = 1, 2, \ldots, n$. Then $\mathbb{E}[Z_i] = 0,\ \mathbb{V}[Z_i] = 1$.

Notice that $U_n = \dfrac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i$

so use $m_{U_n}(t) = \left(m_Z\left(t/\sqrt{n}\right)\right)^n$ Taylor expanding $m_Z(t)$ at $t = 0$:

$$\begin{aligned} m_Z(t) &= m_Z(0) + t m_Z^{(1)}(0) + \frac{t^2}{2} m_Z^{(2)}(t_0) \quad 0 < t_0 < t \\ &= \frac{t^2}{2} m_Z^{(2)}(t_0) \\ \implies m_{U_n}(t) &= \left(1 + \frac{t^2}{2n} m_Z^{(2)}(t_0)\right)^n \quad t_0 \in \left[0, \frac{t}{\sqrt{n}}\right] \to 0 \text{ as } n \to \infty \\ \lim_{n\to\infty} m_{U_n}(t) &= \lim_{n\to\infty} \left(1 + \frac{t^2}{2n} m_Z^{(2)}(0)\right)^n \text{ recall } \mathbb{E}[Z_i^2] = m_Z^{(2)}(0) = 1 \\ &= \lim_{n\to\infty} \left(1 + \frac{t^2}{2n}\right) = e^{t^2/2} \text{ by def of exponential} \\ \implies U_n &\xrightarrow{d} Normal(0,1) \end{aligned}$$

□

Recall that for $X = \sigma Y + \mu$,

$$m_X(t) = e^{\mu t} m_Y(\sigma t)$$

$$\begin{aligned} & \mathbb{E}[Z_i] = 0 \text{ and } \mathbb{V}[Z_i] = 1 \\ \implies\ & m_Z(0) = 1,\ m_Z^{(1)}(0) = \mathbb{E}[Z_i] = 0 \end{aligned}$$