STAT 134: Concepts of Probability

—Final Review Guide—

Disclaimer: This study guide was created by the Statistics Undergraduate Student Association, which is in no way affiliated with the current course staff of Statistics 134. Additionally, this guide is designed merely to *supplement* to your class notes.

Last Updated: Fall 2020

Originally compiled by: Ethan M.

Contents

1 Basics of Probability

- A quick review of sets and set theory may be useful:
  - A **set** is a collection of unordered **elements**. Elements do not need to be numbers; for example, {Blue, Gold} is the set of official Berkeley colors (go bears!)
  - The **union** of two sets is the set containing all the elements of each set, and the **intersection** of two sets is the set containing elements common to both sets. For example, if $A = \{1, 2, 3\}$ and $B = \{2, 3, 4\}$ then $A \cup B = \{1, 2, 3, 4\}$ and $A \cap B = \{2, 3\}$.
  - The **empty set** (denoted $\varnothing$) is the set containing no elements. Two sets are said to be **mutually exclusive** (or **disjoint**) if $A \cap B = \varnothing$.
  - A **subset** A of a set B is a set containing some (possibly all) of the elements in B. For example, $\{2, 4\} \subseteq \{1, 2, 3, 4\}$. Two sets A and B are said to be **equal** if $A \subseteq B$ and $B \subseteq A$.
  - Here is a summary of some set-related concepts:

Union	$A \cup B$	$:= \{x : x \in A \text{ or } x \in B\}$
Intersection:	$A \cap B$	$:= \{x : x \in A \text{ and } x \in B\}$
Difference:	$A \backslash B$	$:= \{x : x \in A \text{ and } x \notin B\}$
Subset:	$A \subseteq B$	$x \in A \implies x \in B$
Equality:	$A = B$	$A \subseteq B$ and $B \subseteq A$
Proper Subset:	$A \subset B$	$A \subseteq B$ and $A \neq B$

- The **outcome space** (denoted Ω) is the set containing all possible outcomes of a particular setup. **Events** are simply subsets of the outcome space.
  - If all events $A \subseteq \Omega$ are **equally likely**, we define the **probability of the event** A to be

$$\mathbb{P}(A) = \frac{\#(A)}{\#(\Omega)}$$

    Here $\#(\cdot)$ denotes the number of elements in a set.
  - A set of pairwise disjoint events $\{B_1, \ldots, B_n\}$ (that is, $B_i \cap B_j = \varnothing$ for any $i \neq j$) is said to **partition** the event B if

$$\bigcup_{i=1}^{n} B_i = B_1 \cup \cdots \cup B_n = B$$

- The **three axioms of probability** state

  (a) $\mathbb{P}(A) \geq 0$ for any $A \subseteq \Omega$

  (b) $\mathbb{P}(\Omega) = 1$

  (c) For mutually exclusive events A and B, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

  - Define the **complement** of an event A to be the unique event $\overline{A}$ (sometimes notated $A^{\complement}$) such that $\{A, \overline{A}\}$ partitions the outcome space Ω. Then, by axioms (b) and (c), we have that

$$\mathbb{P}(\overline{A}) = 1 - \mathbb{P}(A)$$

- The **inclusion-exclusion rule** provides a way to compute the probability of the union of events, even if the events are not mutually exclusive. For 2 events A and B, we have

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

  More generally, for n events $A_1, \ldots, A_n$ we have

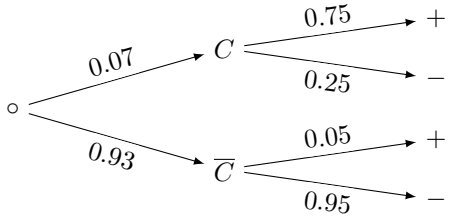
$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j) + \sum_{i<j<k} \mathbb{P}(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n+1} \mathbb{P}\left(\bigcap_{i=1}^{n} A_i\right)$$

- **Conditional probabilities** are probabilistic quantities that reflect some change to the outcome space.

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

  The **multiplication rule** states that $\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\mathbb{P}(B)$

  - Two events A and B are said to be **independent** (notated $A \perp B$) if $\mathbb{P}(A \mid B) = \mathbb{P}(B)$. Alternatively, $A \perp B$ if and only if $\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$.
  - **Probability Trees** can be useful in keeping track of conditional probabilities.
    * For example, suppose 7% of a population has a disease. Of those who have the disease, a test correctly identifies them as disease-positive 75% of the time. Of those who do not have the disease, the test correctly identifies them as disease-negative 95% of the time. The tree for this situation would be as follows:



      Here, C denotes the event {person is actually a carrier}, $+$ denotes the event {the test tests positive}, and $-$ denotes the event {the test tests negative}.

- The **Rule of Average Conditional Probabilities** (also known as the **Law of Total Probability**) states that, for a partition $\{B_1, \ldots, B_n\}$ of the outcome space Ω,

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A \mid B_i)\mathbb{P}(B_i) = \mathbb{P}(A \mid B_1)\mathbb{P}(B_1) + \cdots + \mathbb{P}(A \mid B_n)\mathbb{P}(B_n)$$

  That is, the probability of any event A can be computed as a weighted average of the probabilities of each event in a partition of Ω.

- **Bayes' Rule** provides another tool for evaluating conditional probabilities:

$$\begin{aligned}\mathbb{P}(B \mid A) &= \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A \mid B)\mathbb{P}(B)}{\mathbb{P}(A)} \\ &= \frac{\mathbb{P}(A \mid B)\mathbb{P}(B)}{\sum_{i=1}^{n} \mathbb{P}(A \mid B_i)\mathbb{P}(B_i)}\end{aligned}$$

  where $\{B_1, \ldots, B_n\}$ is a partition of Ω.

2 Random Variables and Distributions

- A **random variable** can be thought of as a measure of some random process. For example, if X denotes the number of heads in 2 tosses of a fair coin, then X is a random variable. The key idea is that X can take on different values, each with different probabilities.
- The **support** of a random variable is the set of all values the random variable is allowed to attain. For example, in the coin-tossing example above, X can be either 0, 1, or 2; it is impossible to toss 2 coins and observe more than 2 heads (or negative heads, for that matter).
- A **p.m.f.** (probability mass function) is an enumeration of the values of $\mathbb{P}(X = k)$ where X is a random variable and k is a value within the support of X. For instance, in the coin-tossing example:

$\boldsymbol{k}$	0	1	2
$\mathbb{P}(\boldsymbol{X = k})$	$(1/2)^2$	$(1/2)$	$(1/2)^2$

  The key to constructing tables (like the one above) is to translate each event into words. For example, $\{X = 2\}$ means "I toss two heads in two tosses of a fair coin." In this wording, it is clearer how to compute the associated probability.

  - The table above can be equivalently expressed as

$$\mathbb{P}(X = k) = \begin{cases} \binom{2}{k} (1/2)^k & \text{if } k = 0, 1, 2 \\ 0 & \text{otherwise} \end{cases}$$

  - The **cumulative mass function** (CMF; notated $F_X(x)$) is defined to be $\mathbb{P}(X \leq x)$; the **survival** (sometimes notated $\overline{F_X}(x)$) is defined to be $\mathbb{P}(X > x)$.

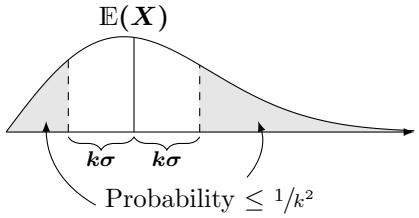
- A **joint PMF** quantifies the probabilities associated with two related random variables, and is denoted $\mathbb{P}(X = x, Y = y)$.

  - Random variables X and Y are said to be **independent** (denoted $X \perp Y$) if $\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y)$.
  - A series of random variables $X_1, \ldots, X_n$ are said to be **pairwise-independent** if $X_i \perp X_j$ for $i \neq j$. Note that pairwise independence does *not* imply independence, whereas independence *does* imply pairwise independence.
  - The **discrete convolution** provides a way of identifying the PMF of a sum of two random variables:

$$\mathbb{P}(X + Y = s) = \sum_{k=0}^{s} \mathbb{P}(X = k, Y = s - k)$$

- The **expected value** (or **expectation**) of a random variable is a measure of central tendency, and is defined to be
$$\mathbb{E}(X) := \sum_{k \in \text{ support}} k \cdot \mathbb{P}(X = k)$$
The **variance** of a random variable is a measure of how "wide" a distribution is, and is defined to be
$$\text{Var}(X) := \mathbb{E}\left\{[X - \mathbb{E}(X)]^2\right\} = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2$$
The **standard deviation** is simply the square-root of variance: $\text{SD}(X) := \sqrt{\text{Var}(X)}$.
  - Expectation is linear: $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$. Variance is not: $\text{Var}(aX + b) = a^2\text{Var}(X)$.
  - The expectation of a function of a random variable is given by the **Law of the Unconscious Statistician** (or LOTUS):
$$\mathbb{E}[g(X)] = \sum_{k \in \text{ support}} g(k)\mathbb{P}(X = k)$$
A proof may be found in **Appendix 13.1**.
  - For independent events $X_1, \ldots, X_n$, we have
$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i)$$
If the events are not independent, the formula becomes a bit more complicated and requires material from chapter 6.

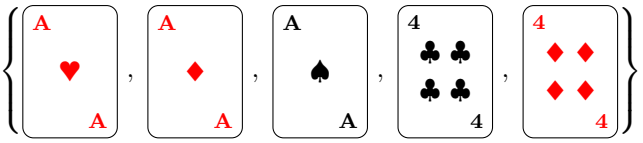
- There are two inequalities which can be used to identify an upper bound of probabilities *without* any knowledge of the underlying distribution:
  - **Markov's Inequality:** $\mathbb{P}(X \geq a) \leq \dfrac{\mathbb{E}(X)}{a}$ if $X \geq 0$, and if $a > 0$.
  - **Chebyshev's Inequality:** $\mathbb{P}\left[|X - \mathbb{E}(x)| \geq k \cdot \text{SD}(X)\right] \leq \dfrac{1}{k^2}$, for $k > 0$, and provided that the support of X contains only nonnegative numbers.



- An **indicator random variable** is a random variable defined as
$$\mathbb{1}_A = \begin{cases} 1 & \text{if } A \text{ occurs} \\ 0 & \text{If } A \text{ does not occur} \end{cases}$$
In this way, $\mathbb{P}(\mathbb{1}_A = 1) = \mathbb{E}(\mathbb{1}_A) = \mathbb{P}(A \text{ occurs})$.
  - Indicators are particularly useful in measuring counts. For example, let X denote the number of heads in 10 tosses of a $p-$coin. Then
$$X = \sum_{i=1}^{n} \mathbb{1}_{T_i} \quad \text{where} \quad \mathbb{1}_{T_k} = \begin{cases} 1 & \text{if } i\text{th toss lands heads} \\ 0 & \text{if } i\text{th toss lands tails} \end{cases}$$
  - More abstractly, say $X = \mathbb{1}_A + \mathbb{1}_B + \mathbb{1}_C + \mathbb{1}_D$. Further suppose that events A and C have occurred, whereas B and D have not. Then $\mathbb{1}_A = \mathbb{1}_C = 1$ and $\mathbb{1}_B = \mathbb{1}_D = 0$, so $X = 1 + 0 + 1 + 0 = 2$, which is precisely the number of events that have occurred.

3 Counting and Combinatorics

- Suppose we wish to pick k objects from a total of n objects. For illustrative purposes, say we wish to pick 3 letters from the set of $n = 5$ letters $\{a, b, c, d, e\}$.
  - If order matters (i.e. $\{a, b, c\}$ is not considered the same thing as $\{b, c, a\}$) then the number of ways to do this is
  $$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$
  - If order does not matter (i.e. $\{a, b, c\}$ is considered the same thing as $\{b, c, a\}$), then the number of ways to do this is
  $$(n)_k = \frac{n!}{(n-k)!} = n \times (n-1) \times \cdots \times (n-k+1)$$
- Always pick like objects together! It may be useful to demonstrate this through example. Given a poker hand of 5 cards drawn from a standard 52-card deck, we wish to compute the number of **full houses**. A full house is defined to be 3 cards of one rank, and 2 cards of another rank. For example,



We first find the number of ways to pick 3 cards from the first rank (in our example above this would be the number of ways to pick 3 aces from the deck): this number is $\binom{4}{3}$. Then we find the number of ways to pick 2 cards from the second rank (in our example above this would be the number of ways to pick 2 four's from the deck): this number is $\binom{4}{2}$.

Finally, we need to count the number of possible ranks we could have chosen for the three-of-a-kind: this is $\binom{13}{1}$. Then, from the remaining 12 ranks we pick one to be the rank of the two-of-a-kind: $\binom{12}{1}$. Putting everything together, the number of full houses is

$$\underbrace{\binom{13}{1}}_{\substack{\text{pick the rank} \\ \text{of the three-of-a-kind}}} \times \underbrace{\binom{4}{3}}_{\substack{\text{pick the cards} \\ \text{in the three-of-a-kind}}} \times \underbrace{\binom{12}{1}}_{\substack{\text{pick the rank} \\ \text{of the two-of-a-kind}}} \times \underbrace{\binom{4}{2}}_{\substack{\text{pick the cards} \\ \text{in the two-of-a-kind}}}$$

4 Approximations to the Binomial Distribution

- The **Standard Normal Distribution** is an example of a *continuous* distribution (continuous distributions will be discussed further after the midterm). The **standard normal distribution** (notated $\mathcal{N}(0, 1)$) has probability density function (the continuous analog of p.m.f's)
$$\phi(z) := \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$
and has cumulative density function (the continuous analog of c.m.f's)
$$\Phi(x) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}x^2} \, \mathrm{d}x$$

The **normal distribution** (notated $\mathcal{N}(\mu, \sigma^2)$) is a nonstandardized version of the standard normal distribution with p.d.f.

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

If $X \sim \mathcal{N}(\mu, \sigma^2)$ then

$$\left(\frac{X-\mu}{\sigma}\right) \sim \mathcal{N}(0,1)$$

- Suppose $X \sim \text{Bin}(n,p)$. If p is not too small and if n is very large, then X is well approximated by the $\mathcal{N}(np, np(1-p))$ distribution.
- When using the normal approximation, it is advised to use the **continuity correction** to account for the fact that we are approximating a discrete random variable with a continuous one. Letting $X \sim \text{Bin}(n,p)$, we have

$$\mathbb{P}(X \leq a) \approx \Phi\left(\frac{[a+0.5]-np}{\sqrt{np(1-p)}}\right)$$

$$\mathbb{P}(X \geq b) = 1 - \mathbb{P}[X \leq (b-1)] \approx 1 - \Phi\left(\frac{[b-0.5]-np}{\sqrt{np(1-p)}}\right)$$

- Quantiles of the normal distribution cannot be obtained analytically; the use of a table (or computing software) is required.

- The **Poisson Distribution** (notated $\text{Pois}(\mu)$) is a discrete distribution with p.m.f.

$$\mathbb{P}(X = x) = e^{-\mu} \cdot \frac{\mu^x}{x!} \qquad x \in \{0, 1, 2, \dots\}$$

- If $X \sim \text{Bin}(n,p)$ and p is very small or very large, then X is **not** well-approximated by a normal distribution and is better approximated by a $\text{Pois}(np)$ distribution.

- **Example:** Consider a coin that lands heads with probability $p = 0.4$. If I toss this coin 100 times and let X denote the number of heads in these 100 tosses, then X is approximately $\mathcal{N}(40, 24)$ and the probability of tossing 30 or less heads is approximately

$$\Phi\left(\frac{30.5-40}{\sqrt{24}}\right) \approx 0.02623975$$

The exact answer, using the binomial distribution directly, is 0.02061342 so we see the error in approximation is quite small.

- **For the Mathematically Curious:** You might ask what we mean when we say that a distribution "approximates" another distribution. This is actually a deeper question that delves into topics relating to **notions of convergence**, and will be discussed further in Stat 135. If you're curious, you can look up the topics of **convergence in distribution** and **convergence in probability**.

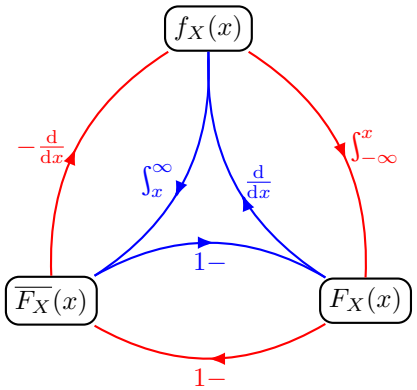
5 Continuous Random Variables, PDF's, and CDF's

- Certain random variables take a continuous set as support; such random variables are said to be **continuous**.

- Continuous random variables are described by their **probability density functions**, which are defined as
$$f_X(x) = \frac{\mathbb{P}(X \in \mathrm{d}x)}{\mathrm{d}x}$$
where $\mathrm{d}x$ represents an infinitesimally small interval. **Note: PDF's do not inherently represent probability!** Rather, probabilities in the context of continuous random variables take the form of areas underneath the PDF.:
$$\mathbb{P}(X \in [a,b]) = \int_a^b f_X(x)\ \mathrm{d}x$$

- The **cumulative distribution function** (or CDF) of a random variable X *does* represent a probability:
$$F_X(x) := \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f_X(t)\ \mathrm{d}t$$
The **survival** of a continuous distribution is
$$\overline{F_X}(x) := \mathbb{P}(X \geq x) = \int_{x}^{\infty} f_X(t)\ \mathrm{d}t$$
Note that $\mathbb{P}(X = x) = 0$ if X is continuous.

  - We can obtain a relation between CDF's and PDF's:
$$f_X(x) = \frac{\mathrm{d}}{\mathrm{d}x}\left[F_X(x)\right]$$
This can be proven using the Fundamental Theorem of Calculus:
$$\frac{\mathrm{d}}{\mathrm{d}x}\left[F_X(x)\right] = \frac{\mathrm{d}}{\mathrm{d}x}\left[\int_{-\infty}^{\infty} f_X(t)\ \mathrm{d}t\right] = f_X(x)$$
Since $\overline{F_X}(x) = 1 - F_X(x)$, it follows that
$$f_X(x) = -\frac{\mathrm{d}}{\mathrm{d}x}\left[\overline{F_X}(x)\right]$$
These relations can be visualized using the following schematic:



  - It is useful to note that PDF's, CDF's, and Survivals uniquely determine distributions. That is, if $f_X(x) = f_Y(y)$ or $F_X(x) = F_Y(y)$ or $\overline{F_X}(x) = \overline{F_Y}(y)$, then $X \overset{\mathrm{d}}{=} Y$.

6 Moment Generating Functions

- The **moment generating function** (or MGF, for short) of a random variable X, denoted $\psi_X(t)$, is defined as
$$\psi_X(t) := \mathbb{E}\left(e^{Xt}\right)$$
In other words,
$$\psi_X(t) = \begin{cases} \sum_x e^{xt}\mathbb{P}(X = x) & \text{if } X \text{ is discrete} \\ \int_x e^{xt} f_X(x)\ \mathrm{d}x & \text{if } X \text{ is continuous} \end{cases}$$
  - For example, if $X \sim \text{Exp}(\lambda)$,
$$\psi_X(t) = \int_0^\infty \lambda e^{xt} e^{-\lambda x}\ \mathrm{d}x = \frac{\lambda}{\lambda - t}\int_0^\infty (\lambda - t)e^{-(\lambda - t)x}\ \mathrm{d}x = \frac{\lambda}{\lambda - t}$$
- MGF's are desirable because **they uniquely determine a distribution:** if $\psi_X(t) = \psi_Y(t)$ for random variables X and Y, then $X \overset{\mathrm{d}}{=} Y$. Additionally, they satisfy the following properties:
  - $\psi_{aX+b}(t) = e^{at}\psi_X(bt)$
  - $\psi_{X+Y}(t) = \psi_X(t)\psi_Y(t)$
  - $\frac{\mathrm{d}^n}{\mathrm{d}t^n}\big|_{t=0}\psi_X(t) = \mathbb{E}(X^n)$
- A list of a few common distributions and their MGF's is displayed below:

Distribution	**MGF**
$\text{Bern}(\alpha)$	$1 - \alpha + \alpha e^t$
$\text{Exp}(\lambda)$	$\dfrac{\lambda}{\lambda - t}$
$\text{Gamma}(r, \lambda)$	$\left(\dfrac{\lambda}{\lambda - t}\right)^r$
$\mathcal{N}(0, 1)$	$e^{\frac{t^2}{2}}$

- **Example:** Let $X, Y \overset{\text{i.i.d.}}{\sim} \text{Exp}(\lambda)$, and define $Z := X + Y$. The MGF of Z is given by (using property 1 above):
$$\psi_Z(t) = \psi_X(t)\psi_Y(t) = \frac{\lambda}{\lambda - t} \times \frac{\lambda}{\lambda - t} = \left(\frac{\lambda}{\lambda - t}\right)^2$$
Consulting the table above, this proves that $Z \sim \text{Gamma}(2, \lambda)$ [this proof is far simpler than the proof involving the convolution formula, discussed in section 8 below].

7 Functions of Random Variables

- The **Law of the Unconscious Statistician** (or **LOTUS**) provides a formula for the expectation of a function of a random variable:
$$\mathbb{E}\left[g(X)\right] = \int_{-\infty}^{\infty} g(x) f_X(x)\ \mathrm{d}x$$

- The discrete version of the LOTUS states

$$\mathbb{E}[g(X)] = \sum_x g(x)\mathbb{P}(X = x)$$

- Given a random variable X and an invertible function g, the **change of variable** formula provides a way to identify the PDF of $Y := g(X)$:

$$f_Y(y) = f_X\left[g^{-1}(y)\right] \cdot \left|\frac{\mathrm{d}}{\mathrm{d}y}\left[g^{-1}(y)\right]\right|$$

A proof of this can be found in **Appendix 13.2**.

- If g is non-invertible, break up the support of X into multiple regions over which g is invertible, then apply the change of variable formula.

- Instead of using the change of variable formula, it is sometimes helpful to use the **c.d.f. method** (sometimes known as the **method of distribution functions**). When using the c.d.f. method to identify the distribution of $Y := g(X)$, we first consider $F_Y(y)$, and later differentiate with respect to y to obtain the p.d.f. of Y.

- **Example:** If $Z \sim \mathcal{N}(0,1)$ and $T := Z^2$, then

$$F_T(t) = \mathbb{P}(T \le t) = \mathbb{P}(Z^2 \le t) = \mathbb{P}(-\sqrt{t} \le Z \le \sqrt{t}) = \mathbb{P}(Z \le \sqrt{t}) - \mathbb{P}(Z \le -\sqrt{t}) = \Phi(\sqrt{t}) - \Phi(-\sqrt{t})$$

Differentiate with respect to t and use the chain rule [as well as the fact that $\Phi'(x) = \phi(x)$] to see

$$f_T(t) = \frac{\mathrm{d}}{\mathrm{d}t}\left[\Phi(\sqrt{t}) - \Phi(-\sqrt{t})\right] = \frac{1}{2\sqrt{t}}\phi(\sqrt{t}) + \frac{1}{2\sqrt{t}}\phi(-\sqrt{t})$$

Since ϕ is even we have $\phi(\sqrt{t}) = \phi(-\sqrt{t})$ and our answer simplifies to

$$f_T(t) = \frac{1}{\sqrt{t}}\phi(\sqrt{t}) = \frac{1}{\sqrt{t}} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\sqrt{t})^2} = \frac{\frac{1}{\sqrt{2}}}{\sqrt{\pi}} t^{-\frac{1}{2}} e^{-\frac{1}{2}t} = \frac{\left(\frac{1}{2}\right)^{\frac{1}{2}}}{\Gamma\left(\frac{1}{2}\right)} t^{\frac{1}{2}-1} e^{-\frac{1}{2}t}$$

which shows that $T \sim \text{Gamma}\left(\frac{1}{2}, \frac{1}{2}\right)$. In Stat 135, you will come to recognize this as a special case of the **χ^2 distribution**.

8 Poisson Point Processes

- A **poisson point process** is an example of a **stochastic process** (for more on stochastic processes, we recommend taking Stat 150). A common example of a Poisson Point Process (or PPP) is that of cars arriving at a tollbooth, where cars arrive at a rate of λ per unit time. In this example:

  - The number of arrivals in a time interval $[0, t]$ follows a $\text{Pois}(\lambda t)$ distribution.
  - The time between consecutive arrivals follows a $\text{Exp}(\lambda)$ distribution.
  - The time until the r^{th} arrival follows a $\text{Gamma}(r, \lambda)$ distribution.

- **Poisson Thinning:** Consider a PPP with rate λ where arrivals are now categorized into two types: A and B. Let p_A denote the probability that any given arrival is of type A, and let $p_B := 1 - p_A$ denote the probability that any given arrival is of type A. Then the arrivals of type A follow a PPP with rate $\lambda \cdot p_A$, and the arrivals of type B follow a PPP with rate $\lambda \cdot p_B = \lambda(1 - p_A)$.

9 Multivariate Probability and Conditional Distributions

- The **joint PDF**, or simply the **joint density**, of two random variables X, Y is

$$f_{X,Y}(x,y) = \frac{\mathbb{P}(X \in \mathrm{d}x,\ Y \in \mathrm{d}y)}{\mathrm{d}x \cdot \mathrm{d}y}$$

- Two continuous random variables X and Y are said to be **independent** (notated $X \perp Y$) if $f_{X,Y}(x,y) = f_X(x) f_Y(y)$.

- In a multivariate setting, X and Y are referred to as the **marginal distributions**, and their PDF's are referred to as the **marginal p.d.f's**.

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\ \mathrm{d}y$$
$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\ \mathrm{d}x$$

- More generally, the joint PDF of random variables $X_1, \ldots, X_n$ is given by

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \frac{\mathbb{P}(X_1 \in \mathrm{d}x_1,\ \ldots\ , X_n \in \mathrm{d}x_n)}{\prod_{i=1}^n \mathrm{d}x_i}$$

- There is a multidimensional analog of the **LOTUS**:

$$\mathbb{E}\left[g(X_1, X_2, \ldots, X_n)\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} g(x_1,x_2,\ldots,x_n) f_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n)\ \mathrm{d}x_1\ \mathrm{d}x_2 \ldots \mathrm{d}x_n$$

- The **convolution formula** provides a way to identify the PDF of the sum of two random variables:

$$f_{X+Y}(s) = \int_{-\infty}^{\infty} f_{X,Y}(x, s-x)\ \mathrm{d}x = \int_{-\infty}^{\infty} f_{X,Y}(s-y, y)\ \mathrm{d}y$$

  Watch out for the limits of integration! Often times, one of the limits will involve s.

  – This formula can be derived using the c.d.f. method: letting $S := X + Y$ we have

$$F_S(s) := \mathbb{P}(S \le s) = \mathbb{P}(X + Y \le s) = \mathbb{P}(X \le s - Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{s-y} f_{X,Y}(x,y)\ \mathrm{d}y\ \mathrm{d}x$$
$$f_S(s) = \frac{\mathrm{d}}{\mathrm{d}s}F_S(s) = \frac{\partial}{\partial s}\left(\int_{-\infty}^{\infty}\int_{-\infty}^{s-y} f_{X,Y}(x,y)\ \mathrm{d}y\ \mathrm{d}x\right)$$
$$= \int_{-\infty}^{\infty}\frac{\partial}{\partial s}\left(\int_{-\infty}^{s-y} f_{X,Y}(x,y)\ \mathrm{d}y\right)\ \mathrm{d}x = \int_{-\infty}^{\infty} f_{X,Y}(s-y,y)\ \mathrm{d}y$$

  (The justification for swapping differentiation and integration is outside the scope of this class, and can be taken for granted).

- Given two random variables X and Y with joint PDF $f_{X,Y}(x,y)$, we can extract **conditional distributions** as follows:

$$f_X(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{f_{X,Y}(x,y)}{\int_{-\infty}^{\infty} f_{X,Y}(x,y)\ \mathrm{d}x}$$

$$f_Y(y \mid x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{f_{X,Y}(x,y)}{\int_{-\infty}^{\infty} f_{X,Y}(x,y)\ \mathrm{d}y}$$

These definitions follow from Bayes' Rule.

- Given a series of n i.i.d. random variables $X_1, \ldots, X_n$ with shared PDF $f_X(x)$ and shared CDF $F_X(x)$, we can examine the distribution of the k^{th} smallest RV, denoted $X_{(k)}$. This is what we call the **k^{th} order statistic**, and it has PDF given by

$$f_{X_{(k)}}(x) = \binom{n}{1, k-1, n-k} f_X(x) \cdot [F_X(x)]^{k-1} - [1 - F_X(x)]^{n-k}$$

Note that $X_{(1)} := \min_{1 \le i \le n} \{X_i\}$ and $X_{(n)} := \max_{1 \le i \le n} \{X_i\}$.

10 Covariances and Conditionals

- **Covariance** is a measure of how related two random variables are to each other. It is defined as

$$\begin{aligned}\mathrm{Cov}(X,Y) &:= \mathbb{E}\left\{[X - \mathbb{E}(X)]\,[Y - \mathbb{E}(Y)]\right\} \\ &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)\end{aligned}$$

Note that if $X \perp Y$, $\mathrm{Cov}(X,Y) = 0$. [**The converse is not necessarily true!** See **Appendix 13.3**]

  - **Correlation** is a standardized measure of covariance:

$$\mathrm{Corr}(X,Y) := \frac{\mathrm{Cov}(X,Y)}{\mathrm{SD}(X)\mathrm{SD}(Y)}$$

Note that $-1 \le \mathrm{Corr}(X,Y) \le 1$.

- The variance of the sum of two random variables is given by

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y)$$

More generally,

$$\begin{aligned}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) &= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{Cov}(X_i, X_j) \\ &= \sum_{i=1}^{n} \mathrm{Var}(X_i) + \sum_{i \ne j} \mathrm{Cov}(X_i, X_j)\end{aligned}$$

If $X_1, \ldots, X_n$ are **identically distributed** then the formula above simplifies to:

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = n\mathrm{Var}(X_1) + n(n-1)\mathrm{Cov}(X_1, X_2)$$

11 Bivariate Normal Distribution

- **Bivariate Distributions** arise when we have two **correlated** random variables. We say X and Y follow a **Standard Bivariate Normal Distribution** with correlation ρ if and only if

$$Y = \rho X + Z \cdot \sqrt{1 - \rho^2}$$

where $X, Z \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. The joint PDF of X and Y is then given by

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left(x^2 - 2\rho xy + y^2\right)\right\}$$

- We say U and V follow a **Bivariate Normal Distribution** with parameters $\mu_U, \mu_V, \sigma_U^2, \sigma_V^2$, and ρ if and only only if

$$X = \frac{U-\mu_U}{\sigma_u}; \qquad Y = \frac{V-\mu_V}{\sigma_V^2}$$

follow a standard bivariate distribution with correlation $\rho = \text{Corr}(X,Y) = \text{Corr}(U,V)$.

- Working with the joint PDF of a bivariate normal pair of random variables is unwieldy. As such, most problems related to the bivariate normal distribution require you to find an an equivalent formulation of the problem involving the *independent* normal random variables X and Z.
  - **Example:** Let X, Y follow a standard bivariate normal distribution (SBVND) with correlation ρ; suppose we are tasked with finding $\mathbb{P}(|Y - \rho X| < \alpha)$ for some $\alpha \in (0,1)$. We use the definition of the SBVND, which guarantees the existence of a random variable $Z \sim \mathcal{N}(0,1)$ such that

$$Y = \rho X + Z\sqrt{1-\rho^2}$$

Using this definition, we can write

$$\begin{aligned}
\mathbb{P}(|Y - \rho X| < \alpha) &= \mathbb{P}\left(\left|\rho X + Z\sqrt{1-\rho^2} - \rho X\right| < \alpha\right) \\
&= \mathbb{P}\left(\sqrt{1-\rho^2}|Z| < \alpha\right) = \mathbb{P}\left(|Z| < \frac{\alpha}{\sqrt{1-\rho^2}}\right) \\
&= 2\Phi\left(\frac{\alpha}{\sqrt{1-\rho^2}}\right) - 1
\end{aligned}$$

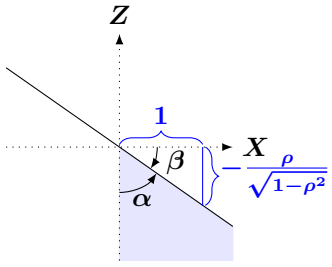
Compare this with trying to use direct integration:

$$\mathbb{P}(|Y - \rho X| < \alpha) = \int_{-\infty}^{\infty}\int_{\rho x - \alpha}^{\rho x + \alpha} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left(x^2 - 2\rho xy + y^2\right)\right\} \, dy \, dx$$

- The Standard Bivariate Normal displays **rotational invariance** (or **rotational symmetry**). This enables us to compute certain probabilities by looking at the ratio of angles, as opposed to integrating the joint density.
  - **Example:** Let $(X,Y) \sim \text{SBVN}(\rho)$ where $\rho > 0$; to find $\mathbb{P}(X \le 0, Y \le 0)$ we write

$$\mathbb{P}(X \le 0, Y \le 0) = \mathbb{P}(X \le 0, \rho X + \sqrt{1-\rho^2}Z \le 0) = \mathbb{P}\left(X \le 0,\ Z \le -\frac{\rho}{\sqrt{1-\rho^2}}X\right)$$

Sketch the region $\mathcal{A} := \left\{(X,Y) \in \mathbb{R}^2 \mid X \le 0,\ Z \le -\frac{\rho}{\sqrt{1-\rho^2}}X\right\}$:



Then, by rotational symmetry, we know

$$\mathbb{P}[(X,Z)\in\mathcal{A}] = \frac{\alpha}{2\pi} = \frac{\frac{\pi}{2}-\beta}{2\pi}$$

To find β, examine the blue triangle shown on the figure above:

$$\tan(\beta) = -\frac{\rho}{\sqrt{1-\rho^2}} \implies \beta = \arctan\left(-\frac{\rho}{\sqrt{1-\rho^2}}\right) = -\arctan\left(\frac{\rho}{\sqrt{1-\rho^2}}\right)$$

Therefore,

$$\mathbb{P}[(X,Z)\in\mathcal{A}] = \frac{\frac{\pi}{2}+\arctan\left(\frac{\rho}{\sqrt{1-\rho^2}}\right)}{2\pi} = \boxed{\frac{1}{4}+\frac{1}{2\pi}\arctan\left(\frac{\rho}{\sqrt{1-\rho^2}}\right)}$$

12 Tips & Tricks

- **Expectations:** When asked to compute the expectation of a quantity, there are three main tricks you can use:
  - **(i) The definition of expectation**. Though sometimes useful, this often leads to a lot of algebra (which in turn can lead to errors!).
  - **(ii) Indicators**. Again, if there's a count involved, see if you can use indicators.
  - **(iii) Relations.** If you're trying to find $\mathbb{E}(X)$, can you write X as the sum of other known random variables? For example, if $X\sim\text{Bin}(2,p)$ you can write $X = B_1+B_2$ where $B_1, B_2 \overset{\text{i.i.d.}}{\sim} \text{Bern}(2,p)$ so $\mathbb{E}(X) = \mathbb{E}(B_1+B_2) = \mathbb{E}(B_1)+\mathbb{E}(B_2) = 2p$. This is a lot easier than using the definition!
- **Maxes go with CMF's, Min's go with Survivals.** Consider random variables X_1, X_2, X_3. If the max of these RV's is less than k, it automatically follows that all three RV's must also be less than k. Similarly, if the minimum is greater than c, all three RV's must be greater than c.
  - Be careful though! A common mistake is to write something like this:

    $$\mathbb{P}(\max\{X_1,X_2,X_3\}\geq k) \implies \mathbb{P}(X_1\geq k, X_2\geq k, X_3\geq k)$$

    This is wrong!!! Suppose $X_1=2$, $X_3=5$, and $X_4=7$. Here, $\max\{X_1,X_2,X_3\}\geq 3$ however not all three RV's are greater than 3!
- **Try to Get Out of Doing Work:** wherever possible, relate algebraic quantities to things pertaining to distributions you know how to work with. For example, if you're asked to evaluate

  $$\int_0^\infty x^{2020}e^{-3x}\,\mathrm{d}x$$

  don't integrate by parts 2,020 times! Rather, multiply by some clever prefactors to turn the integrand into the PDF of the Gamma$(2021,3)$ distribution and use the fact that PDF's integrate to 1 when integrated over their support.
- **Identify the concept(s) being tested.** Often times, problems in Stat 134 test you on *multiple* topics. Identifying what these topics are will help you in creating a plan on how to approach the problem.

- **With or Without Replacement?** Though there are exceptions to this rule, generally speaking "without replacement" means you'll be working with the Hypergeometric Distribution; "with replacement" means you'll be working with the Binomial Distribution.

- **Joint Distributions:** Whenever you're dealing with a joint distribution, **always sketch the support.** This is where students often get the most confused; successfully sketching the support/region of integration is crucial in approaching joint distribution problems.

- **Bivariate Normal:** Dealing with X and Y (which are correlated) is almost always difficult; rather, use the definition of Bivariate Normal Random Variables to write everything in terms of X and Z (which are independent).

13 Appendices

13.1 Appendix 1: Proof of the Law of the Unconscious Statistician (Discrete Case)

Theorem 2: For a discrete random variable X and a differentiable function $g(x)$,

$$\mathbb{E}[g(X)] = \sum_k g(k)\mathbb{P}(X = x)$$

Proof: Definitionally, we have

$$\mathbb{E}[g(X)] := \sum_y y \cdot \mathbb{P}[g(X) = y]$$

Ultimately, we need a way of rewriting the event $\{g(X) = y\}$. Note that we can write $\{g(X) = y\}$ as a union of sets of the form $\{X = k\}$ where $g(k) = y$. That is,

$$\{g(X) = y\} = \bigcup_{k:g(k)=y} \{X = k\}$$

Therefore,

$$\begin{aligned}
\mathbb{E}[g(X)] &:= \sum_y y \cdot \mathbb{P}[g(X) = y] \\
&= \sum_y y \cdot \mathbb{P}\left(\bigcup_{k:g(k)=y} \{X = k\} \right) \\
&= \sum_y y \cdot \sum_{k:g(k)=y} \mathbb{P}(X = k) = \sum_y \sum_{k:g(k)=y} y \cdot \mathbb{P}(X = k)
\end{aligned}$$

In the inner sum, we may write y as $g(k)$:

$$\mathbb{E}[g(X)] = \sum_y \sum_{k:g(k)=y} y \cdot \mathbb{P}(X = k) = \sum_y \sum_{k:g(k)=y} g(k) \cdot \mathbb{P}(X = k)$$

Now, because the inner sum ranges over all values of k where $g(k) = y$, and the outer sum ranges over all values of y, the two sums together will range over all possible values of k. Therefore, we have

$$\mathbb{E}[g(X)] = \sum_k g(k) \cdot \mathbb{P}(X = k)$$

□

13.2 Appendix 2: Proof of the Change of Variable Formula

Theorem 1: Let X be a random variable with PDF $f_X(x)$, and let g be an invertible function. Then, if we define $Y := g(X)$,

$$f_Y(y) = f_X[g^{-1}(y)] \cdot \left| \frac{\mathrm{d}}{\mathrm{d}y} \left[g^{-1}(y) \right] \right|$$

Proof: We will use the CDF method. First assume g is increasing: then

$$F_Y(y) := \mathbb{P}(Y \leq y) = \mathbb{P}[g(X) \leq y] = \mathbb{P}[X \leq g^{-1}(y)]$$

Note that the inequality sign did not change directions because we are assuming g is increasing. Then:

$$F_Y(y) = \mathbb{P}[X \leq g^{-1}(y)] = \int_{-\infty}^{g^{-1}(y)} f_X(x)\ \mathrm{d}x$$

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y}\left[F_y(y)\right] = \frac{\mathrm{d}}{\mathrm{d}y} \int_{-\infty}^{g^{-1}(y)} f_X(x)\ \mathrm{d}x = f_X\left[g^{-1}(y)\right] \cdot \frac{\mathrm{d}}{\mathrm{d}y}\left[g^{-1}(y)\right]$$

where the final equality follows from the Fundamental Theorem of Calculus. Now, consider the case where g is decreasing:

$$F_Y(y) := \mathbb{P}(Y \leq y) = \mathbb{P}[g(X) \leq y] = \mathbb{P}[X \geq g^{-1}(y)] = 1 - \mathbb{P}\left[X \leq g^{-1}(y)\right]$$

where the sign of the inequality switched, because we are assuming that g is decreasing. Substituting and applying the Fundamental Theorem of Calculus as we did before yields

$$f_Y(y) = -f_X\left[g^{-1}(y)\right] \cdot \frac{\mathrm{d}}{\mathrm{d}y}\left[g^{-1}(y)\right]$$

Note, because g is decreasing we have that g^{-1} is decreasing as well, meaning $\frac{\mathrm{d}}{\mathrm{d}y}\left[g^{-1}(y)\right] < 0$ [which, in turn, ensures that our quantity for $f_Y(y)$ above is positive]. As such, we can combine our two cases (g increasing and g decreasing) using an absolute value:

$$f_Y(y) = f_X[g^{-1}(y)] \cdot \left| \frac{\mathrm{d}}{\mathrm{d}y} \left[g^{-1}(y) \right] \right|$$

□

13.3 Appendix 3: Uncorrelated Does Not Imply Independence

Let $X \sim \mathrm{Unif}\{-1, 0, 1\}$ and set $Y := X^2$. Then

$$\begin{aligned}
\mathrm{Cov}(X, Y) &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \\
&= \mathbb{E}(X \cdot X^2) - \mathbb{E}(X)\mathbb{E}(X^2) \\
&= \mathbb{E}(X^3) - \cancelto{0}{\mathbb{E}(X)}\ \mathbb{E}(X^2) \\
&= \mathbb{E}(X^3) = 0
\end{aligned}$$

Therefore, $\mathrm{Corr}(X, Y) = 0$; that is, X and Y are uncorrelated.

However, consider $\mathbb{P}(X = 1, Y = 0)$. Note that this is zero, if $X = 1$ then Y must also be 1. However,

$$\mathbb{P}(X = 1) = \frac{1}{3}; \quad \mathbb{P}(Y = 0) = \frac{1}{3} \implies \mathbb{P}(X = 1)\mathbb{P}(Y = 0) = \frac{1}{9}$$

In other words, $\mathbb{P}(X = 1, Y = 0) \neq \mathbb{P}(X = 1)\mathbb{P}(Y = 0)$, which means X and Y are not independent. Therefore, X and Y form a pair of random variables that are uncorrelated but dependent.

13.4 Appendix 4: Proof of the Law of Iterated Expectation

Theorem 2: $\mathbb{E}_Y\left[\mathbb{E}_{X|Y}(X \mid Y)\right] = \mathbb{E}_X(X)$.

Proof: We outline the proof for the case where X and Y are both discrete. (A rigorous proof of the continuous case requires some additional theory from measure theory and other advanced branches of mathematics).

$$\begin{aligned}
\mathbb{E}_Y\left[\mathbb{E}_{X|Y}(X \mid Y)\right] &= \mathbb{E}_Y\left[\sum_x x \cdot \mathbb{P}(X = x \mid Y)\right] \\
&= \sum_y \left\{\left[\sum_x x \cdot \mathbb{P}(X = x \mid Y)\right] \cdot \mathbb{P}(Y = y)\right\} \\
&= \sum_x \sum_y x \cdot \mathbb{P}(X = x \mid Y)\mathbb{P}(Y = y) \\
&= \sum_x \sum_y x \cdot \mathbb{P}(X = x, Y = y) \\
&=: \mathbb{E}_X(X)
\end{aligned}$$

A Note: If we were being rigorous, we should justify step (3) (in which we swapped the order of summation). Such a justification, though not incredibly time-consuming, would require some results from Real Analysis/Measure Theory (and some nuanced arguments pertaining to convergence of sums) and is therefore outside the scope of the Stat 134. □

13.5 Appendix 5: Additional Resources

A First Course in Probability, by Sheldon Ross (PDF Available online)

This textbook is a popular choice among Probability courses in other universities. The level of mathematical rigor is close to that of Pitman's text, though it aims to cover a (slightly) broader scheme of topics (for example, devoting an entire chapter to the topic of Simulation). Though the order of topics presented is quite different than that of Pitman, Chapters 2 - 7 (with some sections from Chapter 8) are directly applicable to Stat 134.

Additionally, students who have not taken Math 55 may benefit from reading Chapter 1, which provides an overview of basic counting and combinatorial principles (which is pretty much a *de-facto* prerequisite for Stat 134).

Introduction to Probability, by Andersen, Seppäläinen, and Valkó

First published in 2018, this is probably the most up-to-date resource of the textbooks listed here. Many universities (including the University of Washington) have adopted this textbook to be the main text of choice for undergraduate probability courses. It offers a more mathematical approach to Probability, though the topics covered are very similar to Ross's textbook. [Most notably, it includes an optional description of Measure Theory, which is the branch of mathematics more advanced probability courses are built on.]

All of Statistics, by Larry J. Wasserman (Available for free through SpringerLink)

Though concise, this book provides a good overview of many concepts of Probability and Statistics. Specifically, most of Chapters 1 - 4 are directly applicable to Stat 134. We highly recommend this text as a refresher of concepts, as well as a resource for future coursework.

The Probability Lifesaver, by Steven J. Miller

Contrary to *All of Statistics*, this is quite a hefty compendium. However, it contains many worked-out examples, as well as some very useful exercises. This text also includes a section on measure theory, a section that introduces some of the shortcomings of the more basic treatments of probability (including, for example, the Banach-Tarski paradox). On the whole, however, it does not demand an extraordinary level of mathematical expertise.