Contents

2 Random Variables

In this chapter, we discuss discrete and continuous random variables, which are quantities whose values depend on the outcome of a random event, and how to compute associated statistics such as the expected value, variance, and standard deviation. We also discuss joint distributions and independence of random variables and the related notions of covariance and correlation.

We then study several fundamentally important probability distributions, such as the Poisson distribution and the normal distribution, with an ultimate goal of laying the foundation to discussing their applications in statistics. In particular, we describe how these distributions naturally arise in a wide array of practical situations and how to use these random-variable models to provide new information about these phenomena.

2.1 Discrete Random Variables

- When we observe the result of an experiment, we are often interested in some specific property of the outcome rather than the entire outcome itself.
    - For example, if we flip a fair coin 5 times, we may want to know only the total number of heads obtained, rather than the exact sequence of all 5 flips.
    - As another example, if we roll a pair of dice (e.g., in playing the dice game craps) we may want to know the sum of the outcomes rather than the results of each individual roll.

2.1.1 Definition and Examples

- Formally, properties of outcomes can be thought of as functions defined on the outcomes of a sample space:
- Definition: A random variable is a (real-valued) function defined on the outcomes in a sample space. If the sample space is finite (or countably infinite), we refer to the random variable as a discrete random variable.
  - Example: For the experiment of flipping a coin 5 times, with corresponding sample space S, one random variable X is the total number of heads obtained. The value of X on the outcome $HHHHT$ is 4, while the value of X on the outcome $TTTTT$ is 0.
  - Example: For the experiment of rolling a pair of dice, with corresponding sample space S, one random variable Y is the sum of the outcomes. The value of Y on the outcome $(1, 4)$ is 5, while the value of Y on the outcome $(6, 6)$ is 12.
- If X is a random variable, the set of outcomes on which X takes a particular value (or range of values) is a subset of the sample space, which is to say, it is an event.
  - Thus, if we have a probability distribution on the sample space, we may therefore ask about quantities like (i) $P(X = n)$, the probability that X takes the value n, or (ii) $P(X \geq 5)$, the probability that the value of X is at least 5, or (iii) $P(2 < X < 4)$, the probability that the value of X is strictly between 2 and 4.
  - A common way to tabulate all of this information is to make a list or table of all the possible values of X along with their corresponding probabilities. The associated function is called the probability density function of X:
- Definition: If X is a random variable on the sample space S, then the function p_X such that $p_X(E) = P(X \in E)$ for any set of numbers E is called the probability density function (pdf) of X.
  - Explicitly, the value of $p_X(a)$ on a real number a is the probability that the random variable X takes the value a.
  - For discrete random variables with a small number of outcomes, we usually describe the probability density function using a table of values. In certain situations, we can find a convenient formula for the values of the probability density function on arbitrary events, but in many other cases, the best we can do is simply to tabulate all the different values.
- Example: If two standard 6-sided dice are rolled, find the probability distribution for the random variable X giving the sum of the outcomes. Then calculate (i) $P(X = 7)$, (ii) $P(4 < X < 9)$, and (iii) $P(X \leq 6)$.
  - To find the probability distribution for X, we identify all of the possible values for X and then tabulate the respective outcomes in which each value occurs.
  - We can see that the possible values for X are 2, 3, 4, ... , 12, and that they occur as follows:

Value	Outcomes	Probability
$X = 2$	$(1,1)$	$1/36$
$X = 3$	$(1,2)$, $(2,1)$	$2/36$
$X = 4$	$(1,3)$, $(2,2)$, $(3,1)$	$3/36$
$X = 5$	$(1,4)$, $(2,3)$, $(3,2)$, $(4,1)$	$4/36$
$X = 6$	$(1,5)$, $(2,4)$, $(3,3)$, $(4,2)$, $(5,1)$	$5/36$
$X = 7$	$(1,6)$, $(2,5)$, $(3,4)$, $(4,3)$, $(5,2)$, $(6,1)$	$6/36$
$X = 8$	$(2,6)$, $(3,5)$, $(4,4)$, $(5,3)$, $(6,2)$	$5/36$
$X = 9$	$(3,6)$, $(4,5)$, $(5,4)$, $(6,3)$	$4/36$
$X = 10$	$(4,6)$, $(5,5)$, $(6,4)$	$3/36$
$X = 11$	$(5,6)$, $(6,5)$	$2/36$
$X = 12$	$(6,6)$	$1/36$

  - Then we have $P(X = 7) = \frac{6}{36} = \boxed{\frac{1}{6}}$, $P(4 < X < 9) = \frac{4}{36} + \frac{5}{36} + \frac{6}{36} + \frac{5}{36} = \boxed{\frac{5}{9}}$, and $P(X \leq 6) = \frac{1}{36} + \frac{2}{36} + \frac{3}{36} + \frac{4}{36} + \frac{5}{36} = \boxed{\frac{5}{12}}$.

- Remark: In this situation, there is a moderately nice formula for the probability density function: specifically, we have $p_X(n) = P(X = n) = \dfrac{6 - |7 - n|}{36}$ for integers n with $2 \leq n \leq 12$, and $p_X(n) = 0$ for all other values.

- Example: If a fair coin is flipped 4 times, find the probability distributions for the random variable X giving the number of total heads obtained, and for the random variable Y giving the longest run of consecutive tails obtained. Then calculate (i) $P(X = 2)$, (ii) $P(X \geq 3)$, (iii) $P(1 < X < 4)$, (iv) $P(Y = 1)$, (v) $P(Y \leq 3)$, and (vi) $P(X = Y = 2)$.

  - For X, we obtain the following distribution:

Value	Outcomes	Probability
$X = 0$	$TTTT$	1/16
$X = 1$	$TTTH$, $TTHT$, $THTT$, $HTTT$	1/4
$X = 2$	$TTHH$, $THTH$, $THHT$, $HTTH$, $HTHT$, $HHTT$	3/8
$X = 3$	$THHH$, $HTHH$, $HHTH$, $HHHT$	1/4
$X = 4$	$HHHH$	1/16

  - For Y, we obtain the following distribution:

Value	Outcomes	Probability
$Y = 0$	$HHHH$	1/16
$Y = 1$	$THTH$, $THHT$, $THHH$, $HTHT$, $HTHH$, $HHTH$, $HHHT$	7/16
$Y = 2$	$TTHT$, $TTHH$, $THTT$, $HTTH$, $HHTT$	5/16
$Y = 3$	$TTTH$, $HTTT$	1/8
$Y = 4$	$TTTT$	1/16

  - We can then quickly compute $P(X = 2) = \boxed{\dfrac{3}{8}}$, $P(X \geq 3) = \dfrac{1}{4} + \dfrac{1}{16} = \boxed{\dfrac{5}{16}}$, $P(1 < X < 4) = \dfrac{3}{8} + \dfrac{1}{4} = \boxed{\dfrac{5}{8}}$, $P(Y = 1) = \boxed{\dfrac{7}{16}}$, and $P(Y \leq 3) = \dfrac{1}{16} + \dfrac{7}{16} + \dfrac{5}{16} + \dfrac{1}{8} = \boxed{\dfrac{15}{16}}$.
  - To find $P(X = Y = 2)$ we must look at the individual outcomes where X and Y are both equal to 2. There are 3 such outcomes ($TTHH$, $HTTH$, $HHTT$), so $P(X = Y = 2) = \boxed{\dfrac{3}{16}}$.

- If we have a random variable X defined on the sample space, then since X is a function on outcomes, we can define various new random variables in terms of X.

  - If g is any real-valued function, we can define a new random variable $g(X)$ by evaluating g on all of the results of X. Some possibilities include $g(X) = 2X$, which doubles every value of X, or $g(X) = X^2$, which squares every value of X.
  - More generally, if we have a collection of random variables $X_1, X_2, \ldots, X_n$ defined on the same sample space, we can construct new functions in terms of them, such as the sum $X_1 + X_2 + \cdots + X_n$ that returns the sum of the values of $X_1, \ldots, X_n$ on any given outcome.

- A particular random variable is the random variable identifying whether an event has occurred:

- Definition: If E is any event, we define the Bernoulli random variable for E to be $X_E = \begin{cases} 1 & \text{if } E \text{ occurs} \\ 0 & \text{if } E \text{ does not occur} \end{cases}$.

  - The name for this random variable comes from the idea of a Bernoulli trial, which is an experiment having only two possible outcomes, success (with probability p) and failure (with probability $1 - p$). We think of E as being the event of success, while E^c is the event of failure.
  - Many experiments consist of a sequence of independent Bernoulli trials, in which the outcome of each trial is independent from the outcomes of all of the others. For example, flipping a (fair or unfair) coin 10 times and testing whether heads is obtained for each flip is an example of a Bernoulli trial.
  - Using our results on independence of events, we can describe explicitly the probability distribution of the random variable X giving the total number of successes when n independent Bernoulli trials are performed, each with a probability p of success.

- Proposition (Binomial Distribution): Let X be the random variable representing the total number of successes obtained by performing n independent Bernoulli trials each of which has a success probability p. Then the probability distribution of X is the binomial distribution, in which $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$ for integers k with $0 \leq k \leq n$, and $P(X = k) = 0$ for other k.
  - The binomial distribution is so named because of the presence of the binomial coefficients $\binom{n}{k}$. One particular example of this distribution is the total number of heads obtained by flipping n unfair coins each of which has probability p of landing heads.
  - Proof: From our results on binomial coefficients, we can see that there are $\binom{n}{k}$ ways to choose k trials yielding success out of a total of n. Furthermore, since all of the trials are independent, the probability of obtaining any given pattern of k successes and $n-k$ failures is equal to $p^k(1-p)^{n-k}$.
  - Thus, since the probability of obtaining any given one of the $\binom{n}{k}$ outcomes with exactly k successes is $p^k(1-p)^{n-k}$, the probability of obtaining exactly k successes is $\binom{n}{k} p^k (1-p)^{n-k}$, as claimed.
- Example: A baseball player's batting average is 0.378, meaning that she has a probability of 0.378 of getting a hit on any given at-bat, independently of any other at-bat. Find the probability that in her first 100 at-bats that she gets (i) exactly 37 hits, (ii) exactly 40 hits, and (iii) exactly 50 hits.
  - We can view each at-bat as an independent Bernoulli trial (with a hit being considered a success) with $p = 0.378$, so the total number of hits will be binomially distributed.
  - Then the probability of getting 37 hits is $\boxed{\binom{100}{37} \cdot 0.378^{37} \cdot 0.622^{63}} \approx 8.13\%$, the probability of 40 hits is $\boxed{\binom{100}{40} \cdot 0.378^{40} \cdot 0.622^{60}} \approx 7.33\%$, and the probability of 50 hits is $\boxed{\binom{100}{50} \cdot 0.378^{50} \cdot 0.622^{50}} \approx 0.37\%$.

2.1.2 Expected Value

- If we repeat an experiment many times and record the different values of a random variable X each time, a useful statistic summarizing the outcomes is the average value of the outcomes.
  - We would like a way to describe the "average value" of a random variable X.
  - Suppose that the sample space has outcomes $s_1, s_2, \ldots, s_n$ on which the random variable X takes on the values $x_1, x_2, \ldots, x_n$ with probabilities $p_1, p_2, \ldots, p_n$, where $p_1 + \cdots + p_n = 1$.
  - Under our interpretation of these probabilities as giving the relative frequencies of events when we repeat the experiment many times, if we perform the experiment N times where N is large, we should obtain the event $X = x_i$ approximately $p_i N$ times for each $1 \leq i \leq n$.
  - The average value would then be $\dfrac{(p_1 N)x_1 + (p_2 N)x_2 + \cdots + (p_n N)x_n}{N} = p_1 x_1 + p_2 x_2 + \cdots + p_n x_n$.
  - We may use this calculation to give a definition of the "average value" for an arbitrary discrete random variable:
- Definition: If X is a discrete random variable, the expected value of X, written $E(X)$, is the sum $E(X) = \sum_{s_i \in S} P(s_i) X(s_i)$ over all outcomes s_i in the sample space S. In words, the expected value is the average of the values that X takes on the outcomes in the sample space, weighted by the probability of each outcome.
  - The expected value is also sometimes called the mean or the average value of X, and is often also written as μ_X ("mu-X") or as $\overline{X}$ ("X-bar").
  - Example: If a fair coin is flipped once, the expected value of the random variable X giving the number of total heads obtained is equal to $E(X) = \dfrac{1}{2} \cdot 0 + \dfrac{1}{2} \cdot 1 = \boxed{\dfrac{1}{2}}$, because there are two possible outcomes, 0 heads and 1 head, each of probability 1/2.

- In the example above we see that the expected value captures the idea of an "average value" of the random variable when the experiment is repeated many times: if we flip a fair coin N times, we would expect to see about $N/2$ heads (yielding on average 1/2 head per flip) which agrees with the expected value of 1/2.
- Example: If an unfair coin with a probability 2/3 of landing heads is flipped once, the expected value of the random variable X giving the number of total heads obtained is equal to $E(X) = \frac{1}{3} \cdot 0 + \frac{2}{3} \cdot 1 = \boxed{\frac{2}{3}}$, because there are two possible outcomes, 0 heads and 1 head, of respective probabilities 1/3 and 2/3.
- Example: If a standard 6-sided die is rolled once, the expected value of the random variable X giving the result is equal to $E(X) = \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \frac{1}{6} \cdot 3 + \frac{1}{6} \cdot 4 + \frac{1}{6} \cdot 5 + \frac{1}{6} \cdot 6 = \boxed{\frac{7}{2}}$, because each of the 6 possible outcomes 1,2,3,4,5,6 has probability 1/6 of occurring.

- It is very important to note that the expected value of a discrete random variable can be infinite or even not defined at all.
  - Example: If X is the discrete random variable whose value is 2^n occurring with probability 2^{-n} for $n \geq 1$, then its expected value is $E(X) = 2 \cdot \frac{1}{2} + 4 \cdot \frac{1}{4} + 8 \cdot \frac{1}{8} + 16 \cdot \frac{1}{16} + \cdots = 1 + 1 + 1 + 1 + \cdots = \infty$.
  - Example: If Y is the discrete random variable whose value is $(-2)^n$ occurring with the probability 2^{-n} for $n \geq 1$, then its expected value is the sum $2 \cdot (-\frac{1}{2}) + 4 \cdot \frac{1}{4} + 8 \cdot (-\frac{1}{8}) + 16 \cdot \frac{1}{16} + \cdots = -1 + 1 - 1 + 1 - \cdots$. This sum does not converge (since the partial sums alternate forever between the values -1 and 0), and so the expected value of this random variable is not defined.

- A common application of expected value is to calculate the expected winnings from a game of chance:

- Example: In one version of a "Pick 3" lottery, a single entry ticket costs \$1. In this lottery, 3 single digits are drawn at random, and a ticket must match all 3 digits in the correct order to win the \$500 prize. What is the expected value of one ticket for this lottery?
  - From the description, we can see that there is a $\frac{1}{1000}$ probability of winning the prize and a $\frac{999}{1000}$ probability of winning nothing.
  - Since winning the prize nets a total of \$499 (the prize minus the \$1 entry fee), and winning nothing nets a total of $-$\$1, the expected value of the random variable giving the net winnings is equal to $\frac{1}{1000}(\$499) + \frac{999}{1000}(-\$1) = \boxed{-\$0.50}$.
  - The expected value of $-$\$0.50, in this case, indicates that if one plays this lottery many times, on average one should expect to lose 50 cents on every ticket.

- Example: In one version of the game "Chuck-a-luck", three standard 6-sided dice are rolled. Prizes for a bet of \$1 are awarded as follows: \$65 for a roll of three 6s, \$5 for a roll of two 6s, \$1 for one 6, and \$0 for any other roll. What is the expected value of one play of this game?
  - We calculate the probabilities of the various values for the random variable X tallying the net winnings.
  - Observe that the number of sixes obtained will be binomially distributed with $n = 3$ and $p = 1/6$: thus, the probability of getting k sixes will be $\binom{3}{k}(1/6)^k(5/6)^{3-k}$.
  - The probability of getting three sixes is $(1/6)^3 = 1/216$, and in this case, the net winnings total \$65.
  - The probability of getting two sixes is $3(1/6)^2(5/6) = 15/216$, and in this case, the net winnings total \$4.
  - The probability of getting one six is $3(1/6)(5/6)^2 = 75/216$, and in this case, the net winnings total \$0.
  - The probability of getting no sixes is $(5/6)^3 = 125/216$, and in this case, the net winnings total -\$1.
  - Therefore, the expected winnings from one play are $\frac{1}{216} \cdot (\$65) + \frac{15}{216} \cdot (\$4) + \frac{75}{216} \cdot (\$0) + \frac{125}{216} \cdot (-\$1) = \boxed{\$0}$.

- For this game, we can see that the expected winnings are \$0, meaning that the game is fair (in the sense that neither the player nor the person running the game should expect to win or lose money on average over the long term).

- Expected value has several important algebraic properties:

- Proposition (Linearity of Expected Value): If X and Y are discrete random variables defined on the same sample space whose expected values exist, and a and b are any real numbers, then $E(aX+b) = a \cdot E(X) + b$ and $E(X+Y) = E(X) + E(Y)$.

  - Intuitively, if the expected value of X is 4, then it is reasonable to feel that the expected value of $X+1$ should be 5, while the expected value of $2X$ should be 8. These two observations, taken together, form essentially the first part of the statement.
  - Likewise, if the expected value of Y is 3, then it is also reasonable to feel that the expected value of $X+Y$ should be 7, the sum of the expected values of X and Y.
  - Proof: Suppose the outcomes in the sample space are $s_1, s_2, \ldots$ with probabilities $p_1, p_2, \ldots$, where $p_1 + p_2 + \cdots = 1$. Also suppose that the values of X on these outcomes are $x_1, x_2, \ldots$ and the values of Y are $y_1, y_2, \ldots$
  - Then $E(X) = p_1x_1 + p_2x_2 + \cdots$ and $E(Y) = p_1y_1 + p_2y_2 + \cdots$.
  - Since $aX+b$ takes on the values ax_1+b, $ax_2+b, \ldots$ on the respective outcomes $s_1, s_2, \ldots$, we have $E(aX+b) = p_1(ax_1+b) + p_2(ax_2+b) + \cdots = a(p_1x_1 + p_2x_2 + \cdots) + b(p_1 + p_2 + \cdots) = a \cdot E(X) + b$.
  - Also, $X+Y$ takes on the values x_1+y_1, $x_2+y_2, \ldots$ on the respective outcomes $s_1, s_2, \ldots$, we have $E(X+Y) = p_1(x_1+y_1) + p_2(x_2+y_2) + \cdots = (p_1x_1 + p_2x_2 + \cdots) + (p_1y_1 + p_2y_2 + \cdots) = E(X) + E(Y)$.

- By using the linearity of expected value and decomposing random events into simpler pieces, we can compute expected values of more complicated random variables:

- Example: An unfair coin with probability p of landing heads is flipped n times. Find the expected number of heads obtained.

  - One approach would be to let X be the random variable giving the total number of heads, then compute the probability distribution of X and use the result to find the expected value.
  - Since the number of heads is binomially distributed, the probability of obtaining k heads is $\binom{n}{k}p^k(1-p)^{n-k}$, the expected value is $\displaystyle\sum_{k=0}^{n} \binom{n}{k}p^k(1-p)^{n-k} \cdot k$, which can eventually be evaluated (using algebraic identities for the binomial coefficients) as $\boxed{np}$.
  - However, a vastly simpler approach is to observe that we can write X as the sum of random variables $X = X_1 + X_2 + \cdots + X_n$, where X_i is the number of heads obtained on the nth flip.
  - Since $E(X_1) = E(X_2) = \cdots = E(X_n) = p$ since the flips each independently have a probability p of landing heads, by the additivity of expectation we see that $E(X) = E(X_1) + E(X_2) + \cdots + E(X_n) = \boxed{np}$.

- Example: Five cards are randomly drawn from a standard 52-card deck. Find the expected number of cards drawn that are (i) spades, (ii) aces, (iii) aces of spades.

  - For (i), if we let X be the random variable giving the total number of spades, then $X = X_1 + X_2 + X_3 + X_4 + X_5$ where X_i is the random variable that the ith card is a spade. Then we have $E(X_i) = \dfrac{13}{52} = \dfrac{1}{4}$ since each card can be one of 52 equally likely possibilities, 13 of which are spades, and so $E(X) = 5 \cdot \dfrac{1}{4} = \boxed{\dfrac{5}{4}}$.
  - For (ii), if we let Y be the random variable giving the total number of aces, then $Y = Y_1 + Y_2 + Y_3 + Y_4 + Y_5$ where Y_i is the random variable that the ith card is an ace. Then we have $E(Y_i) = \dfrac{4}{52} = \dfrac{1}{13}$ since each card can be one of 52 equally likely possibilities, 4 of which are aces, and so $E(Y) = 5 \cdot \dfrac{1}{13} = \boxed{\dfrac{5}{13}}$.

- In the same way, for (iii), if we decompose the random variable Z giving the total number of aces of spades as a sum over the 5 cards drawn from the deck, we see that $E(Z) = 5 \cdot \frac{1}{52} = \boxed{\frac{5}{52}}$.
- Remark: Notice that the values of the random variables $X_1, \ldots, X_5$ are not independent: for example, if the first card is a spade then the second card is less likely to be a spade (and vice versa, and similarly for the other cards). Nonetheless, the expected value of the total number of spades is still the sum of the individual expectations, even though the actual values of the corresponding random variables are not independent of one another.

2.1.3 Variance and Standard Deviation

- In addition to computing the expected value of a random variable, we also would like to be able to measure how much variation the values have relative to their expected value.
  - For example, if one random variable X is always equal to 2, then there is no variation in its value. If the random variable Y is equal to 0 half the time and 4 the other half the time, then its expected value is also 2, but there is much more variation in the values of Y.
- Definition: If X is a discrete random variable whose expected value $E(X) = \mu$ exists and is finite, we define the variance of X to be $\text{var}(X) = E[(X-\mu)^2]$, the expected value of the square of the difference between X and its expectation. The standard deviation of X, denoted $\sigma(X)$, is the square root of the variance: $\sigma(X) = \sqrt{\text{var}(X)}$.
  - Roughly speaking, the standard deviation measures the "average distance" that a typical outcome of X will be from the expected outcome.
  - We can also give another formula for the variance: by using the linearity of expectation, we can write $E[(X-\mu)^2] = E(X^2 - 2\mu X + \mu^2) = E(X^2) - 2\mu E(X) + \mu^2$, and since $\mu = E(X)$, this formula simplifies to $\text{var}(X) = E(X^2) - \mu^2 = E(X^2) - [E(X)]^2$.
  - It is often faster to compute $E(X)$ and $E(X^2)$ separately than to compute the probability distribution of $(X-\mu)^2$.
- Example: If a coin with probability p of landing heads is flipped once, find the expected value, variance, and standard deviation of the random variable X giving the number of heads.
  - There are two possible outcomes: either $X = 0$, which occurs with probability $1-p$, or $X = 1$, which occurs with probability p.
  - The expected value is then $E(X) = (1-p) \cdot 0 + p \cdot 1 = \boxed{p}$.
  - For the standard deviation, there are two possible outcomes of $(X-\mu)^2 = (X-p)^2$: either $X - \mu = (0-p)^2 = p^2$, which occurs with probability $1-p$, or $(X-\mu)^2 = (1-p)^2$, which occurs with probability p.
  - The variance is then $\text{var}(X) = (1-p) \cdot p^2 + p \cdot (1-p)^2 = \boxed{p(1-p)}$, and the standard deviation is $\sigma(X) = \sqrt{\text{var}(X)} = \boxed{\sqrt{p(1-p)}}$.
  - Alternatively, using the formula $\text{var}(X) = E(X^2) - [E(X)]^2$, we compute $E(X^2) = (1-p) \cdot 0^2 + p \cdot 1^2 = p$, and then $\text{var}(X) = p - p^2 = \boxed{p(1-p)}$, as above.
  - In particular, when $p = 1/2$, we see that the standard deviation is $1/2$. This agrees with the natural idea that although the expected number of heads obtained when flipping a fair coin is $1/2$, the actual outcome is always a distance $1/2$ away from the expectation (since it is either 0 or 1).
- Example: If a standard 6-sided die is rolled once, find the variance and standard deviation of the random variable X giving the result of the die roll.
  - Each of the possible outcomes $X = 1, 2, 3, 4, 5, 6$ occurs with probability $\frac{1}{6}$.

- We compute $E(X) = \frac{1}{6}\cdot 1 + \frac{1}{6}\cdot 2 + \frac{1}{6}\cdot 3 + \frac{1}{6}\cdot 4 + \frac{1}{6}\cdot 5 + \frac{1}{6}\cdot 6 = \frac{7}{2} = 3.5$, and also $E(X^2) = \frac{1}{6}\cdot 1^2 + \frac{1}{6}\cdot 2^2 + \frac{1}{6}\cdot 3^2 + \frac{1}{6}\cdot 4^2 + \frac{1}{6}\cdot 5^2 + \frac{1}{6}\cdot 6^2 = \frac{91}{6}$.
- Thus, $\text{var}(X) = E(X^2) - [E(X)]^2 = \frac{91}{6} - \frac{49}{4} = \boxed{\frac{35}{12}}$, and $\sigma(X) = \sqrt{\text{var}(X)} = \boxed{\sqrt{\frac{35}{12}}} \approx 1.708$.

- Example: A fair coin is flipped 4 times. Find the expected value, variance, and standard deviation of the random variable Y giving the longest run of consecutive tails obtained.
  - We have previously computed the probability distribution of Y:

n	0	1	2	3	4
$P(Y=n)$	1/16	7/16	5/16	1/8	1/16

  - We compute $E(Y) = \frac{1}{16}\cdot 0 + \frac{7}{16}\cdot 1 + \frac{5}{16}\cdot 2 + \frac{1}{8}\cdot 3 + \frac{1}{16}\cdot 4 = \boxed{\frac{27}{16}} = 1.6875$, and also $E(Y^2) = \frac{1}{16}\cdot 0^2 + \frac{7}{16}\cdot 1^2 + \frac{5}{16}\cdot 2^2 + \frac{1}{8}\cdot 3^2 + \frac{1}{16}\cdot 4^2 = \frac{61}{16}$.
  - Thus, $\text{var}(Y) = E(Y^2) - [E(Y)]^2 = \frac{61}{16} - \frac{729}{256} = \boxed{\frac{247}{256}}$, and $\sigma(Y) = \sqrt{\text{var}(Y)} = \boxed{\sqrt{\frac{247}{256}}} \approx 0.9823$.

- As long as the expected value exists, the variance will always exist, because it is computed by summing nonnegative values. However, even when the expected value is finite, the variance can be infinite:

- Example: Show that the random variable that takes the value 2^n with probability $2/3^n$, for integers $n \geq 1$, has a finite expected value but an infinite variance.
  - We compute the expected value $E(X) = 2\cdot\frac{2}{3} + 2^2\cdot\frac{2}{9} + 2^3\cdot\frac{2}{27} + \cdots = 4$ using the formula[1] $a + ar + ar^2 + \cdots = \frac{a}{1-r}$ for the sum of a geometric series.
  - For the variance, we also must compute $E(X^2) = 4\cdot\frac{2}{3} + 4^2\cdot\frac{2}{9} + 4^3\cdot\frac{2}{27} + \cdots = \infty$, since the common ratio in this geometric series is $\frac{4}{3}$, which is greater than 1. But then the variance is $\text{var}(X) = E(X^2) - E(X)^2 = \infty$, which is to say, the variance is infinite.

- The variance and standard deviation also possess some algebraic properties like those of expected value:

- Proposition (Properties of Variance): If X is a discrete random variable and a and b are any real numbers, then $\text{var}(aX+b) = a^2\cdot\text{var}(X)$ and $\sigma(aX+b) = |a|\,\sigma(X)$.
  - Proof: From the linearity of expectation, we know that $E(aX+b) = a\cdot E(X) + b$, and therefore we have $aX + b - E(aX+b) = aX + b - a\,E(X) - b = a\cdot[X - E(x)]$.
  - Then $\text{var}(aX+b) = E[(aX+b-E(aX+b))^2] = E[a^2\cdot(X-E(x))^2] = a^2\text{var}(X)$, and by taking the square root of both sides we then get $\sigma(aX+b) = |a|\,\sigma(X)$.

- Example: If X is a random variable with expected value 1 and standard deviation 3, what are the expected value and standard deviation of $2X+4$?
  - From the properties given above, we have $E(2X+4) = 2E(X)+4 = \boxed{6}$, and $\sigma(2X+4) = |2|\,\sigma(X) = \boxed{6}$.

[1] If $|r| < 1$ and S is the sum of the geometric series $a + ar + ar^2 + ar^3 + \cdots$, notice that $rS = ar + ar^2 + ar^3 + \cdots$ and so $S - rS = a$, meaning that $S = \frac{a}{1-r}$. This manipulation is only valid, however, when $|r| < 1$; for other r the series is either infinite ($r \geq 1$) or nonconvergent (other r).

2.1.4 Joint Distributions

- We now treat more carefully the situation of having several discrete random variables defined on the same sample space, which will allow us to extend some of our analysis of conditional probability and independence into the random-variable setting.
  - If we have a collection of random variables $X_1, X_2, \ldots, X_n$, we can summarize all of the possible information about the behavior of these random variables simultaneously using a joint probability density distribution, which simply lists all the possible collections of values of these random variables together with their probabilities.
  - We often package these values together into a function:
- Definition: If $X_1, X_2, \ldots, X_n$ are discrete random variables on the sample space S, then the function $p_{X_1,X_2,\ldots,X_n}$ defined on ordered n-tuples of events $(a_1, \ldots, a_n) \in S$ such that $p_{X_1,X_2,\ldots,X_n}(a_1, a_2, \ldots, a_n) = P(X_1 = a_1,\ X_2 = a_2, \ldots, X_n = a_n)$ is called the joint probability density function of $X_1, X_2, \ldots, X_n$.
  - The joint probability density function simply measures the probability that the various random variables take particular values, for all combinations of possible values.
  - For the situation of two random variables X and Y, we can display the joint probability density function by tabulating all of the possible values of X and Y in a grid.
- Example: An unfair coin that comes up heads 2/3 of the time is flipped 3 times. Tabulate the joint probability density function of X and Y, where X is the random variable counting the total number of heads and Y is the random variable counting the longest run of tails. Then calculate (i) $P(X = Y = 1)$, (ii) $P(X = 3, Y = 1)$, (iii) $P(Y - X = 1)$, and (iv) $P(X + Y = 3)$.
  - Here is a table of all the possible outcomes, their probabilities, and the values of X and Y on each:

Outcome	Probability	Value of X	Value of Y
HHH	8/27	3	0
HHT, HTH, THH	4/27	2	1
HTT, TTH	2/27	1	2
THT	2/27	1	1
TTT	1/27	0	3

  - We can reorganize this information into the following table, organized by the values of X and Y:

$Y \setminus X$	0	1	2	3
0	0	0	0	8/27
1	0	2/27	12/27	0
2	0	4/27	0	0
3	1/27	0	0	0

  - By looking up the appropriate entries in the second table, we can then compute the probability of any combination of values of X and Y.
  - Thus, we have $P(X = Y = 1) = \boxed{2/27}$, $P(X = 3, Y = 1) = \boxed{0}$, $P(Y - X = 1) = \boxed{4/27}$, and $P(X + Y = 3) = \frac{1}{27} + \frac{4}{27} + \frac{12}{27} + \frac{8}{27} = \boxed{\frac{25}{27}}$.
- Example: The joint probability distribution for the random variables I and O counting the number of diners sitting inside and outside (respectively) at a small cafe in August is given below. Find (i) $P(I = 2, O = 3)$, (ii) $P(I = 1)$, (iii) $P(O = 2)$, (iv) $P(I + O \geq 4)$, (v) $P(I = O)$, and (vi) $P(|I - O| > 2)$. Also, find the individual probability distributions for I and O.

$I \setminus O$	0	1	2	3	4	5
0	0.03	0.10	0.12	0.18	0.10	0.08
1	0	0.01	0.07	0.10	0.08	0.06
2	0	0	0	0.01	0.03	0.02
3	0	0	0	0	0	0.01

  - We simply sum the appropriate entries in the table for each of the underlying results.

- This yields $P(I=2,O=3)=\boxed{0.01}$, $P(I=1)=0.01+0.07+0.10+0.08+0.06=\boxed{0.32}$, $P(O=2)=0.12+0.07=\boxed{0.19}$, $P(I+O\geq 4)=0.10+0.01+0.10+0.08+0.03+0.08+0.06+0.02+0.01=\boxed{0.49}$, $P(I=O)=0.03+0.01=\boxed{0.04}$, and $P(|I-O|>2)=0.18+0.10+0.08+0.08+0.06+0.02=\boxed{0.52}$.
- To find the probability distribution for O by itself, we simply sum over all of the corresponding entries in the table having the same value for O (i.e., down the columns):

O	0	1	2	3	4	5
Probability	0.03	0.11	0.19	0.29	0.21	0.17

- Likewise, to find the probability distribution for I, we sum across the rows:

I	Probability
0	0.61
1	0.32
2	0.06
3	0.01

- In general, just like in the second example above, we can easily recover the individual probability distributions for any of the random variables from their joint distribution by summing over the other variables:
- Proposition (Marginal Densities): If $p_{X,Y}(a,b)$ is the joint probability density function for the discrete random variables X and Y, then for any a and b we may compute the single-variable probability density functions for X and Y as $p_X(a)=\sum_y p_{X,Y}(a,y)$ and $p_Y(b)=\sum_x p_{X,Y}(x,b)$.
  - In general, a probability density function obtained by restricting a given probability distribution to a subset is called a marginal probability distribution. This proposition gives the procedure for computing the marginal probability distribution on the subsets $X=a$ and $Y=b$ (i.e., where one value of one of the random variables is fixed). The word "marginal" is used to evoke the idea of writing the row and column sums in the margins of the probability distribution table.
  - Proof: The first formula follows by observing that the event $\{E : X=a\}$ is the union over all real numbers y of the sets $\{E : X=a, Y=y\}$. Since these sets are disjoint (since the random variable Y can only take one value at a time), we can simply sum the corresponding probabilities by the probability axioms.
  - The second formula follows in the same way upon interchanging the roles of X and Y.
  - Remark: This result can be extended to an arbitrary number of variables using essentially the same argument. In general, if $X_1, X_2, \ldots, X_n$ are discrete random variables with joint pdf $p_{X_1,X_2,\ldots,X_n}(a_1,\ldots,a_n)$ then for any $1\leq k\leq n$ the joint pdf $p_{X_1,X_2,\ldots,X_k}(a_1,\ldots,a_k)$ is given by $p_{X_1,X_2,\ldots,X_k}(a_1,\ldots,a_k)=\sum_{x_{k+1},\ldots,x_n} p_{X_1,\ldots,X_n}(a_1,\ldots,a_k,x_{k+1},\ldots,x_n)$, with a similar formula holding for any subset of the X_i whose values are fixed.

2.1.5 Independence

- We would now like to use joint distributions to describe when two random variables are independent.
  - Intuitively, much as with independence of events in probability spaces, we would say that two random variables X and Y are independent when knowing the value of one gives no additional information about the value of the other.
  - Explicitly, this is the same as saying that $P(X=a|Y=b)=P(X=a)$ for every value of a and b, which (in turn) from our discussion of conditional probability, is equivalent to saying that $P(X=a,Y=b)=P(X=a)\cdot P(Y=b)$.
  - Equivalently, in the language of probability density functions, this says $p_{X,Y}(a,b)=p_X(a)\cdot p_Y(b)$. In other words, the probability that both $X=a$ and $Y=b$ is the product of the probabilities of those two separate events (namely, that $X=a$ and that $Y=b$).
  - In the same way as with conditional probabilities, we may easily extend this notion of independence to more than two random variables. The analogous condition would be that the discrete random variables $X_1, X_2, \ldots, X_n$ are independent when, for any subset $Y_1,\ldots,Y_k$ of the X_i, the joint distribution $p_{Y_1,\ldots,Y_k}(a_1,\ldots,a_k)$ is equal to the product of the individual distributions $p_{Y_1}(a_1)\cdot\cdots\cdot p_{Y_k}(a_k)$.

- However, because we may compute all of these joint distributions using the single joint distribution $p_{X_1,X_2,\ldots,X_n}(a_1,a_2,\ldots,a_n)$ (namely, by summing over all of the possible values of the random variables we are not considering), in fact all of these conditions follow from the single condition that $p_{X_1,X_2,\ldots,X_n}(a_1,a_2,\ldots,a_n) = p_{X_1}(a_1)\cdot p_{X_2}(a_2)\cdot\cdots\cdot p_{X_n}(a_n)$.

- Definition: We say that the discrete random variables $X_1, X_2, \ldots, X_n$ are collectively independent if the joint distribution $p_{X_1,X_2,\ldots,X_n}(a_1,a_2,\ldots,a_n) = p_{X_1}(a_1)\cdot p_{X_2}(a_2)\cdot\cdots\cdot p_{X_n}(a_n)$ for all real numbers $a_1, a_2, \ldots, a_n$.

- Example: If X and Y are random variables with the joint distribution displayed below, determine whether X and Y are independent.

$X \setminus Y$	0	1	2	3
1	0.12	0.18	0.24	0.06
3	0.02	0.03	0.04	0.01
5	0.06	0.09	0.12	0.03

  - We must first compute the probability distributions for X and Y, which we may do by summing the rows and columns, and then we must check whether $p_{X,Y}(a,b) = p_X(a)\cdot p_Y(b)$ for each (a,b) in the table.
  - We obtain the following:

$X \setminus Y$	0	1	2	3	Σ
1	0.12	0.18	0.24	0.06	0.6
3	0.02	0.03	0.04	0.01	0.1
5	0.06	0.09	0.12	0.03	0.3
Σ	0.2	0.3	0.4	0.1	

  - It is then easy to see that each entry is in fact the product of the corresponding row sum and column sum: this means $p_{X,Y}(a,b) = p_X(a)\cdot p_Y(b)$ for each (a,b), and so $\boxed{X \text{ and } Y \text{ are independent}}$.

- Example: A fair coin is flipped 3 times. If X is the total number of heads in the first two flips and Y is the total number of heads in the last two flips, determine whether X and Y are independent.

  - Intuitively, we would expect that these variables should not be independent, since both X and Y will be affected by the outcome of the second coin flip.
  - Indeed, we have $P(X=2, Y=0) = 0$ since $X=2$ requires the middle flip to be heads while $Y=0$ requires the middle flip to be tails.
  - However, $P(X=2) = \frac{1}{4}$ and $P(Y=0) = \frac{1}{4}$ also, and so $P(X=2)\cdot P(Y=0) = \frac{1}{16} \neq P(X=2, Y=0) = 0$. Thus, $\boxed{X \text{ and } Y \text{ are not independent}}$.
  - We could also just evaluate the full joint distribution of X and Y and then use the same analysis as we did in the previous example:

$X \setminus Y$	0	1	2	Σ
0	1/8	1/8	0	1/4
1	1/8	2/8	1/8	1/2
2	0	1/8	1/8	1/4
Σ	1/4	1/2	1/4	

  - We can see that there are four entries (the four corner entries) that are not equal to the product of the corresponding row and column sums, so any of these would yield an appropriate counterexample.

- Under the assumption of independence, we can say a few additional things about expected value and variance:

- Proposition (Variance and Independence): If X and Y are independent discrete random variables whose expected values exist, then $E(XY) = E(X)\cdot E(Y)$, and $\text{var}(X+Y) = \text{var}(X) + \text{var}(Y)$.

  - Note that we do require the hypothesis that X and Y be independent in order for the variance to be additive. The result is not true for non-independent random variables: an easy counterexample occurs for $X = Y$, in which case $\text{var}(X+Y) = \text{var}(2X) = 4\text{var}(X)$ which is not equal to $\text{var}(X)+\text{var}(Y) = 2\text{var}(X)$.
  - Proof: Suppose X and Y are independent discrete random variables, where X takes the values $x_1, x_2, \ldots$ with probabilities $p_1, p_2, \ldots$ and Y takes the values $y_1, y_2, \ldots$ with probabilities $q_1, q_2, \ldots$.

- By the assumption of independence, this means X takes the value x_i and Y takes the value y_j, so that XY takes the value $x_i y_j$, with probability $p_i q_j$. Therefore, $E(XY) = \sum_{i,j} p_i q_j x_i y_j = [\sum_i p_i x_i] \cdot \left[\sum_j q_j y_j\right] = E(X) \cdot E(Y)$, as claimed.
- For the second statement, we may use the result just shown to obtain $\text{var}(X+Y) = E[(X+Y)^2] - [E(X+Y)]^2 = E(X^2 + 2XY + Y^2) - [E(X) + E(Y)]^2 = E(X^2) + 2E(X)E(Y) + E(Y^2) - [E(X)^2 + 2E(X)E(Y) + E(Y)^2] = [E(X^2) - E(X)^2] + [E(Y^2) - E(Y)^2] = \text{var}(X) + \text{var}(Y)$, as claimed.

- Using these properties we can calculate the variance and standard deviation of a binomially-distributed random variable:

- Corollary (Binomial Variance): Let X be the binomially-distributed random variable representing the total number of successes obtained by performing n independent Bernoulli trials each of which has a success probability p. Then $E(X) = np$, $\text{var}(X) = np(1-p)$, and $\sigma(X) = \sqrt{np(1-p)}$.

  - Proof: We write $X = X_1 + X_2 + \cdots + X_n$ where X_i is the random variable representing success on the ith trial for $1 \leq i \leq n$.
  - Observe that $E(X_i) = (1-p) \cdot 0 + p \cdot 1 = p$ and $E(X_i^2) = (1-p) \cdot 0^2 + p \cdot 1^2 = p$, so $\text{var}(X_i) = E(X_i^2) - E(X_i)^2 = p(1-p)$.
  - Each of the X_i is a single Bernoulli trial and they are all collectively independent by assumption, so we have $E(X) = E(X_1) + \cdots + E(X_n) = np$ (as we previously calculated), and also $\text{var}(X) = \text{var}(X_1) + \cdots + \text{var}(X_n) = np(1-p)$ so that $\sigma(X) = \sqrt{\text{var}(X)} = \sqrt{np(1-p)}$ as claimed.

- Example: An unfair coin with a probability 2/3 of landing heads is flipped 450 times. Find the expected number and the standard deviation in the number of tails obtained.

  - Each individual flip can be thought of as a Bernoulli trial, with success corresponding to obtaining tails with probability $p = 1/3$, with a total of $n = 450$ trials.
  - Thus, from our results on the binomial distribution, the expected number of tails is $np = 450 \cdot 1/3 = \boxed{150}$ and the standard deviation is $\sqrt{np(1-p)} = \sqrt{450 \cdot 1/3 \cdot 2/3} = \boxed{10}$.

- Example: A car dealer has a probability 0.36 of selling a car to any individual customer, independently. If 25 customers patronize the dealership, determine the expected number and the standard deviation in the total number of cars sold.

  - Each individual customer can be thought of as a Bernoulli trial, with success corresponding to selling a car with probability $p = 0.36$, with a total of $n = 25$ trials.
  - Thus, from our results on the binomial distribution, the expected number of cars sold is $np = 25 \cdot 0.36 = \boxed{9}$ and the standard deviation is $\sqrt{np(1-p)} = \sqrt{25 \cdot 0.36 \cdot 0.64} = \boxed{2.4}$.

2.1.6 Covariance and Correlation

- Another important quantity, which is moderately related to independence, is known as the covariance:

- Definition: If X and Y are random variables whose expected values exist and are μ_X and μ_Y respectively, then the covariance of X and Y is defined as $\text{cov}(X,Y) = E[(X - \mu_X) \cdot (Y - \mu_Y)]$.

  - Roughly speaking, the covariance measures how well a change in the value of X (relative to its average value) correlates with a change in the value of Y (relative to its average value). If the covariance is large and positive, then when X increases, Y will tend also to increase, and inversely when X decreases, Y will tend also to decrease.
  - Inversely, if the covariance is large and negative, then when X increases Y will tend to decrease, and when X decreases Y will tend to increase. When the covariance is near zero, then a change in the value of X does not tend to correspond to any particular type of change in the value of Y.

- Example: Find the covariance of the random variables X and Y with joint distribution below.

$X \setminus Y$	0	10
0	0.4	0.1
10	0.2	0.3

  - We can compute $\mu_X = 5$, $\mu_Y = 4$, and so $\text{cov}(X,Y) = 0.4 \cdot (-5) \cdot (-4) + 0.1 \cdot (-5) \cdot (6) + 0.2 \cdot (5) \cdot (-4) + 0.3 \cdot (5) \cdot (06) = 10$.
  - We can see that when X is 0, Y is more likely to be 0 than 10, and when X is 10, Y is more likely to be 10 than 0.

- The definition of covariance is somewhat complicated to compute, even in the simple example above. We can give a simpler formula that is more amenable to hand calculations[2]:

- Proposition (Covariance Formula): For any discrete random variables X and Y whose expected values exist, we have $\text{cov}(X,Y) = E(XY) - E(X)E(Y)$.

  - Proof: If the expected values of X and Y are μ_X and μ_Y respectively, then by linearity property of expectation we have $\text{cov}(X,Y) = E(XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y) = E(XY) - \mu_X E(Y) - \mu_Y E(X) + \mu_X \mu_Y = E(XY) - \mu_X \mu_Y = E(XY) - E(X)E(Y)$, as claimed.
  - From this formula, we can see that if X and Y are independent, then their covariance is equal to zero. Do note, however, that the converse is *not* true: if the covariance is zero, it does not imply that X and Y are independent.

- Example: Find the covariance of the random variables X and Y with joint distribution below.

$X \setminus Y$	0	5	10
0	0.3	0.2	0
10	0.4	0	0.1

  - We compute $E(X) = 0.5 \cdot 0 + 0.5 \cdot 10 = 5$, $E(Y) = 0.7 \cdot 0 + 0.2 \cdot 5 + 0.1 \cdot 10 = 2$, and $E(XY) = 0.3 \cdot 0 + 0.2 \cdot 0 + 0 \cdot 0 + 0.4 \cdot 0 + 0 \cdot 50 + 0.1 \cdot 100 = 10$.
  - Therefore, we see $\text{cov}(X,Y) = E(XY) - E(X)E(Y) = 10 - 5 \cdot 2 = \boxed{0}$.
  - Remark: Notice that although their covariance is zero, X and Y are not independent, since for example $P(X = 0) = 0.5$, $P(Y = 0) = 0.7$, but $P(X = 0, Y = 0) = 0.3$ rather than $P(X = 0) \cdot P(Y = 0) = 0.5 \cdot 0.7 = 0.35$.

- We have various algebraic properties involving the covariance:

- Proposition (Properties of Covariance): If X, Y, Z are discrete random variables whose expected values exist, then for any a and b we have $\text{cov}(X,X) = \text{var}(X)$, $\text{cov}(Y,X) = \text{cov}(X,Y)$, $\text{cov}(X+Y,Z) = \text{cov}(X,Z) + \text{cov}(Y,Z)$, $\text{cov}(aX+b,Y) = a \cdot \text{cov}(X,Y)$, and $\text{var}(X+Y) = \text{var}(X) + \text{var}(Y) + 2\text{cov}(X,Y)$.

  - Proof: By definition of the variance, we have $\text{cov}(X,X) = E[(X - \mu_X)^2] = \text{var}(X)$.
  - The second property follows from observing that $(X - \mu_X)(Y - \mu_Y) = (Y - \mu_Y)(X - \mu_X)$, so the corresponding expected values are also equal, and the third property follows in the same way from $(X + Y - \mu_X - \mu_Y)(Z - \mu_Z) = (X + \mu_X)(Z - \mu_Z) + (Y - \mu_Y)(Z - \mu_Z)$.
  - The third property follows by noting that $E(aX + b) = a\mu_X + b$ so that $\text{cov}(aX+b,Y) = E[(aX + b - a\mu_X - b)(Y - \mu_Y) = a \cdot E[(X - \mu_X)(Y - \mu_Y)] = a \cdot \text{cov}(X,Y)$.
  - The last property follows from noting $E(X+Y) = \mu_X + \mu_Y$ so that $\text{var}(X+Y) = E[(X+Y-\mu_X-\mu_Y)^2] = E[(X-\mu_X)^2 + 2(X-\mu_X)(Y-\mu_Y) + (Y-\mu_Y)^2] = E[(X-\mu_X)^2] + 2E[(X-\mu_X)(Y-\mu_Y)] + E[(Y-\mu_Y)^2] = \text{var}(X) + 2\text{cov}(X,Y) + \text{var}(Y)$.

- Per the properties above, the covariance scales linearly with both of the random variables X and Y. In some situations, we prefer to have a "normalized" measure of covariance, which we can obtain by dividing the covariance by the product of the standard deviations:

[2] We will remark that in computational implementation, this formula suffers from issues of numerical instability, since $E(XY)$ and $E(X)E(Y)$ may both be quite large even when the covariance is very small.

- Definition: If X and Y are discrete random variables whose variances exist and are nonzero, the (Pearson) correlation between X and Y is defined as $\text{corr}(X,Y) = \dfrac{\text{cov}(X,Y)}{\sigma(X)\sigma(Y)}$.
  - It is easy to see that the correlation (unlike the covariance) remains unchanged upon scaling and translating the variables: $\text{corr}(aX+b, cY+d) = \text{corr}(X,Y)$.
  - The correlation between two random variables always lies between -1 and 1 inclusive[3], with values near 1 indicating that the variables tend to increase linearly together and decrease linearly together (which agrees with the intuitive notion of two variables being strongly positively correlated) and with values near -1 indicating that the variables tend to increase linearly as the other decreases linearly (which agrees with the intuitive notion of two variables being strongly negatively correlated).
  - A correlation of zero is the same as a covariance of zero. Do note, however (as we saw above) that a correlation of zero is not equivalent to the variables being independent!
  - Remark: This correlation coefficient is also known as the linear regression correlation coefficient, since it represents the closeness by which a linear function can describe the relationship between X and Y. A correlation coefficient near 1 indicates that there is a linear function with a positive slope that models the relationship closely, while a correlation coefficient near -1 indicates that there is a linear function with a negative slope that models the relationship closely. A correlation coefficient near 0 indicates that there is no linear function that models the relationship closely (but of course, this need not mean that the variables are unrelated, merely that any relationship is not linear).
  - Remark (for students who like linear algebra): The covariance is a particular example of an inner product on the vector space of discrete random variables, and the correlation can be interpreted as the cosine of the generalized angle between the associated vectors, giving another reason why its value ranges from -1 to 1.

- Example: A fair coin is flipped 3 times. If X is the total number of heads in the first two flips and Y is the total number of heads in the last two flips, find the covariance and correlation between X and Y.
  - We previously found the joint distribution for X and Y:

$X \setminus Y$	0	1	2
0	1/8	1/8	0
1	1/8	2/8	1/8
2	0	1/8	1/8

  - We can then find $E(X) = E(Y) = 1$ and $E(XY) = \dfrac{3}{8}\cdot 0 + \dfrac{2}{8}\cdot 1 + \dfrac{2}{8}\cdot 2 + \dfrac{1}{8}\cdot 4 = \dfrac{5}{4}$, and so $\text{cov}(X,Y) = E(XY) - E(X)E(Y) = \boxed{\dfrac{1}{4}}$.
  - Also, $E(X^2) = E(Y^2) = \dfrac{2}{8}\cdot 0^2 + \dfrac{4}{8}\cdot 1^2 + \dfrac{2}{8}\cdot 2^2 = \dfrac{3}{2}$, and so $\sigma(X) = \sigma(Y) = \sqrt{\dfrac{3}{2} - 1^2} = \sqrt{\dfrac{1}{2}}$. Thus, $\text{corr}(X,Y) = \dfrac{\text{cov}(X,Y)}{\sigma(X)\sigma(Y)} = \boxed{\dfrac{1}{2}}$.
  - We can see that there is a positive correlation of moderate size between X and Y, which is intuitively reasonable because X and Y each count the number of heads from one independent coin flip and one shared coin flip.

- Example: Suppose X and Y are discrete random variables such that $E(X) = 5$, $\sigma(X) = 1$, $E(Y) = 3$, $\sigma(Y) = 2$, and $E(XY) = 16$. Find $\text{cov}(X,Y)$, $\text{corr}(X,Y)$, $\text{cov}(X+Y, X-Y)$, and $\text{corr}(X+Y, X-Y)$.
  - We have $\text{cov}(X,Y) = E(XY) - E(X)E(Y) = 16 - 5\cdot 3 = \boxed{1}$ and $\text{corr}(X,Y) = \dfrac{\text{cov}(X,Y)}{\sigma(X)\sigma(Y)} = \boxed{\dfrac{1}{2}}$.
  - Also, using property (3) of covariance, we have $\text{cov}(X+Y, X-Y) = \text{cov}(X,X) + \text{cov}(Y,X) - \text{cov}(X,Y) - \text{cov}(Y,Y) = \text{var}(X) - \text{var}(Y) = 1^2 - 2^2 = \boxed{-3}$.

[3] To see that the correlation is always between -1 and 1, observe that $\text{var}(X+tY) = \text{var}(X) + t^2\text{var}(Y) + 2t\text{cov}(X,Y)$ is always nonnegative for any constant t. Setting $t = -\text{cov}(X,Y)/\text{var}(Y)$ and simplifying yields $\text{var}(X) - \text{cov}(X,Y)^2/\text{var}(Y) \geq 0$, so rearranging and taking the square root yields $|\text{cov}(X,Y)| \leq \sigma(X)\sigma(Y)$, whence $-1 \leq \text{corr}(X,Y) \leq 1$.

- To find the correlation we need $\sigma(X+Y) = \sqrt{\text{var}(X+Y)} = \sqrt{\text{var}(X)+\text{var}(Y)+2\text{cov}(X,Y)} = \sqrt{1^2+2^2+2\cdot 1} = \sqrt{7}$ and $\sigma(X-Y) = \sqrt{\text{var}(X-Y)} = \sqrt{\text{var}(X)+\text{var}(Y)-2\text{cov}(X,Y)} = \sqrt{1^2+2^2-2\cdot 1} = \sqrt{3}$.
- Thus, the correlation is given by $\text{corr}(X+Y, X-Y) = \dfrac{\text{cov}(X+Y, X-Y)}{\sigma(X+Y)\sigma(X-Y)} = \boxed{\dfrac{-3}{\sqrt{21}}}$.

2.2 Continuous Random Variables

- Another important class of random variables consists of the random variables whose underlying sample space is the entire real line.
  - Because there are uncountably many possible outcomes, to evaluate probabilities of events we cannot simply sum over the outcomes that make them up.
  - Instead, we must use the continuous analogue of summation, namely, integration.
  - All of the results we will discuss are very similar to the corresponding ones for discrete random variables, with the only added complexity being the requirement to evaluate integrals.

2.2.1 Probability Density Functions, Cumulative Distribution Functions

- Definition: A continuous probability density function $p(x)$ is a piecewise-continuous, nonnegative real-valued function such that $\int_{-\infty}^{\infty} p(x)\,dx = 1$.
  - Note that the integral $\int_{-\infty}^{\infty} p(x)\,dx$ is in general improper. However, since $p(x)$ is by assumption nonnegative, then the value of the integral is always well-defined (although it may be ∞).
  - Example: The function $p(x) = \begin{cases} 1/4 & \text{for } 0 \le x \le 4 \\ 0 & \text{for other } x \end{cases}$ is a continuous probability density function, since the two components of $p(x)$ are both continuous and nonnegative, and $\int_{-\infty}^{\infty} p(x)\,dx = \int_0^4 1/4\,dx = 1$.
  - Example: The function $q(x) = \begin{cases} 2x & \text{for } 0 \le x \le 1 \\ 0 & \text{for other } x \end{cases}$ is a continuous probability density function, since the two components of $q(x)$ are both continuous and nonnegative, and $\int_{-\infty}^{\infty} q(x)\,dx = \int_0^1 2x\,dx = 1$.
  - Example: The function $e(x) = \begin{cases} e^{-x} & \text{for } x \ge 0 \\ 0 & \text{for } x < 0 \end{cases}$ is a continuous probability density function, since the two components of $e(x)$ are both continuous and nonnegative, and $\int_{-\infty}^{\infty} e(x)\,dx = \int_0^{\infty} e^{-x}\,dx = 1$.
- Definition: We say that X is a continuous random variable if there exists a continuous probability density function $p(x)$ such that for any interval I on the real line, we have $P(X \in I) = \int_I p(x)\,dx$.
  - In other words, probabilities for continuous random variables are computed via integrating the probability density function on the appropriate interval.
  - Note that if X is a continuous random variable, then (no matter what the probability density function is), the value of $P(X = a)$ is zero for any value of a, since $P(X = a) = \int_a^a p(x)\,dx = 0$. This means that the probability that X will attain any specific value a is equal to zero.
- Example: If X is the continuous random variable whose probability density function is $p(x) = \begin{cases} x/8 & \text{for } 0 \le x \le 4 \\ 0 & \text{for other } x \end{cases}$, find (i) $P(1 \le X \le 3)$, (ii) $P(X \le 2)$, (iii) $P(X \ge 5)$, (iv) $P(-2 \le X \le 3)$, and (v) $P(X = 2)$.
  - When we set up the integrals, we must remember to break the range of integration up (if needed) so that we are integrating the correct component of $p(x)$ on the correct interval.

- For example, to verify that $p(x)$ is a probability density function, we would compute $\int_{-\infty}^{\infty} p(x)\,dx = \int_{-\infty}^{0} 0\,dx + \int_0^4 \frac{x}{8}\,dx + \int_4^{\infty} 0\,dx = 0 + \frac{1}{16}x^2\Big|_{x=0}^{4} + 0 = 1$, as required.
- For (i), we have $P(1 \le X \le 3) = \int_1^3 p(x)\,dx = \int_1^3 \frac{x}{8}\,dx = \frac{1}{16}x^2\Big|_{x=1}^{3} = \boxed{\frac{1}{2}}$.
- For (ii), we have $P(X \le 2) = \int_{-\infty}^{2} p(x)\,dx = \int_{-\infty}^{0} 0\,dx + \int_0^2 \frac{x}{8}\,dx = \frac{1}{16}x^2\Big|_{x=0}^{2} = \boxed{\frac{1}{4}}$.
- For (iii), we have $P(X \ge 5) = \int_5^{\infty} p(x)\,dx = \int_5^{\infty} 0\,dx = \boxed{0}$.
- For (iv), we have $P(-2 \le X \le 3) = \int_{-2}^{3} p(x)\,dx = \int_{-2}^{0} 0\,dx + \int_0^3 \frac{x}{8}\,dx = \frac{1}{16}x^2\Big|_{x=0}^{3} = \boxed{\frac{9}{16}}$.
- For (v), we have $P(X = 2) = \int_2^2 p(x)\,dx = \int_2^2 \frac{x}{8}\,dx = \frac{1}{16}x^2\Big|_{x=2}^{2} = \boxed{0}$.

- As we saw in the example above, if we are evaluating integrals, we may ignore intervals on which $p(x) = 0$, since their contribution to the integrals will always be 0.
- A useful function related to the probability density function of a continuous random variable is the cumulative distribution function[4], which measures the total probability that the continuous random variable takes a value $\le x$:
- Definition: If X is a continuous random variable with probability density function $p(x)$, its cumulative distribution function (cdf) $c(x)$ is defined as $c(x) = \int_{-\infty}^{x} p(t)\,dt$ for each real value of x. Then $P(X \le a) = c(a)$ and $P(X \ge a) = 1 - c(a)$ for every a, and $P(a \le X \le b) = c(b) - c(a)$ for every a and b.
  - By the fundamental theorem of calculus, we have $c'(x) = p(x)$ for every x, so we may freely convert back and forth between the probability density function and the cumulative distribution function via differentiation and integration.
  - Since $p(x)$ is nonnegative, if $a \le b$ then $c(a) \le c(b)$, and since $\int_{-\infty}^{\infty} p(x)\,dx = 1$, we have $\lim_{x\to\infty} c(x) = 1$, and $\lim_{x\to-\infty} c(x) = 0$.
  - Example: For the random variable with probability density function $p(x) = \begin{cases} 1/4 & \text{for } 0 \le x \le 4 \\ 0 & \text{for other } x \end{cases}$, the cumulative distribution function is $c(x) = \begin{cases} 0 & \text{for } x \le 0 \\ x/4 & \text{for } 0 \le x \le 4 \\ 1 & \text{for } x \ge 4 \end{cases}$.
  - Example: For the random variable with probability density function $q(x) = \begin{cases} 0 & \text{for } x < 1 \\ 1/x^2 & \text{for } x \ge 1 \end{cases}$, the cumulative distribution function is $c(x) = \begin{cases} 0 & \text{for } x < 1 \\ 1 - 1/x & \text{for } x \ge 1 \end{cases}$.
- Example: The probability density function for a continuous random variable X has the form $p(x) = \frac{a}{x^{3/2}}$ for $1 \le x \le 9$ and 0 elsewhere. Find (i) the value of a, (ii) the probability that $1 \le X \le 4$, and (iii) the cumulative distribution function for X.
  - The value of a is determined by the fact that the integral of $p(x)$ over its full domain must equal 1, which is to say that $\int_1^9 \frac{a}{x^{3/2}}\,dx = 1$. Since $\int_1^9 \frac{a}{x^{3/2}}\,dx = -2ax^{-1/2}\Big|_{x=1}^{9} = \frac{4}{3}a$, we must have $a = \boxed{\frac{3}{4}}$.

[4] In fact, one may also define the cumulative distribution function for a discrete random variable; the typical definition is $c(x) = \sum_{n \le x} p(n)$. However, for various reasons in certain cases one may prefer to sum only over all values less than x, rather than less than or equal to x. This issue does not arise when working with continuous random variables, since the probability of obtaining exactly the value x is always zero, so we could in fact use either definition of the cumulative distribution function (as giving the probability of a value $\le x$ or as giving the probability of a value $< x$) since they are the same. For this reason, we will work with cumulative distribution functions only in the context of continuous random variables.

- Next, the probability that $1 \le X \le 4$ is given by $\int_1^4 p(x)\,dx = \int_1^4 \frac{3}{4}x^{-3/2}\,dx = \boxed{\frac{3}{4}}$.
- Finally, the cumulative distribution function is $c(x) = \int_{-\infty}^{x} p(t)\,dt$, yielding $c(x) = \begin{cases} 0 & \text{for } x \le 1 \\ \frac{3}{2}(1 - x^{-1/2}) & \text{for } 1 \le x \le 9 \\ 1 & \text{for } x \ge 9 \end{cases}$.
- Remark: If we had computed the cumulative distribution function first, we could alternatively have calculated $P(1 \le X \le 4) = c(4) - c(1) = \frac{3}{4} - 0 = \boxed{\frac{3}{4}}$.

- A simple class of continuous random variables are those whose probability density functions are constant on an interval $[a, b]$ and zero elsewhere: such random variables are uniformly distributed on the interval $[a, b]$.
  - From the requirement that the integral of the probability density function equals 1, we can see that the probability density function for a uniformly-distributed random variable on $[a, b]$ is given by $p(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \le x \le b \\ 0 & \text{for other } x \end{cases}$ with cumulative distribution function $c(x) = \begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } a \le x \le b \\ 1 & \text{for } x > b \end{cases}$.
  - Using this description we can easily compute probabilities for uniformly-distributed random variables.

- Example: The high temperature in a certain city in May is uniformly distributed between 70°F and 90°F. Find the probabilities that (i) the temperature is between 82°F and 85°F, (ii) the temperature is less than 75°F, and (iii) the temperature is greater than 82°F.
  - From our description of uniformly-distributed random variables, we can see that the probability density function for the temperature is $p(x) = \begin{cases} 1/20 & \text{for } 70 \le x \le 90 \\ 0 & \text{for other } x \end{cases}$.
  - Then for (i), the probability is $\int_{82}^{85} p(x)\,dx = \int_{82}^{85} 1/20\,dx = \boxed{3/20}$.
  - For (ii), the probability is $\int_{-\infty}^{75} p(x)\,dx = \int_{-\infty}^{70} 0\,dx + \int_{70}^{75} 1/20\,dx = \boxed{1/4}$.
  - For (iii), the probability is $\int_{82}^{\infty} p(x)\,dx = \int_{82}^{90} 1/20\,dx + \int_{90}^{\infty} 0\,dx = \boxed{2/5}$.

- Another important class of distributions are the exponential distributions, as we discuss later:

- Definition: The exponential distribution with parameter $\lambda > 0$ is the continuous random variable with probability density function $p(x) = \lambda e^{-\lambda x}$ for $x \ge 0$, and is 0 for negative x.
  - We can easily compute the cumulative distribution function as $c(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 - e^{-\lambda x} & \text{for } x \ge 0 \end{cases}$.

- Example: If X is exponentially distributed with parameter $\lambda = 1/2$, find (i) $P(X < 1)$, (ii) $P(X \ge 3)$, and (iii) $P(1 \le X \le 2)$.
  - Using the cumulative distribution function, we have $P(X < 1) = c(1) = \boxed{1 - e^{-1/2}} \approx 0.3935$, $P(X \ge 3) = 1 - c(3) = \boxed{e^{-3/2}} \approx 0.2231$, and $P(1 \le X \le 2) = c(2) - c(1) = \boxed{e^{-1/2} - e^{-1}} \approx 0.2387$.

2.2.2 Expected Value, Variance, Standard Deviation

- In much the same way as we defined expected value, variance, and standard deviation for discrete random variables, we can also define these quantities for continuous random variables.

- Definition: If X is a continuous random variable with probability density function $p(x)$, we define the expected value as $E(X) = \int_{-\infty}^{\infty} x\,p(x)\,dx$, presuming that the integral converges.

- This formula is the continuous analogue of the expected value formula for a discrete random variable.

- Example: Find the expected values of the continuous random variables X, Y and Z with respective probability density functions $p(x) = \begin{cases} 6(x-x^2) & \text{for } 0 \le x \le 1 \\ 0 & \text{for other } x \end{cases}$, $q(x) = \begin{cases} 2x & \text{for } 0 \le x \le 1 \\ 0 & \text{for other } x \end{cases}$, and $e(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \ge 0 \\ 0 & \text{for } x < 0 \end{cases}$.
  - By definition, we have $E(X) = \int_{-\infty}^{\infty} x\, p(x)\, dx = \int_0^1 x \cdot 6(x-x^2)\, dx = (2x^3 - \frac{3}{2}x^4)\big|_{x=0}^1 = \boxed{\frac{1}{2}}$.
  - Likewise, $E(Y) = \int_{-\infty}^{\infty} x\, q(x)\, dx = \int_0^1 x \cdot 2x\, dx = (\frac{2}{3}x^3)\big|_{x=0}^1 = \boxed{\frac{2}{3}}$.
  - Also, $E(Z) = \int_{-\infty}^{\infty} x\, e(x)\, dx = \int_0^{\infty} \lambda x\, e^{-\lambda x}\, dx = (-xe^{-\lambda x} - e^{-\lambda x}/\lambda)\big|_{x=0}^{\infty} = \boxed{1/\lambda}$.

- Much like in the case of discrete random variables, the expected value of a continuous random variable can be infinite or even undefined:

- Example: Find the expected values of the continuous random variables X and Y with respective probability density functions $p(x) = \begin{cases} 1/x^2 & \text{for } x \ge 1 \\ 0 & \text{for other } x \end{cases}$ and $q(x) = \dfrac{1}{\pi(1+x^2)}$.
  - By definition, we have $E(X) = \int_{-\infty}^{\infty} x\, p(x)\, dx = \int_1^{\infty} x \cdot \dfrac{1}{x^2}\, dx = \ln(x)\big|_{x=1}^{\infty} = \boxed{\infty}$.
  - For Y, we would have $E(Y) = \int_{-\infty}^{\infty} x \cdot \dfrac{1}{\pi(1+x^2)}\, dx$. By using a substitution, we can see that an antiderivative of $\dfrac{x}{\pi(1+x^2)}$ is $\dfrac{1}{2\pi}\ln(1+x^2)$.
  - To evaluate this improper integral, we can split the range of integration at 0 to obtain $\int_0^{\infty} x \cdot \dfrac{1}{\pi(1+x^2)}\, dx = \frac{1}{2\pi}\ln(1+x^2)\big|_{x=0}^{\infty} = \infty$, while $\int_{-\infty}^0 x \cdot \dfrac{1}{\pi(1+x^2)}\, dx = \frac{1}{2\pi}\ln(1+x^2)\big|_{x=-\infty}^{0} = -\infty$.
  - But this tells us that the original integral $E(Y) = \int_{-\infty}^{\infty} x \cdot \dfrac{1}{\pi(1+x^2)}\, dx$ is $\infty - \infty$, which is not defined.
  - Remark: Observe that the probability density function for Y is symmetric about $x = 0$, which would suggest that the expected value is 0. However, this is not the case! This particular probability density function is called the Cauchy distribution, and yields counterexamples to many statements that might seem to be intuitively obvious.

- We would like to define the variance and standard deviation in the same way as for discrete random variables; namely, as $\text{var}(X) = E[(X-\mu)^2]$ where $\mu = E(X)$ is the expected value of X, or equivalently as $\text{var}(X) = E(X^2) - [E(X)]^2$.
  - However, we need to know how to compute the expected value of a function of X.
  - To illustrate the difficulty, suppose X is uniformly distributed on $[0, 2]$. Then we can easily compute $E(X) = 1$, but it is not so obvious how to find $E(X^2)$.
  - Since X^2 is a random variable, it has some probability density function, which we can try to calculate by using the cumulative distribution function.
  - Explicitly, since $X^2 \le a$ is equivalent to $X \le \sqrt{a}$ (at least for X nonnegative), this means that $c_{X^2}(a) = c_X(\sqrt{a})$ for $0 \le a \le 4$.
  - In terms of the probability density functions, this says $\int_0^a p_{X^2}(x)\, dx = \int_0^{\sqrt{a}} p_X(x)\, dx = \int_0^{\sqrt{a}} \dfrac{1}{2}\, dx = \dfrac{\sqrt{a}}{2}$.
  - Then differentiating both sides yields $p_{X^2}(a) = \dfrac{d}{da}\left[\dfrac{\sqrt{a}}{2}\right] = \dfrac{1}{4}a^{-1/2}$ for $0 \le a \le 4$.
  - Thus, we deduce that $p_{X^2}(x) = \begin{cases} x^{-1/2}/4 & \text{for } 0 \le x \le 4 \\ 0 & \text{for other } x \end{cases}$.

- We can then evaluate $E(X^2) = \int_0^4 x \cdot \frac{1}{4}x^{-1/2}\,dx = \frac{1}{6}x^{3/2}\Big|_{x=0}^{4} = \frac{4}{3}$.

- This procedure we used to compute $E(X^2)$ in the example above is quite complicated and difficult. Fortunately, there is a simpler way to compute the expected value of a function of a continuous random variable:

- Proposition (Expected Value of Functions of X): If X is a continuous random variable with probability density function $p(x)$, and $g(x)$ is any piecewise-continuous function, then the expected value of $g(X)$ is $E[g(X)] = \int_{-\infty}^{\infty} g(x)\,p(x)\,dx$.

  - To explain the intuitive reason for this formula, consider instead the case of a discrete random variable X taking values $x_1, x_2, \ldots$ with probabilities $p_1, p_2, \ldots$: then $E[g(X)] = g(x_1)p_1 + g(x_2)p_2 + \cdots = \sum_i g(x_i)p_i$. The continuous analogue of this formula replaces the summation with the corresponding integral, yielding precisely the formula above[5].
  - Proof (special case): Suppose g is increasing and has an inverse function g^{-1}. Then $g(x) \le a$ is equivalent to $x \le g^{-1}(a)$, so by the argument given above, we see that $c_{g(X)}(a) = c_X(g^{-1}(a))$.
  - Differentiating yields $p_{g(X)}(a) = p(g^{-1}(a)) \cdot \dfrac{1}{g'(g^{-1}(a))}$, so $E[g(X)] = \int_{-\infty}^{\infty} x \cdot (g^{-1}(x)) \cdot \dfrac{1}{g'(g^{-1}(x))}\,dx$.
  - Making the substitution $u = g^{-1}(x)$, so that $x = g(u)$ and $dx = g'(u)du$, in the integral and simplifying yields $E[g(X)] = \int_{-\infty}^{\infty} g(u) \cdot p(u)\,du$, as claimed.
  - The argument in the general case is similar.

- Corollary (Linearity of Expected Value): If X and Y are continuous random variables whose expected values are defined, and a and b are any real numbers, then $E(aX+b) = a \cdot E(X) + b$ and $E(X+Y) = E(X) + E(Y)$.

  - Proof (first statement): If X has probability density function $p(x)$, then $E(aX+b) = \int_{-\infty}^{\infty}(ax+b) \cdot p(x)\,dx = a\int_{-\infty}^{\infty} x \cdot p(x) + b\int_{-\infty}^{\infty} p(x)\,dx = a \cdot E(X) + b$.
  - The second statement can be deduced by first finding the probability density function of $X+Y$ and then using an argument similar to that of the proposition above.

- If the expected value of a continuous random variable is defined and finite, we can define the variance and standard deviation in the same way as for discrete random variables:

- Definition: If X is a continuous random variable whose expected value μ exists and is finite, the variance $\mathrm{var}(X)$ is defined as $\mathrm{var}(X) = E[(X-\mu)^2] = E(X^2) - E(X)^2$, and the standard deviation is $\sigma(X) = \sqrt{\mathrm{var}(X)}$.

  - The equality $E[(X-\mu)^2] = E(X^2) - E(X)^2$ follows for continuous random variables by the same argument used for discrete random variables.

- Example: Find the variance and standard deviation for the continuous random variables X, Y, and Z with density functions $p(x) = \begin{cases} 6(x-x^2) & \text{for } 0 \le x \le 1 \\ 0 & \text{for other } x \end{cases}$, $q(x) = \begin{cases} 2x & \text{for } 0 \le x \le 1 \\ 0 & \text{for other } x \end{cases}$, $e(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \ge 0 \\ 0 & \text{for } x < 0 \end{cases}$.

  - For X, we have $E(X) = \int_0^1 x \cdot 6(x-x^2)\,dx = (2x^3 - \frac{3}{2}x^4)\big|_{x=0}^{1} = \frac{1}{2}$ and $E(X^2) = \int_0^1 x^2 \cdot 6(x-x^2)\,dx = (\frac{3}{2}x^4 - \frac{6}{5}x^5)\big|_{x=0}^{1} = \frac{3}{10}$.
  - Thus $\mathrm{var}(X) = E(X^2) - [E(X)]^2 = \frac{3}{10} - \frac{1}{4} = \boxed{\frac{1}{20}}$ and $\sigma(X) = \boxed{\sqrt{\frac{1}{20}}}$.
  - For Y, we have $E(Y) = \int_0^1 x \cdot 2x\,dx = (\frac{2}{3}x^3)\big|_{x=0}^{1} = \frac{2}{3}$ and $E(Y^2) = \int_0^1 x^2 \cdot 2x\,dx = (\frac{1}{2}x^4)\big|_{x=0}^{1} = \frac{1}{2}$.
  - Thus $\mathrm{var}(Y) = E(Y^2) - [E(Y)]^2 = \frac{2}{3} - \frac{1}{4} = \boxed{\frac{5}{12}}$ and $\sigma(Y) = \boxed{\sqrt{\frac{5}{12}}}$.

[5] In fact, another way to prove the continuous version is to approximate the continuous random variable with a discrete one, and observe that the corresponding expectation for the discrete approximation is a Riemann sum for the integral above.

- For Z, $E(Z) = \int_0^\infty \lambda x\, e^{-\lambda x}\, dx = 1/\lambda$, $E(Z^2) = \int_0^\infty x^2 \lambda\, e^{-\lambda x}\, dx = -(x^2 + 2x/\lambda + 2/\lambda^2)e^{-x}\big|_{x=0}^{\infty} = 2/\lambda^2$.
- Thus $\text{var}(Z) = E(Z^2) - [E(Z)]^2 = 2/\lambda^2 - 1/\lambda^2 = \boxed{1/\lambda^2}$ and $\sigma(Z) = \boxed{1/\lambda}$.
- Remark: The last example shows that the exponential distribution with parameter λ has expected value and standard deviation both equal to $1/\lambda$.

- Proposition (Properties of Variance): If X is a continuous random variable and a and b are any real numbers, then $\text{var}(aX + b) = a^2 \cdot \text{var}(X)$ and $\sigma(aX + b) = |a|\, \sigma(X)$.
  - Proof: The proof follows in the same way as for discrete random variables.

- Example: If X is a continuous random variable with expected value 4 and standard deviation 3, what are the expected value and standard deviation of $3X - 5$?
  - From the properties given above, we have $E(3X-5) = 3E(X)-5 = \boxed{7}$, and $\sigma(3X-5) = |3|\, \sigma(X) = \boxed{9}$.

- If a random variable X is measured in a particular unit (e.g., dollars), then its standard deviation is also measured in the same units and measures the approximate spread of X around its expected value.
  - Intuitively, we should expect that "most" of the distribution for X should be concentrated near its average, provided that we measure in increments of the standard deviation.
  - We can make this statement more precise, as follows:

- Theorem (Chebyshev's Inequality): If X is a random variable with expected value μ and standard deviation σ, then $P(|X - \mu| \geq k\sigma) \leq 1/k^2$ for any positive real number k.
  - In words, Chebyshev's inequality says that the probability that X takes a value at least k standard deviations away from its mean is at most $1/k^2$. (The statement thus only has content for $k > 1$.)
  - Proof: First we show that if Y is a nonnegative random variable and a is any positive real number, it is true that $P(Y \geq a) \leq E(Y)/a$. (This is a result known as Markov's inequality.)
  - For this, if we break the expected value calculation into the two pieces where $0 \leq Y < a$ and $Y \geq a$, we can see that $E(Y) = P(0 \leq Y < a) \cdot E(Y|0 \leq Y < a) + P(Y \geq a) \cdot E(Y|Y \geq a) \geq P(Y < a) \cdot 0 + P(Y \geq a) \cdot a$, since the expected value of Y when $0 \leq Y < a$ is at least 0, and the expected value of Y when $Y \geq a$ is at least a.
  - Thus, since $E(Y) \geq P(Y \geq a) \cdot a$, that means $P(Y \geq a) \leq E(Y)/a$.
  - Now, we apply this result to the random variable $Y = (X - \mu)^2$ and $a = k^2\sigma^2$ (note that $Y \geq 0$ and $a > 0$ here so the result applies): it says $P[(X - \mu)^2 \geq k^2\sigma^2) \leq E[(X - \mu)^2]/(k^2\sigma^2)$.
  - But since $E[(X - \mu)^2] = \sigma^2$, we obtain $P[(X - \mu)^2 \geq k^2\sigma^2) \leq \sigma^2/(k^2\sigma^2) = 1/k^2$.
  - Since $(X - \mu)^2 \geq k^2\sigma^2$ is equivalent to $|X - \mu| \geq k\sigma$, we have obtained the desired result.
  - Remark: Because we only used properties of expected value in the proof, Chebyshev's inequality applies to any random variable, discrete or continuous.

- We can interpret Chebyshev's inequality as giving a precise bound on how far away the probability distribution of the random variable X can be concentrated in terms of the standard deviation.
  - Setting $k = 2$, for example, says that the value of X can be 2 or more standard deviations away from the mean at most 1/4 of the time.
  - Similarly, setting $k = 3$ implies that the value can be 3 or more standard deviations away from the mean at most 1/9 of the time.
  - For almost all distributions, Chebyshev's inequality is very conservative (relative to reality): most distributions actually lie within 2 standard deviations of the mean much more than 75% of the time.
  - However, for the discrete random variable taking the values -1, 0, and 1 with respective probabilities $1/(2t^2)$, $1 - 1/t^2$, and $1/(2t^2)$, the mean is 0 and the standard deviation is $1/t$, so the inequality is sharp for this distribution and $k = t$.

- Example: In a statistics class, the exam scores are distributed with mean 80 and standard deviation 5. Find a lower bound for the percentage of students who must score (i) between 72 and 88 inclusive, (ii) between 70 and 90 inclusive, and (iii) between 67 and 93 inclusive.
  - Since we are given the mean and standard deviation, we can use Chebyshev's inequality.
  - For (i), the range represents a width of $k = \frac{8}{5} = 1.6$ standard deviations away from the mean, and so by Chebyshev's inequality, a proportion of at least $1 - 1/k^2 = \boxed{60.9\%}$ of students must score between 72 and 88 inclusive.
  - For (ii), the range represents a width of $k = \frac{10}{5} = 2$ standard deviations away from the mean, and so by Chebyshev's inequality, a proportion of at least $1 - 1/k^2 = \boxed{75\%}$ of students must score between 70 and 90 inclusive.
  - For (iii), the range represents a width of $k = \frac{13}{5} = 2.6$ standard deviations away from the mean, so by Chebyshev's inequality, a proportion of at least $1 - 1/k^2 = \boxed{85.2\%}$ of students must score between 67 and 93 inclusive.

- Example: In a different statistics class, the exam scores are distributed with mean 75, but no additional information is known. (i) If the exam scores are always nonnegative, find an upper bound on the proportion of students who score a 90 or above, and (ii) if the standard deviation is also known to be 9, find a lower bound for the proportion of students who score strictly between 60 and 90.
  - For (i), since we do not know anything about the standard deviation, the only tool we have available is Markov's inequality. In this case, if X represents the exam score, then $P(X \geq a) \leq E(X)/a$, since X is always nonnegative.
  - For $a = 90$, we see that $P(X \geq 90) \leq E(X)/90 = 75/90 = 5/6$. Thus, at most $\boxed{5/6}$ of the class could score 90 or above.
  - In fact, this is the best possible bound: it could be the case that 5/6 of the class got exactly 90 while the other 1/6 got 0, in which case the average score would indeed be 75.
  - For (ii), since now we also have the standard deviation, we can use Chebyshev's inequality. In this case, if X represents the exam score, then $P(|X - 75| \geq 9k) \leq 1/k^2$.
  - Since the desired score range of 75 ± 15 points represents $k = 15/9 = 5/3$ standard deviations away from the mean, Chebyshev's inequality says that the proportion of students scoring 15 or more points away from the mean is at most $1/k^2 = 0.36$. The remaining students $\boxed{64\%}$ of students must therefore score between 60 and 90.
  - This is the best possible bound here also: it could be the case that 18% of the class got 60, 18% got 90, and the other 64% got 75, in which case the mean would be 75 while the standard deviation would be 9.

2.2.3 Joint Distributions

- Next, we discuss the situation of having several continuous random variables defined on the same sample space.
  - Just as with discrete random variables, if we have a collection of continuous random variables $X_1, X_2, \ldots, X_n$, we can summarize all of the possible information about the behavior of these random variables simultaneously using a joint probability density function.

- Definition: If $X_1, X_2, \ldots, X_n$ are continuous random variables, then the function $p_{X_1,X_2,\ldots,X_n}(a_1, a_2, \ldots, a_n)$ defined on ordered n-tuples of real numbers, such that $\iint_R p_{X_1,X_2,\ldots,X_n}(a_1, a_2, \ldots, a_n)\, da_n da_{n-1} \cdots da_1 = P[(X_1, X_2, \ldots, X_n) \in R]$ for every region R in n-dimensional space is called the joint probability density function of $X_1, X_2, \ldots, X_n$.
  - Although the definition seems somewhat complicated, the idea is the same as a one-variable probability density function: to compute the probability that the values of $X_1, X_2, \ldots, X_n$ land in a particular region R, we simply integrate the probability density function on the domain R.

- To evaluate any of these probabilities, we will need to use multivariable integration.
  - For the situation of two random variables X and Y (which we will primarily focus on), R will be a region in the xy-plane, and the integrals will be double integrals. In this situation, we may visualize $z = p(x, y)$ as a surface lying above the xy-plane and the double integral $\iint_R p(x, y)\, dy\, dx$ as the volume underneath the surface that lies on top of the planar region R.
  - In the event that the region R is the rectangle $a \le x \le b$, $c \le y \le d$, we may evaluate this double integral as the iterated integral $\int_a^b \left[\int_c^d f(x, y)\, dy\right] dx$, where we integrate first (on the inside) with respect to the variable y, and then second (on the outside) with respect to the variable x.
  - When we evaluate the inner integral with respect to y, we view x as a constant and y as the variable and take the antiderivative in y, then evaluate at the two limits of integration and subtract.
  - Example: To evaluate $\int_0^1 \int_0^2 (6 - 2x - 2y)\, dy\, dx$, first we evaluate the inner integral $\int_0^2 (6 - 2x - 2y)\, dy$ as follows:

$$\int_0^2 (6 - 2x - 2y)\, dy = \left[6y - 2xy - y^2\right]\Big|_{y=0}^{2} = [12 - 4x - 4] - [0 - 0 - 0] = 8 - 4x.$$

    Now we can evaluate the "outer" integral $\int_0^1 (8 - 4x)\, dx = \left[8x - 2x^2\right]\Big|_{x=0}^{1} = \boxed{6}$.
  - More generally, if the region R is bounded below by the curve $y = c(x)$ and above by the curve $y = d(x)$, then the iterated integral has the form $\int_a^b \int_{c(x)}^{d(x)} f(x, y)\, dy\, dx$, where now the inner limits of integration depend on the outer variable x. When we evaluate the inner integral in y, we will be left with a function of x, and then we can evaluate the outer integral.

- Example: The continuous random variables X and Y have joint probability density function defined by $p(x, y) = \frac{1}{4}xy$ for $0 \le x \le 2$ and $0 \le y \le 2$, and $p(x, y) = 0$ elsewhere. Find (i) $P(0 \le X \le 1, 0 \le Y \le 1)$, (ii) $P(0 \le X \le 1)$, (iii) $P(X + Y \le 1)$, and (iv) $P(1 < X < Y < 2)$.
  - We simply set up each of the corresponding integrals.
  - First, $P(0 \le X \le 1, 0 \le Y \le 1)$ corresponds to the rectangle bounded by $x = 0$, $x = 1$, $y = 0$, and $y = 1$, so the desired integral is $\int_0^1 \int_0^1 \frac{1}{4}xy\, dy\, dx = \int_0^1 \left[\frac{1}{8}xy^2\right]\Big|_{y=0}^{1} dx = \int_0^1 \frac{1}{8}x\, dx = \left[\frac{1}{16}x^2\right]\Big|_{x=0}^{1} = \boxed{\frac{1}{16}}$.
  - Second, $P(0 \le X \le 1)$ corresponds to the infinite strip bounded by $x = 0$ and $x = 1$. But since the probability density function is zero except when $0 \le y \le 2$, we only want to integrate between those bounds.
  - Thus, the desired integral is $\int_0^1 \int_0^2 \frac{1}{4}xy\, dy\, dx = \int_0^1 \left[\frac{1}{8}xy^2\right]\Big|_{y=0}^{2} dx = \int_0^1 \frac{1}{2}x\, dx = \left[\frac{1}{4}x^2\right]\Big|_{x=0}^{1} = \boxed{\frac{1}{4}}$.
  - Third, inside the rectangle $0 \le x \le 2$, $0 \le y \le 2$, the condition $P(X + Y \le 1)$ represents the triangle with vertices $(0, 0)$, $(1, 0)$, and $(0, 1)$.
  - By drawing a quick sketch, and using the fact that the diagonal side of the triangle is the line $x + y = 1$, we see that we can describe this region as $0 \le x \le 1$ and $0 \le y \le 1 - x$. The desired probability is then $\int_0^1 \int_0^{1-x} \frac{1}{4}xy\, dy\, dx = \int_0^1 \left[\frac{1}{8}xy^2\right]\Big|_{y=0}^{1-x} dx = \int_0^1 \frac{1}{8}(x - 2x^2 + x^3)\, dx = \left[\frac{1}{16}x^2 - \frac{1}{12}x^3 + \frac{1}{32}x^4\right]\Big|_{x=0}^{1} = \boxed{\frac{1}{96}}$.
  - Finally, drawing the condition $1 < X < Y < 2$ shows that it is also a triangle with vertices $(1, 1)$, $(1, 2)$, and $(2, 2)$, and can be described as the region with $1 < x < 2$ and $x < y < 2$. The desired probability is then $\int_1^2 \int_x^2 \frac{1}{4}xy\, dy\, dx = \int_1^2 \left[\frac{1}{8}xy^2\right]\Big|_{y=x}^{2} dx = \int_1^2 \frac{1}{8}(4x - x^3)\, dx = \left[\frac{1}{4}x^2 - \frac{1}{32}x^4\right]\Big|_{x=1}^{2} = \boxed{\frac{9}{32}}$.

- As in the case of discrete random variables, we can also recover the individual probability distributions for any of the random variables from their joint distribution by integrating over the other variables:

- Proposition (Marginal Densities): If $p_{X,Y}(a, b)$ is the joint probability density function for the continuous random variables X and Y, then for any a and b we may compute the single-variable probability density functions for X and Y as $p_X(x) = \int_{-\infty}^{\infty} p_{X,Y}(x, y)\, dy$ and $p_Y(y) = \int_{-\infty}^{\infty} p_{X,Y}(x, y)\, dx$.

- Proof: By the definition of the joint probability density function, we know that $P(a \le X \le b) = P(a \le X \le b, -\infty < Y < \infty) = \int_a^b \int_{-\infty}^{\infty} p_{X,Y}(x,y)\,dy\,dx$.
- Thus, we see that integrating $\int_{-\infty}^{\infty} p_{X,Y}(x,y)\,dy$ with respect to x on the interval $[a,b]$ yields $P(a \le X \le b)$, which means that $\int_{-\infty}^{\infty} p_{X,Y}(x,y)\,dy$ is the probability density function for X.
- The second formula follows in the same way upon interchanging the roles of X and Y and switching the order of integration in the iterated integral (this is always allowable by Fubini's theorem since the integrand is nonnegative).
- Remark: This result can be extended to an arbitrary number of variables using essentially the same argument. In general, if $X_1, X_2, \ldots, X_n$ are continuous random variables with joint pdf $p_{X_1,X_2,\ldots,X_n}(x_1,\ldots,x_n)$ then for any $1 \le k \le n$ the joint pdf $p_{X_1,X_2,\ldots,X_k}(x_1,\ldots,x_k)$ is given by $p_{X_1,X_2,\ldots,X_k}(x_1,\ldots,x_k) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p_{X_1,\ldots,X_n}(x_1,\ldots,x_k,x_{k+1},\ldots,x_n)\,dx_n dx_{n-1}\cdots dx_{k+1}$, with a similar formula holding for any subset of the X_i whose values are fixed.

- Example: The continuous random variables X and Y have joint probability density function defined by $p(x,y) = a \cdot (x+2y)$ for $0 \le x \le 3$ and $0 \le y \le 2$, and $p(x,y) = 0$ elsewhere. Find (i) the value of a, (ii) $P(2 \le X \le 3)$, (iii) $P(2Y < X)$, and (iv) the marginal probability density functions for X and Y.
  - For (i), in order to be a probability density function, the integral of $p(x,y)$ over its domain must equal 1. Setting up the integral yields $\int_0^3 \int_0^2 a(x+2y)\,dy\,dx = \int_0^3 a(xy+y^2)\Big|_{y=0}^{2} dx = \int_0^3 a(2x+4)\,dx = \left[a(x^2+4x)\right]\Big|_{x=0}^{3} = 21a$. Thus, we must have $a = \boxed{1/21}$.
  - For (ii), inside the rectangle $0 \le x \le 3, 0 \le y \le 2$, the condition $P(2 \le X \le 3)$ corresponds to the rectangle with $2 \le x \le 3$ and $0 \le y \le 2$. Thus, the desired integral is $\int_2^3 \int_0^2 \frac{1}{21}(x+2y)\,dy\,dx = \int_2^3 \frac{1}{21}(xy+y^2)\Big|_{y=0}^{2} dx = \int_2^3 \frac{1}{21}(2x+4)\,dx = \left[\frac{1}{21}(x^2+4x)\right]\Big|_{x=2}^{3} = \boxed{\frac{3}{7}}$.
  - For (iii), inside the rectangle $0 \le x \le 2$, $0 \le y \le 2$, the condition $P(2Y < X)$ corresponds to the triangle with vertices $(0,0)$, $(0,1)$, and $(2,1)$, and can be described as the region with $0 \le x \le 2$ and $0 < y < x/2$.
  - The integral is thus $\int_0^2 \int_0^{x/2} \frac{1}{21}(x+2y)\,dy\,dx = \int_0^2 \frac{1}{21}(xy+y^2)\Big|_{y=0}^{x/2} dx = \int_0^2 \frac{1}{28}x^2\,dx = \frac{1}{84}x^3\Big|_{x=0}^{2} = \boxed{\frac{2}{21}}$.
  - For (iv), to find the marginal pdfs we simply integrate the joint pdf with respect to the appropriate variable.
  - The marginal pdf for X is $\int_{-\infty}^{\infty} p(x,y)\,dy = \int_0^2 \frac{1}{21}(x+2y)\,dy = \frac{1}{21}(xy+y^2)\Big|_{y=0}^{2} = \boxed{\frac{1}{21}(2x+4)}$.
  - The marginal pdf for Y is $\int_{-\infty}^{\infty} p(x,y)\,dx = \int_0^3 \frac{1}{21}(x+2y)\,dx = \frac{1}{21}(\frac{1}{2}x^2+2xy)\Big|_{x=0}^{3} = \boxed{\frac{1}{42}(9+8y)}$.
  - Remark: We could have instead used our calculation of the marginal pdf for X to find $P(2 \le X \le 3) = \int_2^3 \frac{1}{21}(2x+4)\,dx = \frac{3}{7}$. Note (of course) the result comes out the same either way!

2.2.4 Independence, Covariance, Correlation

- Just like with discrete random variables, we can also use joint distributions to describe when two continuous random variables are independent.
  - As in the discrete case, two continuous random variables X and Y are independent when knowing the value of one gives no additional information about the value of the other, which we can phrase as saying that $P(a < X < b | c < Y < d) = P(a < X < b)$.
  - By rearranging, this says $P(a < X < b, c < Y < d) = P(a < X < b) \cdot P(c < Y < d)$, which in terms of probabilities says $\int_a^b \int_c^d p_{X,Y}(x,y)\,dy\,dx = \int_a^b p_X(x)\,dx \cdot \int_c^d p_Y(y)\,dy$.
  - But the right-hand side is also equal to the iterated integral $\int_a^b \int_c^d p_X(x) \cdot p_Y(y)\,dy\,dx$. So since both sides are equal on every rectangle $[a,b] \times [c,d]$, we must have $p_{X,Y}(x,y) = p_X(x) \cdot p_Y(y)$ for all x and y.

- Notice (as we might have guessed) that this is exactly the same condition as in the discrete case.
- We can extend to more than two variables, in just the same way:

- Definition: We say that the continuous random variables $X_1, X_2, \ldots, X_n$ are collectively independent if the joint distribution $p_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = p_{X_1}(x_1) \cdot p_{X_2}(x_2) \cdot \cdots \cdot p_{X_n}(x_n)$ for all real numbers $x_1, x_2, \ldots, x_n$.
  - Example: The continuous random variables X and Y with joint probability density function defined by $p_{X,Y}(x, y) = \frac{1}{4}xy$ for $0 \le x \le 2$ and $0 \le y \le 2$ are independent because their marginal distribution functions are $p_X(x) = \int_0^2 \frac{1}{4}xy\,dy = \frac{1}{2}x$ and $p_Y(y) = \int_0^2 \frac{1}{4}xy\,dx = \frac{1}{2}y$, and indeed $p_{X,Y}(x, y) = \frac{1}{4}xy = \frac{1}{2}x \cdot \frac{1}{2}y = p_X(x) \cdot p_Y(y)$.
  - Example: The continuous random variables X and Y with joint probability density function defined by $p(x, y) = \frac{1}{21}(x + 2y)$ for $0 \le x \le 3$ and $0 \le y \le 2$ are not independent, because their marginal distribution functions were previously calculated to be $p_X(x) = \frac{1}{21}(2x + 4)$ and $p_Y(y) = \frac{1}{42}(9 + 8y)$, and $p_{X,Y}(x, y) \neq p_X(x)p_Y(y)$.
  - In the second example above, we did not actually need to evaluate the marginal distribution functions to see that the variables were not independent, because their joint distribution function $p(x, y) = \frac{1}{21}(x+2y)$ cannot be written as the product of a single function of x and a single function of y.
  - In fact, the converse observation holds as well: if we can write $p_{X,Y}(x, y) = q(x) \cdot r(y)$ for some functions $q(x)$ and $r(y)$, then in fact the random variables X and Y will be independent[6].

- We obtain the same results as in the discrete case about variance and independence, and can define the covariance and correlation as well:

- Definition: If X and Y are random variables whose expected values exist and are μ_X and μ_Y respectively, then the covariance of X and Y is defined as $\text{cov}(X, Y) = E[(X - \mu_X) \cdot (Y - \mu_Y)] = E(XY) - E(X)E(Y)$.
  - In order to compute the covariance, we need to know how to compute the expected value of arbitrary functions of X and Y.
  - We can use the same principle discussed earlier for how to find the expected value of an arbitrary function of X: explicitly, for any function $g(X, Y)$, we have $E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) \cdot p_{X,Y}(x, y)\,dy\,dx$.

- Example: If X and Y have joint distribution given by $p_{X,Y}(x, y) = x + y$ for $0 \le x \le 1$, $0 \le y \le 1$, find the covariance of X and Y.
  - We compute $E(X) = \int_0^1 \int_0^1 x(x + y)\,dy\,dx = \int_0^1 (x^2 + \frac{1}{2}x)\,dx = \frac{7}{12}$, $E(Y) = \int_0^1 \int_0^1 y(x + y)\,dy\,dx = \int_0^1 (\frac{1}{2}x + \frac{1}{3})\,dx = \frac{7}{12}$, and $E(XY) = \int_0^1 \int_0^1 xy(x + y)\,dy\,dx = \int_0^1 (\frac{1}{2}x^2 + \frac{1}{3}x)\,dx = \frac{1}{3}$.
  - Thus, the covariance is $\text{cov}(X, Y) = E(XY) - E(X)E(Y) = \boxed{-\frac{1}{144}}$.

- The same properties of variance and covariance also hold in the continuous setting:

- Proposition (Properties of Variance and Covariance): If X, Y, Z are continuous random variables whose expected values exist, then for any a and b we have $\text{cov}(X, X) = \text{var}(X)$, $\text{cov}(Y, X) = \text{cov}(X, Y)$, $\text{cov}(X + Y, Z) = \text{cov}(X, Z) + \text{cov}(Y, Z)$, $\text{cov}(aX + b, Y) = a \cdot \text{cov}(X, Y)$, and $\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2\text{cov}(X, Y)$. Furthermore, if X and Y are independent, then $E(XY) = E(X) \cdot E(Y)$, and $\text{var}(X + Y) = \text{var}(X) + \text{var}(Y)$.
  - Proof: All of these properties follow in the same way as in the discrete case.
  - For example, if X and Y are independent, then $E(XY) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy \cdot p_{X,Y}(x, y)\,dy\,dx = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy \cdot p_X(x)p_Y(y)\,dy\,dx = \int_{-\infty}^{\infty} xp_X(x)dx \cdot \int_{-\infty}^{\infty} yp_Y(y)dy = E(X) \cdot E(Y)$.

- Definition: If X and Y are continuous random variables whose variances exist and are nonzero, the (Pearson) correlation between X and Y is defined as $\text{corr}(X, Y) = \dfrac{\text{cov}(X, Y)}{\sigma(X)\sigma(Y)}$.

[6] This follows because the marginal distribution functions can be computed as $p_X(x) = \int_{-\infty}^{\infty} q(x)r(y)\,dy = q(x) \cdot \int_{-\infty}^{\infty} r(y)\,dy$ and $p_Y(y) = \int_{-\infty}^{\infty} q(x)r(y)\,dx = r(y) \cdot \int_{-\infty}^{\infty} q(x)\,dx$, and then in fact $p_X(x) \cdot p_Y(y) = q(x)r(y) \cdot \int_{-\infty}^{\infty} q(x)r(y)\,dy\,dx = q(x)r(y)$ since the latter double integral is 1 because it is the integral of the joint distribution $p_{X,Y}$ over its entire domain.

- Once again, the correlation in the continuous case has the same interpretation as in the discrete case: it describes the strength to which the relationship between X and Y can be captured by a linear model.

- Example: Suppose that the continuous random variables X and Y have joint distribution given by $p_{X,Y}(x,y) = 2e^{-x-2y}$ for $x \geq 0$ and $y \geq 0$. Find (i) the marginal pdfs of X and Y, (ii) $P(X > 1)$, (iii) $P(X < Y < 2X)$, (iv) whether X and Y are independent, and (v) the covariance and correlation of X and Y.

  - First, we have $p_X(x) = \int_0^\infty 2e^{-x-2y}\,dy = -e^{-x-2y}\Big|_{y=0}^{\infty} = \boxed{e^{-x}}$, and also $p_Y(y) = \int_0^\infty 2e^{-x-2y}\,dx = -2e^{-x-2y}\Big|_{x=0}^{\infty} = \boxed{2e^{-2y}}$.
  - Second, we have $P(X > 1) = \int_1^\infty p_X(x)\,dx = \int_1^\infty e^{-x}\,dx = -e^{-x}\Big|_{x=1}^{\infty} = \boxed{e^{-1}} \approx 0.3679$.
  - Third, inside the region with $x, y \geq 0$, the condition $P(X < Y < 2X)$ yields an infinite triangular region that can be described by $x \geq 0$ and $x < y < 2x$, so the desired integral is $\int_0^\infty \int_x^{2x} 2e^{-x-2y}\,dy\,dx = \int_0^\infty -e^{-x-2y}\Big|_{y=x}^{2x} dx = \int_0^\infty [e^{-3x} - e^{-5x}]\,dx = -\frac{1}{3}e^{-3x} + \frac{1}{5}e^{-5x}\Big|_{x=0}^{\infty} = \boxed{\frac{2}{15}}$.
  - Next, we can see that $p_{X,Y}(x,y) = 2e^{-x} \cdot e^{-2y}$ which is the product of a function of x with a function of y, so by the discussion above, we see that $\boxed{X \text{ and } Y \text{ are independent}}$. Alternatively, we could observe that $p_{X,Y}(x,y) = 2e^{-x-2y} = e^{-x} \cdot 2e^{-2y} = p_X(x) \cdot p_Y(y)$.
  - Finally, since X and Y are independent, their covariance and correlation are both $\boxed{0}$.

- Example: Suppose that the continuous random variables X and Y have joint distribution given by $p_{X,Y}(x,y) = \frac{1}{24}(x+y^2)$ for $0 \leq x \leq 2$ and $0 \leq y \leq 3$. Find the covariance and correlation of X and Y.

  - We have $E(X) = \int_0^2 \int_0^3 \frac{1}{24}x(x+y^2)\,dy\,dx = \int_0^2 \frac{1}{8}(3x+x^2)\,dx = \frac{13}{12}$, $E(Y) = \int_0^2 \int_0^3 \frac{1}{24}y(x+y^2)\,dy\,dx = \int_0^2 \frac{1}{32}(6x+27)\,dx = \frac{33}{16}$, and $E(XY) = \int_0^2 \int_0^3 \frac{1}{24}xy(x+y^2)\,dy\,dx = \int_0^2 \frac{1}{32}(6x^2+27x)\,dx = \frac{35}{16}$.
  - Thus, the covariance is $\text{cov}(X,Y) = E(XY) - E(X)E(Y) = \frac{35}{16} - \frac{13}{12} \cdot \frac{33}{16} = \boxed{-\frac{3}{64}} \approx -0.0469$.
  - For the correlation, we also need to compute $\sigma(X)$ and $\sigma(Y)$.
  - First, $E(X^2) = \int_0^2 \int_0^3 \frac{1}{24}x^2(x+y^2)\,dy\,dx = \int_0^2 \frac{1}{8}(3x^2+x^3)\,dx = \frac{3}{2}$, so $\sigma(X) = \sqrt{E(X^2) - E(X)^2} = \sqrt{47}/12 \approx \boxed{0.5713}$.
  - Also, $E(Y^2) = \int_0^2 \int_0^3 \frac{1}{24}y^2(x+y^2)\,dy\,dx = \int_0^2 \frac{1}{40}(15x+81)\,dx = \frac{24}{5}$, so $\sigma(Y) = \sqrt{E(Y^2) - E(Y)^2} = \sqrt{3495}/80 \approx \boxed{0.7390}$.
  - Thus, $\text{corr}(X,Y) = \dfrac{\text{cov}(X,Y)}{\sigma(X)\sigma(Y)} \approx \boxed{-0.1110}$.

2.3 The Normal Distribution, Central Limit Theorem, and Modeling Applications

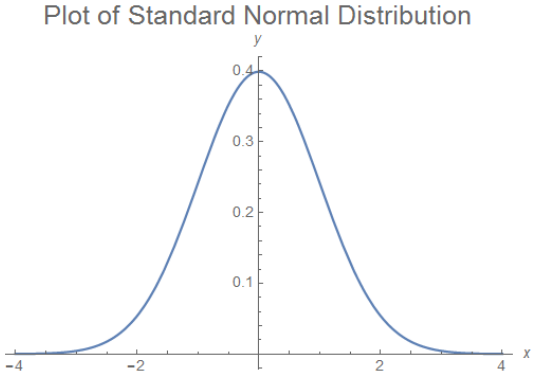
- In this section, we will discuss three important classes of probability distributions: the Gaussian normal distributions, the Poisson distributions, and the exponential distributions.

  - Each of these classes of distributions arises in various practical applications involving phenomena with particular simple properties, and our goal is to describe why these distributions occur so frequently. The normal distribution is by far the most important of these three, but all of them serve as important models for various processes, so we will discuss them together.
  - Specifically, the normal distribution is often used to model quantities arising as sums or averages of a number of small pieces, such as student grades, human heights, errors in measurements, and many other physical phenomena. The reason for this (as we will explain) is because of an extremely important result known as the central limit theorem, which says that the normalized average of repeated sampling from a fixed distribution will always tend to be normally distributed as the sample size grows large.

- The Poisson distribution is used to model the occurrences of discrete rare events such as airplane crashes, mutations in DNA replication, insurance claims, goals during a sports game, and radioactive decay. The reason for this is due to the Poisson limit theorem, which says that if the number of samples from a varying distribution is selected in such a way that the overall expected number of occurrences approaches a fixed limit as the number of samples increases, then the limiting distribution will have a Poisson distribution.
- The exponential distribution is used to model waiting times for "memoryless" processes, in which the distribution of future waiting time is independent of the amount of time already waited.

2.3.1 The Normal Distribution

- One of the most important probability distributions is the normal distribution, often called the "bell curve" due to its shape.
- Definition: A random variable $N_{\mu,\sigma}$ is normally distributed with mean μ and standard deviation σ if its probability density function is $p_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$. The standard normal distribution is $N_{0,1}$, having mean 0 and standard deviation 1.
  - Here is a graph of the standard normal distribution showing its "bell curve" shape:



  - It is not trivial to verify that $p_{\mu,\sigma}(x)$ actually yields a probability density function: one must show that $\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}\,dx = 1$, although this can be done in various ways[7].
  - By manipulating the integrals accordingly using a series of substitutions and the evaluation of $\int_{-\infty}^{\infty} p_{\mu,\sigma}(x)\,dx = 1$, we may eventually compute that $E(N_{\mu,\sigma}) = \mu$ and $\text{var}(N_{\mu,\sigma}) = \sigma^2$ so that $\sigma(N_{\mu,\sigma}) = \sigma$.
- Example: Suppose $N_{11,2}$ is a normally-distributed random variable with expected value 11 and standard deviation 2. Find (i) $P(N \le 11)$, (ii) $P(7 \le N \le 9)$, and (iii) $P(N \ge 13)$.
  - One approach is simply to write down the probability density function and set up the integrals.
  - This yields $P(N_{11,2} \le 11) = \int_{-\infty}^{11} \frac{1}{2\sqrt{2\pi}} e^{-(x-11)^2/8}\,dx$, with $P(7 \le N_{11,2} \le 9) = \int_{7}^{9} \frac{1}{2\sqrt{2\pi}} e^{-(x-11)^2/8}\,dx$, and $P(N_{11,2} \ge 13) = \int_{13}^{\infty} \frac{1}{2\sqrt{2\pi}} e^{-(x-11)^2/8}\,dx$.
  - Unfortunately, except in certain rare cases, integrals involving the normal distribution are very difficult to evaluate exactly, owing to the fact that the function $e^{-(x-\mu)^2/(2\sigma^2)}$ does not have an elementary antiderivative: this means we cannot write down an exact formula for the indefinite integral (i.e., the cumulative distribution function) in terms of polynomials, exponentials, logarithms, or trigonometric functions.

[7] One approach is to substitute $u = (x-\mu)/(\sigma\sqrt{2})$, which converts the problem into showing that the value of the integral $J = \int_{-\infty}^{\infty} e^{-u^2}\,du$ is $\sqrt{\pi}$. This can be done in several ways: the most standard approach is to write $J^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)}\,dy\,dx$ and then convert to polar coordinates to deduce $J^2 = \pi$. There are many other approaches, including differentiation under the integral, interchanging the order of a double integral, integration in the complex plane, and asymptotic analysis.

- Of course, we can use numerical integration procedures (implemented by a calculator or computer) to approximate these integrals to any desired accuracy: this yields $P(N_{11,2} \leq 11) = \boxed{0.5}$, $P(7 \leq N_{11,2} \leq 9) \approx \boxed{0.1359}$, and $P(N_{11,2} \geq 13) \approx \boxed{0.1587}$.

- Another approach to solving problems involving the normal distribution is use a table of computed values for the cumulative distribution function of the standard normal distribution $N_{0,1}$, along with a substitution.
  - Explicitly, it is a straightforward calculation to verify that if we define $z_a = \dfrac{a-\mu}{\sigma}$ and $z_b = \dfrac{b-\mu}{\sigma}$, then $P(a \leq N_{\mu,\sigma} \leq b) = P(z_a \leq N_{0,1} \leq z_b)$.
  - Intuitively, the reason that this change of variables will work is that all of the normal distributions are geometrically similar to one another. Thus, by translating by μ (to center the distribution at 0) and rescaling by $1/\sigma$ (to stretch the distribution so it has standard deviation 1), we may convert any question about areas under an arbitrary normal distribution $N_{\mu,\sigma}$ to one about the standard normal distribution $N_{0,1}$.
  - Once we make this translation and find these (so-called) "z-scores", we can use a table of computed values for the cumulative distribution function of the standard normal $N_{0,1}$ (such as the table given below) to find the desired probabilities:

z	-3	-2	-1.5	-1	-0.5	0	0.5	1	1.5	2	3
$P(N_{0,1} \leq z)$	0.0014	0.0228	0.0668	0.1587	0.3085	0.5	0.6915	0.8413	0.9332	0.9772	0.9987

- Example (again): Suppose $N_{11,2}$ is a normally-distributed random variable with expected value 11 and standard deviation 2. Find (i) $P(N \leq 11)$, (ii) $P(7 \leq N \leq 9)$, and (iii) $P(N \geq 13)$.
  - We have $\mu = 11$ and $\sigma = 2$, so $P(N_{11,2} \leq 11) = P(N_{0,1} \leq 0) = \boxed{0.5}$.
  - Likewise, $P(7 \leq N_{11,2} \leq 9) = P(-2 \leq N_{0,1} \leq -1) = P(N_{0,1} \leq -1) - P(N_{0,1} \leq -2) = 0.1587 - 0.0228 \approx \boxed{0.1359}$.
  - Finally, $P(N_{11,2} \geq 13) = P(N_{0,1} \geq 1) = 1 - P(N_{0,1} \leq 1) = 1 - 0.8413 = \boxed{0.1587}$.

- Example: A certain standardized test is designed so that its score distribution will be normal with a mean of 500 and a standard deviation of 100. Determine the percentage of scores that will be (i) between 450 and 550, (ii) less than 600, and (iii) greater than 700.
  - From the given information, the scores follow the normal distribution $N_{500,100}$.
  - Using the z-score method described above, we can compute $P(450 < N_{500,100} < 550) = P(-0.5 < N_{0,1} < 0.5) = 0.6915 - 0.3085 \approx \boxed{0.3830}$.
  - Next, we have $P(N_{500,100} < 600) = P(N_{0,1} \leq 1) \approx \boxed{0.8413}$, and finally, $P(N_{500,100} > 700) = P(N_{0,1} > 2) = 1 - P(N_{0,1} \leq 2) \approx \boxed{0.0228}$.
  - Remark: Of course, we could also compute these values directly using numerical integration and the original distribution $N_{500,100}$.

- In some situations we want to invert our analysis by starting with a probability and finding the corresponding value or range in the distribution.
  - Analytically, this corresponds to evaluating the inverse function of the normal cumulative density function (which is usually called the inverse normal for short), which can be done efficiently using a calculator or computer.
  - Alternatively, we could look up the needed probabilities in a table to find the associated z-scores.

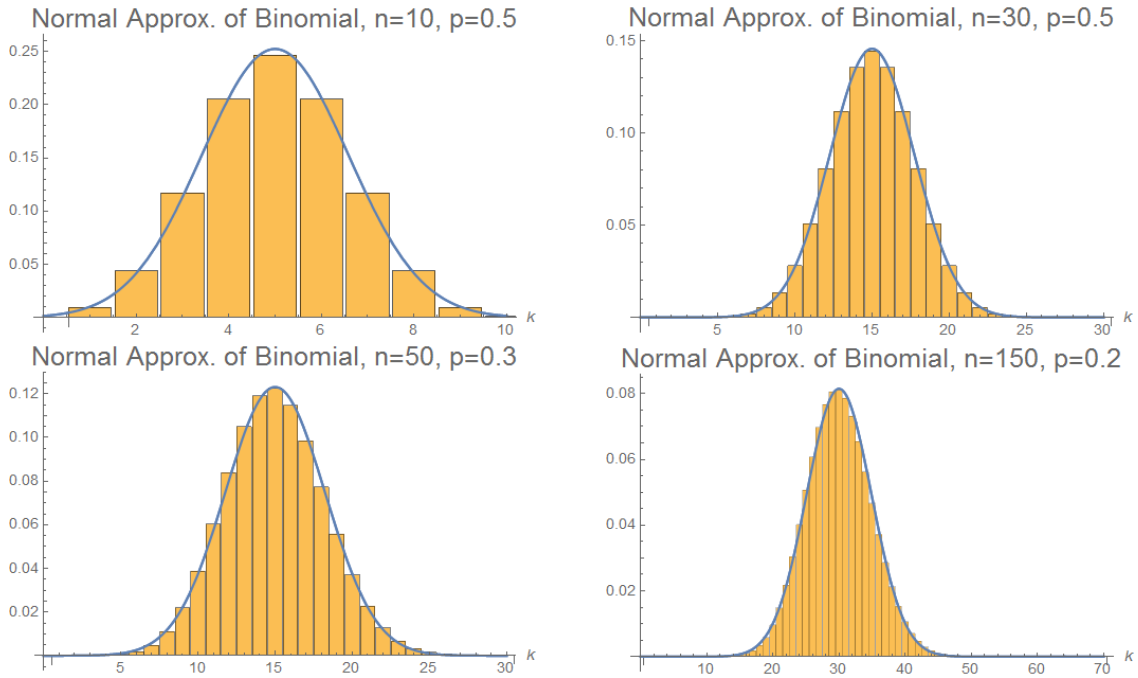
- Example: A certain standardized test is designed so that its score distribution will be normal with a mean of 500 and a standard deviation of 100. Determine the score at (i) the 80th percentile, and (ii) the 99th percentile.
  - Remark: A score is said to be at the Nth percentile of a distribution if a proportion $N/100$ of the other scores are below it. (Thus, the median is at the 50th percentile.)

- The score S_{80} at the 80th percentile will have the property that $P(N_{500,100} \leq S_{80}) = 0.80$. Using a computer, we can find that $P(N_{500,100} \leq 584) \approx 0.7995$ and $P(N_{500,100} \leq 585) = 0.8023$, so the desired 80th-percentile score is $\boxed{584}$ to the nearest integer.
- Alternatively, using a table of z-scores, we could find that the value z with $P(N_{0,1} \leq z) = 0.80$ is $z \approx 0.8416$, and so the desired score is $500 + 100z \approx 584.16$.
- In the same way, the score S_{97} at the 99th percentile will have the property that $P(N_{500,100} \leq S_{99}) = 0.99$. Using a computer, we can find that $P(N_{500,100} \leq 732) \approx 0.9898$ and $P(N_{500,100} \leq 733) = 0.9901$, so the desired 99th-percentile score is $\boxed{733}$ to the nearest integer.
- Alternatively, using a table of z-scores, we could find that the value z with $P(N_{0,1} \leq z) = 0.99$ is $z \approx 2.3263$, and so the desired score is $500 + 100z \approx 732.63$.

2.3.2 The Central Limit Theorem

- As remarked earlier, the normal distribution is seen very commonly in physical applications (e.g., in the distribution of human heights, sizes of parts made by automated processes, blood pressures, measurement errors, and scores on standardized examinations).
- The reason for the common appearance of the normal distribution is the following fundamental result:
- Theorem (Central Limit Theorem): Let X_1, X_2, ... , X_n be a sequence of independent, identically-distributed discrete or continuous random variables each with finite expected value μ and standard deviation $\sigma > 0$. Then the distribution of the random variable $Y_n = \dfrac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$ will approach the standard normal distribution (of mean 0 and standard deviation 1) as n tends to ∞: explicitly, we have $P(a \leq Y_n \leq b) \to \int_a^b \dfrac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx$ for any real numbers $a \leq b$.
  - If $X_1, \ldots, X_n$ are independent and identically-distributed, we may think of the random variable $X_1 + \cdots + X_n$ as being the sum of the results of independently sampling a random variable X a total of n times. (One example of this would be flipping a coin n times and summing the total number of heads; another would be rolling a die n times and summing the outcomes.)
  - From the linearity of expected value, we can see that $E(X_1 + \cdots + X_n) = E(X_1) + \cdots + E(X_n) = n\mu$.
  - Also, since $X_1, \ldots, X_n$ are independent, we also have $\text{var}(X_1 + X_2 + \cdots + X_n) = \text{var}(X_1) + \cdots + \text{var}(X_n) = n\sigma^2$, and so $\sigma(X_1 + \cdots + X_n) = \sqrt{n\sigma^2} = \sigma\sqrt{n}$.
  - Thus, the central limit theorem says that if we "normalize" the summed distribution $X_1 + \cdots + X_n$ by translating and rescaling it so that its expected value is 0 and its standard deviation is 1, the resulting normalized average approaches the standard normal distribution as we average over more and more samples.
  - This result is quite powerful, since it applies to any discrete or continuous random variable whose expected value and standard deviation are defined.
- As one application, since the binomial distribution is obtained by summing n independent Bernoulli random variables, the central limit theorem tells us that if n is sufficiently large, then the binomial distribution will be well approximated by a normal distribution with the same expected value and standard deviation.
  - As a practical matter, the approximation tends to be very good when np and $n(1-p)$ are both at least 5, and increases in accuracy when np and $n(1-p)$ are larger.
  - Since the binomial distribution is discrete while the normal distribution is continuous, we typically make an adjustment when we approximate the binomial distribution by the normal distribution, namely, we approximate the probability $P(B = k)$ that the binomial random variable equals k with the probability $P(k - \frac{1}{2} \leq N \leq k + \frac{1}{2})$ that the normal random variable lands in the interval $[k - \frac{1}{2}, k + \frac{1}{2}]$.
  - We make a similar "continuity correction" whenever we approximate a discrete distribution by a continuous one, because in all such cases we need to compare areas to areas.

- Here are some comparisons of binomial distributions (with various parameters n and p) with their corresponding normal approximations:



- Example: Compute the exact probability that a fair coin flipped 400 times will land heads (i) exactly 200 times, and (ii) between 203 and 208 times inclusive, and compare them to the results of the normal approximation (with continuity correction).

  - The total number of heads is binomially distributed with $n = 400$ and $p = 1/2$, so for (i) the probability is $\boxed{\frac{1}{2^{400}}\binom{400}{200}} \approx 3.99\%$, while for (ii) it is $\boxed{\frac{1}{2^{400}}\left[\binom{400}{203} + \binom{400}{204} + \cdots + \binom{400}{208}\right]} \approx 20.36\%$.
  - The normal approximation to this binomial distribution has expected value $\mu = np = 200$ and standard deviation $\sqrt{np(1-p)} = 10$.
  - Therefore, for (i) we wish to compute $P(199.5 \leq N_{200,10} \leq 200.5)$, which from our discussion of z-scores is also equal to $P(-0.05 \leq N_{0,1} \leq 0.05)$. Using a table of values, we can find $P(N_{0,1} \leq -0.05) = 0.48006$ and $P(N_{0,1} \leq 0.05) = 0.51994$, so the desired probability is $0.51994 - 0.48006 = 0.03988 \approx 3.99\%$.
  - For (ii) we wish to compute $P(202.5 \leq N_{200,10} \leq 208.5)$, which from our discussion of z-scores is also equal to $P(0.25 \leq N_{0,1} \leq 0.85)$. Using a table of values, we can find $P(N_{0,1} \leq 0.25) = 0.59871$ and $P(N_{0,1} \leq 0.85) = 0.80234$, so the desired probability is $0.80234 - 0.59871 = 0.20363 \approx 20.36\%$.
  - As we can see from our calculations, the normal approximation is very good! In fact, the use of the normal distribution to approximate the binomial distribution was, historically speaking, one of the very first applications of the normal distribution.

- Example: Use the normal distribution to estimate the probability that if 420 fair dice are rolled, the total of all the dice rolls will be between 1460 and 1501 inclusive.

  - We saw earlier that $\mu_X = \frac{7}{2}$ and $\sigma_X = \sqrt{\frac{35}{12}}$ for the random variable X giving the outcome of one roll.
  - Since we are summing the results of 420 independent samplings of X, the central limit theorem tells us that the overall distribution of the sum will be closely approximated by a normal distribution with mean $420\mu_X = 1470$ and standard deviation $\sqrt{420}\mu_X = 35$.
  - The desired probability is then $P(1459.5 \leq N_{1420,35} \leq 1501.5) = P(-0.3 \leq N_{0,1} \leq 0.9) \approx \boxed{43.39\%}$.

- Another fundamentally important property of the normal distribution is that it is stable, in the sense that the sum of any number of independent normal distributions is also a normal distribution:

- Proposition (Stability of Normal Distribution): If $X_1, X_2, \ldots, X_n$ are independent normally-distributed random variables with means $\mu_1, \ldots, \mu_n$ and standard deviations $\sigma_1, \ldots, \sigma_n$, then the sum $X_1 + X_2 + \cdots + X_n$ is also normally distributed with mean $\mu_1 + \cdots + \mu_n$ and standard deviation $\sqrt{\sigma_1^2 + \cdots + \sigma_n^2}$.
  - Proof: It is a moderately straightforward calculation using the joint probability distribution to show that the distribution of the sum of two normally-distributed variables is also normally distributed. (Alternatively, this can be derived from the central limit theorem.)
  - Thus, $X_1 + X_2$ is normally-distributed, hence so is $(X_1 + X_2) + X_3$, ... , and hence so is $X_1 + X_2 + \cdots + X_n$.
  - The statements about the mean and standard deviation follows because the expected value and variance for independent random variables are additive.
- We can use this stability property to analyze random variables that are obtained by summing or averaging normal distributions:
- Example: According to an airline's customer research, the weight of its passengers' bags is normally distributed with a mean of 10 kilograms and a standard deviation of 2 kilograms. To maximize efficiency, the airline needs to design the cargo hold to be the smallest size that is able to carry all of its passengers' bags 99.9% of the time. If each flight holds 49 passengers, how much weight should the cargo hold be designed for?
  - Since the total weight is the sum of 49 independent normal distributions each with a mean of 10 kilograms and a standard deviation of 2 kilograms, the total weight will also be normally distributed with a mean of $49 \cdot 10 = 490$ kilograms and a standard deviation of $2\sqrt{49} = 14$ kilograms.
  - In order to be able to carry all its passengers bags 99.9% of the time, we want the maximum capacity M to satisfy $P(N_{49,14} \leq M) = 0.999$. Using a computer or a table of z-scores, we can see that $M \approx 390 + 14 \cdot 3.0902 = \boxed{533.26}$ kilograms.
- Example: In a typical game, a basketball team attempts 15 one-point free throws, 60 two-point field goals, and 25 three-point field goals. If free throws, two-pointers, and three-pointers independently score 75%, 50%, and 35% of the time respectively, find (i) the expected number of points the team scores per game, (ii) the standard deviation in the number of points scored, and (iii) the approximate probability that the team will score at least 110 points.
  - The total number of free throws, two-pointers, and three-pointers will each be binomially distributed.
  - Since the values of np and $n(1-p)$ are fairly large for each of these three distributions, they will be well approximated by the corresponding normal distributions with the same mean and standard deviation. The total number of points will then be a weighted sum of these approximately normal distributions, hence will also be approximately normal.
  - The number of free throws has $n = 15$ and $p = 0.75$ hence the expected value is $15 \cdot 0.75 = 11.25$ with standard deviation $\sqrt{15 \cdot 0.75 \cdot 0.25} \approx 1.6771$.
  - The number of two-pointers has $n = 60$ and $p = 0.50$ hence the expected number of two-pointers is $60 \cdot 0.50 = 30$ with standard deviation $\sqrt{60 \cdot 0.50 \cdot 0.50} \approx 3.8730$. Since two-pointers are worth 2 points each, the expected number of points is $2 \cdot 30 = 60$ with standard deviation 7.7460.
  - In the same way, the number of three-pointers has $n = 25$ and $p = 0.35$, so the expected number of points from three-pointers is $3 \cdot 25 \cdot 0.35 = 26.25$ with standard deviation $3\sqrt{25 \cdot 0.35 \cdot 0.65} = 7.1545$.
  - Thus, the total number of points is (approximately) normally distributed with mean $11.25 + 60 + 26.25 = \boxed{97.5}$ and the standard deviation is $\sqrt{1.6771^2 + 7.7460^2 + 7.1545^2} \approx \boxed{10.6771}$.
  - For (iii), since the distribution is approximately normal, the most obvious estimate for the probability that the team will score at least 110 points is given by $P(N_{97.5,10.6771} \geq 110) = P(N_{0,1} \geq 1.1707) \approx 0.1209$.
  - However, because the distribution of points is discrete, we should use a continuity correction and instead compute $P(N_{97.5,10.6771} \geq 109.5) = P(N_{0,1} \geq 1.1239) \approx 0.1305$: this yields an estimate of roughly 13%.
- Example: A statistics instructor has two classes with 16 and 25 students respectively. She gives an exam to each student in each class, where the student scores are normally distributed with mean 80 and standard deviation 5. Find (i) the expected mean and standard deviation in each class, and also the probabilities that (ii) the average in the 16-student class is at least 81 points, (iii) the average in the 25-student class is less than 79 points, and (iv) that the average in the 16-student class is at least 1 point higher than the average in the 25-student class.

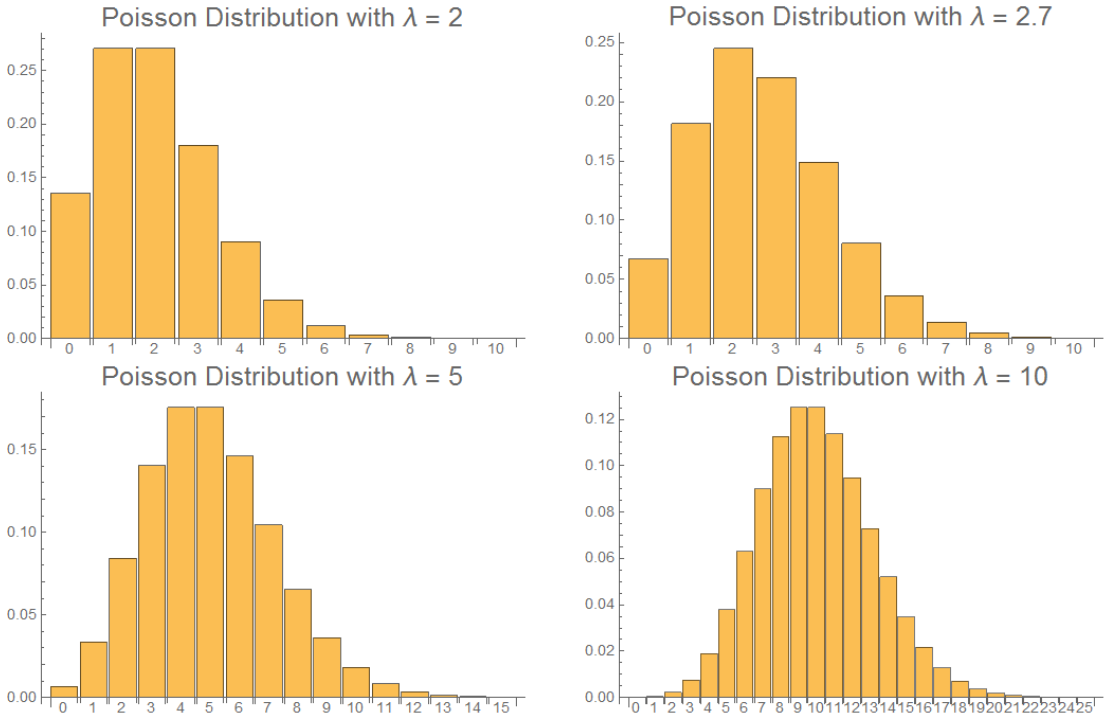
- Suppose the students in the two classes have scores $X_1, X_2, \ldots, X_{16}$ and $Y_1, Y_2, \ldots, Y_{25}$.
- Then the sum $X_1 + X_2 + \cdots + X_{16}$ will be normally distributed with mean $16 \cdot 80$ and standard deviation $5\sqrt{16} = 20$. This means that the average $\frac{1}{16}(X_1 + X_2 + \cdots + X_{16})$ will have mean $\boxed{80}$ and standard deviation $20/16 = \boxed{1.25}$.
- In the same way, $Y_1 + Y_2 + \cdots + Y_{25}$ will be normally distributed with mean $25 \cdot 80$ and standard deviation $5\sqrt{25} = 25$, so the average $\frac{1}{25}(Y_1 + Y_2 + \cdots + Y_{25})$ will have mean $\boxed{80}$ and standard deviation $25/25 = \boxed{1}$.
- For (ii), we want to find $P(N_{80,1.25} \geq 81) = P(N_{0,1} \geq 0.8) = 1 - P(N_{0,1} < 0.8) \approx \boxed{0.2119}$, or about 21.2%.
- For (iii), we want to find $P(N_{80,1} < 79) = P(N_{0,1} < 1) \approx \boxed{0.1587}$, or about 15.9%.
- For (iv), we want to understand the distribution of the difference $X - Y$, where X (the 16-student average) is normally distributed with mean 80 and standard deviation 1.25 and Y (the 25-student average) is normally distributed with mean 80 and standard deviation 1.
- The idea is to recognize that $X - Y = X + (-Y)$ and that $-Y$ is also normally distributed (now with mean -80 and standard deviation 1). Thus, by our results, the random variable $X + (-Y)$ will also be normally distributed with mean $80 + (-80) = 0$ and standard deviation $\sqrt{1.25^2 + 1^2} \approx 1.6008$.
- The desired probability is then equal to $P(N_{0,1.6008} \geq 1) = P(N_{0,1} \geq 0.6247) = 1 - P(N_{0,1} < 0.6247) \approx \boxed{0.2661}$, or about 26.6%.

- The ideas in this last example form the basis for many approaches in statistical testing, since these calculations give a way of determining how likely it is that a difference in sampling averages has occurred by chance, if the means of the distributions were actually equal.
- We will also mention that it is possible to obtain estimates of a similar form even when the underlying distribution is not normal.
  - Instead of relying on the central limit theorem, one must use the (comparatively much weaker) result of Chebyshev's inequality.
- Example: In a different statistics class, the exam scores are distributed with mean 80 and standard deviation 6, but are no longer known to be normally distributed. If the probability that the class average is within 1 point of the mean is at least 84%, find the minimal number of students in the class (i) with no additional information about the distribution, and (ii) if the central limit theorem is assumed to give a good approximation for the average score's distribution.
  - If there are n students in the class, then the average score has mean 80 and standard deviation $6/\sqrt{n}$.
  - For (i), Chebyshev's inequality says that the proportion of students scoring within k standard deviations of the mean is at least $1 - 1/k^2$, which is equal to 84% when $k = 5/2$.
  - Therefore, 1 point must represent 5/2 of a standard deviation in the average score: this means $1 = (5/2) \cdot (6/\sqrt{n})$ so that $n = \boxed{225}$.
  - For (ii), if the central limit theorem is assumed to give a good approximation, then the average score is approximately normally distributed with mean 80 and standard deviation $6/\sqrt{n}$, so we have $P(79 \leq N_{80,6/\sqrt{n}} \leq 81) = 0.84$.
  - Since the two tails of the normal distribution are symmetric, the condition is equivalent to saying that $P(N_{80,6/\sqrt{n}} \leq 79) = 0.08$, or, upon rescaling, that $P(N_{0,1} \leq \frac{79-80}{6/\sqrt{n}}) = 0.08$.
  - Using a computer or table for the inverse normal cdf indicates that $P(N_{0,1} \leq z) = 0.08$ holds for $z \approx -1.4051$, and so $\frac{79-80}{6/\sqrt{n}} = -1.4051$ so that $n \approx 71.07$, meaning that $\boxed{72}$ students would be the minimum.

2.3.3 The Poisson Distribution and Poisson Limit Theorem

- The next class of random variables we will discuss are the Poisson distributions:

- Definition: The Poisson distribution with parameter $\lambda > 0$ is the discrete random variable that takes the nonnegative integer value n with probability $\dfrac{\lambda^n e^{-\lambda}}{n!}$.
  - Here are some plots of Poisson probability distribution functions:



  - Note that this is in fact a valid probability distribution because $\sum_{n=0}^{\infty} \dfrac{\lambda^n e^{-\lambda}}{n!} = e^{-\lambda} \cdot \sum_{n=0}^{\infty} \dfrac{\lambda^n}{n!} = e^{-\lambda} e^{\lambda} = 1$, where we used the Taylor expansion $e^x = \sum_{n=0}^{\infty} \dfrac{x^n}{n!}$.
  - If X has a Poisson distribution with parameter λ, then $E(X) = \sum_{n=0}^{\infty} n \dfrac{\lambda^n e^{-\lambda}}{n!} = \lambda \sum_{n=1}^{\infty} \dfrac{\lambda^{n-1} e^{-\lambda}}{(n-1)!} = \lambda$, and also $E(X^2) = \sum_{n=0}^{\infty} n^2 \dfrac{\lambda^n e^{-\lambda}}{n!} = \sum_{n=1}^{\infty} \lambda \dfrac{\lambda^{n-1} e^{-\lambda}}{(n-1)!} + \sum_{n=2}^{\infty} \lambda^2 \dfrac{\lambda^{n-2} e^{-\lambda}}{(n-2)!} = \lambda + \lambda^2$, and so $\text{var}(X) = E(X^2) - E(X)^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$.
  - Thus, the expected value of a Poisson-distributed random variable is λ, and its variance is also λ.
  - It is also not hard to see that the highest peak of the Poisson distribution occurs at $n = \lfloor \lambda \rfloor$, the greatest integer less than or equal to λ.

- Example: If the random variable X has a Poisson distribution with $\lambda = 4$, find (i) $P(X = 2)$, (ii) $P(X < 4)$, and (iii) $P(X \geq 3)$.
  - Since λ is small, we can make a short table of the initial values of the probability distribution, since $P(X = n) = \dfrac{\lambda^n e^{-\lambda}}{n!}$:

n	0	1	2	3	4	5	6	7	8	9	10
$P(X = n)$	0.0183	0.0733	0.1465	0.1954	0.1954	0.1563	0.1042	0.0595	0.0298	0.0132	0.0053

  - From the table we see $P(X = 2) \approx \boxed{0.1465}$, $P(X < 4) = 0.0183 + 0.0733 + 0.1465 + 0.1954 \approx \boxed{0.4335}$, and $P(X \geq 3) = 1 - 0.0183 - 0.0733 - 0.1465 \approx \boxed{0.7619}$.

- The Poisson distribution arises in the analysis of systems having a large number of independent events each of which occurs rarely.
  - More specifically, suppose we would like to model the probability distribution of how often a rare event will occur in a fixed time window, under the assumption that on average the event will occur λ times in

the window and that occurrences are independent (meaning that the occurrence of one event does not affect the probability that a second will occur).

- We can approximate this situation by dividing the time interval into W possible "small windows" in which a rare event (occurring with probability $p = \lambda/W$) can either occur or not occur: we wish to find the probability distribution for the number of events that do occur.
- With this description, the probability distribution of this approximation will be the binomial distribution with W independent events and event probability $p = \lambda/W$, meaning that the probability of observing exactly n events is equal to $\binom{W}{n}p^n(1-p)^{W-n} = \binom{W}{n}(\lambda/W)^n \cdot (1-\lambda/W)^{W-n}$.
- However, this is an only an approximation to the original problem: to find the answer to the original question, we need to take the limit as $W \to \infty$. Our main result is that taking the limit yields a Poisson distribution:

- Theorem (Poisson Limit Theorem): Suppose $\lambda > 0$ is a fixed constant and $p = \lambda/W$. Then $\lim_{W\to\infty}\binom{W}{n}p^n(1-p)^{W-n} = \frac{\lambda^n e^{-\lambda}}{n!}$. Therefore, the probability distribution of the number of rare independent events occurring in a fixed interval, under the assumption that the average number of events per interval is λ, is Poisson with parameter λ.
  - Proof: We have $\binom{W}{n}p^n(1-p)^{W-n} = \frac{W(W-1)(W-2)\cdots(W-n+1)}{n!} \cdot (\frac{\lambda}{W})^n \cdot (1-\frac{\lambda}{W})^{W-n} = \frac{W(W-1)(W-2)\cdots(W-n+1)}{W\cdot W\cdot W\cdot\cdots\cdot W} \cdot \frac{\lambda^n}{n!} \cdot (1-\frac{\lambda}{W})^W \cdot (1-\frac{\lambda}{W})^{-n}$.
  - As $W \to \infty$, the first term $\frac{W(W-1)(W-2)\cdots(W-n+1)}{W\cdot W\cdot W\cdot\cdots\cdot W}$ has limit 1, the second term $\frac{\lambda^n}{n!}$ is a constant, the third term has limit $e^{-\lambda}$ by a standard application of L'Hôpital's rule, and the last term has limit 1. Thus, the product has limit $1 \cdot \frac{\lambda^n}{n!} \cdot e^{-\lambda} \cdot 1 = \frac{\lambda^n e^{-\lambda}}{n!}$, as claimed.

- The Poisson limit theorem serves as a sort of complement to the central limit theorem for binomial distributions: the central limit theorem says that as $n \to \infty$, the binomial distribution tends to a normal distribution when np and $n(1-p)$ are moderately large, while the Poisson limit theorem says that it tends to a Poisson distribution when np is small.
  - The practical outcome of the Poisson limit theorem is that the Poisson distribution can be used to model the occurrences of independent rare events.
  - In fact, one of the first historical applications of the Poisson distribution was to estimate the number of soldiers killed by horse-kicks each year in the Prussian cavalry. Other situations in which the Poisson distribution arises include the distribution of telephone calls received by a customer service center, the number of mutations created on a DNA strand during replication, the number of customers arriving at a restaurant or shop, the number of insurance claims received during a given month, the number of earthquakes during a given month, the number of goals scored by a hockey team during a game, and the number of decay events observed in a radioactive sample with a long half-life.

- Example: At a call center, customer service calls come in at a rate of 1.2 per hour. Find the probabilities that (i) in the next hour, there are no calls, (ii) in the next hour, there is exactly one call, (iii) in the next two hours, there are no calls, (iv) in the next two hours, there are at least 3 calls, and (v) in the next 30 minutes, there are no calls.
  - First, a Poisson model is reasonable for this problem, because calls are fairly rare (based on the average of 1.2 per hour) and they should be essentially independent of one another.
  - The given information says that the number of calls X in a one-hour window will have a Poisson distribution with parameter $\lambda = 1.2$.
  - Thus, the probability of having no calls is $P(X=0) = \frac{1.2^0 e^{-1.2}}{0!} \approx \boxed{0.3012}$ while the probability of exactly one call is $P(X=1) = \frac{1.2^1 e^{-1.2}}{1!} \approx \boxed{0.3614}$.

- The distribution of the number of calls Y in a two-hour window will also have a Poisson distribution (since the same logic given above still applies), but since the average number of calls in 2 hours is $2 \cdot 1.2 = 2.4$, the corresponding parameter is $\lambda = 2.4$.
- Thus, the probability of having no calls is $P(Y = 0) = \dfrac{2.4^0 e^{-2.4}}{0!} \approx \boxed{0.0907}$ while the probability of at least 3 calls is $P(Y \geq 3) = 1 - P(Y \leq 2) = 1 - \dfrac{2.4^0 e^{-2.4}}{0!} - \dfrac{2.4^1 e^{-2.4}}{1!} - \dfrac{2.4^2 e^{-2.4}}{2!} \approx \boxed{0.4303}$.
- The distribution of the number of calls Z in a 30-minute window will also have a Poisson distribution, but now with parameter $\lambda = 0.5 \cdot 1.2 = 0.6$. The probability of having no calls is therefore $P(Z = 0) = \dfrac{0.6^0 e^{-0.6}}{0!} \approx \boxed{0.5488}$.

- Example: Based on past history, a hospital determines that the average number of patients arriving at the emergency room between 2am and 3am is 5.3. Using a Poisson model, estimate the probability that between 2am and 3am today, (i) no patients arrive, (ii) more than 5 patients arrive, and (iii) 10 or more patients arrive. Next, (iv) describe the distribution of the total number of patients arriving between 2am and 3am over a full 366-day leap year, and find its mean and standard deviation, and (v) estimate the probability of getting at least 2000 patients that year. Finally, estimate the probabilities of seeing (vi) 10 or more patients at least 15 times this year, and (vii) seeing 0 patients at least twice this year.
  - We remark that a Poisson model is reasonable for this problem, because the arrival of patients is fairly rare based on the given average of 5.3 per hour, and it is also reasonable to assume that the arrivals of patients at the emergency room are essentially independent of one another.
  - For (i), the given information says that the number of patients seen on one day will have a Poisson distribution with parameter $\lambda = 5.3$. Thus, the probability of having no patients today is $e^{-5.3} \approx \boxed{0.00499}$, or about 0.5%.
  - For (ii), the probability of having more than 5 patients is $1 - P(X \leq 5) = 1 - e^{-5.3} - \dfrac{5.3e^{-5.3}}{1!} - \dfrac{5.3^2 e^{-5.3}}{2!} - \dfrac{5.3^3 e^{-5.3}}{3!} - \dfrac{5.3^4 e^{-5.3}}{4!} - \dfrac{5.3^5 e^{-5.3}}{5!} \approx \boxed{0.4365}$.
  - For (iii), the probability of having 10 or more patients is $P(X \geq 10) = 1 - P(X \leq 9) = 1 - e^{-5.3} - \cdots - \dfrac{5.3^9 e^{-5.3}}{9!} \approx \boxed{0.0441}$.
  - For (iv), the total number of patients over the full 366-day year is obtained by summing 366 independent Poisson-distributed random variables each with $\lambda = 5.3$. The resulting exact distribution is Poisson with parameter $\lambda' = 366 \cdot 5.3 = 1939.8$, so the mean is $\lambda' = \boxed{1939.8}$ and the standard deviation is $\sqrt{\lambda'} = \boxed{44.04}$.
  - Alternatively, we could invoke the central limit theorem to see that the distribution will be approximately normal with mean $366 \cdot 5.3 = \boxed{1939.8}$ and standard deviation $\sqrt{366 \cdot 5.3} \approx \boxed{44.04}$. (This is a reflection of the fact that for large λ, the Poisson distribution is approximately normal.)
  - For (v), we could use either the Poisson model or the normal model, but the normal model is much easier to calculate with. Using a continuity correction, the desired probability is $P(\# \geq 2000) = P(N_{\mu,\sigma} > 1999.5) = P(N_{0,1} > 1.3555) \approx \boxed{0.0876}$.
  - For (vi), the probability of having 10+ patients on any given day is approximately 0.0441 from (iii) above. The total number of times the hospital will see 10+ patients during the year will be binomially distributed with parameters $n = 366$ and $p = 0.0441$.
  - Since $np = 16.10$, the normal approximation to the binomial will be fairly good (with $\mu = np = 16.10$ and $\sigma = \sqrt{np(1-p)} = 3.9280$), so the desired probability is approximately $P(B_{366,0.0441} \geq 15) \approx P(N_{np,\sqrt{np(1-p)}} > 14.5) = P(N_{0,1} > -0.4177) = \boxed{0.6619}$.
  - For (vii), the probability of having 0 patients on any given day is approximately 0.00499 from (i) above. The total number of times the hospital will see 0 patients during the year will be binomially distributed with parameters $n = 366$ and $p = 0.00499$.
  - Since $np = 1.8269$ and n is large, the Poisson approximation to the binomial will be fairly good (with parameter $\lambda = np = 1.8269$), so the desired probability is approximately $P(B_{366,0.00499} \geq 2) \approx P(P_{1.8269} \geq 2) = 1 - e^{-1.8269} - 1.8269e^{-1.8269} = \boxed{0.5450}$.

2.3.4 The Exponential Distribution and Memoryless Processes

- The third class of random variables we will discuss are the exponential distributions. Recall the definition:
- Definition: The exponential distribution with parameter $\lambda > 0$ is the continuous random variable with probability density function $p(x) = \lambda e^{-\lambda x}$ for $x \geq 0$, and is 0 for negative x.
    - The cumulative distribution function is $c(x) = 1 - e^{-\lambda x}$ (for $x \geq 0$), and we also found earlier that the expected value and standard deviation are both $1/\lambda$.
- The exponential distribution is used to model "memoryless" processes, as follows:
- Definition: Suppose X is a continuous random variable measuring the waiting time for an event (such as the failure of a piece of equipment, the arrival of a customer to a shop, or the decay of a radioactive isotope). We say that X is memoryless if X has the property that the subsequent waiting time is independent of the amount of time already waited.
    - If a represents the total time already waited, and b represents the additional time before the event occurs, then this memoryless condition says that $P(X > a + b | X > a) = P(X > b)$ for every a and b.
    - Equivalently, this means $P(X > a + b) = P(X > a) \cdot P(X > b)$, since the event $X > a + b$ includes the event $X > a$.
    - But now, if X has an exponential distribution, then $P(X > a + b) = e^{-\lambda(a+b)} = e^{-\lambda a} e^{-\lambda b} = P(X > a) \cdot P(X > b)$. This means that exponentially-distributed random variables are memoryless. In fact, the exponential distributions are the only memoryless continuous probability distributions:
- Proposition (Memoryless Distributions): If X is a memoryless continuous random variable, then in fact X has an exponential distribution.
    - Proof: Observe that by continuity, the condition $P(X > a + b) = P(X > a) \cdot P(X > b)$ implies that $P(X > 2) = P(X > 1)^2$, $P(X > 3) = P(X > 1)^3$, and then $P(X > 4) = P(X > 1)^4$, and so forth.
    - By the same logic, we have $P(X > 1/n) = P(X > 1)^{1/n}$ for every integer n, so combining this reasoning with the argument above shows that $P(X > a) = P(X > 1)^a$ for every rational number $a > 0$.
    - But since $P(X > a)$ is a nondecreasing function of a, this means in fact $P(X > a) = P(X > 1)^a$ for every real $a > 0$.
    - Now writing $\lambda = -\ln[P(X > 1)]$ yields $P(X > a) = e^{-\lambda a}$, and so the cumulative distribution function agrees with that of the exponential distribution with parameter λ. This means X must be exponentially distributed with parameter λ, as claimed.
    - Remark: Essentially the same proof shows that the only memoryless discrete random variables are the geometric distributions with parameter p, in which $P(X = n) = p(1-p)^n$ for nonnegative integers n.
- Example: The usage time before a certain refrigerator model needs to be repaired is modeled as an exponential distribution. Customer surveys indicate that 20% of the refrigerators must be repaired within their first year of operation. Find the parameter λ for the distribution, and also the percentage of refrigerators that will last at least 5 years without needing to be repaired.
    - If X is the waiting time for repair, the given information says that $P(X < 1) = 0.20$. If the parameter is λ then since $P(X < 1) = 1 - e^{-\lambda}$ we see $1 - e^{-\lambda} = 0.20$ so that $\lambda = \boxed{-\ln(0.80)} \approx 0.2231$.
    - Then the proportion of refrigerators that will last at least 5 years is $P(X \geq 5) = e^{-5\lambda} = (0.80)^5 \approx 0.3277$, which is about $\boxed{33\%}$.
    - Remark: We could also have calculated the proportion directly using the memoryless property: the given information says that 80% of refrigerators last one year without being repaired, so of these, 80% will last another year, and 80% of those will last a third year, and so forth, for an overall proportion of $(0.80)^5$ that will last 5 years.

- Example: An unreliable ride-share service is supposed to take a customer to the airport. The average waiting time is 45 minutes, but the customer feels that the total amount of time she has waited so far has no relationship to the amount of additional time she will have to wait. If the customer uses the service 40 times a year, estimate the probability that the car actually shows up within 5 minutes at least 6 times out of the 40 uses.
    - The given information is describing a memoryless waiting time, so by our results, the waiting time will be exponentially distributed. Since the expected value is $1/\lambda$, that means $\lambda = 1/45$.
    - Then the probability that the car shows up within 5 minutes (in one use of the service) is $1 - e^{-5/45} = 0.1052$.
    - So, if the customer uses the service 40 times, the total number of times the car shows up within 5 minutes will be binomially distributed with parameters $n = 40$ and $p = 0.1052$.
    - Since $np = 4.2064$, we are in a situation where the Poisson approximation should be better. The resulting probability estimate is $P(P_\lambda \geq 6) = 1 - P(P_\lambda < 6) = \boxed{0.2479}$.
    - Remark: If instead we used the normal approximation, we would get a probability estimate $P(N_{\mu,\sigma} > 5.5) = P(N_{0,1} > 0.6658) = 0.2528$. The exact binomial probability is 0.2406, so we can see that the Poisson estimate is slightly more accurate than the normal estimate.
- As a final remark to finish our discussion, we will note that there is a connection between the Poisson distribution and the exponential distribution that arises from our interpretations of the processes they model.
    - The Poisson distribution models the number of occurrences of independently-occurring rare events in a particular interval of time, while the exponential distribution models the waiting time for a memoryless process.
    - Now suppose we have a Poisson-distributed phenomenon, and we ask: how long do we have to wait between two occurrences of the phenomenon?
    - Because the Poisson events are independent and rare, the occurrence of one does not affect the waiting time for the next one. Since this waiting time is memoryless, the distribution of waiting times between Poisson events will have an exponential distribution.
    - This fact that waiting times between Poisson events have an exponential distribution leads to some unintuitive results.
    - For example, the exponential distribution decreases rapidly, starting from 0: therefore, the distances between Poisson events are more likely to be small rather than big. Specifically, if the average distance is D (so the exponential parameter is $1/D$), the probability of obtaining a distance less than the average is then $1 - e^{-1/D \cdot D} = 1 - e^{-1} \approx 0.6321$.
    - Thus, despite the fact that the Poisson events will be uniformly distributed inside the time interval (since they are, after all, independent and occur randomly), it is nonetheless likely that we will observe "clusters" of occurrences.
    - Although this may seem peculiar, it is really just a reflection of the general fact that randomly-occurring events will tend to appear in clusters: it is, in fact, very unlikely for several independent random events to be spaced evenly apart.

Well, you're at the end of my handout. Hope it was helpful.