Name: ____

ID: ____

On your desk you should have only the exam paper, writing tools, and the cheat-sheet. The cheat-sheet is one page handwritten on both sides.

The exams are color coded. Your exam should have different color than that of your neighbours to the left, right and in front.

There are 11 questions in this exam, totalling 125 points. The final score is determined by summing all the points and taking the min of the sum and 100. For a final grade of A+ it is necessary (but not sufficient) to get more than 100 points in the final.

Be clear and concise. Write your answers **in the space provided** after each question. Whenever possible, **write your answers as expressions, not as numbers**. If an expression is reused later in the problem, you are encouraged to assign it to a letter variable and use the variable in the later expressions. For example, suppose in the first part of the problem you find the answer to be $C(52,5)$ and that in a later part the answer is $22/C(52,5)$. You can define in the first part $a = C(52,5)$ and write the second expression as $22/a$. This will save you time and space and is easier for us to grade.

Do your work at the empty space. If your answer is incorrect, we will look at your work and give you credit if you were going in the right direction.

1	10pt.	
2	10pt.	
3	10pt.	
4	10pt.	
5	10pt.	
6	10pt.	
7	10pt.	
8	10pt.	
9	10pt.	
10	10pt.	
11	10pt.	

Total 110

1.

You roll two fair dice: a red die and a green die. Note: since the two dices are different, a roll with green 4 and red 3 is different from a roll with green 3 and red 4.

1. How many different outcomes are there? ____________________

2. How many different outcomes are there in which the number **6** does not appear, i.e. **6** can NOT appear on either dice? ____________________

3. How many different outcomes are there in which the number **6** appears at least once, i.e. at least one dice has **6**? ____________________

2.

Tossing a coin

Consider an experiment where a fair coin is tossed N times. There is a natural outcome space for the experiment of tossing coins in sequence, where the probability of each outcome is equally likely. For example, if you toss 2 coins, the outcome space is $\{\{$H,H$\}$, $\{$H, T$\}$, $\{$T, H$\}$, $\{$T,T$\}\}$. The size of this outcome space is 4 and the probability of getting each outcome is 1/4.

Suppose a fair coin is tossed 9 times.

- What is the **size** of this outcome space? The **size** is the number of elements in the outcome space. ____________
- What is the **size** of the event set for getting exactly 8 heads? ____________
- What is the **probability** of getting exactly 8 heads? ____________
- What is the **probability** of getting at most 8 heads? ____________
- What is the **probability** of getting exactly 6 heads? ____________

3.

Putting cards in envelopes (Guided Problem)

This is a challenge problem, you have until wed to solve it.It is recommended that you try, this will help you understand what wedo in class.

Suppose we have 7 cards and 3 envelopes to put them in. The envelopes are marked (1,...,3).We will calculate the number of ways to put the cards into the envelopes under different conditions.

1. Suppose all the cards are **distinct** (i.e. are numbered(1,2,....,7)). How many ways are there to place cards intoenvelopes?Don't worry, this is easy! It's only the first part of the question. Think of placing the cards into envelopes in thefollowing way: take the first card and choose an envelope for it,then take the second card and choose an envelope for it etc. Untilall of the cards have been placed. Clearly we can get any possiblecombination this way. It takes a little thought, but you can alsoconvince yourself that there is only **one** way to get eachcombination. In other words, there is no overcounting.

   Counting the number of combinations we get this way is simple. The processtakes 7 steps. For each step, we place a card in one of the 3envelopes. Taking the product over all of these steps we get thatthe number of combinations is __________________ (Note: Envelopes chosen for a card can be chosen for other cards.)

2. Suppose that cards are **identical**. (The envelopes remain distinct)How many combinations are possible in this case?

   Consider the difference between part 1 and part 2. In part 1, each configurationspecified exactly **which** cards were placed in each envelope. Herethe cards are identical, therefore we can only say how **many** cardsare in each envelope, but we cannot identify them.

   Thinking of the problem this way, we realize that it ismathematically equivalent to the problem of choosing 7 candieswhen there are 3 **types** of candy to choose from. As the Candiesare indistinguishable, we are only interested in the number ofcandies chosen from each type. In this case, the cards are candies andthe envelopes are candy types.

   If you go back to this problem, you will recall the formula for it and use it to get theanswer: __________________

3. Suppose the cards are identical as in 2. and, in addition, werequire that each envelope contain at least one card. In thiscase we first have to check that the number of cards is at leastas large as the number of envelopes, otherwise there will be noway to satisfy the requirements. Luckily (thanks to the magic ofWebWork) $7 > 3$.

   OK, good. Now we can proceed in two steps. First, take 3 cardsand put one card into each envelope to satisfy the requirement-that each envelope contains a card. Secondly, place the remainingcards into ANY of the envelopes. We now have the same situationwe had in part 2 except instead of 7 cards, we have7-3=4 cards. We use the same formula that we usedin 2. to find that the answer is __________________

4.

A poker hand consisting of 7 cards is dealt from a standard deck of 52 cards.Find the probability that the hand contains exactly 3 cards of the same suite. It is allowed to have any number of cards in other suites.

First, we know the number of all possible hands of 7 cards is __________ .

Then, we calculate the number of hands that contain exactly 3 cards of the same suite.

We first choose which suite the 3 cards is. Obviously, there are __________ possibilities.

The number of possibilities for the ranks of these cards is __________ .

The other 4 cards in the hand can be any cards that have a different suite than the 3 cards. There are a total of __________ such cards. To choose 4 from them, there are __________ possibilities.

Thus we can compute the number of hands that have exactly 3 cards of the same suite, which is __________ .

Finally we can calculate the probability of such hands, by calculating th ratio of the number of such hands to the number of all hands. This is __________

5.

Remember, the deck you are using has 5 suits and 14 ranks.

1. The number of possibilities for the ranks of the two pairs is ________________ .
2. The number of possibilities for the rank of the single is ________________ .
3. The number of possibilities for the suits of the two pairs is ________________ .
4. The number of possibilities for the suit of the single is ________________ .
5. Thus the number of hands with exactly two pairs is ________________ .
6. The ratio of this number to the number of all hands ________________ .

6.

Straight : Five cards in sequence, but not all from the same suit

Remember, the deck you are using has 5 suits and 11 ranks.

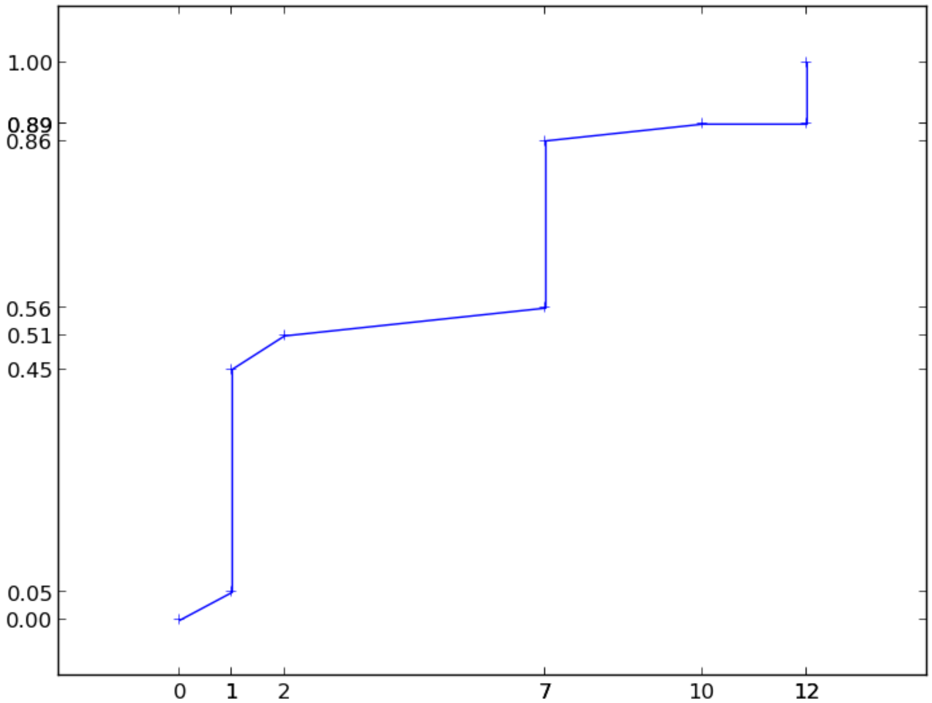
1. In the case of a standard deck, the ranks of a straight is one of (Ace,2,3,4,5) ... (10,J,Q,K,Ace). Similarly, in your deck, the number of possibilities for the ranks of a straight is __________________ .

2. The suits can be anything other than all equal, so the number of possibilities for the suits of a straight is __________________ .

3. Thus the number of hands that is a straight is __________________ .

4. The ratio of this number to the number of all hands __________________ .

7.
Given the following cumulative distribution:



What are the uniform densities and mass distributions that could have been summed to produce this CDF?
Uniform distributions:

- Uniform on (0.00, ________________) of probability density ________________
- Uniform on (1.00, ________________) of probability density ________________

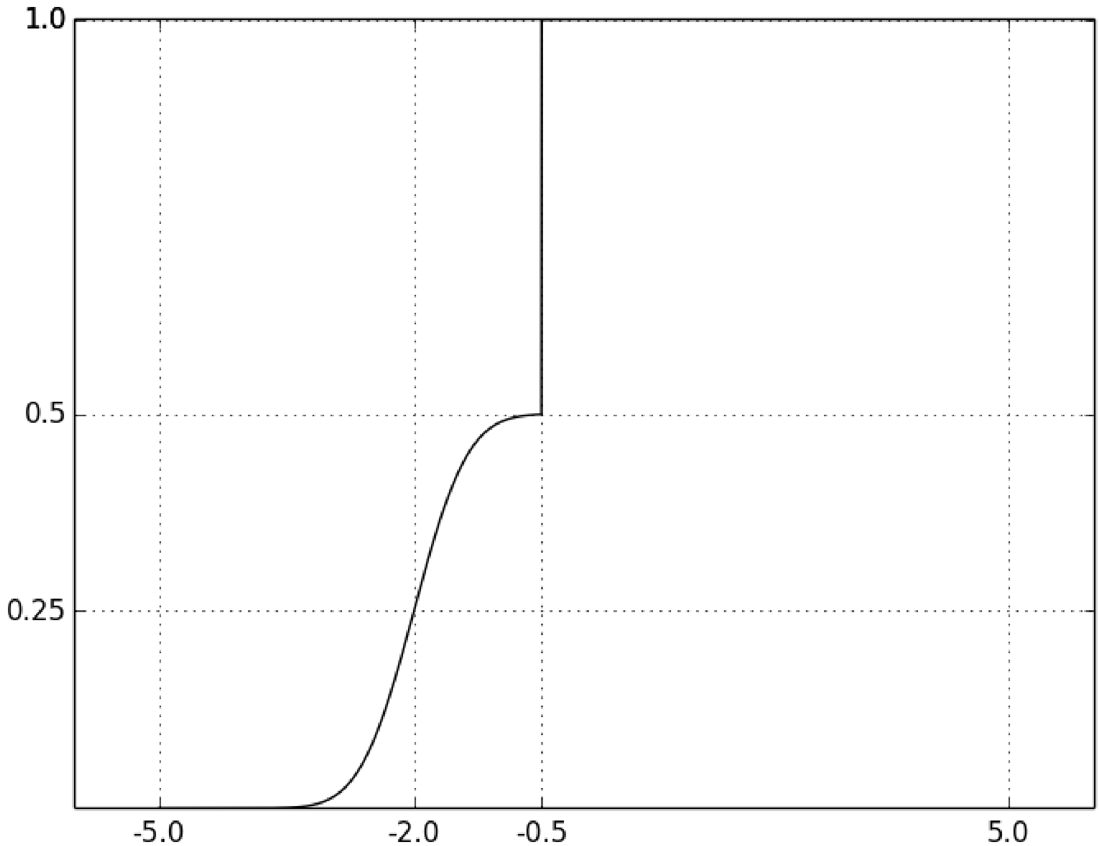
Point masses (ordered by the location):

- Point mass at ________________ with probability ________________
- Point mass at ________________ with probability ________________
- Point mass at ________________ with probability ________________

8.
Below is the CDF of a mixture distribution with **two** components.

All parameters of component distributions are small multiples of 0.5.

Component weights take on multiples of 0.05 and they need to sum to one.



This distribution can be expressed as a mixture of two: a normal distribution and a point mass: $p_1 N(\mu, \sigma) + p_2 PM(a)$

- The normal component $N(\mu, \sigma)$ has parameters $\mu =$ ________________ , and $\sigma = 0.5$. Its component weight is $p_1 =$ ________________
- The point mass $PM(a)$ is at location $a = -0.5$. Its component weight is $p_2 =$ ________________

9.

Suppose you have been dealt 4♡ and 5 ♡. What is the conditional probability that you will get a straight given that you have been dealt these two cards, and that the flop is "2♣, Q♣, K♢"? (Flop: The dealing of the first three face-up cards to the board.)

In this case we need a 3 and either a 6 or A on the turn and river.

- The number of such card pairs, ignoring order, is ____________________
- The conditional probability is ____________________

10.

Find the probability that a poker hand of 5 cards from a standard deck will contain no card smaller than 7 (i.e. 2,3...6) (call this event A), given that it contains at least one face card (i.e. J, Q, K) (call this event B)"? Note 'Ace' is the biggest card

We define two events:

- Event A: The hand contains no card smaller than 7.
- Event B: The hand contains at least one face card.

We first compute the size of the relevant events

- $|A \cap B| =$ __________________
- $|B| =$ __________________

We then use the ratio between the sizes of the events to find the conditional probability:

- The conditional probability $P(A|B) =$ __________________

11.

In Texas Hold'Em, a standard 52-card deck is used. Each player is dealt two cards from the deck face down so that only the player that got the two cards can see them. After checking his two cards, a player places a bet. The dealer then puts 5 cards from the deck face up on the table, this is called the "board" or the "communal cards" because all players can see and used them. The board is layed down in 3 steps: first, the dealer puts 3 cards down (that is called "the flop") followed by two single cards, the first is called "the turn" and the second is called "the river". After the flop, the turn and the river each player can update their bet. The winner of the game is the person that can form the strongest 5-card hand from the 2 hand in their hand and the 5 cards in the board.In previous homework you calculated the probability of getting each 5-card hand.Here we are interested in something a bit more complex: what is the probability of a particular hand given the cards that are currently available to the player.

The outcome space in this kind of problem is the set of 7 cards the user has at her disposal after all 5 board cards have been dealt. The size of this space is $|\Omega| = C(52,7)$

Suppose that A, B are events, i.e. subsets of Ω. We will want to calculate conditional probabilities of the type $P(A|B)$. Given that the probability of each set of seven cards is equal to $1/C(52,7)$ we get that the conditional probability can be expressed as:

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

Therefore the calculation of conditional probability (for finite sample spaces with uniform distribution) boils down to calculating the ratio between the sizes of two sets.**Suppose you have been dealt "4♡, 5♡".**

What is the conditional probability that you will get a straight given that you have been dealt these two cards, and that the flop is "2♣, 6♠, K♢"?

- Define B as the set {7-card hands that contain these 5 cards already revealed}.
- Define A as the set {7-card hands that contain a straight}.

The question is asking for $P(A|B)$. According to the formula above we need to find $|A \cap B|$ and $|B|$.

- In this case $A \cap B$ is the set {7-card hands that contain the 5 revealed cards **AND** contain a straight}. To get a straight, the remaining two cards (turn and river) must either be {7,8} or contain 3. We hence define two subsets within $A \cap B$:
  - S_1: {7-card hands that are in $A \cap B$ **AND** the remaining two cards are 7 and 8, regardless of order}.

$|S_1| =$ ________________

- S_2: {7-card hands that are in $A \cap B$ **AND** its turn and river contain 3}.

$|S_2| =$ ________________

Because there is no overlap in these two cases ($S_1 \cap S_2 = \emptyset$) and these two cases cover all possible valid hands ($A \cap B = S_1 \cup S_2$), by definition S_1 and S_2 form a *partition* of $A \cap B$, and we have $|A \cap B| = |S_1| + |S_2|$.

- Computing $|B|$ should be easy. 5 cards in the hand are already fixed, we have the freedom of choosing the turn and the river from the 47 cards in the deck. $|B| =$ ________________ .
- The conditional probability $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{|A \cap B|}{|B|} = \frac{|S_1|+|S_2|}{|B|}$ is ________________

12.

Like the previous question, suppose you have been dealt "4♡, 5♡".

1. **Suppose you have one opponent. What is the conditional probability that you will win, given these two cards in hand, and that the board is "3♢, 4♢, 4♣, 4♠, 5♢"?**

   - With this board, we have four of a kind. The only way the opponent will beat it is with a straight flush. How many possible two cards does the opponent have that can make a straight flush, regardless of order.? ____________________

   - The conditional probability that you will win is ____________________

13.

Three cards are drawn sequentially from a deck that contains 15 cards numbered 1 to 15 in an arbitrary order. Suppose the first card drawn is a 7, what is the probability that the three cards form an increasing sequence?

Note that unlike the previous question, we are considering a sequence instead of a set of cards, so the order matters. Define the event of interest, A, as the set of all increasing 3-card sequences, i.e. $A = \{(x_1, x_2, x_3) | x_1 < x_2 < x_3\}$, where $x_1, x_2, x_3 \in \{1, \cdots, 15\}$. Define event B as the set of 3-card sequence that starts with 7, i.e. $B = \{(x_1, x_2, x_3) | x_1 = 7\}$ or simply $B = \{(7, x_2, x_3)\}$.

- $|B| =$ __________________ .

It follows that $A \cap B = \{(7, x_2, x_3) | 7 < x_2 < x_3\}$. This set can be partitioned into subsets with different values of x_3, where in each subset x_3 is a constant: $A \cap B = \cup_{x_3=9}^{15} \{(7, x_2, x_3) | 7 < x_2 < x_3\}$.

Let $S_{x_3=t}$ represent the subset $\{(7, x_2, t) | 7 < x_2 < t\}$, then $|A \cap B| = \sum_{t=9}^{15} |S_{x_3=t}|$.

- To compute each $|S_{x_3=t}|$, let's start with a specific case, say, $t = 12$,

$|S_{x_3=12}| = |\{(7, x_2, 12) | 7 < x_2 < 12\}| =$ __________________ .

- Generalize this computation, it should be straightforward to compute $|A \cap B|$ as the sum of $S_{x_3=t}$ over all valid t.

$|A \cap B| =$ __________________

- Now we are ready to compute the conditional probability $P(A|B) =$ __________________

14.

Markov's inequality relate probabilities to expectations, and provide bounds for the cumulative distribution function of a random variable.

The Markov's inequality is stated as follow:

If X is any nonnegative random variable and a > 0, then $\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}$

John has a biased coin with $P(\text{heads}) = 0.5$.He tosses this coin N times and, out of the N times,the coin lands on heads 117 times.Using Markov's inequality that he learned from CSE103,he says that the probabality of seeing at least this manyheads is at most 0.6.

- ◦ Give the best lower bound, using Markov inequality, on the number of times John tossed the coin?N = ________________ (Provide the exact answer, don't round it to the next larger integer number.)

Suppose John lends you this coin. If you flip the coin 355 times,what is the upper bound of the probability of seeing at least 277 headsusing Markov's inequality?

- ◦ $P(\text{Number of heads} > 277) \leq$ ________________

15.
We will use Chebyshev's inequality in this problem. The Chebyshev's theorem is stated as below:

Let X be a random variable with finite expected value and finite non-zero variance ^{2}. Then for any real number $k > 0$, $\Pr(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$, which is the same as $\mathbb{P}(|X - \mathbb{E}(X)| \geq a) \leq \frac{\mathrm{Var}(X)}{a^2}$ for any a>0.

Suppose the mean noon-time temperature for September days in San Diego is 24° and the standard deviation is 3.4. (Temperature in this problem is measured in degrees celsius)

- Using Chebyshev's theorem, what is the **minimal** probability (in percents) that the noon-time temperature of a september days is between 17.2° and 30.8°? ____________________ %.
- On September 26, 1963, the all-time record of noon-time temperature in San Diego of 44° was hit. Assume the temperature distribution is symmetric around the mean, what is the Chebyshev bound for the probability of breaking (or tieing) this record? ____________________

16.

We have 2 9-sided dice. The first die is a normal fair die where each face has a probability of showing of $1/9$. However, the second die is rigged so that the probability of showing the largest face (9) is twice as high as of the other faces and all of the other faces have equal probabilities.

- ○ What is the expected value of the outcome from tossing the fair die? ____________________

- ○ What is the expected value of the outcome from tossingthe rigged die? ____________________

We throw the fair die 2 times, then the rigged die 2 times consecutively and sum up all the outcomes:

- ○ What is the expected value of the sum? ____________________

Let Y denote the sum from the previous part. If we know that

$P(Y > k) \leq 0.5$

- ○ According to Markov's inequality, what is k? ____________________

17.

Some Definitions

The **standard** normal distribution $\mathcal{N}(0,1)$ is a special normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$.

- The CDF of the standard normal distribution is denoted Φ:

$$\Phi(z) = P(Z \leq z).$$

- The complement of the CDF is called the **Q-function**:

$$Q(z) = P(Z > z) = 1 - \Phi(z).$$

While $\Phi(z)$ measures the probability mass of a "head" of the standard normal, $Q(z)$ measures the "tail". The values of Q are often tabulated for commonly used (z). One such table is *http://goo.gl/Szofn1*. One can also use *http://wolframalpha.com* to find numeric values for Q function.

- An alternative representation of the head probability is the **error function**, denoted erf. It is related to Φ by:

$$\Phi(x) = \frac{1}{2} + \frac{1}{2} erf(\frac{x}{\sqrt{2}})$$

For this set of assignment, you can use "Phi", "Q" or "erf" as functions in your answer.

What is the value of Q(0.3)? ____________________

Approximate Binomial Using Normal Distribution

We know that the number of questions the monkey guesses correctly follows a binomial distribution, and we have computed the **exact** probability of a tail of this distribution, by summing up all cases in the tail.

Now, we assume the number of questions is large enough so that by **central limit theorem** we can use a normal distribution to approximate the binomial distribution. This makes computing the probability of a certain part of the distribution much easier.

Again, suppose the monkey is taking a multiple-choice test that consists of 21 questions each with 5 possible answers, let us estimate the probability that it gets at least 8 questions correct, this time using an **approximated normal distribution**.

Suppose X is the number of correct answers.

- What is the mean of X ? ____________________
- What is the standard deviation of X ? ____________________

- What is the z-score of $X = 8$? ____________________

- What is the estimated probability that $X \geq 8$? ____________________

18.
65 numbers are rounded off to the nearest integer and then summed. The roundoff operation introduces an error into the resulting sum. We would like to estimate this error.

We assume that the individual round-off errors are independent and uniformly distributed over $(-.5, .5)$.

Remember: a random variable that follows a uniform distribution over (a,b) has a mean of $(a+b)/2$, and variance of $\frac{1}{12}(b-a)^2$.

What is the mean of the round-off error of one number? ____________________

What is the standard deviation of the round-off error of one number? ____________________

Suppose the **sum** of the round-off errors is Y.

What is the mean of Y? ____________________

What is the standard deviation of Y? ____________________

To compute the probability that the resultant sum of the 65 numbers differs from the exact sum by more than 3, we find the two tails on the distribution of Y.

What is the z-score at $Y = 3$? ____________________ .

What is the probability that the rounded sum is **larger** than the exact sum by more than 3 ? ____________________

What is the probability that the rounded sum **differs** from the exact sum by more than 3 ? ____________________

19.

An airline company is considering a new policy of booking as many as 377 persons on anairplane that can seat only 330.(Past studies have revealed that only 81% of the booked passengers actually arrive for the flight.)

What is the mean of the number of passengers that arrive for the flight ? ____________________

What is the standard deviation ? ____________________

Estimate the probability that if the company books 377 persons, not enough seats will beavailable. ____________________

20.

In this problem, you may use the CLT simulation(http://webwork.cse.ucsd.edu/misc/clt.html) to help you find theanswers.

Suppose a sample average is denoted by $S_n = (\sum_{i=1}^{n} X_i)/n$ and we define $T_n = (S_n - A)/B$.

Find the values of $A$and B under the following conditions so that T_n isdistributed normally with $\mu = 0$ and $\sigma = 1$. Your answers should be correct up to 2 decimal points.

Hint : Use the expectation and variance of the Uniform and Exponential Distributions. For properties of uniform distribution refer to http://en.wikipedia.org/wiki/Uniform_distribution_(continuous) and for exponential distribution refer to http://en.wikipedia.org/wiki/Exponential

- $X_i \sim \text{Uniform}(-1,3)$ and $n = 250$
  - A = ____________________
  - B = ____________________
- $X_i \sim \text{Exponential}(0.4)$ and $n = 160$
  - A = ____________________
  - B = ____________________

21.

$\sum_{i=1}^{\infty} \frac{1}{i} = \infty$

Use the following approximations:

$\sum_{i=1}^{\infty} \frac{1}{i^2} \approx 1.645$

$\sum_{i=1}^{\infty} \frac{1}{i^3} \approx 1.202$

$\sum_{i=1}^{\infty} \frac{1}{i^4} \approx 1.082$

In WebWork, you can use inf in an answer box to denote infinity.And use -1 to denote undefined.

Let X be a random variable over the integers $\mathbb{Z} = \{... - 2, -1, 0, 1, 2, ...\}$.

Let $P(X = 0) = 0$ and for $i \neq 0$ let $P(X = i) = \frac{1}{Z_\alpha |i|^\alpha}$ where $Z_\alpha = \sum_{i=-\infty}^{\infty} 1/|i|^\alpha$. Note that Z_α needs to be finite for this distribution to be well defined.

Notice, by having Z_α in the denominator, we can ensure that $P(X = i)$ is a probability distribution. This is since $\sum_{i=-\infty}^{\infty} P(X = i) = \frac{\sum_{i=-\infty}^{\infty} 1/|i|^\alpha}{\sum_{i=-\infty}^{\infty} 1/|i|^\alpha} = 1$

For $\alpha = 2$:

- What is $E[X]$? ________________
- What is $var[X]$? ________________
- What is $std[X]$? ________________

For $\alpha = 3$:

- What is $E[X]$? ________________
- What is $var[X]$? ________________
- What is $std[X]$? ________________

For $\alpha = 4$:

- What is $E[X]$? ________________

- What is $var[X]$? ____________________

- What is $std[X]$? ____________________

Hint: To compute both the expected value and variances above, split the infinitesummation $\sum_{i=-\infty}^{\infty} ...$ into three parts:

- $\sum_{i=-\infty}^{-1} \cdots$

- $P(X=0)=0$

- $\sum_{i=1}^{\infty} \cdots$

22.

$\sum_{i=1}^{\infty} \frac{1}{i} = \infty$

Use the following approximations:

$\sum_{i=1}^{\infty} \frac{1}{i^2} \approx 1.645$

$\sum_{i=1}^{\infty} \frac{1}{i^3} \approx 1.202$

$\sum_{i=1}^{\infty} \frac{1}{i^4} \approx 1.082$

In WebWork, you can use inf in an answer box to denote infinity.And use -1 to denote undefined.

Let X be a random variable over the integers $\mathbb{Z} = \{... -2, -1, 0, 1, 2, ...\}$.

Let $P(X = 0) = 0$ and for $i \neq 0$ let $P(X = i) = \frac{1}{Z_\alpha |i|^\alpha}$ where $Z_\alpha = \sum_{i=-\infty}^{\infty} 1/|i|^\alpha$. Note that Z_α needs to be finite for this distribution to be well defined.

Notice, by having Z_α in the denominator, we can ensure that $P(X = i)$ is a probability distribution. This is since $\sum_{i=-\infty}^{\infty} P(X = i) = \frac{\sum_{i=-\infty}^{\infty} 1/|i|^\alpha}{\sum_{i=-\infty}^{\infty} 1/|i|^\alpha} = 1$

For this part, let X be defined as above for $\alpha = 4$.

Define the random variable $Y_{53} = \sum_{i=1}^{53} X_i$, X_i independent identically distributed according to X.

Using linearity of expectation, what is $E[Y_{53}]$? ____________________

Using the fact that the X_i random variables are independent and identicallydistributed, what is $var[Y_{53}]$? ____________________

What about the the standard deviation $std[Y_{53}]$? ____________________

Using Chebyshev's inequality, what is a bound on the $P(|Y_{53}| > 30)$? ____________________

RECALL:

Chebyshev's inequality states:

Let X be a random variable with finite expected value μ and finite non-zero variance σ^2. Then for any real number $k > 0$, $\Pr(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$, which is the same as $\mathbb{P}(|X - \mathbb{E}(X)| \geq a) \leq \frac{\mathrm{Var}(X)}{a^2}$ for any a>0.

23.
Let $X_1, X_2, \ldots, X_{100}$ be the outcomes of 100 independent rolls of a fair coin.
$P(X_i = 0) = P(X_i = 1) = 0.5$

1. $\mathbb{E}(X_1) =$ ____________________
2. $var(X_1) =$ ____________________

Define the random variable X to be $X_1 - X_2$.

1. $\mathbb{E}(X) =$ ____________________
2. $var(X) =$ ____________________
   Hint: if X, Y are independent random variables then $var(X+Y) = var(X) + var(Y)$

Define the random variable $Y = X_1 - 2X_2 + X_3$.

1. $\mathbb{E}(Y) =$ ____________________
2. $var(Y) =$ ____________________
   Hint: if a is a constant and X a random variable, then $var(aX) = a^2 var(X)$.

Define the random variable $Z = X_1 - X_2 + X_3 - X_4 + \ldots + X_{87} - X_{88}$.

1. $\mathbb{E}(Z) =$ ____________________
2. $var(Z) =$ ____________________

24.

You will be using Poisson Distribution in this problem. Review: a discrete random variable X is said to have a Poisson distribution with parameter $\lambda > 0$, if for k = 0, 1, 2, ... the probability mass function of X is given by:

$\Pr(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}$

Assume that you live in a district of size 8 blocks by 8 blocks so that the total district is divided into 64 small squares. How likely is it that the square in which you live will receive 1 hits if the total area is hit by 400 bombs. Assume the probability that a particular bomb will hit your square with probability $1/64$.

- ∘ What is λ in the Poisson Distribution? __________________
- ∘ Using Poisson Distribution, what is the approximate probability that your square will receive 1 hits? __________________
- ∘ What is the expected number of squares that will receive exactly 1 hits using the approximate probability from above? __________________

25.

There is a typesetter who, on average, makes one mistake per 900 words. Assume that he is setting a book with 90 words to a page. Let S_{90} be the numberof mistakes that he makes on a single page.

- What is the expected value of S_{90}? ____________________
- What is the exact probability that $S_{90} = 2$ (i.e. using the binomial distribution)? ____________________
- What is the Poisson approximation of $S_{90} = 2$? ____________________

26.

Pick a random permutation of $(1,2,\ldots,n)$. Let X_i be the number that ends up in the$i$th position. For instance, if the permutation is $(3,2,4,1)$ then$X_1 = 3$, $X_2 = 2$, $X_3 = 4$, and $X_4 = 1$.

(a) What is the expected number of positions at which $X_i \neq i$ (i.e. the number of *wrong* positions)?

Let random variable D represents the number of wrong positions, we aim to find $\mathbb{E}(D)$.

If we devise a new r.v. $Y_i = \{0,1\}$ to represent whether or not $X_i \neq i$, then it is easy to see that, $D = Y_1 + Y_2 + \cdots + Y_n$.The linearity of expectation gives:$\mathbb{E}(D) = \mathbb{E}(Y_1 + Y_2 + \cdots + Y_n) = \mathbb{E}(Y_1) + \mathbb{E}(Y_2) + \cdots + \mathbb{E}(Y_n)$. Notice that all positions are equivalent, so all Y_i have the same distribution.

We can easily compute $\mathbb{E}(Y_i) = 0 \cdot \mathbf{Pr}(X_i = i) + 1 \cdot \mathbf{Pr}(X_i \neq i) =$ ________________ .

It follows that,$\mathbb{E}(D) =$ ________________ .

(b) What is the expected number of positions at which $X_i = i+1$?

Similar to the previous question, we let D represent the number of positions at which $X_i = i+1$, and let $Y_i = 0,1$ represent whether or not $X_i = i+1$ at a specific position. Again we use the linearity of expectations. Notice that this time, not all Y_i are the same, because Y_n is always 0, while the other Y_i's have some chances of taking both values.

In other words $\mathbb{E}(Y_n) =$ ________________ (Note: $Y_i = 1$ represents $X_i = i+1$)

and for all $1 \leq i < n; \mathbb{E}(Y_i) =$ ________________

Using the linearity of expectation we get that$\mathbb{E}(D) =$ ________________ .

(c) What is the expected number of positions at which $X_i \geq i$?

In this part, the different Y_i's have different distributions, but you should be able to compute each of $\mathbb{E}(Y_i)$ easily.

$\mathbb{E}(Y_i) =$ ________________ .

$\mathbb{E}(D) =$ ________________ .
Hint: use the equality: $1 + 2 + \cdots + n = \frac{n(n+1)}{2}$

(d) What is the expected number of positions at which $X_i > \mathbf{max}(X_1, \ldots, X_{i1})$?

In this part, $\mathbb{E}(Y_i)$ is not so obvious. We know that $\mathbb{E}(Y_i) = \mathbf{Pr}(X_i > \mathbf{max}(X_1, ..., X_{i1}))$, but how do we compute this probability ?

We are going to use the combinatorial method. Fix the value of i.Let A_i be the set of all permutations which obey the condition $X_i > \mathbf{max}(X_1, ..., X_{i1})$. We will calculate $|A_i|$

Let us design a method for constructing the elements of A_i. We first choose a **set** S_i of i different numbers from 1 to n to put in the bins X_1 through X_i. The largest of these i numbers will be X_i, and the remaining $n-i$ numbers can be assigned arbitrarily to $X_{i+1}, \cdots, X_n$.

1. How large is the sample space, i.e. how many possibilities are there for choosing the **set** S_i when there is no restriction on the values for X_i other than that they are a subset of $\{1, \ldots, n\}$? ________________ (Note that S_i is a set, i.e. order does not matter).
2. How many ways are there to place the elements of S_i into the bins X_1 through X_i? ________________
   Hint: recall that X_i has to be the largest of the i elements, therefor only $i-1$ elements can be freely placed in positions X_1 through X_{i-1}.
3. For positions after i, we can arbitrarily assign the unused $n-i$ numbers, so that gives us ________________ permutations for position $i+1$ through n.
4. taking the product of the numbers we calculated in 1,2,3 we find that $|A_i| =$ ________________
5. Finally we know the size of the entire outcome space is $n!$, dividing by $n!$ we get that

$$\mathbb{E}(Y_i) = P(A_i) = \frac{|A_i|}{n!} = \underline{\hspace{4cm}} \text{ which simplifies to } \mathbb{E}(Y_i) = \underline{\hspace{4cm}}$$

Now you should be able to compute $\mathbb{E}(D) = \sum_{i=1}^{n} \mathbb{E}(Y_i)$

For large n, $\mathbb{E}(D) \approx$ ________________ .

Hint: use the approximation $\sum_{i=1}^{n} 1/i \approx \ln n$

27.

Suppose you have an algorithm A for your problem that always returns the correct answer, but takes different amounts of time each time it runs. Let X_n be the random variable representing the time algorithm A takes to complete for input of size n. Time can't be negative, so $X \geq 0$.
We call A a *Las Vegas Algorithm* if for any input size n there is a $T(n)$ so that $\mathbb{E}(X_n) = T(n)$.

- Side note: it isn't always the case that a random variable has finite expectation, even the values it can take on are finite. The assumption that there is some $T(n)$ for any n is a non-trivial assumption.

Let's say we have algorithm that determines whether any integer is prime. For an integer input that takes up n bits, the algorithm takes n seconds to run with probability $1/2$. With probability $1/2$ the algorithm takes 1 second to run. What is $T(n)$, in seconds? __________________ .

Let's say you have some Las Vegas algorithm A that runs in *expected time* $T(n)$ for problem size n. What you would prefer, however, is an algorithm that *always* finishes in time $O(T(n))$, but may have up to a 5% probability of returning the wrong answer. We will construct an algorithm A' from A that satisfies these properties.

Recall Markov's inequality: for some random variable $Y \geq 0$, $P(Y >= a) \leq \mathbb{E}(Y)/a$.
Fixing some n, apply Markov's inequality to get an upper bound on $P(X_n >= cT(n))$: __________________ .
What is c such that $P(X_n >= cT(n)) <= 0.05$? __________________

Thus an algorithm A' is as follows:

- Run A until time __________________ T(n). If A has completed, return the correct value.
- Else, return a random value.

This type of algorithm we have, where the algorithm completes in deterministic time $Q(n)$ but is correct with some probability is called a *Monte Carlo Algorithm*.

Recap:

- For "Las Vegas" algorithms, uncertainty is in the algorithm *runtime*. The algorithm is always correct.
- For "Monte Carlo" algorithms, uncertainty is in the algorithm *correctness*. The algorithm completes in deterministic time $T(n)$.

28.
You are given a randomized algorithm A for testing whether or not a number is prime. The algorithm is correct with probability $2/3$. More precisely, you can repeatedly run A on the same number m, and each time it outputs the correct answer with probability $2/3$.

To reduce the probability of error, you run A(f) n times, and return the majority answer. What should n be if you want the probability of error to be less than 0.05?

Let C_i be a binary random variable indicating whether the i^{th} execution of algorithm A is correct. Let $C = (C_1 + C_2 ... C_n)/n$.

- What is the minimum value of C such that our method of returning the majority answer will be correct? ____________________
- What is $\mathbb{E}(C)$? ____________________
- What is $var(C)$? ____________________ (Use n as variable in this answer)

Approach 1:Chebyshev's inequality says for random variable Y with mean μ and for any positive number $a > 0$, \$\$P(|Y-\mu| \geq a) \leq var(Y)/a^2\$\$

- Using Chebyshev's inequality, what is an upper bound on the probability your "majority algorithm" is incorrect? ____________________ (Use n as variable in this answer)
- What is the smallest **integer** value for n so that the probability that the "majority algorithm" makes an error is at most 0.05? ____________________ (Give a numerical answer)

Approach 2:Using Central Limit Theorem, approximate the distribution of C as a normal.

- What is the z-score of $C = 0.5$ ____________________ (Use n as variable in this answer)
- What is the smallest **integer** value for n so that the "majority algorithm" makes an error with probability at most 0.05? ____________________ (Give a numerical answer. Use $Q^{-1}(0.05) = 1.6449$, where Q^{-1} is the inverse of Q function)

Notice that n obtained with Approach 2 is much smaller than that obtained with Approach 1, this is because using the normal approximation and Q function give us a much tighter bound than the Chebyshev bound. (The Q function drops exponentially fast as the value deviates from the mean, while the Chebyshev bound drops only quadratically fast)

29.

In this problem, we will analyze the expected running time of a variant to the randomized percentile finding algorithm discussed in lecture. The algorithm is used to select the k^{th} smallest element in an array containing n elements. In other words, the element that would appear in location k if the array is sorted from smallest to largest.

In this version of percentile finding algorithm we change the the way we select the pivot element. Suppose instead of a single pivot we pick 5 numbers, sort them, and select the pivot to be the median of the numbers.

We say that a pivot is "lucky" if it falls in the range of $0.2n$ and $(1-0.2)*n$. When the pivot is lucky we are guaranteed that the size of the array at the following iteration will be at most $(1-0.2)*n$.

Pick one of the 5 numbers that we have sampled. What is the probability that this number is not in the range of $0.2n$ and $(1-0.2)*n$ __________________ .

.

Hint : Pr(A number is not in range) = Pr(The number is below the required range) + Pr(The number is beyond the required range).

.

What is the probability that the median of the 5 numbers that we sample is not in the range of $0.2n$ and $(1-0.2)*n$ __________________ .

.

Hint : Pr(Median of our sample is not in range) = Pr(Median and the numbers smaller than the median are below the required range) + Pr(Median and the numbers greater than the median are beyond the range).

.

What is the probability that the median of the 5 numbers that we sample is in the range of $0.2n$ and $(1-0.2)*n$ __________________ .

.

Now, let us use the results computed above to build a recurrence relation to estimate the expected running time of the algorithm..

.

Let $T(n)$ denote the expected running time of the algorithm with input size n..

.

When we get lucky, our problem reduces to a size of $(1-0.2)*n$.When we are unlucky, our problem size remains n..

.

$T(n) \leq n + aT((1-0.2)*n) + bT(n)$.

.

In the recurrence relation, what is the value of a __________________ In the recurrence relation, what is the value of b __________________ .

.

Solving the recurrence relation we get $T(n) \leq C*n$What is the value of C __________________ .

.

What is the expected number of random splits before we see a lucky split __________________

30.

Let S be a set of n different numbers (think of n as being very large).Suppose we want to find the k^{th} biggest element in S, for any $k \leq n$.Of course, we could just sort S and then pick out the answer, but this would take $n \log n$ time. Can we do better?

In this problem, we will look at an (expected) linear-time algorithmfor finding the k^{th} biggest element in S, for any $k \leq n$.The algorithm is randomized, and we'll specify it soon.

First suppose we pick a random element of S - call it v - and let $S_L = \{x \in S \mid x < v\}$ and $S_U = \{x \in S \mid x > v\}$ be the elements less than and greater than v.

(a) What is the probability that $|S_L|$, the size of S_L, is equal to 7? ____________________ (Hint: what value(s) can v take for this to happen? What's the probability of choosing such v?)

(b) What is the approximate probability that $|S_L|$ is *between* $\lceil n/4 \rceil$ and $\lceil 3n/4 \rceil$? (Round your answer to one decimal place.) ____________________

If v is chosen so that $\mid S_L \mid \in [n/4, 3n/4]$, then this implies $|S_U| \in [n/4, 3n/4]$ as well, so each of the two sets has a significant fraction of the elements.We'll call such a choice of v *lucky*.

(c) We'll consider a randomized algorithm for finding the k^{th} biggest element in S - call this value $f(k,S)$. (For example, $f(1,S)$ would be the maximum element in S, and $f(\frac{1}{2}|S|, S)$ the median.)

The algorithm chooses an element v at random from S and computes $|S_L|$ and $|S_U|$, to determine if this choice of v is lucky. If not, it chooses v again at random and repeats until it chooses a lucky v.

When a lucky v is chosen, the algorithm computes $|S_L|$ and $|S_U|$, and then does one of two things:

1. If $|S_L| \geq k$, then the k^{th} biggest element in S is in S_L. Actually, it must be the k^{th} biggest element in S_L as well, so in this case the algorithm just recursively finds $f(k, S_L)$.

2. If $|S_L| < k$, then the k^{th} biggest element in S is in S_U. Specifically, it must be the $(k - |S_L|)^{th}$ biggest element in S_U, so in this case the algorithm just recursively finds $f(k - |S_L|, S_U)$.

Note that in both cases, we end up recursively working on a set (S_L or S_U) that is at most 3/4 the size of the current set S, because our choice of v is lucky.

However, our lucky choices take a random time to generate because of the randomized way in which they're chosen. Our randomized algorithm continues choosing values of v one at a time independently, until it makes its first lucky choice. Using the answer to part (b) and how our randomized algorithm works, the expected number of random choices of v required to generate a lucky choice is ____________________

(d) Finally, consider the overall runtime of our randomized selection algorithm. This is a random variable. We will look at the *expected* runtime $T(n)$ on problem size n (size of S).

Suppose b is the answer to part (b) above. Then every time the algorithm tries a random choice of v, there is a probability b that the choice is lucky, in which case the remaining runtime is no more than $T(3n/4)$; and a probability $1-b$ that the choice is unlucky, in which case the algorithm basically starts again from scratch with problem size n. Also, the random choice itself takes time n, to verify whether it is lucky or not by building S_L and S_U.

Putting this together,$T(n) \leq n + bT(3n/4) + (1-b)T(n)$. Rearranging terms,$T(n) \leq \frac{n}{b} + T(3n/4)$.Solving this inequality,$T(n) \leq \frac{n}{b} + \frac{3n}{4b} + T((3/4)^2 n) \leq \frac{n}{b}\left(1 + \frac{3}{4} + \left(\frac{3}{4}\right)^2\right) + T((\frac{3}{4})^3 n)$.

Repeating this process,$T(n) \leq \frac{n}{b}\left(1 + \frac{3}{4} + \left(\frac{3}{4}\right)^2 + \dots\right) + T(s)$, where s is some value less than 1.Substituting b, solve for $T(n)$: $T(n) \leq$ ________________

31.
Questions regarding Karger's min-cut algorithm.

1. An undirected graph has 19 nodes and 49 edges. Give a tight upper bound on the size of the min-cut. ________________

2. Suppose a graph has 31 edges and the order of a particular node,node a, is 9 . An edge is chosen uniformly at random. What isthe probability that the chosen edge has a as one of its end points? ________________

3. We have shown that the probability of hitting the min-cut set in asingle iteration is at most $(2/n)$ where n=13 is the number of nodes in the graph. What is the probability of not hitting the min-cut set in anyof the first 3 iterations? ________________
What is the probability of not hitting the min-cut set in any of the last 3 iterations? ________________

32.

To refresh your memory:

- We represent documents by the set of vocabulary indices contained in those documents. If we had an English dictionary indexed by a=0, aah=1, ...the=97124,apple=4337,...zyzzyvas=109581, the sentence "The apple on the tree is red" would be transformed into $\{4377, 47647, 50764, 78524, 97124, 99593\}$ since "the"=97124, "apple"=4337, ... Notice "the"=97124 is only counted once, even though it appears twice in the sentence. Henceforth call the set of vocabulary indices V.
- The Jaccard similarity between two documents represented by sets $A \subseteq V$ and $B \subseteq V$ is $|A \cap B|/|A \cup B|$.
- A minhash h hashes documents to elements of V. For documents A,B, $Pr(h(A) = h(B)) = |A \cap B|/|A \cup B| = J(A,B)$, the Jaccard similarity between A and B.

Thus, we can use a number of independent min-hashes $h_1, ...h_n$ to compute an approximation $H = \frac{\sum_1^n \mathbb{1}\{h_i(A)=h_i(B)\}}{n}$ of $J(A,B)$

We will bound the accuracy of this sampling approximation to the Jaccard similarity in terms of n.

Let p be the jaccard similarity between some documents A,B. In terms of p and n:

- $\mathbb{E}(H) =$ ________________
- $var(H) =$ ________________

What is the smallest upper bound on $var(H)$ in terms of n that holds for any value of p?
Hint: you've computed this result before by taking the derivative with respect to p, setting to 0, and solving. ________________

What is the smallest upper bound on $var(H)$ in terms of n that holds forp > 0.9 ? ________________

Assume n is large enough that H is approximately normal. If p > 0.9, what is an upper bound on $P(H < 0.8)$ in terms of the Q function and n? ________________

Let's say we define two documents A,B with J(A,B) > 0.9 as similar, and those with J(A,B) < 0.8 as dissimilar.

Let's say we use the following algorithm to guess whether documents are similar:

- Compute H using n independent has functions.
- If H > 0.9, we guess the documents are similar.
- If H < 0.8, we guess the documents are dissimilar.
- Else, we output "I don't know"

If $0.8 \leq p \leq 0.9$, the algorithm is judged correct, no matter what it outputs.

Hint: You should be able to quickly answer the following questions using the bounds on the variance calculated above:

- If $p > 0.9$ for two documents A,B, in terms of n, what is an upper bound on the probability our algorithm incorrectly says these documents are dissimilar? ____________________
- If $p < 0.8$, what is an upper bound on the probability our algorithm incorrectly says these documents are similar? ____________________

How many min-hash functions n are needed (as an expression, don't round) to ensure the probability of an error is < 0.03? You can use Qinv. ____________________