Your Name: ______________________

Section: ______________________

36-201 INTRODUCTION TO STATISTICAL REASONING
Computer Lab Exercise – Lab #9
Basic Probability Calculations

Objectives:

1. To review basic calculations and formulas for Probability.
2. To practice applying Venn Diagram and Joint Probability Table reasoning to probability problems.
3. To review the definitions and meaning of "random variable".
4. To work with "binomial distribution" calculations.

Getting Started

For these exercises you will need:

- A good pencil and eraser(s) [ink pens are not recommended!]
- A scientific or business calculator.

Part I: The Language of Probability Calculations

All probability calculations are based on a few simple ideas. These are discussed informally in Moore sections 7.1 and 7.2, and discussed more formally in Siegel and Morgan, Chapter 7. Here is a quick summary of the main ideas:

1. A *random outcome* is the result of any well-specified procedure, that is not completely determined before the procedure[1] is performed. The *sample space* S is the set of all possible outcomes of the procedure. An *event* A is a particular set of outcomes you are interested in. An event *occurs* if one of the outcomes in it occurs.

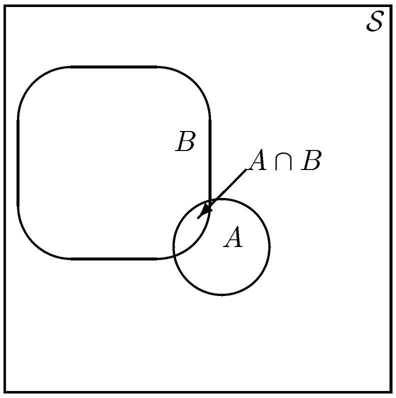
   Example: If you toss a six-sided die, the sample space can be written as the list $S = \{1, 2, 3, 4, 5, 6\}$ *of possible faces that can come up. One of the events in the sample space is the event* $A = \{\text{even face comes up}\}$ *which we can also write as* $A = \{2, 4, 6\}$. *If we perform this procedure and observe that face 4 comes up, we say* A *has occurred, since one of the outcomes in* A*, namely 4, occurred.*

2. The *probability* of an event A is a number between 0 and 1 specifying how likely that event is to occur. We write: $0 \leq P(A) \leq 1$. We also say $P(S) = P(\text{something happened}) = 1$ and $P(\emptyset) = P(\text{nothing happened}) = 0$.
3. $P(A)$ may be calculated as the sum of the probabilities of the outcomes.

   In the example above, if each face is equally likely to come up (probability 1/6), then the event A *has probability* $P(A) = 1/6 + 1/6 + 1/6 = 1/2$.

[1] Some books, like Siegel and Morgan, call the procedure that generates a random outcome a *random experiment*. But I will not use that term since we have already used *experiment* to mean something else in our course.

4. The *complement* of the event A is written A^c; it is the event consisting of all outcomes not in A (so A^c is the "opposite" of A). The probability of A^c is $P(A^c) = 1 - P(A)$.
5. The *union* of two events A and B is the set of all outcomes in A, or in B, or both, written $A \cup B$. The *intersection* of A and B is the set of outcomes in **both** A and B, written $A \cap B$. It is the overlapping part in the Venn Diagram below.
6. The *general "or" rule* is $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. This can easily be seen by equating probabilities with areas in a Venn diagram:



In the calculation $P(A) + P(B)$ the overlapping part of A and B is counted twice and so has to be subtracted out once.

The *disjoint "or" rule* says that if A and B are disjoint (mutually exclusive; no outcomes in $A \cap B$) then $P(A \cup B) = P(A) + P(B)$.

7. The *conditional probability of A given B* is a way of revising the probability of A once you have the new information that B occured. It is the fraction of probability in B that is accounted for by A. We write $P(A|B) = \dfrac{P(A \cap B)}{P(B)}$.
8. The *general "and" rule* says that $P(A \cap B) = P(A) \cdot P(B|A)$ [and also $P(A \cap B) = P(B) \cdot P(A|B)$]. This follows from the definition of conditional probability.
9. Two events A and B are said to be *independent* if either $P(A|B) = P(A)$ or $P(B|A) = P(B)$. In words, the information that B has occurred doesn't change the probability of A, and vice-versa.

The *independent "and" rule* says that if A and B are independent events, then $P(A \cap B) = P(A) \cdot P(B)$.

Example: Applying the Rules to Calculate Probabilities

We will work through an example, to see some of these tools in action.

> *A restaurant has collected data on its customers' orders and so has estimated empirical probabilities about what happens after the main course. It was found that 20% of the customers had dessert only, 40% had coffee only, and 30% had both dessert and coffee [use these percents as probabilities to answer the questions below].*

Question #1: One way to display probability information for a problem like this is with a *joint probability table*, which is like a contingency table with probabilities instead of counts. The joint probability table for this problem is given below. Fill in the missing cells.

		Had Coffee? No	Had Coffee? Yes	Total
Had Dessert?	No	$\boxed{}$	0.40	$\boxed{}$
	Yes	0.20	0.30	0.50
Total		$\boxed{}$	0.70	1.00

Note that the overall probability of having dessert (with or without coffee) is 0.50 (a row total); the overall probability of having coffee (with or without dessert) is 0.70 (a column total).

Question #2: Below is a generic Venn Diagram for problems with two simple events. Write a word or two describing each event in this problem, and fill in the probabilities below, using the joint probability table.

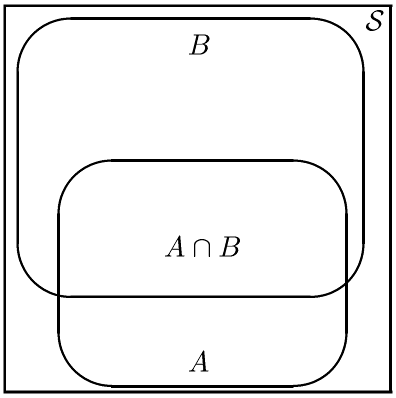
Event A, in words: Had Dessert.

Event B, in words: ______

$P(A) =$ ______

$P(B) =$ ______

$P(A \cap B) =$ ______



♣ **Question #3: In the Venn diagram in Question 2, shade in the area corresponding to "customer had dessert but not coffee". Circle the probability of this event in the joint probability table in Question 1.**

Question #4: Of the customers who order coffee after meal, what percentage also order dessert?

Using the conditional probability formula on p. 2 of this handout:

$$P[A|B] = \frac{P[A \cap B]}{P[B]} = \frac{\boxed{}}{\boxed{}} = \underline{}$$

As a column percent in Question 1:

$$\frac{\boxed{}}{\boxed{}} = \underline{}$$

Question #5: Of the customers who order dessert, what percentage also order coffee?

Using the conditional probability formula on p. 2 of this handout:

$$P[B|A] = \frac{P[A \cap B]}{P[A]} = \frac{\boxed{}}{\boxed{}} = \underline{}$$

As a row percent in Question 1:

$$\frac{\boxed{}}{\boxed{}} = \underline{}$$

♣ **Question #6: Did you get the same answers in Question 4 using either method? How about Question 5? Explain.**

Part II: Random Outcomes and Random Variables

Many random outcomes can be described qualitatively or quantitatively. The language of *random variables* is just another way of describing outcomes quantitatively (numerically), that turns out to be more efficient for math calculations.

> **Example:** *Sally, Bill, George and Bettina make up a small engineering team. Sally makes $75,000 a year, Bill makes $35,000, George makes $100,000 and Bettina makes $75,000.*
>
> *For the procedure "select a person at random from the team", the sample space is* $S = \{\text{Sally, Bill, George, Bettina}\}$ *and if the outcomes are equally likely then each has probability 1/4 of occurring.*
>
> *For the procedure "select a person at random from the team and check their salary" the sample space is* $S = \{35000, 75000, 100000\}$ *(only three outcomes since one salary is repeated). In the language of random variables we would say:*
>
> Let X be a random variable with values 35000, 75000, and 100000. Note that these values are *no longer* equally likely. Rather,
>
> $P(X = 35000) = 1/4$
>
> $P(X = 75000) = 1/2$ (since two people have this salary)
>
> $P(X = 100000) = 1/4$

For each x in the sample space (e.g. $x = 35000$, $x = 75000$, etc.) the probabilities $P(X = x)$ listed above completely describe the possible outcomes of the procedure. Taken together, these probabilities are called the *distribution* of X.

Random variables give us a way of summarizing and predicting the behavior of future replications of our procedure.

In the example above, suppose you repeat 100 times the procedure of selecting a person from the team at random and checking their salary. What would the sample average of your 100 observations be?

Since you expect to see each person about 25 times (1/4 of the 100 trials), you'd expect your sample average to be about

$$\begin{aligned} \overline{X} &\approx \frac{(25)(35000) + (50)(75000) + (25)(100000)}{100} \\ &= (1/4)(35000) + (1/2)(75000) + (1/4)(100000) \\ &= \$71,250. \end{aligned}$$

Note that \$71,250 is the sum of each salary times its probability. This forms the basis of the definition of the *expected value* of a random variable.

> The *expected value* or *mean* of the random variable X with values x is
>
> $$\mu = E(X) = \text{Sum of values times probabilities} = \sum x \cdot P(X = x)$$

The greek letter μ (read: "myew") is used when we want a letter for the mean, but usually we will just talk about it and not worry about the greek.

This is just the value that you expect $\overline{X}$ will be near in data that you gather from repeated independent trials of the same procedure. In the salary example above, we calculated $E(X) = \$71,250$. The expected value is also called the *mean* of the distribution.

You can use the same idea to predict the sample standard deviation in repeated independent trials:

> The *standard deviation* (SD) of the random variable X with values x is
>
> $$\begin{aligned} \sigma &= \sqrt{\text{Sum of squared deviations times probabilities}} \\ &= \sqrt{\sum (x - \mu)^2 P(X = x)} \end{aligned}$$

The greek letter σ (read "sigma") is used when we want a single letter for SD. The *variance* is just the value before taking the square root.

In the salary example above, the variance of the salary random variable X is

$$\begin{aligned} (1/4)(35000 - 71250)^2 \;&+\; (1/2)(75000 - 71250)^2 \;+\; (1/4)(100000 - 71250)^2 \\ &= \quad 542,187,500 \end{aligned}$$

and the SD of X is

$$\sigma = \sqrt{542,187,500} = \$23,285.$$

This is a very large SD, but of course the range of salaries from \$35,000 to \$100,000 is large too.

Example: Working with Random Variables

Imagine a lottery in which you pay \$1.00 for a ticket. You then scratch a silver film off the ticket to see if you win a prize. One ticket in ten pays \$2.00 (net gain $\$1.00$*), another pays \$5.00 (net gain \$4.00). The other eight tickets pay nothing (net loss: the original \$1.00 you paid). If all tickets are equally likely, then each has probability 1/10 of occuring, so we arrive at the table of payoffs listed in Table 1.*

You Pay($-$)	You Win($+$)	Net Gain x	Probability $P(X = x)$
\$1.00	\$2.00	\$1.00	0.10
\$1.00	\$5.00	\$4.00	0.10
\$1.00	\$0.00	$-$\$1.00	0.80

Table 1: Net Gains for a lottery.

♣ **Question #7: Find the expected value (mean) of your net gain in this lottery.**

Question #8: Describe briefly what this expected value represents.

Question #9: Find the standard deviation (SD) of your net lottery gain.

♣ **Question #10: What is the probability of getting a ticket with a net gain of at least $3.00?**

Question #11: What is the probability that a ticket with a net gain of less than $3.00?

Example: The Binomial Distribution [If Time Permits]

A random variable that we encounter repeatedly in opinion surveys, marketing surveys, quality and performances studies, etc., is the *binomial* random variable.

> If a procedure for generating random outcomes is analogous to counting the number X of heads in n independent flips of a loaded coin, then
>
> - The *probability of Heads* on each flip is denoted p.
> - X (the number of heads) is called a *binomial random variable*.
> - $\hat{p} = X/n$ (the fraction of heads) is called a *binomial proportion*.
>
> The formulas for the distribution of a binomial random variable and its mean μ and standard deviation σ are as follows:
>
> $$P[X = k] = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{(for } k = 0, 1, 2, \ldots n\text{)}$$
>
> $$\mu_X = np \qquad \sigma_X = \sqrt{np(1-p)}$$
>
> $$\mu_{\hat{p}} = p \qquad \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

The formula

$$\binom{n}{k} p^k (1-p)^{n-k}$$

has two distinct parts.

- The symbol "$\binom{n}{k}$" stands for "n choose k", the binomial coefficient from high-school algebra:

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

This says how many ways there are to get k heads in n coin tosses.

The symbol "$n!$" is read "n factorial", which is how many ways to arrange n objects. In order to make all the formulas work out nicely we also define $0! = 1$, so for example $\binom{n}{0} = n!/(n!0!) = n!/n! = 1$.

Question #12: Fill out the following table of factorials (see next page):

$n!$	**Formula**	**Value**	**Meaning**
0!	(special case)	1	Number of ways to arrange 0 objects
1!	1	1	Number of ways to arrange 1 object
2!	$2 \cdot 1$	2	Number of ways to arrange 2 objects
3!	$3 \cdot 2 \cdot 1$	6	Number of ways to arrange 3 objects
4!	$4 \cdot 3 \cdot 2 \cdot 1$	24	Number of ways to arrange 4 objects
5!			Number of ways to arrange __ objects
6!			Number of ways to arrange __ objects
7!			Number of ways to arrange __ objects
8!			Number of ways to arrange __ objects

(Notice that, in your table, $n!$ is just $n \cdot (n-1)!$. For example, verify that $8! = 8 \cdot 7!$. This sometimes speeds hand calculations with factorials.)

Question #13: Use your table to compute the following binomial coefficients:

$\binom{4}{1} = \frac{4!}{1!3!} =$ ________. This is the number of ways to get exactly 1 heads in 4 tosses.

$\binom{8}{6} = \frac{\boxed{}}{\boxed{}\,\boxed{}} =$ ________. This is the number of ways to get exactly 6 heads in 8 tosses.

• The formula $p^k(1-p)^{n-k}$ is the probability of seeing exactly k heads in n coin tosses. For example, each of the ways of getting 6 heads in 8 tosses has probability $p \cdot p \cdot p \cdot p \cdot p \cdot p \cdot (1-p) \cdot (1-p) = p^6(1-p)^2$

Here is an example that puts these two pieces together:

> *You are planning to make sales calls at eight firms today. As a rough approximation, you figure that each call has a 15% chance of resulting in a sale and that firms make their buying decisions without consulting each other.*

♣ **Question #14: Find the probability of having exactly 6 sales today. [Hint: use the formula $\binom{n}{k} p^k(1-p)^{n-k}$ with $n = 8$, $k = 6$, and $p = 0.15$.]**

Question #15: Find the probability of having a really terrible day with no sales at all. [Hint: Use the formula again; remember that anything raised to the zero power is 1].

Question #16: What is the mean μ_X and SD σ_X for the number of sales you will make?