Activity 1: German Tanks (continued)

Recall that we are trying to estimate N based on a sample of size n=5 from the integers 1, 2,… N.

(a) Give the formulae for some possible estimators of the total population size, $\widehat{N}$.

(b) Maximum likelihood estimation is a method for determining an estimator (number from the sample) for a parameter (number from the population). The key is to maximize the likelihood for the parameter N, **using the data you are given**. The likelihood here is:

$$\boldsymbol{L(x_1, x_2, \ldots x_n \mid N) = \left(\frac{1}{N}\right)^n \quad 1 \leq x_i \leq N, \quad i = 1, 2, \ldots n}$$

$$\boldsymbol{= 0 \qquad \text{otherwise}}$$

What is the maximum likelihood estimator for N? (Hint: draw a picture of the likelihood as a function of N.)

(c) What makes a good estimator? That is, what qualities of the estimator (that is, the FUNCTION of your data) are you looking for?

Activity 2: AIDS Testing (from Investigating Statistical Concepts, Applications, and Methods; Chance & Rossman)

The ELISA test for AIDS is used in the screening of blood donations. As with most medical diagnostic tests, the ELISA test is not infallible. If a person actually carries the AIDS virus, experts estimate that this test gives a positive result 97.7% of the time. (This number is called the *sensitivity* of the test.) If a person does not carry the AIDS virus, ELISA gives a negative result 92.6% of the time (the *specificity* of the test). Recent estimates are that 0.5% of the American public carries the AIDS virus (the *base rate* with the disease).

(a) Suppose that someone tells you that they have tested positive. Given this information, how likely do you think it is that the person actually carries the AIDS virus?

Imagine a hypothetical population of 1,000,000 people for whom these percentages hold exactly. You will fill in a two-way table as you derive *Bayes' Theorem* to address the question above.

	Positive test	Negative test	Total
Carries AIDS virus	(c)	(c)	(b)
Does not carry AIDS	(d)	(d)	(b)
Total	(e)	(e)	1,000,000

(b) Assuming that 0.5% of the population of 1,000,000 people carries AIDS, how many such carriers are there in the population? How many non-carriers are there? (Record these in the table.)

(c) Consider for now just the carriers. If 97.7% of them test positive, how many test positive? How many carriers does that leave who test negative? (Record these in the table.)

(d) Now consider only the non-carriers. If 92.6% of them test negative, how many test negative? How many non-carriers does that leave who test positive? (Record these in the table.)

(e) Determine the total number of positive test results and the total number of negative test results. (Record these in the table.)

(f) Of those who test positive, what proportion actually carry the disease? How does this compare to your prediction in a)? Explain why this probability is smaller than most people expect.

(g) Of those who test negative, what proportion are actually free of the disease?

Activity 3: After-School Activities (Conditional Probability) (thanks to Roxy Peck)

At the fictional George Washington High School (GWHS), after-school activities can be classified into three types: athletics, fine arts, and other. The following table gives the number of students participating in each of these types of activities by grade. For the purposes of this example, we will assume that every student at the school is in exactly one of these after-school activities.

	9th	**10th**	**11th**	**12th**	**Total**
Athletics	150	160	140	150	**600**
Fine Arts	100	90	120	125	**435**
Other	125	140	150	150	**565**
Total	**375**	**390**	**410**	**425**	**1600**

(a) The principal at GWHS selects students at random and invites them to have lunch with her to discuss various issues that might be of concern to them. She feels that random selection will give her the greatest chance of hearing from a cross-section of the student body. What is the probability that a randomly selected student is a senior athlete?

(b) Now suppose that the principal's secretary records not only the student's name but also the student's grade level. The secretary has indicated that the selected student is a senior. Does this information change our assessment of the likelihood that the selected student is an athlete?

Notation: note that conditional probability is simply the act of reducing the population of interest to a subgroup. That is:

$$P(A|B) = \frac{P(A \,\&\, B)}{P(B)}$$

$$P(Ath\,|12\,) = \frac{P(Ath \,\&\, 12)}{P(12)} = \frac{150/1600}{425/1600} = \frac{150}{425}$$

And really, it doesn't matter which is A and which is B, so

$$P(B|A) = \frac{P(A \,\&\, B)}{P(A)}$$

Note, because of the law of total probability, we can write:

$$P(12) = P(12 \,\&\, Ath) + P(12 \,\&\, FA) + P(12 \,\&\, O)$$

(c) Using **only** the following values, find the probability that the randomly selected student will participate in the fine arts, given we know she is a senior:

P(12 | Ath) = 0.25
P(12 | FA) = 0.2874
P(12 | Other) = 0.2655
P(Ath) = 0.375
P(FA) = 0.2719
P(Other) = 0.3531 (note: I didn't actually need to tell you this value, why not?)

Find: P(FA | 12) = ???

Activity 4: Let's Make a Deal

A popular 1960's game show, Let's Make a Deal, had a particular game where there was a prize behind one of three doors. You chose a door, and the host would open one of the remaining doors (always possible regardless of where the prize was). The host then gave you the option of switching doors. A hotly contested probability question is: do your chances of winning the prize improve if you switch doors?

Define the following:
C_i : the car is behind Door i, for $i \in \{1, 2, 3\}$
H_{ij} : the host opens Door j after the player has picked Door i, for $i, j \in \{1, 2, 3\}$

For example, C_2 is the situation that the car is behind Door 2, and H_{23} denotes the situation that the host opened Door 3 after I chose Door 2. Note that H_{ij} is the information we have (what we'll condition on). Let's say you start off by picking Door 2, and the host opens Door 3.

Find the probability of winning given you switch. [Hint: try writing out the probability you want as a particular scenario… e.g., you pick Door 2 and the host opens Door 3.]

Activity 5: Bayes and Baseball

We can use ideas from Bayes' Theorem to improve estimation and inference (i.e., inferring results from your data to a larger population).

The Setting

You are a statistician employed by On The Ball Consulting. Veteran major-league baseball scout Rocky Chew seeks your advice regarding estimating the probability that amateur baseball player John Spurrier will get a base hit against a major-league pitcher. Rocky has arranged for Spurrier to have ten at bats against a major-league pitcher.

The Background

The traditional batting average is a frequentist estimator in that it makes use of the observed data, but ignores any prior information that might exist. (Some of you baseball enthusiasts will be a bit uncomfortable that we're going to assume that our denominator is # of times up to bat.) If we assume that the at bats are independent Bernoulli trials with a constant probability of getting a base hit, then

X ~ Bin(n = number at bat; p = P(getting a base hit))

$\hat{p}_f = \frac{X}{n}$, is the maximum likelihood estimator, the method of moments estimator, and the minimum variance unbiased estimator of the unknown probability (of getting a base hit.) That makes it a good estimator, but it ignores information we might have about baseball. You have the following prior information:

- John Spurrier appears to be a good but not great player. He is one of the better batters on a somewhat above-average American Legion (high school) baseball team.
- The few major-league scouts who have watched him play do not believe that Spurrier's batting ability is at the professional level.
- A barely adequate major-league hitter has a batting average of about 0.200.
- A very good major-league batter has a batting average of about 0.300.
- Ty Cobb has the all-time best major-league batting average of 0.366.

We're going to use a Beta prior to incorporate our previous knowledge. What should that prior look like?

The Bayesian estimator is:

$$\hat{p}_B = \frac{X + \alpha}{n + \alpha + \beta}$$

The Experiment

1. John Spurrier will have n=10 at bats. The random variable, X, will be the number of base hits that he gets.
2. Determining the prior probability: As a class we will find α and β that are consistent with our prior information.
3. We use Mean Squared Error (MSE) to compare estimators:

MSE($\hat{p}$) = E[$(\hat{p} - p)^2$] = Var($\hat{p}$) + bias^2($\hat{p}$) = Var($\hat{p}$) + [E($\hat{p}$) - p]^2