Coverage of test 1

1. Sampling issues — definition of population and sample, purposes of sampling, types of bias in sampling
2. Data summary and presentation — frequency distribution, histogram, stem–and–leaf display, boxplot; sample mean, median, mode, midrange, range, interquartile range, midhinge, median absolute deviation, variance, standard deviation; appropriateness of statistics for different types of data
3. Probability — types of probability, simple probability calculations, basic properties of probabilities, conditional probability, independence
4. Random variables — definition, probability distribution, mean (expected value), variance, standard deviation
5. Specific random variables
   (a) Binomial — distribution, mean, variance
   (b) Normal — calculating probabilities
   (c) Normal approximation to Binomial
6. Central Limit Theorem

Note: You might find it useful to study for the test by “taking” one of the tests from previous years under actual test conditions.
Note: You should not view the list above as exhaustive. Everything that we have discussed in class, or has been in a handout, is fair game to appear on the test.

Practice problems

1. One of the most serious problems facing research libraries is the preservation of the materials that comprise their collections. This is a particularly critical problem in large libraries, where the age and size of the collections make evaluation and corrective action difficult.

   When faced with a book to be examined, a preservation librarian has four preservation strategies to consider: no repair, restoration, microfilming, and full repair (the latter three being different kinds of repair strategies). These strategy categories are ordered from least to most difficult technically. The following table gives the cost in dollars associated with each strategy per book:

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Cost per book</th>
</tr>
</thead>
<tbody>
<tr>
<td>No repair</td>
<td>0</td>
</tr>
<tr>
<td>Restoration</td>
<td>100</td>
</tr>
<tr>
<td>Microfilming</td>
<td>200</td>
</tr>
<tr>
<td>Full repair</td>
<td>400</td>
</tr>
</tbody>
</table>

   The preservation librarian at a large university library system must decide which library building will undergo a preservation program in the upcoming academic year. Preliminary analysis has yielded the following probabilities of the need for each type of strategy for a randomly selected book in each library:

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Main stacks</th>
<th>Business school library</th>
</tr>
</thead>
<tbody>
<tr>
<td>No repair</td>
<td>.6</td>
<td>.7</td>
</tr>
<tr>
<td>Restoration</td>
<td>.2</td>
<td>0</td>
</tr>
<tr>
<td>Microfilming</td>
<td>.1</td>
<td>.1</td>
</tr>
<tr>
<td>Full repair</td>
<td>.1</td>
<td>.2</td>
</tr>
</tbody>
</table>

   (a) Her budget advisor suggests choosing the library building with smaller mean preservation cost per book examined. Based on this recommendation, which library should she choose?
   (b) The assistant preservation librarian suggests choosing the library with smaller probability of having to do any kind of repair. Based on this recommendation, which library should she choose?
   (c) It occurs to her that she needn’t focus on only one library; rather, she can choose a mixture of books from each library. She has a budget of $95 per book examined to spend on preservation. Let $p$ be the proportion of the total books examined that come from the main stacks (so $1 - p$ is the
proportion that come from the Business school library). What should \( p \) be so that the expected cost per book meets her spending budget exactly?

2. The story “Tough Times in Japan,” which appeared in the February 9, 2003 issue of Parade magazine, contained various statistics purporting to illustrate tough economic times in that country. The article included the following sentence: “A recent survey by Japan’s Ministry of Health, Labor and Welfare indicates that as many as 60% of all households now have annual incomes below the national average of $52,339.” Explain why it would not be surprising to find that more than half of all households have annual incomes below the national average, even if a country was not having tough economic times.

3. In December 1993, the U.S. Census Bureau issued a report on college enrollments in the United States. The report stated that 56% of all college students in 1992 were women. Further, 17% of all college students were aged 35 or older. Overall, 11% of all college students were women aged 35 or older.
   (a) You meet a person on campus who is a member of an organization composed of people who are all under the age of 35. What is the probability that this person is a woman?
   (b) Are the events “being a woman” and “being aged 35 years or older” independent?

4. Consider two investments that you are looking at for your portfolio. The one-month return for each is normally distributed, with a mean \( \mu = 0.01 \) (that is, a 1% return). Investment A’s return has standard deviation \( \sigma_A = 0.005 \), while investment B’s return has standard deviation \( \sigma_B = 0.015 \). What is the probability that investment A’s return next month will be positive? Is investment B’s probability of positive return higher, lower, or the same? Which investment would a “rational” investor prefer?

5. During World War II many economists, mathematicians, and statisticians were members of Columbia University’s Statistics Research Group, which did high-level consulting for the armed services. As part of the group’s work, statistician Abraham Wald was approached to provide guidance on where to place armor on airplanes in order to protect them (the armor was heavy, so it couldn’t be placed over the entire aircraft). The aircraft engineers had taken a large random sample of aircraft that had returned from military action, and had developed a mapping of where (and how often) bullet holes were found. If you were Wald, what would your advice be about where to put the armor? Why?

6. Say the SAT verbal exam is graded along a strict curve, such that the scores in the population of students have mean 500 and standard deviation 100. If a random sample of 20 high school seniors is taken, what is the probability that the mean of their test scores exceeds 530? What assumptions are you making here?

7. Alcohol–related problems on college campuses are widespread, even though most college students are under the national drinking age of 21. A study funded by the Harvard School of Public Health, released on November 4, 1995, examined some issues relating to college alcohol abuse on campuses in the northeastern United States. The study stated that 50% of college men are binge drinkers, where a binge–drinking man is defined as one who consumed five drinks one after the other at least once in the previous two weeks.
   (a) Consider 10 randomly selected northeastern U.S. college men. What is the exact probability that at least 9 of these men are binge drinkers?
   (b) The applicants for a position include 100 northeastern U.S. college men. What is the probability that at least 35 of them are binge drinkers? An approximate answer is good enough here.
Answers to practice problems

1. Let $C$ be the preservation cost of a book.
   (a) The mean preservation cost per book examined equals
   \[
   E(C) = \sum_{\text{strategies}} (\text{Cost of strategy}) P(\text{strategy}).
   \]
   Thus, for the main stacks,
   \[
   E(C) = (0)(.6) + (100)(.2) + (200)(.1) + (400)(.1) = 80,
   \]
   while for the Business school library,
   \[
   E(C) = (0)(.7) + (100)(0) + (200)(.1) + (400)(.2) = 100
   \]
   Thus, the main stacks building has a lower mean preservation cost per book examined.
   (b) Since $P(\text{any repair}) = 1 - P(\text{no repair})$, the values are $1 - .6 = .4$ for the main stacks and $1 - .7 = .3$ for the Business school library. Thus, the Business school library has lower probability of making any repair.
   (c) By (a), the amount spent per book examined would be
   \[
   (80)(p) + (100)(1 - p).
   \]
   So, to match this to the spending budget, solve for $p$:
   \[
   80p + 100 - 100p = 95 \\
   \Rightarrow 20p = 5 \\
   \Rightarrow p = .25
   \]
   So, to meet the budget, choose 25% of the books from the main stacks, and 75% from the Business school library.

   For many years, very little was known about the actual condition of books in the world’s research libraries, although a good deal of folklore was bandied about. The first large–scale, statistically valid, survey of book condition was undertaken at Yale University in the late 1970’s and early 1980’s. For a discussion of the results of this survey, and the issues raised during the study, see the 1985 paper “The Yale Survey: A Large–Scale Study of Book Deterioration in the Yale University Library,” by Gay Walker, Jane Greenfield, John Fox and Jeffrey S. Simonoff, in volume 46 of the journal College and Research Libraries, pages 111–132.

2. Incomes are typically long right-tailed; for such data, the median is usually less than the mean, so more than half of the data will almost invariably fall below the mean.

3. Let $W$ represent a randomly selected student being a woman, and $T$ represent being aged 35 or over. We are given that $P(W) = .56$, $P(T) = .17$, and $P(W \text{ and } T) = .11$. We can use these to construct a “hypothetical 100,000” table to lay out all of the relevant probabilities:

\[
\begin{array}{ccc}
\text{Age 35 and older (T)} & \text{Under age 35 (T')} \\
\hline
\text{Woman (W)} & 11000 & 45000 & 56000 \\
\text{Man (W')} & 6000 & 38000 & 44000 \\
& 17000 & 83000 & 100,000 \\
\end{array}
\]
(a) 
\[ P(W|T') = \frac{P(W \text{ and } T')}{P(T')} = \frac{45000}{83000} = .542 \]

(b) No, the two events are not independent, since \( P(W|T') = .542 \neq P(W) = .56 \); equivalently, \( P(W \text{ and } T) = .11 \neq P(W)P(T) = .095 \). Being under the age of 35 decreases the probability of a randomly selected student being a woman.

4. Let \( A \) be the return of investment \( A \). We’re told that \( A \sim N(.01, .005^2) \), so
\[
P(A > 0) = P\left( Z > \frac{0 - .01}{.005} \right) = P(Z > -2) = 1 - P(Z \leq -2) = 1 - .0228 = .9772.
\]
The corresponding probability for investment \( B \) will be lower, since the probability is more spread out in both directions. You could have calculated it (it’s .7475), but that’s not necessary; all you need to see it is a picture like this:

![Graph showing probability densities of investments A and B]

Thus, a “rational” investor would prefer investment \( A \). Since it has the same expected return as investment \( B \), with less risk (smaller standard deviation), this is just what we would expect. Note also that this simple example demonstrates why modeling the volatility of investments well is a worthy goal: if we were smarter than other investors regarding volatility, we could construct options with positive expected return by going short or long as appropriate.

5. It’s tempting to say that the armor should be put in the places with the most bullet holes, but Wald was smart enough to realize that that is not correct — he suggested that they put the armor in the places with the fewest bullet holes. Why? Because if we assume that bullet holes are roughly evenly
distributed over the airplane (a not unreasonable assumption), then a lack of bullet holes means that airplanes that were hit in those places never made it back from the military action. That is, this is a biased sample, with airplanes that are more seriously damaged less likely to be in the sample.

6. By the Central Limit Theorem, \( \bar{X} \sim N(\mu, \sigma^2/n) \). So,

\[
P(\bar{X} > 530) = P \left( Z > \frac{530 - 500}{100/\sqrt{20}} \right) \\
= P(Z > 1.34) \\
= .0901
\]

We are assuming that the Central Limit Theorem is valid here. Since scores are probably reasonably symmetric, a sample of size 20 is probably large enough to make this a reasonable assumption.

7. The number of binge drinkers \( B \) is distributed as a Binomial random variable with \( n = 100 \) and \( p = .5 \).

(a) We want to calculate an exact Binomial probability, so we use the Binomial probability function,

\[
P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k},
\]

where \( n = 10 \) and \( p = .5 \). In this case, we want

\[
P(B \geq 9) = P(B = 9) + P(B = 10) = \sum_{k=9}^{10} \binom{10}{k} (.5)^k (.5)^{10-k} = .009766 + .000977 = .01074.
\]

(b) Now \( B \sim Bin(100, .5) \). We use a normal approximation to the binomial to estimate this probability. This requires calculating \( \mu = np = 50 \) and \( \sigma^2 = np(1 - p) = 25 \). Then, using the continuity correction, we have

\[
P(B \geq 35) \approx P \left( Z > \frac{34.5 - 50}{\sqrt{25}} \right) \\
= P(Z > -3.1) \\
= 1 - P(Z < -3.1) \\
= 1 - .0009 \\
= .9991.
\]