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, 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 is in the course supplement, is fair game to appear on the test.

Practice problems

(1) As you no doubt know, digital music players include a shuffle feature, whereby the device takes a list of songs, called a playlist, and rearranges them in random order. Since the introduction of the iPod music player some users have questioned the randomness of the shuffle on these devices. Newsweek published an article on this issue in 2005 (“Does Your iPod Play Favorites?” by Steven Levy). The August 2009 issue of The American Statistician contained an article describing several experiments designed to address this question (“Does Your iPod Really Play Favorites?” by Amy Froelich, William Duckworth, and Jessica Culhane).

One of the experiments described in the paper was based on the first author’s iPod. In this experiment, many shuffles of a playlist were undertaken, and the number of times a song by Faith Hill was played among the first 60 songs was recorded. Based on this, consider the following probability distribution for the number of times a Faith Hill song will be played among the first 60 songs:
Based on this probability distribution, what is the expected number of times a Faith Hill song will be played? What is the standard deviation of the number of times a Faith Hill song will be played?

(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) The book The Job-generation Controversy: the Economic Myth of Small Business, written by D. Hirschberg, argued that it is a myth that small firms play a major role in contributing to the growth of employment in the U.S. economy. Consider U.S. firms that were active in both 1989 and 1991. According to figures given in the book, 4.4% of all such firms were large firms in 1989, employing more than 100 people. Of firms that were small (employing no more than 100 people) in 1989, 99.7% were small in 1991; of firms that were large in 1989, 85.7% were large in 1991. Consider a particular firm that was a large firm in 1991. What is the probability that the firm was large in 1989?

(4) The relative poverty line is defined as the income value corresponding to 60% of the median family income over all families in a given area. Say you are told that family incomes (that is, the income of each individual family in a particular area) are normally distributed, with standard deviation being one-half the mean (that is, incomes follow a $N(\mu, [\mu/2]^2)$ distribution, for some $\mu$). Given this assumption, what is the probability that a randomly selected family is below the relative poverty line?

(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) A bank manager who is responsible for stocking the ATM machines at a branch has been keeping track of the withdrawal habits of customers at the ATM machines. She knows from long experience that the amount withdrawn per customer has a mean of $60 with a standard deviation of $15. She also knows that the number of customers who use the ATM machines in a day is approximately 1000. Assume that exactly 1000 customers will use the ATM machines in a day. If the manager puts $61000 in the ATM, what is the probability that the machines run out of money during the day? What assumptions are you making here? Do they seem reasonable?

(7) A study entitled “A Workplace Divided: How Americans View Discrimination and Race on the Job” was released several years ago by Rutgers University and the University of Connecticut.

(a) According to the survey, 8% of white workers in large workplaces (250 or more employees) reported having no African-American coworkers. Consider five white workers randomly chosen from a set of large workplaces (one from each workplace). What is the exact probability that exactly three of them say that they have no African-American coworkers?

(b) According to the same survey, 64% of white workers in small workplaces (25 or fewer employees) reported having no African-American coworkers. Consider 500 white workers randomly chosen from a set of small workplaces (one from each workplace). What is the probability that at least 300 of them say that they have no African-American coworkers? An approximate answer is good enough here.
Answers to practice problems

1. Let $F$ be the number of Faith Hill songs played among the first 60 songs. We thus have that

$$E(F) = (0)(.420) + (1)(.425) + (2)(.140) + (3)(.015) = .75$$

and

$$V(F) = (0 - .75)^2(.420) + (1 - .75)^2(.425) + (2 - .75)^2(.140) + (3 - .75)^2(.015) = .5575,$$

implying that $SD(F) = \sqrt{.5575} = .747$.

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 $S89$ be the event that a firm that was active in both 1989 and 1991 was small in 1989, and $S91$ be the event that it was small in 1991. Also, let $L89$ be the event that the firm was large in 1989, and let $L91$ be the event that it was large in 1991. We are told that $P(L89) = .044$, $P(S91|S89) = .997$, and $P(L91|L89) = .857$. We want $P(L89|L91)$. This can be calculated using a “hypothetical 100,000” table (numbers determined by subtraction or addition are given in bold face):

<table>
<thead>
<tr>
<th></th>
<th>$S91$</th>
<th>$L91$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S89$</td>
<td>(.997)(95600)</td>
<td>287</td>
</tr>
<tr>
<td>$L89$</td>
<td>629</td>
<td>(857)(4400) = 3771</td>
</tr>
</tbody>
</table>

Now we want $P(L89|L91)$, which is obtained from the table as

$$P(L89|L91) = \frac{P(L89 \text{ and } L91)}{P(L91)}$$

$$= \frac{3771}{4058}$$

$$= .929.$$ 

Thus, if a firm was large in 1991, there is a 92.9% chance that it was large in 1989.

4. Let $I$ be the family income of a particular family. We’re told that $I \sim N(\mu, [\mu/2]^2)$. The relative poverty line is defined to be $.6\mu$ (since the distribution of $I$ is symmetric, the median equals the mean), so we want

$$P(I < .6\mu) = P \left( Z < \frac{.6\mu - \mu}{\mu/2} \right)$$

$$= P(Z < -.8)$$

$$= .2119$$

That is, we expect roughly 21.2% of the population to be below the relative poverty line in this circumstance.
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. Let $A$ be the amount of money withdrawn in a single transaction; we are given that $E(A) = 60$ and $SD(A) = 15$. If $T$ is the total amount of money withdrawn from the ATM in a day, $A_i$ is the amount withdrawn by the $i$th customer, and $n$ is the number of customers withdrawing money, $T = \sum_i A_i$, or equivalently $T = n\bar{A}$. By the Central Limit Theorem, we know that (at least approximately) $\bar{A} \sim N(60, 15^2/n)$. We’re told that $n = 1000$, and want $P(T > 61000)$. From the discussion above,

\[
P(T > 61000) = P(\sum_i A_i > 61000) = P(\bar{A} > 61)
\]

\[
= P\left(Z > \frac{61 - 60}{15/\sqrt{1000}}\right)
\]

\[
= P(Z > 2.11) = 1 - P(Z < 2.11) = 1 - .9826 = .0174.
\]

We’re assuming that the Central Limit Theorem conditions hold. We might wonder about whether the sample is large enough, since the money withdrawn is no doubt right-tailed (people can’t withdraw less than 0, but can withdraw hundreds of dollars at a time). I would be very surprised if a sample of size 1000 wasn’t big enough for the CLT to hold. We’re also assuming that the withdrawals are independent of each other, which seems reasonable.

7. (a) Let $A$ be the number of white workers who say that they have no African–American coworkers. We have that $A \sim Bin(5,.08)$, and we want $P(A = 3)$. The formula for binomial probabilities is

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

with $n = 5$, $p = .08$, and $k = 3$ in this case. Thus, we obtain

\[
P(A = 3) = \binom{5}{3}.08^3.92^2
\]

\[
= (10)(.000512)(.8464)
\]

\[
= .0043
\]

(b) Now $A \sim Bin(500,.64)$, and we want $P(A \geq 300)$. We have that $E(A) = (500)(.64) = 320$ and $V(A) = (500)(.64)(.36) = 115.2$, and we use the normal
approximation to the binomial, incorporating the continuity correction.

\[ P(A \geq 300) = P(A > 299.5) \]
\[ \approx P \left( Z > \frac{299.5 - 320}{\sqrt{115.2}} \right) \]
\[ = P(Z > -1.91) \]
\[ = 1 - P(Z < -1.91) \]
\[ = 1 - .0281 = .9719 \]

In fact, if you use Minitab to calculate the exact binomial probability, it turns out to be .9713.