Midterm and Solutions

1) Cloudseeding is an activity in which small crystals of ammonium nitrate (or other particulates that encourage ice formation) are dropped by airplane into clouds with the intent of increasing the amount of rain. Measuring the success of cloudseeding has been a difficult problem. Sometimes, the observed effect of seeding has been to increase rainfall, sometimes to decrease, and sometimes no effect was observed. A difficulty is that once the clouds have been seeded, there's no way of knowing what would have happened if they had not been seeded.

For most of this midterm, we'll study one particular cloud-seeding experiment that took place over a 3000 square-mil area in Florida, near Coral Gables. During the summer of 1975, each day was judged for its "suitability" for seeding. If a day was indeed suitable, then a coin was flipped to determine whether or not seeding was to be performed. Of the 24 days, half were seed days. The pilots did not know the outcome of the coin toss. If there was no seeding, the pilots still flew and released what they thought was cloud-seeding particulate, but was instead a "placebo".

We're going to consider only two of the variables recorded. "Prewetness" is the amount of rain that fell in the hour before the 'seeding' took place. The rainfall is measured in 10^7 cubic meters. The other variable is "rainfall", which measures rainfall, again in 10^7 cubic meters, after the "seeding" took place.

a. Identify the experimental units, the treatment groups, and the response variable.

Experimental Units: these are the things that are assigned to one treatment or the other. In this case, these are days. More precisely, they are "suitable" days, where suitability is determined by an agreed-upon definition. Non-suitable days are ignored and not assigned to treatment or control groups.

Treatment Groups: there were two. One group of days got ammonium nitrate, the other the placebo.

Response Variable: amount of rainfall after seeding. Other answers that combined rainfall with prewetness were accepted if they were meaningful and if they would in fact help us understand if cloudseeding had an effect.

b. According to the definition given in lecture, is this a controlled study or an observational study? Explain.

The distinction between the types of studies has to do with how the experimental units are assigned to the treatment groups. Here, the researchers decide (by coin flip) whether the suitable days are assigned to seeding or placebo and therefore this is a controlled study.

c. Why did the experimenters decide that the pilots couldn't know whether or not they were seeding?

Had pilots known whether or not they were actually seeding, they might have changed their behavior, and this would have induced a source of bias into the study. One can imagine a variety of different scenarios. For example, if pilots believed strongly in the effectiveness of the ammonium nitrate, perhaps they would believe that they didn't need to try very hard for the seeding to be successful, and would not seed as thoroughly. This could bias results against the cloudseeding.

Note that this is not an example of a confounder. A confounder must affect both assignment to treatment group, and the response variable. Assignment has already been made in this case, and so this cannot be a confounder.

2.
a) The goal of the cloud seeding study is to determine whether the seeded days had more rain than the unseeded days. Describe a graphical summary of the variable "rainfall" that you would do to help answer this question. Explain why it would be helpful.

The purpose here is to compare the rainfall amounts for the treatment and placebo groups. Any graphics that allow this would do. One example is a side-by-side boxplot; one boxplot shows rainfall for the seeded days, the other for the placebo days. Other examples: side-by-side histograms, or dotplots, or stem-and-leaf plots.

b) Use the summary below to comment on the shape of the distribrution of prewetness.

```
> summary (prewetness)
        Min.     1st Qu.         Median         Mean  3rd qu.          Max.
        0.0180          0.1405          0.220          0.3271          .0.3298          1.2670
```

The important thing to note is that the mean is larger than the median, which is a sign of a skewed-right distribution. Nothing can be said, from this information, about whether there are several humps or just one.

3. In 1969, famous pediatrician Dr. Benjamin Spock was on trial for conspiracy to violate the Military Service Act. (In addition to writing books on child development, he had spoken out against the Vietnam War.) As one journalist wrote, it was widely believed that Dr. Spock would want women on his jury. In one of the stages of the jury selection, the judge selected a panel of 100 potential jurors, and only 9 were women. In the district in which the trial was held, 53% of the eligible jurors were women.

Suppose 100 eligible jurors were selected WITH replacement. Let X represent the number of women selected. (Note: obviously, selecting with replacement is silly. But if the juror pool is large enough, then it is extremely unlikely that the same person would be selected twice, and so selected *with replacement would not be very different from selecting without.)*

a) Explain why X is a binomial random variable. What is the value for p, the probability of selecting a woman?

There are a fixed number of trials (100), and the trials are independent (since we assume sampling is done with replacement.) The outcome of each trial is success (woman) or failure (man). The probability of a success remains constant at 0.53 (p) for each trial (again because of the sampling with replacement.) Finally, the random variable X is counting the number of successes.

b) What is the expected number of women that should be on the panel of 100?
E(X) = np = 100*.53 = 53.

c) What is the standard deviation of the number of women that should be on the panel of 100?
SD(X) = sqrt(np(1-p)) = sqrt(100*.53*.47) = 4.99

d*) Use the normal distribution to find the approximate probability that 9 or fewer women would be selected on a panel of 100. (See attached tables.)*

The normal distribution can be used to approximate the binomial distribution. We're asked to find
P(X <= 9) where X is approximately N(53,4.99).
P(X <= 9) = P(Z <= (9-53)/4.99)) = P(Z <= -8.8) approximately equals 0.

Sort of a trick question, in that you don't need the tables to tell you that this probability will be exceptionally small.

In short, it is very unlikely that such a woman-shy jury could have arosen purely by chance.

4) In class we've talked about the game of Roulette, but always in the context of betting on a color (red or black). Another bet is to bet on two numbers (any two between 00, 0, 1, 2,...,36). This is called a "split bet". If either of the two numbers appears, you win $17. Otherwise, you lose $1.

Let X represent the amount you win on a single spin of the wheel if you place your bet on two numbers.

a) Sketch the pdf of X.

I can't do this on the computer, but note that this is a DISCRETE RV. And it takes only two values: -1 and +17. Therefore, there should be a "spike" at -1 that indicates a probabilty of 36/38 and a spike at +17 that indicates a probability of 2/38.

b) What's the expected value? Indicate it on the sketch of the pdf.

$E(X) = -1 * (36/38) + 17(2/38) = -0.0526$.

c) What's the standard deviation?

$Var(X) = (-1 - -0.0526)^2 * (36/38) + (17 - -0.0526)^2 * (2/38) = 16.15512$

$SD(X) = 4.019$

d) Suppose this game is played 1000 times and Y represents the average winnings from all 1000 spins. Sketch the pdf of Y, indicating the value at which the mean occurs.

By the Central Limit Theorem, the pdf of Y will be approximately normal. So your sketch should show a bell-shaped curve centered at the mean value of $E(Y) = -0.0526$. The SD of Y is $SD(Y) = 4.019/\text{sqrt}(1000) = .1270$.

e) *Betting on a color, the bet we discussed in class, had an expected value of -0.0526 and a SD of 0.998. Suppose you play both types of bets, and each you play 1000 times. So 1000 times you bet on Red, and let your total winnings be X. Then you play 1000 times by "split betting", and call your total winnings Y. Which of these two is more likely to be a positive number? That is, which are you more likely to have money in your pocket at the end of the game? You do not need to do any calculations (but you can if you want), and might want to use a sketch to support your answer. Be sure to explain.*

If all has gone well for you up to here, you've noted that the means of X and Y are the same, negative 0.0526, but the SDs are different. So the centers of their distributions are both the same distance from 0, but the one with the bigger spread is going to give you a greater chance of seeing positive values. In this case, the biggest spread belongs to Y. Therefore you are more likely to have money in your pocket by playing split bets than the other bets.

If you want to do calculations, then you should note that by the CLT we can assume that both distributions are approximately normal. X's distribution is, approximately, N(-52.6, 31.559).
The 31.559 comes from $SD(X) = \text{sqrt}(n)*SD = \text{sqrt}(1000)*.998$.
Y's distribution is N(-52.6, 126.5).

So $P(X > 0) = P(Z > 52.6/31.6) = P(Z > 1.66) = 1 - \text{pnorm}(1.66) = .05$ (approximately).
$P(Y > 0) = P(Z > 52.6/126.5) = 1 - \text{pnorm}(.4158) = .34$.