6 · Permutation tests; basic probability

Due Monday, March 2, 2015

(1) Testing the effectiveness of a new drug

Complete the walkthrough on 2x2 contingency tables[1] from the class website. Then use what you learn to answer the following question.

[1] http://jgscott.github.io/teaching/r/titanic/titanic_permtest.html

In the dataset "respir.csv" on the course website, you will find the results of a randomized, double-blind placebo-controlled clinical trial for a new drug meant to treat patients with respiratory symptoms. The trial involved 111 patients at two different clinics. Patients were randomized to receive either placebo (denoted P in the "treat" column) or an active treatment (denoted A). Patients were examined at baseline and at four subsequent visits during the course of the trial. At each examination, a doctor determined the respiratory status (1 = good, 0 = poor) of the patient.

Assess the evidence in favor of the claim that the treatment was more effective than placebo at improving patient's respiratory symptoms. Remember the subset command, in case you want to "slice and dice":

```
xtabs(~treat+outcome, data=respir)
xtabs(~treat+outcome, data=subset(respir, visit==4))
xtabs(~treat+baseline, data=subset(respir, visit==4))
```

and so forth. Make sure to address the possibility that any observed differences in clinical outcomes could have arisen due to chance—like an unlucky shuffle of the cards.

(2) Multiple regression and permutation testing

The data in "georgia2000.csv" contains Georgia's county-level voting data from the 2000 presidential election. You might recall that the 2000 election was among the most controversial in history, and turned on an esoteric set of issues surrounding voting machines, vote counts, and the Equal Protection Clause of the Constitution.

This file contains the following information for all 159 counties in Georgia:

votes: number of votes recorded
ballots: number of ballots cast
equip: voting equipment (lever, optical, paper, punch card)
poor: coded 1 if more than 25% of the residents in a county live below 1.5 times the federal poverty line; coded 0 otherwise.

perAA: percent of people in the county who are African-American
urban: indicator of whether county is predominantly urban (1)
atlanta: indicator of whether the county is in Atlanta (1)
gore: number of votes for Gore
bush: number of votes for Bush

Your goal is to investigate the determinants of vote undercount, or the difference between the number of ballots cast and the number of legal votes recorded. There can be many different reasons for undercount. Voters may have chosen not to vote for any presidential candidate; they may have voted for more than one candidate, in which case their votes were disqualified; they may have misunderstood the instructions on the ballot; or the equipment may have simply failed to register their choices.

One possibility that worries state election boards is that certain kinds of voting machines (paper, lever, etc.)will undercount valid ballots at higher rates, and that some precincts are unable to afford better machines. Construct a good statistical model for vote undercount (or some transformation thereof) to address this question.[2] Use this model to assess the marginal effect of voting equipment in explaining the undercount pattern across Georgia in 2000, adjusting for other relevant factors.

[2] Remember that larger counties might have more undercounted ballots just because they are larger, and had more ballots to begin with!

Key questions:

1. Is there evidence for an effect due to voting equipment, adjusting for other factors? Combine what you know about permutation tests and multiple regression. You will need to choose an appropriate test statistic (analogous to the relative risk or odds ratio in 2x2 tables) to measure the relevance of equipment in a single number.

2. If there is an effect due to voting equipment, what is a plausible range of values for its size?

(3) Basic probability

(A) An 18th-century French nobleman offers to play the following game with you. You will pay him $5 for the right to roll a fair, six-sided die four times in a row. (Let's say you provide the die, so you know it's fair.) If at least one of the rolls comes up as a one, you get $10 back (that is, your original bet in addition to $5 in net winnings). If none of the rolls come up one, you get nothing back. Should you agree to play the game?[3]

[3] Often the best way to compute the probability of an event is to compute the probability of its opposite.

(B) Suppose you agree to play the game from Part A ten times in a row—that is, ten distinct games of four rolls each. What is the probability that you will win 6 or more of these games? Hint: remember the binomial distribution.

(C) After ten games, the French nobleman says, "Come now, this game is quite boring. Let's roll two dice at the same time instead. You'll still pay $5 to play. But this time you'll win $10 if we see a pair of ones at least once, and otherwise you'll win nothing. Now, since a single roll of one happens six times more often than a pair of ones, we will need to play the game six times longer to make the odds the same as before—so each game is 24 rolls, instead of four."

Is his reasoning sound? Do you change your mind about whether it's a good idea to play, or is this game essentially the same as the first?

(4) Bayes' rule

A medical condition known as SOS afflicts roughly 1 person out of every 1000. A test for this condition exists, but is imperfect: it gives a positive result for 95% of people who have SOS, and a negative result for 99% of people who do not have SOS.

What is the probability that a patient has SOS, given that he or she tests positive for the disease? Use Bayes' rule:

$$P(A \mid B) = \frac{P(A) \cdot P(B \mid A)}{P(B)}.$$

If you get stuck, follow the steps below, and try mapping each step onto a different term in Bayes' rule.

(1) Recall that 1 in every 1000 people has SOS. In a population of 10,000 people, roughly how many will have the disease?
(2) Of the people in this population of 10,000 that do have SOS, about how many would test positive?
(3) Of the people in this population of 10,000 that do not have SOS, about how many would test positive?
(4) Adding your results from (2) and (3), how many total positive tests will there be among your population of 10,000?
(5) Of the number of positive tests you computed in (4), how many were from the sub-population of people that DID have the disease? Hint: this is the same answer as in (2)!
(6) Combining (4) and (5), deduce the probability that someone who tests positive actually has SOS.