CS 70 FALL 2006 — DISCUSSION #13

D. GARMIRE, L. ORECCHIA & B. RUBINSTEIN

1. Administrivia

Course Information

- Homework #12 is due today at 5pm in the drop box after question 4 was slightly changed on Monday.
- Homework # 13 will be due next Wednesday, it will be the last homework.
- Next week we will post exercises and sample final exams. There will be a review session for the final, probably on the weekend before the exam.

2. A last note on the Normal Distribution

The following are two important properties of the Normal distribution. They can both be proved directly by using the formula for the probability density:

Theorem 1. *If X_1 is Normal with mean μ_1 and variance σ_1^2 and X_2 is also (independently) Normal with mean μ_2 and variance σ_2^2, then:*

- *$X_1 + X_2$ is Normal with mean $\mu_1 + \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$.*
- *for any constant c, cX_1 is Normal with mean $c\mu_1$ and variance $c^2\sigma_1^2$.*

Consider n independent samples $X_1, \cdots, X_n$ from $\mathcal{N}(\mu, \sigma^2)$.

Exercise 1. Conclude that $\frac{X_1+X_2+\cdots X_n}{n}$ is Normal with mean μ and variance $\frac{\sigma^2}{n}$.

3. Power Laws

Last week you were introduced to power laws, a remarkably ubiquitous family of distributions. The first of these we encountered was Zipf's Law: this is very common for analyzing the frequency of single words in large bodies of text. If freq_i is the frequency of the i^{th} most frequent word, then Zipf's Law states that:

$$\text{freq}_i = \frac{\text{freq}_1}{i}$$

This can be generalized to include a fixed exponent:

$$\text{freq}_i = \frac{\text{freq}_1}{i^\beta}$$

Notice that Zipf's Law is an estimate of the value of the i^{th} ranked element in a population and, as such, it does not involve any notion of probability.

However, suppose we have a population of N elements obeying Zipf's Law and we sample a value X uniformly from it. We now have a probability distribution over the values of the element in the population. What is the probability that $X \geq x$?

Date: November 29, 2006.

The authors gratefully acknowledge Chris Crutchfield and Amir Kamil for the use of their previous notes, which form part of the basis for this handout.

Well, if x is the value of the i^{th} ranked element, x will be $\frac{c}{i^\beta}$ and the uniform sample X will be larger or equal to it with probability $\frac{i}{N}$. Hence, for $x = \frac{c}{i^\beta}$, i.e. $i = \left(\frac{c}{x}\right)^{\frac{1}{\beta}}$:

$$\Pr\left(X \geq x\right) = \frac{i}{N} = \frac{c^{\frac{1}{\beta}}}{N} x^{-\frac{1}{\beta}}$$

This is known as the Pareto distribution and takes general form:

$$\Pr\left(X \geq x\right) = Dx^{-\alpha+1}$$

Exercise 2. The Pareto distribution is defined in general on a truncated interval $[1, B]$ for some large B. For which values of α can we define it over the infinite interval $[1, +\infty]$? For which of these values does it have a finite expectation?

Exercise 3. The population of cities within a country is a statistics that is commonly taken to follow the Pareto distribution with α close to 1. What does a larger value of α mean for a specific country (more or less evenly sized cities)? Curiously, different values of α for different countries indicate very different models of development: for examples, most European countries have high values of α (around 1.3 or even 1.4), while African countries tend to be close to 1 or less.

4. Simpson's Paradox

This is another example of Simpson's Paradox, which you saw in class.
Alice and Bob both play Backgammon against their computer. In the first week, Alice wins 10% of the games she plays, while Bob wins only 5%. In the second week, they both do better: Alice wins 20% and Bob wins 15% of the games. Suppose both Alice and Bob played a total of 100 games over the two weeks and they both play at least one game in each week.

Exercise 4. Is it possible that overall percentage of games won over the two weeks is larger for Bob than for Alice? If yes, show an example. Otherwise prove it is not possible.

Exercise 5. How does the answer to the previous question change if Bob wins only 10% of the games in the second week?

By now the catch in all these examples of Simpson's paradox should be clear to you: the success percentages for different categories are not sufficient to reconstruct the overall success percentage, as they must be weighted by the size of each category. To visualize this geometrically consider the example above: construct 4 vectors A_1, A_2, B_1, B_2 in the plane. A_i will have ordinate equal to the number of games won by Alice in the i^{th} week and coordinate equal to the number of games played by Alice in the same week. Define B_i similarly for Bob.

Exercise 6. Show that the slope of A_i corresponds to the percentage of games won by Alice in week i.

Exercise 7. Show that the overall percentage of games won by Alice is the slope of $A_1 + A_2$

Exercise 8. Using this interpretation, construct a set of four vectors showing that Bob can have a larger overall win percentage in the example above.