CSE383M and CS395T Midterm Exam

March 20, 2013

1. An oracle generates Bernoulli i.i.d. answers that are either "yes" or "no". Two of the first three answers are "yes". (Call this circumstance "Event A".) What is the probility that exactly two of the next three answers will be "yes"? (Call this circumstance "Event B".) So you are supposed to calculate the probability of Event B given the data represented by Event A. Answer this question by working the following parts:

a) The nuisance variable in this problem is the oracle's probability $p(\text{"yes"}) \equiv x$. What is your prior probability distribution (p.d.f.) for x? (Write an actual formula, don't just say it in words.)

b) What is the posterior probability distribution (p.d.f.) for x given Event A?

c) What is the probability of Event B given some particular value of x?

d) Marginalize over x to get the final answer to the problem (a simple rational number).

2. You hypothesize that the following list of 20 numbers is drawn from a uniform distribution on the interval $(0, 1)$:

$0.6816, 0.4633, 0.1646, 0.0985, 0.8236, 0.1750, 0.1636, 0.6660, 0.1640, 0.5166,$
$0.1638, 0.1536, 0.9535, 0.5409, 0.1637, 0.0366, 0.8092, 0.7486, 0.1202, 0.1639$

a) Describe exactly what you would do to test this hypothesis. You don't actually have to perform the calculation that you describe, but you must give a complete enough description that I could implement it in computer code and actually try it.

b) State precisely what the possible outcomes of your test would be.

Points (100)

Bill

Midterm Exam Solution Set
3/20/13

(10) 1a) Uniform prior is my choice

$$p(x) = \begin{cases} 1 & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

(10) b) $P(x \mid A) \propto P(A \mid x)\, p(x)$

$$= \frac{x^2(1-x) \cdot 1}{\underbrace{\int x^2(1-x)\,dx}_{1/12}} = 12x^2(1-x)$$

no binomial coeff! We saw exactly some particular data

(10) c) $P(B \mid x) = \binom{3}{2} x^2 (1-x) = 3x^2(1-x)$

yes binomial coefficient because event B actually contains 3 possible, distinguishable events

(20) d) $P(B \mid A) = \int P(B \mid x)\, P(x \mid A)\, dx$

$$= \int_0^1 36\, x^4 (1-x)^2\, dx = \frac{12}{35}$$

2. $H \equiv$ (null) hypothesis that the values $\sim U(0,1)$

(40) a) We will do a p-value test.

We need a test statistic, and its distribution under the null hypothesis.

Hmm. Something weird about those numbers. 6 of them (out of 20) are almost equal to 0.1640.

Bill

So, I'd better think of a test statistic with some ~~ability~~ "power" to see "clumping" like this.

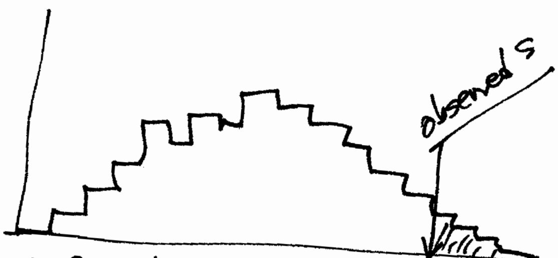
Maybe this: Sort the values, so that X_i, $i = 1..20$ are in ascending order.

$$\text{then} \quad S \equiv \sum_{i=2}^{20} \frac{1}{|X_i - X_{i-1}|}$$

So, close values will give large S.

I would now find the distribution of S under the null hypothesis by simulation:

```
do 100000 times:
    ~~pre~~ generate 20 values iid U(0,1)
    sort them
    calculate S
    record the value.
enddo
plot histogram of S
```



If n is the number of recorded values with $S \geq S_{observed}$, then my p-value would be

$$p = 2\frac{n}{100000}$$

2-tailed test, although one might argue that, since I observed the clumping, 1-tailed is OK.

b) If $p < 0.01$ (my favorite value as a physicist)
Conclude that H is rejected: data not uniform

(10) Otherwise conclude that H not rejected, but not, of course proved to be true.

1. a) $p(x) = \begin{cases} 1 & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$

Jeff

b) $$p(x \mid \text{data}) = \frac{p(\text{data} \mid x)\, p(x)}{\int_0^1 p(\text{data} \mid x')\, p(x')\, dx'}$$

$$= \frac{x^2 (1-x)}{\int_0^1 (x')^2 (1-x')\, dx'}$$

$$= \begin{cases} 12\, x^2 (1-x), & 0 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}$$

c) $$p(B \mid x) = \binom{3}{2} x^2 (1-x)$$

$$= 3 x^2 (1-x)$$

d) $\cancel{p(B) =}$

$$p(B \mid \text{data}) = \int_0^1 p(B \mid x)\, p(x \mid \text{data})\, dx$$

$$= \int_0^1 3x^2(1-x) \cdot 12\, x^2 (1-x)\, dx$$

$$= 36 \int_0^1 x^4 (1-x)^2\, dx = \frac{12}{35}$$

Jeff

2. a) I will construct a test statistic

$$T = \sum_{d=1}^{4} \sum_{i=0}^{9} \left| \left(\sum_{j=1}^{20} \text{digit } d \text{ of number } j \text{ equals } i \right) - 2 \right|$$

I will emperically characterize the distribution of this statistic under the null hypothesis by drawing many sets of real uniform (0,1) deviates.

I will find the fraction of such empirical draws that produce a larger value of T than the real data.

b) If this fraction is less than α, I will rule out that they are uniform at the α level. If it is not, I can't draw any conclusions.