IE300, FALL 2012

RICHARD B. SOWERS

1. 8/28/2012

Reasons for taking IE300

- You will be confronted with *large* datasets. How do you draw conclusions from them and defend your conclusions?
- You are faced with probabilistic (i.e., statistical) descriptions of things. How do you
  - Include observations?
  - Make decisions?
- You have to pick a scale for any model you deal with; behavior on scales smaller than the model can often best be modelled probabilistically.

Skills you will have by the end of the class

- Modelling toolbox
- Understanding how randomness wants to organize (think "the curve")

This class will be *hard*. Why?

- We *assume* mastery of certain calculations
- Often, the answer to a question includes building a probabilist model. This may involve ambiguities (e.g., combinations vs. permutations; we will understand this example later).
- In combinatorial probability, calculations often involve two steps;
  - Figure out *one* way that something can happen
  - Figure out *all* ways that something can happen.

Foundation of probability is an event space, often referred to as Ω, which is the collection of all outcomes of an experiment. For a die $\Omega = \{1, 2, 3, 4, 5, 6\}$, and for a coin toss $\Omega = \{H, T\}$, and for two coin tosses $\Omega = \{(H, H), (H, T), (T, H), (T, T)\}$. An event is a subset of Ω.

Next time (8/31): Chapter 2

Often, we deal with *equally likely* outcomes;

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|}$$

Thus we want to be able to *count*.

Pick gender, first letter of first name, and first letter of last name. There are

$$2 \times 26 \times 26 = 1352$$

that this can occur. *Fundamental principle of counting.* If

- experiment 1 has n_1 possible outcomes
- experiment 2 has n_2 possible outcomes
- experiment 3 has n_3 possible outcomes

(we'll stop at 3 experiments), then the number of possible outcomes of the sequence of experiments is

$$n_1 \times n_2 \times n_3$$

outcomes

Counting:

- Permuntations
- Combinations
- Sampling with replacement
- Sampling without replacement

We want to compute probabilities of *sets* (i.e., events); need to know

- Unions
- Intersections
- Complements

Rules of probability are natural

- ≥ 0
- ≤ 1 (and want to include $\mathbb{P}(\Omega) = 1$)
- don't overcount

2. 8/30/2012

Two parts of a probability model[1]

- Event space Ω
- a *probability* for every event A (event is subset of A); i.e., $\mathbb{P}(A)$

Three rules of probability

- $\mathbb{P}(\Omega) = 1$ (something has to happen)
- $\mathbb{P}(A) \geq 0$ for all events A
- For a countable collection $\{A_n\}_{n=1}^{\infty}$ of disjoint events,

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n)$$

  (often, but not always, a finite sum is sufficient); *don't overcount*

If Ω is finite, we use *equally likely* to describe the probability

$$\mathbb{P}(A) \stackrel{\text{def}}{=} \frac{|A|}{|\Omega|}$$

where $|A|$ is the cardinality of A.

Counting:

- Sampling without replacement (permutations)
- Sampling with replacement
- Combinations

Combinations; How many hands of 5 cards? Call this x. Pick 5 cards in order and place on a pile Two ways to do this;

- Sample 5 times, placing on the pile

$$(52)_5 = \underbrace{52 \times 51 \times 50 \times 49 \times 48}_{5 \text{ times}} = \frac{52!}{47!} = \frac{52!}{(52-5)!}$$

- – Grab 5 cards (can do this x ways)
  – Order them (can do this 5! ways)

Thus $(52)_5 = x5!$;

$$x = \frac{(52)_5}{5!} = \frac{52!}{(52-5)!5!}$$

Did problems

- 2-42, p.31
- 2-47, p. 31
- 2-78, p. 40

Next time (9/4): Chapter 2

- Conditional probability
- Independence

[1] Actually, there is a third part; for uncountable event spaces one can only evaluate $\mathbb{P}(A)$ for a "nice" collection of admissible events.

3. 9/4/2012

Conditional probability:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

If equally likely, this is reduction of event space. If not equally likely, you have to compute numerator and denominator.

- Example 2-24
- Exercise 2-87

Can rearrange $\mathbb{P}(A|B)$ to give

$$\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B).$$

This can be used in *Model Building*!

- Exercise 2-105
- Exercise 2-112

Several thoughts:

- Can be several ways to solve problem
- Don't be slaved to time directionality.

Next time (9/6): Chapter 2

- Independence

- Exercise 2-112.

Independence: $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$, which is equivalent to

$$\mathbb{P}(A|B) = \mathbb{P}(A)$$

(conditioning doesn't help). The following are equivalent:

- A and B are independent
- A and B^c are independent
- A^c and B are independent
- A^c and B^c are independent

Several thoughts:

- Independence is not the same as disjointedness
- Independence can be used to build models.

Exercises:

- 2-125
- 2-127
- 2-131

Next time (9/11): Chapter 2

- Bayes' rule

5. 9/11/2012

Bayes' rule (and possibly implicit conditional probabilities)

- HW example
- Redo exercise 2-127 using Bayes' rule
- 2-142
- Overlook probabilites

Random variables (if possible)

Next time (9/13): Chapter 2

- Random Variables
- Start on Chapter 3 (statistical description of random variables).

6. 9/13/2012

Random variables and probability mass functions; $f_X(3) = \mathbb{P}\{X = 3\}$

- $f_X \geq 0$
- $\sum_x f_X(x)$

Some well-known random variables.

- binomial
- Bernoulli
- discrete uniform
- geometric

Expectations;

$$\mathbb{E}[X] = \sum \text{winnings} \times \text{probabilities} = \sum_{\omega \in \Omega} X(\omega)\mathbb{P}\{\omega\} = \sum_x x \underbrace{\mathbb{P}\{X = x\}}_{f_X(x)}.$$

and

$$\mathbb{E}[G(X)] = \sum G(\text{winnings}) \times \text{probabilities} = \sum_{\omega \in \Omega} G(X(\omega))\mathbb{P}\{\omega\} = \sum_x G(x) \underbrace{\mathbb{P}\{X = x\}}_{f_X(x)}.$$

- mean
- variance
- general expectations
  - moment generating function to get means and variances

Means and variances of some common random variables.

Next time (9/18): Chapter 3 More on random variables

7. 9/18/2012

Discrete Random variables

uniform $$f(j) = \begin{cases} \frac{1}{N} & j \in \{1, 2 \dots N\} \\ 0 & \text{else} \end{cases}$$

Bernoulli (coin flip) $$f(j) = \begin{cases} p & j = 1 \\ 1-p & j = 0 \end{cases}$$

Binomial (N coin flips) $$f(j) = \begin{cases} \binom{N}{j} p^j (1-p)^{N-j} & j \in \{0, 1 \dots N\} \\ 0 & \text{else} \end{cases}$$

Geometric (first heads) $$f(j) = \begin{cases} (1-p)^{j-1} p & j \in \{1, 2 \dots\} \\ 0 & \text{else} \end{cases}$$

Negative Binomial ($k-th$ heads) $$f(j) = \begin{cases} \binom{j-1}{k-1} (1-p)^{j-k} p^k & j \in \{k, k+1 \dots\} \\ 0 & \text{else} \end{cases}$$

Poisson (limit of Binomial) $$f(j) = \begin{cases} e^{-\lambda} \frac{\lambda^k}{k!} & j \in \{0, 1 \dots\} \\ 0 & \text{else} \end{cases}$$

Hypergeometric (pick k red balls) $$f(j) = \begin{cases} \frac{\binom{n}{k}\binom{N-n}{k-j}}{\binom{N}{k}} & j \in \{k, k+1, \dots n\} \\ 0 & \text{else} \end{cases}$$

Expectations; a *linear* operator

- Means
- Variances
- Moment Generating Functions

Next time (9/20): Review

8. 9/27/2012

Expectations and variances of discrete random variables;

- Uniform random variable: $\mu = (N+1)/2$, $\sigma^2 = (n^2-1)/12$.
- Bernoulli $\mu = p$, $\sigma^2 = p(1-p)$.
- Binomial: $\mu = np$, $\sigma^2 = np(1-p)$.
- Geometric: $\mu = \frac{1}{p}$, $\sigma^2 = (1-p)/p^2$.
- Negative binomial: $\mu = \frac{k}{p}$, $\sigma^2 = k(1-p)/p^2$.
- Poisson: $\mu = \lambda$, $\sigma^2 = \lambda$.
- Hypergeometric: complicated

Note:

$$\phi(\theta) = \int_{t=0}^{\infty} e^{-\theta t} dt = \frac{1}{\theta}.$$

Then

$$\phi^{(n)}(\theta) = (-1)^n \int_{t=0}^{\infty} t^n e^{-\theta t} dt = (-1)^n n! \theta^{-n-1};$$

thus

$$\int_{t=0}^{\infty} t^n e^{-t} dt = n!$$

Similar for random variables. Think about Poisson

$$\begin{aligned}\varphi(\theta) = \mathbb{E}[e^{\theta X}] &= \sum_{n=0}^{\infty} e^{\theta k} f(k) = \sum_{n=0}^{\infty} e^{\theta n} e^{-\lambda} \frac{\lambda^n}{n!} \\ &= e^{-\lambda} \exp\left[\lambda e^{\theta}\right] = \exp\left[\lambda(e^{\theta}-1)\right]\end{aligned}$$

Have that

$$\varphi'(\theta) = \sum_{n=0}^{\infty} k e^{\theta k} f(k) = \lambda e^{\theta} \exp\left[\lambda(e^{\theta}-1)\right];$$

then

$$\mathbb{E}[X] = \sum_{n=0}^{\infty} k f(k) = \varphi'(0) = \lambda.$$

Next time (10/2):

More of the same.

9. 10/2/2012

Moment generating functions;

$$\varphi(\theta) \stackrel{\text{def}}{=} \mathbb{E}[e^{\theta X}].$$

$$\mathbb{E}[X] = \varphi'(0) \qquad \text{and} \qquad \mathbb{E}[X^2] = \varphi''(0)$$

so

$$\mu = \mathbb{E}[X] \qquad \text{and} \qquad \sigma^2 = \varphi''(0) - (\varphi'(0))^2.$$

Have

- Bernoulli $\varphi(\theta) = 1 - p + pe^{\theta}$
- Binomial $\varphi(\theta) = (1 - p + pe^{\theta})^n$.
- Geometric $\varphi(\theta) = p/(e^{-\theta} - (1-p))$ (if $\theta < \ln(1-p)^{-1}$)
- Negative binomial $\varphi(\theta) = \big(p/(e^{-\theta} - (1-p))\big)^n$ (if $\theta < \ln(1-p)^{-1}$)
- Poisson $\varphi(\theta) = \exp\big[\lambda(e^{\theta} - 1)\big]$.
- Uniform $\varphi(\theta) = e^{\theta}(e^{n\theta} - 1)/(e^{\theta} - 1)$.

Cumulative distribution function

$$F(t) = \mathbb{P}\{X \leq t\}.$$

Continous random variables;

$$\mathbb{P}\{a \leq X \leq b\} = \int_{t=a}^{b} f(t)dt$$

where f is *density*.

Next time (10/4): Chapter 4 More densities Means and expectations of continuous random variables

10. 10/4/2012

Properties of a density

- nonnegative
- integrates to 1

cumulative distribution function is integral of density
Exponential random variables; 4-95 and 4-99

Next time (10/9): Chapter 4

- Normal random variables
- Uniform random variables

11. 10/9/2012

Last time we discussed exponential random variables. A random variable is, by definition, exponential(λ) (where $\lambda > 0$ is a parameter similar in spirit to the bias on a coin) if it has density

$$f(t) = \begin{cases} \lambda e^{-\lambda t} & \text{if } t \geq 0 \\ 0 & \text{if } t < 0 \end{cases}$$

or alternately if it has cumulative distribution function

$$F(t) = \begin{cases} 1 - e^{-\lambda t} & \text{if } t \geq 0 \\ 0 & \text{if } t < 0 \end{cases}$$

Exponentials are connected to Poissons. Suppose that we have a sequence of events and that the interarrival times are independent and identically distributed exponential(λ) random variables. Fix $t \geq 0$ and let X be the number of events which have occurred by time t. Then X is Poisson with parameter λt; i.e.,

$$\mathbb{P}\{X = k\} = \begin{cases} e^{-\lambda t} \frac{(\lambda t)^k}{k!} & \text{if } k \in \{0, 1 \dots\} \\ 0 & \text{else} \end{cases}$$

Exponentials can be rescaled. Let X be an exponential(1) random variable. Define $Y \stackrel{\text{def}}{=} \frac{1}{3}X$. For any $t \geq 0$,

$$F_Y(t) \stackrel{\text{def}}{=} \mathbb{P}\{Y \leq t\} = \mathbb{P}\left\{\frac{X}{3} \leq t\right\} = \mathbb{P}\{X \leq 3t\} = F_X(t) = 1 - e^{-3t}$$

and since $Y \geq 0$,

$$F_Y(t) = \begin{cases} 1 - e^{-3t} & \text{if } t \geq 0 \\ 0 & \text{if } t < 0 \end{cases}$$

differentiating with respect to t, we see that Y is in fact exponential(3).

We can get more complicated. Define

$$\Phi(x) \stackrel{\text{def}}{=} \begin{cases} 1 - e^{-t} & \text{if } t \geq 0 \\ 0 & \text{if } t < 0 \end{cases}$$

(yes, this *is* the cumulative distribution function of an exponential(1), but let's temporarily forget about that). Let X be exponential(1), and define

$$U \stackrel{\text{def}}{=} \Phi(X).$$

We have that $0 \leq U \leq 1$, and for any $t \in (0, 1)$,

$F_U(t) = \mathbb{P}\{U \stackrel{\text{def}}{=} t\} = \mathbb{P}\{1-e^{-X} \leq t\} = \mathbb{P}\{e^{-x} \geq 1-t\} = \mathbb{P}\{-x \geq \ln(1-t)\} = \mathbb{P}\{X \leq -\ln(1-t)\} = F_X\left(-\ln(1-t)\right) = 1-e^{\ln(}$

Differentiating, we get that

$$f_U(t) = \begin{cases} 1 & \text{if } 0 < t < 1 \\ 0 & \text{else} \end{cases}$$

and U is Uniform$(0, 1)$. Reversing, we get that

$$X = \Phi^{-1}(U)$$

where in fact

$$\Phi^{-1}(x) = \ln \frac{1}{1-x}$$

In other words, *we can generate an exponential by transforming a uniform.*

Standard Gaussians. 4-58, 4-61

Next time (10/11)

12. 10/11/2012

Gaussians. If X is Gaussian with mean 10 and standard deviation 7, X has density

$$f_X(t) = \frac{1}{\sqrt{2\pi \times 49}} \exp\left[-\frac{(t-10)^2}{2 \times 49}\right].$$

Then $Z = \frac{X-10}{7}$ is a standard normal.

Next time (10/18):

13. 10/18/2012

Chapter 5: Joint random variables

- Joint probability mass function
- Conditional mass functions
- marginals

Example; X_1 is first heads, and X_2 is second heads. $\mathbb{P}\{H\} = p$.

$$\underbrace{\mathbb{P}\{X_1 = 4, X_2 = 7\}}_{p_{X_1,X_2}(4,7)} = p^2(1-p)^5.$$

Also,

$$\underbrace{\mathbb{P}\{X_1 = 4 | X_2 = 7\}}_{p_{X_1|X_2}(4|7)} = \frac{\mathbb{P}\{X_1 = 1, X_2 = 7\}}{\mathbb{P}\{X_2 = 7\}} = \frac{p^2(1-p)^5}{6p^2(1-p)^5} = \frac{1}{6}.$$

Thus

$$\mathbb{E}[X_1 | X_2 = 7] = \sum_j j p_{X_1|X_2}(j|7) = \sum_{j=1}^{6} \frac{j}{6} = 3.5.$$

Can recover *marginals*

$$\underbrace{\mathbb{P}\{X_1 = 4\}}_{p_{X_1}(4)} = \sum_j \underbrace{\mathbb{P}\{X_1 = 4, X_2 = j\}}_{p_{X_1,X_2}(4,j)} = \sum_{j=5}^{\infty} p^2(1-p)^{j-2}$$

$$= \sum_{k=0}^{\infty} p^2(1-p)^{k+3} = p^2(1-p)^3 \frac{1}{1-(1-p)} = p(1-p)^3$$

$$\underbrace{\mathbb{P}\{X_2 = 7\}}_{p_{X_2}(7)} = \sum_j \underbrace{\mathbb{P}\{X_1 = j, X_2 = 7\}}_{p_{X_1,X_2}(j,7)} = \sum_{j=1}^{6} p^2(1-p)^5$$

14. 11/1/2012

Transformations of random variables. If X is a random variable (as an example, let X be a standard Gaussian) and ϕ is an increasing invertible map on $\mathbb{R}$ to itself (think $\phi(x) = e^x$, which maps $\mathbb{R}$ to $(0, \infty)$) and we define $Y = \phi(X)$ (in our developing example, Y would be lognormal), then

$$F_Y(t) = \mathbb{P}\{Y \leq t\} = \mathbb{P}\{\phi(X) \leq t\} = \mathbb{P}\{X \leq \phi^{-1}(t)\} = \int_{s=-\infty}^{\phi^{-1}(t)} f_X(s)ds$$
$$= \int_{r=\phi(-\infty)}^{s} f_X(\phi^{-1}(r))(\dot{\phi^{-1}})(r)dr$$

where $(\dot{\phi^{-1}})(r)$ is the derivative of the inverse of ϕ and where we have used here the transformation $r = \phi(s)$. Differentiating, we get that

$$\dot{f}_Y(t) = f_X(\phi^{-1}(t))(\dot{\phi^{-1}})(t)$$

(in our example, $f_Y(t) = \frac{1}{t\sqrt{2\pi}} \exp\left[-\frac{1}{2}(\ln x)^2\right]$). If ϕ is not invertible (i.e., two points can map onto the same point in the range), then one has to divide the calculation into several parts. As an example, let X again be a standard Gaussian and let $\phi(x) = x^2$. Then $Y = \phi(X) = X^2$ is a chi-squared distribution, and

$$f_Y(t) = \begin{cases} \frac{1}{\sqrt{2\pi}} e^{-t/2} & \text{if } t > 0 \\ 0 & \text{if } t < 0 \end{cases}$$

In this case one can invert ϕ on $(0, \infty)$ and $(-\infty, 0)$, but not on all of $\mathbb{R}$ at once.

In the multidimensional case, if $Y = \Phi(X)$ where X is a random vector, then

$$f_Y(t) = f_X(\Phi^{-1}(t)) \det(D\Phi^{-1}(t)).$$

($\det(D\Phi^{-1}(t))$ is essentially the Jacobian).

15. 11/6/2012

Analysis of data

- Sample mean
- Sample variance
- population variance

Data analysis

- histograms
- frequency plot
- scatter plots

Linear regression. Suppose that we have N different instances $\{X_1, X_2 \dots X_N\}$ of a *independent* variable and N different instances $\{Y_1, Y_2 \dots Y_N\}$ of a *dependent* variable. We want to find the "best" and "simplest" way to characterize the dependence of the Y_n's on the X_n's. You might think of X_n as your score on exam 2 and Y_n as your overall score. If you consider all students in the class (i.e., all the n's), what do you expect your overall score to be, based *only* on exam 2?

Let's define the N-dimensional column vectors

$$\vec{X} \stackrel{\text{def}}{=} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{pmatrix} \quad \text{and} \quad \vec{Y} \stackrel{\text{def}}{=} \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix} \quad \text{and} \quad \vec{\mathbf{1}} \stackrel{\text{def}}{=} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

We would like to find two real numbers α and β such that, upon writing the vector equation

$$\vec{Y} = \alpha\vec{X} + \beta\vec{\mathbf{1}} + \mathcal{E} \tag{1}$$

(we'll see in a moment why it is convenient to use $\mathbf{1}$ in our vector representation) where the error term $\mathcal{E}$ is as "small" as possible.

Equation (1) is an equation in $\mathbb{R}^N$; let's measure $\mathcal{E}$ in the standard way; define

$$E(\alpha, \beta) \stackrel{\text{def}}{=} \sum_{n=1}^{N} (\mathcal{E}_n)^2 = \|\mathcal{E}\|^2 = \|\vec{Y} - (\alpha\vec{X} + \beta\vec{\mathbf{1}})\|^2.$$

To make $\mathcal{E}$ as small as possible (i.e., to minimize E), we want to solve the Euler conditions; i.e., we want to solve

$$\begin{aligned} 0 = \frac{\partial E}{\partial \alpha}(\alpha, \beta) &= -\left\langle \vec{X}, \vec{Y} - (\alpha\vec{X} + \beta\vec{\mathbf{1}}) \right\rangle \\ 0 = \frac{\partial E}{\partial \beta}(\alpha, \beta) &= -\left\langle \vec{\mathbf{1}}, \vec{Y} - (\alpha\vec{X} + \beta\mathbf{1}) \right\rangle \end{aligned}$$

where (using some standard notation), $\left\langle \vec{a}, \vec{b} \right\rangle = \vec{a} \cdot \vec{b} = \vec{a}^T\vec{b}$. A bit of linear algebra later, we see that α and β should solve

$$\begin{pmatrix} \left\langle \vec{Y}, \vec{X} \right\rangle \\ \left\langle \vec{Y}, \vec{\mathbf{1}} \right\rangle \end{pmatrix} = \begin{pmatrix} \left\langle \vec{X}, \vec{X} \right\rangle & \left\langle \vec{X}, \vec{\mathbf{1}} \right\rangle \\ \left\langle \vec{X}, \vec{\mathbf{1}} \right\rangle & \left\langle \vec{\mathbf{1}}, \vec{\mathbf{1}} \right\rangle \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \tag{2}$$

Note that $\left\langle \vec{\mathbf{1}}, \vec{\mathbf{1}} \right\rangle = N$. Let's also define the "sample" averages

$$\mathbb{E}_s[X] \stackrel{\text{def}}{=} \frac{1}{N}\sum_{n=1}^{N} X_n = \frac{\left\langle \vec{X}, \mathbf{1} \right\rangle}{\left\langle \vec{\mathbf{1}}, \vec{\mathbf{1}} \right\rangle}$$

$$\mathbb{E}_s[Y] \stackrel{\text{def}}{=} \frac{1}{N}\sum_{n=1}^{N} Y_n = \frac{\left\langle \vec{Y}, \mathbf{1} \right\rangle}{\left\langle \vec{\mathbf{1}}, \vec{\mathbf{1}} \right\rangle}$$

$$\mathbb{E}_s[X^2] \stackrel{\text{def}}{=} \frac{1}{N}\sum_{n=1}^{N} X_n^2 = \frac{\left\langle \vec{X}, \vec{X} \right\rangle}{\left\langle \vec{\mathbf{1}}, \vec{\mathbf{1}} \right\rangle}$$

$$\mathbb{E}_s[XY] \stackrel{\text{def}}{=} \frac{1}{N}\sum_{n=1}^{N} X_n Y_n = \frac{\left\langle \vec{X}, \vec{Y} \right\rangle}{\left\langle \vec{\mathbf{1}}, \vec{\mathbf{1}} \right\rangle}$$

We can then divide (2) by N to get

$$\begin{pmatrix} \mathbb{E}_s[YX] \\ \mathbb{E}_s[Y] \end{pmatrix} = \underbrace{\begin{pmatrix} \mathbb{E}_s[X^2] & \mathbb{E}_s[X] \\ \mathbb{E}_s[X] & 1 \end{pmatrix}}_{M} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

Note that the determinant of M is the sample variance $\sigma_s(X,X) \stackrel{\text{def}}{=} \mathbb{E}_s[X^2] - (\mathbb{E}_s[X]^2)$ Thus

$$\begin{aligned} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} &= \frac{1}{\sigma_s(X,X)} \begin{pmatrix} 1 & -\mathbb{E}_s[X] \\ -\mathbb{E}_s[X] & \mathbb{E}_s[X^2] \end{pmatrix} \begin{pmatrix} \mathbb{E}_s[YX] \\ \mathbb{E}_s[Y] \end{pmatrix} \\ &= \frac{1}{\sigma_s(X,X)} \begin{pmatrix} \mathbb{E}_s[XY] - \mathbb{E}_s[X]\mathbb{E}_s[Y] \\ -\mathbb{E}_s[XY]\mathbb{E}_s[X] + \mathbb{E}_s[X^2]\mathbb{E}_s[Y] \end{pmatrix} \end{aligned}$$

We have that

$$\begin{aligned} \mathbb{E}_s[XY] - \mathbb{E}_s[X]\mathbb{E}_s[Y] &= \sigma_s(X,Y) \\ -\mathbb{E}_s[XY]\mathbb{E}_s[X] + \mathbb{E}_s[X^2]\mathbb{E}_s[Y] &= \mathbb{E}_s[Y]\left\{\mathbb{E}_s[X^2] - \mathbb{E}_s[X]^2\right\} - \mathbb{E}_s[X]\left\{\mathbb{E}_s[XY] - \mathbb{E}_s[X]\mathbb{E}_s[Y]\right\} \\ &= \sigma_s(X,X)\mathbb{E}_s[Y] - \mathbb{E}_s[X]\sigma_s(X,Y) \end{aligned}$$

where $\sigma_s(X,Y)$ is the empirical covariance between X and Y. Setting $\mu_X \stackrel{\text{def}}{=} \mathbb{E}_s[X]$ and $\mu_Y \stackrel{\text{def}}{=} \mathbb{E}_s[Y]$, we have

$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \frac{1}{\sigma_s(X,X)} \begin{pmatrix} \sigma_s(X,X) \\ \sigma_s(X,X)\mu_Y - \sigma_s(X,Y)\mu_X \end{pmatrix} = \begin{pmatrix} \rho_s \\ \mu_Y - \rho_s\mu_X \end{pmatrix}$$

where $\rho_s \stackrel{\text{def}}{=} \sigma_s(X,Y)/\sigma_s(X,X)$ is the empirical correlation between X and Y. Thus our "best" linear estimate of Y on the basis of X is

$$Y \approx \rho_s X - \rho_s\mu_X + \mu_Y = \rho_s(X - \mu_X) + \mu_Y$$

16. 11/8/2012

Estimation. Flip a biased coin (with bias $\mathbb{P}\{H\} = p$) 100 times, and record observations $(x_1, x_2 \dots x_{100})$. We want to reverse-engineer (i.e., *estimate*) p.

16.1. **Moment Estimator.** One way to estimate parameters is to compare empirical (population) moments with theoretical (sample) moments, and to then solve for the parameters. The basis of this is the law of large numbers. In the above case, the empirical (population) average is

$$\bar{x} \stackrel{\text{def}}{=} \frac{1}{100} \sum_{n=1}^{100} x_n$$

(the observed frequency of heads) and the theoretical (sample) average is

$$\mathbb{E}[X] = p.$$

A reasonable estimate of p is given by equating these two averages; i.e., our estimate of the bias is

$$p^* \stackrel{\text{def}}{=} \bar{x}.$$

16.2. **MLE estimator.** The probability that we got the observed sequence is

$$f_p(x_1, x_2 \dots x_{100}) = p^N (1-p)^{100-N}$$

where N is the number of heads. Let's maximize over p. Let's first take the logarithm to make things look nicer;

$$\max_{p \in [0,1]} \ln f_p(x_1, x_2 \dots x_{100}) = \sup_{p \in [0,1]} \{N \ln p + (100 - N) \ln(1-p)\}.$$

The first-order conditions of optimality require that

$$\frac{N}{p} - \frac{100 - N}{1-p} = 0$$

or rather

$$p = \underbrace{\frac{N}{100}}_{\Theta(x_1, x_2 \dots x_{100})}.$$

In fact, $N = \sum_{n=1}^{100} x_n$, so

$$p = \frac{1}{100} \sum_{n=1}^{100} x_n;$$

the MLE estimator is the same as the moment estimator.

If p_* is the true value of p, then

$$\mathbb{E}[\Theta(X_1, X_2 \dots X_{10})] = \frac{10 p_*}{10} = p_*$$

so this estimator is *unbiased*.

To get a better feel for these two estimators, let's next assume that $\{x_1, x_2 \dots x_{100}\}$ are all samples from a uniform distribution on $(0, a)$, where we want to estimate a (which is assumed to be positive). The first moment (mean) of a $U(0, a)$ distribution is

$$\frac{1}{a} \int_{x=0}^{a} x dx = \frac{a^2}{2a} = \frac{a}{2}.$$

The moment estimate of a thus satisfies

$$\bar{x} = \frac{a}{2}$$

where $\bar{x} \stackrel{\text{def}}{=} \frac{1}{100} \sum_{n=1}^{100} x_n$. In other words, the moment estimate of a is given by

$$a^* \stackrel{\text{def}}{=} 2\bar{x}.$$

The MLE estimate of a is given by writing

$$f_a(x) \stackrel{\text{def}}{=} \begin{cases} \frac{1}{a} & \text{if } 0 \le x \le a \\ 0 & \text{else} \end{cases}$$

Thus

$$\ln f_a(x) = \begin{cases} -\ln a & \text{if } 0 \le x \le a \\ -\infty & \text{else} \end{cases}$$

We thus want to maximize

$$\Phi(a) = \sum_{n=1}^{100} \ln f_a(x_n) = \begin{cases} -100 \ln a & \text{if } 0 \le x_n \le a \text{ for all } n \\ -\infty & \text{if } x_n \notin [0, a] \text{ for some } n \end{cases} = \begin{cases} -100 \ln a & \text{if } a \ge \max_n x_n \\ -\infty & \text{if } a < \max_n x_n \text{ for some } n \end{cases}$$

Since $a \mapsto -100 \ln a$ is decreasing, the Φ is maximal when a is as small as possible; the maximum likelihood estimate of a is thus given by

$$a^* \stackrel{\text{def}}{=} \max_n x_n.$$

17. 11/11/2012

Continuing on, let's try to be 90% "confident" of the value of p^*. The Central Limit Theorem tells us that

$$\frac{1}{\sqrt{N}}\sum_{n=1}^{N}\underbrace{\frac{X_n-p^*}{p^*(1-p^*)}}_{Z\text{-score}}\approx \underbrace{\mathfrak{N}(0,1)}_{\text{Standard Gaussian}}$$

From the tables,

$$\mathbb{P}\{\mathfrak{N}(0,1)\le 1.64\}=.95$$

so

$$\mathbb{P}\{|\mathfrak{N}(0,1)|\le 1.64\}=1-\mathbb{P}\{|\mathfrak{N}(0,1)|>1.64\}=1-2\mathbb{P}\{\mathfrak{N}(0,1)>1.64\}=1-2(1-.95)=0.9$$

Thus since 100 is large,

$$\mathbb{P}\left\{\left|\frac{1}{\sqrt{100}}\sum_{n=1}^{100}\frac{X_n-p^*}{p_*(1-p_*)}\right|\le 1.64\right\}\approx 0.9.$$

Rearranging, we get that

$$\mathbb{P}\left\{\left|\frac{1}{100}\sum_{n=1}^{100}X_n-p_*\right|\le\frac{1.64p_*(1-p_*)}{\sqrt{100}}\right\}\approx 0.9$$

or rather

$$\mathbb{P}\left\{\underbrace{\frac{1}{100}\sum_{n=1}^{100}X_n}_{\Theta}-\frac{1.64p_*(1-p_*)}{\sqrt{100}}\le p_*\le\underbrace{\frac{1}{100}\sum_{n=1}^{100}X_n}_{\Theta}+\frac{1.64p_*(1-p_*)}{\sqrt{100}}\right\}\approx 0.9$$

Of course we don't actually know p_*, but we do know that $p_*(1-p_*)\le\frac{1}{4}$; thus

$$\mathbb{P}\left\{\underbrace{\frac{1}{100}\sum_{n=1}^{100}X_n}_{\Theta}-\underbrace{\frac{1.64}{2\sqrt{100}}}_{.041}\le p^*\le\underbrace{\frac{1}{100}\sum_{n=1}^{100}X_n}_{\Theta}+\underbrace{\frac{1.64}{4\sqrt{100}}}_{.041}\right\}\approx 0.9$$

Let's look at exponential random variables. Fix $\lambda_*>0$ and let $\{X_1,\dots X_{100}\}$ be i.i.d. $\exp(\lambda^*)$ random variables. We want to estimate λ_*. The mean of $\exp(\lambda_*)$ is $\frac{1}{\lambda_*}$; that should be close to the sample mean $\frac{1}{100}\sum_{n=1}^{100}X_n$; we should have

$$\frac{1}{100}\sum_{n=1}^{100}X_n\approx\frac{1}{\lambda_*};$$

let's define the estimator

$$\Theta=\frac{100}{\sum_{n=1}^{100}X_n}.$$

Note, however, that there is a bias. It turns out that $\sum_{n=1}^{100}X_n$ has a known distribution; it is a gamma random variable with shape parameter 100 and scale parameter $1/\lambda_*$; it has density

$$f(x)=\begin{cases}\frac{\lambda_*^{100}}{\Gamma(100)}x^{99}e^{-\lambda_* x} & x>0\\ 0 & x\le 0\end{cases}$$

The mean of $100/\sum_{n=1}^{100}X_n$ is thus

$$\begin{aligned}\int_{x=0}^{\infty}\frac{N}{x}f(x)dx&=\int_{x=0}^{\infty}\frac{100\lambda_*^{100}}{\Gamma(100)}x^{98}e^{-\lambda_* x}dx\\&=\frac{100\lambda_*}{\Gamma(100)}\int_{x=0}^{\infty}(\lambda_* x)^{98}e^{-\lambda_* x}\lambda_* dx=\frac{100\Gamma(99)}{\Gamma(100)}\lambda_*=\frac{100\times 98!}{99!}\lambda_*=\frac{100}{99}\lambda_*\approx\lambda_*.\end{aligned}$$

18. 11/15/2012

Hypothesis testing. Suppose we have a collection of coin flips (i.e., Bernoulli random variables) $\{X_1, X_2 \dots X_{10}\}$. The bias μ ($\mathbb{P}\{H\} = \mu$) is either:

- p_N (we'll call this the *null hypothesis*); we'll let $\mathbb{P}_N$ be the probability measure on the coin flips in this case.
- p_A (we'll call this the *alternative hypothesis*); we'll let $\mathbb{P}_A$ be the probability measure on the coin flips in this case.

We would like to determine which value of μ is correct on the basis of the observations. Of course we want $p_A \neq p_N$ for the hypotheses to actually be different.

There are two types of errors

- *Type 1 Error* the null hypothesis is correct, and we decide that the alternative hypothesis is true. Let's take α to be the probability of making a Type I error.
- *Type II Error* the alternative hypothesis is correct, and we decide that the null hypothesis is true. Let's take β to be the probability of making a Type II error.

Let's decide between the hypotheses on the basis of the sample mean

$$\bar{X} \stackrel{\text{def}}{=} \frac{1}{10} \sum_{n=1}^{10} X_n.$$

Of course for N large $\bar{X}$ should be close to the actual value of μ, but for N finite, we still need to understand the tradeoff between the two types of errors.

To break symmetry, let's assume that $p_N < p_A$. We want to fix an x^* and decide

- the null hypothesis if $\bar{X} < x^*$
- the alternative hypothesis if $\bar{X} > x^*$

We want a Type I error to occur with probability of at most $\alpha = 0.05$. Define $\sigma_N \stackrel{\text{def}}{=} p_N(1 - p_N)$ and $\sigma_A \stackrel{\text{def}}{=} p_A(1 - p_A)$, and set

$$Z_N \stackrel{\text{def}}{=} \frac{1}{\sqrt{10}} \sum_{n=1}^{10} \frac{X_n - \mu_N}{\sigma_N} = \frac{\bar{X} - \mu_N}{\sigma_N/\sqrt{10}}$$

$$Z_A \stackrel{\text{def}}{=} \frac{1}{\sqrt{10}} \sum_{n=1}^{10} \frac{X_n - \mu_A}{\sigma_A} = \frac{\bar{X} - \mu_A}{\sigma_A/\sqrt{10}}$$

this should approximately be a under $\mathbb{P}_N$, Z_N is approximately a standard Gaussian, and under $\mathbb{P}_A$, Z_A is approximately a standard Gaussian. "Approximately" would be better if we would replace 10 with 100. We can write

$$\bar{X} = \mu_N + \frac{\sigma_N}{\sqrt{10}} Z_N$$

$$\bar{X} = \mu_A + \frac{\sigma_A}{\sqrt{10}} Z_A.$$

To get a Type I error with probability $\alpha = 0.05$, we want

$$\mathbb{P}_N\{\bar{X} > x^*\} = 0.05.$$

In other words, we want

$$\mathbb{P}_N \left\{ \mu_N + \frac{\sigma_N}{\sqrt{10}} Z_N > x^* \right\} = 0.05;$$

i.e.,

$$\mathbb{P}_N \left\{ Z_N \leq \frac{x^* - \mu_N}{\sigma_N/\sqrt{10}} \right\} = 0.95 = \mathbb{P}\{\mathfrak{G} \leq 1.65\}$$

so let's take

$$\frac{x^* - \mu_N}{\sigma_N/\sqrt{10}} = 1.65;$$

or rather

$$x^* = \mu_N + \frac{1.65\sigma_N}{\sqrt{10}}.$$

We can then compute the probability of a Type II error;

$$\beta = \mathbb{P}_A\{\bar{X} < x^*\} = \mathbb{P}_A\left\{\mu_A + \frac{\sigma_A}{\sqrt{10}}Z_B < \mu_N + \frac{1.65\sigma_N}{\sqrt{10}}\right\} = \mathbb{P}_A\left\{Z_A \leq \frac{\sqrt{10}}{\sigma_A}(\mu_N - \mu_A) + 1.65\frac{\sigma_N}{\sigma_A}\right\}$$

19.1. **Estimation of Variance.** . Let $\{X_n\}_{n=1}^{10}$ be random variables which are i.i.d. samples from a common distribution. We might estimate the variance of the X_n's by first forming

$$\hat{X} \stackrel{\text{def}}{=} \frac{1}{10}\sum_{n=1}^{10} X_n$$

and then forming the population variance

$$S_p \stackrel{\text{def}}{=} \frac{1}{10}\sum_{n=1}^{10}(X_n - \hat{X})^2.$$

Let's understand the bias in S_p. Let's assume that the *real* mean and variance of the X_n's is μ and σ; i.e.,

$$\mathbb{E}[X] = \mu \qquad \text{and} \qquad \mathbb{E}[(X-\mu)^2] = \sigma.$$

Then

$$\begin{aligned}
\mathbb{E}[S_p] &= \mathbb{E}\left[\frac{1}{10}\sum_{n=1}^{10}(X_n - \hat{X})^2\right] = \frac{1}{10}\sum_{n=1}^{10}\mathbb{E}[(X_n - \hat{X})^2] \\
&= \frac{1}{10}\sum_{n=1}^{10}\mathbb{E}\left[\left\{(X_n-\mu) - (\hat{X}-\mu)\right\}^2\right] \\
&= \frac{1}{10}\sum_{n=1}^{10}\left\{\mathbb{E}\left[(X_n-\mu)^2\right] - 2\mathbb{E}\left[(X_n-\mu)(\hat{X}-\mu)\right] + \mathbb{E}\left[(\hat{X}-\mu)^2\right]\right\}
\end{aligned}$$

In order to proceed, we have to recall that if $n \neq n'$, then

$$\mathbb{E}[(X_n-\mu)(X_{n'}-\mu)] = 0 \tag{3}$$

We directly have that $\mathbb{E}\left[(X_n-\mu)^2\right] = \sigma^2$; we also calculate that

$$\begin{aligned}
\mathbb{E}\left[(\hat{X}-\mu)^2\right] &= \mathbb{E}\left[\left(\frac{1}{10}\sum_{n=1}^{10}(X_n-\mu)\right)^2\right] = \frac{1}{(10)^2}\sum_{n=1}^{10}\sum_{n'=1}^{10}\mathbb{E}\left[(X_n-\mu)(X_{n'}-\mu)\right] \\
&= \frac{1}{(10)^2}\sum_{n=1}^{10}\sigma^2 \qquad \text{(the cross terms vanish due to (3))} \\
&= \frac{1}{10}\sigma^2
\end{aligned}$$

We also have that

$$\mathbb{E}\left[(X_n-\mu)(\hat{X}-\mu)\right] = \frac{1}{10}\sum_{n'=1}^{10}\mathbb{E}\left[(X_n-\mu)(X_{n'}-\mu)\right] \quad = \frac{1}{10}\mathbb{E}\left[(X_n-\mu)^2\right] \qquad \text{(we again use (3))}$$

Collecting things together, we get that

$$\mathbb{E}[S_p] = \frac{1}{10}\sum_{n=1}^{10}\left\{\sigma^2 - \frac{2}{10}\sigma^2 + \frac{1}{10}\sigma^2\right\} = \frac{1}{10}\sum_{n=1}^{10}\left\{\sigma^2\left(1-\frac{1}{10}\right)\right\} = \sigma^2\left(1-\frac{1}{10}\right)$$

A 'better' estimate of the variance is thus

$$\left(1-\frac{1}{10}\right)^{-1} S_p = \frac{1}{(1-\frac{1}{10})}\times\frac{1}{10}\sum_{n=1}^{10}(X_n-\hat{X})^2 = \frac{1}{10-1}\sum_{n=1}^{10}(X_n-\hat{X})^2$$

19.2. **Chi-square random variables.** Let X be a standard normal random variable and set $Y \stackrel{\text{def}}{=} X^2$. As we mentioned before,

$$\mathbb{P}\{Y \le 4\} = \mathbb{P}\{X^2 \le 4\} = \mathbb{P}\{-2 \le X \le 2\} = \int_{t=-2}^{2} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt = 2 \int_{t=0}^{2} \frac{2}{\sqrt{2\pi}} e^{-t^2/2} dt$$

and more generally,

$$\mathbb{P}\{Y \le y\} = 2 \int_{t=0}^{\sqrt{y}} \frac{2}{\sqrt{2\pi}} e^{-t^2/2} dt$$

for $y \ge 0$. Thus Y has density

$$f_{\chi_1}(t) = \frac{1}{\sqrt{2\pi t}} e^{-t/2}$$

This is similar to a Gamma distribution.

Let's next look at

$$Y \stackrel{\text{def}}{=} X_1^2 + X_2^2$$

where X_1 and X_2 are i.i.d. standard Gaussians. The distribution of Y now is

$$\begin{aligned} f_{\chi_2}(t) &= \int_{s=-\infty}^{\infty} f_{\chi_1}(s) f_{\chi_2}(t-s) ds = \frac{1}{2\pi} \int_{s=0}^{t} \frac{1}{\sqrt{s}} \frac{1}{\sqrt{t-s}} e^{-s/2} e^{-(t-s)/2} ds \\ &= \frac{1}{2\pi} e^{-t/2} \int_{s=0}^{t} \frac{1}{\sqrt{s}} \frac{1}{\sqrt{t-s}} ds \end{aligned}$$

for $t > 0$. Note that

$$\int_{s=0}^{t} \frac{1}{\sqrt{s}} \frac{1}{\sqrt{t-s}} ds = \int_{r=0}^{1} \frac{1}{\sqrt{r}} \frac{1}{\sqrt{1-r}} dr.$$

Thus

$$f_{\chi_2}(t) = \begin{cases} Ce^{-t/2} & \text{if } t \ge 0 \\ 0 & \text{else} \end{cases}$$

(should probably take $C = 1/2$).

Department of Industrial and Systems Engineering, University of Illinois at Urbana–Champaign, Urbana, IL 61801

E-mail address: r-sowers@illinois.edu