# Lecture Stat 302
# Introduction to Probability - Slides 7

AD

Feb. 2010

# Simpson's Paradox: Sex Bias in Graduate Admissions?

- The University of California at Berkeley was sued for bias against women who had applied for admission to graduate schools there. The admission figures for the fall of 1973 showed that men applying were more likely than women to be admitted, and the difference was so large that it was unlikely to be due to chance.
- The evidence looks really compelling.

| | Applicants | % admitted |
|---|---|---|
| Men | 8442 | 44% |
| Women | 4321 | 35% |

- If we consider the event $A =$"Admitted", $B =$"Be a woman", $B^c =$"Be a man", we have

$$P\left(A\right|B) \leq P\left(A\right|B^c).$$

- Was Berkeley really biased?

# Simpson's Paradox: Sex Bias in Graduate Admissions?

- Let us examine more carefully the data by examining admissions in some representative departments

| Depart. | Men | Women |
|---|---|---|
| | % in | % in |
| 1 | 63 | 68 |
| 2 | 33 | 35 |
| 3 | 6 | 7 |

- Whatever the department to which they apply, women have an higher probability of getting in than men; i.e. we have for any event $C_i = \{\text{apply to department } i\}$

$$P\left(A|\, B \cap C_i\right) \geq P\left(A|\, B^c \cap C_i\right)$$

and still overall

$$P\left(A|\, B\right) \leq P\left(A|\, B^c\right)$$

# Simpson's Paradox: Sex Bias in Graduate Admissions?

- Let us examine even more carefully the data

| Depart. | Men | | Women | |
|---|---|---|---|---|
| | Appli. | % in | Appli. | % in |
| 1 | 560 | 63 | 25 | 68 |
| 2 | 417 | 33 | 375 | 35 |
| 3 | 272 | 6 | 341 | 7 |

# Simpson's Paradox: Sex Bias in Graduate Admissions?

- The mathematical reason is that $P(A|B \cap C_i) \geq P(A|B^c \cap C_i)$ does NOT indeed imply $P(A|B) \geq P(A|B^c)$ as

  $$\begin{aligned} P(A|B) &= \sum_{i=1}^{n} P(A|B \cap C_i)\, P(C_i|B), \\ P(A|B^c) &= \sum_{i=1}^{n} P(A|B^c \cap C_i)\, P(C_i|B^c), \end{aligned}$$

  It highly depends on the proba of applying to department $i$ given you are a woman/man .
- The reason is that women tended to apply to competitive departments with low rates of admission, whereas men tended to apply to less-competitive departments with high rates of admission.

# Proof of expression for conditional proba

- This is very similar to proof of $P\left(A\right)=\sum_{i=1}^{n}P\left(\left.A\right|C_{i}\right)P\left(C_{i}\right).$
- We have

$$P\left(\left.A\right|B\right)=P\left(\left.A\cap\left(\cup_{i=1}^{n}C_{i}\right)\right|B\right)$$

and $A\cap\left(\cup_{i=1}^{n}C_{i}\right)=\cup_{i=1}^{n}\left(A\cap C_{i}\right)$ where $\left(A\cap C_{i}\right)\cap\left(A\cap C_{j}\right)=\varnothing$ for $i\neq j$ so

$$P\left(\left.A\right|B\right)=\sum_{i=1}^{n}P\left(\left.A\cap C_{i}\right|B\right)$$

where by the definition of conditional proba

$$\begin{aligned}P\left(\left.A\cap C_{i}\right|B\right) &= \frac{P\left(A\cap C_{i}\cap B\right)}{P\left(B\right)}\\ &= \frac{P\left(A\cap C_{i}\cap B\right)}{P\left(C_{i}\cap B\right)}\frac{P\left(C_{i}\cap B\right)}{P\left(B\right)}\\ &= P\left(\left.A\right|B\cap C_{i}\right)P\left(\left.C_{i}\right|B\right).\end{aligned}$$

# Simpson's Paradox: A "Smoking" Example

- In 1972-1994 a survey of the electoral roll, largely concerned with smoking habits and survival rates was carried out in Wichkham, a mixed urban and rural district near Newcastle upon Tyne, in the UK. Twenty years later, a follow-up study was conducted.
- Relationship between smoking habits and 20-year survival in 401 women aged 55-74.

| | Smoker | Non-Smoker |
|---|---|---|
| Dead | 80 | 141 |
| Alive | 71 | 109 |

- Introducing $A=$"die", $B=$ "smoking", we have

$$\begin{aligned} P(A|B) &= \frac{80}{80+71} = 0.53, \\ P(A|B^c) &= \frac{141}{141+109} = 0.56. \end{aligned}$$

- Does smoking really help living longer?

# Simpson's Paradox: A "Smoking" Example

- Now lets look more precisely at the data. We have an additional variable related to the age.
- We have

| Age 55-64 | Smoker | Non-Smoker | | Age 65-74 | Smoker | Non-Smoke |
|---|---|---|---|---|---|---|
| Dead | 51=44% | 40=33% | \| | Dead | 29=80% | 101=78% |
| Alive | 64=56% | 81=67% | | Alive | 7=20% | 28=22% |

- For both age class, the survival rate is smaller for the non-smokers. Mathematically, this means that if we introduce $C = \{\text{age between 55-64}\}$ and then $C^c = \{\text{age between 65-74}\}$ then we have as expected

$$\begin{aligned} P\left(A|\, B \cap C\right) &\geq P\left(A|\, B^c \cap C\right), \\ P\left(A|\, B \cap C^c\right) &\geq P\left(A|\, B^c \cap C^c\right) \end{aligned}$$

but still we have

$$P\left(A|\, B\right) \leq P\left(A|\, B^c\right).$$

# Simpson's Paradox: A "Smoking" Example

- This should not be a surprise as

$$\begin{aligned} P(A|B) &= P(A|B\cap C)\,P(C|B) + P(A|B\cap C^c)\,P(C^c|B), \\ P(A|B^c) &= P(A|B^c\cap C)\,P(C|B^c) + P(A|C^c\cap B^c)\,P(C^c|B^c) \end{aligned}$$

  so the results depend on $P(C|B)$, $P(C^c|B)$,$P(C|B^c)$, $P(C^c|B^c)$ i.e. that is the proba of being in a given age given smoking or not!

- The "paradox" occurs as a higher proportion of non-smokers studied belong to older age groups, in which survival rates for both smokers and non-smokers are significantly lower compared to younger age groups.
- In other words, most of the smokers have died off before reaching the older age classes and so the higher number of deaths (in absolute numbers) for the non-smokers in the older age classes has obscured the result.

# Random Variables

- In many scenarios, we are interested in a function of the outcome as opposed to the actual outcome; e.g. we are interested in the sum of two dice and not in the separate values of each die or, when we flip a coin, we want to know the number of tails.
- Real-valued functions defined on the sample space are *random variables*; e.g. number of gold medals by Canadian athletes at Olympics, your score at the SAT test etc.
- **Example**: We are tossing 3 times a fair coin. If we call $X$ the number of heads obtained, then it is a random variables such that

$$\begin{aligned}
P\left\{X=0\right\} &= P\left(\left\{T,T,T\right\}\right) = \left(\tfrac{1}{2}\right)^3 = \tfrac{1}{8},\\
P\left\{X=1\right\} &= P\left(\left\{T,T,H\right\},\left\{T,H,T\right\},\left\{H,T,T\right\}\right) = \tfrac{3}{8},\\
P\left\{X=2\right\} &= P\left(\left\{T,H,H\right\},\left\{H,T,H\right\},\left\{H,H,T\right\}\right) = \tfrac{3}{8},\\
P\left\{X=3\right\} &= P\left(\left\{H,H,H\right\}\right) = \tfrac{1}{8}.
\end{aligned}$$

# Examples

- **Coin Toss**: For a coin toss, the possible events are heads or tails. The number of heads appearing in one fair coin toss can be described using the following random variable:

$$X = \begin{cases} 0 & \text{if head} \\ 1 & \text{if tail} \end{cases} \quad \text{and } P\left(X = 0\right) = P\left(X = 1\right) = \frac{1}{2}$$

  Remark: Note that we could define completely arbitrary random variable such as $X = 10$ if heads and $X = \pi$ is tail. This is of no "practical" interest but possible conceptually.

- **Rolling a fair die**: A random variable can also be used to describe the process of rolling a fair dice and the possible outcomes. The most obvious representation is to take the set $\{1, 2, 3, 4, 5, 6\}$ as the sample space, defining the random variable $X$ as the number rolled. In this case,

$$X = i \text{ if a } i \text{ is rolled and } P\left(X = i\right) = \frac{1}{6} \text{ for } i = 1, ..., 6.$$

# Example

- Assume you have trials consisting of sitting an exam. You have proba $p$ of passing the exam. If you fail, you sit the exam again. These trials are considered independent as you never bother revising the material. We denote $X$ the number of exams you need to sit before passing the subject.
- We have

$$\begin{aligned} P(X=1) &= P(Pass) = p, \\ P(X=2) &= P(Fail, Pass\}) = (1-p)\,p, \\ &\quad .... \\ P(X=n) &= P(\underbrace{Fail, ..., Fail}_{n-1 \text{ times}}, Pass) = (1-p)^{n-1}\,p. \end{aligned}$$

- You can check that

$$\sum_{k=1}^{\infty} P(X=k) = \sum_{k=1}^{\infty} (1-p)^{k-1}\,p = 1.$$

# Example: Roulette Wheel

- The pockets of the roulette wheel are numbered from 1 to 36. In number ranges from 1 to 10 and 19 to 28, odd numbers are red and even are black. In ranges from 11 to 18 and 29 to 36, odd numbers are black and even are red. There is two green pockets numbered 0 and 00.
- *Red/Black*: If you bet 1$ on red, the payout is 1$ if it is red. Let $X_1$ denotes the payout

  $$\begin{aligned} P(X_1 = 1) &= P(\text{red}) = \tfrac{18}{18+18+2} = 0.4737 \\ P(X_1 = -1) &= 1 - P(X = 1) = 0.5263 \end{aligned}$$

  as red$=\{1, 5, 7, 9, 12, 14, 16, 18, 19, 21, 23, 25, 27, 30, 32, 34, 36\}$.
- *Straight up*: you bet 1$ on any pocket 0,00,1,2,...,36. The payout is is 35$ if you are right. Let $X_2$ denotes the payout

  $$\begin{aligned} P(X_2 = 35) &= P(\text{outcome is the pocket bet on}) = \frac{1}{38}, \\ P(X_2 = -1) &= P(\text{outcome is not on the pocket bet on}) = \frac{37}{38}. \end{aligned}$$

# Example: Roulette Wheel

- *Split bet*: you bet 1$ on 11 or 14. The payout is 17$ if you are right. Let $X_3$ denotes the payout

$$\begin{aligned} P\left(X_3 = 17\right) &= P\left(\text{outcome is 11 or 14}\right) = \frac{2}{38}, \\ P\left(X_3 = -1\right) &= P\left(\text{outcome is neither 11 nor 14}\right) = \frac{36}{38}. \end{aligned}$$

- *Street bet*: you 1$ on 1,2,3 or 4,5,6 or... or 19,20,21. Say you bet on 19,20,21. The payout is 11$ if you are right. Let $X_4$ denotes the payout

$$\begin{aligned} P\left(X_4 = 11\right) &= P\left(\text{outcome is 19, 20 or 21}\right) = \frac{3}{38}, \\ P\left(X_4 = -1\right) &= P\left(\text{outcome is not 19, 20 or 21}\right) = \frac{35}{38}. \end{aligned}$$

# Discrete Random Variables

- A random variable (r.v.) $X$ which can take at most a countable number of possible values is said to be *discrete*.
- The *probability mass function* of $X$ is denoted by

$$p\left(a\right)=P\left(X=a\right).$$

- If $X$ can take the values $\left\{x_1, x_2, ...\right\}$ then

$$p\left(x_i\right)\geq 0 \text{ and } \sum_{i=1}^{\infty} p\left(x_i\right)=1$$

- *Cumulative distribution function* is

$$F\left(a\right)=P\left(X\leq a\right)=\sum_{x:x\leq a} p\left(x\right).$$

# Example: Poisson distribution

- Assume $X$ is a discrete r.v. taking values $\{0, 1, 2, ...\}$ where $p(i) = C\lambda^i / i!$ with $\lambda > 0$. 1) Find the expression of $C$, 2) Compute $P(X = 0)$ and $P(X \geq 2)$
- We have

$$\sum_{i=0}^{\infty} p(i) = C \sum_{i=1}^{\infty} \lambda^i / i! = 1$$

but $e^{\lambda} = \sum_{i=1}^{\infty} \lambda^i / i!$ so

$$C = e^{-\lambda}.$$

- It follows that $P(X = 0) = e^{-\lambda} \frac{\lambda^0}{0!} = e^{-\lambda}$ and

$$\begin{aligned} P(X \geq 2) &= 1 - P(X < 2) = 1 - P(X = 0) - P(X = 1) \\ &= 1 - e^{-\lambda} - \lambda e^{-\lambda}. \end{aligned}$$