

36-303: Sampling, Surveys and Society

Statistics of Sampling I
Brian Junker
132E Baker Hall
brian@stat.cmu.edu

28 January 2010

1

Handouts

- Appendix B of Lohr (review of probability)
- Lecture Notes
- HW 01

28 January 2010

2

Outline / Announcements

- TA's Office Hours
- Topics from Project Proposals (Team Project Part I.1)
- Team Project Part I.2 Due Next Tues
 - I will email detailed feedback tonight or tomorrow
 - Revise your proposals and add target pop, frame, mode of data collection
- Statistics of Surveys
 - Part I of an occasional series in the class!
 - Partial review of basic tools
 - Examples related to surveys
 - Foreshadowing: Survey Statistics is Different!
- HW01 Due Next Thurs
- Review: Lohr's Appendix B

28 January 2010

3

TA Office Hours

- Jionglin Wu
FMS 328B
8-1889
jionglin@stat.cmu.edu
- Office Hours:
 - Monday: 2:30-3:30pm
 - Wednesday: 10-11am
- Let me know if additional hours are needed

28 January 2010

4

Project Proposal Ideas

- Post-graduation Plans of Undergraduate Seniors
- Restructuring Pittsburgh: Feelings on moving the border lines between Shadyside and East Liberty, and the "East Side" area.
- How satisfied are our students with CMU dining options (1 of 4!)
- How satisfied are university students with Pittsburgh bus service.

28 January 2010

5

Project Proposal Ideas

- Does using marijuana prevent an individual from doing well in life?
- Do You Believe In The Influence Of Intangible Forces?
- How do Carnegie Mellon University undergraduates perceive Greek Life on campus.
- How do undergraduates at the University of Pittsburgh perceive Carnegie Mellon University undergraduate students.

28 January 2010

6

Project Proposal Ideas

- CMU Perceptions of Academic Integrity & Old Course Material Archives
- Bar Workers' Perspective on Law Requiring Establishments to Close at 2AM
- How well do Carnegie Mellon University Freshmen utilize their meal blocks? (2 of 4!)

28 January 2010

7

Project Proposal Ideas

- Interaction with the Disabled and Its Effects on Self-Esteem (Off Campus)
- Attitudes of non-college denizens toward living in an apartment building with college students
- Best Pizza Place in Pittsburgh (Off Campus) (3 of 4!)
- Attitude towards the new Dining System (On Campus) (4 of 4!)

28 January 2010

8

Project Proposal Ideas

- Which are the most important innovation characteristics and skill that university students need to develop? And which are the most powerful educational elements that promote the innovation skills?
- Graphical versus Numerical Presentation of Quantitative Environmental Risk Information about Unexploded Ordnance (UXO)
- What portion of CMU students know about and/or utilize safe walk and escort services and how could we increase both knowledge and utilization of these services?
- How/why do 28X consumers differ during peak and nonpeak travel times and how does the increase in 28X usage during peak times affect the consumers?

28 January 2010

9

Team Project Part I.2 Due Next Tues

- The projects should to be interesting enough to make an impact (what can someone do about it?)
- I will email detailed feedback tonight or tomorrow
- Revise your proposals and add target pop, frame, mode of data collection; I will ask for the following sections:
 - ☐ Interesting research question? Doable? Previous research/results on this question?
 - ☐ What population do you want to make inferences about?
 - ☐ What population will you sample? Is this different from (b)?
 - ☐ How do you plan to carry out the survey & why?
 - ☐ What variables will you measure?
 - ☐ Sample questions for off-campus project?

28 January 2010

10

Statistics of Surveys

- Survey Statistics is different from other kinds of Statistics
 - ☐ Sampling from a finite population is *different*
 - ☐ Design features (stratification, clustering, weights) increase information at the cost of more complex analysis
- We will get there, in occasional smallish steps
 - ☐ Today:
 - Partial Review of Probability Tools
 - Application: Sample Size Calculations
 - Application: Randomized Response
 - ☐ Future:
 - Urn models
 - What is random about finite population sampling?
 - Accounting for complex survey designs

28 January 2010

11

Partial Review of Probability Tools

- Discrete Random Variables
- Expected Value, Mean, Variance
- More than One Random Variable
 - ☐ Covariances, Independence, Linear Combinations, Normal Approximation (CLT)
 - ☐ *Application: Sample Size Calculations*
- Conditioning
 - ☐ Conditional Probability, Conditional Distribution, Conditional Expectation
 - ☐ *Application: Randomized Response*

28 January 2010

12

Discrete Random Variable

- A **discrete** random variable X has a sample space that you can "count" (1, 2, 3, ...)
- Toss a die, let X be the side that comes "up"
- Toss a coin until "heads" comes up, let X be the number of "Tails" until first "Heads"
- Spin a spinner, let X be the exact angle in degrees at which the spinner comes to rest.
- A **continuous** random variable X has a sample space that includes a continuous interval (so there are uncountably many outcomes)
 - Which of the above X 's is discrete, which is continuous?

28 January 2010

13

Discrete Random Variable

- For us, X usually has a **finite sample space**
 - X can take on only the values $x_1, x_2, \dots, x_K$, with probability $p_1, p_2, \dots, p_K$
- Examples:
 - Biased coin, $X=1$ for "Heads", 0 for "Tails"
 - (this is a _____ random variable!)
 - $P[X=1] = p, P[X=0] = 1-p$
 - Flip a coin n times, let X be the number of "Heads"
 - (this is a _____ random variable!)
 - $P[X=k] = \dots, k=0, 1, 2, \dots, n$
 - Consider a population of 1,000 adults, and let x_k be each adult's annual income, $k=1, \dots, 1000$. Pick one adult at random and let X be that person's income.
 - $P[X=x_k] = \dots, k=1, 2, \dots, 1000$

28 January 2010

14

Expected Value, Mean, Variance

- Let X be a discrete random variable taking on the values $x_1, \dots, x_K$ with probabilities $p_1, \dots, p_K$:

- The probabilities *must* add to 1:

$$\sum_{i=1}^K p_i = 1,$$

- The **mean** of X is defined to be

$$\mu_X = E[X] = \sum_{i=1}^K x_i P(X = x_i) = \sum_{i=1}^K x_i p_i$$

- The **variance** of X is defined to be

$$\sigma_X^2 = \text{Var}[X] = E[(X - \mu_X)^2] = \sum_{i=1}^K (x_i - E[X])^2 P(X = x_i) = \sum_{i=1}^K (x_i - \mu_X)^2 p_i$$

- More generally, for any function $g(x)$, the **expected value** of $g(X)$ is

$$E[g(X)] = \sum_{i=1}^K g(x_i) P(X = x_i).$$

28 January 2010

15

Expected Value Example

- Let X be a Bernoulli random variable, $P[X=1]=.2$, and suppose I pay you \$50 if $X=1$ and you pay me \$10 if $X=0$. What is the expected value of your income?

$g(x) = 50$ if $x = 1$, and $g(x) = -10$ if $x = 0$.

$$E[g(X)] = 50 \times p - 10 \times (1 - p)$$

$$= 50(0.2) - 10(0.8)$$

$$= 2$$

$$\text{Var}(g(X)) = (50 - 2)^2(0.2) + (-10 - 2)^2(0.8)$$

$$= 2304(0.2) + 144(0.8)$$

$$= 576$$

$$SD(g(X)) = \sqrt{576} = 24$$

28 January 2010

16

More Than One Random Variable

x	y	xy	$P[X=x, Y=y]$
1	2	2	$\frac{1}{4}$
2	8	16	$\frac{1}{4}$
4	8	32	$\frac{1}{4}$
3	6	18	$\frac{1}{4}$

Note that

$$E[X]E[Y] = (6)(2.5) = 15 \neq 17 = E[XY]$$

thus X and Y cannot be independent.

$$E[X] = \frac{1}{4}(1 + 2 + 4 + 3) = 2.5$$

$$E[Y] = \frac{1}{4}(2 + 8 + 8 + 6) = 6$$

$$E[XY] = \frac{1}{4}(2 + 16 + 32 + 18) = 17$$

More generally X and Y are **independent** if and only if

$$P[X=x, Y=y] = P[X=x]P[Y=y]$$

for all x and y .

28 January 2010

17

Covariance & Independence

- Recall that $\text{Var}(X) = E[(X - \mu_X)^2]$
- Similarly, $\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$

$$\text{Cov}(X, Y) = \frac{1}{4} \left\{ (1 - 2.5)(2 - 6) + (2 - 2.5)(8 - 6) + (3 - 2.5)(6 - 6) + (4 - 2.5)(8 - 6) \right\}$$

$$= 2$$

- If X and Y are independent, $\text{Cov}(X, Y) = 0$

$$\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

$$= E[(X - \mu_X)]E[(Y - \mu_Y)] = 0 \cdot 0 = 0$$

28 January 2010

18

Linear Combinations

- **Exercise:** Use the definitions so far to show

$$E[aX + bY + c] = aE[X] + bE[Y] + c$$

- **Exercise:** Use this fact to show that for any set of random variables $X_1, X_2, \dots, X_n$ that all have the same mean μ ,

$$E[\bar{X}] = E\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \mu$$

(Definition of $\bar{X}$) (This is the part to show!)

28 January 2010

19

Mean and Variance of Sample Average

- Let $X_1, \dots, X_n$ all have the same mean μ , and let

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

- We know $E[\bar{X}] = \mu$, what about $\text{Var}(\bar{X})$?
 - Use the definitions to show:

$$\text{Var}(aX + bY + c) = a^2\text{Var}(X) + 2ab\text{Cov}(X, Y) + b^2\text{Var}(Y)$$

We use this on the next page to work out $\text{Var}(\bar{X})$.

28 January 2010

20

Mean and Variance of Sample Average

From

$$\text{Var}(aX + bY + c) = a^2\text{Var}(X) + 2ab\text{Cov}(X, Y) + b^2\text{Var}(Y)$$

we can calculate

$$\text{Var}\left[\frac{1}{n}(X_1 + X_2)\right] = \frac{1}{n^2} \left(\text{Var}(X_1) + 2\text{Cov}(X_1, X_2) + \text{Var}(X_2) \right)$$

and applying this to n terms instead of 2 terms (induction!), we get the following mess

$$\text{Var}\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n^2} \left\{ \sum_{i=1}^n \text{Var}(X_i) + 2 \sum_{i=1}^n \sum_{j=1}^{i-1} \text{Cov}(X_i, X_j) \right\}$$

We now assume $X_1, X_2, \dots, X_n$ have the same mean μ , the same variance σ^2 , and covariance $\text{Cov}(X_i, X_j) = 0$ whenever $i \neq j$. Then the "mess" reduces to the more familiar:

$$\text{Var}(\bar{X}) = \frac{1}{n^2} \left\{ n\sigma^2 + 2 \cdot \binom{n}{2} \cdot 0 \right\} = \frac{1}{n} \sigma^2$$

28 January 2010

21

Central Limit Theorem

- We have shown: If $X_1, \dots, X_n$ are independent, identically distributed (iid) with $E[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2$, then

$$E[\bar{X}] = \mu, \quad \text{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

- The Central Limit Theorem then tells us

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

- σ is the SD of X_i ; $\sigma/\sqrt{n}$ is the SE of $\bar{X}$

28 January 2010

22

Application: Sample Size Calculation

- Let $X_1, \dots, X_n$ be an iid sample of people's heights, with a common mean $\mu = 5.75$ ft and SD $\sigma = 0.5$ ft.
- Then $E[\bar{X}] = 5.75$, with SE $0.5/\sqrt{n}$
- CLT: Approx 95% confidence interval for μ : $(\bar{X} - (1.96)(0.5)/\sqrt{n}, \bar{X} + (1.96)(0.5)/\sqrt{n})$
- How large n to have 95% confidence that $\bar{X}$ is within 0.1 of μ ?
 - Roughly, need $0.1 > 1/\sqrt{n}$ or $n > 100$.

28 January 2010

23

Foreshadowing: Survey Statistics is Different!

- In real Survey Sampling work, $\text{Cov}(X_i, X_j)$ is usually not zero!
- Hence

$$E[\bar{X}] = \mu$$

but

$$\text{Var}(\bar{X}) \neq \sigma^2/n$$

- The CLT is not quite true, as stated, either!
- But the basic CLT calculation is often a reasonable "crude guess"...

28 January 2010

24

Conditioning

- The conditional probability of event A, given event B, is

$$P[A|B] = \frac{P[A \cap B]}{P[B]}$$

It is often useful to write this as a formula for $P[A \cap B]$:

$$P[A \cap B] = P[A|B]P[B]$$

- The conditional distribution of X given $Y = y$ is

$$P[X = x|Y = y] = \frac{P[X = x, Y = y]}{P[Y = y]} \quad [\text{comma means "and"!}]$$

- The conditional expected value of X given $Y = y$ is the expected value with respect to the conditional distribution:

$$E[X|Y = y] = \sum_x xP[X = x|Y = y]$$

28 January 2010

25

Conditioning

x	y	xy	$P[X = x, Y = y]$
1	2	2	$\frac{1}{4}$
2	8	16	$\frac{1}{4}$
4	8	32	$\frac{1}{4}$
3	6	18	$\frac{1}{4}$

$$P[X = 2|Y = 8] = \frac{P[X = 2, Y = 8]}{P[Y = 8]} = \frac{1/4}{1/2} = \frac{1}{2}$$

$$P[X = 4|Y = 8] = \dots = \frac{1}{2}$$

$$E[X] = 2.5$$

$$Var(X) = \frac{1}{4}[(1 - 2.5)^2 + (2 - 2.5)^2 + (3 - 2.5)^2 + (4 - 2.5)^2] = 1.25$$

$$E[X|Y = 8] = \frac{1}{2}(2 + 4) = 3$$

$$Var(X|Y = 8) = \frac{1}{2}[(2 - 3)^2 + (4 - 3)^2] = 1$$

Exercise: Show that if X and Y are independent, then $E[X|Y = y] = E[X]$, for any y.

28 January 2010

26

Application: Randomized Response

- "Flip a coin, but don't tell me whether it's heads or tails.
 - "If heads, answer truthfully: have you ever cheated in a CMU class?
 - "If tails, answer truthfully: is the last digit of your SSN odd?"
- Let $p = P[\text{Heads}]$, $\pi = P[\text{Cheat}]$, $\lambda = P[\text{Yes}]$. Then

$$\begin{aligned} \lambda &= P[\text{Yes} \cap \text{Heads}] + P[\text{Yes} \cap \text{Tails}] \\ &= P[\text{Yes}|\text{Heads}]P[\text{Heads}] + P[\text{Yes}|\text{Tails}]P[\text{Tails}] \\ &= \pi \cdot p + (1/2) \cdot (1 - p) \end{aligned}$$

Therefore

$$\pi = \frac{\lambda - (1/2) \cdot (1 - p)}{p}$$

28 January 2010

27

Application: Randomized Response

$$\pi = \frac{\lambda - \frac{1}{2}(1 - p)}{p}$$

Suppose the coin is fair ($p = \frac{1}{2}$) and in our survey we get a fraction $\hat{\lambda}$ of people answering "yes". Then

$$\begin{aligned} \hat{\pi} &= 2(\hat{\lambda} - 1/4) \\ E[\hat{\pi}] &= 2(E[\hat{\lambda}] - 1/4) \\ &= 2(\lambda - 1/4) = \pi \quad (\text{Exercise!}) \end{aligned}$$

So $\hat{\pi}$ is an unbiased estimator of π , and

$$\begin{aligned} Var(\hat{\pi}) &= Var[2(\hat{\lambda} - 1/4)] \\ &= 4Var(\hat{\lambda}) \end{aligned}$$

so $Var(\hat{\pi})$ is inflated, relative to $Var(\hat{\lambda})$: $\hat{\pi}$ is statistically inefficient.

Exercise: The closer $p = P[\text{Answer Cheating Question}]$ is to 1, the closer $Var(\hat{\pi})$ is to $Var(\hat{\lambda})$.

28 January 2010

28

Foreshadowing: Survey Statistics is Different!

- In a regular statistics course we would go on to say $\hat{\lambda} = Y/n$ where Y is the number of "Yes"s among a sample of size n.

Therefore

- $E[\hat{\lambda}] = \lambda$, the true proportion of "Yes"s.
- $SE(\hat{\lambda}) = \sqrt{\lambda(1 - \lambda)/n}$, because Y is a binomial random variable.

- In Survey Sampling

- The expected value part is OK
- The variance will be different; Y is not quite binomial!

28 January 2010

29

Review

- Feedback on Project Proposals
- Statistics of Surveys (Part I of Occasional Series)
- Team Project part I.2 (target pop, frame, mode of data collection) Due Next Tuesday
- HW1 Due Next Thursday
- Read Lohr Appx B (handout today)

28 January 2010

30

Review

- TA's Office Hours
- Topics from Project Proposals (Project Part I.1)
- Team Project Part I.2 Due Next Tues
 - I will email detailed feedback on I.1 tonight or tomorrow
 - Revise your proposals and add target pop, frame, mode of data collection
- Statistics of Surveys (Part I of an occasional series)
- HW01 Due Next Thurs
- Review: Lohr's Appendix B