36-303: Sampling, Surveys and Society

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

Handouts

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

Outline / Announcements

- Project Proposals (Team Project Part I.1)
- Team Project Part I.2 Due Next Tues
  - I will email detailed feedback tonight or tomorrow
  - Revise A,B,C, and add D,E,F,G for *each* project proposal you made
- 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!
- Review: Lohr's Appendix B

Project Proposals

- **I looked at them all**; I will email feedback to each team later today.
- **Grades** (50/project x 2 projects = 100 pts)
  - A: Is this interesting? 20 pts
  - B: General questions/research questions 20 pts
  - C: One article with description from each team member 10 pts
- **Revise everything** – especially the parts where you got less than full credit!
- *Each team proposed at least one doable project!*
  - The project we decide for your team may or may not be the high-scoring project! Depends on *feasibility*, *my interest*, etc.

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
- For *each* project you proposed:
  - **Revise A, B, C**: Interesting topic? General research questions? Articles about past research in the area?
  - **Add D, E, F, G**: Target population? Sampling Frame? Mode of Data Collection? Major Variables?

Pointers for I.2

- E. Target population – What are the individual units that give you information?
  - students? buses? faculty members? times of day? locations? events ("the bus is late" or "10 students walked by", etc.)
- D. Sampling Frame – In most (but not all) cases there will be a real or hypothetical list of units that you could sample from. E.g.:
  - Numbers in the phone book (which one? or maybe random digit dialling? which exchanges? etc)
  - Email addresses in C-Book

  In some cases there will be no natural sampling frame. E.g.:
  - Interview people as they pass by the fence
  - Wait for instances of late buses

  In these cases give a very specific description of what kinds of units you will be looking for, and how you will find them.

Pointers for I.2

- F. Mode of Data Collection – How will you get the data?
  - Invite people to website with online SAQ, using email, postcards, etc.
  - Approach people on the street/sidewalk/etc. and use P&P SAQ, CAPI, etc.
  - Go to a certain intersection at a certain time and observe buses, people, accidents, or other events of interest.
  - Go to a school and interview some/all students

  Give a sense of how many intersections, times, schools, students, etc. might be needed to "represent" the population.
- G. Variables to Measure – List (and define) two to five variables that you must measure to have a successful survey.

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

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*

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?*

Discrete Random Variable

- For us, *X* usually has a ***finite sample space***
  - *X* can take on only the values $x_1, x_2, \ldots, x_K$, with probability $p_1, p_2, \ldots, p_K$
- 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] =* ________________*, k=0, 1, 2, …, n*
  - Consider a population of 1,000 adults, and let x_k be each adult's annual income, *k=1, …, 1000.* Pick one adult at random and let X be that person's income.
    - *P[X=x_k] =* ________________*, k=1, 2, …, 1000*

Expected Value, Mean, Variance

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

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

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

$$E[g(X)] = \sum_{x} g(x) P(X = x).$$

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 $x = 0$.

$$\begin{aligned} E[g(X)] &= 50 \times p - 10 \times (1-p) \\ &= 50(0.2) - 10(0.8) \\ &= 2 \\ Var(g(X)) &= (50-2)^2(0.2) + (-10-2)^2(0.8) \\ &= 2304(0.2) + 144(0.8) \\ &= 576 \\ SD(g(X)) &= \sqrt{576} = 24 \end{aligned}$$

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.

$$\begin{aligned} E[X] &= \frac{1}{4}(1+2+4+3) = 2.5 \\ E[Y] &= \frac{1}{4}(2+8+8+6) = 6 \\ E[XY] &= \frac{1}{4}(2+16+32+18) = 17 \end{aligned}$$

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.

Covariance & Independence

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

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

- If X and Y are independent, Cov(X,Y) = 0

$$\begin{aligned} Cov(X,Y) &= E[(X-\mu_X)(Y-\mu_Y)] \\ &= E[(X-\mu_X)]E[(Y-\mu_Y)] = 0 \cdot 0 = 0 \end{aligned}$$

Linear Combinations

- <u>Exercise:</u> Use the definitions so far to show

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

- <u>Exercise:</u> Use this fact to show that for any set of random variables X_1, X_2, … X_n that all have the same mean μ,

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

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

Mean and Variance of Sample Average

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

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

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

    $Var(aX + bY + c) = a^2Var(X) + 2abCov(X,Y) + b^2Var(Y)$

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

Mean and Variance of Sample Average

From

$$Var(aX + bY + c) = a^2Var(X) + 2abCov(X,Y) + b^2Var(Y)$$

we can calulate

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

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

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

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

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

Central Limit Theorem

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

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

- The Central Limit Theorem then tells us

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

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

Application: Sample Size Calculation

- Let $X_1, \ldots, X_n$ be an iid sample of people's heights, with a common mean μ=5.75 ft and SD σ=0.5ft.
- Then $E[\overline{X}]$ = 5.75, with SE $0.5/\sqrt{n}$
- CLT: Approx 95% confidence interval for μ :

$$\left(\overline{X} - (1.96)(0.5)/\sqrt{n}\,,\ \overline{X} + (1.96)(0.5)/\sqrt{n}\right)$$

- How large n to have 95% confidence that $\overline{X}$ is within 0.1 of μ?
  - Roughly, need $0.1 > 1/\sqrt{n}$ or n > 100.

Foreshadowing: Survey Statistics is Different!

- In real *Survey Sampling* work, Cov(X_i,X_j) is usually not zero!
- Hence

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

but

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

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

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:

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

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	$\frac{1}{4}$

$$\begin{aligned} E[X] &= 2.5 \\ Var(X) &= \frac{1}{4}[(1-2.5)^2 + (2-2.5)^2 \\ &\quad + (3-2.5)^2 + (4-2.5)^2] \\ &= 1.25 \end{aligned}$$

$$\begin{aligned} P[X = 2|Y = 8] &= \frac{P[X = 2, Y = 8]}{P[Y = 8]} \\ &= \frac{1/4}{1/2} = \frac{1}{2} \\ P[X = 4|Y = 8] &= \cdots = \frac{1}{2} \end{aligned}$$

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

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

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[Heads], π=P[Cheat], λ=P[Yes]. Then

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

Therefore

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

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

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

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

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!

Review

- Feedback on Project Proposals
- Team Project part I.2 (target pop, frame, mode of data collection) Due Next Tuesday
  - HW02 due next Tues also!
- Statistics of Surveys (Part I of Occasional Series)
- Read Lohr Appx B (handout today)