Unit 3: Foundations for inference
Lecture 4: Review / Synthesis

Statistics 101

Mine Çetinkaya-Rundel

February 19, 2013

Announcements

- MT on Thursday: cheat sheet, calculator, tables will be provided

Requested topics

- Sampling strategies: random / stratified / cluster
- Conditional probability: Probability trees / Bayes' theorem / Bayesian inference / checking for independence
- Binomial distribution
- HT and CI: One and two sided alternatives / agreement of HT and CI
- Randomization testing
- *Power / Type 1 and Type 2 errors

What is the difference between stratified and cluster sampling? Why might we choose either of these methods over a simple random sample?

Testing for AIDS – with counts

Suppose that the proportion of people infected with AIDS in a large population is 0.01. If AIDS is present, a certain medical test is positive with probability 0.997 (called the sensitivity of the test). If AIDS is not present, the test is negative with probability 0.985 (called the specificity of the test). If a person tests positive, what is the probability that they have AIDS?

- Let's assume there are 1 million individuals in this population.
- How many are expected to have AIDS, and how many are not expected to have AIDS?
  - Have AIDS: $1,000,000 \times 0.01 = 10,000$
  - Don't have AIDS: $1,000,000 \times 0.99 = 990,000$

From *http://www.pitt.edu/~nancyp/stat-1000-s07/week6.pdf* .

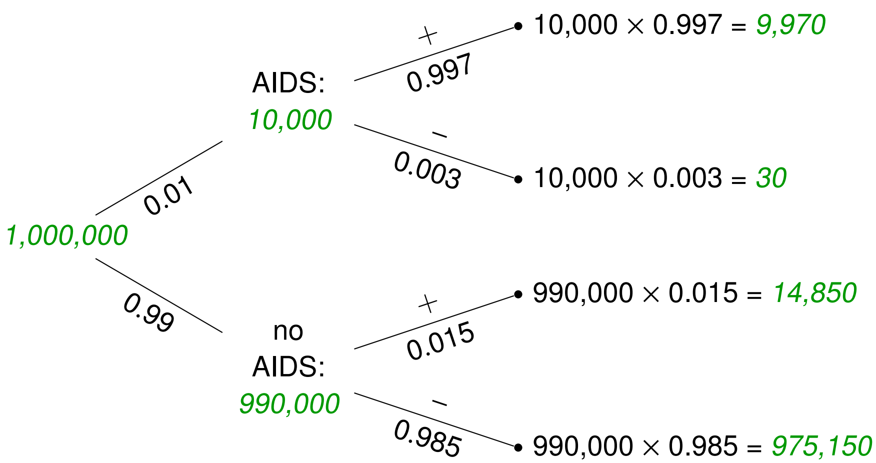
Testing for AIDS – with counts (cont.)

Clicker question

How many of the people with AIDS would we expect to test positive?

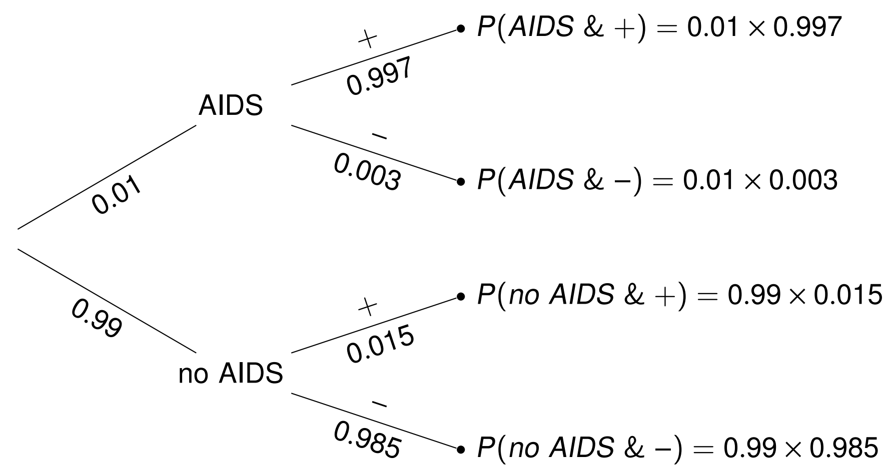
(a) 30
(b) 9,850
(c) 9,970
(d) 987,030
(e) 997,000

Testing for AIDS – with counts (cont.)



$$P(AIDS|+) = \frac{9,970}{9,970 + 14,850} \approx 0.40$$

Testing for AIDS – with probabilities



$$P(AIDS|+) = \frac{0.01 \times 0.997}{0.01 \times 0.997 + 0.99 \times 0.015} \approx 0.40$$

Testing for AIDS – in a Bayesian framework

- In the first stage of testing:
  - Prior: P(AIDS)
    = P(person has AIDS before we collect any data on them) = 0.01
  - Posterior: P(AIDS | test +)
    = P(person has AIDS *given* that they tested positive) = 0.40
- In the second stage of testing:
  - Prior = Posterior from the previous test = 0.40

If the person tests positive for AIDS in the first test, will the prior probability be higher or lower than 1% (prior in the first test)? Why?

Testing for AIDS – independence

Clicker question

Which of the following probabilities should be calculated using the Binomial distribution?

Probability that

(a) a basketball player misses 3 times in 5 shots

(b) train arrives on the time on the third day for the first time

(c) height of a randomly chosen 5 year old is greater than 4 feet

(d) a randomly chosen individual likes chocolate ice cream best

Why Binomial?

Suppose the probability of a miss for this basketball player is 0.40. What is the probability that she misses 3 times in 5 shots?

- One possible scenario is that she misses the first three shots, and makes the last two. The probability of this scenario is:

$$0.4^3 \times 0.6^2 \approx 0.023$$

- But this isn't the only possible scenario:

1. *MMM*HH
2. *MM*H*M*H
3. *M*H*MM*H
4. H*MMM*H
5. H*MM*H*M*
6. H*M*H*MM*
7. HH*MMM*
8. *M*H*M*H*M*
9. *M*HH*MM*
10. *MM*HH*M*

- Each one of these scenarios has 3 *M*s and 2 Hs, therefore the probability of each scenario is 0.023.
- Then, the total probability is $10 \times 0.023 = 0.23$.

... concisely

Suppose the probability of a miss for this basketball player is 0.40. What is the probability that she misses 3 times in 5 shots?

$$\begin{aligned}\binom{5}{3} \times 0.4^3 \times 0.6^2 &= \frac{5!}{3! \times 2!} \times 0.4^3 \times 0.6^2 \\ &= 10 \times 0.023 \\ &= 0.23\end{aligned}$$

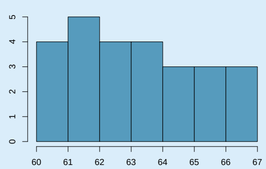
... concisely

Suppose the probability of a miss for this basketball player is 0.40. What is the probability that she misses 3 times in 5 shots?

$$\begin{aligned}\binom{5}{3} \times 0.4^3 \times 0.6^2 &= \frac{5!}{3! \times 2!} \times 0.4^3 \times 0.6^2 \\ &= 10 \times 0.023 \\ &= 0.23\end{aligned}$$

A random sample of 36 female college-aged dancers was obtained and their heights (in inches) were measured. Provided below are some summary statistics and a histogram of the distribution of these dancers' heights. The average height of all college-aged females is 64.5 inches. Do these data provide convincing evidence that the average height of female college-aged dancers is different from this value?

n	36
mean	63.6 inches
sd	2.13 inches

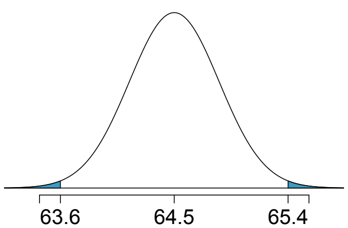


$$H_0 : \mu = 64.5$$
$$H_A : \mu \neq 64.5$$

$$\bar{x} = 63.6, s = 2.13, n = 36, \alpha = 0.05$$

$$\bar{x} \sim N\left(mean = 64.5, SE = \frac{2.13}{\sqrt{36}} = 0.355\right)$$

$$Z = \frac{63.6 - 64.5}{0.355} = -2.54$$



$$p\text{–}value = 2 \times P(Z < -2.54) = 2 \times 0.0055 = 0.011$$

Since p-value < 0.05, reject H_0. The data provide convincing evidence that the average height of female college-aged dancers is different than 64.5 inches.

Clicker question

What is the equivalent confidence level for this two-sided hypothesis test with $\alpha = 0.05$?

(a) 80%
(b) 90%
(c) 95%
(d) 99.7%
(e) 97.5%

Clicker question

If we were to calculate a confidence interval with the equivalent confidence level to this hypothesis test, would this interval include 64.5 (the null value)?

(a) Yes
(b) No
(c) Cannot tell without calculating the interval

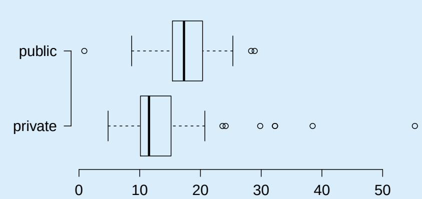
Clicker question

If we were to calculate a confidence interval with the equivalent confidence level to this hypothesis test, would this interval include 64.5 (the null value)?

(a) Yes
(b) No
(c) Cannot tell without calculating the interval

One measure of quality of colleges that people find useful is the student-to-faculty ratio: the number of students enrolled divided by the number of teachers. For schools with small student-to-faculty ratios, we expect class sizes to be small, which is a desired attribute since students can get personalized attention from faculty in smaller classes. The box plots below show the distributions of student-to-faculty ratios in random samples of 57 public and 85 private four-year colleges. Also provided are some sample statistics. Do these data provide convincing evidence that the average (mean) student-to-faculty ratio in public four-year colleges is higher than that of private four-year colleges?

	public	*private*
n	57	85
mean	18	14
sd	4.6	7.3



$$H_0 : \mu_{public} = \mu_{private}$$
$$H_A : \mu_{public} > \mu_{private}$$

We write the student-to-faculty ratio of each public and private college in this sample on a total of ________ index cards. Then, we shuffle these cards and split them into two groups: one group of size ________ representing public colleges, and another group of size ________ representing private colleges. We calculate the difference between the average student-to-faculty ratios in the public and private colleges ($\bar{x}_{public} - \bar{x}_{private}$) and record this value. We repeat this many times to build a randomization distribution, which should be centered at ________. Lastly, we calculate the p-value as the proportion of simulations where the simulated differences in means are ________.