Note Title 10/26/2009

# STT 200 10-26-09

## SUMMARY OF CI INFORMATION

| SETUP | PARAM | POINT ESTIMATOR | CI (ON $\tilde{p}$) | CLAIM |
|---|---|---|---|---|
| WITH REPL $n_x$ (EQUAL PROBABILITY) | POP MEAN $\mu_x$ | $\bar{x}$ (RULE) (MOST SURELY WRONG) | $\bar{x} \pm 1.96 \frac{s_x}{\sqrt{n_x}}$ <br> CI ≡ WIDE BODIED GUESS | |
| " | $p_x$ <br> eg FRACTION OF INDIVIDUALS IN POPULATION HAVING SOME PARTICULAR CHARACTERISTIC | POINT ESTIMATOR <br> $\hat{p}_x$ (or $\bar{p}_x$) <br> SAMPLE PROPORTION <br> $\hat{p}_x = \bar{x}$ | $\bar{x}_n \to \hat{p}_x \pm 1.96 \frac{\sqrt{\hat{p}_x(1-\hat{p}_x)}}{\sqrt{n}}$ | |
| " | | | | |

SCORING
1 PENCIL USER
0 OTHER

NOTE: $\sqrt{\hat{p}_x(1-\hat{p}_x)} = \sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2/n}$

SETUP:

(x, y)

COUPLES.

POP

OR

1 TRANS.

TRANSMISSIONS.

(x = VISCOSITY FLUID

y = TEMP. FLUID).

SETUP IS SPECIAL.

KNOW $\mu_x$.

PARAM $\mu_y$

POINT EST

~~$\bar{y}$ ??~~

REGR-BASED ESTIMATOR OF $\mu_y$

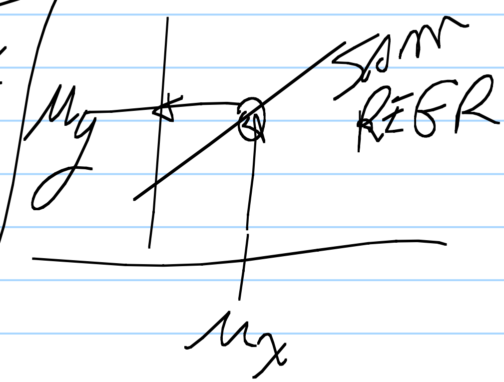
THINKING

$\Rightarrow$ REGR ESTIMATOR OF $\mu_y$ IS $\boxed{\bar{y} + (\mu_x - \bar{x}) R \frac{s_y}{s_x}}$

$$\frac{\text{RISE}}{\text{RUN}} = \frac{y - \bar{y}}{x - \bar{x}} = R \frac{s_y}{s_x}$$

$(x, y)$ ON REGR LINE IF AND ONLY IF

$\mu_x$



SO REGR BASED EST OF $\mu_y$ (CONT)

ESTIMATOR $\quad \bar{y} + (\mu_x - \bar{x}) R \, \sigma_y / \sigma_x$

SAMPLE CORRELATION

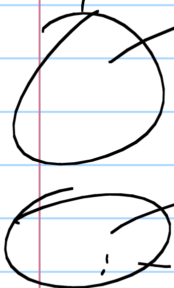
MODIFIES $\bar{y}$ TO TAKE ACCOUNT OF HOW FAR $\bar{x}$ IS FROM KNOWN $\mu_x$.

– THEN ENTITLED TO USE 95% z-BASED CI FOR $\mu_y$

FORM

$$\left(\bar{y} + (\mu_x - \bar{x}) R \, \sigma_y / \sigma_x\right) \pm 1.96 \frac{\sigma_y}{\sqrt{n}} \sqrt{1 - R^2}$$

| DIFF. OF POP MEANS | PARAM | ESTIMATOR | CI |
|---|---|---|---|
| | $\mu_x - \mu_y$ | $\bar{x} - \bar{y}$ | $(\bar{x} - \bar{y}) \pm 1.96\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}$ |



$\mu_x$ $\sigma_x$ $\Rightarrow$

$\mu_y$ $\sigma_y$ $\Rightarrow$

MINUS PLUS.

eg $x$ = HT GAIN ON PLACEBO

$y$ = HT GAIN ON PROTEIN

IF WISH 68% CI $z = 1.0$

| DIFF OF POP PROPORTIONS | $p_x - p_y$ | $\hat{p}_x - \hat{p}_y$ | $\hat{p}_x - \hat{p}_y \pm 1.96\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}}$ |
|---|---|---|---|

NEG POS.

$p_x$ = POP^N FRACTION W/ FEVER DROPPING AFTER PLACEBO.

$p_y$ = POP^N FRACTION W/ MED.

Note Title

STT 200 10-26-09

Assignment due in recitation 10 - 27 - 09.

**1-6. Regression-based CI for $\mu_y$ when we have data $(x_i, y_i)$ for which $\mu_x$ is known.** Each of the students in a with-replacement and equal-probability random sample of 20 students from the class is scored x= exam 1 raw score, y = exam 2 raw score. Suppose we know the class mean $\mu_x = 17.3$ for exam 1 and seek to estimate the class mean raw score on exam 2. The following data are from the random sample of 20 students:

$\bar{x} = 16.8$ (lower than the class exam 1 average)

$\bar{y} = 12.22$ R = sample correlation of (x, y) = 0.64

$s_x = 1.1$ $s_y = 2.73$

We will ignore FPC issues for the present.

1. Point estimate of $\mu_y$ ignoring x data.

2. Regression-based point estimator of $\mu_y$ ("improved" estimator).

3. Usual z-based 95% CI for $\mu_y$ using estimator of (1).

$\bar{y} \pm 1.96\, s_y/\sqrt{n}$

4. 95% z-based CI for $\mu_y$ using estimator of (2).

5. Is the regression estimator (2) raised from (1) and if so why?

YES - BECAUSE $\mu_x - \bar{x} > 0$

6. Is the CI (4) narrower than that of (3) and if so why?

OBS $(x_i, y_i)$

$\mu_x$ = KNOWN 17.3

$\bar{x} = 16.8$

$\bar{y} = 12.22$

$s_x = 1.1$ $s_y = 2.73$

$R = 0.64$

$\approx R^2 = .42$ OF $s_y^2$ IS EXPLAINED BY REGR ON X

POINT ESTR

IMP ESTOR: $\bar{y} + (\mu_x - \bar{x}) \dfrac{R s_y}{s_x} = 12.22$

(*) $12.22 + (17.3 - 16.8) \cdot .64 \dfrac{2.73}{1.11}$

95% CI (*) $\pm 1.96\, s_y/\sqrt{n}\, \sqrt{1 - R^2}$

6/2009

**7-10. z-based CI for $\mu_y$ - $\mu_y$ when independently sampling each of populations x, y with equal-probability and with-replacement. Illustrated by sampling each of exam 1 score and exam 2 scores separately.** Suppose we do not know any of the population parameters and

$\mu_x$

$\overline{x} = 17.4$
$\overline{y} = 14.8$
$n_x = 30$
$n_y = 40$
$s_x = 1.9$
$s_y = 3.8$

EST $\mu_x - \mu_y$

#7. $(14.8 - 17.4) \pm 1.96 \sqrt{\frac{1.9^2}{30} + \frac{3.8^2}{40}}$

TWO SAMPLE PROBLEM

Remember, the samples from x are entirely unrelated to ("independent" of) those from y. Unlike the setting of problems 1-7 where we selected students and took their exam 1 and exam 2 scores (paired data) we now sample potentially different students from each of exam 1 scores and exam 2 scores (unpaired data).

7. Give a 95% z-based CI for $\mu_y$ - $\mu_y$ based on the above data.

8. Now that we have estimates of $s_x = 1.9$ and $s_y = 3.8$ what fraction of our 70 observations should we have allocated to population x in order to make the resulting CI narrowest?

**9-12. z-based CI for $\mu_y$ - $\mu_y$ when independently sampling each of populations x, y with equal-probability and with-replacement. Illustrated by an example discussed in lecture of 10-21-09.**

9. Toss a coin until you see the pattern HH. For example, the sequence TTHH-HTTTT first finds the HH pattern after 4 tosses. Repeat the experiment 20 times, each time recording the score x = number of tosses required to get HH. Calculate sample mean $\overline{x}$ and sample standard deviation $s_x$.

IF WE MINIMIZE $\sqrt{\sigma_x^2/n_x + \sigma_y^2/n_y}$
FOR $n_x + n_y$ FIXED

OPT'L ALLOCATION $f_x = \frac{\sigma_x}{\sigma_x + \sigma_y}$

Twenty rows each of 22 coin tosses.

```
MatrixForm[Table[Flatten[Table[c[[Random[Integer, 1] + 1]], {i, 1, 22}]], {j, 1, 20}]]
```

$x_1 = 10 \rightarrow$

$x_2 = 12 \rightarrow$

| | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| H | T | T | T | T | H | T | T | H | H | T | H | H | H | H | T | H | T | H | T | H | T |
| H | T | H | T | H | T | T | H | T | T | H | H | H | T | H | T | H | T | T | T | T | H |
| T | T | T | H | T | T | H | H | H | T | H | T | T | H | T | T | H | H | T | T | H | H |
| T | T | H | H | H | T | H | H | T | H | H | H | T | H | T | T | H | T | T | T | T | H |
| T | T | T | T | H | T | T | H | H | H | H | H | T | T | H | H | T | T | H | H | H | H |
| T | T | T | H | T | H | T | H | H | H | H | T | H | H | H | H | H | T | T | H | H | T |
| H | H | H | H | T | T | H | T | T | T | H | T | T | H | T | H | T | H | T | H | T | H |
| H | T | T | T | T | H | T | H | H | T | T | H | H | H | T | T | T | H | T | T | T | T |
| H | H | T | T | H | H | T | T | H | T | T | H | T | H | H | T | H | H | H | H | T | H |
| T | T | H | H | H | T | T | T | T | T | H | T | T | T | T | T | T | H | T | H | H | H |
| H | T | H | H | T | H | T | H | H | H | H | T | H | T | H | H | H | T | H | T | H | T |
| H | H | H | H | T | T | H | T | H | T | H | T | T | H | T | H | H | T | T | H | T | H |
| H | T | T | T | H | T | T | H | T | T | T | H | H | T | T | H | H | T | H | H | H | H |
| H | T | T | H | H | H | T | H | T | T | T | H | H | H | H | H | H | T | H | T | T | T |
| T | H | H | T | H | H | H | H | T | T | T | H | T | H | T | T | T | T | H | T | T | T |
| H | T | H | H | H | H | H | H | T | H | T | H | H | H | T | T | T | T | T | T | H | H |
| T | H | T | T | T | H | H | H | H | T | T | H | T | H | H | T | T | T | T | H | T | T |
| H | H | T | H | T | T | T | H | H | T | T | T | T | H | T | H | T | H | H | T | H | T |
| H | H | T | H | T | T | H | T | H | H | H | T | T | H | H | H | T | H | H | T | T | T |
| T | T | T | T | T | H | T | H | H | H | T | T | T | T | H | T | H | T | H | T | H | H |

TOSS COIN + LOOK FOR PATTERN HH.

I FIND $\bar{x} = 4.8$ $s_x = 3.9$ $n_x = 20$

$\bar{x}$ for $n_x = 20 \sim \mu_x$?

CI for $\mu_x$: $4.8 \pm 1.96 \frac{3.9}{\sqrt{20}}$

$n_y = 30$ "HT"

# Thirty rows each of 22 coin tosses.

```
MatrixForm[Table[Flatten[Table[c[[Random[Integer, 1] + 1]], {i, 1, 22}]], {j, 1, 30}]]
```

2
5

| | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| T | H | T | H | H | T | H | T | T | T | H | H | T | T | T | H | H | T | H | T | T | T |
| H | H | H | H | T | H | T | H | H | T | H | H | H | H | H | H | T | T | H | H | H | T |
| T | T | H | T | H | H | H | H | H | H | H | T | T | H | T | T | T | H | T | H | H | T |
| H | T | T | H | T | H | T | T | T | H | H | T | T | H | T | H | H | T | T | T | T | H |
| H | T | H | H | H | T | H | T | H | T | H | T | H | H | H | H | T | H | T | T | H | T |
| H | T | H | T | T | T | T | H | T | T | T | T | H | T | H | T | H | H | H | T | H | H |
| T | T | H | T | H | H | T | T | T | H | T | T | H | H | H | H | H | T | H | T | T | H |
| T | T | H | T | T | H | H | H | T | H | H | H | T | H | T | H | H | T | T | T | T | H |
| H | T | T | T | H | H | T | H | H | H | H | T | H | T | H | H | H | T | T | T | T | H |
| T | H | T | H | T | H | T | H | T | H | T | H | H | T | T | H | H | H | T | T | H | T |
| H | T | H | H | H | H | T | T | H | T | T | T | T | H | T | H | H | H | T | T | H | T |
| H | H | T | H | H | H | T | T | T | T | T | H | H | T | T | T | H | H | T | H | T | H |
| H | H | H | T | T | H | H | T | T | T | T | T | H | T | H | T | T | T | H | T | T | T |
| H | H | T | H | T | T | H | T | T | T | H | T | T | T | T | H | H | T | H | H | T | H |
| T | T | T | T | T | H | T | H | T | T | T | T | H | T | T | T | H | H | H | H | T | T |
| T | T | H | T | T | T | T | H | T | H | T | H | H | T | H | H | T | H | H | H | T | H |
| H | H | T | H | H | T | H | H | T | H | T | T | T | H | H | H | H | T | H | H | T | T |
| H | H | H | H | T | T | H | T | T | T | T | H | T | T | H | H | T | T | T | T | H | H |
| T | H | T | H | H | T | T | T | T | T | T | T | H | H | T | H | H | T | H | T | T | T |
| T | T | H | T | T | H | H | T | H | T | H | H | H | T | T | H | T | T | H | T | H | H |
| H | T | T | T | T | H | H | H | T | H | T | H | H | H | H | T | T | H | T | H | H | H |
| T | H | H | H | T | H | H | H | T | T | T | T | T | H | T | T | T | T | H | H | T | T |
| H | H | H | H | H | H | T | T | H | H | H | H | T | H | T | H | T | T | T | H | T | H |
| H | T | H | T | H | H | T | H | H | T | H | T | H | H | T | H | T | T | T | H | H | H |
| H | H | T | T | T | H | H | H | T | T | T | T | T | H | H | H | H | T | H | H | H | T |
| H | H | H | T | H | H | H | T | H | T | T | T | H | T | H | T | T | T | T | H | H | T |
| H | H | T | H | T | H | T | T | H | T | H | H | H | T | H | H | H | T | H | T | H | H |
| T | H | T | H | H | T | H | T | H | T | H | H | H | H | T | H | H | T | T | T | H | T |
| T | H | H | T | H | H | H | H | H | T | H | H | H | T | H | H | T | T | H | T | T | H |
| T | T | T | H | H | H | T | H | T | T | H | H | H | T | T | T | T | T | H | H | H | T |

IF WANT HT & FAIL. YOU GET HH

IF YOU WANT HH & FAIL. YOU GET HT

NOW YOU NEED WNH

ACTUALLY,

$\mu_x = (HH)\,6$

$\mu_y(HT) = 4$

$\bar{y}(HT) = \boxed{6}$ $s_y = 1.47$

CI USING $p_g$ HH FOR ORIGINAL ASSIGNMENT "SOLUTION"
+ $p_g$ HT FOR EXTRA COINTOSSINGS POSTED.

4 CI for

$$\underbrace{\mu_x - \mu_y}_{\approx 2 \text{ IN THEORY}} \quad (4.8 - 6) \pm 1.96 \sqrt{\frac{3.9^2}{20} + \frac{1.47^2}{30}}$$

WITH THIS DATA OPT'L

$$f_x = \frac{3.9}{3.9 + 1.47}$$

10. As in (9) but instead look for the pattern HT and repeat 30 times. Calculate sample mean $\overline{y}$ and sample standard deviation $s_y$.

11. Does it seem to you that the population mean time $\mu_x$ to get HH should be the same as the population mean time $\mu_y$ to get HT? If they differ which should be the larger (population) mean time? Why?

12. Give a 95% z-based CI for $\mu_x$ - $\mu_y$. Does the CI fall entirely to one side of 0? If so we might take this as some evidence that the population means are not exactly the same, especially if the CI falls far from 0. **Note: In the lecture 10-21-09 we had a REAL FLUKE with a similar example using patterns HTH and HTT when our sample means (obtained from students flipping coins) happened to both EXACTLY equal their theoretical population means for that example (which I know and shared without proof). As you look over the lecture notes do not be confused by the fluke that happened there. Probably ours is the only class ever to have it happen for this example. To flukes!**

**Unrelated to above.**

**13. z-based 95% CI for difference of proportions $p_x - p_y$ when independently sampling populations x, y. See page 561. This will be covered 10-26-09.** From a very large population of patients we sample 30 (equal probability with replacement) to receive a placebo. Of these 30 there are 20 who feel better 2 hours later. From the same population we sample 40 (equal probability with replacement) to receive a prescription medication. Of these 40 there are 32 who feel better 2 hours later. The population is so very large that we consider these samples without-replacement. Give a 95% z-based CI for the difference $p_x - p_y$.

CI for $p_x - p_y$

$$\hat{p}_x - \hat{p}_y \pm 1.96 \sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}}$$

$N_y = 600$

SPSE W/O REPL

$$\frac{20}{30} - \frac{32}{40} \pm 1.96 \sqrt{\frac{20/30 \; 10/30}{30} + \left(32/40 \; 8/40 / 40\right)\frac{600-40}{600-1}}$$