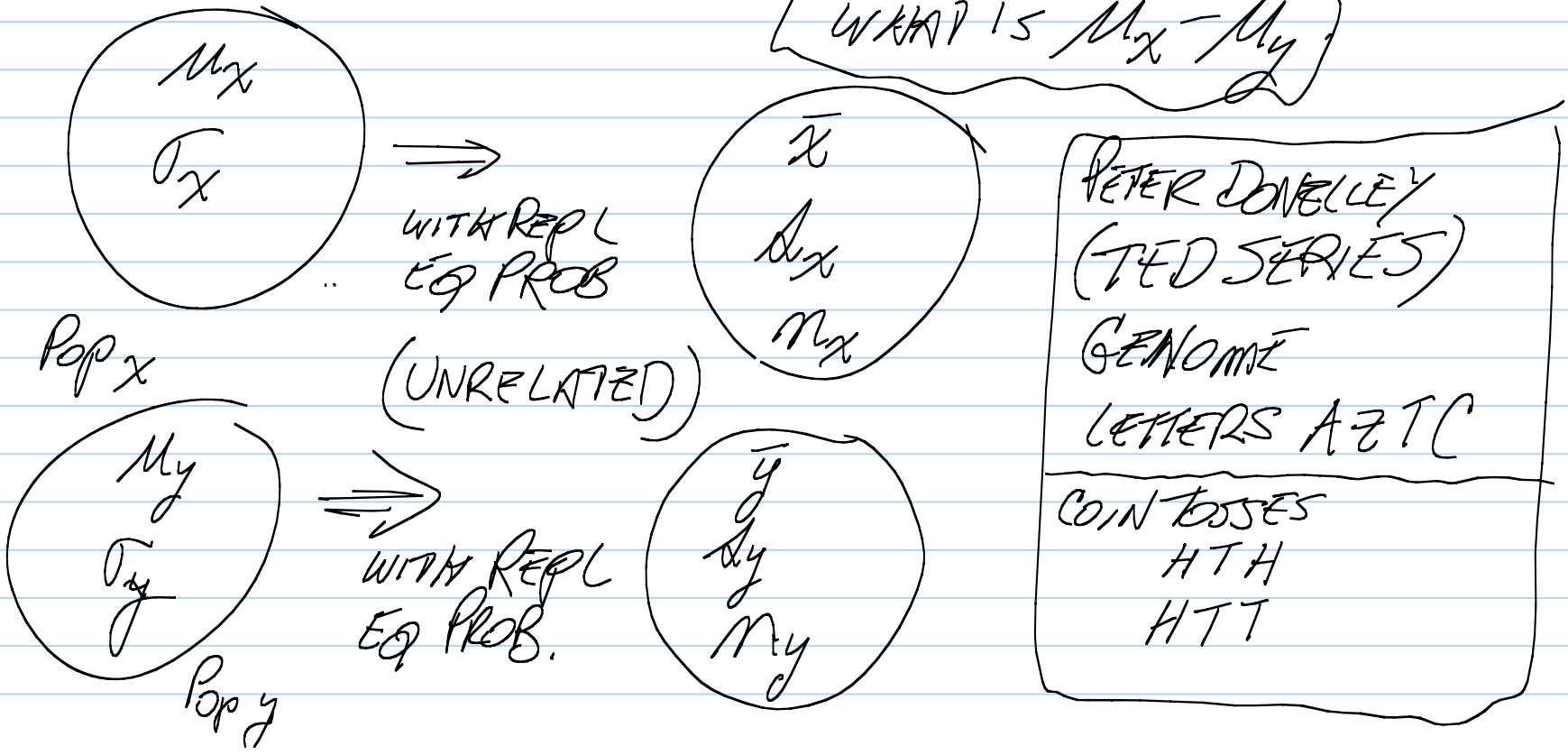
STT 200 10-21-09

Note Title 10/21/2009

TODAY CI FOR DIFF μ_x, μ_y

IDEA



QUESTION: WHICH TAKES LONGER ON AVG.

(a) WAIT FOR HTH

(b) WAIT FOR HTT

(c) SAME.

LET X = #TOSSES TO GET HTH

Y = #TOSSES TO GET HTT.

SAMPLES (STUDENT GENERATED)

$X_1, X_2, \ldots, X_{n_x}$ form $\bar{x}, s_x; n_x$

$Y_1, Y_2, \ldots, Y_{n_y}$ " $\bar{y}, s_y, n_y$

95% CI FOR $\mu_x - \mu_y$:

$$\bar{x} - \bar{y} \pm 1.96 \sqrt{s_x^2/n_x + s_y^2/n_y}$$

{CAUTION: SEPARATE CI FOR X, Y WON'T DO

EASIER TO GET HTT THAN HTH BECAUSE BOTH REQUIRE LEADING OFF WITH HT, BUT IF FAIL TO GET HTT (GETTING HTH) YOU ARE ALREADY AHEAD

CLAIM: (a).

HT(T) HOPE FOR H NEXT WAITING FOR HTH?

HT(H) WAITING FOR HTT?

(a) AVG WAIT TO HTH IS LARGER ($M_x > M_y$)

(b) AVG WAIT TO HTT IS LARGER ($M_x < M_y$)

(c) SAME AVG WAIT

CAN BE PROVEN THAT ON AVG TAKES TEN TOSSES TO GET HTH BUT ONLY 8 TO GET HTT:

HHTTTHTTTTTHHTTTTT(HTT)T

HTTTHHHHTH(T)HTH

MY SCORE $X = 32$

(x) HTH 15 5 8 15 11 7 8 11 $\{x_i\}$ $\bar{x} = 10$

$s_x = 3.66$

$n_x = 8$

(y) HTT 3 4 5 14 9 10 8 16 3 $\bar{y} = 8$

$s_y = 4.74$

$n_y = 9$

95% CI FOR μ_x $10 \pm 1.96 \frac{3.66}{\sqrt{8}}$

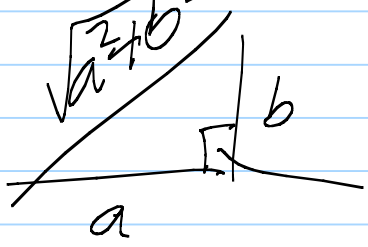
μ_y $8 \pm 1.96\, 4.74/\sqrt{9}$

95% CI FOR $\mu_x - \mu_y = (10 \ominus 8) \pm 1.96 \sqrt{\frac{3.66^2}{8} \oplus \frac{4.74^2}{9}}$

$\bar{x} \ominus \bar{y}$

$\sqrt{\frac{s_x^2}{n_x} \oplus \frac{s_y^2}{n_y}}$

PYTHAGORAS' THEOREM



STATISTICAL INDEPENDENCE

CAVEATS:

REQUIRE n_x, n_y "LARGE ENOUGH"

[LATER WE DISCUSS "BOOTSTRAP" RESAMPLING YOUR DATA TO CHECK UP ON STABILITY OF CI]

RETURN TO OUR CI FOR $\mu_x - \mu_y$:

$$(10-8) \pm 1.96 \sqrt{\frac{3.66^2}{8} \oplus \frac{4.74^2}{9}}$$

$\{ -2 , 6 \}$ WAS (BEFORE YOU SEE IT)

$P($ $\mu_x - \mu_y$ "IN" (COVERED BY) $\left((\bar{x} - \bar{y}) \pm 1.96 \sqrt{\frac{\sigma_x^2}{n_x} \oplus \frac{\sigma_y^2}{n_y}} \right)) \approx .95$

TO GET MORE PRECISION (NARROWER CI)
REQUIRES n_x, n_y LARGER.

TO MAKE $\sqrt{\frac{\sigma_x^2}{n_x} \oplus \frac{\sigma_y^2}{n_y}}$ SMALL

SHOULD ALLOCATE MORE DATA TO
SAMPLING x IF σ_x^2 IS LARGEST ((SEE BELOW))

WHY THE $\frac{\sigma_x^2}{n_x} \oplus \frac{\sigma_y^2}{n_y}$?

TO PREPARE YOU FOR THIS I'LL SPEAK TO
AN ARITHMETICAL FACT THAT UNDERLIES, ↓

LOOK AT TWO LISTS OF SCORES x, y.

x \ y	4	6	9
1			
5			
3			

$x-y$		
$1-4=-3$	$1-6=-5$	$1-9=-8$
$5-4=1$	$5-6=-1$	$5-9=-4$
$3-4=-1$	$3-6=-3$	$3-9=-6$

$$\sigma_{(x-y)} = \sigma\{-3, -5, -8, 1, -1, -4, -1, -3, -6\}.$$

CLAIM: $\sigma_{(x-y)} = \sqrt{\sigma_x^2 \oplus \sigma_y^2}$

ROOTS OF THIS ARE IN FACT THAT FOR x "INDEPENDENT OF y" (ALL POSS COMBOS AS ABOVE)

WE HAVE $\overline{xy} = \bar{x}\bar{y}$

PROOF:

$$\overline{xy} = \frac{\sum_i \sum_j x_i y_j}{n^2}$$

— OCCUR IN ALL POSS COMBOS

$$= \frac{\sum_i x_i \left(\sum_j y_j\right)}{n} \frac{1}{n} = \bar{y}\,\frac{\sum x_i}{n} = \bar{y}\bar{x}$$

FURTHERMORE, LET'S JUST SUPPOSE $\bar{x} = \bar{y} = 0$

THEN $\overline{(x-y)^2} = \overline{x^2 - 2xy + y^2}$

$$\therefore = \overline{x^2} - \underbrace{\overline{2xy}}_{0} \oplus \overline{y^2} = \overline{x^2} \oplus \overline{y^2}$$

$$\Rightarrow \overline{((x-y)-(\bar{x}-\bar{y}))^2} = \overline{(x-\bar{x})^2} \oplus \overline{(y-\bar{y})^2}$$

(n-DIVISOR) $\mathrm{Var}(x-y) = \mathrm{Var}x \oplus \mathrm{Var}y$

↑ CASE OF x INDEP OF y (MEANS ALL COMBOS)

$$\Rightarrow \sigma^2_{(x-y)} = \sigma^2_x \oplus \sigma^2_y$$

SAMPLE ALLOCATION.

95% CI FOR $\mu_X - \mu_Y$ IS $(\bar{x} - \bar{y}) \pm 1.96 \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}$

IF TOTAL SAMPLE SIZE $n = n_X + n_Y$ IS FIXED HOW SHOULD WE ALLOCATE TO MAKE CI NARROWEST?

SEEMS WE SHOULD MAKE n_X LARGER IF $\sigma_X^2 >> \sigma_Y^2$. FOR EXAMPLE. THE ACTUAL PRINCIPLE OF ALLOCATION IS A BIT SUBTLE. LET $f = n_X/(n_X + n_Y)$ BE THE FRACTION OF SAMPLES DEVOTED TO X DATA.

HOW TO CHOOSE f TO MINIMIZE $\frac{\sigma_x^2}{n_x} \oplus \frac{\sigma_y^2}{n_y}$.

THAT IS, CHOOSE f TO MINIMIZE

$$\frac{\sigma_x^2}{fn} \oplus \frac{\sigma_y^2}{(1-f)n}$$

WHERE $n = n_x + n_y$ IS TOTAL SAMPLE SIZE.

THE SOLUTION (REQUIRES CALCULUS) IS

$$f_x = \frac{\sigma_x}{\sigma_x + \sigma_y}$$

= IDEAL FRACTION OF n DEVOTED TO SAMPLING x DATA.

OF COURSE WE DON'T KNOW σ_x, σ_y UNTIL AFTER