Review Notes

This time: final exam review session

Next time: final—Thur, Jan 15, noon-3 in Hodge Theater 1410

Last example of interpreting correlation and regression
[p. 140:141 #41]

Regression line for predicting x from y (slope > 1)

SD line

Regression line for predicting y from x (slope < 1)

\[
\begin{align*}
\text{Answer: } & \text{ wife has 8 years of schooling } = \frac{B-12}{3} = 25D \\
r^2 &= .5 \\
.5 \times 2 &= 1 \text{ SD above average} \\
12 + 1\text{SD} &= 12 + 3 = 15
\end{align*}
\]

\[\text{husband has } 15 \text{ years of education}\]

\[\text{Answer: } \frac{15-2}{3} = 1 \text{SD} \]

\[.5 \times 1\text{SD} = .5 \times 5 = 2.5 \text{ years above average} \]

\[\text{wife has } 13.5 \text{ years of education...?} \]

\[\text{Wrong! Doesn't make sense. This is } \]

\[\text{because we must use the slope of } \]

\[\text{the regression line for predicting } x \text{ from } y = \frac{1}{.5} = 2 \]

Final - Problems, open everything

- probability models for sums of means
- 2 independent samples
- 1 sample, inference on proportions
- correlation and regression
- 2 sample paired comparisons

At least one time, you will be if inference is appropriate.
1. **Pop**
   - All riders at rush hour

**Sample**
   - Observed underground riders

**IDS**
   - Possible sums

\[ \frac{\text{like SES}}{110} \]

\[ n = 200 \]

\[ \text{L-R mean } = \frac{\text{SSE}_{100}}{5} = \frac{n}{4} \]

\[ = 200 \times 155.76 = 31152 \]

\[ \text{L-R SD} = SE_{100}(5) = 0 \sqrt{n} = 301.6 \times \sqrt{200} = 424.16 \]

** Hypothesis **

\[ \text{L-R hint:} \]

\[ \chi^2 = 3100 \]

\[ \frac{9.798}{9.998} \]

\[ 3.75 \]

\[ \chi^2 = (3.75)(424.16) + 310003 = 3250 \]

- Should engineer if to hold an extra 50% of riders every 111 days.
2. a) \( \hat{p}_1 = 49.3\% \), \( \hat{p}_2 = 48.4\% \) - diff not large enough in practical terms. Doesn't even make sense to ask if STAT16 & not PRACTICE, but...

(uuse formula)  
\[ \hat{p}_1 - \hat{p}_2 = 0.9\% \]  
\[ \text{SE}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \]

\[ = \sqrt{\frac{(0.493)(0.507)}{592} + \frac{(0.484)(0.516)}{154}} = 0.9\% \]

\[ 95\% \text{ CI for } (\hat{p}_1 - \hat{p}_2) = (\hat{p}_1 - \hat{p}_2) \pm 2 \text{SE}(\hat{p}_1 - \hat{p}_2) = 0.9 \pm 0.9\% \]

\[ \frac{\hat{p}_1 - \hat{p}_2}{0.9\%} \]

\[ \sqrt{\text{not STAT16}} \]

b) \( \frac{\bar{X}_1 - \bar{X}_2}{s_1} = \frac{16.3 - 24.3}{2.4} = 31 \text{ X drop} \)

STAT16? (model is just like CS about skeletons, but here it is just sketched)

\( \bar{X}_1 = 16.8 \text{ wks} \)
\( \bar{X}_2 = 24.3 \text{ wks} \)
\( s_1 = 15.9 \text{ wks} \)
\( s_2 = 17.3 \text{ wks} \)
\( n_1 = 592 \)
\( n_2 = 154 \)

(uuse sixth formula)

\[ \bar{Y}_1 - \bar{Y}_2 = -7.5 \text{ wks} \]
\[ \text{SE}(\bar{Y}_1 - \bar{Y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{15.9^2}{592} + \frac{17.3^2}{154}} = 2.1 \text{ wks} \]

\[ 95\% \text{ CI for } (\bar{Y}_1 - \bar{Y}_2) = (\bar{Y}_1 - \bar{Y}_2) \pm 2 \text{SE}(\bar{Y}_1 - \bar{Y}_2) = -7.5 \pm 4.8 \text{ wks} \]

\[ \text{STAT16} \]

c) Income support did cause a 31% drop in weeks of paid work (highly p<0.05).
3. Practised?

diff. between 9.87% and 12.5% of 65 in age -

1970  \[ \begin{align*}
1 = 65^+, o = \text{not} \\
\uparrow & 1s \\
\& & n = 203,000,000 \\
\downarrow & 0s \\
\text{mean (p, o) = 9.87} \end{align*} \]

(1990)  \[ \begin{align*}
1 = 65^+, o = \text{not} \\
\uparrow & 1s \\
\& & n' = 479,000,000 \\
\downarrow & 0s \\
\end{align*} \]

2 independent samples (like CS on preferential treatment)

\[
\frac{T - E}{E} \times 100\% = \frac{12.5\% - 9.87\%}{1.8\%} = 18.9\%
\]

\( T \) = proportion of people 65+ in US has
gone up 18.9% in only 20 years (highly fractioned)

\text{STATS16?}

\[
\begin{array}{ccc}
\text{pop} & \text{sample} & \text{1DS} \\
\text{all Americans} & \text{observed people} & \\
1 = 65^+, o = \text{not} & 1 = 65^+, o = \text{not} & \\
\uparrow & \uparrow & \\
1s & 1s & \\
\& & \& & \\
\uparrow & \uparrow & \\
\& & b & b \\
\downarrow & \downarrow & \\
0s & 0s & n = 203,000,000 \\
\end{array}
\]

Sample pop, so \( \text{no uncertainty - no need for inference, STATS16 doesn't make sense.} \)
4. a) pop
   | all eligible
   | jury
   \[ N = \text{big}\]
   \[ \mu = \text{big}\]
   \[ \sigma = \text{big}\]
   \[ \text{mean}(\mu) = 57\%\]
   \[ \text{gender} (1 = f, 0 = m) \frac{15}{295} \]
   \[ \text{if sample} \]
   \[ \text{gender} (1 = f, 0 = m) \frac{15}{295} \]
   \[ \text{possible} \]
   \[ \hat{p} \text{ value} \]

Sample
| observed people
\[ n = 350 \]
\[ \text{mean} (\hat{p}) = ? \]

IDS
L-R Mean = \( E(\hat{p}) = \mu = 57\% \)
L-R SD = \( SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \)
\[ = \sqrt{\frac{0.57 \times 0.43}{350}} = 2.7\% \]
L-R Hist:
CLT
\[ \frac{10.2}{350} = 2.9\% \]
\[ 29\% \]
\[ 53\% \]
\[ -89 \]
\[ 0 \]
\[ 89 \]

b) pop
| Sample
\[ n = 100 \]
| D/M/D

mean (\( p \)) = 29\%
mean (\( \hat{p} \)) = ?

CLT
\[ \text{se}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{100}} = 4.5\% \]
5. $\hat{y} = \hat{a} + \hat{b}x$

$(\text{pred. height}) = 66.75 + 0.25 \text{ in yr. ed.} \text{ (ed level)}$

a) $\hat{y} = 66.75 + 0.25 \text{ in yr. ed.} (12 \text{ yr.}) = 69.75 \text{ in}$

$\hat{y} = 66.75 + 0.25 \text{ in yr. ed.} (16 \text{ yr.}) = 70.75 \text{ in}$

In cross-sectional data, longitudinal question:

very difficult to draw valid long. conclusions
from cross-sectional data.

Associated with each
1 yr of education is
an increased $\frac{1}{4}$ in
of height. If you
compare men 25-34,
whose education level
differs by 4 years,
you would expect
their height to differ
by one inch on average.

(PCF: better nutrition)

b) Association correlation ≠ causation; surved is cross-
sectional and observational, so we can't draw long.
causal conclusions like the one in (b), causing both
income and height to be associated.