*Final exam Review → Sat. 12-2
in Baskin Auditorium

2 June 2005

Last on Regression p. 123 #4
5 number summary

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Husband (M) ed. level</td>
<td>12 yrs</td>
<td>3 yrs</td>
</tr>
<tr>
<td>Wives (W) ed. level</td>
<td>12 yrs</td>
<td>3 yrs</td>
</tr>
</tbody>
</table>

1) If wife has 18 yrs of schooling, what is husband expected to be?
   She is \( \frac{18-12}{3} = 2 \) SD's above average → we expect her husband is only \( r \cdot 2 \) SD's above average in ed. level → \( (0.5)(2 \text{SD}) = 1 \text{SD} \) above average → predict husband has 15 yrs.

2) If husband has 15 yrs of schooling, what is wife? He is \( \frac{15-12}{3} = 1 \text{SD} \) above average → we expect his wife to be only \( r \cdot 1 = (0.5)(1 \text{SD}) = 1 \frac{1}{2} \text{yrs} \) above average → 13.5 years of schooling predicted for wife.

* There are actually 2 different regression lines in any scatterplot: one for predicting \( y \) from \( x \) and a different line for predicting \( x \) from \( y \).

moral: if somebody switches the role of \( x \) and \( y \) in the middle of a problem, watch out.
Review for final (open-book, open-notes)
5 problems:
* correlation and regression
* probability models for sums (like escalator problem)
* 2 independent samples w/ continuous data
* 1 sample problem w/ 0, 1 data
* 2 sample paired comparison

- In one problem, you will be asked to do inference or to say why doing inference would be inappropriate.

Review problems:

5. education = x variable
   height = y variable
   a) predict height of a man with 12 yrs of education:
      height = 66.75 inches + (0.25 inches/year)(12 yrs)
      = 69.75 in.
      predict height for 16 yrs education:
      height = 66.75 in + (0.25 in/yr)(16 yrs)
      = 70.75 in. (1 in. taller)

Q: what does this mean?
   (going back to college and seeing if the guy gets taller is a longitudinal question but data is cross-sectional)
A1: slope came out positive (correlation is positive)
   -> PCF: income ↑ (height ↑ because of nutrition and education ↑)
A2: if we look at two groups of men who differ on average by 1 yr of ed. we can expect them to differ
   on average by 0.25 in. in height.
b) positive correlation is probably due to better
   nutrition as a child (might get a raise with elevator shoes, but simpler explanation of this correlation is
   PCF (nutrition))
 imaginary data set

possible \( \hat{p} \)'s

\[
\begin{bmatrix}
51\% \\
50\% \\
\vdots \\
\end{bmatrix}
\]

long run mean: \( E_{\text{long}}(\hat{p}) = \hat{p} = 53\% \)

long run SD: \( \text{SE}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.53(0.47)}{350}} = 2.7\% \)

long run hist. of \( \hat{p} \):

<table>
<thead>
<tr>
<th>SE</th>
<th>2.7%</th>
</tr>
</thead>
<tbody>
<tr>
<td>z-score</td>
<td>0</td>
</tr>
<tr>
<td>raw</td>
<td>53%</td>
</tr>
<tr>
<td>SU</td>
<td>29%</td>
</tr>
</tbody>
</table>

\[
\frac{29 - 53}{2.7} = -8.9
\]

actual = \( \frac{102}{350} = 29\% \)

Chance of less than 29% \( \approx 0\% \)

b) imaginary data set

possible \( \hat{p} \)'s

\[
\begin{bmatrix}
\end{bmatrix}
\]

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

SD = \( \text{SE} = \sqrt{\frac{0.29(0.71)}{100}} = 4.5\% \)

histogram:

<table>
<thead>
<tr>
<th>SE</th>
<th>4.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>raw</td>
<td>29%</td>
</tr>
<tr>
<td>SU</td>
<td>9%</td>
</tr>
<tr>
<td>z</td>
<td>-4.4</td>
</tr>
</tbody>
</table>

actual = \( \frac{9}{100} = 9\% \)

Chance of 9% or less \( \approx 0\% \)
(3) It always makes sense to think about whether a difference is large in practical terms.

\[
\frac{1970}{9.8\%} \quad \frac{1990}{12.5\%} \quad \Rightarrow \frac{T-C}{C} = \frac{12.5 - 9.8}{9.8} = \frac{2.7\%}{9.8\%} = 28\%
\]

( before ) ( after )

( after-before )

\[
\text{before} \quad \text{after}
\]

i.e., U.S. had 28% more elderly ppl in 1990 than in 1970 (huge practical significance)

\[
\begin{array}{l}
\text{1970} \\
\text{pop} \\
\text{sample} \\
\text{imaginary data set}
\end{array}
\]

\[
\begin{bmatrix}
\text{old}^2 \\
0
\end{bmatrix}
\Rightarrow
\begin{bmatrix}
1^5 \\
1^5 \\
0^5
\end{bmatrix}
\Rightarrow
n = 203 \text{ mill.}
\]

\[
p = 9.8\% \\
\hat{p} = 9.8\%
\]

Not sensible to ask if stat. sig. because pop = sample
so inference is silly.