12th Class **6/23/10**

"In God we trust, all others must use data."--Edward Deming

hand out review sheet, answer, point to old test, answers

now that we've done Ch. 9, for QAC201 folks, spend a little time on chi-squared tests (can look at Ch. 17 for reference)

allows us to consider the whole frequency distribution instead of just one part of it

e.g. look at polling for all candidates instead of just the winning candidate's percentage

e.g. look at the color distribution for all M&Ms instead of just blue

compare the observed frequency to the expected frequency and test the null hypothesis

create the test statistic $\chi^2 \equiv \sum \frac{(O-E)^2}{E}$

this is distributed according to the chi-square distribution (see table VII), similar to the t-distribution in that it is a family of distributions indexed by the d.f. (equal to the number of categories minus one) that converges to the normal distribution as the degrees of freedom increase

H0: the sample comes from the expected distribution
Ha: the sample doesn't come from it
If the sample value exceeds the critical value in table 7, reject the null (the deviations are too great)

note you can also do this if the data are sorted by more than one variable
births example: by season, or by season and part of country
(in two-variable case, d.f. = (rows-1)*(columns-1)
the Swedish case covered in the chapter: note that the four seasons are not of equal length (winter is five months, summer and fall only two each) so different than in the northeastern U.S. Let's do our class:
out of 12 people, expect 3 in each season if births distributed equally (null hypothesis)
compare to test value of 7.81 (at 5% significance level)
for our class, we can't reject the null (3, 4, 1, 4, so sample value is 2)

Ch. 11 and 12: first part of Part III of the course: relating two or more variables

fitting a line to a scatter plot (Y is dependent variable, X is independent variable; so we are making causality assumption)
[draw up picture]

fit by eye or use a criterion

commonly-used criterion (but not the only one): minimize the sum of squared deviations of Ys from the line

short history of the Method of Least Squares:
[all facts as related in Stephen M. Stigler, The History of Statistics: The Measurement of Uncertainty before 1900, Cambridge, MA: Belknap Press, 1986]
1805: Adrien Marie Legendre developed the Method of Least Squares as a method of reconciling a set of measurements; it was rapidly accepted and diffused geographically; used in astronomy and geodesy

$$\hat{Y} = a + bX$$

$$d = Y - \hat{Y}$$

$$\text{min. } d^2 = \sum (Y - \hat{Y})^2$$

[the d^2 above should have a summation sign in front of it]

minimize with respect to choices of a and b:

solution to this (using calculus) gives optimal a (intercept) and b (slope):

$$\text{where } b = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{\sum (X - \overline{X})^2}$$

$$\text{and } a = \overline{Y} - b\overline{X}$$

note also that

$$b = \frac{Cov(XY)}{Var(X)}$$

(actually MSD in denominator, but close enough for large n)

go through and derive this (set S, for sum of squares, equal to d^2):

$$\begin{aligned}\min S &= \Sigma[Y - (a + bX)]^2 \\ &= \Sigma\,[Y^2 - 2(a + bX)Y + (a + bX)^2] \\ &= \Sigma Y^2 - 2a\Sigma Y - 2b\Sigma XY + \Sigma\,(a^2 + 2abX + b^2X^2) \\ &= \Sigma Y^2 - 2a\Sigma Y - 2b\Sigma XY + \Sigma a^2 + 2ab\Sigma X + b^2\Sigma X^2\end{aligned}$$

$$\partial S/\partial a = -2\Sigma Y + 2na + 2b\Sigma X = 0$$
$$\partial S/\partial b = -2\Sigma XY + 2a\Sigma X + 2b\Sigma X^2 = 0$$

get rid of the 2s throughout and reorder terms:

$$na - \Sigma Y + b\Sigma X = 0$$
$$a\Sigma X - \Sigma XY + b\Sigma X^2 = 0$$

$$na = \Sigma Y - b\Sigma X$$
$$a = \Sigma Y/n - b\Sigma X/n$$
$$a = \bar{Y} - b\bar{X}$$

$$(\bar{Y} - b\bar{X})\Sigma X - \Sigma XY + b\Sigma X^2 = 0$$
$$\bar{Y}\Sigma X - b\bar{X}\Sigma X - \Sigma XY + b\Sigma X^2 = 0$$
$$b(\Sigma X^2 - \bar{X}\Sigma X) = \Sigma XY - \bar{Y}\Sigma X$$

$$\text{so } b = [\Sigma XY - n\bar{X}\bar{Y}]/[\Sigma X^2 - n\bar{X}^2] = nCOV(X,Y)/nMSD(X)$$

$$\text{where } b = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sum(X - \bar{X})^2}$$

another way to write this is to set up a notation for the deviations:

$$x = X - \overline{X}$$

$$y = Y - \overline{Y}$$

(notation to remind us these deviations are generally smaller values than the original values X and Y)

now can write b as:

$$b = \frac{\sum xy}{\sum x^2}$$

this is known as ordinary least squares (OLS) estimates of a and b

why do we tend to use this method of line-fitting in preference to other methods?

--simple formulas for a and b
--goes through the means for X and Y
--efficient
--unbiased

still, there are many modifications of this basic methodology which have been developed to adjust for problems with the data, e.g. WLS

examples from economics of these simple linear functions:
— consumption function in macroeconomics:
if GNP = national income, then C = a + b(GNP), where b = MPC
(think about how you would want to expand this to include other variables,
e.g., wealth: C = a + b(GNP) + c(Wealth))
— investment function in macroeconomics: I = f(r); expand to I = f(r,GNP)
— demand curve in microeconomics: Qd = a + bP (where $b \leq 0$)
— supply curve in microeconomics: Qs = c + dP (where $c \leq 0$)
(think again about how you would want to expand these functions to include other variables)

on to Ch. 12

Ch. 11 shows how to summarize data by fitting a line to it. But in order to make inferences about the population from which the sample scatter of observations was drawn, we must build a statistical model (i.e., state our assumptions about the nature of the population and the sample)—then we will be able to construct confidence intervals and test hypotheses.

1809: Carl Friedrich Gauss couched the problem of planetary orbit determiniation in explicitly probabilistic theory

[draw up a scatter diagram]

Note that the Ys are each drawn from <u>different</u> (conditional) probability distributions $P(Y_i \mid X_i)$, as they have different Xs associated with them

Simplifying assumptions about the Ys (about their distributions):

1) homogeneous variance: all Ys have the same variance σ^2
2) the mean of each Y $E(Y_i \mid X_i) = \alpha + \beta Xi$, where our goal is to estimate α and β; we shorthand $E(Y_i \mid X_i) = E(Y_i)$, but remember it is a conditional mean
3) the Ys are independent of each other

The <u>simple regression model</u> can be written as:

$$Y_i = \alpha + \beta X_i + e_i$$

where the e_is are independent <u>error terms</u>, and $E(e_i) = 0$, $Var(e_i) = \sigma^2$ for each e_i

error here signifies "wandering" and inclusion of an error term distinguishes statistical models from other models [give examples of other models]

This was a big insight in economics (from the 1940s) that economic models did not have to be exact and that error need not signify measurement error;

rather measurement error is one component of the error, but the other component is inherent variability (consider whether inclusion of more and more controls in the experiment could reduce this to zero; same as asking whether more and more variables in a multiple regression could reduce this to zero)

since $E(e_i) = 0$, $E(Y_i) = \alpha + \beta X_i$

rather than carry along all the i's, we can write the regression model in shorthand as:

$$E(Y) = \alpha + \beta X$$

so our notation for the true regression is: $E(Y) = \alpha + \beta X$
and for the fitted, or estimated regression, based on the available sample of paired Xs and Ys:

$$\hat{Y} = a + bX$$

we can describe the closeness of the slope of the estimated line to the slope of the true population line:

$$E(b) = \beta$$

$$SE(b) = \frac{\sigma}{\sqrt{\sum x^2}} \text{ where } x = (X - \overline{X})$$

let's decompose the standard error of b:

$$SE(b) = \frac{\sigma}{\sqrt{n\frac{\sum x^2}{n}}} = \frac{\sigma}{\sqrt{n}}\frac{1}{s_x} \text{ (for large n, since } \frac{\sum x^2}{n} \text{ is MSD, not variance)}$$

so standard error can be reduced in three ways:

1) reduce σ, the inherent variability of the Y observations (not generally feasible)
2) increase n (generally possible, but may be expensive)
3) increase s_x, the spread of the X values; s_x is called the <u>leverage</u> of the X values on b

[illustrate with figure like Figure 12-4]

of course, we cannot observe σ in general, so must estimate the standard error of b, using the <u>residual variance</u> s^2 to estimate the variance σ^2, where s^2:

$$s^2 \equiv \frac{1}{n-2}\sum (Y - \hat{Y})^2 \text{ and } SE = \frac{s}{\sqrt{\sum x^2}}$$

note that fitting a regression line leaves only n - 2 degrees of freedom, because you must have 2 points in order to fit a line (which would then be a perfect fit in that both points would lie on the line); hence the denominator in s^2

now we can create a confidence interval around b:

$$\beta = b \pm t_{.025} SE$$

$$\beta = b \pm t_{.025} \frac{s}{\sqrt{\sum x^2}}, \text{ where the d.f. for t are n - 2}$$

and we can test the null hypothesis that $\beta = 0$ (same as saying that X and Y are unrelated—the regression line is horizontal--draw it) by seeing if 0 is in the confidence interval or by setting up a t-ratio for b: $t = \frac{b}{SE}$

we can also set up a confidence interval for the mean of Y_0, given X_0:

$$\mu_0 = (a + bX_0) \pm t_{.025}\, s \sqrt{\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum x^2}}$$

and for an <u>individual</u> Y_0, given X_0:

$$Y_0 = (a + bX_0) \pm t_{.025}\, s \sqrt{\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum x^2} + 1}$$

(motivate why this interval is wider and why both intervals widen away from $\overline{X}$)

interpolation vs. extrapolation -- the latter is riskier [illustrate]
[show extrapolation cartoon]

Do problem 12-14
estimate the OLS line as Y = 16.25 + .625X
so our estimate if X (X_2) = 55 is Y (X_1)=50.625

for the 95% interval: $Y_0 = 50.625 \pm 1.98*SQRT(180)*SQRT[(1/150) + (55\text{-}70)^2/24000 + 1]$
$= 50.625 \pm 26.777$, or about 24 to 77

Note the underlying assumptions we made, namely that these 150 students were regarded as a random sample from a conceptual population of many students where the expected values of Y given X form a straight line, the variance of Y given X is constant along the line (think about this in context), and the student can be regarded as randomly drawn from this population (think about this in context)

Ch 13: multiple regression

go to a multiple regression model for several reasons:

1) are simultaneously interested in the effects of several independent variables on a dependent variable (often because you are testing a theory which involves several variables, or fitting a relationship that involves several variables, e.g. a production function)
2) want to explain as much of the variation in Y as possible (subject to Occam's Razor)
3) are only interested in the effect of one variable, but want to measure its effect accurately
   a. need to eliminate bias on estimating the direct effect of this variable by including confounding factors
   b. want to distinguish between direct and indirect effects of a variable (path analysis)

go on to Ch. 13 tomorrow