SESSION 9 LOGISTIC REGRESSION: AN INTRODUCTION

Binary Outcome Y = 0 or 1

average Y = PROPORTION (of Y's that are 1)

usually denote proportion by
if parameter.. Greek letter π or upper case P
if statistic.. $\hat{\pi}$ ("π-hat") or $\hat{P}$ ("P-hat") or lower case p

Inference concerning a single P {or single Odds = P/(1-P) }

$$p = \frac{\text{\# of observations with } Y=1}{\text{number of observations}} = \frac{\Sigma y}{n} = \bar{y}$$

Tests concerning (and CI's for) P based on ...

Exact Binomial: $\Sigma y \sim$ Binomial(n,P) if n small
$\sim$ Poisson(μ = nP) if n large & P small

or

Gaussian Approxn. to Binomial (n large & P not extreme):

p (= $\bar{y}$) ~ Gaussian(P, SD = sqrt[P(1-P)] / sqrt[n])
~ Gaussian(P, SE = sqrt[p(1-p)] / sqrt[n])

If Y = 0 or 1's, then σ^2(Y) = Var(Y) = P(1-P), where P = proportion of 1's

Inference concerning two P's

Several Comparative Parameters (unlike μ_1 vs μ_2)

$P_2 - P_1$ (Risk Difference RD)

P_2 / P_1 (Risk Ratio RR)

$\dfrac{P_2}{1-P_2} \Big/ \dfrac{P_1}{1-P_1}$ (Odds Ratio OR)

	sample 1	sample 2
# with Y=1	$a = \Sigma y = n_1\bar{y}_1 = n_1p_1$	$b = \Sigma y = n_2\bar{y}_2 = n_2p_2$
# with Y=0	$c = n_1(1-\bar{y}_1) = n_1(1-p_1)$	$d = n_2(1-\bar{y}_2) = n_2(1-p_2)$
total	n_1	n_2

Tests of $P_1 = P_2$ (or, equivalently, RR=1 or OR=1) based on..

Hypergeometric distribution ("Fisher's exact test")
-- conditions on (fixes) BOTH margins

Gaussian Approxn. to distrn. of difference of 2 p's
-- unconditional test ... only ONE fixed margin
-- $Z^2 = X^2$ (chi-square statistic)

CI's for RD, RR and OR ..

RD: - Gaussian Approxn. to distrn. of difference of 2 p's

- test-based method

RR: - Gaussian Approxn. to distrn. of diff. of logs of 2 p's

- test-based method

(see Epi textbooks)

OR: - Gaussian Approxn. to distrn. of diff. of logs of 2 odds (unconditional; "Woolf's method")

- "exact" CI based on Non-Central Hypergeometric distrn.

- test-based method (uses "null" SE)

(see Epi textbooks)

Test for trend (over X) in Proportions

	$X = x_1$	$X = x_2$	...	$X = x_k$
Prop. with Y=1	$\frac{\Sigma y}{n_1} = p_1$	$\frac{\Sigma y}{n_2} = p_2$	...	$\frac{\Sigma y}{n_k} = p_k$

x_1, x_2 , ... , x_k are numerical values or "spacings"

See Armitage and Berry textbook

Only a Test with a p-value..
No measure of actual gradient in proportions

Examples of gradients in Proportions

Illness rates in relation to number of Falls while WindSurfing
Low Birth Weight rates in relation to Altitude
Mortality (in rats) in relation to dose of Cadmium
Unemployment Rates as a function of Age, Education & Gender

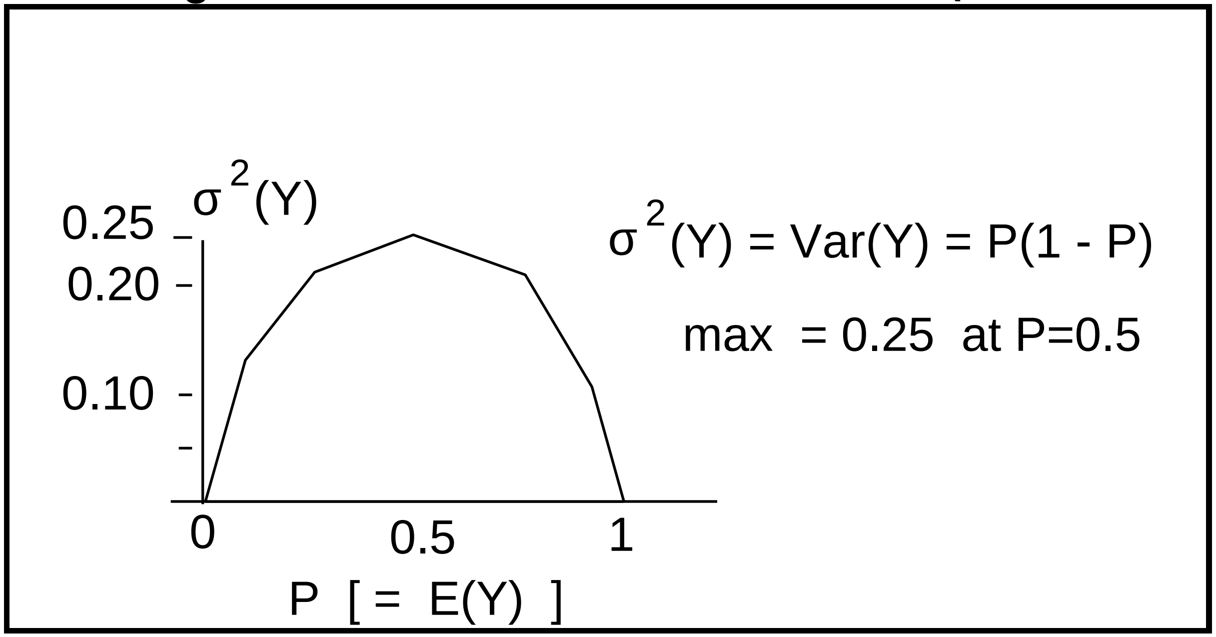
Why not use Y's in "regular" regression?

i.e.

μ**(Y | X1, X2,..) = Prop(Y=1 | X1, X2,..) =** $\beta_0 + \beta_1.X_1 + \beta_2.X_2 + ...$

- **constraints on range of P: 0 < "fitted" P's < 1**

- **Y's arising from P near 0.5 are more variable than Y's arising from P nearer to 0 or 1** **(?? different weight for each obsn.)**



- **P's unlikely to be linear over X if wide P range; more likely to be S-shaped (esp. in toxicology)**

<u>Other options for Binary Regression</u>

- use unequal weights to allow for different variances for Y's ..

  more weight to observations for which P (and thus P[1-P]) is more extreme

  less to observations for which P is more central (near 0.5) (and thus P[1-P]) is larger

  BUT... this still does not fix the issue of the shape of the "P vs X" function, and the fact that P must stay between 0 and 1 and be biologically "sensible"

In 1960 and 70's, statisticians devised ways to fit models where there was some flexibility in the choice of "P vs X" function (<u>Generalized</u> Linear models)

e.g. probit curves in toxicology
(had been around for many decades, but couldn't handle multiple X's very well)

logit curves in epidemiology
(Cornfield / Framingham study)

Generalized Linear Models

- use "link" function to "toggle" between

$$\text{IDENTITY} \quad \mu|X_1, X_2 \ldots = \beta_0 + \beta_1.X_1 + \beta_2.X_2 + \ldots$$

$$\text{LOG} \quad \log(\mu|X_1, X_2 \ldots) = \beta_0 + \beta_1.X_1 + \beta_2.X_2 + \ldots$$

$$\text{LOGIT} \quad \text{logit}(\mu|X_1, X_2 \ldots) = \beta_0 + \beta_1.X_1 + \beta_2.X_2 + \ldots$$

$$\text{PROBIT} \quad \text{probit}(\mu|X_1, X_2 \ldots) = \beta_0 + \beta_1.X_1 + \beta_2.X_2 + \ldots$$

- use "ERROR" distributions to "toggle" between

$Y\ |X_1, X_2 \ldots \sim$ **Gaussian**$(\ \mu[x_1,x_2],\ \sigma)$

$Y\ |X_1, X_2 \ldots \sim$ **Binomial**$(n,\ P[x_1,x_2])$

$Y\ |X_1, X_2 \ldots \sim$ **Poisson**$(\ \mu[x_1,x_2]\ )$

- Typically, there is a "natural" or "canonical" pairing of link and Error, but most software now allows the user to even "mix and match"

"natural" (LINK,ERROR) pairings

(IDENTITY, Gaussian) (LOGIT,Binomial) (LOG,Poisson)

With Binary Y's why logit rather than probit?

very little to choose between them on
quantitative grounds, but...

- influence of epidemiology

  Cornfield

  ODDS RATIO estimable in case-control studies
  without further data

  natural to extend from 2x2 tables to REGRESSION
  and produce ODDS RATIO's for continuous X's

  even if X itself binary, don't have to rely
  on Mantel-Haenszel aggregation of OR estimates
  (cells sparse if multiple X's)

- other reasons..

  see Cox 1970 Analysis of Binary Data

  Cox and Snell 1989 Analysis of Binary Data

With Logistic Regression, What's Different / The Same?

SAME

- the X's, indicator variables for categorical X's
- meaning of linear predictor
- meaning of individual coefficients
- confounding
- interaction
- model fits (variables added last / in order ..)

DIFFERENT

- the scale of Y's and the scale of P
- RATIOS on one scale are differences on another
- having to go forward/back from one scale to another
- error variance tied to mean... σ is function of μ [= P]
- =>can tell if model close to best achievable
- "Individual" Residuals not as meaningful;
- idea of "cells" or covariate patterns
- (tied to this) degrees of freedom available
- the statistics for testing additional terms in model
  => no longer tied to F or t (we don't have to estimate
  σ separately from $\mu[X_1, X_2 ..] = P[X_1, X_2 ..]$