BIOSTATISTICS METHODS

FOR TRANSLATIONAL & CLINICAL RESEARCH

DIAGNOSTIC MEDICINE

Part B: Statistical Issues

Predictive Values

The design in stage I involves two samples, those with the disease and those without. Sensitivity and Specificity are estimated by the corresponding sample proportions.

In stage II, there are no data other than one "incomplete observation", that of the user of which we know the test result but not the disease status. It is not possible to "estimate" the Predictive Values; we could only approximate them - indirectly.

BAYES' RULE

$$\Pr(B \mid A) = \Pr(B \text{ and } A)/\Pr(A)$$

$$\Pr(B \mid A) = \frac{\Pr(A \mid B)\Pr(B)}{\Pr(A \text{ and } B) + \Pr(A \text{ and Not } B)}$$

$$\Pr(B \mid A) = \frac{\Pr(A \mid B)\Pr(B)}{\Pr(A \mid B)\Pr(B) + \Pr(A \mid \text{Not } B)\Pr(\text{Not } B)}$$

We apply the Bayes' Rule twice:

Let B=(D=+) and A=(T=+) to obtain P+

Let B=(D=-) and A=(T=-) to obtain P-

POSITIVE PREDICTIVE VALUE

$$\Pr(B \mid A) = \frac{\Pr(A \mid B)\Pr(B)}{\Pr(A \mid B)\Pr(B) + \Pr(A \mid \text{Not } B)\Pr(\text{Not } B)}$$

Let A = (T=+) and B = (D=+), we have:

$$\Pr(D = + \mid T = +) = \frac{\Pr(T = + \mid D = +)\Pr(D = +)}{\Pr(T = + \mid D = +)\Pr(D = +) + \Pr(T = + \mid D = -)\Pr(D = -)}$$

$$P^{+} = \frac{S^{+}\pi}{S^{+}\pi + (1 - S^{-})(1 - \pi)}$$

NEGATIVE PREDICTIVE VALUE

$$\Pr(B \mid A) = \frac{\Pr(A \mid B)\Pr(B)}{\Pr(A \mid B)\Pr(B) + \Pr(A \mid \text{Not } B)\Pr(\text{Not } B)}$$

Let A = (T=-) and B = (D=-), we have:

$$\Pr(D = - \mid T = -) = \frac{\Pr(T = - \mid D = -)\Pr(D = -)}{\Pr(T = - \mid D = -)\Pr(D = -) + \Pr(T = - \mid D = +)\Pr(D = +)}$$

$$P^{-} = \frac{S^{-}(1-\pi)}{S^{-}(1-\pi) + (1-S^{+})\pi}$$

RESULTS

Both predictive values are functions of disease prevalence, π = Pr(D = +):

$$P^{+} = \frac{S^{+}\pi}{S^{+}\pi + (1 - S^{-})(1 - \pi)}$$

$$P^{-} = \frac{S^{-}(1 - \pi)}{S^{-}(1 - \pi) + (1 - S^{+})\pi}$$

EXAMPLES: AIDS SCREENING

Example A: S^+=.977, S^-=.926, and π=.003:

$$P^+ = \frac{(.977)(.003)}{(.977)(.003)+(.074)(.997)} = .038 \text{ or } 3.8\%$$

Example B: S^+=.977, S^-=.926, and π=.20:

$$P^+ = \frac{(.977)(.20)}{(.977)(.20)+(.074)(.80)} = .767 \text{ or } 76.7\%$$

Note: Current Estimate for USA's AIDS: .3%
and S^+ and S^- are for ELISA in Weiss, 1985.

$$P^+ = \frac{S^+\pi}{S^+\pi + (1 - S^-)(1 - \pi)}$$

$$\frac{dP^+}{d\pi} = \frac{S^+}{[S^+\pi + (1 - S^-)(1 - \pi)]^2}$$

Result: The higher the Prevalence, the higher the Positive Predictive Value.

$$P^- = \frac{S^-(1-\pi)}{S^-(1-\pi)+(1-S^+)\pi}$$

$$\frac{dP^-}{d\pi} = \frac{-S^-(1-S^+)}{[S^-(1-\pi)+(1-S^+)\pi]^2}$$

Result: The higher the Prevalence, the lower the Negative Predictive Value. However, the effect is much weaker here; the derivative is negative but near zero.

Versus :

$$\frac{dP^+}{d\pi} = \frac{S^+}{[S^+\pi+(1-S^-)(1-\pi)]^2}$$

Diagnostic Index

OVERALL AGREEMENT

- **The simplest measure would be the "overall agreement", Pr (T=D); it is an obvious candidate.**
- **However, unlike sensitivity and specificity, the overall agreement is influenced by the disease prevalence: not a good candidate**

$$\Pr(T = D) = \Pr(T = D = +) + \Pr(T = D = -)$$

$$\Pr(T = D) = \Pr(T = + \mid D = +)\Pr(D = +) + \Pr(T = - \mid D = -)\Pr(D = -)$$

$$\mathbf{Pr(T = D) = \pi S^{+} + (1 - \pi) S^{-}}$$

KAPPA STATISTIC

- **Kappa statistic is a popular statistic often used to measure agreement between observers; It adjusts overall agreement for "chance agreement".**
- **Similar to the case of overall agreement, Kappa is also influenced by the disease prevalence: not a good candidate.**

YOUDEN'S INDEX

- **Youden's Index (Cancer, 1950) is defined as:**

$$J = 1 - (\alpha + \beta) = S^{+} + S^{-} - 1$$

- **It does not depend on Prevalence and has been a popular choice.**

$$P^+ = \frac{S^+ \pi}{S^+ \pi + (1 - S^-)(1 - \pi)}$$

$$P^+ > \pi \Leftrightarrow J > 0$$

$$P^+ = \pi \Leftrightarrow J = 0$$

J is a diagnostic measure

	D=+	D=-
T=+	a	b
T=-	c	d

$$J = S^{+} + S^{-} - 1$$

$$= \frac{a}{a+c} + \frac{d}{b+d}$$

$$= \frac{ad - bc}{(a+c)(b+d)}$$

$$OR = \frac{ad}{bc}$$

$$J = 0 \Leftrightarrow OR = 1$$

J is a measure of Independence (on additive scale)

Prevalence Surveys

DESIGN #1

- This is the most simple scenario/design
- We have a screening test T; its sensitivity S^+ and specificity S^- have been independently established.
- A "prevalence survey" is conducted in one target population in order to estimate the disease prevalence, $\pi = Pr(D=+)$.
- Data: x of n subjects found "positive" by the test.

Simple Solution?

It seems a simple solution is to estimate the disease prevalence by the frequency of positive tests: $p_t = x/n$ – ignoring its errors.

This is a good estimate but it is an estimate of $\pi_t = Pr(T=+)$, the "response rate" whereas we want to estimate the disease prevalence, $\pi = Pr(D=+)$. What is the difference?

THE BIAS

$$\pi_t = \Pr(T=+) = \Pr(T=+, D=+) + \Pr(T=+, D=-)$$

$$\pi_t = \Pr(T=+ \mid D=+)\Pr(D=+) + \Pr(T=+ \mid D=-)\Pr(D=-)$$

$$\pi_t = S^+\pi + (1-S^-)(1-\pi)$$

$$= \pi + (1-S^-)(1-\pi) - (1-S^+)\pi$$

$$= E(p_t)$$

$$\textbf{Bias} = \mathbf{E(p_t) - \pi}$$

$$= \boldsymbol{\pi_t - \pi}$$

$$= \mathbf{(1-S^-)(1-\pi) - (1-S^+)\pi}$$

A POINT ESTIMATE, p

$$\pi_t = \Pr(T = +) = \Pr(T = +, D = +) + \Pr(T = +, D = -)$$

$$\pi_t = \Pr(T = + \mid D = +)\Pr(D = +) + \Pr(T = + \mid D = -)\Pr(D = -)$$

$$\boldsymbol{\pi_t = S^+\pi + (1 - S^-)(1 - \pi)}$$

$$\boldsymbol{\pi = \frac{\pi_t + S^- - 1}{J}}; J = S^+ + S^- - 1, \text{leading to}$$

$$\boldsymbol{p = \frac{p_t + S^- - 1}{J}}$$

J is the Youden's Index

A correction, using p instead of p_t , is a substantial improvement; if S^+ and S^- are known apriori, then p is <u>unbiased</u> for π.

$$\pi = \frac{\pi_t + S^- - 1}{J}; J = S^+ + S^- - 1$$

$$p = \frac{p_t + S^- - 1}{J}$$

$$\mathbf{E(p) = \frac{\pi_t + S^- - 1}{J}}$$

$$\mathbf{= \pi}$$

STANDARD ERROR, SE(p)

$$p = \frac{p_t + S^- - 1}{J}$$

$$Var(p) = \frac{Var(p_t)}{J^2}$$

$$\mathbf{SE(p)} = \frac{\mathbf{1}}{\mathbf{J}} \sqrt{\frac{\mathbf{p_t(1-p_t)}}{\mathbf{n}}}$$

DESIGN #2

- This is still a simple scenario/design
- We have a screening test T; but its sensitivity S^+ and specificity S^- are <u>not</u> known; this is the only difference from Design #1.
- A "prevalence survey" is conducted in <u>one target population</u> in order to estimate the disease prevalence, $\pi = Pr(D=+)$.
- <u>Data</u>: x of n subjects found "positive" by the diagnostic test.

Sensitivity and specificity need to be estimated but there are not enough data from this "prevalence survey" to do so.

Sensitivity and specificity are estimated using two other independent samples; S^+ is estimated by the proportion s^+ from a sample of size n_1 , and S^- is estimated by the proportion s^- from a sample of size n_0.

$$p = \frac{p_t + s^- - 1}{s^+ + s^- - 1}$$

We use the same estimator "p" of Design #1 with S^+ and S^- being estimated from two (2) independent samples by proportions s^+ and s^-

Using Taylor series expansion, we can approximate Expected Value and Variance of estimator p

$$E(p) = \pi + \frac{\pi}{J^2}\frac{S^+(1-S^+)}{n_1} - \frac{(1-\pi)}{J^2}\frac{S^-(1-S^-)}{n_0}$$

$$Var(p) = \frac{1}{J^2}\frac{p_t(1-p_t)}{n} + \frac{\pi^2}{J^2}\frac{S^+(1-S^+)}{n_1} + \frac{(1-\pi)^2}{J^2}\frac{S^-(1-S^-)}{n_0}$$

SOURCES OF BIAS

When sensitivity and specificity are unknown and are estimated using two other independent samples, "p" is no longer unbiased; as seen from the last two terms of the following formula, the bias come from the estimation of the sensitivity and specificity. However, the bias is negligible if the other two samples n_1 and n_0 are both large.

$$E(p) = \pi + \frac{\pi}{J^2}\frac{S^+(1-S^+)}{n_1} - \frac{(1-\pi)}{J^2}\frac{S^-(1-S^-)}{n_0}$$

SOURCES OF VARIABILITY

- **The first term of the variance due to the prevalence survey itself.**
- **The last two terms due to our need of estimating sensitivity and specificity.**

$$\mathrm{Var}(p) = \frac{1}{J^2}\frac{p_t(1-p_t)}{n} + \frac{\pi^2}{J^2}\frac{S^+(1-S^+)}{n_1} + \frac{(1-\pi)^2}{J^2}\frac{S^-(1-S^-)}{n_0}$$

PRIORITIES

- **We already knew that positive predictive value is much more affected by the value of specificity.**
- **Now comparing the last two terms in Var(p); in common cases where both S^+ and S^- are high but π is low, the last term is dominating.**
- **That means the contribution of variability in the estimate of the specificity is usually the dominant term in the calculating the precision of the estimated disease prevalence.**

$$Var(p) = \frac{1}{J^2}\frac{p_t(1-p_t)}{n} + \frac{\pi^2}{J^2}\frac{S^+(1-S^+)}{n_1} + \frac{(1-\pi)^2}{J^2}\frac{S^-(1-S^-)}{n_0}$$

OPTIMAL ALLOCATION

- **The sensitivity and the specificity have different effects on the estimated disease prevalence and its precision.**
- **If we regard the sum $m=n_1+n_0$ as fixed and find the choices n_1 and n_0 which minimize the sum of the last two terms in Var(p); result is:**

$$\frac{n_1}{n_0} \cong \frac{\pi}{1-\pi}\sqrt{\frac{S^+(1-S^+)}{S^-(1-S^-)}}$$

EXAMPLE:

$$\frac{n_1}{n_0} \cong \frac{\pi}{1-\pi}\sqrt{\frac{S^+(1-S^+)}{S^-(1-S^-)}}$$

For example, let $S^+=S^-=.93$, and $\pi = .05$, then n_0 should be 19 times as large as n_1

$$\frac{n_1}{n_0} \cong \frac{\pi}{1-\pi}\sqrt{\frac{S^+(1-S^+)}{S^-(1-S^-)}} = \frac{(.05)}{(.95)}\sqrt{\frac{(.93)(.07)}{(.93)(.07)}} = \frac{1}{19}$$

Note: We usually do not do this!

PREDICTIVE VALUES

- **With disease prevalence π has been estimated, say, using Design 1Plus.**
- **Let turn the focus to the next targets, predictive values P^+ and P^-; we focus on P^+ estimated by p^+.**

$$P^+ = \frac{S^+\pi}{S^+\pi + (1-S^-)(1-\pi)} = \frac{S^+\pi}{\pi_t}$$

$$P^+ = \frac{S^+\pi}{\pi_t}$$

$$p^+ = \frac{s^+ p}{p_t}$$

$$= \frac{s^+}{p_t}\left[\frac{p_t + s^- - 1}{s^+ + s^- - 1}\right]$$

$$= \left(\frac{s^+}{j}\right)\left[\frac{p_t + s^- - 1}{p_t}\right]$$

$$= \left(\frac{s^+}{j}\right)\left[1 - \frac{1 - s^-}{p_t}\right]$$

With S^+ and S^- being estimated from two independent samples by proportions s^+ and s^-

Theorem (Gastwirth, 1987): When n, n_1, and n_0 are all large, the sampling distribution of p^+ is approximately normal with mean P^+ and variance:

$$Var(p^+) = \left\{ \frac{S^+(1-S^-)}{J\pi_t} \right\}^2 \frac{\pi_t(1-\pi_t)}{n}$$

$$+ \left\{ \frac{\boldsymbol{\pi}(1-S^-)}{J\boldsymbol{\pi}_t} \right\}^2 \frac{S^+(1-S^+)}{n_1} + \left\{ \frac{(1-\boldsymbol{\pi})S^+}{J\boldsymbol{\pi}_t} \right\}^2 \frac{S^-(1-S^-)}{n_0}$$

PRIORITIES

- Comparing the last two terms in Var(p^+); in common cases where both S^+ and S^- are high but π is low, the last term is dominating.
- That means the contribution of variability in the estimate of the specificity is usually the dominant term in the calculating the precision of the estimated positive predictive value; need larger n_0. This is very similar to the result concerning precision of the estimated disease prevalence.

OPTIMAL ALLOCATION

- **The sensitivity and the specificity have different effects on positive predictive value and the precision of its estimate.**
- **If we regard the sum $m=n_1+n_0$ as fixed and find the choices n_1 and n_0 which minimize the sum of the last two terms in Var(p^+); result is:**

$$\frac{n_1}{n_0} \cong \frac{\pi}{1-\pi}\sqrt{\frac{(1-S^+)(1-S^+)}{S^+S^-}}$$

EXAMPLE:

$$\frac{n_1}{n_0} \cong \frac{\pi}{1-\pi}\sqrt{\frac{(1-S^+)(1-S^-)}{S^+S^-}}$$

For example, let $S^+=S^-=.93$, and $\pi = .05$, then n_0 should be 253 times as large as n_1

$$\frac{n_1}{n_0} \cong \frac{\pi}{1-\pi}\sqrt{\frac{(1-S^+)(1-S^-)}{S^+S^-}} = \frac{(.05)}{(.95)}\sqrt{\frac{(.07)(.07)}{(.93)(.93)}} \cong \frac{1}{253}$$

Note: Few even think of doing this!

EXAMPLE

- **The ELISA test for AIDS is used to screen donated blood for blood banks.**
- **An evaluation of ELISA yielded estimates (Weiss, 1985): s^+=.977 (using n_1=88) and s^-=.926 (using n_0=297) – not optimal!**
- **Tables 1 and 2 present estimated positive predictive value and its standard error for prevalence surveys with n=500 and n=10,000.**

Table 1 (n=500)

		S's as fixed)	S's as estimates	% of Var(p+) due
Prevalence	p+	SE(p+)	SE(p+)	to estimation of S-
0.50	0.930	0.007	0.017	84.9
0.40	0.898	0.009	0.025	85.2
0.20	0.768	0.024	0.057	82.1
0.10	0.595	0.049	0.103	77.0
0.05	0.410	0.082	0.154	72.0
0.03	0.290	0.106	0.190	69.0
0.01	0.118	0.143	0.243	65.2

Table 2 (n=10,000)

		S's as fixed)	S's as estimates	% of Var(p+) due
Prevalence	p+	SE(p+)	SE(p+)	to estimation of S-
0.50	0.930	0.001	0.016	98.5
0.40	0.898	0.002	0.023	98.9
0.20	0.768	0.005	0.052	99.0
0.10	0.595	0.041	0.091	98.5
0.05	0.410	0.018	0.132	98.1
0.03	0.290	0.024	0.160	97.8
0.01	0.118	0.032	0.199	97.4

RESULTS

- In the two tables we treated the sensitivity and specificity as fixed (col. 3) and as estimated (col. 4).
- The results show that when the prevalence is large, say 40% or above, positive predictive value is high and its standard error is small - even for n=500.
- On the other hand, when prevalence is low, 5% or below, positive predictive value falls below 50% and its standard error gets larger - regardless of the size of the prevalence survey; most of its sampling variability is due to the estimation of the specificity.
- If resources are limited, rather spent to get a good estimate of specificity; the sample size n0 is even more important than the size of the main survey.

Estimating & Modeling ROC Curve

If the idea, in the developmental stage, was to classified people as “diseased” (condition present) or “healthy” (condition absent) based on certain continuous measurement (from blood or urinary components); then we need to “dichotomize” the measurement: for example, if the measurement is “high” then he’s classified as “diseased” – if it’s “low”, the subject is “healthy”. But the basic question is “How high is high?” or “How low is low?”.

A SIMPLE PLAUSIBLE MODEL

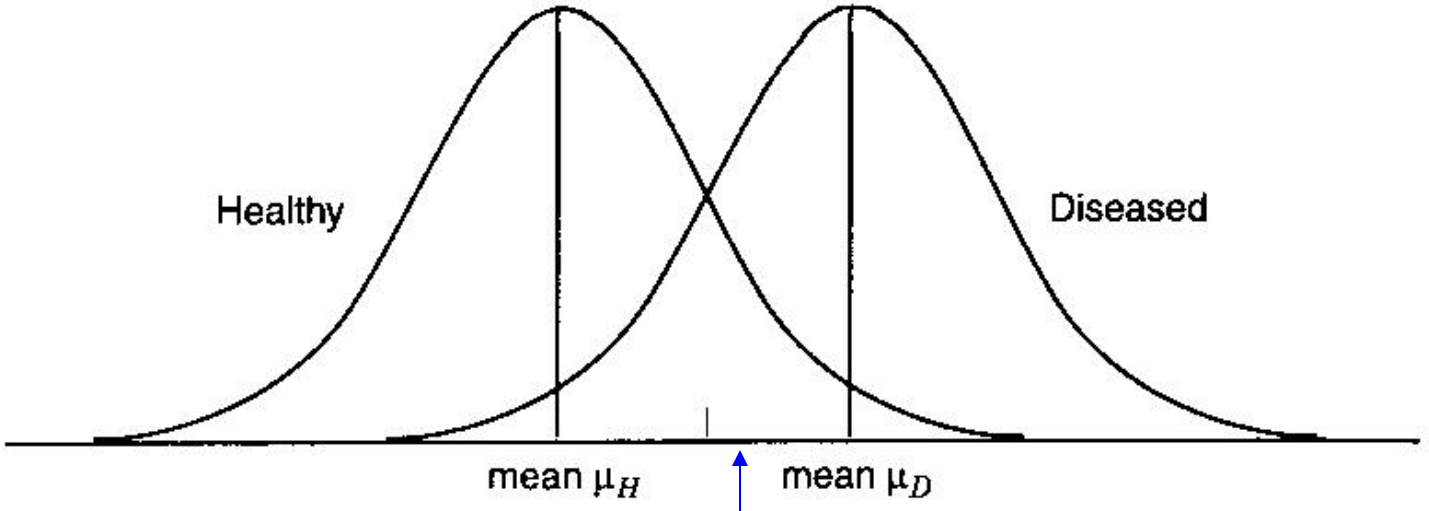


Figure Graphical display of a translational model of diseases.

Separator Y is normally distributed with the same variance, but different means; no matter where you "cut", both errors result! More important, specificity & sensitivity are functions of the "cutpoint" y.

ASSUMPTION

- In the case of many diseases, the larger values of the separator Y are associated with the diseased population (also called “population of the cases”) and smaller values are associated with the control or non-diseased (or healthy) population (e.g. blood glucose for diabetes, PSA for prostate cancer, antibodies for infections),
- For many others, the smaller values of the separator Y are associated with the diseased population and larger values are associated with the non-diseased population (static admittance for Otitis Media, TSH for hyperthyroidism).
- We will assume, without loss of generality, that <u>larger</u> values of Y are associated with the diseased population.

SENSITIVITY

- With our assumption that larger values of Y are associated with the diseased population, the sensitivity, $Pr(T=+|D=+)$, associated with a cutpoint $Y=y$ is:

$$S^+(y) = Pr(Y>y|D=+) = \text{“true positive rate”}$$

$$= 1 - Pr(Y \leq y|D=+) = 1 - F^+(y)$$

- where $F^+(y) = Pr(Y \leq y|D=+)$ is the cumulative distribution function (cdf) of Y for the diseased population (or population of cases).

SPECIFICITY

- With our assumption that larger values of Y are associated with the diseased population, the specificity, $Pr(T=-|D=-)$, associated with a cutpoint $Y=y$ is:

$S^-(y) = Pr(Y \leq y|D=-) = F^-(y)$, or

$1 - S^-(y) = 1 - F^-(y)$ = "false positive rate"

- where $F^-(x)$ is the cumulative distribution function (cdf) of Y for the non-diseased or healthy population.

The sensitivity,
$S^{+}(y) = 1 - F^{+}(y)$,
and the (1-specificity),
$1 - S^{-}(y) = 1 - F^{-}(y)$
are "survival functions".

ROC FUNCTION & ROC CURVE

- A function "R" from [0,1] to [0,1] that "maps" false positive rate to true positive rate, $(1-F^{-}(y))$ to $(1-F^{+}(y))$, is called the "ROC function":

$R[1-F^{-}(y)] = 1-F^{+}(y)$ or $R[1-S^{-}(y)] = S^{+}(y)$

- The graph of R(.) is called the "ROC curve"
- The ROC curve, the graph of sensitivity, $S^{+}(y)$, versus (1-specificity), $(1-S^{-}(y))$, is generated as the "cutpoint" y moves through its range of possible values.

STATISTICAL EXPRESSION

(1-cdf) is called the "Survival Function", S(t); let

$$S_D(t) = 1 - F^+(t)$$

$$S_H(t) = 1 - F^-(t)$$

$$R[S_H(t)] = S_D(t)$$

$$\mathbf{R(u) = S_D[S_H^{-1}(u)];\ u \in [0,1]}$$

THE BINORMAL ROC CURVE

When test result, or separator, Y are assumed to be normally distributed in both diseased and non-diseased populations, we have the so-called "binormal ROC curve"

$$Y_1 \in N(\mu_1, \sigma_1^2), Y_0 \in N(\mu_0, \sigma_0^2)$$

$$ROC(t) = \Phi[\frac{\mu_1 - \mu_0}{\sigma_1} + \frac{\sigma_0}{\sigma_1}\Phi^{-1}(t)]$$

"Slope" is equal to 1 if same variance

LOGISTIC DISTRIBUTION

An alternative to "normal" is the "Logistic Distribution"; the standard logistic curve looks very much like the standard normal curve but with "thicker tails":

Standard Logistic Density:

$$f(x) = \frac{e^x}{(1+e^x)^2}$$

General Logistic Density:

$$f(x) = \frac{\frac{1}{\sigma}\exp(\frac{x-\mu}{\sigma})}{[1+\exp(\frac{x-\mu}{\sigma})]^2}$$

μ is the Mean, σ Standard Deviation

LOG-LOGISTIC DISTRIBUTION

If ln(X) is distributed as logistic, X is distributed as log-logistic; the log-logistic distribution is similar to log-normal distribution but with thicker tails – so fits better "real" non-negative measurements.

$$S(t) = \frac{1}{1+(\rho t)^{\nu}};$$

$$\rho = e^{-\mu}, \text{ where } \mu \text{ is Mean}$$

$$\nu = \frac{1}{\sigma}, \text{ where } \sigma \text{ St Deviation}$$

THE BILOG-LOGISTIC ROC CURVE

$$S(t) = \frac{1}{1 + (\rho t)^{\nu}}$$

$$\sigma_D = \sigma_H = \sigma$$

$$Then:$$

$$\mathbf{R(u)} = \frac{\mathbf{u}}{\mathbf{u} + (\mathbf{1} - \mathbf{u})\mathbf{exp}\left(-\frac{\boldsymbol{\mu}_{\mathbf{D}} - \boldsymbol{\mu}_{\mathbf{H}}}{\boldsymbol{\sigma}}\right)}$$

Issue: **How to estimate the ROC** curve given two independent samples, $\{y_{0i}; i=1,\ldots,n_0\}$ and $\{y_{1j}, j=1, \ldots, n_1\}$ from n_0 controls and n_1 cases?

EMPIRICAL ESTIMATE

- The simplest way to estimate R(.) is to replace cdfs $F^+(y)$ and $F^-(y)$ by their empirical estimates $p^+(y)$ and $p^-(y)$; $p^+(y)$ is the proportion of the n_1 observations y_{1j}'s of the cases which are less than or equal to y, and $p^-(y)$ is defined similarly.
- This is a non-parametric estimate and $\{1-p^-(y), 1-p^+(y)\}$ is an unbiased estimator of $\{1-F^-(y), 1-F^+(y)\}$ but, as Bamber (1975) put it, "the sample ROC (for continuous Y) can never be anything but a finite set of points".
- If there are no ties in the combined sample of y_{0i}'s and y_{1j}'s, there $(n_0 * n_1)$ points.

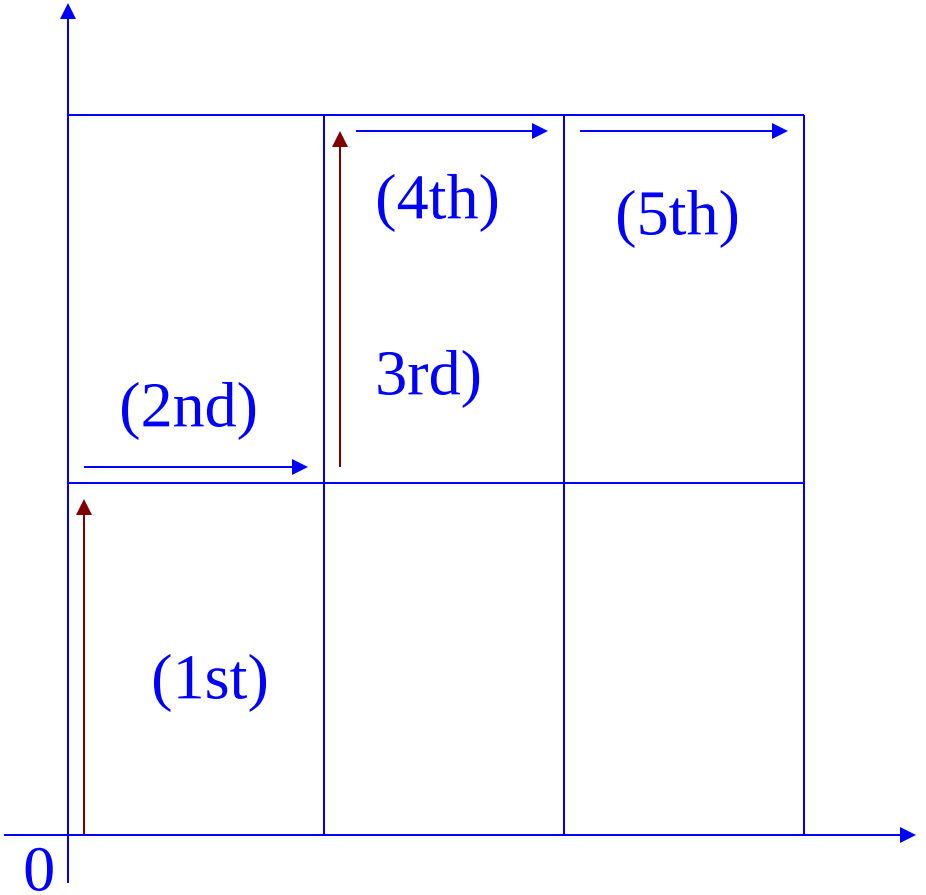
CONNECTING THE DOTS

- Steck (1971) actually made an attempt to connect the dots, turning them into a step function.
- He combined 2 samples & in the usual increasing order.
- He described the empirical estimator as "a random walk from the bottom-left corner (0,0) to the top-right corner (1,1) – and read the combined order sample from largest to smallest - whose next step is $1/n_1$ up or $1/n_0$ to the right according to whether the next observation in the ordered combined sample is a case's measurement (y_1) or a control's measurement (y_0)".

EXAMPLE #1A:

$y_{01} < y_{02} < y_{11} < y_{03} < y_{12}$

Read



It's like an empirical cdf of size 2 with weights 1/2 at points 0 & 1/3; we would read from "small" to "large" if smaller values of Y are associated with the diseased population.

AN ALTERNATIVE VERSION

- **Le (1997) re-formulated the empirical estimator by defining a "score" u_j for each case's observation y_{1j}:**

$$u_j = \frac{n_0 - (S_j - R_j)}{n_0}$$

where R_j is the rank of y_{1j} among the cases and S_j is the rank of y_{1j} among observations in the pooled sample.

- **By treating $\{u_j\}$ as a "pseudo-sample", R(.) is estimated by the empirical step function:**

$$R_{n0}(u) = \frac{\#\text{ of } u_j\text{'s} \le u}{n_0}; \text{ for } 0 \le u \le 1$$

EXAMPLE #1B

- **Each "u" represents the % of control observations greater than that case observation.**
- **Consider the same example as in 1A: $\{y_{01}<y_{02}<y_{11}<y_{03}<y_{12}\}$ with $n_1=2$ and $n_0=3$;**
- **We have $u_1 = 0$, and $u_2 = 1/3$, leading to:**

$$R_2(u) = \begin{cases} 1/2 \text{ for } 0 \le u < 1/3 \\ 1 \quad \text{for } 1/3 \le u \le 1 \end{cases}$$

- **The result is identical to that in Example 1A, an empirical cdf of size 2 with weights 1/2 at points(0,1/3).**

ADVANTAGES OF THE RANK-BASED VERSION

- $R_{n0}(.)$ converges in probability to R(.)
- When there are no ties in the combined sample, it is identical to the empirical estimator. However, the "whole" curve is defined and it has two advantages:

(1) There is an easy option to handle ties: use of "mid-ranks"; can handle graded/ordinal separators.

(2) It provides a "pseudo sample", the $\{u_j$'s$\}$, which can be used for other purposes; for example, comparison of ROC curves

Index for DIAGNOSTIC ACCURACY

- ROC curve is a graphical device to show all possible combinations of sensitivity and specificity but, for simplicity, it is desirable to reduce an entire curve to a single quantitative index of diagnostic accuracy.
- Possibilities include the difference between means of Y for the two populations, those with disease and those without; and the ratio of variances. However, the most popular one has been the area under the ROC curve.
- The area under the curve has a powerful interpretation and it is related to other well-known statistics making it easier to learn its statistical properties.

Suppose that an observation y_1 is randomly sampled from the diseased population and another random observation Y_0 is independently sampled from the non-diseased population; and let $Pr(Y_1>Y_0)$ denote the probability of the event that the Y_1 observation is larger than the Y_0 observation; we have:

$$A = \Pr(Y_1 > Y_0)$$

$$A = \int (1 - F^+) d(1 - F^-)$$

$$A = \int_0^1 R(u) du$$

$$A = \text{Area under ROC curve}$$

INTERPRETATION

- **The area under the ROC curve measures the size and the importance of the difference between two populations, those with the disease and those without it.**
- **It tells how accurately a given diagnostic test differentiates two populations by giving the probability of correct ranking; the probability of separating a case from a control: a measure of "separation power"**
- **If Y_0 and Y_1 are identically distributed, A = 1/2; the maximum value of 1.0 is attained if and only if the cases' distribution lies entirely above that of controls.**

IMPORTANT NOTE

- We can keep the practice of graphing sensitivity versus (1-specificity) in forming the ROC curve
- If the larger values of the separator Y are associated with the diseased population and smaller values are associated with the non-diseased population as we are assuming, the curve is above the main diagonal joining (0,0) and (1,1) and $1/2 \leq A \leq 1$.
- If the smaller values of the separator Y are associated with the diseased population and larger values are associated with the non-diseased population, the curve is below the main diagonal and $0 \leq A \leq 1/2$.

VERSUS **MANN-WHITNEY**'S

- **Given two independent samples, $\{y_{0i}; i=1,\ldots,n_0\}$ and $\{y_{1j}, j=1. \ldots, n_1\}$ from n_0 controls and n_1 cases, there are $n_0\, n_1$ ways of pairing an Y_0 observation with an Y_1 observation.**
- **The Mann-Whitney statistic U counts the number of pairs (y_{0i},y_{1j}), out of $n_0\, n_1$ possible pairs, where $y_{0i} < y_{1j}$; and statistic U is related to Wilcoxon's rank sum W.**
- **From the interpretation of the area "A" under the ROC curve, we have (U is the Mann-Whitney statistics which is related to the Wilcoxon's rank sum W)**

$$\mathbf{A = Pr(Y_1 > Y_0) = U/n_0 n_1}$$

RELATIONSHIP TO WILCOXON'S

- Since the Mann-Whitney statistic U is related to the Wilcoxon's rank sum statistic W, the area A under the ROC curve is related to the statistic W.
- We can also derive such a relationship directly using the area under the $R_{n0}(.)$ curve which is formed from the scores $\{u_j$'s$\}$.

$$\text{Area} = (u_2 - u_1)\frac{1}{n_1} + (u_3 - u_2)\frac{2}{n_1} + \ldots + (u_{n_1} - u_{n_1-1})\frac{n_1 - 1}{n_1} + (1 - u_{n_1})$$

$$A = 1 - \frac{u_1 + u_2 + \ldots + u_{n_1}}{n_1} = 1 - \bar{u}$$

$$u_j = \frac{n_0 - (S_j - R_j)}{n_0} \text{ leading to :}$$

$$\mathbf{A} = \frac{\mathbf{W_1} - \mathbf{(1/2)n_1(n_1 + 1)}}{\mathbf{n_0 n_1}}$$

W_1 is the sum of the ranks for the cases

$$A = U/n_0 n_1$$

$$A = \frac{1}{n_0 n_1}\{W_1 - \frac{1}{2} n_1 (n_1 + 1)\}$$

VARIANCE

- **Var(A) = Var(W_1)/($n_0 n_1$)^2**
- **Few people are familiar with the variance of Wilcoxon's statistic W, because we usually use it only under the Null Hypothesis; but there is a large body of literature dealing with it.**
- **Exact variance and standard error are still complicated, but there are good approximations - both for standard error and confidence intervals.**

"EXACT" STANDARD ERROR

- **The exact standard error of the area A under the ROC curve is given below with:**
- **Q_2 is the probability that y-values for two randomly selected cases will both be smaller than the y-value for a randomly selected control,**
- **Q_1 is the probability that y-value for one randomly selected case will be smaller than both y-values for two randomly selected controls.**
- **Q_1 and Q_2 are estimated by the proportions of triplets satisfying the required properties.**

$$SE(A) = \sqrt{\frac{A(1-A) + (n_0 - 1)(Q_1 - A^2) + (n_1 - 1)(Q_2 - A^2)}{n_0 n_1}}$$

APPROXIMATION BY HANLEY & McNEIL

- Hanley and McNeil (1982), using simulation, showed that standard error of A is only very slightly influenced by the distribution of the separator Y,
- Then they provided, by assuming that X is distributed as “negative exponential”, very simple approximations for Q_0 and Q_1.

$$Q_1 \cong \frac{2A^2}{1+A}$$

$$Q_2 \cong \frac{A}{2-A}$$

CONFIDENCE INTERVALS

- **Using the estimate of either the variance, confidence intervals are formed using asymptotic normality of A.**
- **Ury (1972) suggested an approximate (1-α)100% confidence interval, obtained by applying "Chebyshev inequality", which is as follow:**

$$A \pm \frac{1}{\sqrt{4 \min(n_0, n_1)(1-\alpha)}}$$

EXAMPLE #3

Consider Example #2A with A=.893;

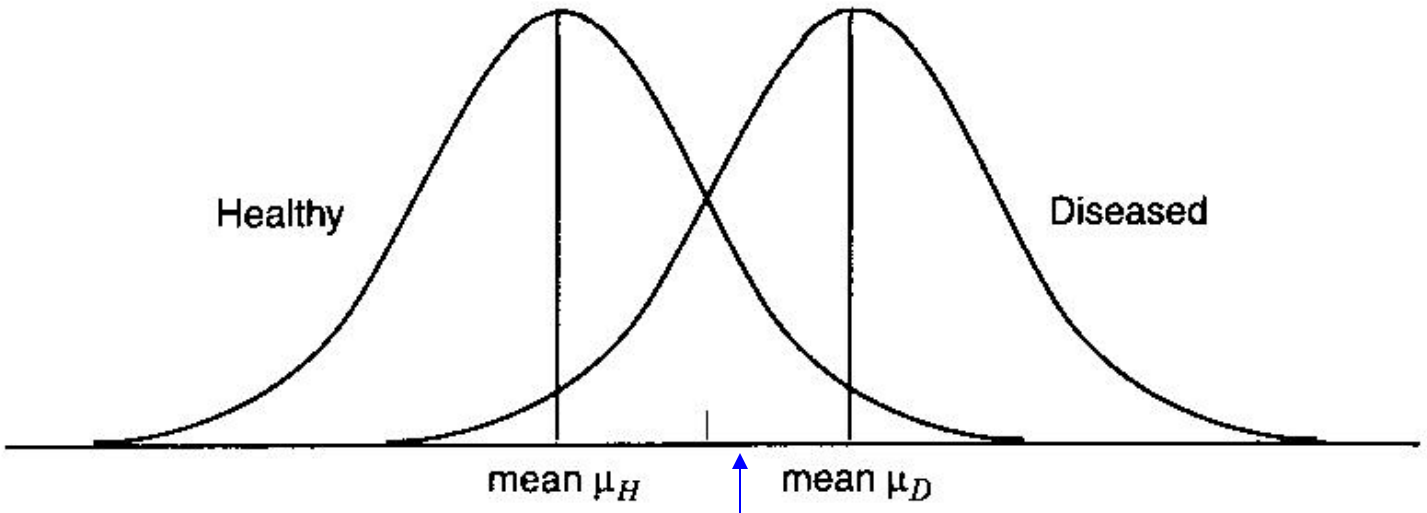
$$Q_1 \cong \frac{2A^2}{1+A} = \frac{(2)(.893)^2}{1+.893} = .843$$

$$Q_2 \cong \frac{A}{2-A} = \frac{.893}{2-.893} = .807$$

$$SE(A) = \sqrt{\frac{(.893)(1-.893)+(58-1)(.843-.893^2)+(51-1)(.807-.893^2)}{(58)(51)}} = .031$$

A 95% confidence interval for A is: .893±(1.96)(.031) =(.832,.954)

AN ALTERNATIVE INDEX



Graphical display of a translational model of diseases.

T=- cutpoint T=+

Separator Y is normally distributed with the same variance, but different means; no matter where you "cut", both errors result!

The sizes of these errors depend on the "standardized distance"

$$\mathbf{d = (\mu_D - \mu_H)/\sigma}$$

THE BILOG-LOGISTIC ROC CURVE

$$S(t) = \frac{1}{1+(\rho t)^{\nu}}$$

$$\sigma_D = \sigma_H = \sigma$$

$$Then:$$

$$\mathbf{R(u)} = \frac{\mathbf{u}}{\mathbf{u} + (\mathbf{1}-\mathbf{u})\mathbf{exp}\left(-\frac{\boldsymbol{\mu}_{\mathbf{D}} - \boldsymbol{\mu}_{\mathbf{H}}}{\boldsymbol{\sigma}}\right)}$$

$$= \frac{\mathbf{u}}{\mathbf{u} + (\mathbf{1}-\mathbf{u})\mathbf{exp}(-\mathbf{d})}$$

Are the two indices, “A” and “d” different?

Yes, different numerical values but “statistically equivalent”. If we let “Φ^{-1}(.) denote “inverse of the standard normal cumulative distribution function”, for example $\Phi^{-1}(.975) = 1.96$, then Simpson and Fitter (1973) showed that:

$$d = \sqrt{2}\,\Phi^{-1}(A)$$

So, what is special about index “d”?

It has a very powerful interpretation in terms of disease development!

LOGISTIC REGRESSION

The probability of disease development and the value Y=y of the separator Y are related by the "Logistic Regression Model":

$$\pi_y = \Pr(D = + \mid Y = y) = \frac{e^{\beta_0 + \beta_1 y}}{1 + e^{\beta_0 + \beta_1 y}}, or$$

$$\ln \frac{\pi_y}{1 - \pi_y} = \beta_0 + \beta_1 y$$

USE OF BAYES' RULE

$$\frac{\pi_y}{1-\pi_y} = \frac{\Pr(D=+\,|\,Y=y)}{\Pr(D=-\,|\,Y=y)}$$

$$\frac{\pi_y}{1-\pi_y} = \frac{\Pr(Y=y\,|\,D=+)\Pr(D=+)/\Pr(Y=y)}{\Pr(Y=y\,|\,D=-)\Pr(D=-)/\Pr(Y=y)}$$

$$\frac{\pi_y}{1-\pi_y} = \frac{\Pr(Y=y\,|\,D=+)\Pr(D=+)}{\Pr(Y=y\,|\,D=-)\Pr(D=-)}$$

$$\mathbf{\ln\frac{\pi_y}{1-\pi_y} = Constant + \ln[\frac{Pr(Y=y\,|\,D=+)}{Pr(Y=y\,|\,D=-)}]}$$

RESULT

Suppose Y is normally distributed with the same variance but different means for Pr(Y=y|D=+) and Pr(Y=y|D=-), we have:

$$\ln \frac{\pi_y}{1-\pi_y} = \text{Constant} + \ln[\frac{\Pr(Y=y \mid D=+)}{\Pr(Y=y \mid D=-)}]$$

$$\ln \frac{\pi_y}{1-\pi_y} = \text{Constant} + \ln[\frac{\exp\{-(y-\mu_D)^2/\sigma^2\}}{\exp\{-(y-\mu_H)^2/\sigma^2\}}]$$

$$\ln \frac{\pi_y}{1-\pi_y} = \text{Constant} + \frac{(\mu_D-\mu_H)}{\sigma^2} y$$

$$\mathbf{d} = \boldsymbol{\beta}_1 \boldsymbol{\sigma}$$

INTERPRETATION OF “d”

Under logistic model and Suppose Y is normally distributed with the same variance but different means for Pr(Y=y|D=+) and Pr(Y=y|D=-), then:

$$d = \beta_1 \sigma$$

The value of Index “d” is equal to the log(Odds Ratio) due to a change of “one SD” in the value of the marker Y

The Optimization Problem

We all know that, for example, high PSA likely indicates prostate cancer; but how high it is to classify a man as having prostate cancer?

If we set the cut-point too high, we would miss cases – that is "low sensitivity"; if we set the cut-point too low, we would have many false positives – that is "low specificity"!

Basic Strategy/Criterion:

To determine an "optimal cutpoint" for a continuous marker by maximizing the "Youden's Index" of the dichotomized test.

Using this strategy, when using the resulting dichotomized test in a prevalence survey we would obtain an estimate with minimal error. There are other gains too.

SOLUTION #1: EMPIRICAL

- Pool the two samples and arrange in increasing order
- At each midway between two data points, calculate the Sensitivity S^+ and Specificity S^-; then the Youden's Index $J = S^+ + S^- - 1$
- Locate the cutpoint corresponding to the maximum value of J.
- Note: It's hard to determine standard error

SOLUTION #2: NON-PARAMETRIC

- The ROC function R(.) maps ($U = 1-S^{-}$) on the horizontal axis to ($V = S^{+}$) on the vertical axis: $V = R(U)$
- The Youden's Index ($J = S^{+} + S^{-} - 1 = R(U) - U$) is maximized when: $0 = R'(U) - 1$, or $R'(U) = 1$.
- **Process:** (i) Smooth empirical estimate by any smoothing technique (eg. Lowess), (ii) Locate the point with (slope = 1) to obtain specificity, then (iii) Go to control sample to get cut-point.
- It may require lots of data & It's still very hard to determine standard error.

Alternative Solution:

Instead of maximizing the Youden Index, one could minimize the distance to the upper left corner (0,1).

Weaknesses:

Characteristics of resulting test is not known and It's hard to determine standard error

SOLUTION #3: SEMI-PARAMETRIC

- **Still looking for the point "on the curve" with (Slope = 1) but, first fitting empirical data with a smooth curve $Y = R(U|\theta)$ because it would take less data to do a better job than nonparametric smoothing (we need a model but can check for goodness-of-fit).**
- **The two components needed are:**
  (i) Choosing a meaningful parameter θ,
  (ii) Choosing a functional form for R(.)

PROPORTIONAL HAZARDS MODEL

Since what we have on the axes of the ROC curve are two survival functions, one possibility is the "Proportional Hazards Model", also called the "Lehmann's Alternatives":

$$1 - F^{+}(y) = (1 - F^{-}(y))^{\theta}, or$$

$$\mathbf{v = R(u) = 1 - (1 - u)^{\theta}; 0 \le u \le 1}$$

$$\text{Area under the ROC curve} : A = \frac{\theta}{\theta + 1}$$

$$\text{Youden's Index} : J(u) = v - u = 1 - u - (1 - u)^{\theta}$$

SUMMARY:
FOUR STEPS FOR SOLUTION #3

(1) Model the ROC function by PHM:
$R(u)=1-(1-u)^{\theta};0 \le u \le 1$

(2) Maximize the Youden's Index:
$J = 1-u-(1-u)^{\theta}$ to obtain $u = F^{-}(y)$

(3) Solve for optimal "cutpoint": $y = (F^{-})^{-1}(u)$

(4) In the result, θ is estimated from $A= \theta/(\theta+1)$; & A is obtained from the Wilcoxon's rank sum statistic.

DETAILS

$$\frac{dJ(u)}{du} = -1 + \theta(1-u)^{\theta-1} = 0$$

$$u = 1 - \frac{1}{\theta^{1/(\theta+1)}}$$

$$\theta = \frac{A}{1-A}$$

$$A = \frac{1}{n_0 n_1}\{W_1 - \frac{1}{2} n_1(n_1+1)\}$$

The value obtained is optimal value for the "cdf" of the control group; knowing the value of u, and having the sample of controls, leads the optimal cutpoint for the marker Y.

First, get the "rank sum" W by SAS, say

Then the Area A under ROC curve

Then θ, parameter of the PHM

Then u which is the (1- cdf) of the controls

Then the optimal cutpoint

Note: Still not easy to obtain standard error, but possible by "Delta method".

MEDICAL IMAGING STUDY

	Rating by Reader					
Disease Status	Def. Normal (1)	Prob. Normal (2)	Questionanle (3)	Prob. Abnormal (4)	Def. Abnormal (5)	Total
Normal	33	6	6	11	2	58
Abnormal	3	2	2	11	33	51
cdf F-	0.569	0.672	0.776	0.966	1	

$$A = Area = .014$$

$$\theta = \frac{A}{1-A} = .014 \text{ approximat ely}$$

$$u = 1 - \frac{1}{\theta^{1/(\theta-1)}} = F^{-}(x) = .986$$

$F^{-}(x)$ is .986; optimal cutpoint is between 4 and 5 (56/58=.966): classify as “abnormal” only those with rating “5” (Resulting test is 99% specific but only 63% sensitive)

SOLUTION #4: PARAMETRIC

$$S_D(t) = 1 - F^+(t) = S^+$$

$$S_H(t) = 1 - F^-(t) = 1 - S^-$$

$$R[S_H(t)] = S_D(t)$$

$$R(u) = S_D[S_H^{-1}(u)]; u \in [0,1]$$

$$J = R(u) - u; u = 1 - S^-$$

$$J' = R'(u) - 1$$

$$= 0 \Leftrightarrow 1 = R'(u)$$

LOG-LOGISTIC DISTRIBUTION

If ln(X) is distributed as logistic, X is distributed as log-logistic; the log-logistic distribution is similar to log-normal distribution but with thicker tails – so fits better "real" non-negative measurements.

$$S(t) = \frac{1}{1+(\rho t)^{\nu}};$$

$$\rho = e^{-\mu}, \text{ where } \mu \text{ is Mean}$$

$$\nu = \frac{1}{\sigma}, \text{ where } \sigma \text{ St Deviation}$$

BOTH LOG-LOGISTIC DISTRIBUTIONS

$$S(t) = \frac{1}{1+(\rho t)^{\nu}}$$

$$\sigma_D = \sigma_H = \sigma$$

$$Then:$$

$$R(u) = \frac{u}{u + (1-u)\exp(-\frac{\mu_D - \mu_H}{\sigma})}$$

$$= \frac{u}{u + (1-u)\beta}$$

$$\beta = \exp\left(-\frac{\mu_D - \mu_H}{\sigma} \right)$$

$$0 < \beta < 1 \text{ for } (\mu_D > \mu_H)$$

$$R(u) = \frac{u}{u + (1-u)\beta}$$

$$R'(u) = \frac{\beta}{[u + (1-u)\beta]^2}$$

$$R'(u) = 1 \Leftrightarrow u = \frac{-\beta + \sqrt{\beta}}{1-\beta}$$

$$Optimal: S^- = 1 - u = S^+ = \frac{1-\sqrt{\beta}}{1-\beta}$$

$$where: \beta = \exp\left(-\frac{\mu_D - \mu_H}{\sigma}\right) = \exp(-d)$$

SCREENING VALUE OF BIOMARKERS

d	S-=S+
1	62%
2	73%
3	82%
4	88%

Evaluation of Screening Tests

EVALUATION

In an "evaluation", we want to see if certain test is "acceptable"; i.e. meeting certain pre-specified criterion. For example, we may want to accept and use only tests with specificity of at least 90%.

Tests' evaluation is simpler when the endpoint is binary and more complicated for continuous endpoints.

A TEST'S ACCEPTABILITY

- For the case of a continuous endpoint, we can also impose a condition; say, to define a "good test" - for a particular disease, as one which detects 90% or more of the diseased while misclassifying no more than 5% of the well.
- Here an optimal dichotomization is not needed; the question is if we could find a cut-point, any cut-point, so that the resulting dichotomized test meets the condition.
- Given data from the developmental stage, we can construct a "z test" as follows.

CRITERION

- Let W_{95} be the 95th percentile of the distribution for controls (or **W**ell) and D_{10} the 10th percentile of the distribution for the cases (or **D**iseased).
- Assume that large values of test associate with disease, like the case of blood glucose.
- Then, the (alternative) hypothesis we wish to test is:

$$\theta = D_{10} - W_{95} > 0$$

RATIONALE

- Let “C” be the cut-point; in order to misclassify no more than 5% of the well , we must have (1) $C \geq W_{95}$
- Similarly, in order to detects 90% or more of the diseased , we must have (2) $C \leq D_{10}$
- In order to satisfy both conditions (1) and (2), we must have $D_{10} \geq W_{95}$ or $D_{10} - W_{95} \geq 0$

Let (μ_0,σ_0) and (μ_1,σ_1) be the mean and standard deviation of the populations of controls and cases, respectively. Then we have:

$$W_{95} = \mu_0 + 1.645\sigma_0$$

$$D_{10} = \mu_1 - 1.282\sigma_1$$

$$\theta = \mu_1 - 1.282\sigma_1 - \mu_0 - 1.645\sigma_0$$

$$\hat{\theta} = \bar{y}_1 - 1.282 s_1 - \bar{y}_0 - 1.645 s_0$$

$$Var(\hat{\theta}) = \frac{\sigma_1^2}{n_1}\{1 + \frac{(1.282)^2}{2}\} + \frac{\sigma_0^2}{n_0}\{1 + \frac{(1.645)^2}{2}\}$$

$$SE(\hat{\theta}) = \sqrt{\frac{s_1^2}{n_1}\{1 + \frac{(1.282)^2}{2}\} + \frac{s_0^2}{n_0}\{1 + \frac{(1.645)^2}{2}\}}$$

$$z = \frac{\hat{\theta}}{SE(\hat{\theta})} \text{ versus } \mathrm{N}(0,1) \text{ under } \mathrm{H}_0$$

Comparison of Screening tests

In a “comparison” we want to see if two tests, usually a new versus a more established one, have the same performance using “statistical test or tests of significance”. The comparison is easy statistically; the more difficult problem is how “express” the “level of difference” if the two screening tests do not have the same performance (i.e. statistical test is significant).

STUDY DESIGNS

- Decisions about which test or tests to recommend for widespread use and which to abandon, assuming that more than one are acceptable, are made on the basis of research studies that compare the accuracies of the tests.
- If each study subject is tested by all tests, we refer to as “paired design”, even more than two tests are under consideration.
- If each study subject is tested by one test, we will refer to the design as “unpaired”.

UNPAIRED DESIGNS

- **Follow the same design principles of multi-arm randomized clinical trials.**
- **Those include well-defined inclusion-exclusion criteria, clear apriori definition of disease and test result - including measurement scale, preparation of study protocol, and randomization to ensure that study arms are balanced with regards to factors affecting test performance and/or result; analysis plan must be in place.**
- **Blinding – if feasible- may be needed to ensure integrity of disease and test assessments.**

DATA LAYOUT

CASES	Test Result		
Test	negative (T=-)	positive (T=+)	Total
A	n10(A)	n11(A)	n1(A)
B	n10(B)	n11(B)	n1(B)
CONTROLS	Test Result		
Test	negative (T=-)	positive (T=+)	Total
A	n00(A)	n01(A)	n0(A)
B	n00(B)	n01(B)	n0(B)

There are four groups of subjects

COMPARISON OF TESTS WITH BINARY ENDPOINT

We can perform two separate Chi-square tests, one for cases and one for controls; for an overall level of α, each test is performed at α /2 (Also, degree of freedom depends on the number of diagnostic tests involved).

MEASURING DIFFERENCES

- **If the difference between two diagnostic tests are found to be significant; the level of difference should be summarized and presented.**
- **The two commonly used parameters are the ratio of two sensitivities (RS^+) and the ratio of two specificities (RS^-); these are ratio of independent proportions, variances are calculated as follows.**

RATIO OF PROPORTIONS

$$r = \frac{p_2}{p_1}$$

$$\ln r = \ln p_2 - \ln p_1$$

$$Var(\ln r) = Var(\ln p_2) + Var(\ln p_1)$$

$$Var(\ln r) \cong \frac{1}{p_2^2}\frac{p_2(1-p_2)}{n_2} + \frac{1}{p_1^2}\frac{p_1(1-p_1)}{n_1}$$

$$\cong \frac{1-p_2}{n_2 p_2} + \frac{1-p_1}{n_1 p_1}$$

RATIOS OF SENSITIVITIES & SPECIFICITIES

$$RS^{+}(A,B) = \frac{S_A^{+}}{S_B^{+}}$$

$$Var\{\ln RS^{+}(A,B)\} = \frac{1-S_A^{+}}{n_1(A)} + \frac{1-S_B^{+}}{n_1(B)}$$

$$RS^{-}(A,B) = \frac{S_A^{-}}{S_B^{-}}$$

$$Var\{\ln RS^{-}(A,B)\} = \frac{1-S_A^{-}}{n_0(A)} + \frac{1-S_B^{-}}{n_0(B)}$$

EXAMPLE

Data from Pepe's book: a randomized study of chronic villus sampling (CVS: Test B) versus early amniocentesis (EA: Test A) for fetus abnormality (Disease D)

CASES	Test Result		
Test	negative (T=-)	positive (T=+)	Total
EA	6	116	122
CVS	13	11	124
CONTROLS	Test Result		
Test	negative (T=-)	positive (T=+)	Total
EA	4844	34	4878
CV	4765	111	4876

TESTS OF SIGNIFICANCE

CASES	Test Result		
Test	negative (T=-)	positive (T=+)	Total
EA	6	116	122
CVS	13	111	124
CONTROLS	Test Result		
Test	negative (T=-)	positive (T=+)	Total
EA	4844	34	4878
CV	4765	111	4876

$$\text{Cases}: \chi^2 = \frac{246\{(6)(111)-(13)(116)\}^2}{(19)(127)(122)(124)} = 4.78$$

$$\text{Controls}: \chi^2 = \frac{9754\{(4844)(11)-(4765)(34)\}^2}{(9609)(145)(4878)(4876)} = 41.54$$

CONFIDENCE INTERVALS

CASES	Test Result		
Test	negative (T=-)	positive (T=+)	Total
EA	6	116	122
CVS	13	111	124
CONTROLS	Test Result		
Test	negative (T=-)	positive (T=+)	Total
EA	4844	34	4878
CV	4765	111	4876

$$RS^{+}(A,B)=\frac{S_A^{+}}{S_B^{+}}$$

$$Var\{\ln RS^{+}(A,B)\}=\frac{1-S_A^{+}}{n_1(A)}+\frac{1-S_B^{+}}{n_1(B)}$$

$$RS^{-}(A,B)=\frac{S_A^{-}}{S_B^{-}}$$

$$Var\{\ln RS^{-}(A,B)\}=\frac{1-S_A^{-}}{n_0(A)}+\frac{1-S_B^{-}}{n_0(B)}$$

$$RS^{+}=\exp\{\ln\frac{.951}{.895}\pm(2.28)\sqrt{\frac{1-.951}{122}+\frac{1-.895}{124}}\}=(.981,1.151)$$

$$RS^{-}=\exp\{\ln\frac{.993}{.977}\pm(2.28\sqrt{\frac{1-.993}{4878}+\frac{1-.977}{4876}}\}=(1.007,1.024)$$

We should use +/-1.96 if each test is at 5%

PAIRED DESIGNS

- **If feasible, paired designs are more desirable.**
- **Most important, only valid if tests do not interfere with each other; be cautious because interference can be subtle.**
- **Also paying attention to cooperation of the subjects; "order" should/may be randomized.**

ADVANTAGES

- More efficient because impact of between-subject variability is minimized.
- Possibilities of confounding are eliminated,
- One can examine characteristics of subjects where tests yield different results; this can lead to insight about test performance and, sometimes, strategies for improving tests.
- One can assess the value of applying combinations of tests compared to single tests.

DATA LAYOUT

CASES		Test A Result		Total
		negative	positive	
Test B Result	negative	a1	b1	n10(B)
	positive	c1	d1	n11(B)
Total		n10(A)	n11(A)	n1
CONTROLS		Test A Result		Total
		negative	positive	
Test B Result	negative	a0	b0	n00(B)
	positive	c0	d0	n01(B)
Total		n00(A)	n01(A)	n0

One group of cases and one group of controls

CASES		Test A Result		Total
		negative	positive	
Test B Result	negative	a1	b1	n10(B)
	positive	c1	d1	n11(B)
Total		n10(A)	n11(A)	n1
CONTROLS		Test A Result		Total
		negative	positive	
Test B Result	negative	a0	b0	n00(B)
	positive	c0	d0	n01(B)
Total		n00(A)	n01(A)	n0

The marginal frequencies for a paired design correspond to the entries for an unpaired design.

COMPARISON OF TESTS

We can perform two separate McNemar's Chi-square tests, one for the set of cases and one for the set of controls.

MEASURING DIFFERENCES

- **If the difference between two diagnostic tests are found to be significant; the level of difference should be summarized and presented.**
- **The two commonly used parameters are still the ratio of two sensitivities (RS^+) and the ratio of two specificities (RS^-). However, these are no longer ratio of independent proportions, the method becomes a little more complicated.**

CASES		Test A Result		Total
		negative	positive	
Test B Result	negative	a1	b1	n10(B)
	positive	c1	d1	n11(B)
Total		n10(A)	n11(A)	n1
CONTROLS		Test A Result		Total
		negative	positive	
Test B Result	negative	a0	b0	n00(B)
	positive	c0	d0	n01(B)
Total		n00(A)	n01(A)	n0

Note: term d_1 are in both numerator and denominator.

$$S_A^+ = n_{11}(A)/n_1$$

$$S_B^+ = n_{11}(B)/n_1$$

$$RS^+ = \frac{S_A^+}{S_B^+} = \frac{b_1 + d_1}{c_1 + d_1}$$

Results: Cheng and Macaluso. Epidemiology 8: 104-106, 1997

CASES		Test A Result		Total
		negative	positive	
Test B Result	negative	a1	b1	n10(B)
	positive	c1	d1	n11(B)
Total		n10(A)	n11(A)	n1
CONTROLS		Test A Result		Total
		negative	positive	
Test B Result	negative	a0	b0	n00(B)
	positive	c0	d0	n01(B)
Total		n00(A)	n01(A)	n0

$$RS^{+}(A,B)=\frac{b_1+d_1}{c_1+d_1};Var\{\ln\ RS^{+}(A,B)\}=\frac{b_1+c_1}{(b_1+d_1)(c_1+d_1)}$$

$$RS^{-}(A,B)=\frac{b_0+a_0}{c_0+a_0};Var\{\ln\ RS^{-}(A,B)\}=\frac{b_0+c_0}{(b_0+a_0)(c_0+a_0)}$$

EXAMPLE

Data from Pepe's book: a paired study of exercise stress test (EST: Test B) versus chest pain history (CPH: Test A) for diagnosing coronary artery (Disease D)

CASES	Test A Result		Total
	negative	positive	
negative	25	183	208
positive	29	786	815
	54	969	1023
CONTROLS	Test A Result		Total
	negative	positive	
negative	151	176	327
positive	46	69	115
	197	245	442

TESTS OF SIGNIFICANCE

CASES	Test A Result		Total
	negative	positive	
negative	25	183	208
positive	29	786	815
	54	969	1023
CONTROLS	Test A Result		Total
	negative	positive	
negative	151	176	327
positive	46	69	115
	197	245	442

$$\text{Cases}: \chi^2 = \frac{(183-29)^2}{183+29} = 111.87$$

$$\text{Controls}: \chi^2 = \frac{(176-46)^2}{176+46} = 76.13$$

CONFIDENCE INTERVALS

CASES	Test A Result		Total
	negative	positive	
negative	25	183	208
positive	29	786	815
	54	969	1023
CONTROLS	Test A Result		Total
	negative	positive	
negative	151	176	327
positive	46	69	115
	197	245	442

$$RS^{+}(A,B)=\frac{183+786}{29+783}=1.189;Var\{\ln RS^{+}(A,B)\}=\frac{183+29}{(183+786)(29+786)}=(.016)^{2}$$

$$RS^{-}(A,B)=\frac{176+69}{46+69}=2.130;Var\{\ln RS^{-}(A,B)\}=\frac{176+46}{(176+69)(46+69)}=(.089)^{2}$$

$$RS^{+}=\exp\{\ln(1.189)\pm(2.28)(.016)\}=(1.146,1.232)$$

$$RS^{-}=\exp\{\ln(2.130)\pm(2.28)(.089)\}=(1.738,2.609)$$

Again, we should use +/- 1.96 if each test is at 5%

Consideration of Subjects' Characteristics

EFFECTS OF COVARIATES

- Various factors can influence the performance of a diagnostic test: the environment in which it is performed, the characteristics of technician, and especially the characteristics of the test subject.
- It may be important to identify and understand the influence of such factors in order to optimize conditions for using a test.
- In general “regression analysis” can be used to make inferences.

EXAMPLES

- **The ability of mammography to detect breast cancer depends on the woman's age; younger women have denser breast which renders the mammogram more difficult to "read" and to interpret.**
- **Men and women differ in their abilities to perform physical exercise; gender should be considered in evaluating exercise stress test.**
- **The experience of the pathologist makes a difference in reading histologic slides.**

MODELING

- If the endpoint of a test is on the continuous scale, we can simply use (regular regression analysis with the usual "Normal Error Regression Model")
- For diagnostic tests with binary endpoint, for example the presence or absence of a bacterium (infections) or some specific DNA sequence (genetic tests), we can "model" the "True Positive Probability" (Pr(T=+|D=+), or sensitivity) or the "False Positive Probability" (Pr(T=+|D=-), or (1-specificity)).
- Note that, in both, the event (T=+) is "random" but the effects of a covariate may be different; for example, a co-morbidity may only affect Pr(T=+|D=+) but not Pr(T=+|D=-).

MODELING FOR BINARY TEST

- We need to model a “probability”, Pr(T=+|D=+) or Pr(T=+|D=-), as a function of one or a set of several covariates – each is binary or continuous.
- Some prefer to model the “log” of the probability as a “linear function” of covariates, some prefer the conventional use of Logistic Regression.

COMBINED MODEL

- When we have common set of covariates, we can consider fitting a composite model to the combined data instead of fitting separate models to Pr(T=+|D=+) and Pr(T=+|D=-).
- The advantage is that one can test if a covariate is common to both models by including an “interaction term” (product of covariate and “D”).
- If so, a combined model requires estimation of fewer parameters leading to greater precision.

EXAMPLE

Suppose X = 0/1 is a binary covariate and D is coded as (+=1,-=0). Then in the model

$$\ln\{\Pr(T=+)\} = \beta_0 + \beta_1 X + \beta_2 D + \beta_3 XD,$$

β_1 represents the effect of X on "false positive probability" whereas ($\beta_1 + \beta_3$) represents the effect of X on "true positive probability" sensitivity); β_2 only tells the difference of test responses from the diseased and the non-diseased populations.

SUGGESTED EXERCISES

#1. Express Kappa statistic as a function of Disease Prevalence, Sensitivity and Specificity.

#2. Refer to Data set on next slide (Prostate Cancer) and focus on "Acid Phosphatase" in blood serum, form the ROC curve and estimate the area under it, including 95% confidence intervals.

#3. Model the ROC curve according to the "proportional hazards model" (also called Lehmann's alternatives):

$$1 - F^{+}(y) = \{1 - F^{-}(y)\}^{\theta}, \text{ or}$$

$$R(u) = 1 - \{1 - u\}^{\theta}; \; 0 \le u \le 1$$

Show how θ is related to the are under the curve. Then show how to determine specificity, sensitivity, and cutpoint Y=y so as to maximize the Youden's Index.

DATA: PROSTATE CANCER

There were 53 patients with prostate cancer; 20 of them with nodal involvement and 33 without. We examined level of acid phosphatase in blood serum (x100). Data are reproduced from Miller et al (1980) and are as follows:

Patients without Nodal Involvement: 40, 40, 46, 47, 48, 48, 49, 49, 50, 50, 50, 50, 50, 52, 52, 55, 55, 56, 59, 62, 62, 63, 65, 66, 71, 75, 76, 78, 83, 95, 98, 102, 187.

Patients with Nodal Involvement: 48(6), 49(9), 51(16), 56(21.5), 67(30), 67(30), 67(30), 70(32,5), 70(32.5), 72(35), 76(37.5), 78(40), 81(41), 82(42.5), 82(42.5), 84(45), 89(46), 99(49), 126(51), 136(52); numbers in parentheses are the ranks in the combined sample, mid-ranks are used for tied observations.

SUGGESTED EXERCISES #4-#6

In a previous study (Anderson et al, 2001) of environmental tobacco smoke, we compared two groups of non-smoking women, $n_1 = 23$ women had male partners who smoke in the home and $n_0 = 22$ women who had male partners who did not smoke. Urine samples were obtained and analyzed and the comparison based on a number of chemicals, among then cotinine (a metabolite of nicotine, in nmol/mL) and NNAL and its glucuronide, NNAL-Gluc (NNAL and NNAL-Gluc are metabolites of the tobacco-specific lung carcinogen called NNK, in pmol/mL). Data (cotinine, NNAL+NNAL-Gluc) are given in the following page (ND is for "not detectable", the limit of detection for cotinine is .003 nmol/mL and for NNAL and NNAL-Gluc is .005 pmol/mL; one case has missing value for NNAL+NNAL-Gluc):

Non-exposed women: (ND,ND), (ND,ND), (ND,ND), (ND,ND), (ND,ND), (ND,.008), (.003,ND), (.003,.015), (.006,ND), (.007,ND), (.007,ND), (.007,.018), (.008,ND), (.008,ND), (.009,ND), (.01,ND), (.012,ND), (.016,ND), (.017,ND), (.019,.047), (.025,ND), and (.03,ND).

Exposed women: (ND,.067), (.003,.009), (.003,.012), (.007,.039), (.008,ND), (.008,.010), (.008,.011), (.009,ND), (.011,.037), (.017,.072), (.018,-), (.021,.083), (.036,.022), (.037,.032), (.042,.063), (.046,ND), (.053,.210), (.076,.041), (.099,.018), (.101,.031), (.111, .018), (.122,.282), and (.200,.027).

#4. Determine the optimal cutpoint for cotinine and the corresponding sensitivity and specificity.

#5. Determine the optimal cutpoint for NNAL+NNAL-Gluc and the corresponding sensitivity and specificity.

#6. Compare the two resulting tests.