Diagnostic screening

Department of Statistics, University of South Carolina

Stat 506: Introduction to Experimental Design

Ties together several things we've discussed already...

- The consideration of dichotomous tests results in a 2×2 table!
- Continuous tests can classify binary outcomes using logistic regression.

Two possibilities: diseased or not diseased

- We assume a state loosely termed **diseased** $D+$ or **not diseased** $D-$, but any event of interest works.
- **Examples**:
  - $D+ =$ cardiovascular disease
  - $D+ =$ hepatitis B
  - $D+ =$ Parkinson's disease
  - $D+ =$ recent use of illegal drugs
- Notice shades of gray and differences in these outcomes.
  - Cardiovascular disease is an umbrella term and can be tested for many different ways: exercise stress test, MRI, X-ray, Echocardiogram, CT scan, PET, SPECT, plus various blood tests. Usually diagnosis takes multiple tests into account.
  - Drug use is known to the person being tested!
  - Hepatitis B is either there or not.

Binary tests

Binary tests: result in one of two outcomes, either $T+$ or $T-$.
Examples:

- over the counter pregnancy tests
- rapid strep test
- cultures (either something grows or it doesn't)
- direct microscopic examination of body fluid (either see it or not)
- asking a potential employee if they've recently used illegal drugs

Continuous tests

Continuous tests: result in a number Y. Typically as the number increases the likelihood of $D+$ increases.

Examples:

- Enzyme-Linked ImmunoSorbent Assay (ELISA) measures an inferred amount of antigen in a blood sample
- minutes of briskly walking on a treadmill before discomfort
- pathologist classifying a slide as (1) negative, (2) atypical squamous hyperplasia, (3) carcinoma *in situ* (not metastasized), (4) invasive carcinoma (metastasized)

Often a continuous test is made into a binary one by *dichotomizing*:

$$T+ \Leftrightarrow Y > k \text{ and } T- \Leftrightarrow Y \leq k.$$

Diagnostic screening

Binary tests

An individual from a population will fall into one of four categories:

$$(D+, T+), (D+, T-), (D-, T+), \text{ or } (D-, T-).$$

These are ‘true positive’, ‘false negative’, ‘false positive’, and ‘true negative’.

Diagnostic screening

Two common measures of *binary* test accuracy are sensitivity and specificity:

$$Se = \Pr\{T+|D+\} \quad Sp = \Pr\{T-|D-\}.$$

- How well does the test do identifying those that really are $D+$? The *sensitivity* of a test, denoted Se, is the probability that a diseased person tests positive.
- How well does the test do identifying those that really are $D-$? The test's *specificity* is the probability that a nondiseased person tests negative.

Note, *gold standard* tests have perfect sensitivity and specificity.
For example, western blot test for HIV; culture for strep.
A measure for dichotomized tests that considers sensitivity and specificity over all possible cutoffs k will be discussed shortly.

Example: Rapid strep test

Sheeler et al. (2002) describe a modest prospective trial of $n = 232$ individuals complaining of sore throat who were given the rapid strep (*streptococcal pharyngitis*) test. Each individual was also given a gold standard test, a throat culture.

	D+	D−	Total
T+	44	4	48
T−	19	165	184
Total	63	169	232

Estimating sensitivity, specificity, and prevalence

	D+	D−	Total
T+	44	4	48
T−	19	165	184
Total	63	169	232

- An estimate of Se is $\widehat{Se} = \widehat{\Pr}\{T + |D+\} = \frac{44}{63} = 0.70$.
- An estimate of Sp is $\widehat{Sp} = \widehat{\Pr}\{T - |D-\} = \frac{165}{169} = 0.98$.
- The estimated prevalence of strep among those complaining of sore throat $\Pr\{D+\}$ is $p = \widehat{\Pr}\{D+\} = \frac{63}{232} = 0.27$.

Odds ratio for these data...

```
m=matrix(c(44,19,4,165),nrow=2)
rownames(m)=c("test.positive","test.negative")
colnames(m)=c("strep","no strep")
m # check that table is correct
fisher.test(m)
```

Odds of strep are 92 times greater when the test comes up positive vs. negative.

PVP

If we have a sore throat, and test positive, we may be interested in the probability we have strep

$$\begin{aligned}
\Pr\{D+|T+\} &= \frac{\Pr\{T+|D+\}\Pr(D+)}{\Pr\{T+|D+\}\Pr\{D+\}+\Pr\{T+|D-\}\Pr\{D-\}} \\
&= \frac{Se \times p}{Se \times p + (1-Sp)\times(1-p)} \\
&\approx \frac{0.70 \times 0.27}{0.70 \times 0.26 + (1-0.98)\times(1-0.27)} \\
&= 0.92.
\end{aligned}$$

This is called the *predictive value positive* (PVP).

PVN

Similarly,

$$\begin{aligned}
\Pr\{D-|T-\} &= \frac{\Pr\{T-|D-\}\Pr(D-)}{\Pr\{T-|D-\}\Pr\{D-\}+\Pr\{T-|D+\}P\{D+\}} \\
&= \frac{Sp\times(1-p)}{Sp\times(1-p)+(1-Se)\times p} \\
&\approx \frac{0.98\times(1-0.27)}{0.98\times(1-0.27)+(1-0.70)\times 0.27} \\
&= 0.90.
\end{aligned}$$

This is called the *predictive value negative* (PVN).

Sensitivity, specificity, PPV, and NPV

- These four numbers summarize how useful a test T is: sensitivity $\Pr\{T+|D+\}$, specificity $\Pr\{T-|D-\}$, positive predictive value $\Pr\{D+|T+\}$ and negative predictive value $\Pr\{D-|T-\}$.
- PPV and NPV are tied to how prevalent $\Pr\{D+\}$ the disease is in the population – useful to an individual.
- *Se* and *Sp* not tied to prevalence. Useful for picking a test in terms of cost of making a mistake.
- We ignored variability here and only reported *point estimates*. How reliable these estimates are depends on how many people were sampled. For example, $\widehat{Se} = 0.70$ but a 95% CI is $(0.57, 0.81)$; that's a large range. Similarly, $\widehat{Sp} = 0.97$ with 95% CI $(0.94, 0.99)$.

Which test is best?

Comparing tests

Say we have two tests, T_1 and T_2, with:

$$Se_1 = 0.8,\ Sp_1 = 0.99,\ Se_2 = 0.99, Sp_2 = 0.8.$$

Which is better?

It depends which is worse: a false negative or a false positive.

- If a false positive is worse – perhaps resulting in unnecessary surgery or a regimen of pharmaceuticals with harmful side effects – then we want the false positive rate to be as small as possible $\Leftrightarrow$ want specificity to be high. Here we'd pick T_1.
- If a false negative is worse – perhaps letting a toxically diseased (think mad cow) proceed to slaughter, or a home pregnancy test – we want the false negative rate to be as small as possible $\Leftrightarrow$ want sensitivity to be high. Here's we'd pick T_2.

How to evaluate continuous tests?

Evaluating continuous tests: ROC Curves
Recall that *dichotomizing* a continuous test Y makes a new binary test T:

$$Y > k \Rightarrow T+ \text{ and } Y \leq k \Rightarrow T-.$$

- Magnitude of the individual test scores ignored ⇒ information loss
- Predictive probability of disease is same for *all* $T+$ (or $T-$) individuals regardless of actual test scores
- Subjects w/ very large scores Y are identical to those barely above the cutoff
- BUT, expect probability of disease to be an increasing function of Y...

Picking one cutoff has implications...

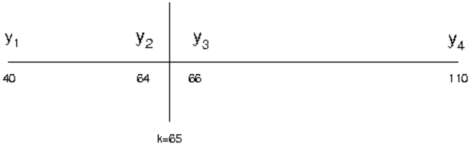

Figure: Four serology scores dichotomized using cutoff $k = 65$.

- Individuals 1 & 2 are $T-$; individuals 3 & 4 are $T+$.
- Individuals 1 and 2 $T-$, test scores differ by 24 units.
  Individuals 3 and 4 $T+$, test scores differ by 44 units.
- Individuals 2 and 3 different although differ by only 2 units.

Underlying densities of Y for diseased and non-diseased

Dichotomizing can oversimplify the analysis but gives easily interpretable parameters: Se, Sp, PVP, and PVN.

Let G_0 and G_1 be distribution of Y from non-diseased and diseased populations.

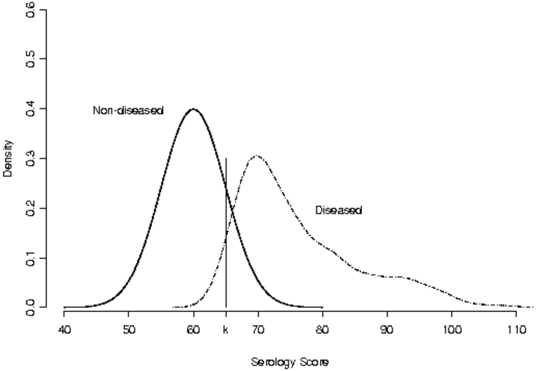

Figure: Cutoff $k = 65$ used to dichotomize continuous serology scores distributed according to G_0 (non-diseased) or G_1 (diseased).

ROC curve

The receiver operator characteristic (ROC) curve plots $(1 - Sp(k), Se(k))$ for all cutoff values k.

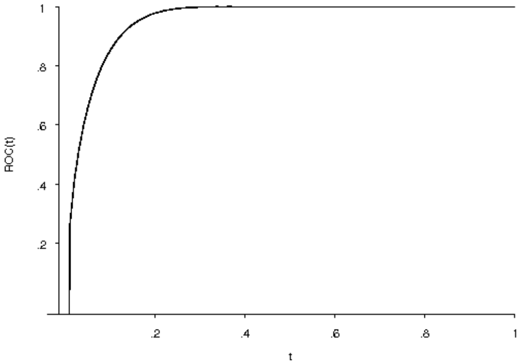

Figure: ROC curve corresponding to the distributions G_0 and G_1.

Overall test accuracy

- ROC curve graphically illustrates a continuous test's Y usefulness in terms of all error rates.
- Good tests have $Se(k)$ close to one and $1 - Sp(k)$ close to 0 for most k – translates into a concave curve with area underneath close to one.
- Area under the curve (AUC) is measure of tests overall diagnostic accuracy. Often reported in publications.
- The AUC is the probability of an infected having a larger Y than a non-infected – for reasonable tests, this should be larger than 0.5.

ROC and logistic regression

- Can use logistic regression to *predict* or *model* $D+$ vs. $D-$ as a function of continuous Y.
- Can have multiple predictors of $D+$ or $D-$, continuous or categorical! Gives one overall "test" predicting $D+$ or $D-$.
- Doesn't necessarily have to be a disease; can be any dichotomous outcome, e.g. "metastasized" vs. "not metastasized", etc.

Esophageal tumor size and metastasis

Recall $n = 31$ patients with esophageal cancer studied; looked at size of patients tumor size Y & whether cancer had spread (metastasized) to lymph nodes ($D+$ or $D-$). Let's see how well tumor size classifies whether the cancer spreads.

```
library(ModelGood) # has Roc function
size=c(6.5,6.3,3.8,7.5,4.5,3.5,4.0,3.7,6.3,4.2,8.0,5.2,
5.0,2.5,7.0,5.3,6.2,2.0,9.0,4.0,3.0,6.0,4.0,4.0,
4.0,5.0,9.0,4.5,3.0,3.0,1.7)
spread= c(1,0,1,1,1,1,0,0,1,1,0,1,1,0,1,0,1,0,1,0,1,1,
0,0,0,1,1,1,0,1,0)
d=data.frame(size,spread)
f=glm(spread˜size,family=binomial,data=d)
plot(Roc(f),auc=T)
```

$T_{1\rho}$ to detect Parkinson's disease

A newly developed continuous measure $T_{1\rho}$ is derived from an MRI scan.

It is postulated that $T_{1\rho}$ is related to neuronal loss. This loss is focused in the substantia nigra part of the brain in Parkinson's disease (PD) patients.

- Case/control study looked at 9 PD patients (PD=1) and 10 controls (PD=0). $T_{1\rho}$ measured on all 19 subjects. (Other covariates also recorded: UPSIT (smell), age, etc.)
- Of interest is to determine if significant differences exist between the PD=0 and PD=1 groups. Dotplot shows $T_{2\rho}$ tends to be higher (more neuronal loss) in PD group.
- *t*-test gives $p = 0.000$ for $H_0 : \mu_0 = \mu_1$: $T_{1\rho}$ values are significantly different in PD=0 and PD=1 groups.

Dichotomizing...

Let's define a formal *binary* test based on $k = 172,500$.

	PD+	PD−	Total
$T_{1\rho}+$	8	1	9
$T_{1\rho}-$	1	9	10
Total	9	10	19

$k = 172,500 \Rightarrow \widehat{Se} = 8/9 \approx 0.89$ and $\widehat{Sp} = 0.90$.

Dichotomizing...

If instead $k = 171,000$ we get

	PD+	PD−	Total
$T_{1\rho}+$	9	1	10
$T_{1\rho}-$	0	9	9
Total	9	10	19

Our estimates change to $\widehat{Se} = 1.00$ and $\widehat{Sp} = 0.90$.

Sensitivity and specificity change with k; a measure that summarizes accuracy over all values of k is the ROC curve and the area under the curve.

$T_{1\rho}$ data

```
pd=c(1,0,1,1,0,1,1,1,1,1,0,0,0,0,0,0,0,0,1)
t1rho=c(178745,165850,182821,172052,172708,176209,174769,174976,
 174655,180869,163760,164660,162285,167675,151261,169693,160504,
 170219,173043)
t2rho=c(63147,67666,64033,59079,73077,61439,63367,64488,67261,
 70754,68670,73119,71357,73881,69354,70111,74136,72173,64101)
plot(t1rho~pd)
MRI=data.frame(pd,t1rho,t2rho)
f=glm(pd~t1rho,family=binomial,data=MRI)
plot(Roc(f),auc=T)
```

Does adding another test help?

Another measure derived from an MRI scan is $T_{2\rho}$ which measures iron content – also linked to Parkinson's disease.

Neither test alone perfectly discriminates PD$=$0 versus PD$=$1; both together do a perfect job, at least on the sample. A linear discriminant rule (i.e. a line) separates the PD$=$0 from the PD$=$1 perfectly.

$T_{1\rho}$ and $T_{2\rho}$ used together...

```
plot(t1rho,t2rho,pch=pd)
legend(152000,65000,legend=c("PD-","PD+"),pch=c(0,1))
MRI=data.frame(pd,t1rho,t2rho)
f=glm(pd˜t1rho+t2rho,family=binomial,data=MRI)
plot(Roc(f),auc=T)
```