

Sections: 2.1.2, 6.2.6

1. Estimating test accuracy: dichotomous tests.
2. Estimating test accuracy: continuous tests.
3. Does adding additional tests help?

- 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: 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: 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*:

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

Binary tests

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

$(D+, T+)$, $(D+, T-)$, $(D-, T+)$, or $(D-, T-)$.

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

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

$$Se = P(T + | D+) \quad Sp = P(T - | D-).$$

- 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

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

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

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

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

Similarly,

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

- These four numbers summarize how useful a test T is: sensitivity $P(T + |D+)$, specificity $P(T - |D-)$, positive predictive value $P(D + |T+)$ and negative predictive value $P(D - |T-)$.
- PPV and NPV are tied to how prevalent $P(D+)$ 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 $(0.94, 0.99)$.

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 .

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 \Rightarrow 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 ...

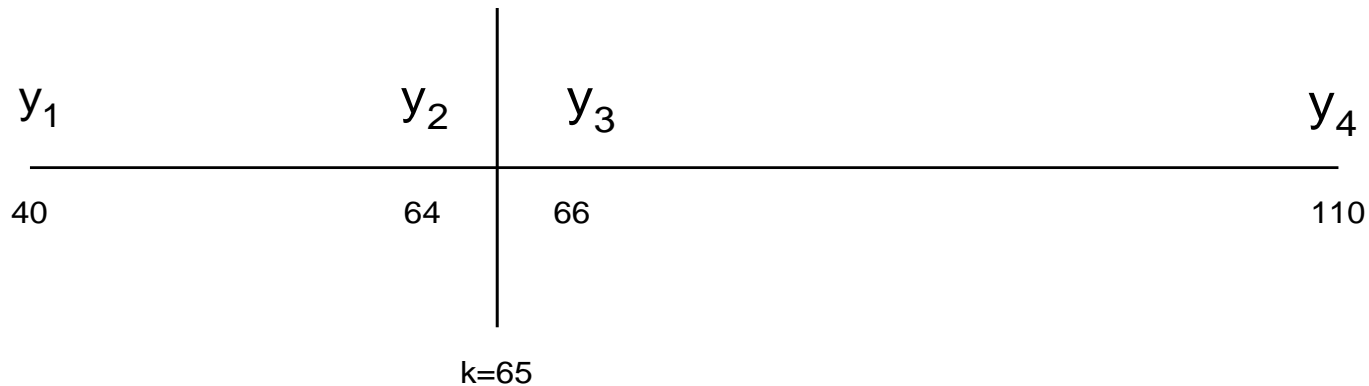


Figure 1: Four serology scores dichotomized using cutoff $k = 65$.

- 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.

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.

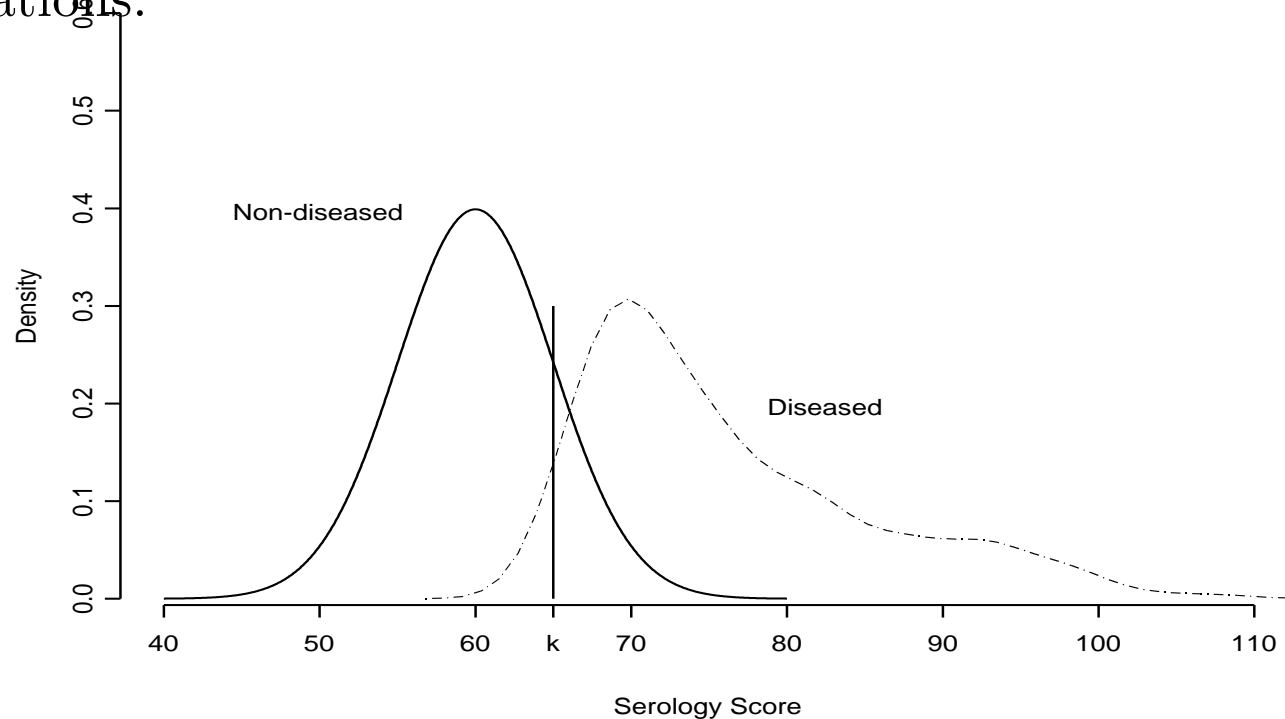


Figure 2: Cutoff $k = 65$ used to dichotomize continuous serology scores distributed according to G_0 (non-diseased) or G_1 (diseased).

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

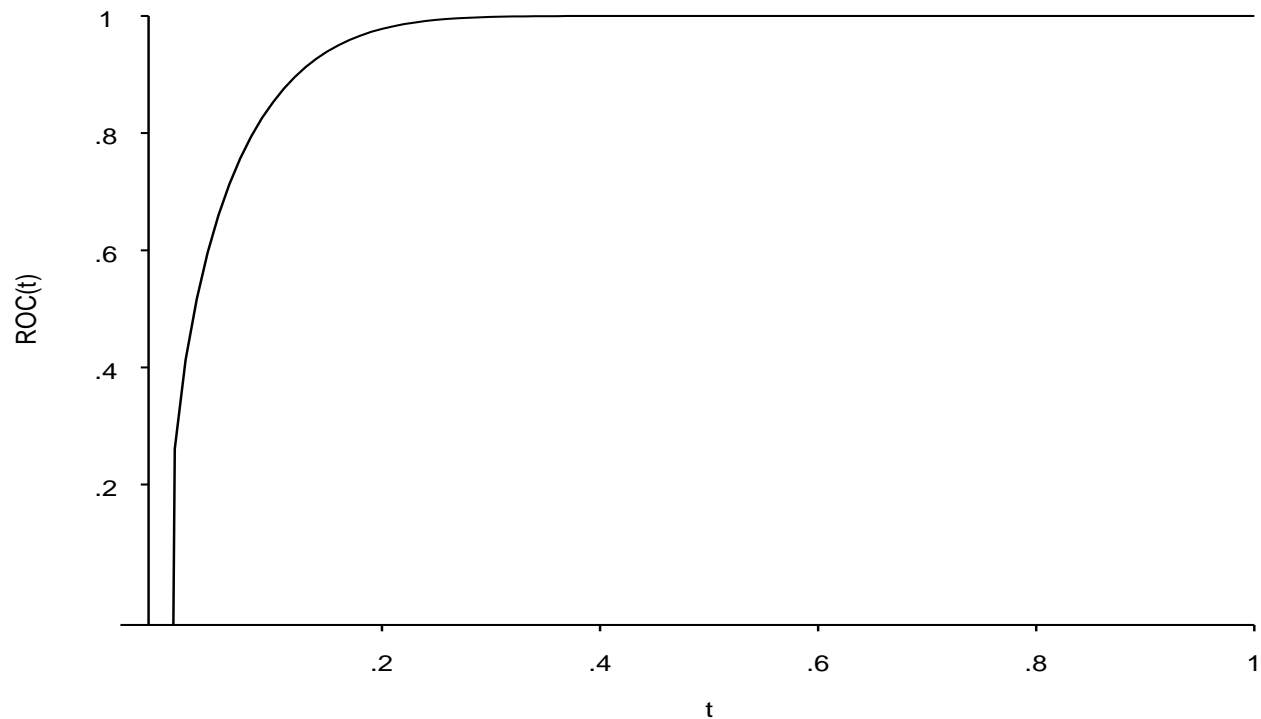


Figure 3: ROC curve corresponding to the distributions G_0 and G_1 .

- ROC curve graphically illustrates a continuous test's X 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.
- $AUC = P(Y^+ > Y^-)$ where $Y^+ \sim G_1$ and $Y^- \sim G_0$.
- Treating Y as continuous means G_0 & G_1 need to be estimated.

Example: 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. Let's look at a dotplot. $T_{1\rho}$ tends to be higher (more neuronal loss) in PD group.
- A t -test gives $p = 0.000$ on 18 df . We strongly reject $H_0 : \mu_0 = \mu_1$. That is, average $T_{1\rho}$ values are different in PD=0 and PD=1 groups.

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 $\widehat{Sp} = 0.90$.

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 $\widehat{Sp} = 0.90$.

These are small sample sizes! Variability? Se CI is $(0.81, 1.00)$, Sp $(0.63, 0.99)$ – almost useless.

Note that: estimates of Se and Sp change w/ k , written $Se(k)$ and $Sp(k)$.

Let's look at the (nonparametric) estimated ROC curve from SAS PROC LOGISTIC. $\widehat{AUC} = 0.99$ w/ CI (0.96, 1.00) – SE obtained from MACRO.

ROC curve interpretation:

- A test that perfectly discriminates between non-diseased and diseased individuals has ROC curve $ROC(t) = 1$ w/ $AUC = 1$. G_0 and G_1 are completely separated.
- Diagnostic tests that are equivalent to a coin toss, and hence are worthless, have $ROC(t) = t$ w/ $AUC = 0.5$. G_0 and G_1 are the same.
- Typically there's overlap between G_0 and G_1 and the ROC will show this. Let's look at normal fits of G_0 and G_1 in each group.

```

data d; input obs pd id t1 t2; datalines;
          1    1   201  178745.0  63147.5
          2    0   202  165850.0  67666.5
          3    1   204  182821.0  64033.5
          4    1   205  172052.5  59079.0
          5    0   206  172708.5  73077.5
          6    1   207  176209.5  61439.5
          7    1   208  174769.0  63367.0
          8    1   211  174976.0  64488.0
          9    1   213  174655.5  67261.5
         10    1   214  180869.0  70754.0
         11    0   215  163760.0  68670.5
         12    0   217  164660.5  73119.0
         13    0   218  162285.5  71357.0
         14    0   220  167675.0  73881.0
         15    0   221  151261.5  69354.0
         16    0   222  169693.0  70111.0
         17    0   223  160504.5  74136.5
         18    0   224  170219.0  72173.0
         19    1   225  173043.0  64101.0
;
ods html; ods graphics on;
proc logistic data=d;
  model pd(event='1')=t1 / lackfit outroc=roc1;
ods graphics off; ods html close;

ods html; ods graphics on;
proc logistic data=d;
  model pd(event='1')=t1 t2 / lackfit outroc=roc1;
ods graphics off; ods html close;

```

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.

A plot of $(T_{1\rho}, T_{2\rho})$ from each individual shows...

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. The second call to PROC LOGISTIC produces AUC estimated to be 1.00.

We may be interested in whether one test is better:

$H_0 : AUC_1 = AUC_2$. There are SAS MACROs available to perform these sorts of tests as well as obtain standard errors and CIs for the AUC.