STATISTICS 305/605: Intro to Biostatistics

October 4, 2021

Lecture 8: Diagnostic Testing

(Reading: PB Section 6.4)

1 Goals of Lecture

- We often use 2×2 tables to measuring properties of tests for diseases
  - There are particular parameters and statistics that are important in this setting
  - We will learn these definitions and see how they are determined and interpreted
- The whole situation is very similar to hypothesis testing!

2 Diagnostic tests

- Many diseases and conditions have tests to confirm presence or absence in an individual.
  - Some are definitive, like a biopsy showing cancer cells
    - (But does failing to find cancer cells in a biopsy *guarantee* no cancer???)
  - Some are uncertain, like using PSA to "test" for Prostate cancer
    - PSA is a protein produced naturally in the prostate
    - Cancer cells often produce excess PSA, which circulates in the blood.
      - Not all cancers do this!
      - Other conditions can produce excess PSA!
- We want tests that give the right answer all the time, but biology is complex
  - Tests are often based on indirect measures (like PSA)
  - Even tests based on direct measures require good sampling (like biopsy or covid swab)
  - Other factors can affect how well the test works

    - A urine-based test for STDs works better on men than on women
- Instead we must live with tests that have *probabilities* of giving the right answer

2.1 Definitions

- We study the most basic and common situation, where the outcome is binary: presence or absence of a condition
  - **Truth**: not known, possibly will never be known unless it can be verified later
  - **Test result**: the binary outcome of a test
    - Called "positive" and "negative"
    - Sometimes based on a numeric measurement with a threshold for positivity (like PSA>3).
- Before a test is used in a public setting, it must be determined to be "accurate"
  - Best understood in the context of a 2×2 table:

		Test Result		
		Positive	Negative	Total
Truth	Present	*True Pos*	*False Neg*	All Present
	Absent	*False Pos*	*True Neg*	All Absent
	Total	All Positive	All Negative	

- "Accurate" means giving the right answer, but this has two inherently different components:
  - If a unit is truly positive, the test should be positive very often
    - It should be *sensitive* to detecting the condition
    - The probability that the test reports positive on a positive person is called SENSITIVITY
  - If a unit is truly negative, the test should be negative very often
    - It should detecting the condition *specifically*, and not other conditions as well
    - The probability that the test reports negative on a negative person is called SPECIFICITY
- These are both conditional probabilities!
  - Sensitivity is *the conditional probability that a test result is positive, given that the true condition is present*
    - Proportion of people with the condition present who test positive
    - True Pos / All Present
  - Specificity is *the conditional probability that a test result is negative, given that the true condition is absent*

    - Proportion of people with the condition absent who test negative
    - True Neg / All Absent

- Ideally, both sensitivity and specificity should be very high.
  - But these are in opposition to each other
  - If you declare everyone positive, then sensitivity is 1, but specificity is 0
  - If you declare everyone negative, then specificity is 1, but sensitivity is 0
  - This is very much the same problem we faced with hypothesis testing with Type 1 and Type 2 errors!

- Often can't estimate test parameters in actual testing practice
  - We don't know the truth!
  - Sometimes can use multiple tests
  - Usually calculated from lab studies where sample properties are known
    - But Lab≠Life
    - Usually describe test *potential* that is not achievable in practice.

- The other important quantity is PREVALENCE of the disease, which is defined as the proportion of the population who have the condition.
  - i.e., the probability that a randomly selected person has "true condition = present".
  - The "population" in question is typically interpreted as "everyone who meets conditions for testing"
    - This prevalence might be different—usually higher—than the general population, because seeking a test is unlikely if you believe you have no reason to.

- Note that sensitivity and specificity are properties of the *test*, conditioned on the truth
  - Changing the prevalence has no effect on these parameters.

2.2 Interpreting diagnostic tests

- What does a positive test mean?
  - Not certain that it means Truth=Present!
  - It means that there is a *probability* that the condition is present
    - This is called POSITIVE PREDICTIVE VALUE
  - A negative test similarly implies a probability that the condition is absent
    - This is called NEGATIVE PREDICTIVE VALUE

- These are more conditional probabilities, but reading the table in the opposite direction!
  - Positive predictive value *the conditional probability that the true condition is present, given that the test is positive*
    * Proportion of people testing positive who actually have the condition present
    * True Pos / All Positive
  - Negative predictive value *the conditional probability that the true condition is absent, given that the test is negative*
    * Proportion of people testing negative who actually have the condition absent
    * True Neg / All Negative
- Positive and negative predictive values are properties of the *prevalence*, conditioned on the test result.
  - They can change dramatically when population prevalence changes!

Estimating prevalence from test data

Importantly, prevalence is usually unknown, and we cannot necessarily judge the prevalence of a disease in the test-seeking population by looking at the proportion of positive tests. However, we can estimate the prevalence using only quantities that we can actually measure: the proportion of positives (say $\hat{p}_+$) and the sensitivity and specificity of the test. This required an application of rules of conditional probability that I don't expect you to be able to duplicate, but the result is

$$\widehat{Prev} = \frac{\hat{p}_+ + spec - 1}{sens + spec - 1}$$

Example: Simulation of positive and negative predictive values (Lecture 8 Scripts.R)

An early study on reverse transcriptase polymerase chain reaction (RT-PCR) tests for SARS-COV2 suggested laboratory values of sensitivity to be about 0.974 and specificity to be 0.985. These are very high numbers, suggesting a remarkably accurate test! I recently checked for current estimates in practice but struggled to find anything clear. According to College of American Pathologists: "clinical performance of testing depends on biology and pre-analytic factors and only approaches 80% sensitivity and 98-99% specificity." So it seems that in real settings, sensitivity is reduced—people who actually have Covid are missed [test negative] due, for example, to failure to obtain virus from the swabbed location. Specificity remains very high, suggesting that nothing else in the body gets mistaken for SARS-COV2.

So suppose you take a test. How should you interpret the results? Let's see what happens if I use the lab values for sensitivity and specificity to see how the prevalence of Covid among test-takers influences the interpretation of the results.

To start with, suppose that 1% of people who get tested for Covid actually have it. Notice that this is WAY higher than what is happening in the population at any given time...at

least I *hope* so! But it is reasonable to think that most of us don't seek a test unless we are sick, have been exposed, or are required to for another reason.

I start by considering a population where 1% of 38 million people are infected. So I have a vector of 380,000 infected (binary =1), and the rest not (binary=0). Just to have a large sample that we can study, I pretend that 100,000 people get tested, so I sample 100,000 of the 0's and 1's. To represent test errors, I "flip" the result for some people at random according to the probabilities (1−sensitivity) if a 1 is sampled, and (1−specificity) if a 0 is sampled. Then I make a table similar to the one above showing the test results against the truth. The results are below.

```
> tab = xtabs(rep(1,samp.size) ~ truth + test)
> tab
     test
truth     0     1
    0 97593  1440
    1    26   941
> (prev.props = rowSums(tab)/100000)
      0       1
0.99033 0.00967
> (test.props = colSums(tab)/100000)
      0       1
0.97619 0.02381
```

Since the prevalence is low, most people have `truth=0`, meaning SARS-COV2 is absent. The `rowSums()` gives the row marginal totals, which I have divided by the total sample size obtained through `sum(tab)`. They shows us that the proportion of sampled people who actually had SARS-COV2 was 0.00967, which rounds to the expected 1% prevalence. However, `colSums()` shows that almost 2.4% of people actually tested positive! Does this mean that the tests didn't "work right"?

We can estimate the test sensitivity and specificity from the conditional probabilities of test outcomes, given true status, found using `prop.table()` as before. For reference, I print the true test sensitivity and specificity below the results.

```
> prop.table(x=tab, margin="truth")
     test
truth          0          1
    0 0.98545939 0.01454061
    1 0.02688728 0.97311272
> c(spec,sens)
[1] 0.985 0.974
```

The estimated sensitivity (conditional proportion of positive tests [`test=1`], given infection [`truth=1`]) is 0.973. The estimated specificity (conditional proportion of negative tests [`test=0`], given no infection [`truth=0`]) is 0.985. These are quite close to what we expect

them to be, so the tests worked exactly as designed. However, a huge fraction of the sample consists of people who had no disease. Even though a tiny fraction of them accidentally test positive, the number testing positive (1440 from the table) drastically increases the proportion of positive tests in the sample.

This is why positive predictive value is so important. If we condition the other way—focusing only on people who tested positive—we can determine what fraction of them actually had the virus. We can compute negative predictive value as well, by conditioning on those who test negative. This is simple in R—we just condition on the `test` margin in `prop.table()`.

```
> prop.table(x=tab, margin="test")
     test
truth             0             1
    0 0.9997336584 0.6047879042
    1 0.0002663416 0.3952120958
```

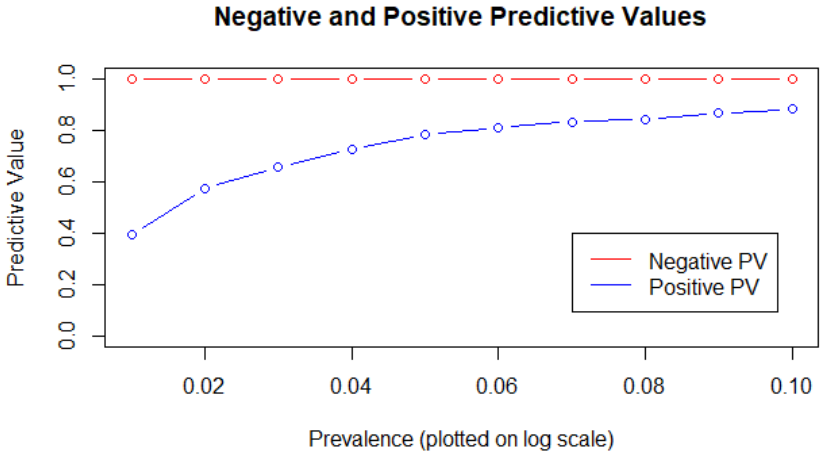
The results are very disappointing. Given a positive test result, the proportion of people who actually had SARS-COV2 in this simulated example was only 39.5%. Testing positive in this scenario identifies more *un*infected people than it does infected people. Again, this is because there were so many uninfected people that the small number of false positives dominated the actual number of people with disease. This phenomenon is likely to occur any time the prevalence among the tested population is very low. On the other hand, the negative predictive value is very, very high. A negative test result leaves one 99.97% confident that they do not have the disease.

What would happen if the actual prevalence were lower or higher? I am no expert, but I have seen proportions of positive tests for Covid reported around 0.05 frequently, with some above and some below. The actual prevalence numbers are probably lower and vary across time. So I reran the calculations of predictive values assuming prevalences 0.01, 0.02, ..., 0.1. The results are in Figure 1.

The negative predictive values remain consistently very high—one can be confident in a negative result—the positive predictive values grow steadily as the true prevalence increases. Intuitively, this makes sense: the more people there are with a disease, the more likely it is that a positive test finds one of them.

I should close this section by pointing out that all of these examples are overly simplistic. In reality, there are many tests for SARS-COV2, and even when based on similar technology, not all perform identically. These tests' properties change when moved from lab to life, and indeed, it is conceivable that different staff are better or worse at obtaining a good sample from a patient. The point of this section and these examples is that you understand the concepts of test performance and interpretation, and how factors beyond our control may influence the interpretation of a test result.

Figure 1: Positive and negative predictive values for various levels of prevalence for a test with sensitivity 0.974 and specificity 0.985.



3 Test Calibration and ROC Curves

- Many diagnostic tests and screens are based on DICHOTOMIZING a test with a numeric result into a binary ("dichotomous") positive/negative decision
  - The PSA$\geq$ 3 screen described previously is like this
  - *Hypothesis tests are like this!*
- The advantage of doing this is that we don't have to think too hard about what a number means
- However, you lose some information about the strength of evidence in either direction
  - PSA= 0.1 and PSA= 2.98 are treated as equivalent
  - PSA= 3.1 and PSA= 12 are treated as equivalent
- *You have to choose a threshold!*
  - Why PSA$\geq$ 3? It used to be 4!
  - **What are the test properties???**
- Different thresholds will exhibit different sensitivities and specificities, and hence different positive and negative predictive values
  - Make the threshold too easy to pass, and there are a lot of false positives, few false negatives
    - High sensitivity (if you have it, it flags you),
    - Low specificity (If you don't have it, it often flags you anyway!)
  - Make the threshold too hard to lass, and the opposite happens

Figure 2: Data and description from a study on choosing a threshold for a screening test for organ rejection following kidney transplant

Consider Table 6.1. This table displays data from a kidney transplant program in which renal allografts were performed [9]. The level of serum creatinine, a chemical compound found in the blood and measured in milligrams percent, was used as a diagnostic tool for detecting potential transplant rejection. An increased creatinine level is often associated with subsequent organ failure.

TABLE 6.1
Sensitivity and specificity of serum creatinine level for predicting transplant rejection

Serum Creatinine (mg %)	Sensitivity	Specificity
1.2	0.939	0.123
1.3	0.939	0.203
1.4	0.909	0.281
1.5	0.818	0.380
1.6	0.758	0.461
1.7	0.727	0.535
1.8	0.636	0.649
1.9	0.636	0.711
2.0	0.545	0.766
2.1	0.485	0.773
2.2	0.485	0.803
2.3	0.394	0.811
2.4	0.394	0.843
2.5	0.364	0.870
2.6	0.333	0.891
2.7	0.333	0.894
2.8	0.333	0.896
2.9	0.303	0.909

Example: Serum Creatinine and Rejection of Kidney Transplant (Lecture 8 Scripts.R)

I don't have a good data set on setting thresholds for a test, so I borrow the one from our book. As Figure 2 explains, serum creatinine (SC) is used as a measure of potential rejection of the organ following kidney transplant. The higher the SC, the greater the danger of rejection. But patients do not line up perfectly with these levels. Some with lower levels experienced rejection while some with higher levels did not (full data not shown).

I assume that if doctors feared organ rejection, they would take some measures to prevent it, where they wouldn't take these steps for patients who were safe from rejection. So we take "high" SC value as a diagnostic screening test for potential rejection. But how high is "high"? The table in Figure 2 shows what would happen if we set the threshold for sample proportions of patients. Estimated sensitivity is taken as the proportion of patients who had SC above this threshold, given that they eventually rejected their organs. Estimated specificity is the proportion of patients who had SC *below* this threshold, given that they eventually *did not reject* their organs.

It is clear that, the higher we set the threshold, the less sensitive the test becomes, because more patients with lower SC eventually rejected organs. But specificity becomes higher, because fewer of the people who never suffered organ rejection are above the higher thresholds.

- How can we use estimated sensitivity and specificity values to identify a good threshold for a test?

  - Although numerous methods have been discussed over the years, there is no single perfect answer.

- Where to set a threshold depends on the relative "costs" of each mistake
  - False positives create needless extra tests and care; anxiety in patient
  - False negatives may cost lives!
  - Requires high-level professional expertise (beyond this course)

- What we *can* do is visualize how the sensitivity changes relative to change specificity

- The standard way to do this is with something called a Receiver-Operator Characteristic (ROC) curve.
  - Plot sensitivity vs. (1−specificity)
  - Fraction of cases correctly detected (true positives) vs. fraction of non-cases wrongly detected (false positives)

- Want large fraction of true cases detected correctly, and small fraction of non-cases detected falsely
  - So want curve that sits far above 1:1 line

- A "perfect" test would detect all cases while finding none falsely, so there would be sensitivity=1 and (1−specificity)=0.
  - "Good" thresholds might be where sensitivity increases slow down considerably as (1−specificity) continues to increase (an "elbow")

Example: Serum Creatinine and Rejection of Kidney Transplant (Lecture 8 Scripts.R)

I don't have a good data set on setting thresholds for a test, so I borrow the one from our book. This data set contains threshold values for SC ranging from 1.2 to 2.9, and has estimated proportions for sensitivity and specificity already computed.

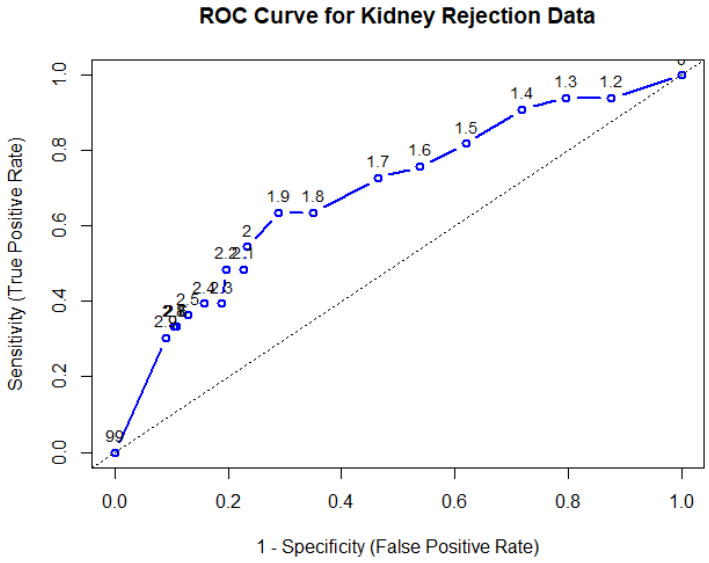
First I enter the data, adding points to anchor the plot at the lower left and upper right corners. Then I plot both the data points and connecting line (`plot(...type="b"...)`). I add a 1:1 line using `abline(a=0, b=1,...)`, which adds a line with intercept=0 and slope=1. Finally, I use `text()` to add the SC values above the points. The full code is below.

```
> # This set did not have values for 0 and 1 for sens and spec
> #   so I am adding them.  This forces plot to connect
> #   to the bottom and top corners.  Assuming that
> #     SC=0 has no positives of any type
> #     SC=99 has all positives
> #
```

```
> SC = c(0,seq(from=1.2, to=2.9, by=0.1),99)
> Sens = c(1,.939,.939,.909,.818,.758,.727,.636,.636,.545,.485,
+          .485,.394,.394,.364,.333,.333,.333,.303,0)
> Spec = c(0,.123,.203,.281,.380,.461,.535,.649,.711,.766,.773,
+          .803,.811,.843,.870,.891,.894,.896,.909,1)
>
> plot(x=1-Spec, y=Sens, type="b", xlim=c(0,1), ylim=c(0,1),
+      col="blue", lwd=2,
+      main="ROC Curve for Kidney Rejection Data",
+      xlab="1 - Specificity (False Positive Rate)",
+      ylab="Sensitivity (True Positive Rate)")
> # Add 1:1 diagonal line for reference
> abline(a=0, b=1, lty="dotted")
> # Add SC values as labels above points using text()
> #  formula=Y~X is where to plot labels.  Can't handle
> #     "1-" in formula so have to create new object named FN
> #  labels=  controls what to print
> #  cex=  is font size relative to standard
> #  pos=3 puts labels above points
> FP = 1-Spec
> text(formula=Sens~FP, labels=SC, cex=0.9, pos=3)
```

The plot produced by this code is in Figure 3. We see that when we move from left to right on the plot—corresponding to decreasing the threshold to make it easier to create positives—the sensitivity (ability to accurately detect potential organ rejections) quickly rises to a point, around SC=1.9, then starts to level off slowly. While it's not a sharp "elbow", it is typical that this is as close to a bend as we can find. The threshold *should* be set based on medical considerations and not purely statistical ones. But this information can help developers to identify good points to create thresholds, and help others to understand the diagnostic risks better.

Figure 3: Data and description from a study on choosing a threshold for a screening test for organ rejection following kidney transplant



4 What to learn from this

1. Diagnostic tests are not perfect instruments and are subject to errors
   (a) False positives when non-cases are detected, relating to "specificity"
   (b) False negatives where actual cases are missed, relating to "sensitivity".
2. Predictive values can help to understand test results
   (a) Prevalence, sensitivity, and specificity all influence this
3. Tests are often developed by creating dichotomous rules around numerical measurements
   (a) Requires sensible thresholds
   (b) ROC curves depict test properties for different thresholds and help to select good thresholds

Table 1: Covid Outcomes by age distribution, through Aug 28, 2021

Table 4: Age distribution: COVID-19 cases, hospitalizations, ICU admissions, deaths, and BC population by age group January 15, 2020 (week 3) – August 28, 2021 (week 34) (N= 166,262)[a]

Age group (years)	Cases n (%)	Hospitalizations n (%) [b]	ICU n (%)	Deaths n (%)	General BC population n (%)
<10	9,903 (6)	104 (1)	8 (<1)	2 (<1)	470,017 (9)
10-19	18,235 (11)	78 (<1)	18 (1)	0 (<1)	529,387 (10)
20-29	38,543 (23)	481 (6)	58 (3)	2 (<1)	699,476 (13)
30-39	31,179 (19)	893 (10)	173 (9)	16 (1)	750,054 (14)
40-49	23,903 (14)	981 (11)	226 (11)	31 (2)	648,377 (12)
50-59	19,959 (12)	1,354 (16)	389 (19)	78 (4)	711,930 (14)
60-69	12,862 (8)	1,629 (19)	495 (25)	179 (10)	686,889 (13)
70-79	6,549 (4)	1,602 (18)	456 (23)	386 (21)	454,855 (9)
80-89	3,525 (2)	1,161 (13)	171 (8)	633 (35)	193,351 (4)
90+	1,604 (1)	408 (5)	18 (1)	495 (27)	52,885 (1)
Total	**166,262**	**8,691**	**2,012**	**1,822**	**5,197,221**
Median age [c]	**34**	**62**	**62**	**84**	**41**

a. Among those with available age information only.
b. Data sources: health authority case line lists and a subset of PHSA Provincial COVID19 Monitoring Solution (PCMS) data for children <20 years of age. PCMS data were included as of June 8 2021. Due to this change in data source, additional admissions that occurred since the start of the pandemic are now included in age groups 0-9 and 10-19 years.
c. Median ages calculated are based on health authority case line lists only.

5 Exercises

Use R for all calculations, unless otherwise specified.

1. Recall the data in Table 1 from the Aug 22–28 Situation Report described previously, and also the coding done in Exercise 1 from Lecture 7. I keep hearing people claim, "Only old people are hospitalized with Covid" as if age were a "test" for whether someone will be hospitalized. Let's examine the properties of a test that used age to "diagnose" the true status of hospitalization.

   From Lecture 7, Exercise 1, we made a table of hospitalization status against age group, so you already have objects counting the number of hospitalized and non-hospitalized cases for each age group. Now we need to dichotomize age group into two "test age" groups, say, "old" and "young". We will define "old" as age ≥ 60, since this still leaves me (barely) hanging on to youth. We need to create a 2×2 table where hospitalization is tabulated against test age. From this table we can compute all of our test property values.

   (a) Start with the 20×3 data frame created in Lecture 7 Exercise 1a. Create one new variable to the data frame called `test.age`, that takes the value 1 if the age group is ≥ 60 and 0 otherwise. You can do this with manual typing using `c()` if you want, since we already have the rest of the data, or you can get clever with `rep()` or `ifelse()`. Use `data.frame(***,test.age)` to add the variable to the previous data frame, where "`***`" is the name of your old data frame. **Print the code you used for the new variable and the resulting 20×4 data frame.**

   (b) Use this data frame to create a cross-tabulation, but this time tabulate the test variable as (Y) and the truth as (X). **Print the `xtab()` results.**

   (c) Refer to the R script, **Lecture 8 Scripts - Analysis Portion.R**. Use this code to estimate the sensitivity and specificity. **Print out code and results, and add an written explanation reporting the sensitivity and specificity**

values. Present them rounded to 3 decimal places. Comment on these values—do they seem good for a diagnostic test?

(d) Estimate the positive and negative predicted values. **Print out code and results, and add an written explanation reporting the two values, correctly labeled. Present them rounded to 3 decimal places. Comment on these values—If your "test" for being old is positive, are you very likely to be hospitalized, and if it is negative, are you very likely to be safe from hospitalization?**

2. BONUS: Notice that we could have set the threshold for "old" vs. "young" in some other place.

   (a) Repeat the process from Exercise 1 above for each of the 9 possible thresholds. Estimate sensitivity, and specificity from each threshold. **Display one printout resembling the table in Figure 2, showing the different thresholds for age, the sensitivity and the specificity for the test using each threshold.**

   (b) **Create an ROC curve.**

   (c) (not marked) Use your best judgment to guess at where the best threshold for this "test" might be.