

DATA605 - Assignment 5

Keith Colella

2023-09-28

```
library(tidyverse)
```

1. (Bayesian)

A new test for multinucleoside-resistant (MNR) human immunodeficiency virus type 1 (HIV-1) variants was recently developed. The test maintains 96% sensitivity, meaning that, for those with the disease, it will correctly report “positive” for 96% of them. The test is also 98% specific, meaning that, for those without the disease, 98% will be correctly reported as “negative.” MNR HIV-1 is considered to be rare (albeit emerging), with about a .1% or .001 prevalence rate. Given the prevalence rate, sensitivity, and specificity estimates, what is the probability that an individual who is reported as positive by the new test actually has the disease? If the median cost (consider this the best point estimate) is about \$100,000 per positive case total and the test itself costs \$1000 per administration, what is the total first-year cost for treating 100,000 individuals?

For the first question, if A is the probability that one has the disease, and B is the probability they test positive, then we are looking for $P(A|B)$. We can calculate this conditional probability as $P(A|B) = P(B|A) * P(A) / P(B)$.

We know that $P(A)$ (the prevalence) is equal to 0.001, that $P(B'|A')$ (the specificity) is equal to 0.98, and that $P(B|A)$ (the sensitivity) is 0.96. We don't yet know the value for $P(B)$, but we can use an alternative definition of the conditional probability that using Bayes Rule:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B|A) * P(A) + (P(B|A') * (P(A'))}$$

We can calculate $P(B|A')$ as $1 - P(B|A)$ and $P(A)$ as $1 - P(A)$. I put this all together below.

```
sensitivity = 0.96
specificity = 0.98
prevalance = 0.001
```

```
(sensitivity * prevalence) / (sensitivity * prevalence + (1 - specificity) * (1 - prevalence))
```

```
## [1] 0.04584527
```

The resulting probability is ~4.58%.

I'll also test this out with a more “empirical” approach and put together a contingency table. We'll assume we're working with a test population of 100,000. We fill values using the probabilities we know ($P(A)$, $P(B|A)$, and $P(B'|A')$), then use the converse of these values to fill in the remainder of the table.

```

n = 100000

contingency <- table(
  factor(levels = c('sick','not sick','margin')),
  factor(levels = c('positive','negative','margin')))
)

contingency['sick', 'margin'] <- prevalence * n
contingency['not sick', 'margin'] <- (1 - prevalence) * n

contingency['sick', 'positive'] <- sensitivity * contingency['sick', 'margin']
contingency['sick', 'negative'] <- (1 - sensitivity) * contingency['sick', 'margin']

contingency['not sick', 'negative'] <- specificity * contingency['not sick', 'margin']
contingency['not sick', 'positive'] <- (1 - specificity) * contingency['not sick', 'margin']

contingency['margin', 'positive'] <- sum(contingency[, 'positive'])
contingency['margin', 'negative'] <- sum(contingency[, 'negative'])

round(contingency)

```

```

##
##           positive negative margin
##  sick           96          4    100
##  not sick      1998      97902  99900
##  margin        2094      97906     0

```

We can then convert these to probabilities (marginal probabilities along the margins, joint probabilities in the top-left square).

```

contingency_freq <- contingency / n

contingency_freq

```

```

##
##           positive negative  margin
##  sick        0.00096  0.00004 0.00100
##  not sick     0.01998  0.97902 0.99900
##  margin       0.02094  0.97906 0.00000

```

Finally, we divide $P(AB)$ (i.e. the joint probability that one is sick and tests positive) by $P(B)$ (i.e. the marginal probabilities of testing positive) to get the probability that one is truly sick *given* that they have tested positive (i.e. $P(A|B)$).

```

contingency_freq['sick','positive'] / contingency_freq['margin','positive']

```

```

## [1] 0.04584527

```

The result matches with ~4.58%.

The second question asks for the expected cost if we test 100,000 individuals. The cost of testing is simply the cost per test (\$1000) times the number of tests (100,000). To get the cost of treatment, there are two approaches. If we assume that anyone who tests positive will require full treatment, then we multiply the *marginal* probability of testing positive ($P(A)$) by the number of individuals tested (100,000).

```
test_cost = 100000 * 1000

treatment_cost1 = contingency['margin', 'positive'] * 100000

total_cost = test_cost + treatment_cost1

cat(
  'Test cost: $', format(test_cost, big.mark = ',', scientific = FALSE),
  '\nTreatment cost: $', format(treatment_cost1, big.mark = ',', scientific = FALSE),
  '\nTotal cost: $', format(total_cost, big.mark = ',', scientific = FALSE)
)
```

```
## Test cost: $ 100,000,000
## Treatment cost: $ 209,400,000
## Total cost: $ 309,400,000
```

However, if we assume that only individuals who are *truly* sick require the full \$100,000 treatment, then we multiply the *conditional* probability of testing positive ($P(A)$) by the number of individuals tested (100,000).

```
treatment_cost2 = contingency['sick', 'positive'] * 100000

total_cost = test_cost + treatment_cost2

cat(
  'Test cost: $', format(test_cost, big.mark = ',', scientific = FALSE),
  '\nTreatment cost: $', format(treatment_cost2, big.mark = ',', scientific = FALSE),
  '\nTotal cost: $', format(total_cost, big.mark = ',', scientific = FALSE)
)
```

```
## Test cost: $ 100,000,000
## Treatment cost: $ 9,600,000
## Total cost: $ 109,600,000
```

2. (Binomial)

The probability of your organization receiving a Joint Commission inspection in any given month is .05. What is the probability that, after 24 months, you received exactly 2 inspections?

```
dbinom(2, size = 24, prob = 0.05)
```

```
## [1] 0.2232381
```

What is the probability that, after 24 months, you received 2 or more inspections?

```
1 - pbinom(2, size = 24, prob = 0.05)
```

```
## [1] 0.1159446
```

What is the probability that you received fewer than 2 inspections?

```
pbinom(1, size = 24, prob = 0.05)
```

```
## [1] 0.6608173
```

What is the expected number of inspections you should have received?

```
n = 24  
p = 0.05  
n*p
```

```
## [1] 1.2
```

What is the standard deviation?

```
sqrt(n*p*(1-p))
```

```
## [1] 1.067708
```

3. (Poisson)

You are modeling the family practice clinic and notice that patients arrive at a rate of 10 per hour. What is the probability that exactly 3 arrive in one hour?

```
lambda = 10  
dpois(3, lambda = lambda)
```

```
## [1] 0.007566655
```

What is the probability that more than 10 arrive in one hour?

```
1 - ppois(11, lambda = lambda)
```

```
## [1] 0.3032239
```

How many would you expect to arrive in 8 hours?

```
lambda * 8
```

```
## [1] 80
```

What is the standard deviation of the appropriate probability distribution?

```
sqrt(lambda)
```

```
## [1] 3.162278
```

If there are three family practice providers that can see 24 templated patients each day, what is the percent utilization and what are your recommendations?

```
current_appointments = 24 * 3  
capacity = lambda * 8 * 3  
current_appointments / capacity
```

```
## [1] 0.3
```

This response assumes there are 8 working hours a day per clinic. Based on that, all three clinics should be able to handle a total of $\lambda * 8 * 3 = 24$ patients a day. This implies a low utilization rate of 30%. I would therefore recommend increasing the number of templated patients per day. The clinics may wish to retain some extra capacity for walk-ins, but they could increase the number of templated patients to 50 per clinic and put the utilization rate at only 62.5%. This level would still keep reasonable capacity for walk-ins.

4. (Hypergeometric)

Your subordinate with 30 supervisors was recently accused of favoring nurses. 15 of the subordinate's workers are nurses and 15 are other than nurses. As evidence of malfeasance, the accuser stated that there were 6 company-paid trips to Disney World for which everyone was eligible. The supervisor sent 5 nurses and 1 non-nurse.

```
x = 5  
m = 15  
n = 15  
k = 6
```

If your subordinate acted innocently, what was the probability he/she would have selected five nurses for the trips?

```
dhyp(x, m, n, k)
```

```
## [1] 0.07586207
```

How many nurses would we have expected your subordinate to send?

```
m*k / (m+n)
```

```
## [1] 3
```

How many non-nurses would we have expected your subordinate to send?

```
n*k / (m+n)
```

```
## [1] 3
```

5. (Geometric)

The probability of being seriously injured in a car crash in an unspecified location is about .1% per hour. A driver is required to traverse this area for 1200 hours in the course of a year.

```
p = 0.001  
n = 1200
```

What is the probability that the driver will be seriously injured during the course of the year?

```
pgeom(q = n, prob = p)
```

```
## [1] 0.6992876
```

In the course of 15 months?

```
pgeom(q = (n/12*15), prob = p)
```

```
## [1] 0.7772602
```

What is the expected number of hours that a driver will drive before being seriously injured?

```
1 / p
```

```
## [1] 1000
```

Given that a driver has driven 1200 hours, what is the probability that he or she will be injured in the next 100 hours?

```
(pgeom(q = 1300, prob = p) - pgeom(q = 1200, prob = p)) / pgeom(q = 1200, prob = p)
```

```
## [1] 0.04094193
```

6. (Binomial)

You are working in a hospital that is running off of a primary generator which fails about once in 1000 hours.

```
p = 1/1000
```

(What is the probability that the generator will fail more than twice in 1000 hours?)

```
1 - pbinom(3, prob = p, size = 1000)
```

```
## [1] 0.01892683
```

What is the expected value?

```
p * 1000
```

```
## [1] 1
```

7. (Uniform)

A surgical patient arrives for surgery precisely at a given time. Based on previous analysis (or a lack of knowledge assumption), you know that the waiting time is uniformly distributed from 0 to 30 minutes. What is the probability that this patient will wait more than 10 minutes?

```
1 - punif(10, 0, 30)
```

```
## [1] 0.6666667
```

If the patient has already waited 10 minutes, what is the probability that he/she will wait at least another 5 minutes prior to being seen?

```
(1 - punif(15, 0, 30)) * punif(10, 0, 30)
```

```
## [1] 0.1666667
```

What is the expected waiting time?

```
(30 + 0) / 2
```

```
## [1] 15
```

8. (Exponential)

Your hospital owns an old MRI, which has a manufacturer's lifetime of about 10 years (expected value). Based on previous studies, we know that the failure of most MRIs obeys an exponential distribution.

```
rate = 1 / 10
```

What is the expected failure time?

```
1 / rate
```

```
## [1] 10
```

What is the standard deviation?

```
1 / rate
```

```
## [1] 10
```

What is the probability that your MRI will fail after 8 years?

```
1- pexp(q = 9, rate = rate)
```

```
## [1] 0.4065697
```

Now assume that you have owned the machine for 8 years. Given that you already owned the machine 8 years, what is the probability that it will fail in the next two years?

```
(pexp(q = 10, rate = rate) - pexp(q = 8, rate = rate)) / pexp(q = 8, rate = rate)
```

```
## [1] 0.1479096
```