Exercise Solutions - Statistical Methods for Diagnostic Studies

Alireza Ghorbani

2025-03-15

Contents

1 Exercise 1 (Predictive Values)

It is estimated that the prevalence of the HIV virus in the heterosexual German population is 0.1% (Dt. rzteblatt 85, Issue 37). An HIV test used for screening has a sensitivity of 98% and a specificity of 99%.

(a) Derive the corresponding 2×2 contingency table (with real disease status in columns and test result in rows) for a sample of 1,000,000 heterosexual German citizens assuming the above probabilities.

(b) Compute the predictive values and interpret them.

Answer:

(a) 2×2 contingency table:

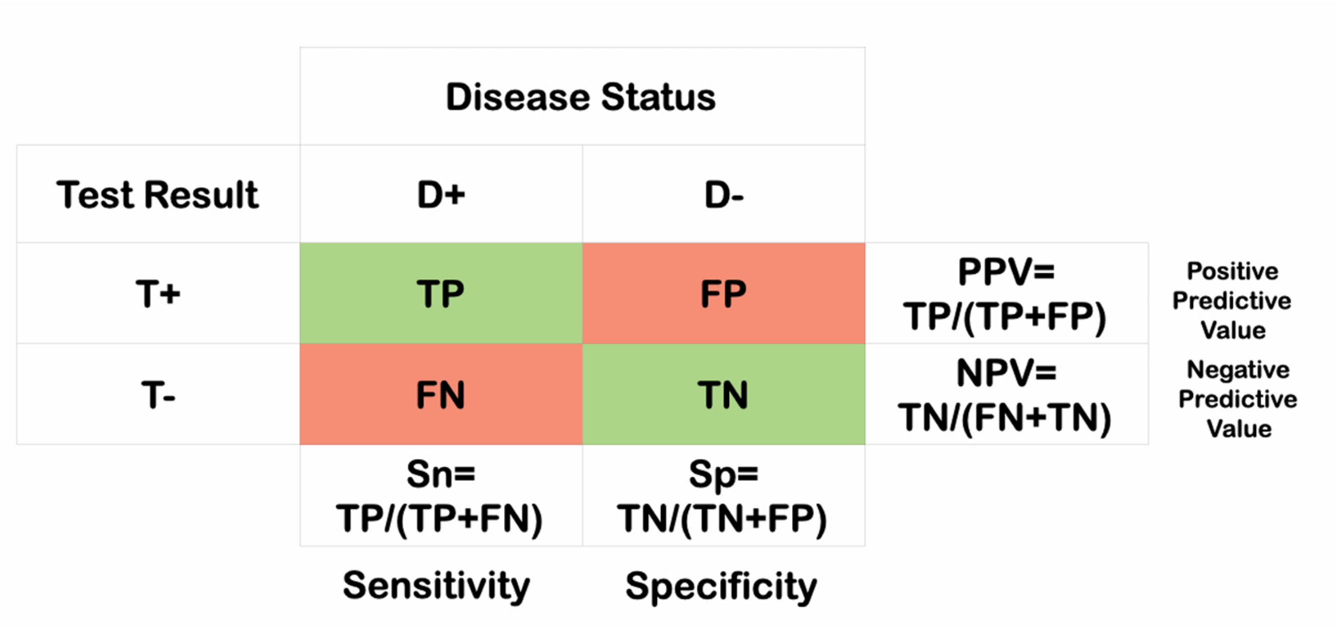
		Disease Status		
		-	+	
Test	-	989010	20	989030
	+	9990	980	10970
		999000	1000	1000000

(b) Positive & Negative predictive values:

$$PPV = P(D^+|T^+) = \frac{TP}{TP+FP} = \frac{980}{10970} = 0.089$$
$$NPV = P(D^-|T^-) = \frac{TN}{TN+FN} = \frac{989010}{989030} = 0.999$$

The probability that a person is actually having a disease while tested positive is 8.9 percent. Despite the high sensitivity and specificity of the test, the low prevalence of HIV causes a high number of false positives. This highlights the importance of confirmatory testing.

Also the probability that a person is not diseased while they're tested negative is almost 1.



	Disease Status			
Test Result	D+	D-		
T+	TP	FP	PPV= TP/(TP+FP)	Positive Predictive Value
T-	FN	TN	NPV= TN/(FN+TN)	Negative Predictive Value
	Sn= TP/(TP+FN)	Sp= TN/(TN+FP)		
	Sensitivity	Specificity		

2 Exercise 2 (TPF/FPF and PPV/NPV)

It can be shown that $PPV = \frac{\rho \times TPF}{\rho \times TPF + (1-\rho) \times FPF}$.

(a) Define all quantities involved in this equality.

(b) Interpret this equality.

Answer:

(a)
- PPV: Positive Predictive Value $P(D^+|T^+)$
- TPF: True Positive Fraction $P(T^+|D^+)$ or sensitivity
- FPF: False Positive Fraction $P(T^+|D^-)$ or 1-specificity
- ρ: prevalence of disease in population

(b) The PPV formula reflects how prevalence (ρ), sensitivity (TPF), and false positive rate (FPF) interact:

Role of Prevalence (ρ):

- High ρ: PPV is dominated by sensitivity (TPF). True positives outweigh false positives.

- Low ρ: PPV is highly sensitive to FPF. False positives overwhelm true positives in rare diseases.

Practical Implications:

- Common disease (ρ high): Prioritize sensitivity (TPF).

- Rare disease (ρ low): Prioritize specificity (minimize FPF).

- Moderate ρ: Balance TPF and FPF improvements.

EXTRA:

The formula for the Negative Predictive Value (NPV) is: $NPV = \frac{(1-\rho) \times TNF}{(1-\rho) \times TNF + \rho \times FNF}$

Quantities involved:

- NPV: Negative Predictive Value $P(D^-|T^-)$
- TNF: True Negative Fraction $P(T^-|D^-)$ (specificity)
- FNF: False Negative Fraction $P(T^-|D^+)$ (1 - sensitivity)
- ρ: Prevalence of the disease in the population

The NPV formula reflects how disease prevalence (ρ), specificity (TNF), and false negative rate (FNF) interact:

Role of Prevalence (ρ)

- High ρ (common disease): NPV **decreases** because the term $\rho \times FNF$ (false negatives) becomes significant.

- Low ρ (rare disease): NPV is dominated by specificity (TNF). True negatives outweigh false negatives, leading to high NPV.

Practical Implications

- Common disease (ρ high): Prioritize sensitivity.
- Rare disease (ρ low): Prioritize specificity.
- Moderate ρ: Balance improvements to sensitivity and specificity.

3 Exercise 3 (Cost)

Consider the following formula:

$$Cost(Screening) = C + \rho * C_D^+ * TPF + \rho * C_D^- * (1 - TPF) + (1 - \rho) * C_{\bar{D}}^+ * FPF$$

$$Cost(NoScreening) = \rho * C_D^-$$

(a) Define all parameters involved in this equality.

(b) Briefly explain what the following statements imply with respect to these parameters and whether they are arguments in favor of testing or no testing.

- The test is very expensive.
- The disease is frequent.
- The test by far does not detect all cases.
- Diseased subjects that are classified as non-diseased ultimately induce very high costs.
- Work-up for subjects testing positive is very expensive.

Answer:

(a)
- C: Cost of the test itself.
- ρ: Prevalence of the disease.
- C_D^+: Cost of treatment for true positives (diseased subjects correctly identified).
- C_D^-: Cost of disease morbidity for false negatives (diseased subjects incorrectly classified as non-diseased).
- $C_{\bar{D}}^+$: Cost of work-up and unnecessary treatment for false positives (non-diseased subjects incorrectly classified as diseased).
- TPF: True Positive Fraction (sensitivity of the test).
- FPF: False Positive Fraction (1 - specificity of the test).

(b) Implications of Statements on Parameters and Testing Decision

1. **"The test is very expensive."**
   Directly increasing C (cost of the test) will impact on $Cost(Screening)$. $Cost(Screening) \uparrow$ due to higher fixed cost of C. This argument is **Against testing** (higher upfront cost).

2. **"The disease is frequent."**
   When ρ (prevalence) increases, impact on costs will be:

   - $Cost(No\,Screening) = \rho \times C_D^- \uparrow$
   - $Cost(Screening) \uparrow$ via $\rho \times [C_D^+ \times TPF + C_D^- \times (1 - TPF) - C_{\bar{D}}^+ * FPF]$

   This argument is **For testing** if $C_D^+ < C_D^-$ (screening reduces untreated disease costs). High ρ favors testing **only if** the cost savings from treating true positives ($C_D^+ \cdot TPF$) outweigh the added costs of testing (C) and false positives ($C_{\bar{D}}^+ \cdot FPF$).

3. **"The test by far does not detect all cases."**

- **Parameter affected:** Reduces TPF (sensitivity).
- **Impact on** $Cost(Screening)$:

$$\rho \times C_D^- \times (1 - TPF) \uparrow \quad \text{(higher false negative costs)}$$

- **Argument: Against testing** (poor sensitivity increases morbidity costs).

4. **"Diseased subjects classified as non-diseased induce very high costs."**

- **Parameter affected:** Increases C_D^- (cost of false negatives).
- **Impact on Costs:**
  - $Cost(No\,Screening) = \rho \times C_D^- \uparrow$
  - $Cost(Screening) \uparrow$ but less severely than no screening if $TPF > 0$

- **Argument: For testing** (avoids worst-case C_D^- under no screening).

5. **"Work-up for subjects testing positive is very expensive."**

- **Parameter affected:** Increases $C_{\bar{D}}^+$ (cost of false positives).
- **Impact on** $Cost(Screening)$:

$$(1 - \rho) \times C_{\bar{D}}^+ \times FPF \uparrow \quad \text{(higher false positive costs)}$$

- **Argument: Against testing** (expensive work-up outweighs benefits).

EXTRA:

The cost is a measures for assessing binary tests, which is less commonly used, but important. This measure balances the two dimensions in an appropriate way considering the substantive context.

We can compare $Cost(Testing)$ to $Cost(NoTesting)$ like mentioned above or, in the case of the comparison of two tests A and B, $Cost(Test_A)$ and $Cost(Test_B)$.

---

4 Exercise 4 (Estimation of TPF/FPF and PPV/NPV)

We consider a diagnostic **case-control study** including 120 early stage multiple sclerosis cases and 240 non diseased controls that aims at evaluating a new blood test for detecting multiple sclerosis for screening purposes in young adults. The data are given in the following table.

	$D = 0$	$D = 1$	
$Y = 0$	172	36	208
$Y = 1$	68	84	152
	240	120	360

(a) Can we obtain valid estimates of TPF, FPF, PPV and NPV using this data? If no, why and which type of study would we need to obtain them? If yes, compute the estimates.

(b) Considering the additional information that the disease prevalence is 1% in the population intended to undergo the test, compute an estimate of PPV.

(c) Derive and calculate a 95% confidence interval for TPF using the normal approximation.

(d) Estimate DLR+ and DLR- and interpret them

Answer:

(a) We can compute the estimates of TPF and FPF. Because this is a case-control study (subjects selected based on disease status). The prevalence in the study does not reflect the true population prevalence. Therefore, the estimates for PPV and NPV cannot be calculated

$$\widehat{\text{TPF}} = \frac{\text{True Positives}}{\text{Total Diseased}} = \frac{84}{120} = 0.7$$

$$\widehat{\text{FPF}} = \frac{\text{False Positives}}{\text{Total Non-Diseased}} = \frac{68}{240} = 0.2833$$

(b)

$$\widehat{\text{PPV}} = \frac{0.01 \times \frac{84}{120}}{0.01 \times 0.7 + 0.99 \times 0.2833} = \frac{0.007}{0.007 + 0.280467} \approx 0.0243$$

So, the estimated PPV considering a disease prevalence of 1% in the population is approximately 2.43%.

(c) The confidence interval

$$\text{CI}_{\text{logit}} = \text{logit}(\widehat{\text{TPF}}) \pm 1.96\sqrt{\frac{1}{n_D \cdot \widehat{\text{TPF}} \cdot (1 - \widehat{\text{TPF}})}}$$

First,

$$\text{logit}(\widehat{\text{TPF}}) = \log\left(\frac{\widehat{\text{TPF}}}{1 - \widehat{\text{TPF}}}\right) = \log\left(\frac{0.7}{1 - 0.7}\right) = \log(2.333) \approx 0.8473$$

Next,

$$\text{SE} = \sqrt{\frac{1}{n_D \cdot \widehat{\text{TPF}} \cdot (1 - \widehat{\text{TPF}})}} = \sqrt{\frac{1}{120 \cdot 0.7 \cdot 0.3}} \approx 0.199$$

Now, calculate the confidence interval on the logit scale:

$$\text{CI}_{\text{logit}} = 0.8473 \pm 1.96 \cdot 0.199 = [0.457, 1.237]$$

Back-transform the confidence interval to the original scale:

$$\text{CI} = \left[\frac{e^{0.457}}{1+e^{0.457}}, \frac{e^{1.237}}{1+e^{1.237}}\right] = [0.612, 0.775]$$

So, the 95% confidence interval for TPF is approximately $[0.612, 0.775]$.

(d)

$$\widehat{\text{DLR+}} = \frac{\widehat{\text{TPF}}}{\widehat{\text{FPF}}} = \frac{0.7}{0.2833} \approx 2.47$$

$$\widehat{\text{DLR-}} = \frac{1-\widehat{\text{TPF}}}{1-\widehat{\text{FPF}}} = \frac{0.3}{0.7167} \approx 0.42$$

$\widehat{DLR+} = 2.47$, A positive test result is approximately 2.47 times more likely in a person with the disease compared to a person without the disease. $\widehat{DLR-} = 0.42$ A negative test result is 0.42 times as likely in a person with the disease compared to a person without the disease. These values suggest that the test is effective in both confirming and ruling out the disease.

EXTRA:

Diagnostic Likelihood Ratios: DLR+ and DLR-

Positive DLR (DLR+):

$$\text{DLR+} = \frac{P[Y=1|D=1]}{P[Y=1|D=0]} = \frac{\text{TPF}}{\text{FPF}}$$

Negative DLR (DLR-):

$$\text{DLR-} = \frac{P[Y=0|D=1]}{P[Y=0|D=0]} = \frac{1-\text{TPF}}{1-\text{FPF}}$$

Measures of Diagnostic Accuracy: Diagnostic Likelihood Ratios

Post-odds and Pre-odds: The pre-test odds are defined as the odds that a subject has disease before the test is performed, i.e., in the absence of test result Y:

$$\text{Pre-test odds} = \frac{P[D=1]}{P[D=0]}$$

The post-test odds are defined as the odds of disease after the test is performed, i.e., with knowledge of the test result:

$$\text{Post-test odds} = \frac{P[D=1|Y]}{P[D=0|Y]}$$

Relationship between DLR, Post-odds, and Pre-odds are as following:

$$\text{Post-test odds (Y = 1)} = \text{DLR+} \times \text{Pre-test odds}$$

$$\text{Post-test odds (Y = 0)} = \text{DLR-} \times \text{Pre-test odds}$$

Thus, the $(\text{DLR+}, \text{DLR-})$ parameters quantify the change in the odds of disease obtained by knowledge of the result of the diagnostic test. They are also called Bayes factors, since the DLR is the Bayesian multiplication factor relating the prior and posterior distributions.

---

5 Exercise 5 (Comparison of Two Tests)

The dataset *lplaudio_b_subset.csv* contains (simulated) data on audiology tests for neonates. For each child, one of the two new hearing tests "A" or "B" was performed. The test type is stored in the variable test (0 standing for A or 1 for B). The variable d indicates whether the child is affected by hearing loss according to the gold standard test (1=yes, 0=no), while y indicates whether the performed test (A or B) was suggestive of hearing loss or not (1=yes, 0=no). The variable sev indicates the severity of the hearing loss. The variable loc indicates whether the test was performed in an isolated sound both (loc=1) or in a normal hospital room (loc=0). The age of the child is also considered as a covariate.

(a) Which type of design is it: paired or unpaired?

(b) Compute estimates for $rTPF(A, B)$ and $rFPF(A, B)$ as well as corresponding confidence intervals. To answer this question you may consult the course materials to find the needed formulas and use R.

(c) Are these confidence intervals valid for a paired design? The paired design is said to be often "more efficient". What does it mean? Explain qualitatively why it is often more efficient.

(d) Specify a log-link regression model that can alternatively be used to obtain estimates of $rTPF(A, B)$ and $rFPF(A, B)$ and to derive their confidence intervals. For this, use the notation xtestB to denote the binary variable equaling 1 if test B and 0 if test A. Fit these models to obtain estimates of $rTPF(A, B)$ and $rFPF(A, B)$ (do not care about confidence intervals) and compare the results to (b).

(e) We now consider the covariate loc in addition to test type. Fit the model

$$logTPF(test, location) = \beta_0 + \beta_1 x_{testB} + \beta_2 x_{loc} + \beta_3 x_{testB} \times x_{loc}$$

for TPF in R and interpret the regression coefficients.

Answer:

(a) Unpaired. Each child received only one test (either A or B), so the results aren't directly compared within the same child.

(b) Relative True Positive Fraction (rTPF) between two tests (A) and (B) and its' variance:

$$rTPF(A, B) = \frac{TPF_A}{TPF_B}, \mathrm{var}(\log rTPF(A, B)) = \frac{1 - TPF_A}{n_{\bar{D}}(A) TPF_A} + \frac{1 - TPF_B}{n_{\bar{D}}(B) TPF_B}$$

Relative False Positive Fraction (rFPF) between two tests A and B and its' variance:

$$rFPF(A, B) = \frac{FPF_A}{FPF_B}, \mathrm{var}(\log rFPF(A, B)) = \frac{1 - FPF_A}{n_{\bar{D}}(A) FPF_A} + \frac{1 - FPF_B}{n_{\bar{D}}(B) FPF_B}$$

Confidence Intervals:

$$\log rFPF(A, B) \pm z_{1-\alpha/2} \sqrt{\mathrm{var}(\log rFPF(A, B))}$$

$$\log rTPF(A, B) \pm z_{1-\alpha/2} \sqrt{\mathrm{var}(\log rTPF(A, B))}$$

```r
# Load data and display the first rows
audiodata = read.csv("lplaudio_b_subset.csv"); head(audiodata,5)
```

```
##   test loc sev d y
## 1    0   0  NA 0 1
## 2    1   1  NA 0 0
## 3    0   0  NA 0 0
## 4    1   0  NA 0 1
## 5    1   1 492 1 1
```

```r
# Transform the binary variables into factors
audiodata$test = as.factor(audiodata$test)
audiodata$loc = as.factor(audiodata$loc)
audiodata$d = as.factor(audiodata$d)
audiodata$y = as.factor(audiodata$y)
audiodataD<-subset(audiodata,d==1); audiodataND<-subset(audiodata,d==0)
nD<-nrow(audiodataD); nND<-nrow(audiodataND)
## rTPF(A,B)
audiodataD_testA<-subset(audiodataD,test==0); TPF_A<-mean(audiodataD_testA$y==1)
audiodataD_testB<-subset(audiodataD,test==1); TPF_B<-mean(audiodataD_testB$y==1)
rTPF_AB<-TPF_A/TPF_B; rTPF_AB
```

```
## [1] 0.8508464
```

```r
# CI for rTPF(A,B)
nD_A<-nrow(audiodataD_testA); nD_B<-nrow(audiodataD_testB)
varlog_rTPF_AB<-(1-TPF_A)/(nD_A*TPF_A)+(1-TPF_B)/(nD_B*TPF_B)
CIlog_rTPF_AB<-c(log(rTPF_AB)-qnorm(0.975)*sqrt(varlog_rTPF_AB),
                 log(rTPF_AB)+qnorm(0.975)*sqrt(varlog_rTPF_AB))
CI_rTPF_AB<-exp(CIlog_rTPF_AB);CI_rTPF_AB
```

```
## [1] 0.7374081 0.9817352
```

```r
# rFPF(A,B)
audiodataND_testA<-subset(audiodataND,test==0); FPF_A<-mean(audiodataND_testA$y==1)
audiodataND_testB<-subset(audiodataND,test==1); FPF_B<-mean(audiodataND_testB$y==1)
rFPF_AB<-FPF_A/FPF_B; rFPF_AB
```

```
## [1] 0.9825919
```

```r
# CI for rFPF(A,B)
nND_A<-nrow(audiodataND_testA); nND_B<-nrow(audiodataND_testB)
varlog_rFPF_AB<-(1-FPF_A)/(nND_A*FPF_A)+(1-FPF_B)/(nND_B*FPF_B)
CIlog_rFPF_AB<-c(log(rFPF_AB)-qnorm(0.975)*sqrt(varlog_rFPF_AB),
                 log(rFPF_AB)+qnorm(0.975)*sqrt(varlog_rFPF_AB))
CI_rFPF_AB<-exp(CIlog_rFPF_AB); CI_rFPF_AB
```

```
## [1] 0.7774384 1.2418820
```

$rTPF_{AB}$ and $rFPF_{AB}$ with 95% CIs show if Test A is better/worse than Test B.

(c) No, these confidence intervals are not valid for a paired design. A paired design is often “more efficient” because it controls for inter-subject variability by comparing the tests within the same subjects. This reduces random variation and provides more precise estimates.

(d)

```
# Models to estimate rTPF(A,B) and rFPF(A,B)
rTPF_model = glm(y ~ test, data = audiodataD, family = binomial(link = "log"))
summary(rTPF_model); exp(-rTPF_model$coefficients[2]) # = value computed in (b)
```

```
##
## Call:
## glm(formula = y ~ test, family = binomial(link = "log"), data = audiodataD)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.52078    0.05816  -8.954   <2e-16 ***
## test1        0.16152    0.07301   2.212   0.0269 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 549.77  on 423  degrees of freedom
## Residual deviance: 544.74  on 422  degrees of freedom
## AIC: 548.74
##
## Number of Fisher Scoring iterations: 5
```

```
##     test1
## 0.8508464
```

```
rFPF_model = glm(y ~ test, data = audiodataND, family = binomial(link = "log"))
summary(rFPF_model); exp(-rFPF_model$coefficients[2]) # = value computed in (b)
```

```
##
## Call:
## glm(formula = y ~ test, family = binomial(link = "log"), data = audiodataND)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.01160    0.08703 -11.623   <2e-16 ***
## test1        0.01756    0.11948   0.147    0.883
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 637.62  on 484  degrees of freedom
## Residual deviance: 637.60  on 483  degrees of freedom
## AIC: 641.6
##
## Number of Fisher Scoring iterations: 5
```

```
##     test1
## 0.9825919
```

(e)

```r
# Model for rTPF(A,B) with adjustment for loc
rTPF_model_adj = glm(y ~ test*loc, data = audiodataD, family = binomial(link = "log")); summary(rTPF_mod
```

```
##
## Call:
## glm(formula = y ~ test * loc, family = binomial(link = "log"),
##     data = audiodataD)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.55702    0.08812  -6.321  2.6e-10 ***
## test1        0.11921    0.11295   1.055    0.291
## loc1         0.06796    0.11711   0.580    0.562
## test1:loc1   0.08216    0.14726   0.558    0.577
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 549.77  on 423  degrees of freedom
## Residual deviance: 541.51  on 420  degrees of freedom
## AIC: 549.51
##
## Number of Fisher Scoring iterations: 5
```

The coefficients can be interpreted as follows:

- Intercept (β_0): Baseline log risk of a positive test result for Test A in a normal hospital room.
- Test B (β_1): Log risk ratio of a positive test result for Test B compared to Test A.
- Location (β_2): Log risk ratio of a positive test result in an isolated sound booth compared to a normal hospital room.
- Test B * Location (β_3): Interaction effect between test type and location on the log risk of a positive test result.

Non of the coefficients are significant except for intercept.

---

6 Exercise 6 (ROC Curve)

Let Y be a continuous diagnostic test. In the following D refers to the diseased subjects and $\bar{D}$ to the non diseased subjects. Here are the test results of five non-diseased subjects and five diseased subjects:

$$Y_{\bar{D}} : 0, 2, 4, 5, 5$$

$$Y_D : 1, 2, 3, 7, 8$$

(a) Compute the coordinates of the corresponding points of the ROC curve and draw (approximately) the empirical ROC curve, $\widehat{ROC}_e(t)$.

(b) Compute the empirical AUC, $\widehat{AUC}_e$.

(c) Briefly explain the principle of the two other main families of methods for ROC estimation (2-3 sentences each)

(d) Explain the two following formulas (what are they for, how are they used in practice):

For large samples:

$$\text{var}(\widehat{AUC}_e) \approx \frac{var(\widehat{FPF}(Y_{D_i}))}{nD} + \frac{var(\widehat{TPF}(Y_{\bar{D}_j}))}{n\bar{D}}, (1)$$

For unpaired samples:

$$var(\Delta\widehat{AUC}_e) = var(\widehat{AUC}_{A_e}) + var(\widehat{AUC}_{B_e}), (2)$$

Answer:

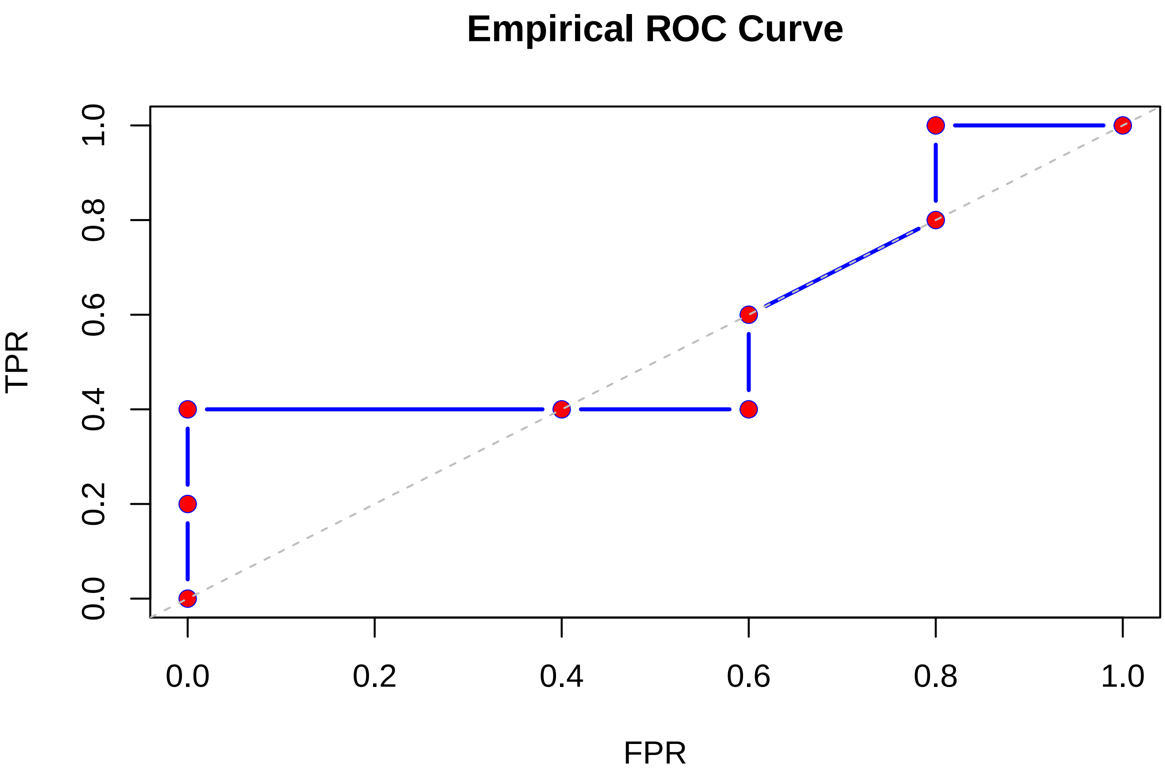
(a) To compute the ROC curve coordinates, we need to compute the TPR and FPR at various thresholds.

Threshold	TPR	FPR
0	1	1
1	1	0.8
2	0.8	0.8
3	0.6	0.6
4	0.4	0.6
5	0.4	0.4
7	0.4	0
8	0.2	0
8>	0	0

The empirical ROC curve can be drawn by plotting the points (FPR, TPR) from the table above.

```r
# (FPR, TPR)
fpr <- c(0,   0, 0.0, 0.4, 0.6, 0.6, 0.8, 0.8, 1)
tpr <- c(0, 0.2, 0.4, 0.4, 0.4, 0.6, 0.8, 1.0, 1)

# Plot ROC curve
plot(fpr, tpr, type = "b", col = "blue", lwd = 2, xlim = c(0, 1),
     ylim = c(0, 1), xlab = "FPR", ylab = "TPR", main = "Empirical ROC Curve")
# Add points
points(fpr, tpr, pch = 19, col = "red")
# Add diagonal line
abline(a = 0, b = 1, lty = 2, col = "gray")
```



(b) The empirical AUC:

$$AUCe = \frac{1}{n_D \bar{n}_D} \sum_{i=1}^{n_D} \sum_{j=1}^{\bar{n}_D} \Big((I(Y_{D_i} > Y_{\bar{D}_j}) + \frac{1}{2}(I(Y_{D_i} = Y_{\bar{D}_j}) \Big)$$

The calculattion based on the formula would be as following:

```r
# Calculate empirical AUC (AUCe)
n_D <- length(tpr); n_bar_D <- length(fpr)

AUCe <- 0
for (i in 1:n_D) {
  for (j in 1:n_bar_D) {
    if (tpr[i] > fpr[j]) { AUCe <- AUCe + 1}
    else if (tpr[i] == fpr[j]) { AUCe <- AUCe + 0.5}
  }
}

AUCe <- AUCe / (n_D * n_bar_D); print(AUCe)
```

```
## [1] 0.5432099
```

(c) Main Families of ROC Estimation Methods

1. Empirical Estimation:

   The empirical estimation method involves directly calculating True Positive Fraction (TPF) and False Positive Fraction (FPF) at various thresholds. The empirical ROC curve is then plotted as points (FPF, TPF).

   Key Features:

   - Does not assume any specific distribution for Y_D or $Y_{\bar{D}}$.
   - Allows straightforward computation directly from data.
   - The Area Under the Curve (AUC) is computed using numerical integration of the empirical ROC curve or via the Mann-Whitney U-statistic.

   Strengths and Weaknesses:

   - Strength: Flexible and relies only on observed data.
   - Weakness: Sensitive to sample size and variability, especially for small datasets.

2. Estimation by Modeling Test Result Distributions:

   This family assumes specific statistical distributions for diseased (Y_D) and non-diseased ($Y_{\bar{D}}$) populations.

   (a) Fully Parametric Models

   - Assumes a known distribution (e.g., normal) for Y_D and $Y_{\bar{D}}$.
   - The ROC curve is calculated based on the fitted parametric models. For example, the binormal model defines the ROC curve as:

     $$\text{ROC}(t) = \Phi(a + b\Phi^{-1}(t))$$

     where $a = \frac{\mu_D - \mu_{\bar{D}}}{\sigma_{\bar{D}}}$ and $b = \frac{\sigma_D}{\sigma_{\bar{D}}}$.

     Strengths and Weaknesses:

     - Strength: Efficient and robust when assumptions hold.
     - Weakness: Results can be biased if the assumed model is incorrect.

   (b) Semi-Parametric Models

   - Makes weaker assumptions, such as requiring that Y_D and $Y_{\bar{D}}$ differ only by location and/or scale parameters (e.g., location-scale models).

     Strengths and Weaknesses:

     - Strength: More flexible than fully parametric methods.
     - Weakness: Requires some parametric assumptions, which may not always hold.

3. Parametric Distribution-Free Estimation

   This family assumes no specific distribution for Y_D and $Y_{\bar{D}}$, but imposes a mathematical form on the ROC curve.

   Example Method: LABROC

   - LABROC fits a binormal ROC curve without directly modeling the distributions of Y_D and $Y_{\bar{D}}$.

   Strengths and Weaknesses:

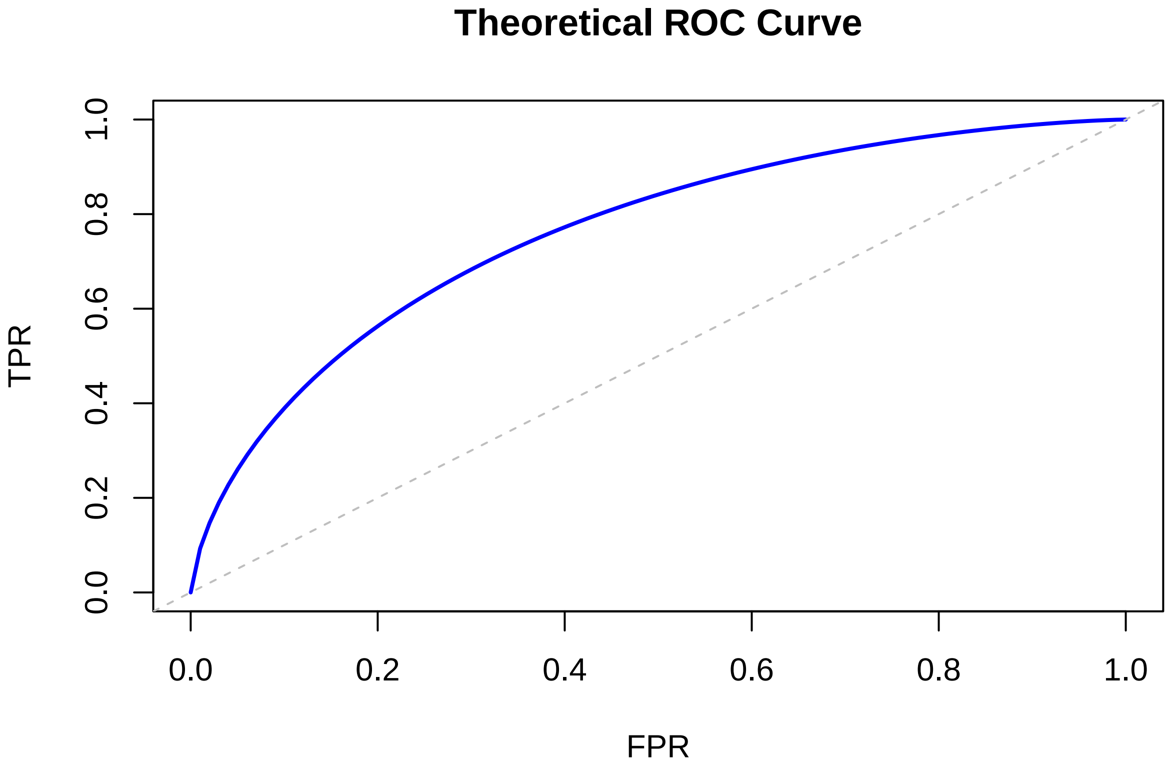
   - Strength: Allows flexibility in distribution assumptions while retaining structure in the ROC curve.
   - Weakness: Requires strong assumptions on the functional form of the ROC curve.

```
# Define parameters
mu_D <- 1
mu_bar_D <- 0
sigma_D <- 1
sigma_bar_D <- 1

# Generate FPR values
fpr <- seq(0, 1, length.out = 100)

# Compute TPR values
tpr <- 1 - pnorm(qnorm(1 - fpr, mu_bar_D, sigma_bar_D), mu_D, sigma_D)

# Plot ROC curve
plot(fpr, tpr, type = "l", col = "blue", lwd = 2, xlab = "FPR", ylab = "TPR",
     main = "Theoretical ROC Curve")
abline(a = 0, b = 1, lty = 2, col = "gray")
```



(d) Explanation of the Formulas

- for large samples:

$$\mathrm{var}(\widehat{AUC_e}) \approx \frac{\mathrm{var}(\widehat{FPF}(Y_{D_i}))}{n_D} + \frac{\mathrm{var}(\widehat{TPF}(Y_{\bar{D}_j}))}{n_{\bar{D}}}$$

this formula estimates the variance of the empirical AUC ($\widehat{AUC_e}$). it reflects the uncertainty of the estimate based on the variability of the false positive fraction (FPF) and true positive fraction (TPF) across the samples. in practice, this formula is used to compute confidence intervals
for $\widehat{AUC_e}$. The variances of $\widehat{FPF}$ and $\widehat{TPF}$ are estimated from the data, and the resulting values are scaled by the sample sizes (n_D for diseased subjects and $n_{\bar{D}}$ for non-diseased subjects) to support statistical inference.

- for unpaired samples:

$$\mathrm{var}(\Delta\widehat{AUC_e}) = \mathrm{var}(\widehat{AUC_{A_e}}) + \mathrm{var}(\widehat{AUC_{B_e}})$$

  this formula calculates the variance of the difference between two empirical AUCs, $\widehat{AUC_{A_e}}$ and $\widehat{AUC_{B_e}}$, derived from unpaired samples. it assumes the two diagnostic tests being compared are independent. in practice, this formula is used in hypothesis testing (e.g., with a z-statistic) to assess whether the performance of one test significantly exceeds that of another. the variances for $\widehat{AUC_{A_e}}$ and $\widehat{AUC_{B_e}}$ are estimated from the respective data sets.

7 Exercise 7 (Verification Bias)

Consider a screening test for a particular disease which is applied to all newborns. If the test suggests the presence of hearing impairment, then follow-up testing with the time consuming gold standard test (yielding D) is clinically indicated and performed. If the screening test suggests that the child does not suffer from hearing impairment, then there are no clinical reasons to perform the gold standard test. For research purposes, however, it is necessary to perform the gold standard test on some subjects who screen negative, so one decides that a fraction of 5% of the negatively screening subjects are selected for testing with the gold standard. In the following, we consider the data from such a study:

	$D = 1$	$D = 0$
$Y = 1$	100	50
$Y = 0$	2	50

(a) Compute the TPF "naively" and explain why this estimator should not be used.

(b) Which type of bias is present here?

(c) Explain the idea of one alternative estimator of TPF that is not affected by this bias, compute it and explain whether it makes sense in the two following situations:

(i) the 5% of negatively screening subjects undergoing the gold standard test are selected completely randomly;

(ii) the 5% of negatively screening subjects undergoing the gold standard test are selected among those with a case in their family history (knowing that the considered disease is partly hereditary).

Answer:

Note: Always pay attention to the study design before starting to calculate things.

(a) The naive true positive fraction (TPF) is calculated as:

$$\text{TPF}_{\text{naive}} = \frac{\text{Number of True Positives (TP)}}{\text{Number of True Positives (TP)} + \text{Number of False Negatives (FN)}} = \frac{100}{100+2} = 0.980$$

This estimator is biased because only 5% of $Y = 0$ subjects are verified with the gold standard test. The observed false negatives (FN = 2) must be scaled up by the inverse of the sampling fraction (1/0.05) to estimate the total number of false negatives. The naive approach neglects this scaling, leading to an overestimation of sensitivity (TPF).

(b) The type of bias present is **verification bias** (also known as work-up bias, referral bias, or selection bias). Verification bias arises because only a subset of negatively screening subjects ($Y = 0$) undergoes verification with the gold standard test, leading to biased estimates of sensitivity and specificity.

(c) To address verification bias, the **inverse probability weighting (IPW)** method can be used. This approach corrects the observed counts by scaling them with the inverse of the sampling probabilities. For the negatively screening group ($Y = 0$), the number of subjects verified is scaled by the inverse of the sampling fraction ($1/0.05 = 20$).

Corrected counts for $Y = 0$:

- For $D = 1$: $2/0.05 = 40$
- For $D = 0$: $50/0.05 = 1000$

The corrected true positive fraction is:

$$\text{TPF}_{\text{corrected}} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{Corrected False Negatives (FN)}} = \frac{100}{100+40} = 0.714$$

Evaluation of corrected estimator:

1. Random sampling of 5% of $Y = 0$ subjects:
   If the 5% of negatively screening subjects ($Y = 0$) are selected completely randomly, the IPW method is valid because the sampling fraction is independent of disease-related factors. In this case, the corrected TPF estimate is unbiased.

2. Family history-based sampling of 5% of $Y = 0$ subjects:
   If the 5% of negatively screening subjects ($Y = 0$) are selected based on family history, the sampling fraction depends on disease-related factors. This violates the missing at random (MAR) assumption, making the IPW method invalid. The corrected TPF estimate would be biased due to confounding introduced by the family history criterion.