Chapter 8: Hypothesis Testing

Table of Contents

Online Supplementary Material

`Six sigma.pdf`

8.1: Introduction

The problem addressed in this chapter is how to decide whether an estimate of AUC is consistent with a pre-specified value. One example of this is when a single-reader rates a set of cases in a single-modality, from which one estimates AUC, and the question is whether the estimate is statistically consistent with a pre-specified value. From a clinical point of view, this is generally not a useful exercise, but its simplicity is conducive to illustrating the broader concepts involved in this and later chapters. The clinically more useful analysis is when multiple readers interpret the same cases in two or more modalities. With two modalities, for example, one obtains an estimate AUC for each reader in each modality, averages the AUC values over all readers within each modality, and computes the inter-modality difference in reader-averaged AUC values. The question forming the main subject of this book is whether the observed difference is consistent with zero.

Each situation outlined above admits a binary (yes/no) answer, which is different from the estimation problem that was dealt with in connection with the maximum likelihood method in **Chapter 06**, where one computed numerical estimates (and confidence intervals) of the parameters of the fitting model.

Hypothesis testing is the process of dichotomizing the possible outcomes of a statistical study and then using probabilistic arguments to choose one option over the other.

The two competing options are termed the null hypothesis (NH) and the alternative hypothesis (AH). The hypothesis testing procedure is analogous[1] to the jury trial system in the US, with 20 instead of 12 jurors, with the null hypothesis being the presumption of innocence and the alternative hypothesis being the defendant is guilty, and the decision rule is to assume the defendant is innocent unless all 20 jurors agree the defendant is guilty. If even one juror disagrees, the defendant is deemed innocent (equivalent to choosing an α – defined below - of 0.05, or 1/20).

8.2: Hypothesis testing for a single-modality single-reader ROC study

The binormal model described in **Chapter 06** can be used to generate sets of ratings to illustrate the methods being described in this chapter. To recapitulate, the model is described by:

$$\left.\begin{array}{l} Z_{k_1 1} \sim N\left(0,1\right) \\ Z_{k_2 2} \sim N\left(\mu,\sigma^2\right) \end{array}\right\} . \qquad \textbf{(8.1)}$$

Set $\mu = 1.5$ and $\sigma = 1.3$ (these values were selected arbitrarily; the reader may wish to experiment with different ones) and simulate $K_1 = 50$ non-diseased cases and $K_1 = 52$ diseased cases. For debugging purposes the author likes to keep the sizes of the two arrays slightly different; this allows one to quickly check, with a glance at the **Environment** panel, that array dimensions are as expected. The **R**-code in **mainHT1R1M.R** follows (the file name stands for "Hypothesis Testing One-Reader One-Modality"):

8.2.1: Code Listing

```
rm(list = ls()) #mainHT1R1M.R
source("Wilcoxon.R")

seed <- 1;set.seed(seed)
mu <- 1.5;sigma <- 1.3;K1 <- 50;K2 <- 52

# cheat to find the population mean and std. dev.
AUC <- array(dim = 10000)
for (i in 1:length(AUC)) {
  zk1 <- rnorm(K1);zk2 <- rnorm(K2, mean = mu, sd = sigma)
  AUC[i] <- Wilcoxon(zk1, zk2)
}
sigmaAUC  <-  sd(AUC);meanAUC   <-  mean(AUC)
cat("mean AUC = ", meanAUC, ", sigma AUC = ", sigmaAUC, "\n")
```

```
# one more trial, this is the one we want to compare to meanAUC, i.e., get p-value
zk1 <- rnorm(K1);zk2 <- rnorm(K2, mean = mu, sd = sigma) # sqrt(DeLong(zk1,zk2))
AUC <- Wilcoxon(zk1, zk2)
cat("New AUC = ", AUC, "\n")

z <- (AUC - meanAUC)/sigmaAUC
#z <- qnorm(0.05/2)
cat("z-statistic = ", z, "\n")

p2tailed <- pnorm(-abs(z)) + (1-pnorm(abs(z))) # p value for two-sided AH
p1tailedGT <- 1-pnorm(z) # p value for one-sided AH > 0
p1tailedLT <- pnorm(z)# p value for one-sided AH < 0
alpha  <- 0.05

z2tailed <- -qnorm(alpha/2) # critical value for two-sided AH: AUC not equal to meanAUC
z1tailedGT <- qnorm(1-alpha) # critical value for one-sided AH: AUC > meanAUC
z1tailedLT <- qnorm(alpha) # critical value for one-sided AH: AUC < meanAUC

cat("alpha of test = ", alpha, "\n")
cat("\nTwo-sided AH: AUC not equal to meanAUC", "\n")
cat("Critical value for two-sided AH:", z2tailed, "\n")
cat("p value for two-sided AH:", p2tailed, "\n")

cat("\nOne-sided AH: AUC > meanAUC", "\n")
cat("Critical value for one-sided AH:", z1tailedGT, "\n")
cat("p value for two-sided AH:", p1tailedGT, "\n")

cat("\nOne-sided AH: AUC < meanAUC", "\n")
cat("Critical value for one-sided AH:", z1tailedLT, "\n")
cat("p value for two-sided AH:", p1tailedLT, "\n")
```

`Sourcing` the code yields the following output.

8.2.2: Code Output

```
> source('~/book2/03 B Statistics of ROC analysis/B1 Hypothesis Testing/software/mainHT1R1M.R')
pop mean AUC =  0.819178 , pop sigma AUC =  0.04176683
New AUC =  0.8626923
z-statistic =  1.04184
alpha of test =  0.05

Two-sided AH: AUC not equal to meanAUC
Critical value for two-sided AH: 1.959964
p value for two-sided AH: 0.297486

One-sided AH: AUC > meanAUC
Critical value for one-sided AH: 1.644854
p value for two-sided AH: 0.148743

One-sided AH: AUC < meanAUC
Critical value for one-sided AH: -1.644854
p value for two-sided AH: 0.851257
```

Lines 7 – 14 use the simple (if unimaginative) approach of sampling 10,000 times to estimate the population mean and standard deviation of empirical *AUC*, denoted below by AUC_{pop} and σ_{AUC}, respectively, (similar to that done in **Chapter 07** to validate the different variance estimation methods). Based on the 10,000 simulations, $AUC_{pop} = 0.8192$ and $\sigma_{AUC} = 0.04177$.

Lines 17-18 simulate one more independent ROC study with the same numbers of cases, and the resulting area under the empirical curve is denoted AUC , **AUC** in the code. Is the new value (0.8627) sufficiently different from the population mean 0.8192 to reject the null hypothesis $NH: AUC = AUC_{pop}$? *Note that the answer to this question can be either yes or no: equivocation is not allowed.*

The new value is "somewhat close" to the population mean, but how does one decide if "somewhat close" is close enough? Needed is the statistical distribution of the random variable AUC under the hypothesis that the true mean is AUC_{pop} . In the asymptotic limit of a large number of cases (this is an approximation), one can assume that the *pdf* of AUC under the null hypothesis is the normal distribution $N\left(AUC_{pop},\sigma_{AUC}^2\right)$:

$$pdf_{AUC}\left(AUC \mid AUC_{pop},\sigma_{AUC}\right)=\frac{1}{\sqrt{2\pi}\sigma_{AUC}}\exp\left(-\frac{1}{2}\left(\frac{AUC-AUC_{pop}}{\sigma_{AUC}}\right)^2\right) \quad . \qquad \textbf{(8.2)}$$

The translated and scaled value is distributed as a unit normal distribution, i.e.,

$$Z=\frac{AUC-AUC_{pop}}{\sigma_{AUC}}\sim N\left(0,1\right) \quad . \qquad \textbf{(8.3)}$$

[The *Z* notation here should not be confused with z-sample, decision variable or rating of a case in an ROC study; the latter, when sampled over a set of non-diseased and diseased cases, yield a realization of AUC. The author trusts the distinction will be clear from the context.] The observed *magnitude* of *z* is 1.042 (fourth line in code output above).

The ubiquitous *p-value* is the probability that the *observed magnitude of z, or larger*, occurs under the null hypothesis (NH), that the true mean of Z is zero.

The *p-value* corresponding to an observed *z* of 1.042 is given by (as always, uppercase Z is the random variable, while lower case z is a realized value):

$$\left.\begin{aligned}&P\left(|Z| \geq |z| \,\middle|\, Z \sim N(0,1)\right) = P\left(|Z| \geq 1.042 \,\middle|\, Z \sim N(0,1)\right)\\&= P\left(Z \geq 1.042 \,\middle|\, Z \sim N(0,1)\right) + P\left(Z \leq -1.042 \,\middle|\, Z \sim N(0,1)\right)\\&= 2\Phi(-1.042) = 0.2975\end{aligned}\right\} \quad . \qquad \textbf{(8.4)}$$

To recapitulate statistical notation, $P\left(|Z| \geq |z| \,\middle|\, Z \sim N(0,1)\right)$ is to be parsed as $P(A|B)$, that is, the probability $|Z| \geq |z|$ given that $Z \sim N(0,1)$. The last line in Eqn. (8.4) follows from the symmetry of the unit normal distribution, i.e., the area above 1.042 must equal the area below -1.042.

Since *z* is a continuous variable, the probability that a sampled value will *exactly* equal the observed value is *zero*. Therefore, one must pose the question as stated above, namely what is the probability that *Z* is at least as extreme as the observed value (by "extreme" the author means further from zero, in either positive or negative directions). If the observed was *z* = 2.5 then the corresponding p-value would be $2\Phi(-2.5) = 0.01242$, which is smaller than 0.2975 (**`2*pnorm(-2.5)`** = 0.01241933). This is cited below as the "*second example*".

Under the zero-mean null hypothesis, the larger the magnitude of the observed value of *Z*, the smaller the p-value, and the more unlikely that the data supports the NH. *The p-value can be interpreted as the degree of unlikelihood that the data supports the NH.*

By convention one adopts a *fixed* value of the probability, denoted α and usually $\alpha = 0.05$, which is termed the *size of the test* or the *significance level of the test*, and the decision rule is to reject the null hypothesis if the observed *p-value* < α .

$$p < \alpha \Rightarrow \text{Reject NH} \quad . \qquad \textbf{(8.5)}$$

In the first example, with observed p-value equal to 0.2975, one would not reject the null hypothesis, but in the second example, with observed p-value equal to 0.01242, one would. If the p-value is exactly 0.05 (unlikely with ROC analysis, but one needs to account for it) then one does not reject the NH. In the 20-juror analogy, of one juror insists the defendant is not guilty, then observed P is 0.05, and one does not reject the NH that the defendant is innocent (the double negatives, very common in statistics, can be confusing; in plain English, the defendant goes home).

According to the previous discussion, the critical magnitude of *z* that determines whether to reject the null hypothesis is given by:

$$z_{\alpha/2} = -\Phi^{-1}\left(\alpha/2\right) \quad . \qquad \textbf{(8.6)}$$

For $\alpha = 0.05$ this evaluates to 1.95996 (which is sometimes rounded up to two, good enough for "government work" as the saying goes) and the decision rule is to reject the null hypothesis only if the observed *magnitude* of z is larger than $z_{\alpha/2}$.

> *The decision rule based on comparing the observed z to a critical value is equivalent to a decision rule based on comparing the observed p-value to* α. It is also equivalent*, as will be shown later, to a decision rule based on a* $\left(1-\alpha\right)$ *confidence interval for the observed statistic. One rejects the NH if the closed confidence interval does not include zero.*

8.3: Type-I errors

Just because one rejects the null hypothesis, as in the second example, does not mean that the null hypothesis is false. Following the decision rule "caps", or puts an upper limit on, the probability of incorrectly rejecting the null hypothesis at α. In other words, by agreeing to reject the NH only if p < α, one has set an upper limit, namely α, on errors of this type, termed *Type-I* errors. These could be termed false positives in the hypothesis testing sense, not to be confused with false positive occurring on individual case-level decisions. According to the definition of α:

$$P\left(\text{Type I error} \middle| NH\right) = \alpha \quad . \qquad \textbf{(8.7)}$$

To demonstrate the idea one needs to have a very cooperative reader interpreting new sets of independent cases not just one more time, but 2000 more times (the reason for the 2000 trials will be explained below). The code for this is in file **`mainTypeIErrors.R`**:

8.3.1: Code Listing

```
rm(list = ls()) # mainTypeIErrors.R
source("Wilcoxon.R")
```

```
seed <- 1;set.seed(seed)
mu <- 1.5;sigma <- 1.3;K1 <- 50;K2 <- 52

# cheat to find the population mean and std. dev.
AUC <- array(dim = 10000)
for (i in 1:length(AUC)) {
  zk1 <- rnorm(K1);zk2 <- rnorm(K2, mean = mu, sd = sigma)
  AUC[i] <- Wilcoxon(zk1, zk2)
}
sigmaAUC  <-  sqrt(var(AUC));muAUC   <-  mean(AUC)

nTrials <- 2000
alpha <- 0.05 # size of test
reject = array(0, dim = nTrials)
for (trial in 1:length(reject)) {
  zk1 <- rnorm(K1);zk2 <- rnorm(K2, mean = mu, sd = sigma)
  AUC <- Wilcoxon(zk1, zk2)
  z <- (AUC - muAUC)/sigmaAUC
  p <- 2*pnorm(-abs(z)) # p value for individual trial
  if (p <= alpha) reject[trial] = 1
}

CI <- c(0,0); width <- -qnorm(alpha/2)
ObsvdTypeIErrRate <- sum(reject)/length(reject)
CI[1] <- ObsvdTypeIErrRate - width*sqrt(ObsvdTypeIErrRate*(1-ObsvdTypeIErrRate)/nTrials)
CI[2] <- ObsvdTypeIErrRate + width*sqrt(ObsvdTypeIErrRate*(1-ObsvdTypeIErrRate)/nTrials)
cat("alpha = ", alpha, "\n")
cat("ObsvdTypeIErrRate = ", ObsvdTypeIErrRate, ", 95% confidence interval = ", CI, "\n")
```

The first 13 lines are identical to the corresponding lines in **`mainHT1R1M.R`**. Line 15 initializes **`NTrials`** to 2000 and line 16 initializes α to 0.05. Line 17 initializes all 2000 elements of a vector named **`reject`** to zeroes. Line 18-24 describes our captive reader interpreting independent sets of cases 2000 times. Each completed interpretation of 102 cases is termed a *trial*. For each trial line 20 calculates the observed value of AUC, the next line calculates the observed *z* statistic and the next line the observed *p-value*. Line 23 tests the observed *p-value* against the fixed value α and sets the corresponding **`reject`** flag to unity if $p < \alpha$. In other words, if the trial-specific *p-value* is less than α one counts an instance of rejection of the null hypothesis. The process is repeated 2000 times.

Line 27 calculates the observed Type-I error rate, denoted **`ObsvdTypeIErrRate`** by summing the **`reject`** array and dividing by the number of trials. Lines 28-29 calculate a 95% confidence interval for **`ObsvdTypeIErrRate`** based on the binomial distribution, as in **Chapter 03**. Finally, lines 30-32 print out the relevant results. If one sources the file one, after a brief delay, gets the following output:

8.3.2: Code Output

```
> source('~/book2/03 B Statistics of ROC analysis/B1 Hypothesis
Testing/software/mainTypeIErrors.R')
alpha =  0.05
ObsvdTypeIErrRate =  0.049 , 95% confidence interval =  0.03953934 0.05846066
exact 95% CI =  0.03995676 0.05939265
```

The first line reminds us that the chosen value for the size of the test is $\alpha = 0.05$. The next line is the observed value of the Type-I error rate (which is a realization of a random variable). The last line is the 95% confidence interval for the observed Type-I error rate (which is also a realization of a random range variable). The fact that this confidence interval includes the chosen value $\alpha = 0.05$ is no coincidence; it shows that the hypothesis testing procedure is working as expected. To distinguish between the selected α (a fixed value) and that observed in a simulation study (a realization of a random variable), the term *empirical* α *alpha* is used to denote the observed α.

It is a mistake to state that one wishes to minimize the Type-I error probability (the author has seen this comment from a senior researcher, which is the reason for bringing it up). The minimum value of α (a probability) is zero. Run the software with this value of α: *the software will never reject the NH. The downside of minimizing the expected Type-I error rate is that the software will never reject the NH, even when the NH is patently false*. The aim of a valid method of analyzing the data is not minimizing the Type-I error rate, rather, the observed Type-I error rate should equal the specified value of α (0.05 in our example), allowance being made for the inherent variability in it's estimate, i.e., its confidence interval. This is the reason 2000 trials were chosen for testing the validity of the NH testing procedure. With this choice, the 95% confidence interval, assuming that observed is close to 0.05, is roughly ±0.01 as explained next.

Following analogous reasoning to **Chapter 03**, Eqn. (3.10.10), and defining *f* as the *observed rejection fractio*n over *T* trials, and as usual, *F* is a random variable and *f* a realized value,

$$\begin{aligned} &\sigma_f = \sqrt{f\left(1-f\right)/T} \\ &F \sim N\left(f,\sigma_f^2\right) \end{aligned} \quad . \qquad \textbf{(8.8)}$$

An approximate $\left(1-\alpha\right)$ CI for f is:

$$CI_f = \left[f - z_{\alpha/2}\sigma_f, f + z_{\alpha/2}\sigma_f\right] \quad . \qquad \textbf{(8.9)}$$

If *f* is close to 0.05, then for 2000 trials, the 0.95 or 95% CI for *f* is *f* ± 0.01, i.e., **`-qnorm(alpha/2) * sqrt(.05*(.95)/2000)`** = 0.009551683 ~ 0.01.

The only way to reduce the width of the CI, and thereby run a more stringent test of the validity of the analysis, would be to increase the number of trials *T*. Since the width of the CI depends on the inverse square root of the number of trials, one soon reaches a point of diminishing returns. Usually *T* = 2000 trials are enough for most statisticians and the author, but examples of more simulations have been published.

8.4: One sided vs. two sided tests

In the preceding example, the null hypothesis was rejected anytime the *magnitude* of the observed value of *z* exceeded the cutoff value $z_{\alpha/2} = -\Phi^{-1}(\alpha/2)$. This is a statement of the *alternative hypothesis* $AH: AUC \neq AUC_{pop}$, in other words too high or too low values of *z* both result in rejection of the null hypothesis. This is referred to as a *two-sided* AH and the resulting p-value is termed a *two-sided* p-value. This is the most common one used in the literature.

Now suppose that the additional interpretation performed by the radiologist, in lines 17 - 18 in file **`mainHT1R1M.R`**, was performed after an intervention following which the radiologist's performance is expected to increase. To make matters clearer, assume the interpretations in the 10,000 trials used to estimate AUC_{pop} were performed with the radiologist wearing an old pair of eye-glasses, possibly out of proper strength, and the additional interpretation is performed after the radiologist gets a new set of prescription eye-glasses. Because the radiologist's eyesight has improved, the expectation is that performance should increase. In this situation, it is appropriate to use the one-sided alternative hypothesis $AH: AUC > AUC_{pop}$. Now excessively large values of z result in rejection of the null hypothesis, *but excessively small values of z do not*. The critical value of z is defined by $z_\alpha = \Phi^{-1}(1-\alpha)$, which for $\alpha = 0.05$ is 1.645 (**`qnorm(1-alpha)`** = 1.644854). Compare 1.64 to the value $z_{\alpha/2} = -\Phi^{-1}(\alpha/2)$ =1.96 for a two-sided test. If the change is in the expected direction, it is easier to reject the NH with a one-sided than with a two-sided test. The *p-value* for a one-sided test is given by (the relevant file is **`mainHT1R1M.R`**)

$$P(Z \geq 1.042 | NH) = \Phi(-1.042) = 0.1487 \quad . \qquad \mathbf{(8.10)}$$

Notice that this is half the corresponding two-sided test *p-value*; this is because one is only interested in the area under the unit normal that is above the observed value of *z*. If the intent is to obtain a significant finding, it is tempting to use one-sided tests. The down side of a one-sided test is that even with a large excursion of the observed *z* in the *other* direction one cannot reject the null hypothesis. So if the new eye-glasses are so bad as to render the radiologist practically blind (think of a botched cataract surgery) the observed z would be large and negative, but one could not reject the null hypothesis $NH : AUC_{true} = \mu_{\theta}$.

The one-sided test could be run the other way, with the alternative hypothesis being stated as follows $AH : AUC < AUC_{pop}$. Now large negative excursions of the observed value of AUC cause rejection of the null hypothesis, but large positive excursions do not. The critical value is defined by $z_{\alpha} = \Phi^{-1}(\alpha)$, which for $\alpha = 0.05$ is -1.645. The *p-value* is given by (note the reversed sign compared to Eqn. (8.10)):

$$P(Z \le 1.042 | NH) = \Phi(1.042) = 1 - 0.1487 = 0.8513 \quad . \qquad \textbf{(8.11)}$$

This is the complement of the value for a one-sided test with the alternative hypothesis going the other way: obviously the probability that *Z* is smaller than the observed value (1.042) plus the probability that *Z* is larger than the same value must equal one.

8.5: Statistical power

So far, focus has been on the null hypothesis. The Type-I error probability was introduced, defined as the probability of incorrectly rejecting the null hypothesis, the control, or "cap" on which is α, usually set to 0.05. What if the null hypothesis is actually false and the study fails to reject it? This is termed a Type-II error, the control on which is denoted β, the probability of a Type-II error. The complement of β is called *statistical power*.

Table 8.1 summarizes the two types of errors and the two correct decisions that can occur in hypothesis testing. In the context of hypothesis testing, a Type-II error could be termed a false negative, not to be confused with false negatives occurring on individual case-level decisions.

Table 8.1: This table illustrates the two types of errors and correct decisions that can occur in hypothesis testing. The probability of incorrectly rejecting the NH, a Type-I error, the cap on which is α, is usually set to 0.05. A Type-II error occurs when the NH

is false and the study fails to reject it, the control on which is denoted β. The complement of β is statistical power. The two types of errors could be termed FPF (declaring a non-existent difference) and FNF (failing to declare an existing difference), but these terms should not be confused with the results of a single binary paradigm ROC study.

Truth	Decision	
	Fail to reject NH	Reject NH
NH is true	$1-\alpha$	α (FPF)
NH is false	β (FNF)	$\text{Power} = 1-\beta$

This resembles the 2 x 2 table encountered in **Chapter 02**, Table 2.1, which led to the concepts of FPF, TPF and the ROC curve. Indeed, it is possible think of an analogous plot of empirical (i.e., observed) power vs. empirical α, which looks like an ROC plot, with empirical α playing the role of FPF and empirical power playing the role of TPF, see below. If α = 0, then power = 0; i.e., if Type-I errors are minimized all the way to zero, then power is zero and one makes Type-II errors all the time. On the other hand, if α = 1 then Power = one and one makes Type-I errors all the time.

A little history is due at this point. The author's first FROC study, which led to his entry into this field[2], was published in Radiology in 1986 after a lot of help from a reviewer, who we correctly guessed was the late Prof. Charles E. Metz. Prof. Gary T. Barnes (the author's mentor at that time at the University of Alabama at Birmingham) and the author visited Prof. Charles Metz in Chicago for a day ca. 1986, to figuratively “pick Charlie’s brain”. Prof. Metz referred to the concept outlined in the previous paragraph, as an "*ROC within an ROC*".

This curve does not summarize the result of a single ROC study. Rather it summarizes the probabilistic behavior of the two types of errors that occur when one conducts thousands of such studies, under both NH true and NH false conditions, each time with different values of α, with each trial ending in a decision to reject or not reject the null hypothesis. The long sentence is best explained with an example. Open the file **`mainRocWithinRoc.R`**:

8.5.1: Metz's "ROC within an ROC"

8.5.1.1: Code Listing

```
rm(list = ls()) # mainRocWithinRoc.R
library(ggplot2)
source("Wilcoxon.R")

seed <- 1;set.seed(seed)
muNH <- 1.5;muAH <- 2.1;sigma <- 1.3;K1 <- 50;K2 <- 52#;K1 <- K1*2;K2 <- K2*2

# cheat to find the population mean and std. dev.
AUC <- array(dim = 10000)
```

```
for (i in 1:length(AUC)) {
  zk1 <- rnorm(K1);zk2 <- rnorm(K2, mean = muNH, sd = sigma)
  AUC[i] <- Wilcoxon(zk1, zk2)
}
sigmaAUC <- sqrt(var(AUC));meanAUC <- mean(AUC)

T <- 2000
mu <- c(muNH,muAH)
alphaArr <- seq(0.05, 0.95, length.out = 10)
EmpAlpha <- array(dim = length(alphaArr));EmpPower <- array(dim = length(alphaArr))
for (a in 1:length(alphaArr)) {
  alpha <- alphaArr[a]
  reject <- array(0, dim = c(2, T))
  for (h in 1:2) {
    for (t in 1:length(reject[h,])) {
      zk1 <- rnorm(K1);zk2 <- rnorm(K2, mean = mu[h], sd = sigma)
      AUC <- Wilcoxon(zk1, zk2)
      obsvdZ <- (AUC - meanAUC)/sigmaAUC
      p <- 2*pnorm(-abs(obsvdZ)) # p value for individual t
      if (p < alpha) reject[h,t] = 1
    }
  }
  EmpAlpha[a] <- sum(reject[1,])/length(reject[1,])
  EmpPower[a] <- sum(reject[2,])/length(reject[2,])
}
EmpAlpha <- c(0,EmpAlpha,1)
EmpPower <- c(0,EmpPower,1)

pointData <- data.frame(EmpAlpha = EmpAlpha, EmpPower = EmpPower)
zetas <- seq(-5, 5, by = 0.01)
muRoc <- 1.8
curveData <- data.frame(EmpAlpha = pnorm(-zetas), EmpPower = pnorm(muRoc - zetas))
alphaPowerPlot <- ggplot(mapping = aes(x = EmpAlpha, y = EmpPower)) + geom_point(data = pointData,
shape = 1, size = 3) + geom_line(data = curveData)
print(alphaPowerPlot)
```

Line 6 creates two variables, **muNH** = 1.5 (the binormal model separation parameter under the NH) and **muAH** = 2.1 (the separation parameter under the AH). Under either hypotheses, the same diseased case standard deviation **sigma** = 1.3 and 50 non-diseased and 52 diseased cases are assumed. As before, lines 8 – 14 use the "brute force" technique to determine population AUC and standard deviation of AUC under the NH condition. Line 16 defines the number of trials **T** = 2000. Line 16 creates a vector **mu** containing the NH and AH values defined at line 5. Line 18 creates **alphaArr**, a sequence of 10 equally spaced values in the range 0.05 to 0.95, which represent 10 values for α. Line 19 creates two arrays of length 10 each, named **EmpAlpha** and **EmpPower**, to hold the values of the observed Type-I error rate, i.e., empirical α, and the empirical power, respectively. The program will run **T** = 2000 NH and **T** = 2000 AH trials using as α each successive value in **alphaArr** and save the observed Type-I error rates and observed powers to the arrays **EmpAlpha** and **EmpPower**, respectively.

The action begins in line 20, which begins a **for**-loop in **a**, an index into **alphaArr**. Line 21 selects the appropriate value for **alpha** (0.05 on the first pass, 0.15 on the next pass, etc.). Line 21 initializes **reject[2,2000]** with zeroes, to hold the result of each trial; the first index corresponds to hypothesis **h** and the second to trial **t**. Line 23 begins a for-loop in **h**, with **h** = 1 corresponding to the NH and **h** = 2 to the AH.

Line 24 begins a for-loop in **`t`**, the trial index. The code within this block is similar to previous examples. It simulates ratings, computes AUC, calculates the p-value, and saves a rejection of the NH as a one at the appropriate array location **`reject[h,t]`**. Lines 32 – 33 calculate the empirical α and empirical power for each value of α in **`alphaArr`**. After padding the ends with zero and ones (the trivial points), the remaining lines plot the "ROC within an ROC". **`Source`** this code to get Fig. 8.1.

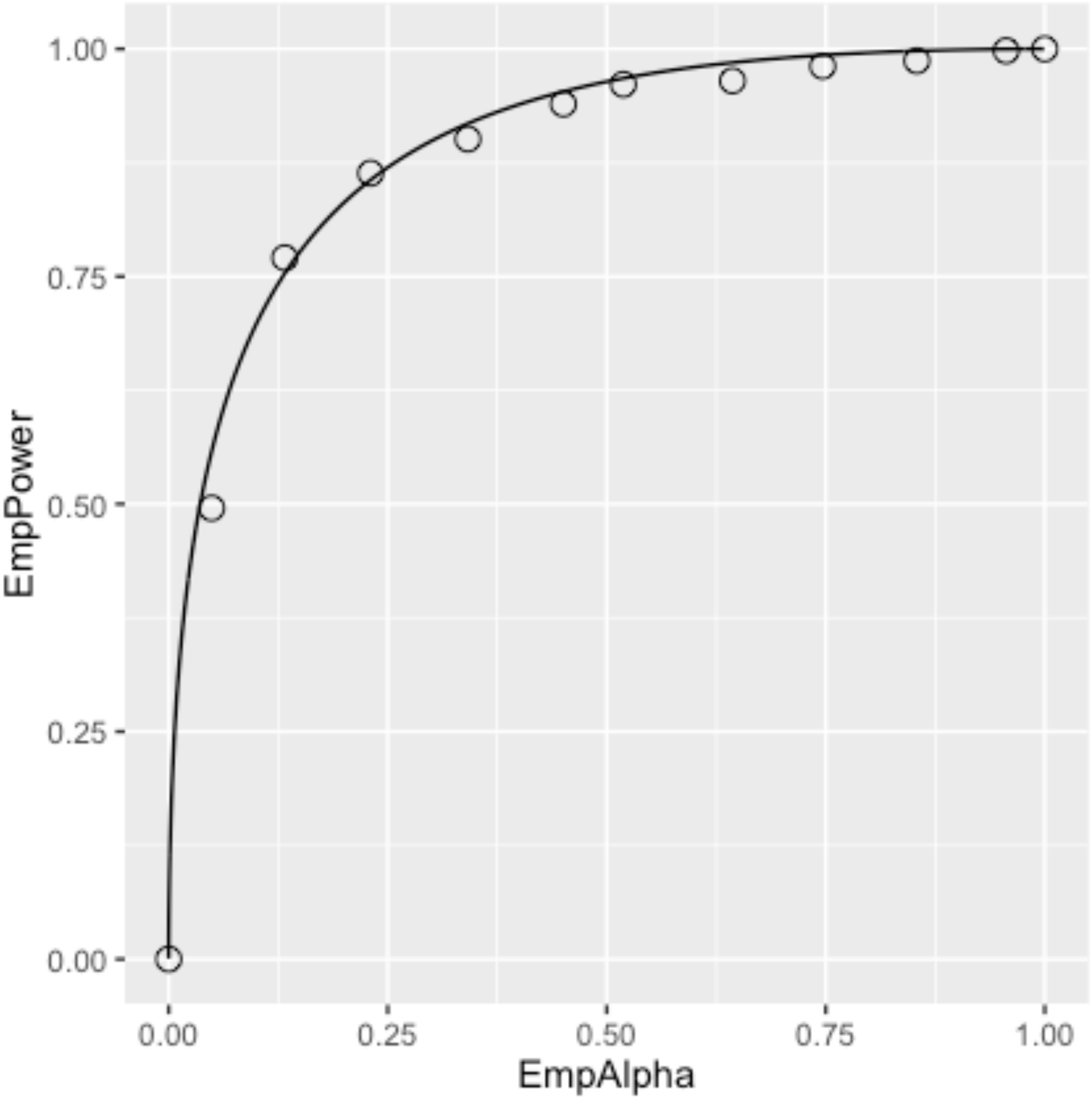

Fig. 8.1: Metz's "ROC within an ROC": each circle corresponds to a specific value of true α, which increases as the operating point moves up the plot, leading to corresponding increases in empirical alpha (x-axis) and empirical power (y-axis). This figure was generated by the code in **`mainRocWithinRoc.R`**.

Each of the circles in Fig. 8.1 corresponds to a specific value of α. For example the lowest non-trivial corresponds to $\alpha = 0.05$, for which the empirical α is 0.049 and the corresponding empirical Power is 0.4955.

True α increases as the operating point moves up the plot, with empirical α and empirical power increasing correspondingly. The AUC under this curve is determined by the effect size, defined as the difference between the AH and NH values of the separation parameter. If the effect size is zero, then the circles will scatter around the chance diagonal; the scatter will be consistent with the 2000 trials used to generate each coordinate of a point. As the effect size increases, the plot approaches the perfect "ROC", i.e., approaching the top-left corner. One could use AUC under this "ROC" as a measure of the incremental performance, the advantage being that it would be totally independent of α, but this would not be practical as it requires replication of the study under NH and AH conditions about 2000 times each and the entire process has to be repeated for several values of α. The purpose of this demonstration was to illustrate the concept behind Metz's profound remark.

It is time to move on to factors affecting statistical power in a single study.

8.5.2: Factors affecting statistical power

(1) **Effect size:** effect size is defined as the difference in AUC_{pop} values between the alternative hypothesis condition and the null hypothesis condition. Recall that AUC_{pop} is defined as the true or population value of the empirical ROC-AUC for the relevant hypothesis. One can use the "cheat method" to estimate it under the alternative hypothesis. The formalism is easier if one assumes it is equal to the asymptotic binormal model predicted value using Eqn. (6.49). The binormal model yields an *estimate* of the parameters, which only approach the population values in the asymptotic limit of a large number of cases. In the following, it is assumed that the parameters on the right hand side of Eqn. (8.12) are the population values):

$$AUC_{pop} = \Phi\left(\frac{\mu}{\sqrt{1+\sigma^2}}\right) \quad . \qquad \mathbf{(8.12)}$$

It follows that effect size (ES) is given by (all quantities on the right hand side of Eqn. (8.13) are population values):

$$ES = \Phi\left(\frac{\mu_{AH}}{\sqrt{1+\sigma^2}}\right) - \Phi\left(\frac{\mu_{NH}}{\sqrt{1+\sigma^2}}\right) \quad . \qquad \mathbf{(8.13)}$$

This formula is coded in a function **`EffectSize()`**. It is called by the code in **`mainStatPower.R`**. The latter is a stripped down version of **`mainRocWithinRoc.R`**. The listing follows:

8.5.2.1: Code Listing

```
rm(list = ls())# mainStatPower.R
source("Wilcoxon.R");source("EffectSize.R")

seed <- 1;set.seed(seed)
mu <- 1.5;muAH <- 2.1;sigma <- 1.3;K1 <- 50;K2 <- 52#;K1 <- K1*2;K2 <- K2*2

# cheat to find the population mean and std. dev.
AUC <- array(dim = 10000)
for (i in 1:length(AUC)) {
  zk1 <- rnorm(K1);zk2 <- rnorm(K2, mean = mu, sd = sigma)
  AUC[i] <- Wilcoxon(zk1, zk2)
}
sigmaAuc  <-  sqrt(var(AUC));muAuc   <-  mean(AUC)

T <- 2000
alpha <- 0.05 # size of test
reject = array(0, dim = T)
for (t in 1:length(reject)) {
  zk1 <- rnorm(K1);zk2 <- rnorm(K2, mean = muAH, sd = sigma)
  AUC <- Wilcoxon(zk1, zk2)
  obsvdZ <- (AUC - muAuc)/sigmaAuc
  p <- 2*pnorm(-abs(obsvdZ)) # p value for individual t
  if (p <= alpha) reject[t] = 1
}

ObsvdTypeIErrRate <- sum(reject)/length(reject)
CI <- c(0,0);width <- -qnorm(alpha/2)
CI[1] <- ObsvdTypeIErrRate - width*sqrt(ObsvdTypeIErrRate*(1-ObsvdTypeIErrRate)/T)
CI[2] <- ObsvdTypeIErrRate + width*sqrt(ObsvdTypeIErrRate*(1-ObsvdTypeIErrRate)/T)
cat("alpha = ", alpha, "\n")
cat("#non-diseased images = ", K1, "\t#diseased images = ", K2, "\n")
cat("obsvdPower = ", ObsvdTypeIErrRate, "\n")
cat("95% confidence interval = ", CI, "\n")
cat("Effect Size = ", EffectSize(mu, sigma, muAH, sigma), "\n")
```

Ensure that **`alpha`** at line 16 is set to 0.05. Sourcing **`mainStatPower.R`** yields the following output:

8.5.2.2: Code Output

```
> source('~/book2/03 B Statistics of ROC analysis/B1 Hypothesis Testing/software/mainStatPower.R')
alpha =  0.05
#non-diseased images =  50       #diseased images =  52
obsvdPower =  0.509
95% confidence interval =  0.4870905 0.5309095
Effect Size =  0.08000617
```

The ES for the code above is 0.08 (in AUC units). It should be obvious that if effect size is zero, then power equals α. This is because then there is no distinction between the null and alternative hypotheses conditions (this choice would yield points scattered around the chance diagonal in Fig. 8.1). Conversely, as effect size increases, statistical power increases, the limiting value being unity, when every trial results in rejection of the null hypothesis. The reader should experiment with different values of **`muAH`** to be convinced of the truth of these statements.

(2) **Sample size**: uncomment the last two statements on line 5, i.e., increase the number of cases by a factor of two, i.e., $K_1 = 100; K_2 = 104$ and **`source`** the file. After waiting a short time, one gets:

8.5.2.3: Code Output

```
> source('~/book2/03 B Statistics of ROC analysis/B1 Hypothesis Testing/software/mainStatPower.R')
alpha =  0.05
#non-diseased images =  100      #diseased images =  104
obsvdPower =  0.8435
95% confidence interval =  0.8275767 0.8594233
Effect Size =  0.08000617
```

So doubling the numbers of cases (both non-diseased and diseased) results in statistical power increasing from 0.509 to 0.844. Increasing the numbers of cases decreases σ_{AUC}, the standard deviation of the empirical AUC curve, Eqn. (8.2). The reader can confirm by looking at the **`Environment`** panel that the new value of σ_{AUC} is 0.02947, which should be compared to the value 0.04177 for **`K1`** = 50, **`K2`** = 52. Recall that σ_{AUC} enters the denominator of the Z-statistic, Eqn. (8.3), so decreasing it will increase the probability of rejecting the null hypothesis.

(3) **Alpha**: Statistical power depends on α as shown in Fig. 8.1: return the sample size to the original values $K_1 = 50; K_2 = 52$. The results below are for two runs of the code, the first with the original value $\alpha = 0.05$, set at line 16, the second with $\alpha = 0.01$:

8.5.2.4: Code Output

```
> source('~/book2/03 B Statistics of ROC analysis/B1 Hypothesis Testing/software/mainStatPower.R')
alpha =  0.05
#non-diseased images =  50       #diseased images =  52
obsvdPower =  0.509
95% confidence interval =  0.4870905 0.5309095
Effect Size =  0.08000617

> source('~/book2/03 B Statistics of ROC analysis/B1 Hypothesis Testing/software/mainStatPower.R')
alpha =  0.01
#non-diseased images =  50       #diseased images =  52
obsvdPower =  0.1915
95% confidence interval =  0.1688365 0.2141635
Effect Size =  0.08000617
```

Statistical power decreases as α decreases, Fig. 8.1.

8.6: Comments on the code

The reader may have noticed the Wilcoxon statistic was used to estimate the area under the ROC curve. One could have used **`RocfitR.R`**, introduced in **Chapter 06**, to obtain maximum likelihood estimates of the area under the binormal model fitted ROC curve. The reasons for choosing the simpler empirical area are as follows. (1) With continuous ratings and 102 data points, the area under the empirical ROC curve is expected to be a close approximation to the fitted area. (2) With maximum likelihood estimation, the code would be more complex – in addition to the fitting routine one would require a binning routine and that would introduce yet another variable in the analysis, namely the number of bins. (3) The maximum likelihood fitting code can sometimes fail to converge, while the Wilcoxon method is always guaranteed to yield a result (this is a mixed blessing, see §7.10). The non-convergence issue is overcome by modern methods of curve fitting described in later chapters. (4) The aim was to provide an understanding of null hypothesis testing and statistical power without being bogged down in the details of curve fitting.

8.7: Why alpha is chosen to be 5%

One might ask why is α traditionally chosen to be 5%. It is not a magical number, rather a cost benefit tradeoff. Choosing too small a value of α would result in greater probability ($1-\alpha$) of the NH not being rejected, even when it is false, i.e., decreased power, Table 8.1. This would correspond to operating at the low-end of the plot shown in Fig. 8.1. Sometimes it is important to detect a true difference between the measured AUC and the postulated value. For example, a new eye-laser surgery procedure is invented and the number of patients is necessarily small as one does not wish to subject a large number of patients to an untried procedure. One seeks some leeway on the Type-I error probability, possibly increasing it to 0.1, in order to have a reasonable chance of success in detecting an improvement in performance due to better eyesight after the surgery. If the NH is rejected and the change is in the right direction, then that is good news for the researcher. One might then consider a larger clinical trial and set alpha at the traditional 0.05, making up the lost statistical power by increasing the number of patients on which the surgery is tried.

If a whole branch of science hinges on the results of a study, such as discovering the Higg's Boson in particle physics, statistical significance is often expressed in multiples of the standard deviation (σ) of the normal distribution, with the significance threshold set at a much stricter level (e.g. 5σ). This corresponds to an alpha of about 1 in 3.5 million (**`1/pnorm(-5)`** = 3.5 x 10^6, a one-sided test of significance). There is an article in Scientific American (https://blogs.scientificamerican.com/observations/five-sigmawhats-that/) on the use of $n\sigma$, where *n* is an integer, e.g. 5, to denote the significance level of a study, and some interesting anecdotes on why such high significance levels (low alpha) are used in some fields of research.

Similar concerns apply to manufacturing where the cost of a mistake could be the very expensive recall of an entire product line. For background on Six Sigma Performance, see http://www.six-sigma-material.com/Six-Sigma.html. An article downloaded 3/30/17 from https://en.wikipedia.org/wiki/Six_Sigma is included as supplemental material to this chapter (**`Six Sigma.pdf`**). It has an explanation of why 6σ translates to one defect per 3.4 million opportunities (it has to do with short-term and long-term drifts in a process). In the author's opinion, looking at other fields offers a deeper understanding of this material than simply stating that by tradition one adopts alpha = 5%.

Most observer performance studies, while important in the search for better imaging methods, are not of such "earth-shattering" importance, and it is somewhat important to detect true differences (AH is true) at a reasonable alpha, so alpha = 5% and beta = 20% represent a good compromise. If one adopted a 5σ criterion, the NH would never be rejected, and progress in image quality optimization would come to a grinding halt. That is not to say that a 5σ criterion cannot be used; rather if used, the number of patients needed to detect a reasonable difference (effect size) with 80% probability would be astronomically large. Truth-proven cases are a precious commodity in observer performance studies. Particle physicists working on discovering the Higg's Boson can get away with 5σ criterion because the number of independent observations and/or effect size is much larger than corresponding numbers in observer performance research.

8.8: Discussion / Summary

In most statistical books, the subject of hypothesis testing is demonstrated in different (i.e., non-ROC) contexts. That is to be expected since this field is a very small subspecialty of statistics (Prof. Howard E. Rockette, private communication, ca. 2002). Since this book is about ROC analysis, the author decided to use a demonstration using ROC analysis. Using a data simulator, one is allowed to "cheat" by conducting a very large number of simulations to estimate the population AUC under the null hypothesis. This permitted us to explore the related concepts of Type-I and Type-II errors within the context of ROC analysis. Ideally, both errors should be zero, but the nature of statistics leads one to two compromises. Usually one accepts a Type-I error capped at 5% and a Type-II error capped at 20%. These translate to α = 0.05 and desired statistical power = 80%. The dependence of statistical power on α , the numbers of cases and the effect size was explored. Statistical power increases with the effect size, it increases with α and it increases with the sample size (numbers of cases).

In **Chapter 11** sample-size calculations are described that allow one to estimate the numbers of readers and cases needed to *detect* a specified difference in inter-modality AUCs with an expected statistical power 1 - β . The word "*detect*" in the preceding sentence is shorthand for "reject the NH with probability capped at α *while* also rejecting the alternative hypothesis with probability capped at β "; see Table 8.1.

This chapter also gives the first example of validation of a hypothesis testing method. Statisticians sometimes refer to this as showing a proposed test is a "5% test". What is meant is that one needs to be assured that when the NH is true the probability of NH rejection equals the expected value, namely α, typically chosen to be 5%. Since the observed NH rejection rate over 2000 simulations is a random variable, one does not expect the NH rejection rate to exactly equal 5%, rather the constructed 95% confidence interval (also a random interval variable) should include the NH value with probability α.

As noted in the introduction, comparing a single reader's performance to a specified value is not a clinically interesting problem. The next two chapters describe methods for significance testing of multiple-reader multiple-case (MRMC) ROC datasets, consisting of interpretations by a group of readers of a common set of cases in typically two modalities. It turns out that the analyses yield variability estimates that permit sample size calculation. After all, sample size calculation is all about estimation of variability, the denominator of the z-statistic, i.e., Eqn. (8.3), in the context of this chapter. The formulae will look more complex, as interest is not in determining the standard deviation of AUC, but in the standard deviation of the inter-modality reader-averaged AUC difference. However, the basic concepts remain the same.

8.9: References

1. Larsen RJ, Marx ML. *An Introduction to Mathematical Statistics and Its Applications.* 3rd ed. Upper Saddle River, NJ: Prentice-Hall Inc; 2001.
2. Chakraborty DP, Breatnach ES, Yester MV, Soto B, Barnes GT, Fraser RG. Digital and Conventional Chest Imaging: A Modified ROC Study of Observer Performance Using Simulated Nodules. *Radiology.* 1986;158:35-39.