# ANSWER SHEET 5

**Assignment 1.** (a) $\widehat{\mu} = \overline{X}$.
(b) Using the CLT, it follows that $\sqrt{n}(\widehat{\mu} - \mu) \xrightarrow{d} N(0, 1)$ as $n \to \infty$.
(c) Since $\widehat{\mu}$ is the MLE of $\mu$, using a theorem done in the class, it follows that $\sqrt{n}(\widehat{\mu} - \mu) \xrightarrow{d} N(0, \mathcal{I}_1^{-1}(\mu))$, where $\mathcal{I}_1(\mu)$ is the Fisher information of $\mu$ from a single sample. Thus, it follows from part (b) that $\mathcal{I}_1(\mu) = 1$. So, $\mathcal{I}_n(\mu) = n\mathcal{I}_1(\mu) = n$. Hence, the Cramer-Rao lower bound for the variance of an unbiased estimator of $\mu$ is $\mathcal{I}_n^{-1}(\mu) = n^{-1}$.
(d) Since $\text{Var}(\overline{X}) = n^{-1}$, it follows that $\widehat{\mu}$ satisfies the Cramer-Rao lower bound for all $n \geq 1$.
(e) $g(\mu) = \mathbb{P}[X_1 \leq 2] = \Phi(2 - \mu)$, where $\Phi$ is the cdf of the $N(0, 1)$ distribution.
(f) Since $g$ is a bijective function from $\mathbb{R}$ to $(0, \infty)$, it follows from the equivariance property of MLEs that the MLE of $g(\mu)$ is $g(\widehat{\mu}) = \Phi(2 - \overline{X})$.
(g) Note that $g'(\mu) = -\phi(2 - \mu)$, where $\phi$ is the density function of the $N(0, 1)$ distribution. So, using the delta method, it follows that $\sqrt{n}\{g(\widehat{\mu}) - g(\mu)\} \xrightarrow{d} N(0, [g'(\mu)]^2) \equiv N(0, \phi^2(2 - \mu))$ as $n \to \infty$.

**Assignment 2.** (a) Note that the likelihood $L(\alpha)$ is zero outside of the set $\mathbf{1}(x_{(1)} \geq \pi)$, where $x_{(1)} = \min\{x_1, x_2, \ldots, x_n\}$. So, it is good enough to consider the maximization of $L(\alpha)$ when the sample points satisfy this condition. Then,

$$
\begin{aligned}
L(\alpha) &= \prod_{i=1}^{n} \left\{ \frac{\alpha \pi^\alpha}{x_i^{\alpha+1}} \right\} = \frac{\alpha^n \pi^{n\alpha}}{\left(\prod_{i=1}^{n} x_i\right)^{\alpha+1}} \\
\Rightarrow \quad \log L(\alpha) &= n \log \alpha + n\alpha \log \pi - (\alpha + 1) \sum_{i=1}^{n} \log x_i \\
\Rightarrow \quad \frac{\partial}{\partial \alpha} \log L(\alpha) &= \frac{n}{\alpha} + n \log \pi - \sum_{i=1}^{n} \log x_i.
\end{aligned}
$$

Setting $\Delta_\alpha \log L(\alpha) = 0$ yields the solution

$$
\widehat{\alpha} = \frac{n}{\sum_{i=1}^{n} \log x_i - n \log \pi}.
$$

Since $\partial^2 \log L(\alpha)/\partial \alpha^2 = -n/\alpha^2 < 0$, it follows that the $\widehat{\alpha}$ is the unique maximizer and hence the MLE of $\alpha$.
(b) Observe that

$$
\mathcal{I}_n(\alpha) = \mathbb{E}\left[-\frac{\partial^2 \log L(\alpha)}{\partial \alpha^2}\right] = \frac{n}{\alpha^2}.
$$

So, $\mathcal{I}_1(\alpha) = \alpha^{-2}$. Thus, using a theorem done in the class, it follows that

$$
\sqrt{n}(\widehat{\alpha} - \alpha) \xrightarrow{d} N(0, \alpha^2)
$$

as $n \to \infty$.
(c) Note that for any $y > 0$, we have

$$
\mathbb{P}[Y \leq y] = \mathbb{P}[X \leq \pi \exp(y)] = \int_{\pi}^{\pi \exp(y)} \frac{\alpha \pi^\alpha}{x^{\alpha+1}} dx = \pi^\alpha \left[\pi^{-\alpha} - (\pi \exp(y))^{-\alpha}\right] = 1 - \exp(-\alpha y).
$$

So, the density of $Y$ is given by $f_Y(y) = \alpha \exp(-\alpha y)$ if $y > 0$, and equals zero otherwise. Thus, $Y \sim Exp(\alpha)$.
(d) We know that the mean and the variance of the $Exp(\alpha)$ distribution are $\alpha^{-1}$ and $\alpha^{-2}$, respectively. So, using the CLT, we have

$$\sqrt{n}\left(\frac{T(Y_1, Y_2, \ldots, Y_n)}{n} - \frac{1}{\alpha}\right) \xrightarrow{d} N\left(0, \frac{1}{\alpha^2}\right)$$

as $n \to \infty$.
(e) $\widehat{\alpha} = n/T(Y_1, Y_2, \ldots, Y_n)$.
(f) Define $h(x) = x^{-1}$ on $(0, \infty)$. So, $h'(x) = -x^{-2}$. Using the delta method, it follows that

$$\sqrt{n}\left\{h\left(\frac{T(Y_1, Y_2, \ldots, Y_n)}{n}\right) - h\left(\frac{1}{\alpha}\right)\right\} \xrightarrow{d} N\left(0, \left[h'\left(\frac{1}{\alpha}\right)\right]^2 \frac{1}{\alpha^2}\right)$$

$$\Rightarrow \quad \sqrt{n}(\widehat{\alpha} - \alpha) \xrightarrow{d} N\left(0, \alpha^2\right)$$

as $n \to \infty$. This is the same asymptotic distribution as that obtained in part (b).

**Assignment 3.** (a) The Neyman–Pearson most powerful test is given by the rejection region

$$\frac{(\sqrt{2\pi}\sigma_1)^{-n} \exp\{-\sum_{i=1}^n X_i^2/(2\sigma_1^2)\}}{(\sqrt{2\pi}\sigma_0)^{-n} \exp\{-\sum_{i=1}^n X_i^2/(2\sigma_0^2)\}} > k$$

which is equivalent to

$$\sum_{i=1}^n X_i^2 < c$$

since $\sigma_1^2 < \sigma_0^2$.
To determine the critical value $c$ it is enough to realize that the $\sum_{i=1}^n X_i^2/\sigma_0^2 \sim \chi_n^2$ under $H_0$. Denote the $\alpha$-quantile of the $\chi_n^2$ distribution by $c_\alpha$, i.e., $c_\alpha = H_n^{-1}(\alpha)$, where $H_n$ is the cdf of the $\chi_n^2$ distribution. Therefore, the critical value is $c = \sigma_0^2 c_\alpha$.
No, the critical value of the test does not depend on $\sigma_1^2$.
(b) The power against $\sigma_1^2 < \sigma_0^2$ equals

$$P_{\sigma_1^2}\left(\frac{\sum_{i=1}^n X_i^2}{\sigma_0^2} < c_\alpha\right) = P_{\sigma_1^2}\left(\frac{\sum_{i=1}^n X_i^2}{\sigma_1^2} < c_\alpha \frac{\sigma_0^2}{\sigma_1^2}\right) = H_n\left(c_\alpha \frac{\sigma_0^2}{\sigma_1^2}\right).$$

(c) The minimal sample size needed to reject $H_0$ with probability $\beta$ when the true variance is $\sigma_1^2$ is given implicitly as the solution to

$$H_n\left(c_\alpha \frac{\sigma_0^2}{\sigma_1^2}\right) \geq \beta.$$

**Assignment 4.** (a) The Neyman–Pearson test rejects for

$$\frac{p_1^T (1 - p_1)^{n-T}}{p_0^T (1 - p_0)^{n-T}} > k,$$

where $T = \sum_{i=1}^n X_i$. Equivalently, it rejects for

$$\left[\frac{p_1(1 - p_0)}{p_0(1 - p_1)}\right]^T \left[\frac{1 - p_1}{1 - p_0}\right]^n > k.$$

Since $\frac{p_1(1-p_0)}{p_0(1-p_1)} > 1$ (because $p_1 > p_0$), the critical region can be further simplified to $T > c$. Under $H_0$, the statistic $T = \sum_{i=1}^n X_i \sim Bin(n, p_0)$. Let $c = c_{1-\alpha}$ be the $(1-\alpha)$-quantile of this distribution, i.e., $c_{1-\alpha} = \inf\{x : G_{n,p_0}(x) \geq 1-\alpha\} = G^-_{n,p_0}(1-\alpha)$, where $G_{n,p_0}$ denotes the cdf of the $Bin(n, p_0)$ distribution. If $1 - G_{n,p_0}(c_{1-\alpha}) = \alpha$, then the test $T > c_{1-\alpha}$ is the most powerful test of significance level $\alpha$ of the test. Otherwise, when

$$\mathbb{P}_{H_0}(T > c_{1-\alpha}) < \alpha < \mathbb{P}_{H_0}(T \geq c_{1-\alpha}),$$

we do not get a most powerful test.
No, when a most powerful test exists, the critical value of the test does not depend on $p_1$.
(b) For $p_0 = \frac{3}{10}$, $n = 3$, we have $P(T=3) = \binom{3}{3}(\frac{3}{10})^3 = \frac{27}{1000}$, $P(T=2) = \binom{3}{2}(\frac{3}{10})^2(1-\frac{3}{10}) = \frac{189}{1000}$. Thus, rejecting for $T > 2$ would give level 0.027 while rejecting for $T > 1$ would give level $0.027+0.189 = 0.216$. Therefore, there does not exist a most powerful test at significance level $\alpha = 0.05$.
(c) However, there exists a most powerful test at the significance level $\alpha = 0.027$. It is given by $\delta(X_1, X_2, \ldots, X_n) = \mathbf{1}(T > 2)$.
(d) The statistic

$$\frac{T - np_0}{\sqrt{np_0(1-p_0)}}$$

is asymptotically standard normal under $H_0$. Hence, the asymptotic significance level-$\alpha$ test rejects when this statistic exceeds the $(1-\alpha)$-quantile of the $N(0,1)$ distribution.

**Assignment 5.** (a) The likelihood ratio is

$$\Lambda(X_1, \ldots, X_n) = \frac{5^n e^{-5\sum X_i}}{4^n e^{-4\sum X_i}} = \exp(n\log(5/4) - \sum X_i).$$

We reject $H_0$ if $\Lambda(X_1, \ldots, X_n)$ is large; equivalently, if $T(X_1, \ldots, X_n) = \sum X_i$ is small. The test function is therefore $\delta(X_1, \ldots, X_n) = 1$ if $T \leq q$ and 0 otherwise, where $q$ is such that $\mathbb{P}(\sum X_i \leq q) = \alpha$.
(b) The moment generating function of the sum is

$$\prod_{i=1}^n \frac{\lambda}{\lambda - t} = \left(\frac{\lambda}{\lambda - t}\right)^n$$

which is the moment generating function of a $Gamma(n, \lambda)$ random variable.
(c) The test function is the same for all $\lambda_1 > 4$. This test can be shown to be uniformly optimal. (If $\lambda_1 < 4$, we would reject when $\sum X_i$ is large.)
(d) Under $H_0$, $T \sim Gamma(n, 4)$. Therefore $q$ is the $\alpha$-quantile of the $Gamma(n, 4)$ distribution. This is a continuous distribution, so $q$ exists, and it is unique because the density is positive on $[0, \infty)$.
(e)–(g) We may use the following code :

```
set.seed(18102017)
lambda <- 4
n <- 17
REP <- 1000
alpha <- 0.05
```

```
rej <- logical(REP)
q <- qgamma(alpha, shape = n, rate = 4)
for(i in 1:REP)
{
  X <- rexp(n, rate = lambda)
  rej[i] <- (sum(X) <= q)  #### returns 1 if the condition is satisfied, 0 otherwise
}
mean(rej)
```

(h) When $\lambda = 4$, we indeed reject approximately 50 times, namely 5%. When $\lambda = 3$ we reject less ; the test is conservative and the type I error is smaller than 5%. When $\lambda = 5$ we reject more (209 times in this particular example), so the power is approximately 0.209 ; as $\lambda$ becomes larger we reject more and more and the power increases and approaches one. These two phenomena (increase of power and decrease of type I error) occur more rapidly the larger $n$ is.

**Assignment 6.** The optimal value of $a$ is $n-2 = 1$. The assignment can be carried out using the following code :

```
set.seed(18102017)
n <- 3
REP <- 1000
mu <- c(-1, 0, 1)
a <- n-2
MSE.mle <- MSE.stein <- numeric(REP)
for(i in 1:REP)
{
Y <- rnorm(n, mean = mu, sd = 1)
Y.norm <- sum(Y^2)
stein <- Y * (1 - a/Y.norm)
MSE.mle[i] <- sum((Y - mu)^2)
MSE.stein[i] <- sum((stein - mu)^2)
}
mean(MSE.mle)
mean(MSE.stein)
```

**Assignment 7.** (a) We have

$$\mathbb{E}\,\frac{1}{X} = \int_0^\infty \frac{\lambda^k x^{k-2} e^{-\lambda x}}{\Gamma(k)} dx = \frac{\lambda}{k-1} \int_0^\infty \frac{\lambda^{k-1} x^{k-2} e^{-\lambda x}}{\Gamma(k-1)} dx = \frac{\lambda}{k-1},$$

since $k > 1$ and the last integrand is the density of a $Gamma(k-1, \lambda)$ distribution. If $k \leq 1$, then $\mathbb{E}\,\frac{1}{X} = \infty$.
(b) Put $\lambda = 1/2$ and $k = n/2 > 1$ because $n > 2$.
(c) Up to constants, the log likelihood is the negative of this sum of squares.
(d) The additive nature of the objective function allows for minimisation each $\mu_i$ separately. The first derivatives with respect to $\mu_i$ are

$$2\mu_i - 2y_i + 2\lambda\mu_i = 2[(1+\lambda)\mu_i - y_i]; \qquad \text{and} \qquad 2(1+\lambda) > 0$$

so the unique minimum is attained at $\tilde{\mu}_i = y_i/(1+\lambda)$. In vector form, this can be written $\tilde{\mu}_\lambda = y/(1+\lambda)$.
(e) The mean squared error can be written as the expected value of

$$\sum_{i=1}^n (\tilde{\mu}_i - \mu_i)^2 = (1+\lambda)^{-2} \sum_{i=1}^n (y_i - \mu_i - \lambda\mu_i) = (1+\lambda)^{-2} \sum_{i=1}^n (y_i - \mu_i)^2 + \lambda^2 \mu_i^2 - 2\lambda\mu_i(y_i - \mu_i).$$

Since $\mathbb{E}\, y_i = \mu_i$ the last term vanishes and since $\operatorname{Var} y_i = 1$ the first sum is $n$. Thus the mean squared error equals

$$\frac{1}{(1+\lambda)^2}\left(n + \lambda^2 \sum_{i=1}^n \mu_i^2\right) = \frac{1}{(1+\lambda)^2}\left(n + \lambda^2 \|\mu\|^2\right).$$

(f) The derivative of the mean squared error with respect to $\lambda$ is

$$\frac{2\lambda\|\mu\|^2(1+\lambda)^2 - 2(1+\lambda)(n+\lambda^2\|\mu\|^2)}{(1+\lambda)^4} = \frac{2}{(1+\lambda)^3}\left(\lambda\|\mu\|^2 - n\right).$$

This is negative for small $\lambda$, so small but positive values of $\lambda$ have a lower mean squared error than that of $\widehat{\lambda} = \tilde{\lambda}_0$.
(g) Since the derivative is negative for small $\lambda$ and positive for for large $\lambda$, the unique minimum is attained when $\lambda = n/\|\mu\|^2$ (if $\mu \neq 0$). What this means is that the smaller $\|\mu\|$ is, the better it is to penalise it by choosing a high value of $\lambda$. In the extreme case where $\mu = 0$, the mean squared error is $n/(1+\lambda)^2$, which is strictly decreasing; the more we penalise, the better.
The problem with this choice of $\lambda$ is that it depends on the unknown value of $\mu$. We will later see some ways of choosing $\lambda$ in practice, most notably *cross-validation.* Note that this problem does not arise with the James–Stein estimator.

**Assignment 8.** The likelihood function for the sample is the joint probability function of all $x_i$'s and $y_i$'s and is given by

$$L(\theta_1, \theta_2) = \prod_{i=1}^n \frac{\theta_1^{x_i} e^{-\theta_1}}{x_i!} \prod_{i=1}^n \frac{\theta_2^{y_i} e^{-\theta_2}}{y_i!} = \left(\frac{1}{k}\right) \theta_1^{\sum_{i=1}^n x_i} e^{-n\theta_1} \theta_2^{\sum_{i=1}^n x_i} e^{-n\theta_2}$$

where $k = x_1! \dots x_n! y_1! \dots y_n!$ and $n = 100$.
We can see that $L(\theta_1, \theta_2)$ is maximised when both $\theta_1$ and $\theta_2$ are equal to their m.l.e. $\theta_1 = \bar{x}$ and $\theta_2 = \bar{y}$.
Moreover under $H_0$ the likelihood is

$$L(\theta) = \left(\frac{1}{k}\right) \theta^{\sum_{i=1}^n x_i + \sum_{i=1}^n y_i} e^{-2n\theta},$$

a function of only one parameter $\theta = \theta_1 = \theta_2$ maximised in

$$\hat{\theta} = \frac{1}{2n}\left(\sum_{i=1}^n x_i + \sum_{i=1}^n y_i\right) = \frac{1}{2}(\bar{x} + \bar{y}).$$

In this example the parameter space is $\Theta = \{(\theta_1, \theta_2) : \theta_1 > 0, \theta_2 > 0\}$, and we can write the likelihood ratio as

$$\Lambda = \frac{L(\hat{\theta})}{L(\theta_0)} = \frac{\bar{x}^{n\bar{x}} \bar{y}^{n\bar{y}}}{\hat{\theta}^{n\bar{x}+n\bar{y}}}.$$

(i). The value for $\Lambda$ would be intractable to compute, but we can compute $\log(\Lambda) = 4.76$.

(ii). $2\log\Lambda$ is an approximate $\chi_1^2$ distribution, therefore we would reject the null hyphothesis for value of $2\log\Lambda$ larger than the $k = 7.879$, where $k$ is such that $\mathbb{P}_{\not\vdash}[\Lambda \geq k] = \alpha$. In our case $2\log\Lambda = 9.52$ hence we reject the null hypothesis $\theta_1 = \theta_2$.

**Assignment 9.** The table mimics table on Slide 192. If we wanted to test $H_0$ : patient $i$ is healthy, then the cell corresponding to FP is related to the Type I error (N.B. not exactly type 1 error !).
For our case study the table would be

| | healthy | desease |
|---|---|---|
| $GLI < s$ | 95 | 2 |
| $GLI > s$ | 7 | 8 |

To write down the standardize table we need to divide each cell in column $i$, $i = 1, 2$ by the total number of elements in that column, yielding

| | healthy | desease |
|---|---|---|
| $GLI < s$ | 0.95 | 0.2 |
| $GLI > s$ | 0.07 | 0.8 |

(i). If we were screening for a cold, we could take a higher $s$. When $s$ increase, the probability of the false alarm and the detection power decrease.
On the contrary if we were to screen for a serious desease, we would tollerate a higher percentage of false positive and a lower of false negative, because missing the desease would have more serious consequence than performing another test. We would increase the significance level of our test, losing power on the way.

(ii). For $s \to \infty$ the true negative rate will decay to 0 and the table for our data will be

| | healthy | desease |
|---|---|---|
| $GLI < s$ | 100 | 10 |
| $GLI > s$ | 0 | 0 |

which standardized becomes

| | healthy | desease |
|---|---|---|
| $GLI < s$ | 1 | 1 |
| $GLI > s$ | 0 | 0 |

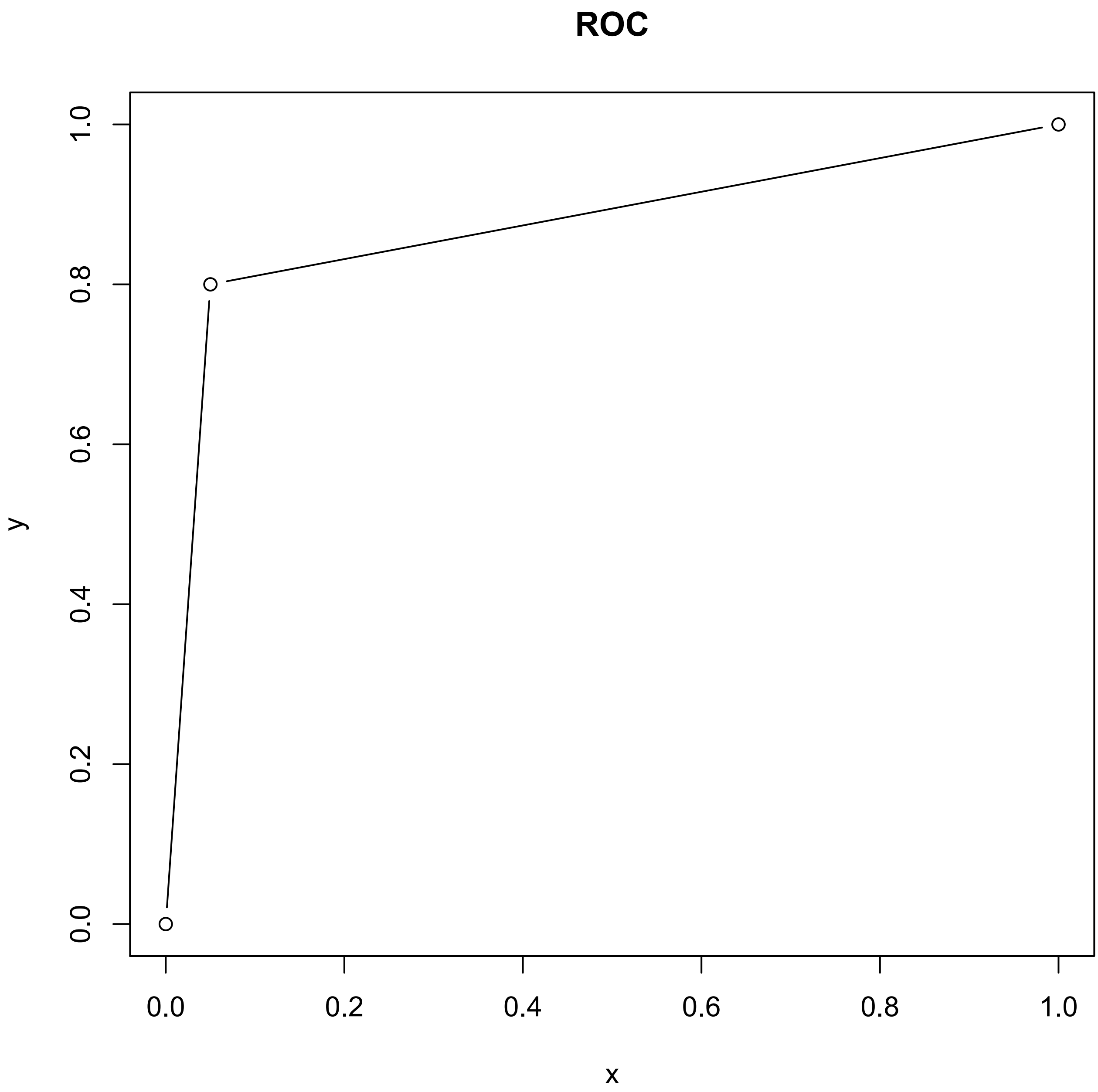
.

For $s \to -\infty$ instead the situation is the following

| | healthy | desease |
|---|---|---|
| $GLI < s$ | 0 | 0 |
| $GLI > s$ | 100 | 10 |

which once standardized becomes

| | healthy | desease |
|---|---|---|
| $GLI < s$ | 0 | 0 |
| $GLI > s$ | 1 | 1 |




(iii).

The best possible prediction method would yield a point in the upper left corner. Ideally it would have coordinate (0,1). The diagonal divides the ROC space. Points above the diagonal represent good classification results, points along the diagonal represent the classification we would have by randomly throwing a coin.