Homework_2

Håkon Sandaker Vincent Wilmet

3/25/2021

Comparison of estimators

a) MLE with expectation and variance

We have the Binomial Distribution

$$Bin(n, \theta) \sim \binom{n}{k} \theta^k (1-\theta)^{n-k}$$

Log-Likelihood:

$$\begin{aligned} \text{log-likelihood} &= log(Bin(n, \theta)) \\ &= log(\binom{n}{k} \theta^k (1-\theta)^{n-k}) \\ &= log(\binom{n}{k}) + k * log(\theta) + (n-k) * log(1-\theta) \end{aligned}$$

MLE:

$$\begin{aligned} \frac{\partial \text{ log-likelihood}}{\partial \theta} &= 0 \\ \frac{\partial log(\binom{n}{k}) + k * log(\theta) + (n-k) * log(1-\theta)}{\partial \theta} &= 0 \\ \frac{k}{\theta} - \frac{n-k}{1-\theta} &= 0 \\ \frac{k}{\theta} &= \frac{n-k}{1-\theta} \\ \frac{1-\theta}{\theta} &= \frac{n-k}{k} \\ \frac{1}{\theta} - 1 &= \frac{n}{k} - 1 \\ \frac{1}{\theta} &= \frac{n}{k} \\ \theta &= \frac{k}{n} \\ \hat{\theta}_{MLE} &= \frac{k}{n} \end{aligned}$$

Expectation and Variance:

$$E[\hat{\theta}_{MLE}] = E[\frac{k}{n}] = \frac{1}{n} E[k] = \frac{\theta n}{n} = \theta$$

$$
\begin{aligned}
Var(\hat{\theta}_{MLE}) &= Var(\frac{k}{n}) \\
&= \frac{1}{n^2}Var(k) \text{ (by variance property)} \\
&= \frac{1}{n^2}n\theta(1-\theta) \\
&= \frac{\theta(1-\theta)}{n}
\end{aligned}
$$

b)

$$
\hat{\theta}_{alt.} = \frac{X+1}{n+2}
$$

$$
\begin{aligned}
E[\hat{\theta}_{alt.}] &= E[\frac{X+1}{n+2}] \\
&= \frac{E[X]+1}{n+2} \text{ (by linearity of expectation)} \\
&= \frac{n\theta+1}{n+2}
\end{aligned}
$$

$$
\begin{aligned}
Var(\hat{\theta}_{alt.}) &= Var(\frac{X+1}{n+2}) \\
&= \frac{1}{(n+2)^2}Var(X+1) \\
&= \frac{Var(X)}{(n+2)^2} \\
&= \frac{n\theta(1-\theta)}{(n+2)^2}
\end{aligned}
$$

c) MSE

MLE

$$
\begin{aligned}
MSE(\hat{\theta}_{MLE}) &= (E[\hat{\theta}_{MLE}]-\theta)^2 + Var(\hat{\theta}_{MLE}) \\
&= (\theta-\theta)^2 + \frac{\theta(1-\theta)}{n} \\
&= \frac{\theta(1-\theta)}{n}
\end{aligned}
$$

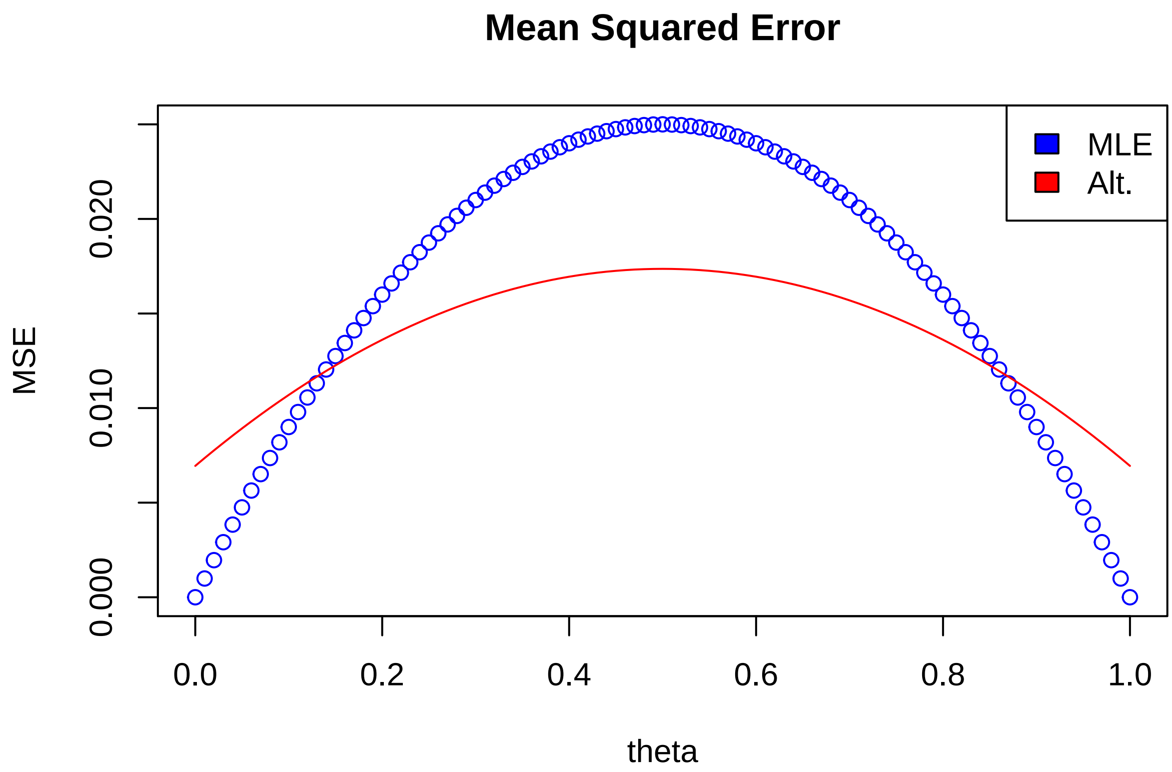
Alt

$$
\begin{aligned}
MSE(\hat{\theta}_{alt.}) &= (E[\hat{\theta}_{alt.}] - \theta)^2 + Var(\hat{\theta}_{alt.}) \\
&= (\frac{n\theta + 1}{n+2} - \theta)^2 + \frac{n\theta(1-\theta)}{(n+2)^2} \\
&= (\frac{n\theta + 1}{n+2} - \frac{\theta(n+2)}{n+2})^2 + \frac{n\theta(1-\theta)}{(n+2)^2} \\
&= (\frac{n\theta + 1}{n+2} - \frac{n\theta + 2\theta)}{n+2})^2 + \frac{n\theta(1-\theta)}{(n+2)^2} \\
&= (\frac{1 - 2\theta}{n+2})^2 + \frac{n\theta(1-\theta)}{(n+2)^2} \\
&= \frac{(1-2\theta)^2}{(n+2)^2} + \frac{n\theta(1-\theta)}{(n+2)^2} \\
&= \frac{(1-2\theta)^2 + n\theta(1-\theta)}{(n+2)^2} \\
&= \frac{(1-2\theta)^2 + n\theta(1-\theta)}{(n+2)^2} \\
&= \frac{1 - 4\theta + 4\theta^2 + n\theta - n\theta^2}{(n+2)^2} \\
&= \frac{1 - \theta(n-4)(\theta-1)}{(n+2)^2}
\end{aligned}
$$

MSE vs Alt. The maximum likelihood estimator of $\hat{\theta}_{MLE}$ is unbiased as it is equal to its true variance. Hence, we would rather use the MSE MLE over the MSE Alt. approach.

d) Comparisson

```r
n <- 10
theta <- seq(0, 1, 0.01)

MSE_MLE <- function(n, theta)
{
  return (theta*(1 - theta)/n)
}

MSE_Alt <- function(n, theta)
{
  return ((1 - theta*(n-4)*(theta-1))/ (n+2)^2)
}

# Plot the MLE & Alt
plot(theta, MSE_MLE(n, theta), col="blue", main="Mean Squared Error", ylab="MSE")
lines(theta, MSE_Alt(n, theta), col="red", main="Mean Squared Error", ylab="MSE")
legend("topright", c("MLE","Alt."), fill=c("blue", "red"))
```



2. Robustness of the estimators

a)

Find the MLE of the Guassian Distribution. Likelihood:

$$\begin{aligned}\text{likelihood} &= \prod_n \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\theta}{\sigma})^2} \\ &= \prod_n \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2}, \text{from the fact that } \sigma \text{ is } 1 \\ &= \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2}\end{aligned}$$

Log-Likelihood:

$$\begin{aligned}\text{log-likelihood} &= log(\frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2}) \\ &= -\frac{n}{2} log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2\end{aligned}$$

MLE:

$$\frac{\partial \text{ log-likelihood}}{\partial\theta} = 0$$

$$\frac{\partial - \frac{n}{2}log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2}{\partial\theta} = 0$$

$$\sum_{i=1}^{n}(x_i-\theta) = 0$$

$$\sum_{i=1}^{n}x_i - n\theta = 0$$

$$\theta = \frac{\sum_{i=1}^{n}x_i}{n} = \bar{x}$$

Thus, $\hat{Q}^{MV} \sim N(\bar{X}_n, 1)$

b)

Of course, the data might not follow the Gaussian distribution. If thats the case, the MLE will not work. However, if the size of sample n is sufficiently large, by CLT we can say that the distribution will converge towards a Gaussian distribution, at which point $\hat{Q}^{MV}$ is an appropriate estimator.

c)

Expectation:

$$E[Q] = 0.99 * E[P_0] + 0.01 * E[P_{300}] = 0 + 0.01 * 300 = 3$$

Variance:

$$Var(Q) = Var(0.99 * P_0) + Var(0.01 * P_{300}) = 0.99^2 * Var(P_0) + 0.01^2 * Var(P_{300}) = 0.99^2 + 0.01^2 = 0.9802.$$

Since the variance of Q is not equal to 1 it does not belong to the model $N(\theta, 1)$. On the contrary, Q is a mixture of the two distributions.

Density of Q:

$$f_x(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-3)^2}$$

d)

Yes, but $\hat{Q}^{MV}$ is not identical to Q. $\hat{Q}^{MV} \sim N(\bar{X}_n, 1) \sim N(3, 1)$, but $Q \sim N(3, 0.99)$ with a different variance.

e)

We are given

$$\Phi^{-1}(\frac{1}{2 * 0.99}) = 0.013$$

Additionally, we have

$$Q = 0.99P_0 + 0.01P_{300}$$

This tells us that 99% of the mixed distribution is weighted towards the P_0 distribution. We have that $P_0 \sim N(0, 1)$. Furthermore, we have the property that the median of the standard normal distribution is the same as the mean. Hence the median of the quantile $\Phi^{-1}(\frac{1}{2*0.99}) = 0.013$ is supposed to be close to zero, which it is.

3. Hypothesis testing and doping controls

a)

Hematocrit levels in the blood:

$$Y \sim N(45, 2)$$

Observed values:

$$X \sim N(\mu_1, 2)$$

$$X \sim N(\mu_2, 2)$$

Want to check if expectation of X is equal or greater than the mean of 45.

b)

Hypothesis.

$$H_0 : \mu_i = 45$$
$$H_1 : \mu_i > 45$$

c)

Estimator.

$$Z = \frac{\bar{x}_i - 45}{\sqrt{2}}, Z \sim N(0, 1)$$

With H_0 Z follows N(0, 1). With H_1 Z does not follow N(0, 1).

d)

```
qnorm(.95)
```

```
## [1] 1.644854
```

With $p - value = .05$ and a right tail we get a Z-score of 1.64. Hence, we reject the null hypothesis if Z value is greater than 1.64.

e)

i.

```r
pnorm(45, mean=45, sd=2, lower.tail=FALSE)
```

```
## [1] 0.5
```

ii.

```r
pnorm(60, mean=45, sd=2, lower.tail=FALSE)
```

```
## [1] 3.190892e-14
```

Very unlikely.

iii.

```r
cv <- qnorm(.95, mean=45, sd=2)
cv
```

```
## [1] 48.28971
```

f)

i.

Reject if Z is above 1.645.

ii.

J.C: $Z - score = \frac{48-45}{\sqrt{2}} = 2.121$. Reject. S.R: $Z - score = \frac{50-45}{\sqrt{2}} = 3.53$. Reject.

iii. using the student population t test

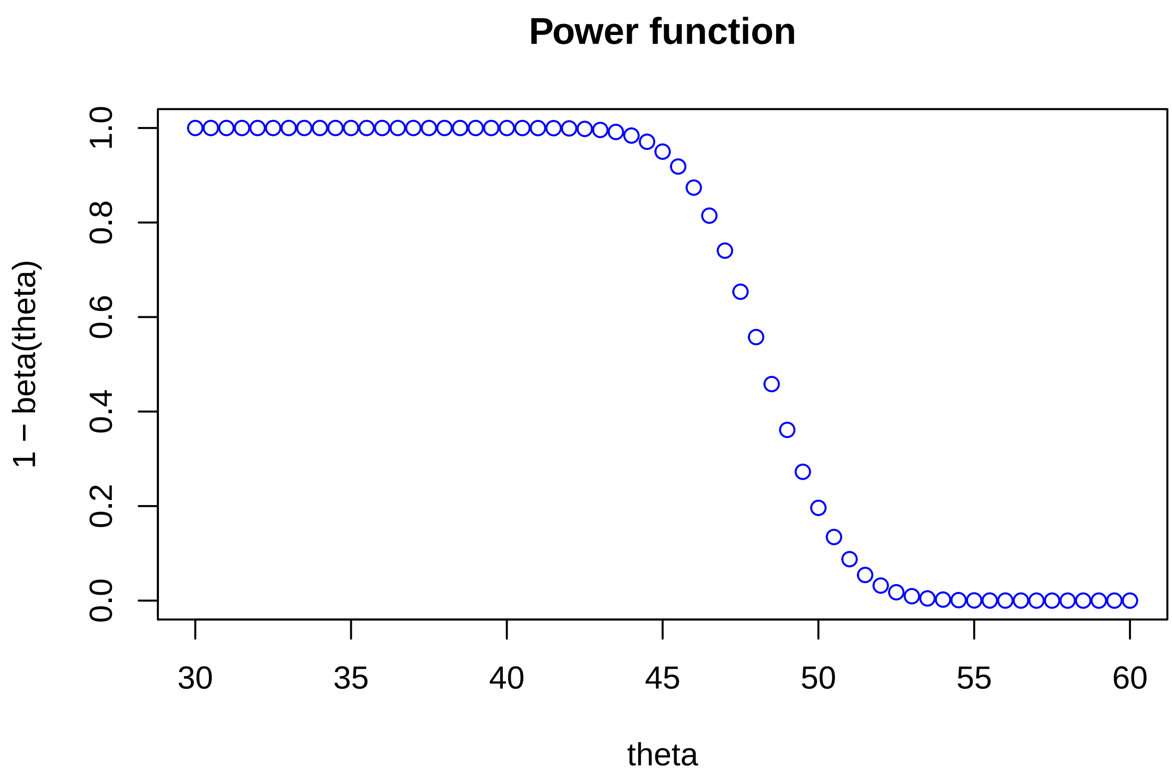
```r
typeI.test <- function(mu0, sigma, n, alpha, iterations = 10000) {
  pvals <- rep(NA, iterations)
  for(i in 1 : iterations){
    temporary.sample <- rnorm(n = n, mean = mu0, sd = sigma)
    temporary.mean <- mean(temporary.sample)
    temporary.sd <- sd(temporary.sample)
    pvals[i] <- 1 - pt((temporary.mean - mu0)/(temporary.sd / sqrt(n)), df = n-1)
  }
  return(mean(pvals < alpha))
}

typeI.test(mu0 = 45, sigma = 2, n = 5, alpha = 0.05)
```

```
## [1] 0.0499
```

g)

i.

```
theta <- seq(30, 60, 0.5)
# changed from mean=45 to critical value=48.28971
plot(theta, 1 - pnorm(theta, mean=cv, sd=2), col="blue", main="Power function", xlab="theta", ylab="1 -
```



ii.

By looking at the Power function, the probability of detecting an abnormal hematocrit level is around 19%.

```
1 - pnorm(50, mean=cv, sd=2)
```

```
## [1] 0.1962351
```

4.

a)

```r
df <- read.csv("poulpeF.csv")
head(df)
```

```
##   Poids
## 1  1300
## 2   300
## 3   900
## 4   120
## 5   180
## 6    80
```

b)

Mean and Variance:

```r
mean <- mean(df$Poids)
variance <- var(df$Poids)
mean
```

```
## [1] 639.625
```

```r
variance
```

```
## [1] 198823.7
```

Under the MLE:

```r
# MASS library with fitdistr
library(ggplot2)
library(MASS)
library(stats4)
MLE <- fitdistr(as.numeric(df$Poids), "normal")
MLE$estimate["mean"]
```

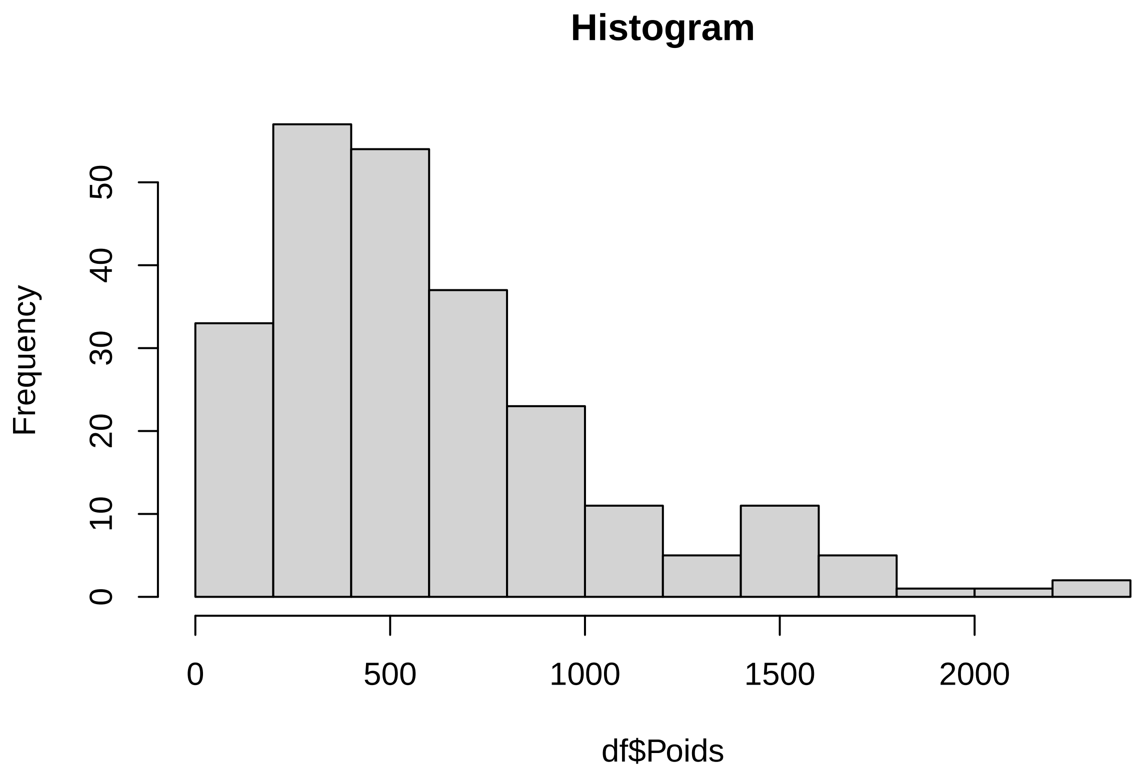
```
##    mean
## 639.625
```

```r
MLE$estimate["sd"]^2
```

```
##       sd
## 197995.3
```

c)

```r
hist(df$Poids, main="Histogram")
```



No, it is skewed towards the right. Hence, basing it on a Gaussian can be a problem.

d)

95% Confidence interval

$$\bar{X} \pm Z * \frac{std}{\sqrt{n}}$$

```
err <- qnorm(.975) * sqrt(variance) / sqrt(length(df$Poids))
# left, lower bound
mean - err
```

```
## [1] 583.2123
```

```
# right, upper bound
mean + err
```

```
## [1] 696.0377
```