**Math-661: Assignment 1 – Sample Solution**

1. **Exercise 1: Comparison of drugs**

   In a crossover trial comparing a new drug to a standard, $\pi$ denotes the probability that the new one is judged better. It is desired to estimate $\pi$ and test $H_0 : \pi = 0.50$ agains $H_1 : \pi \neq 0.50$. The new drug is found to be better in 15 out of 20 independent observations.

   (a) **Find and sketch the log-likelihood function. Is it close to the quadratic shape that large-sample normal approximations utilize?**

   Let $X$ = number of times the new drug is better in 20 independent observations.

$$X \sim \text{ Binomial}(20, \pi) \qquad p(x|\pi) = \binom{20}{x} \pi^x (1-\pi)^{20-x}, \quad x = 0, \ldots, 20$$

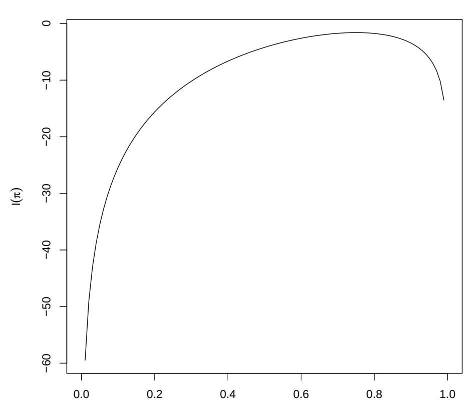
   Given that the new drug is better in 15 out of the 20 trials, the likelihood function and the log-likelihood function of $\pi$ are

$$\mathcal{L}(\pi) = \binom{20}{15} \pi^{15} (1-\pi)^5 \qquad \text{and} \qquad l(\pi) = \log \binom{20}{15} + 15 \log(\pi) + 5 \log(1-\pi), \qquad 0 \leq \pi \leq 1$$

   The plot of the log-likelihood function is given in Figure 1

```
fun.logl = function(x) log(choose(20,15))+15*log(x) + 5*log(1-x)
curve(fun.logl, 0, 1, xlab=expression(pi), ylab=expression(l(pi)))
```

   Figure 1: Log-likelihood functions of $\pi$ with15 successes out of 20 trials




   The log-likelihood function has a convex shape that is somewhat close to the quadratic shape used by large-sample normal approximations.

(b) **Give the ML estimate of $\pi$.**

The score function is

$$U(\pi) = \frac{d}{d\pi} l(\pi) = \frac{x}{\pi} - \frac{(n-x)}{1-\pi}$$

setting it to 0, we get the MLE, $\hat{\pi}$

$$\frac{x}{\pi} = \frac{n-x}{1-\pi} \qquad \Rightarrow \qquad \hat{\pi} = \frac{x}{n} = \frac{15}{20} = 0.75$$

(c) **Wald test** We want to test

$$H_0 : \pi = 0.5 \qquad \text{vs.} \qquad H_1 : \pi \neq 0.5$$

– **Conduct a Wald test, report the $p$-value and state your conclusion.**

The Wald test is given by

$$z = \frac{\hat{\pi} - \pi_0}{\sqrt{\widehat{Var}(\hat{\pi})}} \sim \; N(0,1)$$

where

$$\widehat{Var}(\hat{\pi}) = [I(\pi)|_{\pi=\hat{\pi}}]^{-1} = \frac{\hat{\pi}(1-\hat{\pi})}{n}$$

since

$$l(\pi) = x\log(\pi) + (n-x)\log(1-\pi) \quad \Rightarrow \quad U(\pi) = \frac{dl(\pi)}{d\pi} = \frac{x}{\pi} - \frac{n-x}{1-\pi} \quad \Rightarrow \quad \frac{d^2 l(\pi)}{d\pi^2} = -\frac{x}{\pi^2} - \frac{n-x}{(1-\pi)^2}$$

and the Fisher information is

$$I(\pi) = -E\left[\frac{d^2}{d\pi} l(\pi)\right] = \frac{n\pi}{\pi^2} + \frac{n(1-\pi)}{(1-\pi)^2} = \frac{n(1-\pi) + n\pi}{\pi(1-\pi)} = \frac{n}{\pi(1-\pi)}$$

So, the Wald statistic for the observed data is

$$z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}} = \frac{0.75 - 0.5}{\sqrt{\frac{0.75(1-0.75)}{20}}} = 2.582$$

which yields a a $p$-value $= 2P(Z \geq 2.582) \approx 0.0098$. Thus, we reject $H_0$ at $\alpha = 0.05$.

```
# two-sided p-value
> 2*(1-pnorm(2.582))
[1] 0.009822958
```

– **Construct a 95% Wald confidence interval for $\pi$ and interpret it.**

A 95% Wald confidence interval is given by

$$\hat{\pi} \pm z_{0.025}\sqrt{\widehat{Var}(\hat{\pi})} = 0.75 \pm 1.96\sqrt{\frac{0.75(1-0.75)}{20}} = (0.560, 0.940)$$

We are 95% confident the interval $(0.56, 0.94)$ contains the true $\pi$. That is, the drug is judged better at least 56% and at most 94% of the times. We note that the confidence interval does not contain the hypothesized 0.5 value.

(d) **Score test**

- **Conduct a score test, report the $p$-value and state your conclusion.**

The score test is given by

$$Q = \frac{U(\pi_0)^2}{I(\pi_0)} = \frac{\left[\frac{x}{\pi_0} - \frac{n-x}{1-\pi_0}\right]^2}{\frac{n}{\pi_0(1-\pi_0)}} = \frac{\left[\frac{x(1-\pi_0)-(n-x)\pi_0}{\pi_0(1-\pi_0)}\right]^2}{\frac{n}{\pi_0(1-\pi_0)}} = \frac{(x - n\pi_0)^2}{n\pi_0(1-\pi_0)} \sim \chi_1^2$$

For the observed data, we have

$$Q = \frac{\left(\frac{x}{n} - \pi_0\right)^2}{\frac{\pi_0(1-\pi_0)}{n}} = \frac{(0.75 - 0.5)^2}{\frac{0.5^2}{20}} = 5$$

which yields a $p$-value $= P(\chi_1^2 \geq 5) = 0.025$. Therefore, we reject $H_0$ at $\alpha = 0.05$.

```
> 1-pchisq(5,1)
[1] 0.02534732
```

- **Construct a 95% score confidence interval and interpret it.**

A 95% score confidence interval corresponds to the $\pi_0$ values such that

$$\frac{U(\pi_0)^2}{I(\pi_0)} < \chi^2_{1,0.05} = 3.84$$

that is,

$$\frac{\left[\frac{x}{n} - \pi_0)\right]^2}{\frac{\pi_0(1-\pi_0)}{n}} = \frac{(0.75 - \pi_0)^2}{\frac{\pi_0(1-\pi_0)}{20}} < 3.84$$

$$\Rightarrow \quad 20(0.75 - \pi_0)^2 < 3.84\pi_0(1-\pi_0) \qquad \Rightarrow \qquad 23.84\pi_0^2 - 33.84\pi_0 + 11.25 < 0$$

The solutions to the quadratic equation are

$$\pi_0 = \frac{33.84 \pm \sqrt{33.84^2 - 4 \times 23.84 \times 11.25}}{2 \times 23.84}$$

which correspond to 0.5313 and 0.8881.
Thus, a 95% score confidence interval for $\pi$ is $(0.5313, 0.8881)$. That is, we are 95% confidence that this interval contains the true probability, $\pi$, that the new drug is better.

(e) **Likelihood ratio test**

– **Conduct a likelihood ratio test, report the $p$-value and state your conclusion.**

The likelihood ratio test is given by

$$-2\log\Delta = 2\left[l(\hat{\pi}) - l(\pi_0)\right] \sim \ \chi^2_1$$

$$\begin{aligned}
-2\log\Delta &= 2\left[\{x\log\hat{\pi} + (n-x)\log(1-\hat{\pi})\} - \{x\log\pi_0 + (n-x)\log(1-\pi_0)\}\right] \\
&= 2\left[x\log\frac{\hat{\pi}}{\pi_0} + (n-x)\log\frac{1-\hat{\pi}}{1-\pi_0}\right] \\
&= 2\left[15\log\frac{0.75}{0.5} + 5\log\frac{0.25}{0.5}\right] = 5.232
\end{aligned}$$

The $p$-value $= P(\chi^2_1 \geq 5.232) = 0.022$. Therefore, we reject $H_0$ at $\alpha = 0.05$.

```
> 1-pchisq(5.232481,1)
[1] 0.02216888
```

– **Construct a likelihood-based 95% confidence interval and interpret it.**

A likelihood-based 95% confidence interval is the range of $\pi_0$ values such that

$$-2\log\Delta < \chi^2_{1,0.05} = 3.84$$

that is,

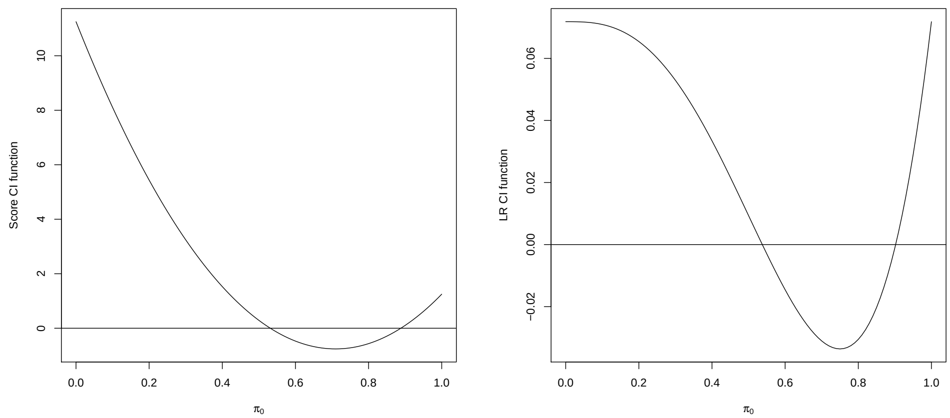
$$\begin{aligned}
& 2\left[15\log(0.75) + 5\log(0.25) - 15\log\pi_0 - 5\log(1-\pi_0)\right] < 3.84 \\
\Rightarrow\ & -22.49341 - 3.84 < 30\log\pi_0 + 10\log(1-\pi_0) \\
\Rightarrow\ & -26.33341 < 10\log(\pi_0^3(1-\pi_0)) \\
\Rightarrow\ & \exp(-2.633341) < \pi_0^3(1-\pi_0) \\
\Rightarrow\ & \pi_0^4 - \pi_0^3 + e^{-2.633341} < 0
\end{aligned}$$

We can use the function `uniroot.all` in the R package `rootSolve` to find the solutions of the quartic equation, which are 0.5376 and 0.9022.

```
fun.LRT = function(x) x^4-x^3+exp(-2.633341)
curve(fun.LRT, 0, 1, ylab="LR CI function"); abline(h=0)
uniroot.all(fun.LRT, c(0,1))
[1] 0.5375809 0.9021643
```

Thus, a likelihood-based 95% confidence interval for $\pi$ is $(0.5376, 0.9022)$. That is, we are 95% confident that the true probability $\pi$ that the new drug is better is between 53.8% and 90.2%.

Figure 2: CI functions based on score test and likelihood ratio test




2. **Exercise 2: Urea formaldehyde foam insulation**

Data were collected to check whether the presence of urea formaldehyde foam insulation (UFFI) has an effect on the ambient formaldehyde concentration ($CH_2O$) inside the house. Twelve homes with and 12 homes without UFFI were studied, and the average weekly $CH_2O$ concentration (in parts per billion) was measured. It was thought that the $CH_2O$ concentration was also influenced by the amount of air that can move through the house via windows, cracks, chimneys, etc. A measure of "air tightness", on a scale of 0 to 10, was determined for each home. $CH_2O$ concentration is the response variable ($Y$) that we try to explain through the two explanatory variables: air tightness of the home ($X_1$) and the absence/presence of UFFI ($X_2$). The data are provided in the file `UFFI.txt`

(1) **Give the $X$ matrix needed to fit the regression model**

$$\mathbf{Y} = X\boldsymbol{\beta} + \varepsilon \qquad \text{with} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

$$\mathbf{Y} = \begin{bmatrix} 31.33 \\ 28.57 \\ 39.95 \\ 44.98 \\ 39.55 \\ 38.29 \\ 50.58 \\ 48.71 \\ 51.52 \\ 62.52 \\ 60.79 \\ 56.67 \\ 43.58 \\ 43.30 \\ 46.16 \\ 47.66 \\ 55.31 \\ 63.32 \\ 59.65 \\ 62.74 \\ 60.33 \\ 53.13 \\ 56.83 \\ 70.34 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 4 & 0 \\ 1 & 4 & 0 \\ 1 & 5 & 0 \\ 1 & 7 & 0 \\ 1 & 7 & 0 \\ 1 & 8 & 0 \\ 1 & 8 & 0 \\ 1 & 8 & 0 \\ 1 & 9 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 2 & 1 \\ 1 & 4 & 1 \\ 1 & 4 & 1 \\ 1 & 5 & 1 \\ 1 & 5 & 1 \\ 1 & 6 & 1 \\ 1 & 6 & 1 \\ 1 & 7 & 1 \\ 1 & 9 & 1 \\ 1 & 10 & 1 \end{bmatrix}$$

(2) **Compute $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'Y$.**

$$X'X = \begin{bmatrix} 24 & 123 & 12 \\ 123 & 823 & 61 \\ 12 & 61 & 12 \end{bmatrix} \qquad X'Y = \begin{bmatrix} 1215.81 \\ 6776.22 \\ 662.35 \end{bmatrix}$$

```
> n=nrow(UFFI)
> Y = UFFI[,1]
> X = matrix(cbind(rep(1,n), as.matrix(UFFI[,-1])), ncol=3)
> XpY = crossprod(X,Y)
> XpX.inv = solve(crossprod(X))
> XpX.inv
            [,1]          [,2]          [,3]
[1,]  0.22194577 -0.0268282129 -0.0855690177
[2,] -0.02682821  0.0051925573  0.0004327131
```

```
[3,] -0.08556902  0.0004327131  0.1667027261
> beta.hat = XpX.inv%*%XpY
> beta.hat
          [,1]
[1,] 31.373371
[2,]  2.854509
[3,]  9.312042
```

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 0.22195 & -0.02683 & -0.08557 \\ -0.02683 & 0.00519 & 0.00043 \\ -0.08557 & 0.00043 & 0.16670 \end{bmatrix} \begin{bmatrix} 1215.81 \\ 6776.22 \\ 662.35 \end{bmatrix} = \begin{bmatrix} 31.373371 \\ 2.854509 \\ 9.312042 \end{bmatrix}$$

(3) **Calculate $SSE$, $SSR$, $SST$, and construct the ANOVA table for this regression model.**

$$\mathbf{Y}'\mathbf{Y} = \sum_{i=1}^{n} y_i^2 = 64227.56$$

$$\mathbf{Y}'X\hat{\boldsymbol{\beta}} = (X'Y)'\hat{\boldsymbol{\beta}} = \begin{bmatrix} 1215.81 & 6776.22 & 662.35 \end{bmatrix} \begin{bmatrix} 31.373371 \\ 2.854509 \\ 9.312042 \end{bmatrix} = 63654.67$$

$$\begin{aligned} SSE &= \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'X\hat{\boldsymbol{\beta}} = 64227.56 - 63654.67 = 572.89 \\ SSR &= \mathbf{Y}'X\hat{\boldsymbol{\beta}} - n\bar{y}^2 = 63654.67 - 24 \times 50.65875^2 = 2063.255 \\ SST &= SSE + SSR = \mathbf{Y}'\mathbf{Y} - n\bar{y}^2 = 64227.56 - 24 \times 50.65875^2 = 2636.149 \end{aligned}$$

ANOVA table:

| Source | df | SS | MS | F |
|---|---|---|---|---|
| Regression | $p = 2$ | $SSR = 2063.255$ | $MSR = \frac{SSR}{p} = 1031.627$ | $\frac{MSR}{MSE} = 37.816$ |
| Error | $n - p - 1 = 21$ | $SSE = 572.89$ | $MSE = \frac{SSE}{n-p-1} = 27.28$ | |
| Total | $n - 1 = 23$ | $SST = 2636.149$ | | |

(4) **Compute the global $F$-test statistic and state your conclusion.**

The global $F$-test evaluates

$$H_0 : \beta_1 = \beta_2 = 0 \qquad \text{vs.} \qquad H_1 : \text{at least one } \beta_j \neq 0$$

$$F = \frac{MSR}{MSE} = \frac{1031.627}{27.28048} = 37.81559$$

$p$-value $= P(F_{2,21} \geq 37.81559) = 1.1 \times 10^{-7}$.

```
> 1-pf(37.81559,2,21)
[1] 1.095396e-07
```

We reject $H_0$. There is strong evidence that at least one $\beta_j$ is different from 0.

(5) **Provide the estimator of $\sigma^2$.**

$$s_e^2 = MSE = \frac{SSE}{n-p-1} = \frac{572.89}{24-3} = 27.28048$$

(6) **Compute $R^2$ and interpret it.**

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{572.89}{2636.149} = 0.7827$$

The regression model on air tightness of the home ($X_1$) and the absence/presence of UFFI ($X_2$) accounts for 78.27% of the variability in $CH_2O$ concentration.

(7) **Find the standard errors of $\hat{\beta}_1$ and $\hat{\beta}_2$.**

$$\begin{aligned}
\widehat{Var}(\hat{\boldsymbol{\beta}}) &= s_e^2(X'X)^{-1} = 27.28048 \begin{bmatrix} 0.22195 & -0.02683 & -0.08557 \\ -0.02683 & 0.00519 & 0.00043 \\ -0.08557 & 0.00043 & 0.16670 \end{bmatrix} \\
&= \begin{bmatrix} 6.0547870 & -0.73188653 & -2.33436388 \\ -0.7318865 & 0.14165546 & 0.01180462 \\ -2.3343639 & 0.01180462 & 4.54773039 \end{bmatrix}
\end{aligned}$$

Thus,

$$SE(\hat{\beta}_1) = \sqrt{0.14165546} = 0.3763714 \qquad SE(\hat{\beta}_2) = \sqrt{4.54773039} = 2.132541$$

(8) **Perform the hypothesis test $H_0 : \beta_2 = 0$ versus $H_1 : \beta_2 \neq 0$ and state your conclusion.**

The $t$-test statistic is given by:

$$t = \frac{\hat{\beta}_2}{SE(\hat{\beta}_2)} = \frac{9.312042}{2.132541} = 4.367$$

and yields a $p$-value $= 2 \times P(t_{21} \geq 4.367) = 0.00027$.

```
> 2*(1-pt(4.366641,21))
[1] 0.0002703892
```

Thus, we reject $H_0$. There is strong evidence of an association between the absence/presence of UFFI ($X_2$) and $CH_2O$ concentration after controlling for air tightness of the home ($X_1$).

(9) **Interpret the regression coefficient $\hat{\beta}_2$.**

After adjusting for air tightness of the home, the $CH_2O$ concentration in a home with UFFI is expected to be 9.31 ppb more than in a home without UFFI, on average.

(10) **Construct a 95% Wald confidence interval for $\beta_2$ and interpret it in context.**
A 95% Wald confidence interval for $\beta_2$ is given by

$$\hat{\beta}_2 \pm t_{21,0.025} SE(\hat{\beta}_2) = 9.312042 \pm 2.079614 \times 2.132541 = (4.877, 13.747)$$

```
> qt(0.025, 21)
[1] -2.079614
```

We are 95% confidence that the $CH_2O$ concentration in a home with UFFI is expected to be at least 4.88 ppb and at most 13.75 ppb higher in a home without UFFI, on average, controlling for the home's air tightness.

(11) **Fit the regression model in R and show that you get the same answers for $\hat{\boldsymbol{\beta}}$, $s_e^2$, $SE(\hat{\beta}_1)$, $SE(\hat{\beta}_2)$, $t$-test for $H_0 : \beta_2 = 0$, $R^2$ and the $F$-test statistic.**

```
fit = lm(Y...CH2O ~ X1...Air.Tightness+X2...UFFI.present)

> summary(fit)

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          31.3734     2.4607  12.750 2.36e-11 ***
X1...Air.Tightness    2.8545     0.3764   7.584 1.92e-07 ***
X2...UFFI.present     9.3120     2.1325   4.367  0.00027 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.223 on 21 degrees of freedom
Multiple R-squared:  0.7827,Adjusted R-squared:  0.762
F-statistic: 37.82 on 2 and 21 DF,  p-value: 1.095e-07
```

We note that the results match those we calculated

$$\begin{array}{lll} \hat{\beta}_0 = 31.3734 & s_e^2 = (5.223)^2 = 27.28 & t_{\beta_2} = 4.367 \\ \hat{\beta}_1 = 2.8545 & SE(\hat{\beta}_1) = 0.3764 & R^2 = 0.7827 \\ \hat{\beta}_2 = 9.3120 & SE(\hat{\beta}_2) = 2.1325 & F = 37.82 \text{ with df} = (2, 21) \end{array}$$

(12) **Obtain the regression ANOVA table in R and show that it matches your results in (3).**

```
> anova(fit)
Analysis of Variance Table

Response: Y...CH2O
                  Df  Sum Sq Mean Sq F value    Pr(>F)
X1...Air.Tightness  1 1543.08 1543.08  56.563 2.185e-07 ***
X2...UFFI.present   1  520.17  520.17  19.067 0.0002704 ***
Residuals          21  572.89   27.28
```

We note that we get the regression *df* and $SSR$ by summing the elements of the first two rows corresponding to each covariate in the R output, ant the *df* and $SST$ for the total source of variability is obtained by summing the 3 rows of the R output:

ANOVA table:

| Source | df | SS | MS | F |
|---|---|---|---|---|
| Regression | $2 = 1 + 1$ | $SSR = 1543.08 + 520.17 = 2063.255$ | $MSR = 1031.627$ | $\frac{MSR}{MSE} = 37.816$ |
| Error | 21 | $SSE = 572.89$ | $MSE = 27.28$ | |
| Total | $23 = 1 + 1 + 21$ | $SST = 1543.08 + 520.17 + 572.89 = 2636.149$ | | |