1. [10 marks] Circle the letter corresponding to the correct answer.

(a) Which statement is **FALSE**?

A: For Negative Binomial data the likelihood ratio statistic is a discrete random variable.

B: The distribution of the likelihood ratio statistic based on a random sample $Y_1, Y_2, \ldots, Y_n$ is approximately $\chi^2(1)$ for large n.

C: For Exponential data, the likelihood ratio statistic is a continuous random variable.

$\boxed{\text{D:}}$ For Binomial(n, θ) data, an approximate 95% confidence interval for θ based on the asymptotic Normal pivotal quantity only contains values inside the interval $[0, 1]$.

E: For Exponential(θ) data, an approximate 95% confidence interval for θ based on a 15% likelihood interval only contains values of θ greater than zero.

(b) Which of the following statements is **TRUE**?

A: A large observed value of the likelihood ratio test statistic indicates good agreement between the data and the null hypothesis.

$\boxed{\text{B:}}$ For Binomial(n, θ) data, the $p - value$ for testing $H_0 : \theta = \theta_0$ using the likelihood ratio test statistic can be approximated by the $G(0, 1)$ distribution for large n.

C: For Binomial(n, θ) data and $H_0 : \theta = \theta_0$, the $p - value$ obtained using the likelihood ratio test is the same $p - value$ obtained using the test statistic based on the asymptotic Normal pivotal quantity.

D: If $-2 \log R(\theta_0) = 5$ then the value $\theta = \theta_0$ is inside a 15% likelihood interval.

(c) Let $y_1, y_2, \ldots, y_{25}$ be a random sample from Poisson(θ). Suppose $[7.8, 9.6]$ is a 15% likelihood interval for the unknown parameter θ. If we use the likelihood ratio test statistic to test $H_0 : \theta = 10$, then we can conclude

A: the approximate $p - value$ is larger than 0.15.

B: the approximate $p - value$ is larger than 0.05.

$\boxed{\text{C:}}$ the approximate $p - value$ is smaller than 0.05.

D: nothing about the $p - value$ because it is not related to the likelihood interval.

(d) Suppose that a data set $y_1, y_2, \ldots, y_{36}$ is assumed to be an observed random sample from a $G(\mu, \sigma)$ distribution where μ and σ are unknown. Suppose also that the data set is stored in the variable y and that the command

```
t.test(y,0,conf.level=0.90)
```

has been run in R and the following output obtained:

```
One Sample t-test
data:  y
t = 3.0374, df = 35, p-value = 0.004488
alternative hypothesis:  true mean is not equal to 0
90 percent confidence interval:
0.4700055 1.6483278
sample estimates:
mean of x
1.059167
```

Based on this information the sample standard deviation to 3 decimal places is equal to

A: 3.037

$\boxed{\text{B:}}$ 2.092

C: 4.378

D: Not enough information to determine.

(e) In the simple linear regression model with a single covariate x, which of the following random variables has a Gaussian distribution?

A: $\frac{\tilde{\beta}-\beta}{S_e/\sqrt{S_{xx}}}$

B: $\frac{(n-2)S_e^2}{\sigma^2}$

C: $\frac{\tilde{\mu}(x)-\mu(x)}{S_e\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}}$

$\boxed{\text{D:}}$ $\tilde{\mu}(x) = \tilde{\alpha} + \tilde{\beta}x$ for a given x

(f) Suppose for data (x_i, y_i), $i = 1, 2, \ldots, n$ we assume the model $Y_i \sim G(\alpha + \beta x_i, \sigma)$, $i = 1, 2, \ldots, n$ independently. Which statement is **FALSE**?

A: $\sum\limits_{i=1}^{n} \left(y_i - \hat{\alpha} - \hat{\beta}x_i\right)^2 = S_{YY} - \hat{\beta}S_{XY}$

B: $S_{YY} = \sum\limits_{i=1}^{n} (y_i - \bar{y})^2$

C: $S_{YY} = \sum\limits_{i=1}^{n} (y_i - \bar{y})\, y_i$

$\boxed{\text{D:}}$ $\hat{\alpha} = \bar{y} + \hat{\beta}\bar{x}$

E: The least squares estimate of α and β, and the maximum likelihood estimates of α and β both minimize the function $g(\alpha, \beta) = \sum\limits_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$.

(g) Suppose for data (x_i, y_i), $i = 1, 2, \ldots, n$ we assume the model $Y_i \sim G(\alpha + \beta x_i, \sigma)$, $i = 1, 2, \ldots, n$ independently. Which statement is **FALSE**?

A: The parameter σ represents the variability in the response variate in the study population for each value of the explanatory variate x.

B: The parameter β represents the change in the mean of the response variate in the study population for a one unit increase in the explanatory variate.

$\boxed{\text{C:}}$ The parameter α represents the intercept of the least squares line.

D: The parameter $\mu(x) = \alpha + \beta x$ represents the mean response in the study population for units with explanatory variate equal to x.

(h) Which of the following statements about the simple linear regression model is **TRUE**?

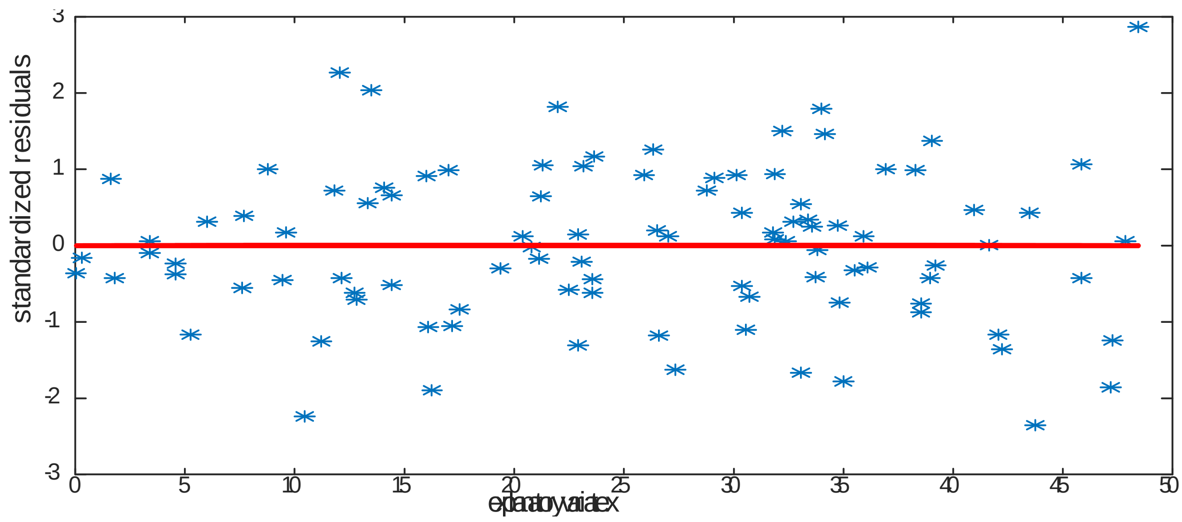
A: If $\hat{\beta} \neq 0$, then we can conclude that there is a linear relationship between the explanatory variate x and the response variate Y.

$\boxed{\text{B:}}$ The relationship between the least squares estimate of the slope $\hat{\beta}$ and the sample correlation r is $\hat{\beta} = r\,(S_{yy}/S_{xx})^{1/2}$.

C: S_e is the maximum likelihood estimator of σ.

D: The least squares estimates $\hat{\alpha}$ and $\hat{\beta}$ maximize $\sum\limits_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$.

(i) Suppose the simple linear regression model has been fit to the data (x_i, y_i), $i = 1, 2, \ldots, 100$. The standardized residual plot $(x_i, \frac{y_i - \hat{\alpha} - \hat{\beta} x_i}{s_e})$, $i = 1, 2, \ldots, 100$ for these data is:



Based on this plot we would conclude that:

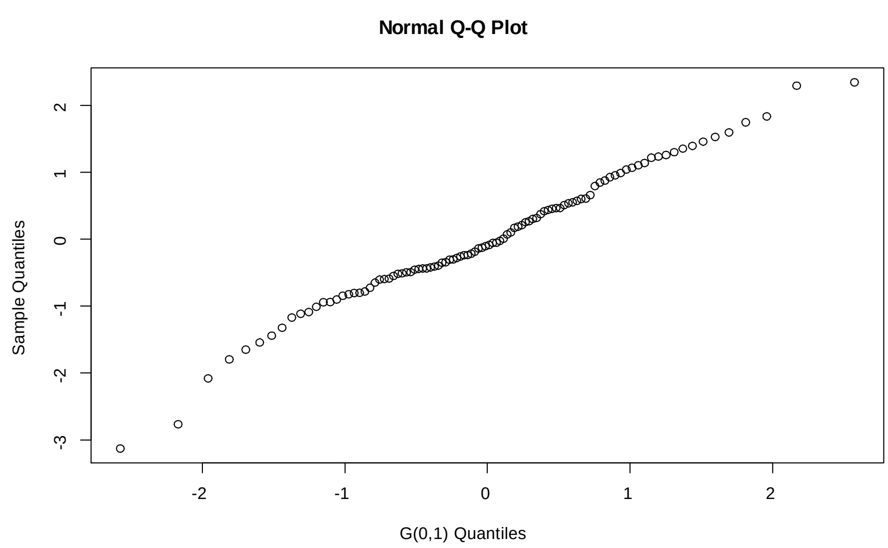
A: the simple linear regression model is not an appropriate model for these data because the assumption of constant standard deviation does not hold.

B: the simple linear regression model is not an appropriate model for these data because the assumption that the mean of the response variate is a linear function of the explanatory variate does not hold.

C: the simple linear regression model is not an appropriate model for these data because the sample size is too small.

$\boxed{\text{D}}$: the simple linear regression model is an appropriate model for these data.

(j) Suppose the simple linear regression model has been fit to the data (x_i, y_i), $i = 1, 2, \ldots, 100$. The qqplot of the standardized residuals $\hat{r}_i^* = \left(y_i - \hat{\alpha} - \hat{\beta} x_i\right) / s_e$, $i = 1, 2, \ldots, 100$ for these data is:



Based on this plot we would conclude that:

A: the simple linear regression model is not an appropriate model for these data because the assumption of constant standard deviation does not hold.

B: the simple linear regression model is not an appropriate model for these data because the assumption that the mean of the response variate is a linear function of the explanatory variate does not hold.

C: the simple linear regression model is not an appropriate model for these data because the Gaussian distribution assumption for the residuals does not hold

$\boxed{\text{D}}$: the simple linear regression model is an appropriate model for these data.

2. [11 marks] Suppose $y_1, y_2, \ldots, y_n$ is an observed random sample from the distribution with probability function

$$f(y;\theta) = (y+1)(1-\theta)^2\theta^y \quad \text{for } y = 0, 1, \ldots; \quad \theta \in (0,1)$$

where θ is an unknown parameter.

(a) [4] Find the maximum likelihood estimate $\hat{\theta}$ for θ. Show your steps clearly.

The likelihood function, after dropping the constants, is

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} (1-\theta)^2 \theta^{y_i} = (1-\theta)^{2n} \theta^{\sum_{i=1}^{n} y_i} \\ &= (1-\theta)^{2n} \theta^{n\bar{y}} \quad \text{for} \quad \theta \in (0,1). \end{aligned}$$

The log likelihood function is

$$l(\theta) = 2n \log(1-\theta) + n\bar{y} \log \theta \quad \text{for} \quad \theta \in (0,1).$$

Since

$$\begin{aligned} \frac{dl(\theta)}{d\theta} &= \frac{-2n}{1-\theta} + \frac{n\bar{y}}{\theta} = \frac{n}{\theta(1-\theta)} \left[-2\theta + \bar{y}(1-\theta)\right] \\ &= \frac{n}{\theta(1-\theta)} \left[-\theta(2+\bar{y}) + \bar{y}\right] = 0 \end{aligned}$$

if

$$\theta = \frac{\bar{y}}{2+\bar{y}}$$

therefore the maximum likelihood estimate of θ is

$$\hat{\theta} = \frac{\bar{y}}{2+\bar{y}}$$

(b) [2] If $n = 30$ and the observed sample mean is $\bar{y} = 2.5$, show that $R(0.5) = 0.4339$ where $R(\theta)$ is the relative likelihood function.

The relative likelihood function

$$\begin{aligned} R(\theta) &= \frac{L(\theta)}{L(\hat{\theta})} = \left(\frac{1-\theta}{1-\hat{\theta}}\right)^{2n} \left(\frac{\theta}{\hat{\theta}}\right)^{n\bar{y}} \\ &= \left[\frac{(1-\theta)(2+\bar{y})}{2}\right]^{2n} \left[\frac{\theta(2+\bar{y})}{\bar{y}}\right]^{n\bar{y}} \end{aligned}$$

Since $n = 30$, $\bar{y} = 2.5$,

$$R(0.5) = \left[\frac{(1-0.5)(2+2.5)}{2}\right]^{60} \left[\frac{0.5(2+2.5)}{2.5}\right]^{75} = 0.4339$$

(c) [5] Given that $n = 30$ and $R(0.5) = 0.4339$, use the likelihood ratio test statistic to test the null hypothesis $H_0 : \theta = 0.5$. Show your work. Write your final numerical answers to 3 decimal places in the space provided.

(i) [1] The observed value of the likelihood ratio test statistic is 1.67 .

$$\lambda(0.5) = -2\log R(0.5) = -2\log(0.4339) = 1.670343 = 1.67$$

(ii) [2] The approximate $p-value$, using the Normal table, is 0.197.

$$p - value = 2(1 - P(Z \le \sqrt{1.670343})) = 2(1 - 0.90147) = 0.19706,$$

where $Z \sim N(0, 1)$.

(iii) [2] State your conclusion regarding the hypothesis $H_0 : \theta = 0.5$ in a sentence.

Since $p - value > 0.1$, we conclude that there is no evidence based on the data against the null hypothesis $H_0 : \theta = 0.5$.

Note: The $p - value$ must be referred to in the conclusion.

3. [9 marks] Suppose the data set $x_1, x_2, \ldots, x_{35}$ are stored in the vector x and the data set $y_1, y_2, \ldots, y_{35}$ are stored in the vector y in R. These data are to be analyzed using the simple linear regression model

$$Y_i \sim G(\alpha + \beta x_i, \sigma) \quad i = 1, 2, \ldots, 35 \text{ independently}$$

where α, β, σ are unknown parameters and the x_i's are known constants.

The following code was run in R:

```
RegModel<-lm(y~x)
summary(RegModel)
```

The output obtained was:

```
Call:
lm(formula = y ~x)

Residuals:
Min      1Q      Median  3Q      Max
-8.1989 -2.9508 0.2016 3.3276 6.4129

Coefficients:
              Estimate  Std.  Error  t value   Pr(>|t|)
 (Intercept)   3.6246     1.3712      2.643    0.0125 *
      x        0.3743     0.1109      3.375    0.0019 **

Signif.  codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'  0.1 ' ' 1

Residual standard error:  4.202 on 33 degrees of freedom
Multiple R-squared:  0.2566,  Adjusted R-squared:  0.2341
F-statistic:  11.39 on 1 and 33 DF, p-value:  0.001903
```

Answer the following questions based on this information. Use all the decimals given in the output.

(a) [1] The least squares estimate of β is 0.3743.

(b) [1] The maximum likelihood estimate of α is 3.6246.

(c) [1] The equation of the fitted least squares line is $y = 3.6246 + 0.3743x$.

(d) [1] The value of the test statistic for testing $H_0 : \beta = 0$ is equal to 3.375.

(e) [1] The $p - value$ for testing $H_0 : \beta = 0$ is equal to 0.0019 or 0.001903.

(f) [2] State your conclusion with justification regarding the hypothesis $H_0 : \beta = 0$ in a sentence.

Since $0.001 < p - value < 0.01$, we conclude that there is strong evidence based on the data against the null hypothesis $H_0 : \beta = 0$.

Note: The $p - value$ must be referred to in the conclusion.

(g) [2] The following additional code was run:

```
xbar<-mean(x)
Sxx<-(n-1)*var(x)
se<-summary(RegModel)$sigma
cat("xbar = ", xbar,", Sxx = ", Sxx, ", se = ", se)
```

The output obtained was:

```
xbar = 10.57429 , Sxx = 1435.367 , se = 4.202233
```

Based on this information and the information from the output on the previous page determine a 95% prediction interval for a response at $x = 2$ is (show your work).

$[-4.545, 13.291]$.

We want is $P(T \leq a) = 0.975$ where $T \backsim t(33)$. From t tables the closest value is $P(T \leq 2.0423) = 0.975$ where $T \backsim t(30)$.

Predicted value for $x = 2$ is $\hat{\mu}(2) = 3.6246 + 0.3743(2) = 4.3732$.

The 95% prediction interval for a response at $x = 2$ is

$$
\begin{aligned}
& 3.6246 + 0.3743(2) \pm 2.0423(4.202233)\sqrt{1 + \frac{1}{35} + \frac{(2 - 10.57429)^2}{1435.367}} \\
= \; & 4.3732 \pm 8.918041 \\
= \; & [-4.544841, 13.29124]
\end{aligned}
$$