

Stat312: Sample Final Exam

Moo K. Chung, Yulin Zhang
mchung@stat.wisc.edu, yulin@stat.wisc.edu

May 16, 2003

1. Let $X_1, \dots, X_n$ be a random sample from Bernoulli distribution with parameter p .

(a) Find the maximum likelihood estimator of $(1 - p)^2$ (20pts).

(c) Is your estimator in (a) unbiased? If it is biased, find an unbiased estimator of $(1 - p)^2$ (20pts).

Solution. (a) First find the MLE for p . Note that the Bernoulli probability function is $P(X_i = x) = p^x(1 - p)^{1-x}$ for $x = 0, 1$ (5pts). The likelihood function is then $L(x_1, \dots, x_n; p) = p^{\sum_{i=1}^n x_i} (1 - p)^{\sum_{i=1}^n 1 - x_i}$ (5pts). By solving $\partial \log L / \partial p = 0$, we get $\hat{p} = \bar{x}$ (5pts). From the invariance principle, $\widehat{(1 - p)^2} = (1 - \bar{x})^2$ (5pts). (b) $\mathbb{E}(1 - \bar{X})^2 = 1 - 2\mathbb{E}\bar{X} + \mathbb{E}\bar{X}^2$. Note that $\mathbb{E}\bar{X}^2 = \text{Var}\bar{X} + (E\bar{X})^2 = p(1 - p)/n + p^2$. So $\mathbb{E}(1 - \bar{X})^2 = 1 - 2p + p(1 - p)/n + p^2$. Hence it is biased (5pts). We only need to find an unbiased estimator of p^2 . See Midterm I, where it is given by $\bar{X} - S^2$, where S is the sample variance. So the unbiased estimator of $(1 - p)^2$ is $1 - 2\bar{X} + \bar{X} - S^2 = 1 - \bar{X} - S^2$ (15pts).

2. Gross sales before and after a training program is given by

Salesperson :	1	2	3	4	5	6
Sales before:	90	83	105	97	110	78
Sales after :	97	80	110	93	123	84

Determine if the training program is effective at $\alpha = 0.1$. Carefully do your analysis stating all the relevant assumptions (30 pts).

Solution. This problem can be also solved using regression analysis. The points will be given in three categories. **Model specification** (10 pts). We assume that the gross sale X_i of i -th sales person before the training to follow normal, i.e. $X_i \sim N(\mu_b, \sigma_b^2)$. We also assume that the gross sale Y_i of i -th sales person after the training to follow normal, i.e. $X_i \sim N(\mu_a, \sigma_a^2)$. **Inference setup** (10 pts): the null hypothesis of interest would be $H_0 : \mu_a = \mu_b$ vs. $H_1 : \mu_a > \mu_b$. The appropriate test statistic would be the one sample t -statistic (this is not a two sample t -test since X_i and Y_i are paired and correlated). Note that the sales difference $W_i = X_i - Y_i \sim N(\mu_a - \mu_b, \sigma^2)$ for some σ (figure out the relationship between σ, σ_a and σ_b). So the t -statistic is $T = \bar{W} / (S / \sqrt{6})$, where S is the sample variance of W and it is distributed as $T \sim t_5$. **Actual computation** (10 pts). You need to compute S by yourself and actual R output either in pt or qt will be given.

3. Consider a linear model $Y = \beta_0 + \beta_1 x + \epsilon$. There is a reason to believe that $\beta_1 = 0$. Following sample data are given: $(x_1, y_1), \dots, (x_n, y_n)$. We assume ϵ to follow zero mean normal distribution.

(a) Estimate β_0 by minimizing the sum of the residuals. No point given for not showing the minimization steps (10pts).

(b) Determine if the estimator in (a) is an unbiased maximum likelihood estimator. Please derive everything (20pts).

(c) When $n = 2$, show that the estimator in (a) is the minimum variance unbiased estimator (MVUE) among the estimators of the form $c_1 Y_1 + c_2 Y_2$ (10pts).

Solution. This problem tests your total understanding of linear model and estimation. Since $\beta_1 = 0$, we are dealing with a little dumber version of linear model $Y = \beta_0 + \epsilon$. (a) The sum of squared residuals $SSE = \sum_{i=1}^n (y_i - \beta_0)^2$ (3pts). Solving $\partial SSE / \partial \beta_0 = 0$, we get $\beta_0 = \bar{y}$ (7pts). (b) The estimator is $\hat{\beta}_0 = \bar{Y}$ (3pts). Do not mix up little y_i and big Y_i . It will cost you 2 points. Then it can be shown that $\mathbb{E}\hat{\beta}_0 = \beta_0$ (see Lecture 20 derivations) (7pts). We need to check if $\hat{\beta}_0$ is MLE. Note that $Y_i \sim N(\beta_0, \sigma^2)$ for some σ . The maximum likelihood estimator of β_0 is $\bar{Y}$ (see

Lecture 4 normal distribution example). Hence $\hat{\beta}_0$ is the unbiased MLE (10pts). (c) To make $c_1 Y_1 + c_2 Y_2$ unbiased, we need $\mathbb{E}(c_1 Y_1 + c_2 Y_2) = c_1 \beta_0 + c_2 \beta_0 = \beta_0$. So $c_2 = 1 - c_1$. Now minimize the variance $\text{Var}(c_1 Y_1 + c_2 Y_2) = c_1^2 \text{Var} Y_1 + c_2^2 \text{Var} Y_2 = (c_1^2 + (1 - c_1)^2) \sigma^2$ for some σ . By differentiating it w.r.t. c_1 , we get $c_1 = 1/2$. Hence, $\bar{Y}$ is MVUE when $n = 2$.

4. Data for the average blood lead level (blood) of children age 6 months to 5 years and the amount of lead used in gasoline production (gasoline) for ten 6-months periods are available.

blood	48	59	79	80	95	95	97	102	102	107
gasoline	9.3	11.0	12.8	14.1	13.6	13.8	14.6	14.6	16.0	18.2

R output for analyzing this data is given below.

```
>blood
[1] 48 59 79 80 95 95 97 102 102 107
>gasoline
[1] 9.3 11.0 12.8 14.1 13.6 13.8 14.6 14.6 16.0 18.2
>summary(lm(blood~gasoline))
Call:
lm(formula = blood ~ gasoline)
Residuals:
    Min       1Q   Median       3Q      Max
-11.5112  -6.6140  -0.2788   7.6404  10.0596
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -14.312     16.101  -0.889  0.399984
gasoline       7.298       1.150   6.344  0.000222 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 8.508 on 8 degrees of freedom
Multiple R-Squared:  0.8342,    Adjusted R-squared:  0.8135
F-statistic: 40.25 on 1 and 8 DF,  p-value: 0.0002220
```

- (a) Is there any linear relationship between blood lead level and the amount of lead used in gasoline production? What is the statistical significance of your conclusion? (20 pts).
- (b) What is the correlation coefficient between two variables? (10pts).

Solution. (a) We test if the intercept slope is zero, i.e. $H_0 : \beta_1 = 0$ (see Lecture 22 Concept 2). The P -value is extremely small so we conclude that there is a linear relationship with 99% significance. (b) The coefficient of determination $r^2 = 0.8442$. If you take the square root, $r = \pm\sqrt{0.844}$. From the negative regression slope, we see that the correlation must be negative. So $\rho = -\sqrt{0.844}$ is the correlation.

5. Suppose that 20 out of 50 females and 10 out of 40 males are depressed in a sample. Determine whether the frequency of depression is related to sex (30 pts).

Solution. This was an in-class example we studied. Let us formulate this problem in terms of categorical data. **Model specification** (10 pts). Let p_{11} and p_{21} be the respective proportions of depression in females and males. p_{12} and p_{22} are the respective proportions of non-depression in females and males. We are interested in testing $H_0 : p_{11} = p_{21}, p_{12} = p_{22}$. 4 pts will be deducted if you do not clearly specify what these p_{ij} 's are. **Inference setup** (10 pts). Under H_0 , we compute the expected number of elements in each cell and compute the χ^2 test statistic value. **Actual computation** (10pts). Then the cutoff value $\chi_{\alpha,1}^2$ for α -level test can be computed from $N(0,1)$. Note that $\alpha = P(\chi_1^2 > \chi_{\alpha,1}^2) = P(Z^2 > \chi_{\alpha,1}^2) = 2P(Z > \chi_{\alpha,1})$. So $\chi_{\alpha,1}^2 = Z_{\alpha/2}^2$. For $\alpha = 0.05$, $\chi_{\alpha,1}^2 = 1.96^2$. For this problem, pnorm or qnorm values will be given.