

Stat324: Final Exam Solutions

Moo K. Chung
mchung@stat.wisc.edu

May 11, 2005

You may use the following R output for the answering exam problems.

```
> qnorm(c(0.005, 0.01, 0.025, 0.05, 0.1))
[1] -2.575829 -2.326348 -1.959964 -1.644854 -1.281552
> qt(0.025, 5:11)
[1] -2.57 -2.44 -2.36 -2.30 -2.26 -2.22 -2.20
> qt(0.05, 5:11)
[1] -2.01 -1.94 -1.89 -1.85 -1.83 -1.81 -1.79
> qt(0.01, 5:11)
[1] -3.36 -3.14 -2.99 -2.89 -2.82 -2.76 -2.71
> qf(0.025, 1, 5:11)
[1] 0.001084778 0.001067109 0.001054611 0.001045308 0.001038115
0.001032389 [7] 0.001027723
> qf(0.05, 1, 5:11)
[1] 0.004344768 0.004273760 0.004223537 0.004186155 0.004157256
0.004134252 [7] 0.004115506
> qf(0.01, 1, 5:11)
[1] 0.0001735012 0.0001706780 0.0001686808 0.0001671942 0.0001660447
[6] 0.0001651297 0.0001643840
```

1. Two players A and B flip a biased coin alternately and the first player to obtain a head wins. The probability of obtaining a head is $p > 0$ at each toss. Suppose A flips the coin first.
- (a) Find the probability that A wins the game. Derive your result [10pts].
- (b) Find the expected number of tosses A must make till A wins. Derive your result [10pts].
- Solution.* (a) Let X be the number of tosses A makes until he gets head.

$$P(A \text{ wins}) = \sum_{i=1}^{\infty} P(X = i) = p + (1-p)^2 p + (1-p)^4 p + \dots = \frac{p}{1 - (1-p)^2}.$$

(b)

$$\mathbb{E}X = p + 2(1-p)^2 p + 3(1-p)^4 p + \dots = \sum_{x=1}^{\infty} x(1-p)^{2(x-1)} p.$$

$$\text{let } r = (1-p)^2. \text{ Then } \mathbb{E}X = p \sum_{x=1}^{\infty} x r^{x-1} = p \frac{d}{dr} \sum_{x=1}^{\infty} r^x = p \frac{d}{dr} \left(\frac{r}{1-r} \right) = \frac{p}{(1-r)^2} = \frac{p}{(1-(1-p)^2)^2}.$$

2. (a) Suppose that X and Y are Bernoulli random variables with parameter $0 < p < 1$ such that $\text{cov}(X, Y) = 0$. Prove that X and Y are independent [10pts].
- (b) Let $X_1, \dots, X_n$ be a random sample from the Bernoulli distribution with parameter p . Find an unbiased estimator for $(1-p)^2$ [10pts].

Solution. (a) Note $\text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}X\mathbb{E}Y = P(X=1, Y=1) - P(X=1)P(Y=1) = 0$. Hence $P(X=1, Y=1) = P(X=1)P(Y=1) = p^2$. Then it follows $P(X=1, Y=0) = P(X=1) - P(X=1, Y=1) = p(1-p) = P(X=1)P(Y=0)$ and similarly for other cases. So X and Y are independent.

(b) Note that $\mathbb{E}S^2 = p(1-p)$ and $\mathbb{E}\bar{X} = p$. Hence $\mathbb{E}(1 - \bar{X} - S^2) = p^2 - 2p + 1 = (1-p)^2$.

3. 47 out of 102 doctors knew the generic name for the drug methadone. Compute a 98% confidence interval for the proportion of all doctors who **do not know** the generic name for methadone. [10 pts].

Solution. Problem setup [5pts]. Let X_i be 1 if the i -th doctor knows the name and 0 otherwise. Then $X_i \sim \text{Bernoulli}(p)$. The unbiased point estimate for p is $\bar{X}$. Since the sample size is large, we use the central limit theorem and the statistic is

$$Z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \sim N(0, 1).$$

Computation [5pts]. This requires solving a quadratic equation so we will approximate this by substituting p in the denominator with estimate $\hat{p} = 47/102$. Then a CI for p becomes $\hat{p} \pm z_{\alpha/2} \sqrt{\hat{p}(1-\hat{p})/n}$. So CI is $[0.35, 0.58]$. The proportion of doctors who do not know is $q = 1 - p$. Hence a CI for q is $[1 - 0.58, 1 - 0.35] = [0.42, 0.65]$.

4. We wish to fit observations $(x_1, y_1), \dots, (x_n, y_n)$ with a linear model $Y = cx + \epsilon$. We assume that $\epsilon \sim N(0, \sigma^2)$.

- Estimate c by the least squares estimation. Determine if it is unbiased [10pts].
- Estimate c by the maximum likelihood method and compute the variance of the estimator [15pts].
- If the sample correlation coefficient of the above paired data is denoted by r , what is the sample correlation coefficient of data $(x_1 - \bar{x}, y_1 - \bar{y}), \dots, (x_n - \bar{x}, y_n - \bar{y})$? $\bar{x}$ and $\bar{y}$ are the respective sample means of x_i 's and y_i 's. Prove your result [10pts].
- Prove the sample correlation r is a number between -1 and 1 [10pts].

Solution. (a) The sum of squared residuals $SSE = \sum_{i=1}^n (y_i - cx_i)^2$ (2pts). Letting $\frac{\partial SSE}{\partial c} = 0$, we get $\sum_{i=1}^n x_i(y_i - cx_i) = 0$. Solving this, $\hat{c} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$ (3pts). $\mathbb{E}\hat{c} = \sum_{i=1}^n x_i \mathbb{E}Y_i / \sum_{i=1}^n x_i^2 = c$ (5pts).

(b) Note that $\mathbb{E}Y_i = cx_i$ and $\text{Var}Y_i = \sigma^2$. So $Y_i \sim N(cx_i, \sigma^2)$. Then the likelihood function is given by

$$L(c, \sigma) = \frac{\text{const.}}{\sigma^n} \exp\left(-\frac{\sum_{i=1}^n (y_i - cx_i)^2}{2\sigma^2}\right).$$

The log-likelihood function is then

$$\log L(c, \sigma) = \text{const.} - n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - cx_i)^2.$$

Letting $\frac{\partial \log L}{\partial c} = 0$, we get $\sum_{i=1}^n x_i(y_i - cx_i) = 0$ so we get the same answer as in (a). $\hat{c} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$. $\text{Var}\hat{c} = \frac{\sum_{i=1}^n x_i^2 \text{Var}Y_i}{[\sum_{i=1}^n x_i^2]^2} = \frac{\sigma^2}{\sum_{i=1}^n x_i^2}$.

(c) The sample correlation is given by $r_{x,y} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$, where $S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$. Let $(z_i, w_i) = (x_i - \bar{x}, y_i - \bar{y})$. Note that $\bar{z} = \bar{w} = 0$. Also $S_{zz} = S_{xx}$ and $S_{ww} = S_{yy}$. $S_{zw} = \sum_{i=1}^n (z_i - \bar{z})(w_i - \bar{w}) = \sum_{i=1}^n z_i w_i = S_{xy}$. So the sample correlation is the same, i.e. $r_{z,w} = r_{x,y}$.

(d) We show that $0 \leq r^2 \leq 1$. First we show that

$$SSE = \min_{\beta_0, \beta_1} \sum_{j=1}^n (y_j - \beta_0 - \beta_1 x_j)^2 \leq \min_{\beta_0} \sum_{j=1}^n (y_j - \beta_0)^2 \leq \sum_{j=1}^n (y_j - \bar{y})^2 \leq SST.$$

Hence $0 \leq r^2 = 1 - \frac{SSE}{SST} \leq 1$ and the proof follows.

5. Three different types of boxes underwent compression strength test. The following data are compression strength (lb). Assume the normality of data.

Type I	65	79	72	68
Type II	78	77	68	77
Type III	60	62	74	

The following R output may be used to answer the following questions; however, you are not required to use it.

```

> type1<-c(65,79,72,68)
> type2<-c(78,77,68,77)
> type3<-c(60,62,74)
> z<-c(type1,type2,type3)
> w<-c(0,0,0,0,1,1,1,1,2,2,2)
> summary(lm(z~w))

```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	73.158	3.092	23.663	2.05e-09 ***
w	-2.474	2.564	-0.965	0.36

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

Residual standard error: 6.738 on 9 degrees of freedom Multiple
R-Squared: 0.09376, Adjusted R-squared: -0.006933 F-statistic:
0.9312 on 1 and 9 DF, p-value: 0.3598

(a) Test if the true compression strength for box types I and III are equal at $\alpha = 0.1$ Assume **equal** variance for different box types [10pts].

(b) Test if the true compression strength for all box types I, II and III are equal at $\alpha = 0.1$. Assume **equal** variance for different box types. Clearly write down the null and alternate hypotheses, the test statistic, and the statistical model you are using [20pts].

Solution. (a) Let $X_i \sim N(\mu_X, \sigma_X^2)$ be the compression strength for type I and $Y_i \sim N(\mu_Y, \sigma_Y^2)$ be the compression strength for type II. The pooled estimator of σ^2 is

$$S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2} = 44.93,$$

where $S_X^2 = 36.67$ and $S_Y^2 = 57.33$. The proper test statistic for testing

$$H_0 : \mu_X = \mu_Y \text{ vs. } H_1 : \mu_X \neq \mu_Y$$

is

$$T = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{S_p \sqrt{1/n + 1/m}} \sim t_{n+m-2}.$$

Reject H_0 if $|T| > t_{\alpha/2, n+m-2}$. Since $t = 0.165$ and $t_{0.05, 5} = 2.015$, we do not reject H_0 at $\alpha = 0.1$.

(b) **Problem Setup.** Combine x_i and y_i into a single variable z_i with $1 \leq i \leq 11$. Introduce a dummy variable $w_i = 0$ for type I, $w_i = 1$ for type II, and $w_i = 2$ for type III. Then we set up a linear model $Z_i = \beta_1 + \beta_2 w_i + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$. Then we test $H_0 : \beta_2 = 0$ vs. $H_1 : \beta_2 \neq 0$ [15pts]. **Computation.** The test statistic is $T = \frac{\hat{\beta}_2}{S_{\hat{\beta}}} \sim t_{11-2}$. Since $t = -0.965$ and $t_{0.05, 9} = 1.83$, we do not reject H_0 at $\alpha = 0.1$ [5pts].