MATH 476 & 563 Statistics

Fred J. Hickernell **Spring 2011**
Final Exam **Wednesday, May 4**

Instructions:

i. This test consists of SIX *questions. For MATH 476 students the scores for your* BEST FIVE *problems will count. For MATH 563 students the scores for the* FIRST THREE *problems plus the* BEST TWO OF THE LAST THREE *problems will count.*

ii. The time allowed for this test is 75 minutes.

iii. This test is closed book, but you may use FOUR *double-sided letter-size sheets of notes.*

iv. Calculators, even of the programmable variety, are allowed. Computers using JMP or MATLAB, are also allowed. No internet access, web browsing, email, chat, etc. is allowed.

v. Show all your work to justify your answers. Answers without adequate justification will not receive credit.

1. (20 marks)
Let $X_1, \ldots, X_n$ be independent and identically distributed (i.i.d.) random variables with mean μ and variance σ^2. For each statement below, give a brief answer with explanation.

a) Must the *sample mean*,
$$\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n),$$
be an *unbiased* estimator of μ?

Answer: Yes, because
$$E[\bar{X}] = E\left[\frac{1}{n}(X_1 + \cdots + X_n)\right] = \frac{1}{n}(E[X_1] + \cdots + E[X_n]) = \frac{n\mu}{n} = \mu.$$

b) Must the *sample variance*,
$$S^2 = \frac{1}{n-1}[(X_1 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2],$$
be a *maximum likelihood estimator* of σ^2?

Answer: No. The maximum likelihood estimator can be defined even for $n = 1$, but the sample variance is undefined for $n = 1$. For example, consider $n = 1$ and $X_1 \sim$ Bernoulli(θ)*, so the likelihood function is*
$$L(x_1|\theta) = f(x_1) = \theta^{x_1}(1-\theta)^{1-x_1}, \quad x_1 = 0, 1, \ 0 \le \theta \le 1$$
and the log likelihood function is
$$\ell(x_1|\theta) = \log\left(\theta^{x_1}(1-\theta)^{1-x_1}\right) = x_1 \log(\theta) + (1 - x_1)\log(1-\theta), \quad x_1 = 0, 1, \ 0 \le \theta \le 1.$$
This is maximized when
$$0 = \frac{\mathrm{d}\ell(x_1|\theta)}{\mathrm{d}\theta} = \frac{x_1}{\theta} - \frac{1-x_1}{1-\theta} = \frac{x_1 - \theta}{\theta(1-\theta)}, \iff \theta = x_1.$$

Since the variance is $\theta(1-\theta)$, then the maximum likelihood estimate of the variance is $x_1(1-x_1) = 0$. For $X_i \sim \text{Bernoulli}(\theta)$ and arbitrary n, the maximum likelihood estimate of θ is $\bar{x}$, and the maximum likelihood estimate of the variance is $\bar{x}(1-\bar{x})$, which is not the same as the sample variance.

c) Suppose that the sample size is $n = 100$, and one observes $\bar{x} = 47$, and $s^2 = 121$. Construct a two-sided 95% confidence interval for the mean.

Answer:

$$\bar{x} \pm \frac{1.96s}{\sqrt{n}} = 47 \pm \frac{1.96\sqrt{16}}{\sqrt{100}} = 47 \pm \frac{1.96 \times 4}{10} = 47 \pm 2.156 = [44.844, 49.156].$$

d) Before the data was observed, Chris had claimed that the true mean of the random variable was at least 50. What null hypothesis and and alternative hypothesis would be set up, if you wanted to test this claim? What would be the conclusion assuming 2.5% type I error.

Answer: The hypotheses are $H_0 : \mu \geq 50$ and $H_a : \mu < 50$. The probability of having $\bar{x} = 47$, when $\mu = 50$ and $\sigma^2 = 121$ is

$$p = \text{Prob}(\bar{X} \leq 47 | \mu = 50) = \text{Prob}\left[Z = \frac{\bar{X} - 50}{\sqrt{121/100}} \leq \frac{47 - 50}{\sqrt{121/100}} = -2.7273\right]$$
$$\approx \Phi(-2.7273) = 0.32\%,$$

where Φ is the normal cumulative distribution function, since Z is approximately $N(0,1)$. Since $p < 2.5\%$, we would reject H_0 in favor of H_a.

2. (20 marks)
Let $X_1, \ldots, X_n$ be i.i.d. Exponential(θ) random variables, i.e., with *probability density function*

$$f(x) = \frac{1}{\theta} e^{-x/\theta}, \quad 0 \leq x < \infty,$$

and *moment generating function*

$$M(t) = E[e^{tX}] = \frac{1}{1 - \theta t}, \quad 0 \leq t < \frac{1}{\theta}.$$

a) Compute $E(X^k)$ for $k = 0, 1, 2, 3, 4$.

Answer: From the moment generating function,

$$E(X^k) = M^{(k)}(0) = \frac{\mathrm{d}}{\mathrm{d}t}\left[\frac{1}{1-\theta t}\right]\Bigg|_{t=0} = \frac{k!\theta^k}{(1-\theta t)^{k+1}}\Bigg|_{t=0} = k!\theta^k,$$

so

$$E(X^0) = 1, \quad E(X^1) = \theta, \quad E(X^2) = 2\theta^2, \quad E(X^3) = 6\theta^3, \quad E(X^4) = 24\theta^4.$$

b) Suppose that $\theta = 1$. Let $Y = (X_1^2 + \cdots + X_{320}^2)/320$. Use the Central Limit Theorem to approximate the probability that $Y \geq 2.5$.

Answer: First, note that $Z_i = X_i^2$ are also i.i.d. Moreover,

$$\mu_Z = E[Z_i] = E[X_i^2] = 2,$$
$$\sigma_Z^2 = E[Z_i^2] - \{E[Z_i]\}^2 = E[X_i^4] - \{E[Z_i]\}^2 = 24 - 2^2 = 20.$$

from part a. Noting that Y is the sample mean of the Z_i, by the Central Limit Theorem it follows that

$$\frac{Y - \mu_Z}{\sigma_Z/\sqrt{n}} = \frac{Y-2}{\sqrt{20/320}} = 4(Y-2) = 4Y - 8 \approx N(0,1),$$

and so

$$\text{Prob}(Y \geq 2.5) = \text{Prob}(4Y - 8 \geq 2) = 1 - \Phi(2) = 2.28\%.$$

3. (20 marks)
Consider the i.i.d. random variables $X_1, \ldots, X_n$ with uniform distribution on $[-\theta, \theta]$, i.e., X_i has the *probability density function*

$$f(x) = \begin{cases} \frac{1}{2\theta}, & -\theta \leq x \leq \theta, \\ 0, & \text{otherwise.} \end{cases}$$

a) Let $Y = \max(|X_1|, \ldots, |X_n|)$. Show that Y is a *sufficient statistic* for estimating θ.

Answer: The likelihood function is

$$\begin{aligned} L(x_1, \ldots, x_n|\theta) &= f(x_1) \cdots f(x_n) \\ &= \begin{cases} \frac{1}{(2\theta)^n}, & |x_1| \leq \theta \ \& \ \cdots \ \& \ |x_n| \leq \theta, \\ 0, & \textit{otherwise}, \end{cases} \\ &= \begin{cases} \frac{1}{(2\theta)^n}, & y = \max(|x_1|, \ldots, |x_n|) \leq \theta, \\ 0, & \textit{otherwise}. \end{cases} \end{aligned}$$

Since the likelihood is a function of y and θ only, Y is a sufficient statistic.

b) Find, θ_{MLE}, the *maximum likelihood estimate* for θ.

Answer: The likelihood function is This likelihood function increases as θ decreases from ∞ towards 0, until θ reaches $y = \max(|x_1|, \ldots, |x_n|)$. If θ becomes any smaller than y, then the likelihood drops to zero. Thus, the maximum likelihood estimator of θ is Y.

c) Is the maximum likelihood estimator *unbiased*?

Answer: The cumulative distribution function of Y is given by

$$\begin{aligned} F_Y(y) &= \text{Prob}(Y \le y) = \text{Prob}(|X_1| \le y \ \& \ \cdots \ \& \ |X_n| \le y) \\ &= \frac{2y}{2\theta} \times \cdots \times \frac{2y}{2\theta} = \frac{y^n}{\theta^n}, \qquad 0 \le y \le \theta. \end{aligned}$$

Thus,

$$\begin{aligned} E(Y) &= \int_0^\theta y f_Y(y)\,\mathrm{d}y = \int_0^\theta y \frac{ny^{n-1}}{\theta^n}\,\mathrm{d}y = \int_0^\theta \frac{ny^n}{\theta^n}\,\mathrm{d}y \\ &= \left.\frac{ny^{n+1}}{(n+1)\theta^n}\right|_0^\theta = \frac{n\theta^{n+1}}{(n+1)\theta^n} = \frac{n\theta}{(n+1)} \ne \theta. \end{aligned}$$

Thus, Y is biased.

4. (20 marks)
The battery lives in months for 20 different car batteries each for two different battery models are recorded below:

Power Charge	45	32	28	40	46	51	29	37	36	43
Extreme Shock	56	41	65	60	59	63	70	66	65	49

Power Charge	50	32	41	38	46	44	50	48	45	39
Extreme Shock	56	64	72	67	57	56	63	61	64	58

Compute a 95% confidence interval for the difference in the mean battery lives of the two models.

Answer: Putting all the battery lives in one column and the models in another column, and then choosing `Fit Y by X` *with battery life as* `Y` *and model as* `X` *yields a confidence interval of* $[-24, -15]$.

5. (20 marks)
A survey is taken of American, European, and Pakistani citizens as to whether they feel less safe, the same, or safer after the death of Osama Bin Laden.

	Less Safe	Same	Safer
Americans	52	276	342
Eurpopeans	56	281	265
Pakistanis	152	143	41

Is there a significant difference in feelings across different nationalities?

Answer: Putting nationality in one column, safety in another column, and frequency in a third column, we then choose `Fit Y by X` *with safety as* `Y`*, nationality as* `X`*, and the frequency as* `Frequency`*. This leads to a p-value less than* 0.0001*, providing strong evidence of a difference in perception.*

6. (20 marks)
Find a curve that explains well the following data:

x	1.3	1.8	2.4	3.0	3.6	4.1	5.2	5.4	6.1	6.5	7.3	7.6	8.7	9.0
y	−3.2	−2.0	0.0	1.5	2.7	5.0	6.3	6.1	4.9	5.2	3.2	2.1	1.1	1.4

Explain from your analysis why the curve you finally choose is the appropriate one and why at least one other possibility is not.

Answer: Put the x and y data in respective columns. Choosing `Fit Y by X` *the data does not look linear. A quadratic fit looks good, with the p-values for the model (< 0.001) and the coefficients ($0.0013, 0.0002, < 0.0001$) all being small enough. A cubic model gives a cubic term with a $p = 0.03759$, which is too large to justify keeping the cubic term. Therefore, the best fitted model is*

$$y = 2.55 + 0.51x - 0.46(x - 5.14)^2 + \varepsilon.$$