

Queen's University
Faculty of Arts and Sciences
Department of Mathematics and Statistics
STAT 261 Winter 2006, Midterm Exam Solutions
Professor David Steinsaltz

- 1) (12 pts) We have six independent observations of a random process: $(X_1, \dots, X_6) = (4, 4, 2, 7, 5, 8)$.
a) Compute the mean and standard deviation of the observations.

$$\text{Mean} = \frac{1}{6}(4 + 4 + 2 + 7 + 5 + 8) = 5.$$

$$\text{SD} = \sqrt{\frac{1}{6}((4 - 5)^2 + (4 - 5)^2 + (2 - 5)^2 + (7 - 5)^2 + (5 - 5)^2 + (8 - 5)^2)} = 2.$$

- b) Suppose you know that the X_i are normally distributed. What is your best estimate for the expectation and variance of X_i ?

Estimate mean by sample mean = 5.

Estimate variance by the unbiased variance estimator $S^2 = \frac{1}{6}5SD_{obs}^2 = 4.8$.

- c) Suppose now you are told that the expectation of X_i is 6. What is your best estimate for the variance?

If we know the expectation, we can write $Var(X) = E[(X - 6)^2]$, so that we estimate

$$\hat{\sigma}_X^2 = \frac{1}{6} \sum (X_i - 6)^2 = \frac{1}{30}6 = 5.$$

- 2) (12 points) We take a simple random sample of 100 single-person households from the population of Kingston, and ask their annual income. We find that the average income is \$30,000, and the standard deviation is \$20,000. Say which of the following statements is true or false, and explain. If you need more information to decide, say what you need, and why.

- a) About 68% of the households in the sample had incomes in the range \$10,000 to \$50,000.

FALSE (or not enough information, since it is not actually impossible). This would be true if the distribution of incomes were normal, but that is generally not the case, and it can't be true here, because it would mean that about 7% of the people had negative incomes.

- b) About 68% of the single-person households in Kingston have incomes in the range \$28,000 to \$32,000.

FALSE. The SD of \$2000 has nothing to do with the income distribution in the population. It's about the distribution of the random sample mean.

c) We can be 68% sure that the real average income of all single-person households in Kingston is in the range \$28,000 to \$32,000.

TRUE. This is what we mean when we say that the SD of the sample mean of 100 incomes is $\$20,000/\sqrt{100} = \2000 , and that a 68% confidence interval for the population mean is $\$30,000 \pm 1SD$.

3) (12 points) a) What does it mean for an estimator to be consistent?

Let $\hat{\theta}$ estimate θ from n independent observations X_1, \dots, X_n . Then $\hat{\theta}$ is consistent if $\hat{\theta} \xrightarrow{n \rightarrow \infty} \theta$ (in probability).

b) Give an example of an estimator that is biased but consistent.

If we use the sample variance as an estimator for the population variance, it is biased by a factor of $(n-1)/n$. It is nonetheless consistent, since the factor converges to 1 as $n \rightarrow \infty$.

c) Give an example of an estimator that is unbiased but not consistent.

Suppose we estimate the distribution mean with the estimator $\hat{\mu} = X_1$. This is unbiased, since $E\hat{\mu} = E[X_1]$ is the true expectation, but it doesn't converge.

4) (12 points) Systolic blood pressure in a certain population is normally distributed, with mean 115 mmHg and SD 12 mmHg.

a) What is the probability that an individual selected at random has blood pressure between 109 and 118?

$$P\{109 \leq BP \leq 118\} = P\left\{\frac{109 - 115}{12} \leq \frac{BP - \mu}{\sigma} \leq \frac{118 - 115}{12}\right\} \leq \Phi(0.25) - \Phi(-0.5) = 0.29.$$

b) Suppose we select 4 individuals at random. What is the probability that the average of their blood pressures is between 109 and 118?

The average of 4 BPs is normally distributed with mean 115 and $SD = 12/\sqrt{4} = 6$. So

$$P\{109 \leq \bar{X} \leq 118\} = P\left\{\frac{109 - 115}{6} \leq \frac{\bar{X} - \mu}{\sigma} \leq \frac{118 - 115}{6}\right\} \leq \Phi(0.5) - \Phi(-1) = .53.$$

5) (8 points) Statistics Canada wants to estimate the fraction of people in different cities whose incomes are below the poverty level. They will interview a random sample of 1000 people in each city. Assuming things in Toronto are roughly the same as in Kingston:

- (1) The accuracy in Toronto (population 3 million) will be about the same as the accuracy in Kingston (population 100,000).
- (2) The accuracy in Toronto will be quite a bit higher than the accuracy in Kingston.
- (3) The accuracy in Toronto will be quite a bit lower than the accuracy in Kingston.

Choose one, and explain briefly.

1 is right. The population and sample size affect the random fluctuations in the estimate through the SD for the sample mean, which is

$$\frac{\sigma}{\sqrt{n}} \sqrt{1 - \frac{n-1}{N-1}},$$

where n is the sample size and N the population size. Unless we sample a large fraction of the population, this SD is essentially independent of the population size. In this case, the finite-population correction varies from 0.995 in Kingston to 0.9998 in Toronto: not “quite a bit”.

6) (20 points) We have n independent observations, X_1, \dots, X_n , from the distribution with density $f_\alpha(x) = (\alpha - 1)x^{-\alpha}$ on $1 \leq x < \infty$, where $\alpha > 1$ is unknown.

a) Compute a method-of-moments estimate for α .

$$\begin{aligned} \mu_1 = E[X] &= \int_1^\infty x f_\alpha(x) dx \\ &= \int_1^\infty (\alpha - 1)x^{-\alpha+1} dx \\ &= \frac{\alpha - 1}{\alpha - 2}. \end{aligned}$$

(Note that the first moment — and all moments — is infinite for $\alpha \leq 2$.) We rearrange this to $\alpha = (\mu_1 - 1)/(2\mu_1 - 1)$. This gives us a MoM estimator

$$\hat{\alpha} = \frac{\bar{X} - 1}{2\bar{X} - 1}.$$

b) Compute the maximum-likelihood estimate for α .

[3mm]

The loglikelihood function is

$$\ell(\alpha) = \sum_{i=1}^n \sum_{i=1}^n (\log(\alpha - 1) - \alpha \log X_i) = n \log(\alpha - 1) - \alpha \sum_{i=1}^n \log X_i.$$

We maximize by setting the derivative to 0:

$$0 = \ell'(\hat{\alpha}) = \frac{n}{\hat{\alpha} - 1} - \sum_{i=1}^n \log X_i,$$

so

$$\hat{\alpha} = 1 + \left(\frac{1}{n} \sum_{i=1}^n \log X_i \right)^{-1}.$$

c) Suppose you have 100 observations, and you have computed the maximum-likelihood estimate $\hat{\alpha} = 2$. Compute an approximate 95% confidence interval for the true parameter α .

Applying the asymptotic normality of MLEs, we have an approximate CI $(\hat{\alpha} - 2SD, \hat{\alpha} + 2SD)$, where the SD is the standard deviation for the sampling distribution of $\hat{\alpha}$. This may be approximated by $1/\sqrt{n\mathcal{I}(\hat{\alpha})}$. We compute

$$\mathcal{I}(\alpha) = -E \left[\frac{d^2}{d\alpha^2} f_\alpha(X) \right] = -E \left[-\frac{1}{(\alpha - 1)^2} \right] = \frac{1}{(\alpha - 1)^2}.$$

Thus $\mathcal{I}(2) = 1$, and the SD is about 0.1, so the 95% CI is about (1.8, 2.2).

d) Suppose you have 100 observations, and you have computed the maximum-likelihood estimate $\hat{\alpha} = 2$. Describe how you would use computer simulation (bootstrap approach) to estimate a 95% confidence interval for $\hat{\alpha}$.

Two approaches:

- (1) Normal confidence interval: Here we only need to estimate the SD. We simulate 100 observations from the density f_2 , and compute the MLE $\hat{\alpha}$. We repeat this 1000 times, giving us estimators $\hat{\alpha}_1, \dots, \hat{\alpha}_{1000}$. We take the SD of these simulated estimators for the SD of $\hat{\alpha}$.
- (2) The “real bootstrap”: Draw 100 samples with replacement from our 100 observations. Compute the MLE from these 100 samples. Repeat this procedure 1000 times, yielding estimators $\hat{\alpha}_1, \dots, \hat{\alpha}_{1000}$. We put these in order, so $\hat{\alpha}_1 \leq \dots \leq \hat{\alpha}_{1000}$. The 95% confidence interval is any interval from $\hat{\alpha}_i$ to $\hat{\alpha}_{949+i}$: for instance, from $\hat{\alpha}_{26}$ to $\hat{\alpha}_{975}$.

7) (8 points) If a penny is spun on its edge, the probability p of coming up heads differs appreciably from $1/2$. We wish to estimate p to within 0.01, at a 90% confidence level by spinning the penny n times, and observing how many times it comes up heads. How large must n be?

We want a 90% CI to have a width of ± 0.01 . Since a 90% CI is $\hat{p} \pm 1.64\sigma_{\hat{p}}$, we want the SD of the sample mean to be $\sigma_{\hat{p}} = .01/1.64 = 0.006$. Since $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$, we have

$$n = (\sqrt{p(1-p)}/.006)^2 = 26000p(1-p).$$

An upper bound for the required n is thus $26000/4 = 6500$.

8) (8 points) We are trying to estimate a parameter θ_0 . We observe X_1, \dots, X_n which are independent observations from the distribution \mathcal{D} , determined by the parameter $\theta = \theta_0$. On the basis of the data, we compute an exact 95% confidence interval for θ_0 .

Now, we repeat the experiment a total of 20 times, so we have 20 independently determined confidence intervals. What is the probability that the true θ_0 is in every one of the 20 intervals? If you don't have enough information to answer, explain what information you would need.

Each experiment has probability 0.95 of producing a confidence interval that contains θ_0 . The probability that all 20 CIs contain θ_0 is $(0.95)^{20} = 0.36$.

9) (8 points) The following report appeared in the Feb. 17, 2006 issue of *Science*:

A British education researcher is causing a stir with his report indicating that U.K. children are getting a lot less sharp than they were 30 years ago.

In a study submitted last month to the Economic and Social Research Council, psychologist Michael Shayer of King's College London reports that performance by children of both sexes has plummeted on a test that involves perceptions of weight and volume. Shayer compared the 1976 performance of 2350 11- and 12-year-olds in a representative sample of British schools with that of students from the years 2001-04.[...] In 2004, only 5.7% of boys could equal scores made by the top third in 1976.

The test features questions such as whether the volume of water stays the same when it is poured into different shaped vessels. Psychologist Jim Ridgway of Durham University, U.K., calls it a "fairly robust indicator of cognitive development." Shayer blames the falling scores partly on computer games. Children, especially boys, are playing more in virtual worlds instead of "outdoors, with tools and things," he says.

Durham education researcher Peter Tymms calls the findings "something to be worried about," but says they need confirmation as they are belied by rises in IQ and other test scores.

Suppose that the test scores of boys in 1976 were normally distributed, and were normalized so that the average score was 100 points and the standard deviation was 10 points. Suppose, too, that the scores of boys in 2004 remained normal, and that the standard deviation was still 10 points. On the basis of the above information, what was the average score of the boys tested in 2004? If you don't have enough information, say what additional information you would need.

Let x be the score attained by one third of boys in 1976, but only 5.7% in 2004. Then

$$1 - \Phi\left(\frac{x - 100}{10}\right) = 0.33.$$

The normal table tells us that $(x - 100)/10 = 0.43$, so that $x = 104.3$.

Now set μ be the average score in 2004. We know that

$$1 - \Phi\left(\frac{x - \mu}{10}\right) = 0.057.$$

Thus $(x - \mu)/10 = 1.58$, so that $\mu = x - 15.8 = 88.5$.