

Stat 400 : Discussion section BD3 and BD4 Handout 8

Subhadeep Paul

April 9, 2013

1 Estimation

- Method of moments conceptually the easiest, Population moments=sample moments. Population moments are something that you get by doing nice math on the probability models, e.g $\mu = E(X)$, $E(X^2)$ etc. Recall, I spoke about r th order moment as well and defined it as $E(X^r)$. So what are the sample analogues? $\bar{x} = \frac{1}{n} \sum x_i$, $\frac{1}{n} \sum x_i^2$, and so on. So we equate : $\bar{x} = E(X)$, $\frac{1}{n} \sum x_i^2 = E(X^2)$ and then solve for the unknown parameters
- Maximum likelihood estimation on the other hand tries to find the parameters for which the observed data is most *likely*, which is measured by something called the *likelihood function*. Likelihood function is nothing but the joint pdf seen as a function of the unknown parameters instead of as a function of x . So,

$$L(\theta) = f(x; \theta) = \prod f(x_i; \theta)$$

why? $f(x; \theta)$ is the joint pdf of the sample given that the value of the unknown parameter is θ . Since all the elements of a random sample are independent, the joint pdf is simply product of the pdfs. For computational purpose its easy to work with $l(\theta) = \log L(\theta)$

- OK, so now we have a function of the unknown parameter θ (which can be more than one as well, e.g in normal distribution, we have 2 unknown parameters μ and σ). And we want to find the value of θ that maximizes it. Great! we know how to do that (from algebra & calculus). The most general method is ofcourse to equate the first derivative to 0. If θ has multiple parameters then we set the partial derivative wrt each of them to 0.

Problem 1. Let X_1, X_2, \dots, X_n be a random sample of size n from distributions with the given probability density functions. In first two cases $0 < x < \infty$ and $0 < \theta < \infty$. In the last case $k \leq x < \infty$ and $1 \leq \theta < \infty$. Find Method of moments and maximum likelihood estimators

$$f(x, \theta) = \frac{1}{\theta^2} x \exp\left(-\frac{x}{\theta}\right)$$

$$f(x, \theta) = \frac{1}{2} \exp(-|x - \theta|)$$

$$f(x, \theta) = \theta k^\theta \left(\frac{1}{x}\right)^{\theta+1}$$

- If θ has multiple parameters then we set the partial derivative wrt each of them to 0.

Problem 2. Let X_1, X_2, \dots, X_n be a random sample of size n from $N(\mu, \sigma^2)$ distribution. $-\infty < \mu < \infty$ and $0 < \sigma^2 < \infty$. Find MoM and MLE estimators.

2 Goodness of estimators : Unbiased and MSE

- Recall in the last class I said some estimators are good and some are not so good. The question is how do we measure if something is good or not. One such criterion is *unbiasedness* which says that we want our estimator to give us the *correct* parameter value at an average (pretty logical demand). Now here is the reason I spent so much time last class to tell you that an estimator is a function of sample/data x . If its a function of x , say $T(x)$, then we know how to find its mean and variance. So intuitively we take say 100 such samples x and calculate 100 such estimators $T(x)$ and we want their average to be same as θ . So we want $E(T(X)) = \theta$. This is the condition for unbiasedness. Clearly variance of an estimator $T(x)$ is $var(T(x)) = E(T(x)^2) - (E(T(x)))^2$

Problem 3. Let's verify unbiasedness of the estimators obtained in problem 2.

Also note that sample mean and sample variance are unbiased estimators of *mean* and *variance* for any distribution.

Problem 4. (midterm 3, last fall) Suppose a distribution exists with the first two moments as a function of some unknown parameter θ . $E(X) = \frac{3\theta}{5}$ and $E(X^2) = \frac{2\theta^2}{5}$. Also let the MoM estimator from a sample of size n be $\hat{\theta} = \frac{5\bar{X}}{3}$. (a) Is $\hat{\theta}$ an unbiased estimator of θ ? (b) What is the variance of this estimator?

3 Confidence intervals

- *Point estimation* is not enough, we would like to control the *error* we are making. Contrary to popular perception, 95 % CI says nothing about the probability of the true parameter lying in the interval, it just comment on how good the method of obtaining the CI is i.e how much are we *confident* about our interval. True parameter value is a constant and not a random variable so we can't talk about assigning probabilities to it. Interpretation of Confidence interval is that if we repeat the experiment 100 times, 95 times we will come up with an interval that will contain the true value of the parameter.

3.1 CI for mean μ

The estimator for mean is \bar{x} . So for *two sided* $100(1-\alpha)$ % CI we need an interval around the point estimator with coverage probability of $1-\alpha$ i.e we need $P(\bar{x} - a < \mu < \bar{x} + a) = 1-\alpha$, i.e $P(|\bar{x} - \mu| < a) = 1 - \alpha$. which implies $P(\frac{\sqrt{n}|\bar{x}-\mu|}{\sigma} < \frac{\sqrt{na}}{\sigma}) = 1-\alpha$. so $\frac{\sqrt{na}}{\sigma} = z_{\alpha/2}$. so $a = \frac{\sigma}{\sqrt{n}}z_{\alpha/2}$. The $100(1-\alpha)$ % *one sided* lower limit confidence interval is $(\frac{\sigma}{\sqrt{n}}z_{\alpha}, \infty)$ and $100(1-\alpha)$ % *one sided* Upper limit confidence interval is $(\infty, \frac{\sigma}{\sqrt{n}}z_{\alpha})$. So for 2 sided CI you need look into z value corresponding to $\alpha/2$ and for one sided α .

3.2 Minimum sample size required

In estimating the population mean μ to within ϵ with $(1-\alpha)$ 100% confidence is

$$n = \left[\frac{\sigma}{\epsilon} z_{\alpha/2} \right]^2$$

Problem 5. As part of a Department of Energy (DOE) survey, American families will be randomly selected and questioned about the amount of money they spent last year on home heating oil or gas. Of particular interest to the DOE is the average amount μ spent last year on heating fuel. Suppose the DOE estimates the population standard deviation to be \$72. If the DOE wants the estimate of μ to be correct within \$7 with a 91% confidence level, how many families should be included in the sample?