

Math 218 Mathematical Statistics

Prof. D. Joyce, Clark University

6 Feb 2009

Due Monday. From page 229, exercises 1, 2ab, 5, 7.

For the next meeting or two. Read chapter 15.1 (p. 613) on maximum likelihood estimators through the end of example 15.4 on page 617.

Last meeting. Student's t -distribution and Snedecor-Fisher's F -distribution.

Today. The main job of statistics is to make inferences, specifically inferences about parameters based on data from a sample.

We assume that a sample X_1, \dots, X_n comes from a particular distribution, called the *population distribution*, and although that particular distribution is not known, it is assumed that it is one of a family of distributions parametrized by one or more parameters.

For example, if there are only two possible outcomes, then the distribution is a Bernoulli distribution parametrized by one parameter p , the probability of success.

For another example, many measurements are assumed to be normally distributed, and for a normal distribution, there are two parameters μ and σ^2 .

Point estimation. The first kind of inference that we'll look at is estimating the values of parameters. A *point estimator* $\hat{\theta}$ of a parameter θ is some function of the sample X_1, \dots, X_n . Any function of a sample is called a *sample statistic*.

For example, for the Bernoulli distribution, a typical estimator for p is the sample mean \bar{X} , that is, \hat{p} is often taken to be \bar{X} .

For another example, for the normal distribution, $\hat{\mu}$ is often taken to be \bar{X} , and $\hat{\sigma}^2$ is often taken

to be the sample variance $S^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2$, although sometimes $\hat{\sigma}^2$ is taken to be the sample variance $\frac{1}{n} \sum (X_i - \bar{X})^2$

Sometimes, there are many different choices for estimators. To estimate the population mean μ , besides (1) the sample mean \bar{X} , you might instead take (2) the midrange value $\frac{1}{2}(X_{\min} + X_{\max})$, or (3) the median, or (4) just about any other statistic that has "central tendencies." Which of these is best, and why?

Desirable criteria for point estimators.

The MSE. Well, of course, we want our point estimator $\hat{\theta}$ of the parameter θ to be close to θ . We can't expect them to be equal, of course, because of sampling error. How should we measure how far off the random variable $\hat{\theta}$ is from the unknown constant θ ? One standard measure is what is called the *mean squared error*, abbreviated MSE and defined by

$$\text{MSE}(\hat{\theta}) = E((\hat{\theta} - \theta)^2)$$

the expected square of the error. If we have two different estimators of θ , the one that has the smaller MSE is closer to the actual parameter (in some sense of "closer").

Variance and bias. The MSE of an estimator $\hat{\theta}$ can be split into two parts, the estimator's variance and a "bias" term. We're familiar with the variance of a random variable X ; it's

$$\text{Var}(X) = \sigma_X^2 = E((X - \mu_X)^2) = E(X^2) - \mu_X^2.$$

Right now, though, our random variable is $\hat{\theta}$, so its variance is

$$\text{Var}(\hat{\theta}) = E((\hat{\theta} - E(\hat{\theta}))^2) = E(\hat{\theta}^2) - (E(\hat{\theta}))^2.$$

The $\text{MSE}(\hat{\theta})$ is the sum of this variance and one other component. Let's see what that other component is.

$$\begin{aligned} \text{MSE} &= E((\hat{\theta} - \theta)^2) \\ &= E(\hat{\theta}^2 - 2\hat{\theta}\theta + \theta^2) \\ &= E(\hat{\theta}^2) - (E(\hat{\theta}))^2 + (E(\hat{\theta}))^2 - 2E(\hat{\theta})\theta + \theta^2 \\ &= \text{Var}(\hat{\theta}) + (E(\hat{\theta}))^2 - 2E(\hat{\theta})\theta + \theta^2 \\ &= \text{Var}(\hat{\theta}) + (E(\hat{\theta}) - \theta)^2 \end{aligned}$$

The expression $E(\hat{\theta}) - \theta$ is called the *bias* of the estimator $\hat{\theta}$.

$$\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta.$$

If $\text{Bias}(\hat{\theta})$ is positive that means that you expect the estimator $\hat{\theta}$ to be too large, if negative, then too small. From the computation above, we see that the MSE is the sum of the variance of $\hat{\theta}$ and the square of the bias of $\hat{\theta}$:

$$\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + (\text{Bias}(\hat{\theta}))^2.$$

So, what makes a good estimator? We've only seen a couple of measures of the goodness of an estimator, and people disagree just based on them. The authors say the best estimator is the unbiased one with the smallest variance. Others say the smallest MSE is most important since it's even closer to the parameter. Others say that there are other criteria that are more important than the ones we've just discussed. Example 6.4 discusses the sample variance S^2 as an estimator for variance σ^2 for a normal distribution. Dividing by $n - 1$ leads to an unbiased estimator, but dividing by n leads to an estimator with a much smaller MSE, at least for small n , but as n increases the difference between the MSE's for the two estimators (the one involving $n - 1$ and the one involving n) approaches 0.

Standard error and estimated standard error of an estimator. The variance of $\hat{\theta}$ is one measure of its error, and that's often reported as its square root, the standard deviation of $\hat{\theta}$, called the *standard error* of the estimator $\hat{\theta}$, abbreviated

SE. Unfortunately, the actual value of this standard error is not known because the parameters of the distribution are unknown. But they can be estimated and so can the standard error be estimated. So, what's usually reported is the *estimated standard error*.

An example. Suppose the population has two unknown parameters μ and σ^2 , as is the case for normal populations. The sample mean \bar{x} is often used to estimate μ . The standard deviation of \bar{X} , that is, the SE of the estimator \bar{X} , is σ/n . But σ is estimated by the sample standard deviation S . Thus, the estimated sample error for \bar{X} is S/\sqrt{n} . This particular estimated sample error is so commonly used, it's got its own name—the *standard error of the mean*—abbreviated SEM.

The method of moments. There are other ways of coming up with estimators. One is called the method of moments. The idea is that to estimate a moment μ_k of the population distribution, just use the corresponding moment $\hat{\mu}_k$ of the sample. They're analogous, anyway.

$$\mu_k = E(X^k) \quad \text{while} \quad \hat{\mu}_k = \frac{1}{k} \sum_{i=1}^n X_i^k.$$

The method of moments goes further than that, though. If what you want to estimate k parameters $\theta_1, \dots, \theta_k$, and those parameters aren't moments themselves, then find those parameters in terms of the moments.

Example 6.6, page 202, shows how this is done when a random sample is taken from a uniform distribution on an unknown interval $[\theta_1, \theta_2]$. Using the method of moments, it turns out that

$$\hat{\theta}_1, \hat{\theta}_2 = \hat{\mu}_1 \pm \sqrt{3(\hat{\mu}_2 - \hat{\mu}_1^2)}$$

where $\mu_1 = \bar{X} = \frac{1}{k} \sum X_i$, and $\mu_2 = \frac{1}{k} \sum X_i^2$.