

Math 217 Probability and Statistics

Prof. D. Joyce, Clark University

Wednesday, 28 Nov 2007

Due Friday. From 7.1: 3, 6; from 7.2: 2a, 3a, 5a. within ϵ of μ . Then

Last time. The law of large numbers, and how it follows from the Chebyshev inequality.

Today. The Chebyshev inequality. Sample statistics and estimators, sample variance.

The Chebyshev inequality. There are various ways to express this inequality. Let X be a random variable X with a finite mean μ and variance σ^2 . The Chebyshev inequality can be expressed as

$$P(|X - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}$$

for any $\epsilon > 0$, but it's more usually expressed as

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

for any $k > 0$. The connection, of course, is that $\epsilon = k\sigma$ or $k = \sigma/\epsilon$. In the second form, the inequality states that the probability that X takes on a value further than k standard deviations from the mean is at most $1/k^2$.

Here's a proof of the inequality in the discrete case. Suppose that

$$P(|X - \mu| \geq \epsilon) = p.$$

Then the variance σ^2 of X is

$$\begin{aligned} \sigma^2 &= V(X) \\ &= E(|X - \mu|^2) \\ &= \sum_x |x - \mu|^2 P(X = x) \end{aligned}$$

We'll break this last sum into two parts—one where x is further from μ than ϵ , the other where x is

$$\begin{aligned} \sigma^2 &= \sum_{|x-\mu| \geq \epsilon} |x - \mu|^2 P(X = x) \\ &\quad + \sum_{|x-\mu| < \epsilon} |x - \mu|^2 P(X = x) \end{aligned}$$

Now, the first sum is greater than or equal to

$$\sum_{|x-\mu| \geq \epsilon} \epsilon^2 P(X = x) = \epsilon^2 P(|X - \mu| \geq \epsilon) = \epsilon^2 p$$

while the second sum is at least 0. Therefore,

$$\sigma^2 \geq \epsilon^2 p,$$

so $p \leq \frac{\sigma^2}{\epsilon^2}$. Thus,

$$P(|X - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}.$$

Sample statistics and estimators. We've spent a little time now about how to estimate the mean μ_X of a random variable X . You take a sample X_1, X_2, \dots, X_n of some size n , then form the sample mean $\bar{X} = (X_1 + X_2 + \dots + X_n)/n$. This sample mean \bar{X} will probably be pretty close to μ_X , but how close will depend on the the particular distribution of X and how big n is. The law of large numbers said that if X had a finite variance, then you can make \bar{X} as close as you like to μ_X by taking n large enough.

A *sample statistic* is any function of a sample X_1, X_2, \dots, X_n . So far, we've only looked at one sample statistic, the sample mean \bar{X} , and we can use it to estimate the mean μ .

Sample variance. The second most important thing about a distribution, after its mean μ_X , is probably its variance σ_X^2 . The mean μ_X can be estimated by a sample mean \bar{X} , but what can we use to approximate the variance σ^2 ? You take a sample X_1, X_2, \dots, X_n of some size n , but then what do you do to that sample to estimate σ^2 ? In other words, what is a good sample statistic to estimate the variance σ^2 ?

Here's what might make sense. Take the average of the squares of the distances of the value X_i from the sample mean \bar{X} . That would be

$$\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2.$$

In fact, that's sometimes done, but for many purposes, instead of dividing by n , we divide by $n - 1$. Let's call that statistic the sample variance. The *sample variance*, denoted S^2 , of a sample X_1, X_2, \dots, X_n , is defined as

$$S^2 = \frac{1}{n - 1} \sum_{i=1}^n (X_i - \bar{X})^2.$$