Math 3215: Lecture 22

Will Perkins

April 12, 2012

1 Goodness of Fit Testing

Say we don't know what type of distribution our random data is coming from, but we have a guess. How can we test whether or not we are right?

Example: We see a sequence of independent random digits from 1 to 6. We want to know if the distribution is uniform, i.e. like a fair die. We know the complete specification of the hypothesized distribution: $p_1 = p_2 = \cdots = p_6 = \frac{1}{6}$. Our observed data consists of $x_1, \ldots x_6$ where x_1 is the number of 1's that appeared and so on.

If the null hypothesis (the distribution is $\{p_i\}$) is true, then we would expect, as n gets large, for $\frac{x_i}{n}$ to be close to p_i for each i. How can we devise a test to measure this?

- Exercise: come up with a test (i.e. a statistic) and a method of calculating p-values that will tell us whether or not the true distribution is close to the hypothesized distribution

2 Goodness of Fit with Parameter Estimation

Here is a more complicated example:

You are a physicist and you use a Geiger counter to measure the number of radioactive particles emitted by a piece of plutonium in 10 milleseconds. You repeat the experiment and you see the following counts: $6, 2, 8, 6, 4, 4, 1, 9, 7, 6, 3, 5$.

How can you test whether or not the number of particles has a Poisson distribution? The difference between this and the last scenario is that here you do not have a complete specification of the hypothesized distribution, only its type.

3 Regression

Often we have data that comes in pairs and we want to understand the relationship between the two quantities. In particular, we would like to have a formula to predict one quantity given the other.

Examples:

- Age and income
- A country's education level and GDP
- Rainfall and crop yield

- Stock market level and election results for incumbants
- Temperature and crime rate
- etc.

How can we visualize the data in such situations? A scatterplot.

- Always be careful about the distinction between correlation and causation.
- Q: If two random variables have strong correlation does one of them necessarily cause the other? why or why not?

We want to use statistics to come up with the best possible prediction of one variable given the other. For example, knowing someone's age, what is your best guess of their income? And as always, we want to quantify our prediction. How confident are we in the predicted value?

4 Linear Model

The simplest predictive model is a *linear model*: let x be the independent variable and y the dependent variable. In other words, we are trying to predict the value of y given the value of x. In a linear model, our prediction function is a linear function of x:

$$y = \alpha + \beta(x - \overline{x})$$

How do we find α and β? Parameter estimation! ... but what is the probability model?

We have to make some assumptions. The most common assumption about the distribution of y:

$$y_i = \alpha + \beta(x_i - \overline{x}) + \epsilon_i$$

where the error random variable, ϵ_i, has a $N(0, \sigma^2)$ distribution and these errors are independent.

Under that assumption, we can compute maximum likelihood estimators for α, β, and σ^2 and compute confidence intervals for them. Steps:

- What is the distribution of y_i?
- Write the likelihood function for the observations $y_1, \ldots y_n$.
- Take the partial derivative with respect to α
- Take the partial derivative with respect to β
- Take the partial derivative with respect to σ^2