STAT 802 Data Analysis September 25, 2003

Today: Some examples and theory from Ch 5

T^2 as generalization of t^2 - Why not T? Note t^2 is $F_{1,n-1}$

When we use t, we are using the distance in one direction of x from □ to assess the credibility of □. But the ordering that exists in 1 dimension is lost in more than one dimension, and there is no longer a positive and negative direction that x can be from □. All we have is distance of x from □, and this is always positive. Thus, for p>1, there is no
essential difference is using |T| or $T^{2.}$

As long as we are not interested in the sign of x-□, this deviation can be assessed by studying t^2. The usual strategies for assessing t^2, CI and HT, are generalized by extending the Euclidean distance of 1 dimension to the statistical distance in p-dimensions. (In 1 dim, statistical distance is Euclidean once standard units are used. Since we summarize the data location using $\overline{X}$, the standard deviation used to "circularize" the distance from □ is s/$\sqrt{n}$). T^2 is the squared statistical distance from $\overline{X}$ to □.

The generalization from t^2 to T^2 described on pp 211-212. The key result is that when X ~ N_p(□,□), then $T^2 \sim \frac{(n□1)p}{n□p} F_{p,n□p}$. Note that if p=1, this is $F_{1,n-1}$. Note also that while t assumes Normality of X, this assumption is very weak for large n (because of CLT).

The □ that are credible in view of $\overline{X}$ form a hyperellispsoid (assuming Normality of X). This set consists of all values of □ that are within a certain statistical distance of $\overline{X}$. The hyperellipsoid is the generalization of the contour ellipses we see for the bivariate normal, that are the locus of "equidistant" points of $\overline{X}$ from □. As usual, we turn a distance of $\overline{X}$ from □ into a distance of □ from $\overline{X}$ to construct CIs.

T^2 is a Likelihood ratio test – this is discussed on pp 216-220 and we will not say more about this. (Wilks □ on p 217 is equivalent to T^2.) The argument for T^2 using statistical distance is compelling enough!

Confidence intervals for component distributions

The confidence hyperellispoid for the multivariate □ can provide component confidence intervals for the mean of each component variable by projection on the component axis. (Fi 5.2). But if one wants simultaneous CIs for the component means, smaller intervals are justified. See the argument in the last para of p 231. A statement about the location of the multivariate □ is more stringent than the comparable statement about all the component means. Even the conservative Bonferroni CIs are smaller than the ellipsoid shadows (Fig 5.4).

The college test data is used to illustrate multivariate CIs. We use it also to consider visualization of this trivariate data.
Brushing, Spinning are demo-ed, and also smoothing X_3 as a function of X_1 and X_2.

Section 5.5 extends the discussion of the normal theory inference in the usual way using the CLT.

5.6 Control Charts: univariate and multivariate.

General intro to Quality Control:
Monitoring Processes. Management by exception. Reduction of variability to increase quality and profit. Common causes (uncontrollable) and special causes (controllable).

p =2 Ellipse Chart Use Chronological label?
Larger p. Use T^2 Chart. (Fig 5.8, p 244)
Can ignore other details …

5.7 E-M Algorithm

General idea should be known – no details necessary for this topic in this course.

With missing data:

Start with a tentative estimate of □, and use this to predict the missing data from the non-missing data. (In a normal context this uses conditional distributions). With the data completed in this way, re-estimate □. Repeat the process until convergence. This works if the missing data is missing at random.

Exercises: (not to hand in) 5.5, 5.8, 5.9, 5.18, …

Mini-lectures: Each student choose one topic. Let me know what it is. You will present when we get to that topic. 15 minutes allocated. Be sure you have a strategy for completing in 15 minutes. You should include one exercise for the class to do relating to your presentation topic. Please submit proposals to me by next Tuesday Sept 30.