Written problems to be handed in Wednesday, January 25, 2012.

1. Let $X_1, X_2, \ldots$ be independent random variables having mean μ, variance σ^2, and finite fourth central moment μ_4. Further let $\gamma_2 = \mu_4/\sigma^4 - 3$ be the coefficient of excess kurtosis and $\phi = \gamma_2/2 + 1$, which can be viewed as an alternative coefficient of excess kurtosis. Based on the first n observations, define the first sample moment $\overline{X}_n$ and second central sample moment $\hat{\sigma}_n^2$ as

$$\overline{X}_n \equiv \frac{1}{n}\sum_{i=1}^{n} X_i \qquad\qquad \hat{\sigma}_n^2 \equiv \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

and define sample variance $s_n^2 \equiv n\hat{\sigma}_n^2/(n-1)$.

   a. Show that both $\hat{\sigma}_n^2$ and s_n^2 are asymptotically unbiased estimators of σ^2.

   b. Derive expressions for the variance of $\hat{\sigma}_n^2$ and s_n^2. Express the variances both in terms of γ_2 and ϕ.

   c. Which estimator has lower variance? Which estimator has lower mean squared error?

   d. Without appealing to the weak law of large numbers, show that both $\hat{\sigma}_n^2$ and s_n^2 are consistent estimators of σ^2.

2. In the setting of problem 1 with identically distributed random variables, the Levy Central Limit Theorem provides that

$$Z_n \equiv \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \rightarrow_d \mathcal{N}(0,1).$$

In practice, however, we rarely know σ and thus typically compute

$$T_n \equiv \sqrt{n}\frac{\overline{X}_n - \mu}{s_n}.$$

   a. Show that $T_n \rightarrow_d Z$ for some random variable Z. What is the distribution of Z?

   b. Show that inference based on using a t distribution for T_n is asymptotically correct for arbitrary distributions in this setting.

   c. Suppose we want to know that T_n is arbitrarily close to Z_n. Derive an expression that would allow you to put bounds on the probability thatZ_n/T_n might be far from 1 for an arbitrary distribution.

   d. Find an expression for the smallest sample size n such that $Pr(\,|1 - Z_n/T_n| \geq \delta) \leq \epsilon$ for arbitrary $\delta > 0$ and arbitrary $\epsilon > 0$ as a function of γ_2. Also provide your expression as a function of ϕ.

   e. For what distribution does the expression in part d achieve its minimum?

   f. Suppose that the X_i's are known to be normally distributed. How does the bound you found in part c compare with the probability computed using the exact distribution?

   g. What do the above results suggest about the sample sizes for which approximate statistical inference based on the asymptotic distribution of T_n will be accurate?

3. Again, consider the setting of problem 1 with identically distributed random variables.

   a. Show that

$$W_n \equiv \sqrt{n}\left(s_n^2 - \sigma^2\right) \rightarrow_d W$$

   for some random variable W, and specify the distribution of W.

b. Suppose again that the X_i's are normally distributed. What is the exact distribution of W_n? Compare the accuracy of using an approximate distribution based on W to the exact distribution for W_n for a range of quantiles and a range of sample sizes.

c. Now consider non-normal X_i's. Compare the accuracy of using an approximate distribution based on W to using the exact distribution for W_n when the data are normally distributed. Consider several distributions covering a range of departures from the normal distribution.

4. Suppose we now have two groups of interest. Let $X_1, X_2, \ldots$ be independent random variables having mean μ, variance $\sigma^2 < \infty$ and $Y_1, Y_2, \ldots$ be independent random variables having mean ν, variance $\tau^2 < \infty$. Based on the first n observations of the X_i's and the first m observations of the Y_i's define sample means $\overline{X}_n$ and $\overline{Y}_m$ and sample variances s^2_{Xn} and s^2_{Ym}. Further define pooled sample variance

$$s^2_{Pmn} = \frac{(n-1)s^2_{Xn} + (m-1)s^2_{Ym}}{m+n-2}.$$

We consider the following variations of t tests comparing μ to ν:

$$T^{(e)}_{mn} = \frac{(\overline{X}_n - \overline{Y}_m) - (\mu - \nu)}{s_{Pmn}\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

$$T^{(u)}_{mn} = \frac{(\overline{X}_n - \overline{Y}_m) - (\mu - \nu)}{\sqrt{\frac{s^2_{Xn}}{n} + \frac{s^2_{Ym}}{m}}}$$

$$T^{(x)}_{mn} = \frac{(\overline{X}_n - \overline{Y}_m) - (\mu - \nu)}{s_{Xn}\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

a. Derive asymptotic distributions for each of the above statistics as $n \to \infty$ and $m/n \to \lambda$, a positive constant. Provide sufficient conditions that each of those limiting distributions would have variance 1. Under what conditions would each statistic out perform the others to detect a difference between μ and ν?

b. Derive asymptotic distributions for each of the above statististics as $\min(n, m) \to \infty$.