

Handed out: Wednesday October 7, 2009

Due: Wednesday October 14, 2009 by 5pm (NOTE NEW TIME)

Reading: Oct. 7 Finish discussion of two sample tests
 Oct. 7 – Oct. 14 ANOVA Basics (Chap 16-18)
 Oct. 16 – Oct. 19 Pairing/blocking (Chap 21)

1. **Two sample tests and outliers** – Nutritionists studied the metabolic expenditures (the amount of energy expended on basic metabolism) for 20 hospital emergency room patients (10 trauma patients and 10 nontrauma patients). The data they obtained (measurements are in kcal/kg/day) were:
 nontrauma: 13.9, 15.4, 15.8, 17.9, 18.3, 19.9, 20.6, 21.4, 21.7, 23.1
 trauma: 20.0, 20.6, 24.0, 25.1, 26.2, 30.0, 30.6, 30.9, 33.8, 44.1
 - (a) For the nontrauma group the mean is 18.80 and the s.d. is 3.05. For the trauma group the mean is 28.53 and the s.d. is 7.12. There is evidence that the variances differ between the two populations. Use an appropriate two-sample t -test to assess the hypothesis that the mean metabolic expenditures are the same for trauma and nontrauma patients.
 - (b) The distribution of data values in each sample appears skewed. (Of course, it is hard to assess the normality of small samples.) Given the small sample sizes and this evidence of possible nonnormality, we might worry about the validity of our t -test. Carry out a non-parametric test to compare the median metabolic expenditures of the two populations.
 - (c) Suppose that the last value in the non-trauma group was mistakenly recorded as 231.0. How would the tests you carried out in (a) and (b) be affected by such an outlier?

2. **SAS: Hospital says** – The Study on the Efficacy of Nosocomial Infection Control (SENIC) was a study of U.S. hospitals focused on studying hospital-acquired (nosocomial) infection rates and the factors associated with them. (The data set is further described in Appendix C in the text.) For this homework we focus on whether a hospital being associated with a medical school leads to higher infection rates. The data file `senic.txt` / `senic.xls` / `senic.csv` provides information about 113 hospitals. The data file contains 12 columns:
 - id = hospital identification number
 - stay = average length of stay per patient (in days)
 - age = average patient age (in years)
 - inf = infection risk (infections per 100 patients)
 - cult = routine culturing ratio
 - xray = routine x-ray ratio
 - beds = number of beds
 - m = medical school affiliation (1=yes, 2=no)
 - r = region (1=NE, 2=NC, 3=S, 4=W)
 - pat = average number of patients
 - nur = average number of full-time equivalent nurses
 - facil = percent of facilities/services offered

SAS Hints: The programs used to demonstrate SAS (rainfall data) in discussion are available on the course website. You must read in all of the variables even though we are only interested in columns 4 (inf) and 8 (m).

 - (a) Do the assumptions required for an analysis based on the pooled t -distribution appear to be satisfied? Explain and provide supporting evidence.
 - (b) Regardless of your answer in (a), carry out a pooled two-sample t -test of the hypothesis that the mean infection rate for hospitals with a medical school is the same as the mean infection rate for hospitals without one.
 - (c) Report a 95% confidence interval for the difference in mean infection rates.
 - (d) Infection rates (and other proportions) are prone to non-constant variance (proportions p that are near 0.5 have higher variance those those near 0 or 1). A common solution is to transform the proportion to the logit scale ($\log(p/1-p)$). Since our variable (inf) is expressed as a percentage it would be $\log(\text{inf}/(100-\text{inf}))$. Create the logit infection rate. Does this improve the fit of the data to the t assumptions? (Note: It does not have to!) How do the t -test results compare to those obtained with the untransformed data?
 - (e) Would this comparison of the two types of hospitals be considered an experiment or an observational study? Explain. How does this impact the interpretation of the results you found?
 - (f) Write a 2-4 sentence summary of the results of your investigation. (Address what you were studying, what you found, and any concerns)

3. Theory: Effect of non-independence

- (a) Assume that Y_1, \dots, Y_n are independent identically distributed (iid) random variables with mean μ and σ^2 . Show that $\bar{Y}_n = \frac{1}{n} \sum_i Y_i$ has mean μ and variance σ^2/n .
- (b) Suppose that Y_1, \dots, Y_n are identically distributed random variables with mean μ and variance σ^2 but they are no longer independent. Instead assume that Y_i and Y_j have correlation $\rho^{|i-j|}$ (equivalent to covariance $\sigma^2 \rho^{|i-j|}$); observations nearer each other in the sequence are more highly correlated. Find $E(\bar{Y}_n)$ and $\text{Var}(\bar{Y}_n)$ in this case. (Hint: The variance expression is ugly!)
- (c) The correlation structure in (b) is the type of correlation we expect for time series data following an autoregressive model. The parameter ρ measures the degree of correlation between consecutive values. To see the impact of this correlation on the variance of \bar{Y}_n , compute the variance if:
- $n = 10$ and $\rho = 0.9$
 - $n = 10$ and $\rho = 0.5$
 - $n = 10$ and $\rho = -0.9$

NOTE 1: You will need to use formulas from probability/statistics theory ($E(X+Y) = E(X) + E(Y)$ and $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2 \text{Cov}(X,Y)$).

NOTE 2: As $n \rightarrow \infty$ the variance of \bar{Y} behaves like $\frac{\sigma^2(1+\rho)}{n(1-\rho)}$. Negative correlation leads to more precise sample means than independent data. Positive correlation leads to less precise sample means than independent data.

4. **Theory: non-constant variance** - We described in class the Behrens-Fisher test for comparing two means when the underlying population variances are not equal. That test relies on an approximation for the distribution of the pooled variance; we investigate that approximation here. Recall that $s_1^2 \sim \sigma_1^2 \chi_{n_1-1}^2 / (n_1 - 1)$ (i.e., the sample variance has a chi-squared distribution with $n_1 - 1$ degrees of freedom multiplied by σ_1^2 and divided by the degrees of freedom). This can also be written as $(n_1 - 1)s_1^2 / \sigma_1^2 \sim \chi_{n_1-1}^2$. A similar result holds for s_2^2 . Also recall that we define the pooled variance as $s_p^2 = ((n_1 - 1)s_1^2 + (n_2 - 1)s_2^2) / (n_1 + n_2 - 2)$.

- (a) A chi-squared random variable with ν degrees of freedom has mean ν and variance 2ν . Show that $E(s_1^2) = \sigma_1^2$ and $E(s_2^2) = \sigma_2^2$ and find $E(s_p^2)$.
- (b) Referring back to the chi-square result given above (specifically the fact that $(n_1 - 1)s_1^2 / \sigma_1^2 \sim \chi_{n_1-1}^2$), it seems natural to try and find a degrees of freedom ν such that $W = \nu s_p^2 / E(s_p^2)$ behaves like a chi-square random variable with ν degrees of freedom. Find $E(W)$ and $\text{Var}(W)$.
- (c) Setting $\text{Var}(W) = 2\nu$ yields a formula for ν . Solve for ν and show that this approach yields the Satterthwaite approximation to the degrees of freedom.

5. **Paired t procedures** - A study was carried out to assess the effect of alcohol on pilots. Ten pilots performed a series of routine tasks at a simulated altitude of 25,000 feet. Each pilot performed the tasks in a completely sober condition and then three days later the pilot repeated the tasks after drinking alcohol. The response variable is the time (in seconds) of appropriate performance of the tasks. Bigger times are better. The null hypothesis is that alcohol has no effect, the alternative is that performance degrades after alcohol use. The data are provided below along with summaries.

pilot	no alcohol	alcohol	difference
1	261	185	76
2	565	375	190
3	900	310	590
4	630	240	390
5	280	215	65
6	365	420	-55
7	400	405	-5
8	735	205	530
9	430	255	175
10	900	900	0
mean	546.6	351.0	195.6
s.d.	238.8	210.9	230.5

- (a) Explain why the paired design in which the same pilot was used in both the no alcohol and alcohol conditions was a good idea.
- (b) Obtain a 95% confidence interval for the difference in mean performance under the two conditions. What do you conclude about the hypothesis of equal means based on your confidence interval?
- (c) Is there quantitative evidence of the benefits of pairing in this case? If so, describe the evidence.
- (d) If one believes the null hypothesis is true in this case then you would allow pilots to use alcohol. Discuss the type I and type II errors in this case and argue why using $p < .05$ as a cutoff might be a bad idea!