Solutions to Final STAT 421, Fall 2008

Fritz Scholz

1. (8) Two treatments A and B were randomly assigned to 8 subjects (4 subjects to each treatment) with the following responses: $0, 1, 3, 6$ and $5, 7, 9, 10$ for treatments A and B, respectively. Based on the randomization reference distribution find the p-value (corresponding to the above sets of observations) for the **two-sided** test of the hypothesis H_0 of no treatment effect when rejecting for large values of $|\bar{X}_B - \bar{X}_A|$. Would you reject at level $\alpha = .05 = 1/20$? (Hint: you don't need to compute the full randomization reference distribution! Find the most extreme value possible for $|\bar{X}_B - \bar{X}_A|$ and reason from there to the next most extreme.)

The most extreme values for $|\bar{X}_B - \bar{X}_A|$ arise when the 8 data values are split as $0, 1, 3, 5$ and $6, 7, 9, 10$ (i.e., smallest four versus largest four, or in reverse order. The next largest value of $|\bar{X}_B - \bar{X}_A|$ arises when we interchange 5 and 6 in the above splits. This configuration was given as the observed situation. Thus there are $2 + 2$ such splits with values of $|\bar{X}_B - \bar{X}_A|$ that are $\geq |(0+1+3+6)/4 - (5+7+9+10)/4|$. Since there are $\binom{8}{4} = 70$ possible splits of the 8 numbers into two samples of 4 it follows that the p-value of teh observed split is $4/70 = 2/35 > 1/20 = .05$, i.e., we should not reject at level .05.

2. (6) In the context of the previous problem, why is the randomization reference distribution of $\bar{X}_B - \bar{X}_A$ symmetric around which value?

The reference distribution is symmetric around zero, since for any split with positive (negative) $\bar{X}_B - \bar{X}_A$ there is a corresponding split (with X's under A exchanged against X's under B, possible because $m = n$) with the same negative (positive) value.

3. (6) Complete the following vectors $\mathbf{c} = (c_1, \ldots, c_t)$ such that they become proper contrast vectors in a 1-way ANOVA with $t = 6$ treatment levels.

a) $c1 = c(0, 0, 1, -.5, 0, ?)$ $\quad ? = -.5$

b) $c2 = c(-1, 0, ?, -.5, 0, .5)$ $\quad ? = 1$

c) $c3 = c(-1/6, 0, -1/6, 0, ?, 1/2)$ $\quad ? = -1/6$

4. (10) The `Fmin.test` uses **which** test statistic T to test **which** hypothesis H_0, rejecting H_0 when T is too ...? **What** are the crucial assumptions underlying the null distribution of T? **How** is the null distribution of T obtained? **Which** other test may be preferable when which one of these crucial assumptions is not satisfied?

$T = \min(s_1^2, \ldots, s_t^2)/\max(s_1^2, \ldots, s_t^2)$ can be used to test the hypothesis $H_0 : \sigma_1^2 = \ldots = \sigma_t^2$, i.e., of equal variances across t samples (still assuming equal variance within each sample). We reject H_0 when T is too small and the sampling distribution of T under H_0 can be obtained by generating random samples from the standard normal distribution and computing T each time, i.e, we simulate the null distribution. However, the assumption of normality is important for this null distribution, since it may change otherwise. When the assumption of normality is not reasonable it is advisable to use the modified Levene test which is more robust in that regard.

5. (5) If `x = 1:5` what are the responses to the following two R commands?

```
> x[x > 3]              [1] 4 5

> mean(x[-c(1,3)])      [1] 3.666667
```

6. (5) What are the responses to the following two R commands?

```
> c(8,6,4,4)/c(2,2,1,2)     [1] 4 3 4 2

> c(8,6,4,4)/c(2,2)         [1] 4 3 2 2
```

7. (6) What is the output of the following R function:

```
mystery.function=function(n){
out=NULL
for(k in 1:n){
   out=c(out,1:k)
}
out
}
```

when you call `mystery.function(3)`?

```
[1] 1 1 2 1 2 3
```

8. (9) Assume that for three samples of respective sizes $n_1 = 4$, $n_2 = 4$ and $n_3 = 8$ you are given the pooled standard deviation estimate $s = 4$. What is the standard error (SE) of the contrast estimate $\hat{C} = \frac{1}{2}\bar{X}_{1\bullet} + \frac{1}{2}\bar{X}_{2\bullet} - \bar{X}_{3\bullet}$? Also state which contrast is being estimated.

$$SE = s \times \sqrt{\frac{(1/2)^2}{4} + \frac{(1/2)^2}{4} + \frac{(-1)^2}{8}} = s \times \sqrt{\frac{1}{16} + \frac{1}{16} + \frac{1}{8}} = \frac{s}{2} = 2$$

and the contrast being estimated is: $C = \frac{1}{2}\mu_1 + \frac{1}{2}\mu_2 - \mu_3$.

9. (8) For the 2-sample Student-t based confidence interval for $\Delta = \mu_Y - \mu_X$ show how you would convert a 95% confidence interval $[A, B] = [.5, 3.5]$ into a 99% confidence interval, if I give you the

value for the ratio $t_{.995,m+n-2}/t_{.975,m+n-2} = \texttt{qt}(.995, \texttt{m}+\texttt{n}-2)/\texttt{qt}(.975, \texttt{m}+\texttt{n}-2) = 1.35$. Would this change the conclusion when testing the hypothesis $H_0 : \Delta = 0$ at the two levels $\alpha = .05$ and $\alpha = .01$? Justify your answer.

The 95% t-based interval is

$$\bar{Y} - \bar{X} \pm t_{.975,m+n-2}\, s\sqrt{\frac{1}{m}+\frac{1}{n}} = \frac{A+B}{2} \pm \frac{B-A}{2}$$

The 99% t-based interval is

$$\bar{Y} - \bar{X} \pm t_{.995,m+n-2}\, s\sqrt{\frac{1}{m}+\frac{1}{n}} = \frac{A+B}{2} \pm \frac{t_{.995,m+n-2}}{t_{.975,m+n-2}}\frac{B-A}{2} = 2 \pm 1.35 \times 1.5 = 2 \pm 2.025.$$

Since the first interval does not contain zero we would reject H_0 at level $\alpha = .05$, but at level $\alpha = .01$ we would not reject H_0 since the 99% interval contains zero.

10. (15) Give the definitions of the noncentral t, the noncentral χ^2, and the noncentral F distributions, respectively. State these definitions in terms of functions of more basic random variables (name the random variables together with the appropriate parameters and relationships to each other, and give the relevant function).

Let $Z \sim \mathcal{N}(\delta, 1)$ and $V \sim \chi^2_f$ (central chi-square with f degrees of freedom) be independent, then $Z/\sqrt{V/f}$ is said to have a noncentral t-distribution with f degrees of freedom and noncentrality parameter δ.

Let $Z_i \sim \mathcal{N}(\mu_i, 1),\ i = 1, \ldots, f$ be independent, then $Z_1^2 + \ldots + Z_f^2$ is said to have a noncentral χ^2 distribution with f degrees of freedom and noncentrality parameter $\lambda = \mu_1^2 + \ldots + \mu_f^2$.

Let $V_1 \sim$ noncentral χ^2 distribution with f_1 degrees of freedom and noncentrality parameter λ and $V_2 \sim$ central χ^2 distribution with f_2 degrees of freedom, with V_1 and V_2 independent, then $(V_1/f_1)/(V_2/f_1)$ is said to have a noncentral F distribution with f_1 (numerator) and f_2 (denominator) degrees of freedom and with noncentrality parameter λ.

11. (8) In a 1-way ANOVA, comparing means for $k = 3$ groups, each with sample size 5, at significance level $\alpha = .05$, what happens to the type II error probability when $\mu_1 = 1, \mu_2 = 2$ and $\mu_3 = 3$ and

a) the common σ for all k groups becomes very small?

b) the common σ for all k groups becomes very large?

(Think of the noncentrality parameter).

The noncentrality parameter is $\lambda = \sum_i \mu_i^2/\sigma^2$. Since the power $\beta = \beta(\lambda)$ is strictly increasing to 1 as $\lambda \to \infty$ and since the type II error probability is $1 - \beta(\lambda)$ we get $1 - \beta(\lambda) \to 0$ under a) and $1 - \beta(\lambda) \to 1 - \alpha$ under b) since then $\lambda \to 0$, i.e., we approach the null distribution.

12. (9) In a 1-way ANOVA with t treatments and n observations per treatment indicate your answer for a)-c) by True, False or Undecidable (without additional information) using T, F, U. Assume $\bar{Y}_{i\bullet} = \sum_{j=1}^n Y_{ij}/n$.

a) $\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_{i\bullet}) = 0$ (T) b) $\sum_{i=1}^{t}(Y_{ij} - \bar{Y}_{i\bullet}) = 0$ (U) c) $\sum_{j=1}^{n}\sum_{i=1}^{t}(Y_{ij} - \bar{Y}_{i\bullet}) = 0$ (T)

13. (6) Assume a 1-way ANOVA situation with t treatments.

a) If we multiply a contrast vector $\mathbf{c} = (c_1, \ldots, c_t)$ by 3, is the resulting vector again a contrast vector? (T) since $\sum_i c_i = 0 \Rightarrow \sum_i 3c_i = 0$, i.e., it is still a contrast vector

b) If instead of multiplying by 3 we add 3 to the contrast vector $\mathbf{c}$, do you get again a contrast vector? (F) since $\sum_i c_i = 0 \Rightarrow \sum_i (c_i + 3) \neq 0$, i.e., it is no longer a contrast vector

Answer True (T), False (F) or Undecidable (U) (without additional information).

14. (18) The following table gives the responses Y_{ijk} from a 2 factor experiment with 3 observations per factor level combination. Give the 2×3 table of $\bar{Y}_{ij\bullet}$, $i = 1, 2$, $j = 1, 2, 3$ (leave answers in fractional form). Also give $\bar{Y}_{1\bullet\bullet}, \bar{Y}_{2\bullet\bullet}, \bar{Y}_{\bullet 1\bullet}, \bar{Y}_{\bullet 2\bullet}, \bar{Y}_{\bullet 3\bullet}$, and $\bar{Y}_{\bullet\bullet\bullet}$.

Data Table

	level	Factor 2: 1	2	3
Factor 1	1	1, 4, 3	2, 2, 4	3, 1, 2
	2	6, 1, 2	3, 2, 1	4, 5, 2

Table of $Y_{ij\bullet}$

	level	Factor 2: 1	2	3
Factor 1	1	8/3	8/3	6/3
	2	9/3	6/3	11/3

$\bar{Y}_{1\bullet\bullet} = 22/9$, $\bar{Y}_{2\bullet\bullet} = 26/9$, $\bar{Y}_{\bullet 1\bullet} = 17/6$, $\bar{Y}_{\bullet 2\bullet} = 14/6$, $\bar{Y}_{\bullet 3\bullet} = 17/6$, $\bar{Y}_{\bullet\bullet\bullet} = 48/18 = 8/3$.

15. (8) When comparing $k = 5$ independent normal samples with equal variances, we can compute confidence intervals for **how many pairwise** differences of means $\mu_i - \mu_j$, $i < j$? Using Bonferroni's method, how would you adjust the confidence levels of these individual confidence intervals so that they simultaneously cover their respective targets $\mu_i - \mu_j$ with probability $\geq .90$.

There are $\binom{5}{2} = 10$ possible paired differences. Thus the individual confidence levels need to be adjusted to $1 - \alpha/10 = 1 - .10/10 = .99$.

16. (15) In a 2-way (2 factor) ANOVA rewrite the Sum of Squares Decomposition $SS_T = SS_A + SS_B + SS_{AB} + SS_E$ in terms of the responses Y_{ijk} using the dot-bar notation, e.g., $\bar{Y}_{i\bullet\bullet}$, etc.

$SS_{\mathrm{T}} = SS_{\mathrm{A}} + SS_{\mathrm{B}} + SS_{\mathrm{AB}} + SS_{\mathrm{E}}$ with $SS_{\mathrm{T}} = \sum_{ijk}(Y_{ijk} - \bar{Y}_{\bullet\bullet\bullet})^2$

$$SS_{\mathrm{A}} = \sum_{ijk}(\bar{Y}_{i\bullet\bullet} - \bar{Y}_{\bullet\bullet\bullet})^2 = \sum_{ijk}\hat{a}_i^2 \ , \qquad SS_{\mathrm{B}} = \sum_{ijk}(\bar{Y}_{\bullet j\bullet} - \bar{Y}_{\bullet\bullet\bullet})^2 = \sum_{ijk}\hat{b}_j^2$$

$$SS_{\mathrm{AB}} = \sum_{ijk}(\bar{Y}_{ij\bullet} - \bar{Y}_{i\bullet\bullet} - \bar{Y}_{\bullet j\bullet} + \bar{Y}_{\bullet\bullet\bullet})^2 = \sum_{ijk}\hat{c}_{ij}^2 \quad \text{and} \quad SS_{\mathrm{E}} = \sum_{ijk}(Y_{ijk} - \bar{Y}_{ij\bullet})^2 = \sum_{ijk}\hat{\epsilon}_{ijk}^2$$

$$\sum_{ijk}(Y_{ijk} - \bar{Y}_{\bullet\bullet\bullet})^2 = \sum_{ijk}(\bar{Y}_{i\bullet\bullet} - \bar{Y}_{\bullet\bullet\bullet})^2 + \sum_{ijk}(\bar{Y}_{\bullet j\bullet} - \bar{Y}_{\bullet\bullet\bullet})^2 + \sum_{ijk}(\bar{Y}_{ij\bullet} - \bar{Y}_{i\bullet\bullet} - \bar{Y}_{\bullet j\bullet} + \bar{Y}_{\bullet\bullet\bullet})^2 + \sum_{ijk}(Y_{ijk} - \bar{Y}_{ij\bullet})^2$$

17. (6) What role (distinct in character) did `SIR` and `FLUX` play in the circuit board experiment?

`SIR` was the response variable and `FLUX` was the treatment or factor.

18. (10) The testing of which hypothesis is addressed by the modified Levine test and how does it use a 1-way ANOVA analysis?

The modified Levene test tests the hypothesis of equal variance across several samples by subtracting the respective sample medians from each sample, and performing a 1-way ANOVA on these absolute differences.

19. (12) In the context of an experiment with two factors A and B with t_1 and t_2 levels, respectively, and with n replications of the response Y per factor level combination, what is the difference in performed analyses in the following three commands?

`anova(lm(Y ~ A*B))` and `anova(lm(Y ~ A+B))` and `anova(lm(Y ~ A:B))` ?

In the case of `A*B` a full model is fitted consisting of an additive model and interactions, with appropriate constraints. In the case of `A+B` an additive model is fitted and in the case of `A:B` a simple 1-way ANOVA model is fitted to all treatment level combinations for factors `A` and `B`, without any implicit additive and interaction structure.

20. (8) What type of constraints does `R` use when calculating and presenting the least squares estimates of the model coefficients in the 1-way and 2-way ANOVA? (answer in a short phrase)

`R` uses "set-to-zero" constraints.

21. (12) In the full 2-factor ANOVA model $\mu_{ij} = \mu + a_i + b_j + c_{ij}$ ($i = 1, \ldots, t_1$, $j = 1, \ldots, t_2$, and with equal number $n > 1$ of replications per factor level combination) what is the null distribution of the F-test for testing $H_0 : b_1 = \ldots = b_{t_2}$? What is the null distribution of the F-test for testing $H_0 : c_{ij} = 0 \ i = 1, \ldots, t_1, \ j = 1, \ldots, t_2$?

The respective null distributions are $F_{t_2-1, t_1 t_2(n-1)}$ and $F_{(t_1-1)(t_2-1), t_1 t_2(n-1)}$, respectively.

22. (10) In the one-way or two-way ANOVA what is the underlying assumption about data variability? If that assumption is violated what technique can correct for this? Give two distinct examples for applying the technique and explain how the technique was tuned to each situation.

It is assumed that the variance of all observations is constant. If the variance changes with the mean for each treatment (combination) then a variance stabilizing transform is often useful. One plots $\log(s_i)$ against $\log(\hat{\mu}_i)$ and hopes to see a strong linear pattern. Then one fits a straight line and its slope α, rounded to an integer or nearest multiple of $1/2$, is used in the $\lambda = 1 - \alpha$ power transformation of the data. Such transformed data will then have a more stable variance across all groups. When $\alpha = 1$ the transformation should be the logarithm. We saw two examples, the insecticide example with a reciprocal transform and the hermit crab count data with a log-transform.