Midterm STAT 421, Fall 2008

Fritz Scholz

1. (8) Give 2 reasons **why** we randomize **which** aspects in an experimental design.

We randomize the assignment of treatments to experimental units to give us a mathematical basis for testing the hypothesis of no treatment effect, i.e., to be justified to use the randomization reference distribution.
We randomize all other (nuisance) factors to avoid possible bias.

2. (6) **Name** the three basic principles in experimental design (3 words suffice).

replication, randomization and blocking.

3. (5) If we increase the type I error probability, how does that affect the type II error probability.

We increase the proportion of samples for which we reject, thus we decrease the proportion of samples for which we accept, whether falsely or correctly. Thus the type II error probability is decreased.

4. (6) If `x = c(5,4,1,1,1)` what are the responses to the following two R commands?

```
> x[x <= 3]
1 1 1
> x[-c(1,4)]
4 1 1
```

5. (6) What is the response to the following R command?

```
> mean(c(1:3,c(1,5),12))

24/6 = 4
```

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

```
mystery.function=function(x){
n=length(x)
out=NULL
for(k in 1:n){
   out=c(out,x[x <= k])
}
sum(out)
}
```

when you call `mystery.function(1:3)`?

`1+1+2+1+2+3=10`

7. (5) If X_1, X_2, X_3 are i.i.d. $\mathcal{N}(5,1)$ random variables, what is the distribution of $X_1 - X_2 + X_3$? $X_1 - X_2 + X_3 \sim \mathcal{N}(5-5+5, 1+1+1) = \mathcal{N}(5,3)$.

8. (6) Describe the two advantages of a simulated randomization reference distribution, when compared to the full randomization reference distribution.

The advantage is that computer memory storage and execution time are controlled to manageable amounts by the number of simulations run, and we get reasonable estimates of the full reference distribution if we run enough such simulations.

9. (9) Suppose, in a 2-sided 2-sample t-test of $H_0 : \mu_X = \mu_Y$ we get a 2-sided p-value of .03. Would the 95% confidence interval for $\mu_Y - \mu_X$ contain 0? Briefly explain why.

We would rejects H_0 at level $\alpha = .05$ since $.03 \leq .05$. Thus $\mu_Y - \mu_X = 0$ is not an acecptable hypothesis at level .05, i.e., 0 cannot be in the 95% confidence interval for $\mu_Y - \mu_X$.

10. (6) If $Z_1 \ldots, Z_5 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, name the distribution of $(Z_1 - \bar{Z})^2 + \ldots + (Z_5 - \bar{Z})^2$ where $\bar{Z} = (Z_1 + \ldots + Z_5)/5$, indicating the effect of 5, and give its mean and variance?

It is the chi-square distribution with $5 - 1 = 4$ degrees of freedom, with mean 4 and variance $2 \times 4 = 8$.

11. (10) Assume as given that `qt(.025,55)=-2.00` and `qt(.05,55)=-1.67`. Based on a sample of size $n = 56$ a 95% t-based confidence interval for the mean μ is $[-.2, 1.8]$, would you be justified in rejecting the hypothesis $H_0 : \mu = 0$ at significance level $\alpha = .10$? Briefly give your reasoning.

The interval center is $\bar{X} = (1.8 + (-.2))/2 = .8$ and the interval half width is $(1.8 - (-.2))/2 = 1$. Thus the 90% confidence interval for μ is $.8 \pm 1 \times 1.67/2 = .8 \pm .835$, which does contain zero. Thus we should not reject $H_0 : \mu = 0$ at significance level $\alpha = .10$.

12. (9) In a 1-sided one sample t-test explain how the sample size n affects the power function in three different ways.

n affects the critical point t_{crit}, obtained from the central t-distribution with degrees of freedom $n - 1$, it affects the degrees of fredom $n - 1$ of the noncentral t-distribution used to calculate the power, and the noncentrality parameter $\delta = \sqrt{n}(\mu - \mu_0)/\sigma$.

13. (8) What was done in the WHI study concerning hormone replacement therapy to come up with results different from previous observational studies. Why are the previous observational results suspect?

The treatments (placebo and HRT) were randomly assigned, rather than by each woman's choosing, possibly confounding that choice with some form of health consciousness. Blocking was used to make the experiment more sensitive.

14. (10) Explain the utility of a single boxplot and that of several boxplots side by side.

A single boxplot conveys basis information (median, spread, skewness, outliers) about a possibly very large sample in a succinct way. Side by side boxplots are useful in comparing many samples simultaneously, are they on the same level, do they have comparable spread, etc. This is more useful than pairwise QQ-plots, which compare just two samples each.

15. (10) If $X_1, \ldots, X_9$ are distributed like $U(-.5, .5)$ (uniform over (-.5,.5)) and X_{10} has a uniform distribution over the interval $(-20, 20)$, describe the **approximate shape** of the distribution of $X_1 + \ldots + X_{10}$ assuming that all random variables are independent. Describe the **approximate distribution shape** of $X_1 + \ldots + X_9$. In each case give the mean and variance of each sum and evaluate the important **variance ratio criterion** (you may assume as given that the $U(-.5, .5)$ distribution has variance 1/12).

The distribution shape of $X_1 + \ldots + X_{10}$ will look approximately uniform over $(-20, 20)$ with rounded shoulders, and centered at zero and with variance $9/12 + 40^2/12 = 1609/12$. The distribution shape of $X_1 + \ldots + X_9$ will look approximately normal with mean zero and variance 9/12. The variance ratio criteria for these two scenarios are respectively

$$\max_i \left\{ \frac{\sigma_i^2}{\sum_j \sigma_j^2} \right\} = \frac{40^2/12}{1609/12} = \frac{1600}{1609} \approx 1$$

and

$$\max_i \left\{ \frac{\sigma_i^2}{\sum_j \sigma_j^2} \right\} = \frac{1/12}{9/12} = \frac{1}{9} \ll 1$$