**A SAMPLE COMP QUESTION FROM DR. STROUP**

**Preface to students taking the comp**: The purposes of the comprehensive exam are 1) to test your overall comprehension of essential principles your should have mastered entering your 2nd year of the Stat MS program and 2) to identify any major weaknesses while you still have a full academic year to address them. The comprehensive exam is ***not*** a repetition of final exam questions from each of the core courses. Nor should you expect questions that neatly fit material from only one of your core courses. The comp will include questions that call for you to integrate concepts from the core courses and you will see problems that go considerably beyond the nice, tidy classroom examples you saw last year – even though they do not require anything beyond the *principles* you learned in your core courses.

In other words, the comp requires that you take an *above and beyond* approach. Memorize and regurgitate won’t suffice.

The following is an example of a perfectly do-able problem that I would consider to be fair game for the open-book portion of the final. Consider it a summer project. In addition to core principles, this is the kind of preparing you need to spend at least some of your time doing.

**Basic description of the problem:**

A researcher wants to test 4 genotypes (labeled A\_1, A\_2, A\_3 and A\_4 on Figure 1 below). Refer to these genotypes as Factor A. There are 2 diseases under study that afflict the plant species these 4 genotypes represent. Genotype A\_1 is the most commonly used and it is susceptible to both diseases. Genotypes A\_2, A\_3 and A\_4 are each claimed to be resistant to one or both of these diseases. Factor B represents a treatment that protects against one of the diseases (call it Disease B) and Factor C represents a treatment that protects against the other disease (call it Disease C). Level 1 of factors B and C imply no protection of that type applied. Level 2 denotes protection applied.

The response variable, Y, is a measure of the plant’s productiveness. The greater the response, Y, the better. Higher response is possible only if the plant is resistant to the disease or if the protection for the disease to which the plant is susceptible has been applied.

Figure 1 shows the study design used to evaluate the genotypes and disease-protection treatments. A growing area (it could be a field or a greenhouse) is divided into a 4 x 4 grid. Each row is subdivided into 2 strips; the levels of B are randomized so that one strip within a row gets B\_1 and the other gets B\_2. That is applied *all the way across the row*. Similarly, columns are subdivided into strip and levels of C are applied *all the way down the column*. The genotypes (levels of A) are assigned so that each genotype appears 4 times, exactly once per row and once per column.

1. Develop a “skeleton ANOVA” – sources of variation and degrees of freedom – giving the basic architecture of an analysis of these data.
2. Write a statistical model for these data. Clearly identify each element of the model (e.g. don’t make the reader guess what is supposed to denote – explain it so even your younger sister can understand). State *all* relevant probability assumptions (and any other assumptions deemed essential).
3. Describe how the parameters of your model in (2) are estimated. You do not need to go into extreme detail, but your write-up should be sufficient to convince the graders that you understand the general approach(es) involved.
4. The main goals of this experiment are
	1. to find out whether genotypes A\_2, A\_3 and A\_4 are resistant to either disease B or C – or both, and
	2. to find out how their productiveness compares to A\_1
	3. to make policy recommendations. The disease protection treatments (B\_2 and C\_2) have risks and potentially undesirable side-effects. If switching to any of the alternative genotype could allow foregoing theses treatment without lost productivity, policy makers want to know

For each of above objectives, write the estimable function – or set of estimable functions – that would specifically address the objective. Describe explicitly how you will use each estimable function in addressing the objective. Your write-up should be succinct. If you find yourself writing more that a paragraph or two per objective, you are writing too much (and you should use part of your study time to teach yourself to write clearly and to the point without going on and on). I tend to deduct points for each superfluous estimable function, so this is a problem that requires clear and focused thinking. Sloppiness is the most common way students get into trouble on these kinds of problems – it’s not an option here.

1. By the way, what is an estimable function? How do you know the functions you gave in (4) are estimable? What’s the definition and what criteria need to be satisfied?
2. For the model you gave in (2) is estimability a concern? Explain. Why does estimability matter?
3. Implement the analysis, consistent with your model in (2) and your plan as given by the estimable functions you gave in (4). Based on your analysis, write a report summarizing your findings.

You can use SAS, R, HLM, Stata, iTunes ...whatever you think you need to do the appropriate computing. I *do* want it done *via computer* because that’s what your eventual employer will *demand*. You need to make sure that the analysis you compute is consistent with the model you gave in (2) and your plan in (4). I tend to be merciless about three things: 1) if the analysis you do does not match the model you describe, big loss of credit; 2) reporting superfluous & irrelevant results – not everything on a SAS or R output is worth reporting – when someone reads your report, they want the stuff that’s relevant, not a complete memory dump of everything in the output. Remember that all that stuff is there because it’s useful in *some* applications. BUT there is *no* application for which *all* of the output is relevant. Part of the test is to see if you know the difference. Finally, 3) annotated computer output is not a report. You need to write the results up so that the answers to the questions posed in (4) are easy to discern, relevant evidence is provided, and the reader (who is probably your boss) doesn’t have to wade through a verbal swamp. You need to make sure important statistical concerns have been addressed. This includes “why did you use this particular software?” But the report should not lapse into unintelligible “stat-speak.”

1. By the way – why do you think the study was designed and conducted to way it was?

**Figure 1**. *Picture of Design*



data practice\_comp;

 input col row a b c y;

datalines;

1 1 1 1 1 38.7

1 1 1 1 2 41.8

1 1 1 2 1 43.5

1 1 1 2 2 58.7

1 2 2 1 1 33.7

1 2 2 1 2 41.2

1 2 2 2 1 45.6

1 2 2 2 2 46.4

1 3 3 1 1 45.6

1 3 3 1 2 66.5

1 3 3 2 1 31.6

1 3 3 2 2 53.9

1 4 4 1 1 48.3

1 4 4 1 2 49.3

1 4 4 2 1 43.8

1 4 4 2 2 45.4

2 1 2 1 1 30.2

2 1 2 1 2 31.9

2 1 2 2 1 45.5

2 1 2 2 2 43.3

2 2 4 1 1 50.9

2 2 4 1 2 47.6

2 2 4 2 1 42.4

2 2 4 2 2 36.7

2 3 1 1 1 33.4

2 3 1 1 2 30.2

2 3 1 2 1 31.1

2 3 1 2 2 44.6

2 4 3 1 1 27.4

2 4 3 1 2 30.1

2 4 3 2 1 28.6

2 4 3 2 2 35.2

3 1 3 1 1 40.3

3 1 3 1 2 53.5

3 1 3 2 1 42.4

3 1 3 2 2 52.8

3 2 1 1 1 34.3

3 2 1 1 2 35.6

3 2 1 2 1 34.1

3 2 1 2 2 48

3 3 4 1 1 39.9

3 3 4 1 2 52.9

3 3 4 2 1 33.8

3 3 4 2 2 41.4

3 4 2 1 1 24.4

3 4 2 1 2 30.1

3 4 2 2 1 36.2

3 4 2 2 2 42.6

4 1 4 1 1 55.4

4 1 4 1 2 45.4

4 1 4 2 1 55.7

4 1 4 2 2 41

4 2 3 1 1 46.3

4 2 3 1 2 43.5

4 2 3 2 1 38.1

4 2 3 2 2 45.9

4 3 2 1 1 41.7

4 3 2 1 2 37.2

4 3 2 2 1 56.2

4 3 2 2 2 48.9

4 4 1 1 1 34.1

4 4 1 1 2 28.4

4 4 1 2 1 36.4

4 4 1 2 2 41.5

;