Suggestions in Reviewing Stat421

Basic R programming constructs.

Observational Studies and Controlled Experiments

The Why's of Randomization, Replication, Blocking

Example cases that were presented, what were their points?

Hypergeometric Distribution, its relevance w.r.t. randomization in contingency tables. What can give roughly same answers w.r.t. P-values. Why is randomization required for either approach.

Steps in designing an experiment.

What did Einstein say w.r.t. Repeating? (not verbatim)

Role of SIR and FLUX in the circuit board experiment.

Describe benefits of rejecting the hypothesis of no difference. Describe benefits of not rejecting that hypothesis.

What was randomized in the FLUX experiment, and why?

Why were nuisance factors randomized individually?

Definition of a box plot. Utility of box plots. What can a QQ-plot tell you? What can you do to help you in judging patterns in QQ-plots.

QQ-plots, what are they, utility, limitations?

What are the most extreme values of $\bar{Y} - \bar{X}$ when computing it for a full randomization reference distribution. What might be the next most extreme value (think about it)?

The role of `combn`, how many evaluations does it perform?

What does a p-value derived from the randomization reference distribution represent.

Why is the randomization reference distribution of $\bar{Y} - \bar{X}$ symmetric around which value when $m = n$?

When D is the vector obtained from `combn`, representing the randomization reference distribution of $\bar{Y} - \bar{X}$, and if D_o is the actually observed value of $\bar{Y} - \bar{X}$, what R command would give you the two-sided p-value corresponding to D_o (I say command not program).

How is a normal QQ-plot constructed and what can it tell us.

Why is the randomization reference distribution of $\bar{Y} - \bar{X}$ usually normal with what mean and which variance, and why is that not true in certain situations (which?).

What is the impact on the p-value when the scatter of $(X_1, \ldots, X_m, Y_1, \ldots, Y_n) = (Z_1, \ldots, Z_N)$ is on the wide (narrow) side compared to the observed difference $\bar{Y} - \bar{X}$.

If a calculation in R produces a number like $-3.330669\text{e}-16$, what is most likely the real answer and why?

Simulated estimate for the full randomization reference distribution, how is it done, what is the benefit?

Why is the full randomization reference distribution for $\bar{Y} - \bar{X}$ equivalent to that of the two-sample t-statistic $t(\underline{X}, \underline{Y})$ as far as p-value calculation is concerned?

What well-known distribution usually approximates the randomization reference distribution for $t(\underline{X}, \underline{Y})$ when $m = 4$ and $n = 7$. What could throw a monkey wrench into that.

Critical values, significance levels, p-values, type I and II errors.

Advantages of significance levels, advantages of p-values?

Distinction between sampled and sampling distribution.

Power function and type I and II error probabilities.

Means, variances, quantiles, covariances, independence, random samples.

Rules for means (expectations) and variances of sums/averages of (independent?) random variables.

Normal density, cdf, CLT, conditions for the CLT. Why does linearization and the CLT explain an approximate normal distribution for many measured random phenomena.

Definitions of distributions that arise as sampling distribution of statistics computed from normal random samples. Normal distribution, χ^2-distribution (mean and variance), t-distribution, F-distribution, noncentral versions of the latter three.

Sampling distributions (and independence) of $\bar{X}$ and $\sum(X_i - \bar{X})^2$ for normal random samples.

One and two sample t-test statistics, their null distribution and their distribution under alternative hypotheses.

Sample size impact on power of the t-tests, how, ultimate effect.

Why is expressing $\mu - \mu_0$ in relation to σ more relevant and thus fortunate.

One- and two-sided t-tests.

Confidence intervals and bounds, relation to tests of hypothesis. What does the confidence level express operationally about confidence intervals.

Assumptions behind the two-sample t-test, pooled variance, what to do when the equal variance assumption is suspect. Rationale of Satterthwaite approximation. Effect of $m = n$.

Confidence intervals for σ_X^2/σ_Y^2.

Checking normality via QQ-plots and EDF tests. Know discrepancy metrics but not computational formulas. Sensitivities of the three tests.

Why is the fitted normal distribution closer to the EDF than the sampled normal distribution?

The acronym ANOVA and how it relates to testing for differences in means.

Why do we prefer a single overall F-test to all possible pairwise mean comparison tests.

What is the overall type-I error probability when doing several significance tests at once. How can it be bounded using Boole's inequality? What is the downside?

In the one-way ANOVA compare treatment means model with treatment effects model. In this context what is the reduced model.

What does the least squares method do in the previous context, what are the least squares estimates, their means and variances (if you know and understand the rules of means and variances of sums/averages you can easily derive them on the spot, i.e., no further memorization)?

SS_{E}, SS_{Treat}, SS_{T} and relationship between them.

The dot notation as in $\bar{Y}_{i\bullet}$, etc.

Unbiasedness of s^2, for a any random sample from a population (not just normal) with finite variance.

F-statistic in terms of SS_{E} and SS_{Treat}, its distribution under hypothesis and alternatives. The effect of sample sizes on the power and on the noncentrality parameter in particular.

Optimal allocation of sample sizes to t different treatments subject to $n_1 + \ldots + n_t = N$ for which two rationales.

The null F-distribution as approximation to the randomization reference distribution of the F-statistic.

Geometric view of data vector in ANOVA, sum of squares decomposition, degrees of freedom, orthogonality, Pythagoras.

Confidence intervals for means using individual sample standard deviations and the pooled version, advantage of the latter, lurking danger.

Standard error, definition, general purpose use, two examples given.

Simultaneous confidence intervals for means, simultaneous and individual coverage probabilities.

Boole's and Bonferroni's Inequalities, utility in simultaneous inference.

Aspects of the mean vector that captures differences in means.

Contrasts (definition), contrast estimates, their means and variances, confidence intervals.

Four methods for dealing with simultaneous confidence intervals for all pairwise mean difference contrasts C_{ij}.

Kramer's modification of Tukey's method.

Orthogonal contrasts (definition), simple example for $t = 3$.

Model diagnostics. Checking normality via QQ-plots. Checking homoscedasticity via Levene's test and `Fmin.test`, caution about the latter about normality.

Plots of residuals against fitted values, to see possible patterns. Variance stabilizing transform. The log-example, motivated by multiplicative error/variation model. The log leads to sums which may lead to CLT effect. How did we deal with the log(0) issue?

Variance stabilizing in more general setting for what type of mean/sigma relations? Box-Cox transformations. How do we find the appropriate transformation exponent?

The remaining issues concern the two-factor experiment.

Other than randomized treatment assignment what was initially done with the observed responses in the insecticide experiment to take care of which issue?

Is it possible that neither factor alone (i.e., we ignore the presence of the other factor) is significant, but the factor levels in combination are significant? What could obscure significance in the single factor tests?

What is the additive effects model (full model) for the means and what conditions are imposed to make the various parameters identifiable?

In the additive effects model (full model) with the sum-to-zero constraints, how can the parameters be expressed in terms of the treatment combination means μ_{ij}?

What constraints does R use in computing the least squares estimates of the model parameters?

Explain the degrees of freedom in an additive model (full model) ANOVA table in relation to the constraints.

Identities involving SS_{T}, SS_{A}, SS_{B}, SS_{AB} and SS_{E}, in the additive or full models.

Least squares estimates of the additive model parameters, their means.

Understand the various R ANOVA commands, what do they give you.

Distributions underlying the various F-tests.

Connection between interaction parameters and additive model.

What graphical view/pattern of the $\bar{Y}_{ij\bullet}$ might indicate an additive model.

What does blocking do for us? How can you explain it looking at the noncentrality parameters (slide 97)?

Cite examples of blocking opportunities. If you ignore such opportunities, what might happen?

What is randomized in a blocked experiment?

Go over the HW solutions (posted).