QBS 120 - Problem Set 8

Rob Frost

1. Pancreatic Cancer Example: Consider two pancreatic cancer gene expression experiments (A and B). Both experiments measure transcript abundance for a set of human protein coding genes in bulk tissue samples taken from pancreatic cancer patients with either non-metastatic (M0) or metastatic disease (M1). The goal of these experiments is to identify a set of genes whose expression differs between M0 and M1 tumors and may therefore represent potential therapeutic targets and/or prognostic biomarkers.

   - **Experiment A:** Measures expression of 1000 randomly selected protein coding genes. Assume that 10 of these genes are differentially expressed (DE) between M0 and M1 pancreatic tumors.
   - **Experiment B:** Measures expression of 100 genes with prior evidence of DE between M0 and M1 solid human cancers. Assume that 20 of these genes are DE between the M0 and M1 pancreatic tumors.

   For both experiments, assume that:

   - A total of n bulk tissue samples from separate individuals are analyzed ($n/2$ from M0 tumors and $n/2$ from M1 tumors).
   - The non-DE genes have expression values that follow a $\mathcal{N}(\mu_{M0}, 1)$ distribution in both M0 and M1 samples. (*Note: a normal distribution is being assumed to simplify the computations in the questions below; transcript abundance computed using methods like RNA-seq is count data and is typically modeled by a Poisson or negative binomial distribution.*)
   - The DE genes have expression values that follow a $\mathcal{N}(\mu_{M0}, 1)$ distribution among M0 samples and a $\mathcal{N}(\mu_{M1}, 1)$ distribution among M1 samples.
   - Researchers are interested in analyzing the expression values for each gene to test the following null and alternative hypotheses:
     - $H_0 : \mu_{M0} = \mu_{M1}$
     - $H_A : \mu_{M0} \neq \mu_{M1}$

   Questions:

   (a) What is the minimum number (i.e., n) of tumor samples required to achieve a power of 0.8 for testing a single DE gene in Experiment A with a type I error rate of $\alpha = 0.05$ and the specific H_A of $\mu_{M0} = 0.5$ and $\mu_{M1} = 1.0$? (*OK to use pwr R package*)

   (b) Does the required n calculated for Experiment A in a) differ for Experiment B?

   (c) If only 50 samples are available, i.e., $n = 50$, what is the power to detect the effect size of $\mu_{M0} = 0.5$ and $\mu_{M1} = 1.0$ with $\alpha = 0.05$?

(d) Confirm the theoretical power computed in c) via simulation. Hint: remember that power is defined under H_A.

(e) How could a researcher increase the power for the analysis of a single DE gene?

(f) Estimate the empirical power if the value of μ_{M1} for each DE gene is modeled as a random draw from $U(0,1)$, $n = 50, \alpha = 0.05$ and $\mu_{M0} = 0.1$.

2. For the coefficient of skewness question in Problem Set 6, calculate the empirical power of your normality test for the following cases:

   (a) Type I error rate of 0.05 and H_A of 100 iid Poisson RVs (i.e., part d)) with λ values of 1 to 10 (incrementing by 1). Plot empirical power vs. λ. Do the results match your expectations? Explain.

   (b) Type I error rate values between 0.01 and 0.1 (incrementing by 0.01) and H_A of 100 iid Poisson RVs with $\lambda = 1$. Plot empirical power vs. type I error rate. Do the results match your expectations? Explain.

3. (Based on Rice 12.1) Simulate observations like those of Figure 12.1 under the H_0 of no treatment effects. That is, simulate seven batches of ten normally distributed random numbers with mean 4 and variance 0.0037. Make parallel boxplots of the seven batches like those of Figure 12.1. Do this twice. Your figures display the kind of variability that random fluctuations can cause; do you see any pairs of labs that appear different in either mean level or dispersion?

4. (Based on Rice 12.3) For the one-way analysis of variance with I = 2 treatment groups, show that the F statistic is t^2, where t is the usual t statistic for a two-sample case.

5. (Based on Rice 12.21) During each of four experiments on the use of carbon tetrachloride as a worm killer, ten rats were infested with larvae. Eight days later, five rates were treated with carbon tetrachloride; the other five were kept as controls. After two more days, all the rats were killed and the numbers of worms were counted. The data.frame below contains the counts of worms for the four control groups:

```
> worms = data.frame(
+         group=as.factor(c(rep("Group I", 5), rep("Group II", 5),
+                 rep("Group III", 5), rep("Group IV", 5))),
+         count=c(279, 338,334,198,303,
+                 378,275,412,265,286,
+                 172,335,335,282,250,
+                 381,346,340,471,318))
> summary(worms)

       group       count
 Group I  :5   Min.   :172.0
 Group II :5   1st Qu.:278.0
 Group III:5   Median :326.0
 Group IV :5   Mean   :314.9
               3rd Qu.:341.5
               Max.   :471.0
```

Significant differences, although not expected, might be attributable to changes in experimental conditions. A finding of significant differences could result in more carefully controlled experimentation and thus greater precision in later work. Use both graphical techniques and the F test to test whether there are significant differences among the four groups. Use a nonparametric technique as well.