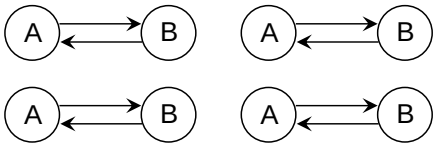
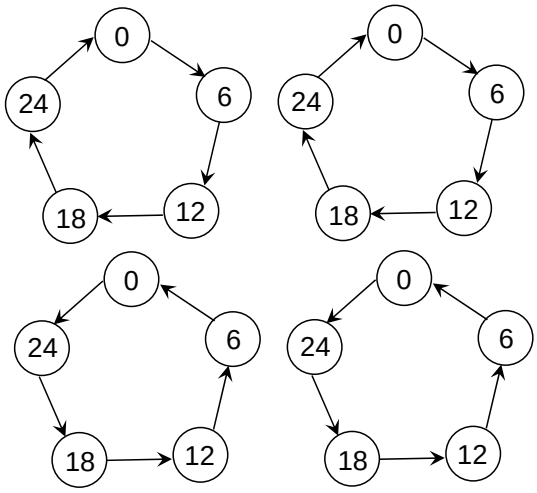
Stat 416 Homework 1 Solutions

1. *(15 points)* Many of you had difficulties with this problem. The notation calls for one circle for each experimental unit. The most common mistake was to use one circle for each sample rather than one for each experimental unit. If multiple samples from a single experimental unit are measured on multiple slides, there is still only one circle for that experimental unit. The circle will be connected to another circle or other circles using multiple arrows to indicate that the experimental unit is measured with multiple slides.

1 a)



1 b)



2. (a) *(8 points)* $y_{ij} = \mu + \tau_i + e_{ij}$ where y_{ij} denotes the normalized log signal intensity for the j^{th} mouse in the i^{th} treatment group, μ denotes the overall mean normalized log signal intensity, τ_i denotes the effect of the i^{th} treatment, and e_{ij} denotes the random residual effect associated with the j^{th} mouse in the i^{th} treatment group. In this case i is either 1 (treatment A) or 2 (treatment B). The subscript j ranges from 1 to 5 because there are 5 mice in each treatment group.

   (b) *(3 points)* $H_0 : \tau_1 = \tau_2$

   (c) *(15 points)* $\bar{y}_{1\cdot} = \frac{9.4+7.8+10.8+9.4+9.0}{5} = 9.28 \quad \bar{y}_{2\cdot} = \frac{6.3+9.2+7.1+7.2+7.5}{5} = 7.46$

$$s_p^2 = \frac{(5-1)s_1^2 + (5-1)s_2^2}{5+5-2} = \frac{(5-1)1.152 + (5-1)1.143}{5+5-2} = 1.1475$$

$$t = \frac{9.28 - 7.46}{\sqrt{1.1475(\frac{1}{5} + \frac{1}{5})}} \approx 2.6864$$

This t-statistic has 5+5-2=8 d.f. The probability that a t random variable with 8 d.f. would be greater in absolute value than 2.6864 is approximately 0.02765. Thus the p-value is 0.02765.

The problem asked for a written explanation of what this p-value means in this case. A reasonable answer would go something like this:

> *If this gene were not differentially expressed (i.e., if the treatment effects were identical), the chance of obtaining data that would yield a t-statistic that would be greater than or equal to 2.6864 in absolute value is 0.02765.*

Because this chance is small, we would typically choose to doubt the null hypothesis and conclude that the gene is differentially expressed. In this case our p-value tells us that if the gene were not differentially expressed, data like we saw would be somewhat unusual. Therefore we reject the null hypothesis and conclude that the gene is differentially expressed. Most of you know how to use a p-value to decide whether to reject or fail to reject (accept) a null hypothesis. That is not the same thing as understanding what a p-value means. That was the point of this question.

Some of you said that the p-value is the chance that the null hypothesis is true. That is a common misinterpretation of a p-value. The p-value tells us about the probability of the data under the assumption that the null hypothesis is true. It does not tell us the probability that the null hypothesis is true.

(d) *(10 points)* $9.28 - 7.46 \pm t^{(0.975)}_{5+5-2}\sqrt{1.1475(\frac{1}{5}+\frac{1}{5})} \Longrightarrow (0.26, 3.38)$

(e) *(8 points)* Estimated Fold Change $= e^{9.28-7.46} = 6.17$ 95% C.I. for fold change: $(e^{0.26}, e^{3.38}) = (1.3, 29.4)$

The computations for this question can be done in R. Try the commands below.

```
y1=c(9.4,7.8,10.8,9.4,9)
y2=c(6.3,9.2,7.1,7.2,7.5)
t.test(y1,y2,var.equal=T)
exp(c(9.28-7.46,0.26,3.38))
```

3. (a) *(8 points)* $y_{ij} = \mu + \tau_i + e_{ij}$ where y_{ij} denotes the normalized log signal intensity for the j^{th} mouse in the i^{th} treatment group, μ denotes the overall mean normalized log signal intensity, τ_i denotes the effect of the i^{th} treatment, and e_{ij} denotes the random residual effect associated with the j^{th} mouse in the i^{th} treatment group.

(b) *(3 points)* $H_0 : \tau_1 = \tau_2 = \tau_3$

(c) *(15 points)* The F-test can be obtained using the SAS code below. The relevant results are F=0.25, d.f.=2 and 12, and p-value=0.7812. The problem asked for a written explanation of what this p-value means in this case. A reasonable answer would go something like this:

> *If all treatment effects were identical, the chance of obtaining data that would yield an F-statistic greater than or equal to 0.25 is 0.7812.*

Because this chance is quite large, we cannot rule out the possibility that all treatment effects are identical in this case. Therefore we would fail to reject the null hypothesis. Most of you know how to use a p-value to decide whether to reject or fail to reject (accept) a null hypothesis. That is not the same thing as understanding what a p-value means. That was the point of this question.

Some of you said that the p-value is the chance that the null hypothesis is true. That is a common misinterpretation of a p-value. The p-value tells us about the probability of the data under the assumption that the null hypothesis is true. It does not tell us the probability that the null hypothesis is true.

R code for this problem is as follows:

```
trt=rep(c("A","B","C"),each=5)
y=c(5.9,3.8,4.2,4.0,5.5,4.9,6.9,4.5,5.5,4.1,3.4,5.6,6.7,4.0,4.6)
anova(lm(y~as.factor(trt)))
```

4. (a) *(10 points)* p-value=0.002205

```
x1=c(3.8,4.1,4.6,5.7,6.1)
x2=c(6.7,6.9,7.2,7.9,8.9)
t.test(x1,x2,var.equal=T)
```

(b) *(5 points)* When a gene is not differentially expressed, the probability that its p-value will be less than or equal to any number α between 0 and 1 is simply α. (This assumes that our model for the data is correct. It will be approximately true when our model is approximately correct.) For example, if a gene is not differentially expressed, the probability that its p-value will be less than or equal to 0.002205 is simply 0.002205. Thus the sort of thing we saw in part (a) is not likely to happen. (Of course it will happen sometimes, especially when we test many genes for which the null hypothesis is true.) On the other hand, if a gene is differentially expressed, its p-value is likely to be small. The more differentially expressed a gene is, the smaller its p-value is expected to be. For example, it may be that for a differentially expressed, the probability that its p-value will be less than 0.002205 is 0.9 rather than 0.002205. (The exact probability will depend on the degree of differential expression.) Thus when we see a small p-value we can either decide that the gene is not differentially expressed and something unusual happened by chance, or we can choose to believe that the small p-value is due to differential expression. We will undoubtedly be wrong from time to time no matter which choice we make. We will see later in the course that the p-value can provide a useful ranking of genes that can do a reasonable job of separating the differentially expressed and non-differentially expressed genes.