Stat 516 Homework 1 Solutions

1. (a) GAUUACACGUGCCUUGGA

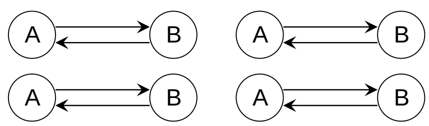
   (b) asp tyr thr cys gly

   (c) The amino acid cys would change to the stop codon. Thus, we would end up with the sequence asp tyr thr.

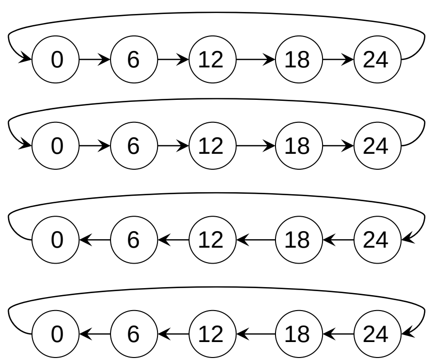
2. See slide 9 of slide set number 3.

3. 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. Correct answers are as follows:

a)



b)



4. $(2^3)^{(16\cdot16)} = 2^{768}$

5. The following SAS code can be used to obtain answers to parts (a), (b), and (d).

```
proc mixed;
  class slide dye trt;
  model y=trt dye;
  random slide;
  estimate 'trt 2 - trt 1' trt -1 1 / cl;
  ods output estimates=estimates;
run;
```

```
data estimates; set estimates;
  fc=exp(estimate);
  lfc=exp(lower);
  ufc=exp(upper);
run;

proc print data=estimates;
run;
```

Alternatively, R can be used. Cut and paste the data to a text file called pr2data.txt with column headings slide, dye, trt, and y. Open an R workspace and set the working director to the directory containing the file using the R function setwd. Then issue the following commands

```
d=read.table("pr2data.txt",header=T)
library(nlme)
out=lme(y~factor(trt)+factor(dye),data=d,random=~1|factor(slide))
summary(out)
anova(out)
intervals(out)
exp(intervals(out)$fixed[2,])
```

(a) Treatment: F=7.35, d.f.= 1 and 6, p-value=0.0350
Dye: F=28.84, d.f.=1 and 6, p-value=0.0017

The p-values provide convincing evidence of differing treatment effects and differing dye effects.

(b) $\widehat{\tau_2 - \tau_1} = 1.3$
95% Confidence interval for $\tau_2 - \tau_1$ is $(0.1268, 2.4732)$

$$
\begin{array}{rclclcl}
2.7 & = & 7.9\text{-}5.2 & = & y_{221} - y_{111} & = & \tau_2 - \tau_1 + \delta_2 - \delta_1 + e_{221} - e_{111} \\
3.2 & = & 6.5\text{-}3.3 & = & y_{222} - y_{112} & = & \tau_2 - \tau_1 + \delta_2 - \delta_1 + e_{222} - e_{112} \\
4.8 & = & 5.9\text{-}1.1 & = & y_{223} - y_{113} & = & \tau_2 - \tau_1 + \delta_2 - \delta_1 + e_{223} - e_{113} \\
4.8 & = & 7.6\text{-}2.8 & = & y_{224} - y_{114} & = & \tau_2 - \tau_1 + \delta_2 - \delta_1 + e_{224} - e_{114} \\
-1.3 & = & 6.3\text{-}7.6 & = & y_{215} - y_{125} & = & \tau_2 - \tau_1 + \delta_1 - \delta_2 + e_{215} - e_{125} \\
-0.2 & = & 6.0\text{-}6.2 & = & y_{216} - y_{126} & = & \tau_2 - \tau_1 + \delta_1 - \delta_2 + e_{216} - e_{126} \\
-0.1 & = & 4.0\text{-}4.1 & = & y_{217} - y_{127} & = & \tau_2 - \tau_1 + \delta_1 - \delta_2 + e_{217} - e_{127} \\
-3.5 & = & 7.2\text{-}10.7 & = & y_{218} - y_{128} & = & \tau_2 - \tau_1 + \delta_1 - \delta_2 + e_{218} - e_{128}
\end{array}
$$

(c)

$$
\begin{array}{rclclcl}
2.7 & = & 7.9\text{-}5.2 & = & y_{221} - y_{111} & = & (\tau_2 - \tau_1) + (\delta_2 - \delta_1)(1) + e_{221} - e_{111} \\
3.2 & = & 6.5\text{-}3.3 & = & y_{222} - y_{112} & = & (\tau_2 - \tau_1) + (\delta_2 - \delta_1)(1) + e_{222} - e_{112} \\
4.8 & = & 5.9\text{-}1.1 & = & y_{223} - y_{113} & = & (\tau_2 - \tau_1) + (\delta_2 - \delta_1)(1) + e_{223} - e_{113} \\
4.8 & = & 7.6\text{-}2.8 & = & y_{224} - y_{114} & = & (\tau_2 - \tau_1) + (\delta_2 - \delta_1)(1) + e_{224} - e_{114} \\
-1.3 & = & 6.3\text{-}7.6 & = & y_{215} - y_{125} & = & (\tau_2 - \tau_1) + (\delta_2 - \delta_1)(-1) + e_{215} - e_{125} \\
-0.2 & = & 6.0\text{-}6.2 & = & y_{216} - y_{126} & = & (\tau_2 - \tau_1) + (\delta_2 - \delta_1)(-1) + e_{216} - e_{126} \\
-0.1 & = & 4.0\text{-}4.1 & = & y_{217} - y_{127} & = & (\tau_2 - \tau_1) + (\delta_2 - \delta_1)(-1) + e_{217} - e_{127} \\
-3.5 & = & 7.2\text{-}10.7 & = & y_{218} - y_{128} & = & (\tau_2 - \tau_1) + (\delta_2 - \delta_1)(-1) + e_{218} - e_{128}
\end{array}
$$

Note that the differences in residual random effects are independent and normally distributed with mean 0 and constant variance as long as these original assumptions hold for the original random effects. Thus, we can regress the differences 2.7, 3.2, 4.8, 4.8, -1.3, -0.2, -0.1, -3.5 against the x values 1, 1, 1, 1, -1, -1, -1, -1 to obtain estimates and standard errors of the intercept=$\tau_2 - \tau_1$ and slope=$\delta_2 - \delta_1$. This regression is easy to carry out in SAS or R. Using the estimates and standard errors, it is straightforward to obtain results identical to those obtained in parts (a) and (b). Example SAS and R code follows.

```
/*
Read data from a tab delimited text file that
has the same format as the data in the homework
assignment.
*/

PROC IMPORT file="C:\d1.txt" out=one
            DBMS=TAB REPLACE;
RUN;

data a; set one;
  if trt=1;
  a=y;
run;

data b; set one;
  if trt=2;
  b=y;
run;

data two; merge a b;
  y=b-a;
  dye=2*(dye-1.5);
run;

proc reg;
  model y=dye;
run;
```

Assuming that the data frame d has been created as described above, R commands are as follows:

```
y=d$y[d$trt=="B"]-d$y[d$trt=="A"]
x=rep(c(1,-1),c(4,4))
summary(lm(y~x))
```

(d) Exponentiating the point estimate and confidence interval endpoints above yields: estimated fold change = 3.66930, 95% Confidence interval for the fold change $e^{\tau_2-\tau_1}$ is $(1.13515, 11.8607)$.

6. (a) Note that $\log(X_1/Y_1) = \log(X_1) - \log(Y_1) \sim N(\mu_x - \mu_y, 2\sigma^2)$. Thus, X_1/Y_1 is log normally distributed with mean $\exp\{\mu_x - \mu_y + \sigma^2\}$.

(b)

$$E(X_1)/E(Y_1) = \exp\{\mu_x + \sigma^2/2\}/\exp\{\mu_y + \sigma^2/2\} = \exp\{\mu_x - \mu_y\}.$$

(c) Let $U_i = \log(X_i)$ and $V_j = \log(Y_j)$ for all $i = 1, \ldots, m$ and $j = 1, \ldots, n$. Let $\bar{U} = \frac{1}{m}\sum_{i=1}^{m} U_i$, and let $\bar{V} = \frac{1}{n}\sum_{j=1}^{n} V_j$. The log likelihood function (aside from a constant that does not depend on parameters) is

$$-\frac{m+n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\left\{\sum_{i=1}^{m}(U_i - \mu_x)^2 + \sum_{j=1}^{n}(V_j - \mu_y)^2\right\}.$$

Standard calculus can be used to show that the maximum likelihood estimators of μ_x, μ_y, and σ^2 are $\bar{U}$, $\bar{V}$, and

$$\frac{1}{m+n}\left\{\sum_{i=1}^{m}(U_i - \bar{U})^2 + \sum_{j=1}^{n}(V_j - \bar{V})^2\right\},$$

respectively. Thus, the the maximum likelihood estimator for part (a) is

$$\exp\left[\bar{U} - \bar{V} + \frac{1}{m+n}\left\{\sum_{i=1}^{m}(U_i - \bar{U})^2 + \sum_{j=1}^{n}(V_j - \bar{V})^2\right\}\right],$$

and the maximum likelihood estimator for part (b) is

$$\exp\{\bar{U} - \bar{V}\}.$$