1a) Assuming all of the hypotheses are independent and exactly level α=0.05, then if all 19 of the H_0 are true the various reject decisions are independent Bernoulli random variables with p=0.05. The chance of rejecting will be higher if any of the H_1 are true. We can thus test H_0* vs. H_1* by testing the null hypothesis p=0.05 vs. the alternate p>0.05 using a binomial distribution with n=19 and p=0.05 where 3 (the number of rejections ("successes") out of 19 at the α=0.05 level) is our test statistic. The p-value would be P[Number of successes >= 3 | n=19, p=0.05]. From R, this can be found with 1-pbinom(2,19,.05)= 0.0665. With a p-value > alpha we fail to reject the null hypothesis and do not have significant evidence of discrimination at the 0.05 level.

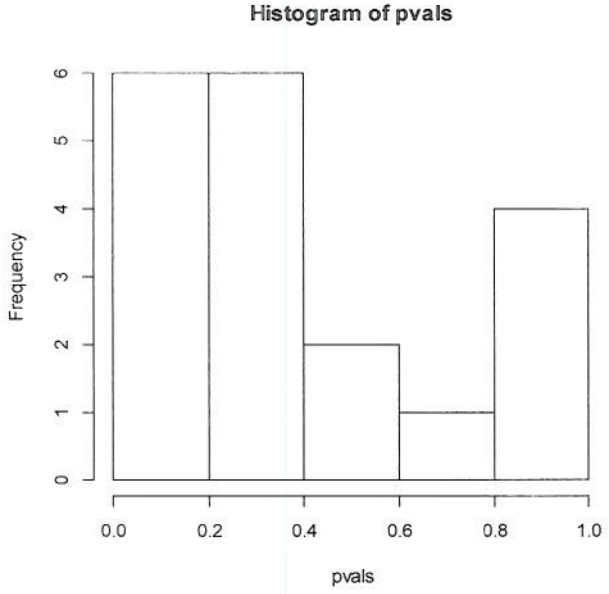
b) The rejection region is the largest right tail of the binomial with probability 5% or less. As seen in (a), rejecting for 3 or more successes gives a 0.0665 level test. Rejecting for 4 or more gives a 0.0132 level test (1-pbinom(3,19,.05)). So the rejection region would be 4 or more and it would be a level 0.0132 test.

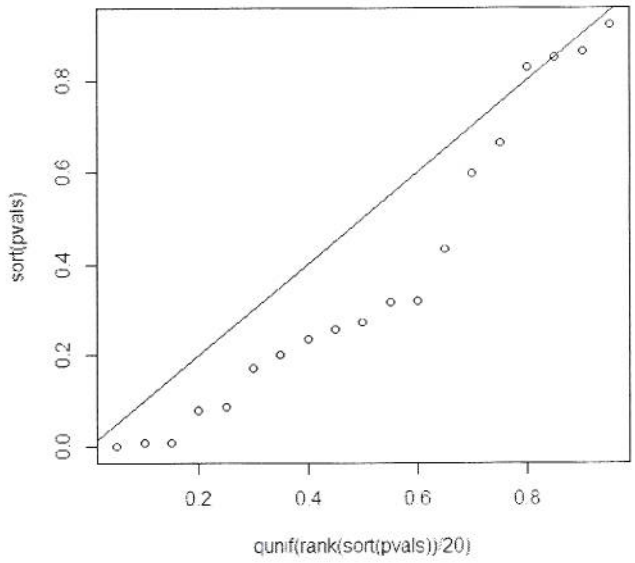
The power for this test is P[Number of successes >= 4 | n=19, p=0.2]. From R, this is 1 - pbinom(3, 19,.2) ≈ 0.545.

Note: For a randomized α=0.05 level test , you would reject for 4 or more successes and reject .6903 of the time with 3 successes. (.0132 + .6903 (.0533) = 0.05). This would change part (a) so that you would need to generate a uniform (0,1) random variable and reject if it was 0.6903 or less. The power for this randomized test would be:
P[Number of successes >= 4 | n=19, p=0.2] + 0.6903 P[Number of successes = 3 | n=19, p=0.2]
From R, this is 0.545 + 0.6903*dbinom(3,19,.2) ≈ 0.696.

c) Assuming the 19 p-values come from a test with a continuous test statistic, the p-values under the null hypothesis will follow a uniform distribution. It looks like the p-value distribution is piled up near the small values, and lacks many between 0.4 and 0.8.
pvals<-c(0.000, 0.006, 0.007, 0.078, 0.086, 0.171, 0.201, 0.234, 0.255, 0.273, 0.318, 0.319, 0.435, 0.597, 0.664, 0.828, 0.850, 0.862, 0.921)

```
#Histogram
hist(pvals)
#Quantile quantile plot against a uniform distribution
plot(qunif(rank(sort(pvals))/20),sort(pvals))
lines(c(0,1),c(0,1))
```





d) Simply testing the mean of the distribution doesn't get the entire shape. Since the p-values should tend to take small values under H_A, one possibility is a a one sided Kolmogorov-Smirnov test ks.test(pvals,qunif,0,1,alternative="greater") gives a p-value of 0.01896 and we can reject the null hypothesis at the α=0.05 level, concluding that the p-values support H_1*. (The t-test that the mean p-value is 0.5 versus the alternate that it is less gives a p-value of 0.0501 and would fail to find statistically significant evidence of lack of fit at the α=0.05 level.)

#2 $X \sim N(0,1)$ $\quad X \perp\!\!\!\perp Y$.
$Y \sim N(0,1)$

(a) The mgf of $Z_1 = X+Y$ is

$$M_{Z_1}(t) = M_X(t) M_Y(t) \quad \leftarrow \text{mgf of } N(0,2)$$
$$= e^{t^2/2} e^{t^2/2} = e^{2t^2/2}.$$

Therefore $Z_1 \sim N(0,2)$.

The mgf of $Z_2 = X-Y$ is

$$M_{Z_2}(t) = M_X(t) M_Y(-t)$$
$$= e^{t^2/2} e^{(-t)^2/2} = e^{2t^2/2}$$

Therefore $Z_2 \sim N(0,2)$. Thus, $Z_1 = X+Y \overset{d}{=} X-Y = Z_2$.

(b) $Z = \min(X,Y)$. Clearly $Z^2 \geq 0$. The cdf of Z^2 is

$$F_{Z^2}(z) = P(Z^2 \leq z)$$
$$= P(-\sqrt{z} \leq \min(X,Y) \leq \sqrt{z})$$
$$= P(\min(X,Y) \leq \sqrt{z}) - P(\min(X,Y) \leq -\sqrt{z})$$
$$= [1 - P(\min(X,Y) > \sqrt{z})] - [1 - P(\min(X,Y) > -\sqrt{z})]$$
(*) $$= P(\min(X,Y) > -\sqrt{z}) - P(\min(X,Y) > \sqrt{z})$$

Note: $\{\min(X,Y) > -\sqrt{z}\} = \{X > -\sqrt{z} \text{ and } Y > -\sqrt{z}\}$
$\{\min(X,Y) > \sqrt{z}\} = \{X > \sqrt{z} \text{ and } Y > \sqrt{z}\}$.

(*) $\overset{X \perp\!\!\!\perp Y}{=} P(X > -\sqrt{z}) P(Y > -\sqrt{z}) - P(X > \sqrt{z}) P(Y > \sqrt{z})$
$= [P(X > -\sqrt{z})]^2 - [P(X > \sqrt{z})]^2$
↑ identical distribution

$$= [1 - F_X(-\sqrt{z})]^2 - [1 - F_X(\sqrt{z})]^2$$

Note: $X \sim N(0,1) \Rightarrow 1 - F_X(\sqrt{z}) = F_X(-\sqrt{z})$.

Therefore, the last expression equals

$$[1 - F_X(-\sqrt{z})]^2 - [F_X(-\sqrt{z})]^2$$

$$= 1 - 2F_X(-\sqrt{z}) + [F_X(-\sqrt{z})]^2 - [F_X(-\sqrt{z})]^2$$

$$= 1 - 2F_X(-\sqrt{z}).$$

Therefore, the pdf of Z^2, for $z^2 \geq 0$, is

$$f_{Z^2}(z) = \frac{d}{dz} F_{Z^2}(z)$$

$$= \frac{d}{dz} [1 - 2F_X(-\sqrt{z})]$$

$$= 0 - 2 f_X(-\sqrt{z}) \cdot (-1) \frac{1}{2} z^{-1/2}$$

(chain rule)

$$= \frac{1}{z^{1/2}} f_X(-\sqrt{z})$$

$$= \frac{1}{z^{1/2}} \frac{1}{\sqrt{2\pi}} e^{-(-\sqrt{z})^2/2}$$

$$= \frac{1}{z^{1/2}} \frac{1}{\sqrt{2\pi}} z^{\frac{1}{2}-1} e^{-z/2}.$$

$\Gamma(\frac{1}{2}) = \sqrt{\pi}$

$$= \frac{1}{\Gamma(\frac{1}{2}) 2^{1/2}} z^{\frac{1}{2}-1} e^{-z/2}.$$

This is the pdf of a χ^2_1 r.v. Therefore, the result.

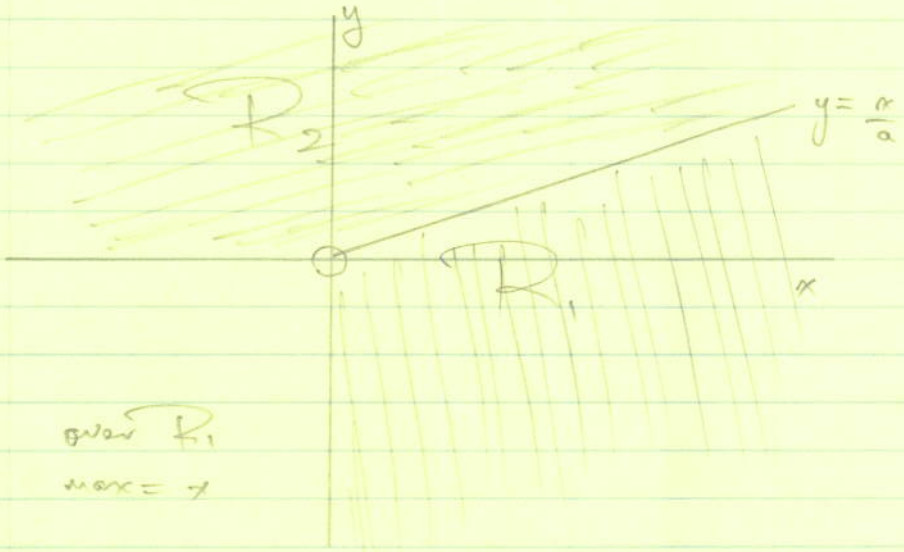
(c) The joint pdf of (X, Y), for all $(x, y) \in \mathbb{R}^2$, is

$$f_{X,Y}(x, y) = f_X(x) f_Y(y) \quad \text{(indep)}$$

$$= \frac{1}{2\pi} e^{-(x^2+y^2)/2}.$$

To find $E(X \max(0, X, aY))$, we need to integrate over the correct region.

Picture:



over R_2, $\max = ay$

over R_1, $\max = x$

Therefore, $E(X \max(0, X, aY))$ is given by:

$$\iint_{R_1} x^2 \frac{1}{2\pi} e^{-(x^2+y^2)/2}\, dx\, dy + \iint_{R_2} axy\, e^{-(x^2+y^2)/2}\, dx\, dy$$

We calculate each integral separately. In both, we use polar coordinates:

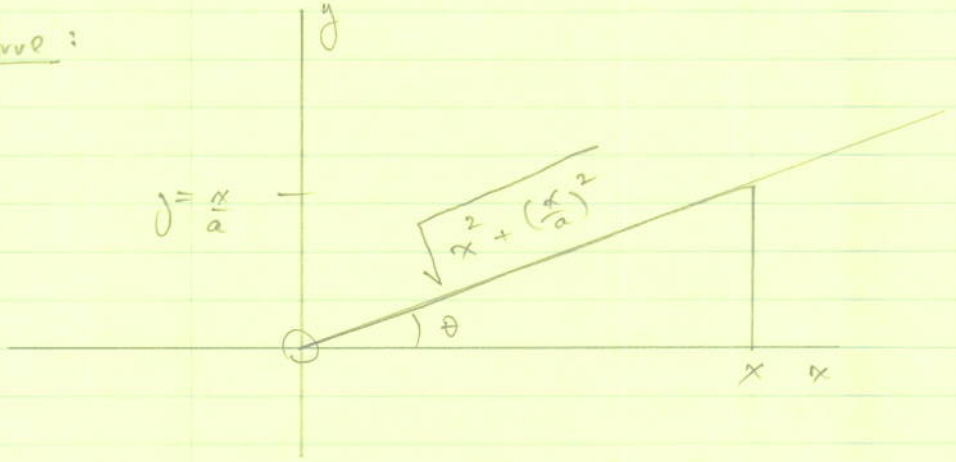
$$x = r\cos\theta$$
$$y = r\sin\theta$$

The integral over R_1 is

$$I_1 = \int_{\theta=-\pi/2}^{\tan^{-1}(\frac{1}{a})} \int_{r=0}^{\infty} r^2 \cos^2\theta\, \frac{1}{2\pi} e^{-r^2/2}\, r\, dr\, d\theta$$

To see why these limits are correct, consider the picture:

Picture:



$$\tan\theta = \frac{x/a}{x} = \frac{1}{a} \iff \theta = \tan^{-1}\left(\frac{1}{a}\right).$$

Therefore,

$$I_1 = \frac{1}{2\pi} \int_{\theta=-\pi/2}^{\tan^{-1}(\frac{1}{a})} \int_{r=0}^{\infty} r^2 \cos^2\theta \; e^{-r^2/2} \; dr \; d\theta$$

$$= \frac{1}{2\pi} \int_{\theta=-\frac{\pi}{2}}^{\tan^{-1}(\frac{1}{a})} \cos^2\theta \; d\theta \int_{r=0}^{\infty} r^2 e^{-r^2/2} \; dr$$

$$= 2$$

Note:

$$\int_{r=0}^{\infty} r^2 e^{-r^2/2} \; dr = \int_{u=0}^{\infty} r^2 e^{-u/2} \frac{du}{2r}$$

$u = r^2$
$du = 2r \, dr$

$$= \frac{1}{2} \int_{u=0}^{\infty} u \, e^{-u/2} \; du$$

$$= \frac{1}{2} \Gamma(2) \, 2^2 = 2.$$

Therefore,

$$I_1 = \frac{1}{\pi} \int_{\theta=-\pi/2}^{+\tan^{-1}(\frac{1}{a})} \cos^2\theta \, d\theta \qquad (*)$$

Recall: $\cos(2\theta) = 2\cos^2\theta - 1$

$$\Rightarrow \cos^2\theta = \frac{1+\cos(2\theta)}{2}.$$

$$(*) = \frac{1}{\pi} \int_{\theta=-\pi/2}^{+\tan^{-1}(\frac{1}{a})} \left(\frac{1}{2} + \frac{\cos(2\theta)}{2} \right) d\theta$$

$$= \frac{1}{\pi} \left[\frac{\theta}{2} + \frac{\sin(2\theta)}{4} \right]_{\theta=-\pi/2}^{+\tan^{-1}(\frac{1}{a})}$$

$$= \frac{1}{4\pi} \left[2\theta + \sin(2\theta) \right]_{\theta=-\pi/2}^{+\tan^{-1}(\frac{1}{a})}$$

Recall: $\sin(2\theta) = 2\sin\theta\cos\theta$

See Figure on last page:

$$\sin\theta = \frac{x/a}{\sqrt{x^2 + \left(\frac{x}{a}\right)^2}} = \frac{1}{a\sqrt{1+\frac{1}{a^2}}}$$

$$\cos\theta = \frac{x}{\sqrt{x^2 + \left(\frac{x}{a}\right)^2}} = \frac{1}{\sqrt{1+\frac{1}{a^2}}}$$

$$\Rightarrow \sin(2\theta) = \frac{2}{a}\left(\frac{1}{1+\frac{1}{a^2}}\right) = \frac{2a}{1+a^2}.$$

Therefore,

$$I_1 = \frac{1}{4\pi}\left[2\tan^{-1}\left(\frac{1}{a}\right) + \frac{2a}{1+a^2} + \pi - \underbrace{\sin(-\pi)}_{=0}\right]$$

Recall: $\tan^{-1}(a) + \tan^{-1}\left(\frac{1}{a}\right) = \frac{\pi}{2}$

Therefore,

$$I_1 = \frac{1}{4\pi}\left[2\left(\frac{\pi}{2} - \tan^{-1}(a)\right) + \frac{2a}{1+a^2} + \pi\right]$$

$$= \frac{1}{4\pi}\left[2\pi - 2\tan^{-1}(a) + \frac{2a}{1+a^2}\right]$$

The integral over R_2 is

$$I_2 = \int_{\theta=\tan^{-1}\left(\frac{1}{a}\right)}^{\frac{\pi}{2}} \int_{r=0}^{b} a\, r\cos\theta \sin\theta \, \frac{1}{2\pi} e^{-r^2/2}\, r\, dr\, d\theta$$

$$= \frac{1}{2\pi}\int_{\theta=\tan^{-1}\left(\frac{1}{a}\right)}^{\frac{\pi}{2}} \cos\theta \sin\theta \, d\theta \underbrace{\int_{r=0}^{\infty} a\, r^3 e^{-r^2/2}\, dr}_{=2a}$$

$$= \frac{2a}{2\pi}\int_{\theta=\tan^{-1}\left(\frac{1}{a}\right)}^{\frac{\pi}{2}} \cos\theta \sin\theta \, d\theta \qquad (*)$$

Recall: $\sin(2\theta) = 2\sin\theta\cos\theta \iff \cos\theta\sin\theta = \frac{\sin(2\theta)}{2}$

$$* = \frac{2a}{2\pi} \int_{\theta = \tan^{-1}\left(\frac{1}{a}\right)}^{\pi} \frac{\sin(2\theta)}{2} \, d\theta$$

$$= \frac{a}{2\pi} \int_{\theta = \tan^{-1}\left(\frac{1}{a}\right)}^{\pi} \sin(2\theta) \, d\theta$$

$$= \frac{a}{2\pi} \left[-\frac{\cos(2\theta)}{2} \Big|_{\theta = \tan^{-1}\left(\frac{1}{a}\right)}^{\pi} \right]$$

$$= \frac{a}{4\pi} \left[\cos(2\theta) \Big|_{\theta = \pi}^{\tan^{-1}\left(\frac{1}{a}\right)} \right]$$

Recall: $\cos(2\theta) = 2\cos^2\theta - 1$

$$= 2\left(\frac{1}{\sqrt{1+\frac{1}{a^2}}}\right)^2 - 1$$

$$= \frac{2a}{1+a^2} - 1 = \frac{a^2 - 1}{1+a^2}.$$

Therefore,

$$I_2 = \frac{a}{4\pi}\left(\frac{a^2-1}{1+a^2} - 1\right) = \frac{1}{4\pi}\left(-\frac{2a}{1+a^2}\right)$$

Finally,

$$I_1 + I_2 = \frac{1}{4\pi}\left[\pi - 2\tan^{-1}(a) + \frac{2a}{1+a^2} - \frac{2a}{1+a^2}\right]$$

$$= \frac{1}{2} - \frac{1}{\pi}\tan^{-1}(a).$$

```
## R simulation
B = 10000 # number of MC simulations
a = 1

x = rnorm(B,0,1)
y = rnorm(B,0,1)
z = x*0
for (i in 1:B){
    z[i] = max(0,x[i],a*y[i])
    }

est = sum(x*z)/B
true = 0.5 - (1/(2*pi))*atan(a)
est
est - true

> est
[1] 0.3755757
> est - true
[1] 0.0005756839
```

#3 (a) $X_1, X_2, \ldots, X_n$ are iid $F_X(x)$, where

$$F_X(x) = p F_1(x) + (1-p) F_2(x),$$

where

$$F_1(x) = \begin{cases} 0, & x \le -\frac{1}{2} \\ x + \frac{1}{2}, & -\frac{1}{2} < x < \frac{1}{2} \\ 1, & x \ge \frac{1}{2} \end{cases} \qquad F_2(x) = \begin{cases} 0, & x \le 0 \\ x, & 0 < x < 1 \\ 1, & x \ge 1 \end{cases}$$

Therefore,

$$F_X(x) = \begin{cases} 0, & x \le -\frac{1}{2} \\ p(x+\frac{1}{2}), & -\frac{1}{2} < x < 0 \\ x + \frac{p}{2}, & 0 \le x < \frac{1}{2} \\ p + (1-p)x, & \frac{1}{2} \le x < 1 \\ 1, & x \ge 1 \end{cases}$$

← CDF of X_1

The pdf of X_1 is

$$f_X(x) = \begin{cases} p, & -\frac{1}{2} < x < 0 \\ 1, & 0 \le x < \frac{1}{2} \\ 1-p, & \frac{1}{2} \le x < 1 \\ 0, & \text{otherwise} \end{cases}$$

We can also write $f_X(x)$ as

$$f_X(x) = p\, I(-\tfrac{1}{2} < x < 0) + I(0 \le x < \tfrac{1}{2}) + (1-p)\, I(\tfrac{1}{2} \le x < 1)$$

(b) The likelihood function is

$$L(p \mid \underset{\sim}{x}) = \prod_{i=1}^{n} f_X(x_i)$$

$$= \prod_{i=1}^{n} \left[p\, I(-\tfrac{1}{2} < x_i < 0) + I(0 \le x_i < \tfrac{1}{2}) + (1-p)\, I(\tfrac{1}{2} \le x_i < 1) \right]$$

$$= p^{y_1 + y_2} (1-p)^{y_3} = p^{y_1} (1-p)^{y_3}$$

where $y_1 = \sum_{i=1}^{n} I(-\frac{1}{2} < x_i < 0) = \#\, X_i$ in $(-\frac{1}{2}, 0)$

$y_2 = \sum_{i=1}^{n} I(0 \le x_i < \frac{1}{2}) = \#\, X_i$ in $[0, \frac{1}{2})$

and

$$y_3 = \sum_{i=1}^{n} \mathbb{I}\left(\tfrac{1}{2} \le x_i < 1\right) = \#\, X_i \text{ in } \left[\tfrac{1}{2}, 1\right).$$

The log-likelihood function is

$$\ln L(p \mid \underset{\sim}{x}) = y_1 \ln p + y_3 \ln(1-p)$$

The score function is

$$S(p \mid \underset{\sim}{x}) = \frac{\partial \ln L(p \mid \underset{\sim}{x})}{\partial p} = \frac{y_1}{p} - \frac{y_3}{1-p} \overset{\text{set}}{=} 0$$

$$\Rightarrow \frac{y_1(1-p) - y_3 p}{p(1-p)} = 0$$

$$\Rightarrow y_1 - p y_1 - p y_3 = 0$$

$$\Rightarrow \hat{p} = \frac{y_1}{y_1 + y_3}.$$

Note:

$$\frac{\partial^2 \ln L(p \mid \underset{\sim}{x})}{\partial p^2} = -\frac{y_1}{p^2} - \frac{y_3}{(1-p)^2} < 0.$$

Therefore, the MLE of p is

$$\hat{p} = \frac{Y_1}{Y_1 + Y_3}, \quad \text{where } Y_1 = \sum_{i=1}^{n} \mathbb{I}\left(-\tfrac{1}{2} < X_i < 0\right)$$
$$Y_3 = \sum_{i=1}^{n} \mathbb{I}\left(\tfrac{1}{2} \le X_i < 1\right).$$

Applying standard results for MLEs,

$$\sqrt{n}\,(\hat{p} - p) \xrightarrow{d} N(0, \sigma_{\hat{p}}^2), \quad \text{as } n \to \infty$$

where

$$(\sigma_{\hat{p}}^2)^{-1} = E\left[\frac{\partial^2 \ln f_X(X \mid p)}{\partial p^2}\right] \leftarrow \text{Fisher Information based on 1 observation.}$$

The Fisher information based on n observations is

$$I_n(p) = E\left[-\frac{\partial^2 \ln L(p \mid \underset{\sim}{X})}{\partial p^2}\right]$$

$$= E\left(\frac{Y_1}{p^2} + \frac{Y_2}{(1-p)^2}\right)$$

Note:

$$E(Y_1) = E\left(\sum_{i=1}^n I\left(-\tfrac{1}{2} < X_i < 0\right)\right)$$
$$= \sum_{i=1}^n P\left(-\tfrac{1}{2} < X_i < 0\right)$$
$$= n\left(\frac{p}{2}\right)$$

$$E(Y_2) = E\left(\sum_{i=1}^n I\left(\tfrac{1}{2} \le X_i < 1\right)\right)$$
$$= \sum_{i=1}^n P\left(\tfrac{1}{2} \le X_i < 1\right)$$
$$= n\left(\frac{1-p}{2}\right).$$

Therefore,

$$I_n(p) = \frac{np}{2p^2} + \frac{n(1-p)}{2(1-p)^2}$$

$$= \frac{n}{2}\left(\frac{1}{p} + \frac{1}{1-p}\right) = \frac{n}{2p(1-p)}.$$

Therefore,

$$(\sigma^2_{\hat{p}})^{-1} = \frac{1}{2p(1-p)} \longleftarrow \text{Fisher information based on 1 observation.}$$

Finally,

$$\sqrt{n}\,(\hat{p} - p) \xrightarrow{d} N\left(0,\ 2p(1-p)\right); \quad \text{as } n \to \infty.$$

i.e., $\hat{p} \sim AN\left(p,\ \frac{2p(1-p)}{n}\right)$ for "large n."

(c) We have

$$E(X) = \int_{\mathbb{R}} x f_X(x)\, dx$$

$$= \int_{-\frac{1}{2}}^{0} x p\, dx + \int_{0}^{\frac{1}{2}} x\, dx + \int_{\frac{1}{2}}^{1} x(1-p)\, dx$$

$$= p \frac{x^2}{2}\Big|_{-\frac{1}{2}}^{0} + \frac{x^2}{2}\Big|_{0}^{\frac{1}{2}} + (1-p)\frac{x^2}{2}\Big|_{\frac{1}{2}}^{1}$$

$$= p\left(-\frac{1}{8}\right) + \frac{1}{8} + (1-p)\left(\frac{3}{8}\right)$$

$$= \frac{1}{2}(1-p).$$

So

$$\bar{X} = E(X) = \frac{1}{2}(1-p)$$

$$\Rightarrow 2\bar{X} = 1-p$$

$$\Rightarrow \hat{p}_{MOM} = 1-2\bar{X}.$$

We first find the asymptotic distribution of $\bar{X}$.
By CLT,

$$\sqrt{n}(\bar{X}-\mu) \xrightarrow{d} N(0,\sigma^2)$$

where

$$\mu = E(X) = \frac{1}{2}(1-p)$$
$$\sigma^2 = \mathrm{var}(X)$$

$$E(X^2) = \int_{\mathbb{R}} x^2 f_X(x)\, dx$$

$$= \int_{-\frac{1}{2}}^{0} x^2 p\, dx + \int_{0}^{\frac{1}{2}} x^2\, dx + \int_{\frac{1}{2}}^{1} x^2(1-p)\, dx$$

$$= p\left.\frac{x^3}{3}\right|_{-\frac{1}{2}}^{0} + \left.\frac{x^3}{3}\right|_{0}^{\frac{1}{2}} + (1-p)\left.\frac{x^3}{3}\right|_{\frac{1}{2}}^{1}$$

$$= p\left(\frac{1}{24}\right) + \frac{1}{24} + (1-p)\left(\frac{7}{24}\right) = \frac{1}{3} - \frac{1}{4}p$$

Thus,

$$\sigma^2 = \text{var}(X) = \frac{1}{3} - \frac{1}{4}p - \left[\frac{1}{2}(1-p)\right]^2$$

$$= \frac{1}{3} - \frac{1}{4}p - \frac{1}{4} + \frac{1}{2}p - \frac{1}{4}p^2$$

$$= \frac{1}{12} + \frac{1}{4}p - \frac{1}{4}p^2$$

$$= \frac{1}{12}(1 + 3p - 3p^2).$$

By the CLT,

$$\sqrt{n}\left(\bar{X} - \tfrac{1}{2}(1-p)\right) \xrightarrow{d} N\left(0, \frac{1}{12}(1+3p-3p^2)\right)$$

Now,

$$\hat{p}_{MOM} = 1 - 2\bar{X} = g(\bar{X}), \text{ say.}$$

The Delta Method gives

$$\sqrt{n}\left(\hat{p}_{MOM} - p\right) \xrightarrow{d} N\left(0, [g'(p)]^2 \frac{1}{12}(1+3p-3p^2)\right)$$

Note: $g(p) = 1 - 2p$

$g'(p) = -2$

$[g'(p)]^2 = 4.$

$$\xrightarrow{d} N\left(0, \frac{1}{3}(1+3p-3p^2)\right).$$

In other words,

$$\hat{p}_{MOM} \sim AN\left(p, \frac{1}{3n}(1+3p-3p^2)\right)$$

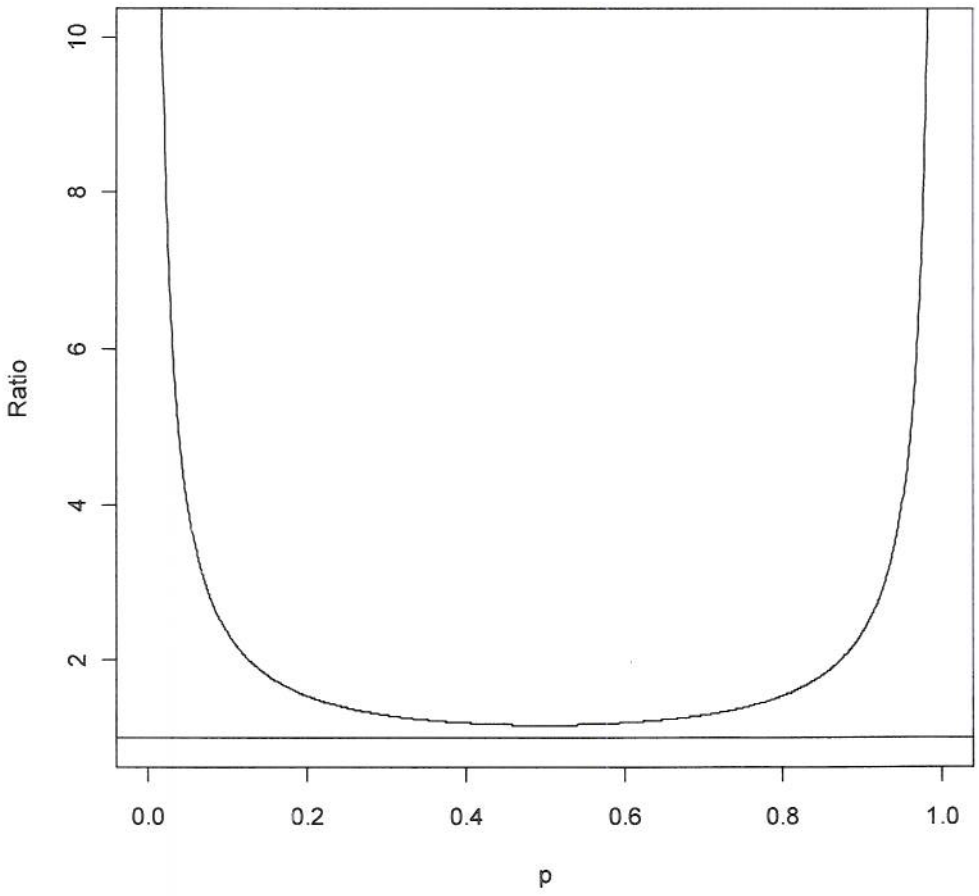
for "large n."

(d) Compare the asymptotic variances of $\hat{p}$ & $\hat{p}_{MOM}$ ($\hat{p}$: MLE)

$$\frac{avar(\hat{p}_{MOM})}{avar(\hat{p})} = \frac{\frac{1}{3}(1+3p-3p^2)}{2p(1-p)} = f(p), \text{ say.}$$

A plot of $f(p)$ versus p is given on the next page.

```
p = seq(0.001,0.999,0.001)
ratio = ((1/3)+p-p^2)/(2*p*(1-p))
plot(p,ratio,type="l",xlab="p",ylab="Ratio",ylim=c(1,10))
abline(h=1)
```



The ratio of the asymptotic variances (MOM to MLE) is always larger than 1. Therefore, the MLE is more efficient.

#4

I've included SAS code first, followed by answers and supporting graphs and tables for each section of the question.

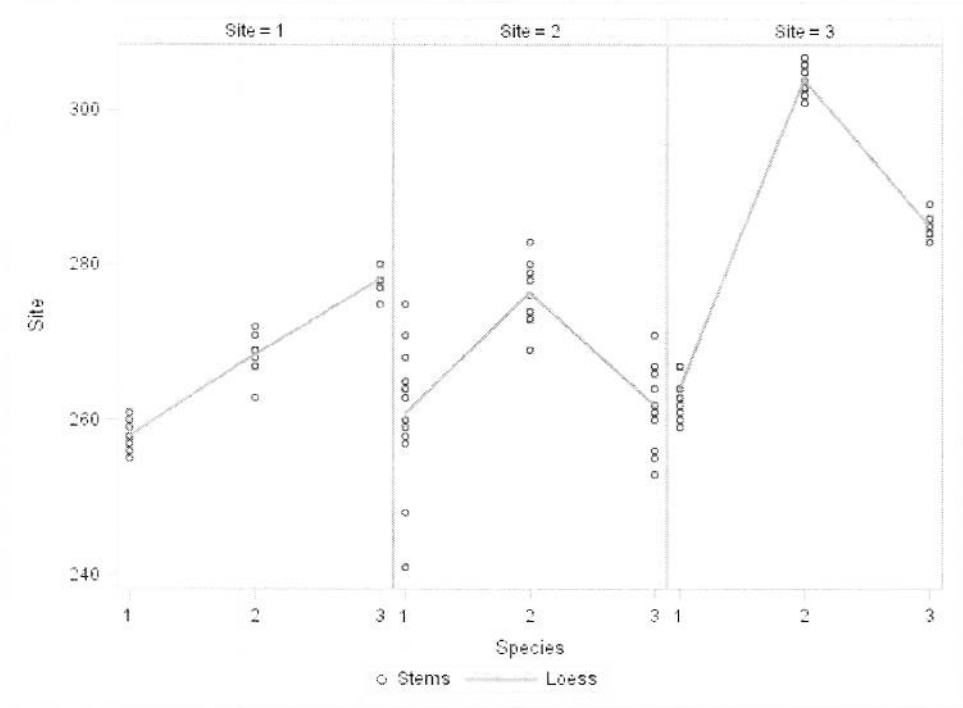
SAS code:

```
*Import Q1;
PROC IMPORT OUT= WORK.Q1
            DATAFILE= "Z:\QUAL
2014\QualJG2014.xlsx"
            DBMS=EXCEL REPLACE;
     RANGE="Q1$";
     GETNAMES=YES;
     MIXED=NO;
     SCANTEXT=YES;
     USEDATE=YES;
     SCANTIME=YES;
RUN;

*Q1;
*Interaction Graph with data;
proc sgpanel data=q1;
panelby Site/columns=3;
scatter x=Species y=Stems;
loess x=Species y=Stems;
rowaxis label="Site" integer;
run;

*A more typical version of the interaction graph;
proc sgplot data=q1;
scatter x=Species y=Stems/group=Site;
loess x=Species y=Stems/group=Site nomarkers;
xaxis integer;
run;

*Flip the axes for the above in case students code
it differently;
proc sgplot data=q1;
scatter x=Site y=Stems/group=Species;
loess x=Site y=Stems/group=Species nomarkers;
xaxis integer;
run;
```

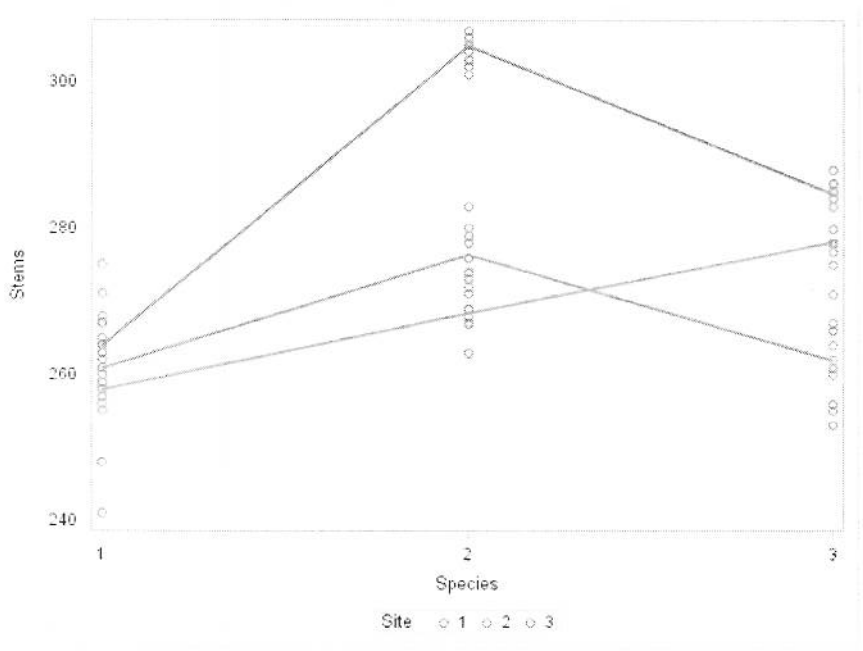
```
*Test equality of variances;
*Create single factor combining A and B;
data q1; set q1; AB=3*(Site-1)+Species; run;

ods graphics on;
proc glm data=q1;
class AB;
model Stems=AB;
means AB/hovtest=levene hovtest=BF;
run;

*Testing to handle unequal variances;
ods graphics on;
proc mixed data=q1 plots=all;
class Site Species;
model Stems=Site Species
Site*Species/ddfm=satterthw;
repeated/group=Site*Species;
lsmeans Site*Species/slice=Site adjust=simulate
adjdfe=row;
lsmeans Site*Species/slice=Species adjust=simulate
adjdfe=row;
run;
```

Q1a. I used a couple different methods for my interaction graphs—I'm not sure students would choose any of these. Their graphs should clearly demonstrate both interaction and variation within cell. In general, they should note that interaction is present and variance looks markedly different in some of the cells.

Interaction Plot Version 1

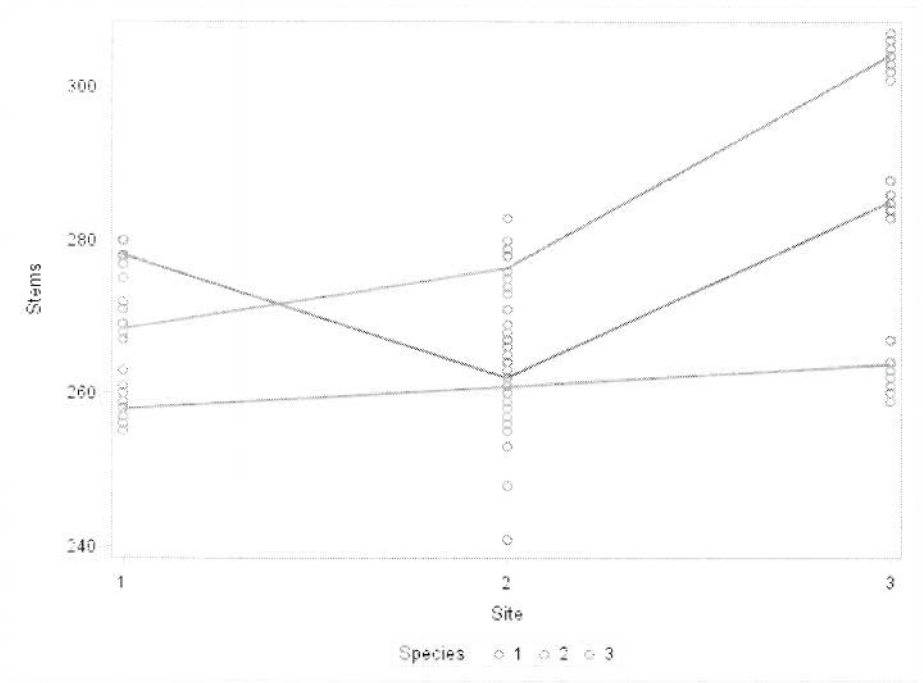


Q1a. Interaction Plot Version 2

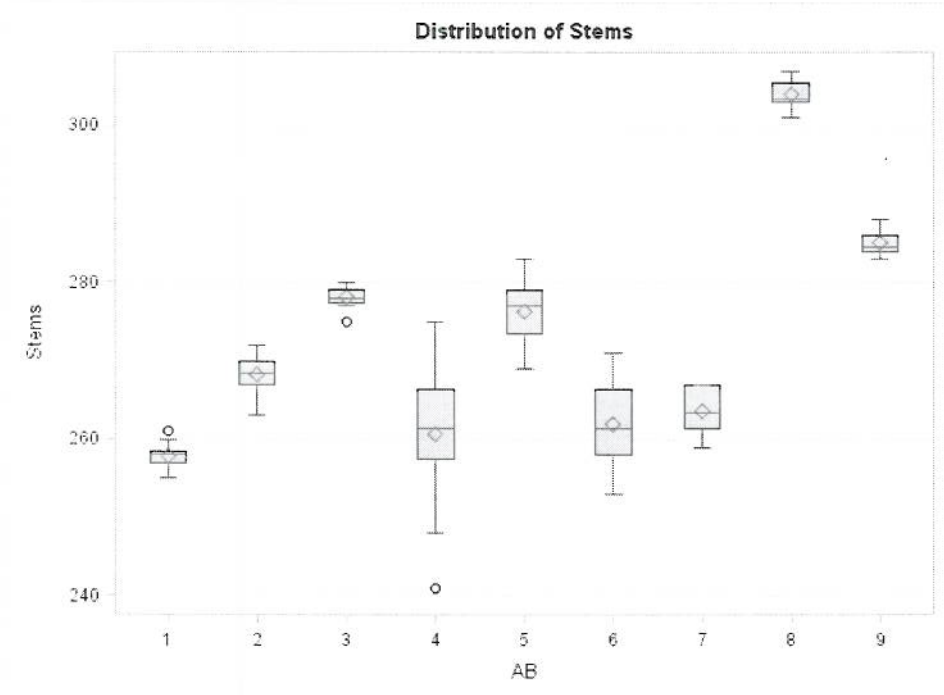


4

Q1a. Interaction Plot Version 2a (roles of Site and Species switched)



Q1b. Output from test of variances. (Could be used to help with Q1a). I needed to create a single factor named AB; students may know a slicker way of doing this.



Q1b. Tests of Equality of Variances. The variances are unequal (by design, as it turns out). I think these are both reasonable choices for variance tests. 6

Levene's Test for Homogeneity of Stems Variance
ANOVA of Squared Deviations from Group Means

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
AB	8	63169.1	7896.1	4.60	<.0001
Error	99	170097	1718.2		

Brown and Forsythe's Test for Homogeneity of Stems Variance
ANOVA of Absolute Deviations from Group Medians

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
AB	8	360.5	45.0579	6.70	<.0001
Error	99	665.4	6.7214		

Q1c&d. Students may try Weighted Least Squares, but they did learn an interesting use of REPEATED (REPEATED/GROUP=SITE*SPECIES) in PROC MIXED that should take care of the unequal variances. They need to make some attempt to handle the unequal variances.

Model Information

Data Set	WORK.Q1
Dependent Variable	Stems
Covariance Structure	Variance Components
Group Effect	Site*Species
Estimation Method	REML
Residual Variance Method	None
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Between-Within

Variance estimates:

Covariance Parameter Estimates

Cov Parm	Group		Estimate
Residual	Site*Species	1 1	2.6970

Covariance Parameter Estimates

Cov Parm	Group	Estimate
Residual	Site*Species 1 2	5.8788
Residual	Site*Species 1 3	2.0833
Residual	Site*Species 2 1	88.3864
Residual	Site*Species 2 2	14.9697
Residual	Site*Species 2 3	29.3561
Residual	Site*Species 3 1	8.2424
Residual	Site*Species 3 2	3.2727
Residual	Site*Species 3 3	2.9091

Main effects and interactions are strongly significant.

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
Site	2	37.9	589.55	<.0001
Species	2	36.4	227.15	<.0001
Site*Species	4	40.3	210.72	<.0001

To study multiple comparisons in the presence of interactions, students can use a SLICE option, or build the comparisons by hand using the ESTIMATE statement. The SLICE option can be over-rated, but there's no harm in using it. You can see that difference between Species within Sites 1 and 3, and differences between Sites within Species 2 and 3 are strongly significant.

They should adjust for multiple comparisons, though the correct choices are subtle; I selected options for ADJUST (ADJUST=SIMULATE) and ADJDFE (ADJDFE=ROW) that are among the most suitable for typical analyses with unequal variances. I chose to highlight pairwise differences that were "apples to apples"; I would hope students focus on similar pairwise differences. They may do so with CONTRAST or ESTIMATE statements, rather than through LSMEANS. You can see that most pairwise differences of interest are flagged as strongly significant, with only a few exceptions.

8

Tests of Effect Slices

Effect	Site	Species	Num DF	Den DF	F Value	Pr > F
Site*Species	1		2	25.8	515.42	<.0001
Site*Species	2		2	16.7	35.14	<.0001
Site*Species	3		2	17.9	912.53	<.0001
Site*Species		1	2	13.9	18.81	0.0001
Site*Species		2	2	26.3	912.18	<.0001
Site*Species		3	2	16.1	128.74	<.0001

Site	Species	Site	Species	Estimate	Standard Error	DF	t Value	Pr > |t|	Adj P
1	1	1	2	-10.5000	0.8454	19.3	-12.42	<.0001	<.0001
1	1	1	3	-20.2500	0.6312	21.6	-32.08	<.0001	<.0001
1	1	2	1	-2.9167	2.7550	11.7	-1.06	0.3112	0.9711
1	1	2	2	-18.5000	1.2134	14.8	-15.25	<.0001	<.0001
1	1	2	3	-4.0833	1.6343	13	-2.50	0.0267	0.2647
1	1	3	1	-5.8333	0.9548	17.5	-6.11	<.0001	0.0004
1	1	3	2	-46.1667	0.7053	21.8	-65.46	<.0001	<.0001
1	1	3	3	-27.1667	0.6835	22	-39.75	<.0001	<.0001
1	2	1	3	-9.7500	0.8146	17.9	-11.97	<.0001	<.0001
1	2	2	1	7.5833	2.8028	12.5	2.71	0.0186	0.1901
1	2	2	2	-8.0000	1.3181	18.5	-6.07	<.0001	0.0004
1	2	2	3	6.4167	1.7135	15.2	3.74	0.0019	0.0276
1	2	3	1	4.6667	1.0848	21.4	4.30	0.0003	0.0106
1	2	3	2	-35.6667	0.8733	20.3	-40.84	<.0001	<.0001
1	2	3	3	-16.6667	0.8558	19.7	-19.48	<.0001	<.0001
1	3	2	1	17.3333	2.7457	11.5	6.31	<.0001	0.0003
1	3	2	2	1.7500	1.1921	14	1.47	0.1642	0.8361
1	3	2	3	16.1667	1.6186	12.6	9.99	<.0001	<.0001
1	3	3	1	14.4167	0.9276	16.2	15.54	<.0001	<.0001

Site	Species	Site	Species	Estimate	Standard Error	DF	t Value	Pr > |t|	Adj P
1	3	3	2	-25.9167	0.6681	21	-38.79	<.0001	<.0001
1	3	3	3	-6.9167	0.6450	21.4	-10.72	<.0001	<.0001
2	1	2	2	-15.5833	2.9348	14.6	-5.31	<.0001	0.0013
2	1	2	3	-1.1667	3.1324	17.6	-0.37	0.7140	1.0000
2	1	3	1	-2.9167	2.8377	13	-1.03	0.3227	0.9764
2	1	3	2	-43.2500	2.7637	11.8	-15.65	<.0001	<.0001
2	1	3	3	-24.2500	2.7583	11.7	-8.79	<.0001	<.0001
2	2	2	3	14.4167	1.9219	19.9	7.50	<.0001	<.0001
2	2	3	1	12.6667	1.3908	20.3	9.11	<.0001	<.0001
2	2	3	2	-27.6667	1.2330	15.6	-22.44	<.0001	<.0001
2	2	3	3	-8.6667	1.2206	15.1	-7.10	<.0001	<.0001
2	3	3	1	-1.7500	1.7701	16.7	-0.99	0.3369	0.9817
2	3	3	2	-42.0833	1.6490	13.4	-25.52	<.0001	<.0001
2	3	3	3	-23.0833	1.6397	13.2	-14.08	<.0001	<.0001
3	1	3	2	-40.3333	0.9796	18.5	-41.17	<.0001	<.0001
3	1	3	3	-21.3333	0.9640	17.9	-22.13	<.0001	<.0001
3	2	3	3	19.0000	0.7177	21.9	26.47	<.0001	<.0001

#5 X_1, X_2 indep Poisson (λ). $\lambda > 0$.

$$H_0: \lambda \le 1$$
$$H_1: \lambda > 1$$

(a) Test

$$\phi_1 = \phi_1(X_1, X_2) = \begin{cases} 1, & X_1 \ge 2 \\ 0, & \text{o.w.} \end{cases}$$

The power function is

$$\begin{aligned} \beta_1(\lambda) &= E_\lambda[\phi_1(X_1, X_2)] \\ &= E_\lambda[I(X_1 \ge 2)] \\ &= P_\lambda(X_1 \ge 2) \\ &= 1 - P_\lambda(X_1 = 0) - P_\lambda(X_1 = 1) \\ &= 1 - e^{-\lambda} - \lambda e^{-\lambda}. \end{aligned}$$

Note that $\beta_1(\lambda)$ is an increasing function of λ because

$$\frac{d}{d\lambda}\beta_1(\lambda) = \lambda e^{-\lambda} > 0. \quad \forall \lambda > 0.$$

Therefore,

$$\text{Size} = \sup_{0 < \lambda \le 1} \beta_1(\lambda)$$

$$= \beta_1(1) = 1 - e^{-1} - e^{-1} \doteq 0.264.$$

(b) $\phi_2(x_1, x_2) = E[I(X_1 \ge 2) \mid X_1 + X_2 = x_1 + x_2]$

$$= P(X_1 \ge 2 \mid X_1 + X_2 = t), \qquad t = x_1 + x_2.$$

We need to find the conditional pmf of X_1, given $X_1 + X_2 = t$.

Note that $X_1 + X_2 \sim \text{Poisson}(2\lambda)$. For $x_1 = 0, 1, \ldots, t$, we have

$$f_{X_1 | X_1+X_2}(x_1 | t) = \frac{P_\lambda(X_1 = x_1, X_1 + X_2 = t)}{P_\lambda(X_1 + X_2 = t)}$$

$$= \frac{P_\lambda(X_1 = x_1, X_2 = t - x_1)}{P_\lambda(X_1 + X_2 = t)}$$

(indep) $$= \frac{P_\lambda(X_1 = x_1) P_\lambda(X_2 = t - x_1)}{P_\lambda(X_1 + X_2 = t)}$$

$$= \frac{\frac{\lambda^{x_1} e^{-\lambda}}{x_1!} \cdot \frac{\lambda^{t-x_1} e^{-\lambda}}{(t-x_1)!}}{\frac{(2\lambda)^t e^{-2\lambda}}{t!}}$$

$$= \frac{t!}{x_1!(t-x_1)!} \left(\frac{1}{2}\right)^t$$

$$= \binom{t}{x_1} \left(\frac{1}{2}\right)^{x_1} \left(\frac{1}{2}\right)^{t-x_1}.$$

Therefore $X_1 | X_1 + X_2 = t \sim b\left(t, \frac{1}{2}\right)$.

Therefore

$$P(X_1 \geq 2 | X_1 + X_2 = t) = 1 - P(X_1 = 0 | X_1 + X_2 = t) - P(X_1 = 1 | X_1 + X_2 = t)$$

$$= 1 - \binom{t}{0}\left(\frac{1}{2}\right)^t - \binom{t}{1}\left(\frac{1}{2}\right)^t$$

$(t = x_1 + x_2)$

$$= 1 - \left(\frac{1}{2}\right)^t - t\left(\frac{1}{2}\right)^t.$$

(c) The power function of the test that uses $\phi_2(X_1, X_2)$ is

$$\beta_2(\lambda) = E_\lambda [\phi_2(X_1, X_2)]$$

$$= E_\lambda \{ E[\phi_1(X_1, X_2) \mid T] \}$$

iterated expectation

$$= E_\lambda [\phi_1(X_1, X_2)]$$

$$= P_\lambda(X_1 \geq 2) = 1 - e^{-\lambda} - \lambda e^{-\lambda}$$

same as $\beta_1(\lambda)$.

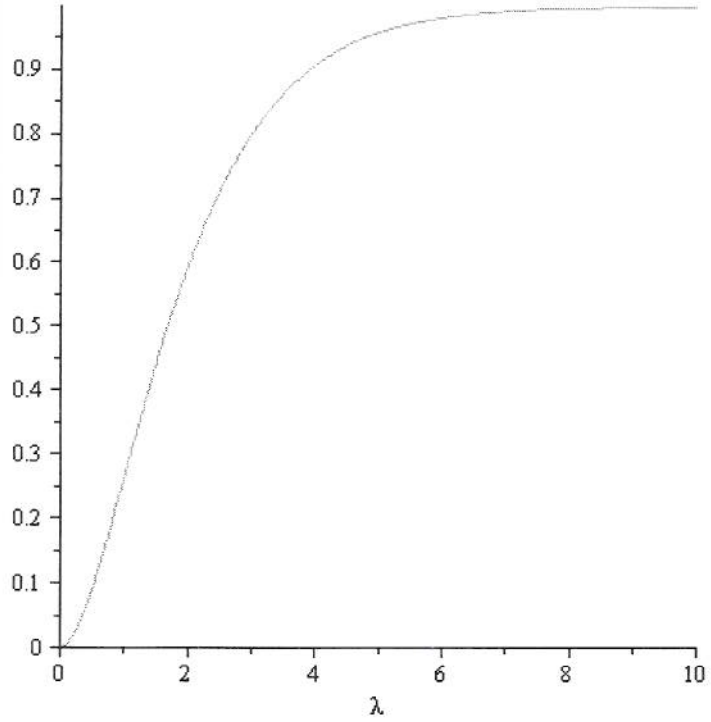
Therefore, the power functions are identical. A graph of the (common) power function is on the next page.

4

```
# Power function
> f := 1 - exp(-lambda) - lambda·exp(-lambda);
```

$$1 - e^{-\lambda} - \lambda e^{-\lambda}$$

```
plot(f, lambda = 0..10);
```



Problem 6.

a) For Model 1, the intercept and the slope are estimated as 19.356 and 6.210, respectively. Thus,

the fitted model is $\hat{y}_{ij} = 19.356 + 6.210 * x_j = 19.356 + 6.210 * j$.

The error variance is estimated as 4.392^2=19.290.

The t test for the slope (equivalently, the F test in the ANOVA table) is significant with p-value <2e-16, showing that the week number is significantly useful for predicting pig's weight.

However, from the residual plots, we can tell that the error assumptions do not hold well. The residual plot (residuals v.s. fitted weights) shows that the constant error variance across all measurements does not hold. It shows the larger the fitted weight (the larger the week number) is, the larger the error variance is. The other residual plot (residuals v.s. id) shows that residuals (and corresponding errors) are not independent. The residuals in the same pig intend to be correlated. For example, some pigs have all residuals positive and some pigs have all residuals negative.

Use R:

Codes:

```
PigWeight=read.table("pigWeight.txt",header=F, col.names=c('id','week','weight'))

attach(PigWeight)

weight.reg<-lm(weight~week)

summary(weight.reg)

anova(weight.reg)

plot(week, weight); abline(weight.reg)

par(mfrow=c(1,2))
plot(fitted(weight.reg), resid(weight.reg)); abline(h=0)
plot(id, resid(weight.reg)); abline(h=0)
```

Output

```
Call:

lm(formula = weight ~ week)

Residuals:

   Min     1Q  Median     3Q    Max

-11.9051  -2.5348  -0.1952   2.5949  13.1751

Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept) 19.35561    0.46054   42.03   <2e-16 ***

week         6.20990    0.08184   75.88   <2e-16 ***
```

Residual standard error: **4.392** on 430 degrees of freedom

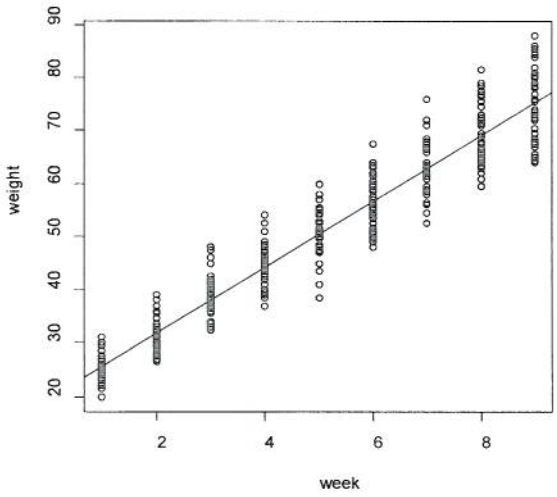
Multiple R-squared: 0.9305, Adjusted R-squared: 0.9303

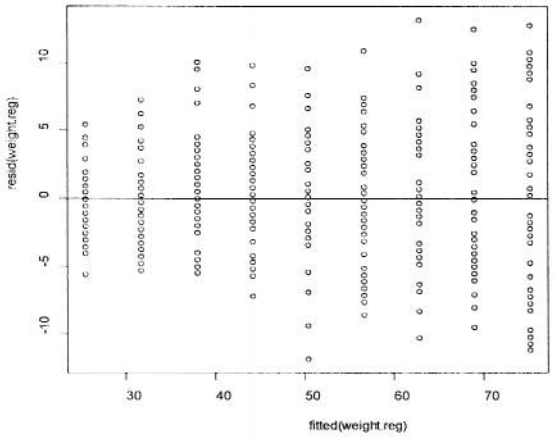
F-statistic: 5757 on 1 and 430 DF, p-value: < 2.2e-16

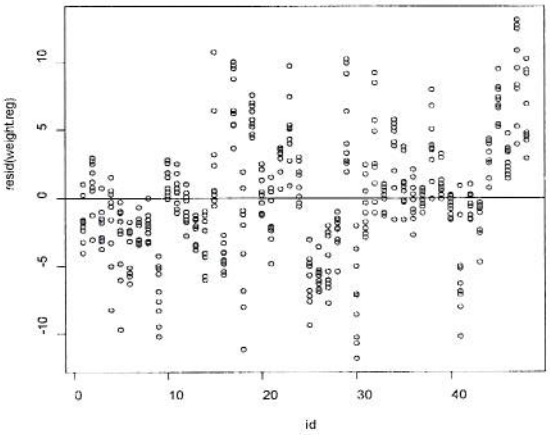
Analysis of Variance Table

Response: weight

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
week	1	111061	111061	5757.4	< 2.2e-16 ***
Residuals	430	8295	19		







3

(b)

Model 1:

$cov(y_{ij}, y_{ij'}) = cov(\varepsilon_{ij}, \varepsilon_{ij'}) = 0.$

So, $corr(y_{ij}, y_{ij'}) = 0.$

Model 2:

$cov(y_{ij}, y_{ij'}) = cov(u_i + \varepsilon_{ij}, u_i + \varepsilon_{ij'}) = var(u_i) = \sigma_u^2.$

$var(y_{ij}) = var(y_{ij'}) = var(u_i + \varepsilon_{ij}) = var(u_i) + var(\varepsilon_{ij'}) = \sigma_u^2 + \sigma_\varepsilon^2.$

So, $corr(y_{ij}, y_{ij'}) = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\varepsilon^2}.$

Model 3:

$$cov(y_{ij}, y_{ij'}) = cov(u_i + v_i x_j + \varepsilon_{ij}, u_i + v_i x_{j'} + \varepsilon_{ij'}) = var(u_i) + x_j x_{j'} var(v_j)$$
$$= \sigma_u^2 + x_j x_{j'} \sigma_v^2.$$

$$var(y_{ij}) = var(u_i + v_i x_j + \varepsilon_{ij}) = var(u_i) + x_j^2 var(v_i) + var(\varepsilon_{ij})$$
$$= \sigma_u^2 + x_j^2 \sigma_v^2 + \sigma_\varepsilon^2.$$

$$var(y_{ij'}) = \sigma_u^2 + x_{j'}^2 \sigma_v^2 + \sigma_\varepsilon^2.$$

So, $corr(y_{ij}, y_{ij'}) = \frac{\sigma_u^2 + x_j x_{j'} \sigma_v^2}{\sqrt{(\sigma_u^2 + x_j^2 \sigma_v^2 + \sigma_\varepsilon^2)(\sigma_u^2 + x_{j'}^2 \sigma_v^2 + \sigma_\varepsilon^2)}}.$

(c) All of the three models model a straight line linear relationship between mean weight and the number of week. Model 1 is a simple linear regression. It assumes that all pigs have the same straight line relationship between mean weight and the number of week. It ignores the fact of repeated measurements pertaining to the same pig. Model 2 and Model 3 are mixed models with both fixed effects and random effects. Models 2 adds random effects u_i (which are centered around 0). Random effects u_i account for possible variability in the intercept of the growth curve for the whole population of the pigs. It allows for possibly different intercepts for each pig. Besides a random intercept, Model 3 adds random effects v_i (which are centered around 0). Random effects v_i account for possible variability in the slope of the growth curve for the whole population of the pigs. It allows for possibly different slopes for each pig. With random effects in, these two models naturally incorporate the within-pig correlation.

In practice, we can check the scatterplot (weight v.s. the number of week) with lines connecting points that belong to the same pig. If we see all lines are roughly parallel with random errors, then Model 2 is suggested to be used; If we see lines with different intercepts and slopes, Model 3 is suggested to be used; If all lines are randomly scattered around, then Model 1 may be used.

(d) According to the data description, it's longitude data, where each pig has 9 measurements. Therefore, within-pig correlation should be considered. Plus, from the plot, we can tell that besides the random errors, there's some variability in the intercepts and slopes of the connecting lines. For example, pig 1 and pig 48 have obvious different intercepts. The microphone shaped residual plot (residual v.s. fitted weight) also indicates different slopes for different pigs. Therefore, Model 3 is most natural to use for analyzing these data.