Problem 1

(a) We need to show that

(i) $\lim_{x\to+\infty} F_X(x) = 1$

(ii) $\lim_{x\to-\infty} F_X(x) = 0$

(iii) $F_X(x)$ is right continuous

(iv) $F_X(x)$ is non-decreasing.

Because $F_0(x)$ is valid, it has these properties above.

(i) $$\lim_{x\to+\infty} F_X(x) = \lim_{x\to+\infty} \{(1+\delta)F_0(x) - \delta[F_0(x)]^2\}$$
$$= (1+\delta)\lim_{x\to+\infty} F_0(x) - \delta\left[\lim_{x\to+\infty} F_0(x)\right]^2$$
$$= (1+\delta)(1) - \delta(1)^2 = 1.$$

(ii) $$\lim_{x\to-\infty} F_X(x) = \lim_{x\to-\infty} \{(1+\delta)F_0(x) - \delta[F_0(x)]^2\}$$
$$= (1+\delta)\lim_{x\to-\infty} F_0(x) - \delta\left[\lim_{x\to-\infty} F_0(x)\right]^2$$
$$= (1+\delta)(0) - \delta(0)^2 = 0.$$

(iii) Suppose $a \in \mathbb{R}$. It suffices to show $\lim_{x\to a^+} F_X(x) = F_X(a)$.

$$\lim_{x\to a^+} F_X(x) = \lim_{x\to a^+} \{(1+\delta)F_0(x) - \delta[F_0(x)]^2\}$$
$$= (1+\delta)\lim_{x\to a^+} F_0(x) - \delta\left[\lim_{x\to a^+} F_0(x)\right]^2$$
$$= (1+\delta)F_0(a) - \delta[F_0(a)]^2 = F_X(a).$$

(iv) It suffices to show $x_2 \geq x_1 \Rightarrow F_X(x_2) \geq F_X(x_1)$.
Suppose $x_2 \geq x_1$. We will show $F_X(x_2) - F_X(x_1) \geq 0$ $\forall \delta \in [-1,1]$. Note that

$$F_X(x_2) - F_X(x_1) = (1+\delta)[F_0(x_2) - F_0(x_1)] - \delta\{[F_0(x_2)]^2 - [F_0(x_1)]^2\}$$

call this $*$ $\longrightarrow$ $$= [F_0(x_2) - F_0(x_1)]\{1+\delta - \delta[F_0(x_2) + F_0(x_1)]\}.$$

Note that

$$(*) = \underbrace{[F_0(x_2) - F_0(x_1)]}_{\geq 0} \underbrace{\{1 + \delta - \delta\,[F_0(x_2) + F_0(x_1)]\}}_{\geq 0 \ \forall \delta \in [-1,1]}$$

(because $F_0(x)$ is valid)

Therefore $F_X(x_2) - F_X(x_1) \geq 0$. Because x_1 and x_2 are arbitrary ($x_2 \geq x_1$), the result follows.

(b) For $x \leq 0$, clearly $f_X(x|\underset{\sim}{\theta}) = 0$. For $x > 0$, we have

CDF $\longrightarrow$ $F_X(x|\underset{\sim}{\theta}) = (1+\delta)(1 - e^{-x/\lambda}) - \delta(1 - e^{-x/\lambda})^2$

$$\Rightarrow f_X(x|\underset{\sim}{\theta}) = \frac{d}{dx} F_X(x|\underset{\sim}{\theta})$$

PDF

$$= (1+\delta)\left[0 - \left(-\tfrac{1}{\lambda}\right) e^{-x/\lambda}\right] - \delta(2)(1 - e^{-x/\lambda})\left[0 - \left(-\tfrac{1}{\lambda}\right) e^{-x/\lambda}\right]$$

$$= (1+\delta)\left(\frac{1}{\lambda}\right) e^{-x/\lambda} - 2\delta(1 - e^{-x/\lambda})\left(\frac{1}{\lambda}\right) e^{-x/\lambda}$$

$$= \frac{1}{\lambda} e^{-x/\lambda}\left[1 + \delta - 2\delta(1 - e^{-x/\lambda})\right]$$

$$= \frac{1}{\lambda} e^{-x/\lambda}\left(1 + \delta - 2\delta + 2\delta e^{-x/\lambda}\right)$$

$$= \frac{1}{\lambda} e^{-x/\lambda}\left(1 - \delta + 2\delta e^{-x/\lambda}\right).$$

Therefore,

$$f_X(x|\underset{\sim}{\theta}) = \frac{1}{\lambda} e^{-x/\lambda}\left(1 - \delta + 2\delta e^{-x/\lambda}\right) I(x > 0),$$

as claimed.

(c) We have

$$E(X^k) = \int_0^\infty x^k f_X(x \mid \underset{\sim}{\theta})\, dx$$

$$= \int_0^\infty x^k \frac{1}{\lambda} e^{-x/\lambda} (1 - \delta + 2\delta e^{-x/\lambda})\, dx$$

$$= \int_0^\infty \frac{(1-\delta)}{\lambda} x^k e^{-x/\lambda}\, dx + \int_0^\infty \frac{2\delta}{\lambda} x^k e^{-2x/\lambda}\, dx$$

$$= I_1 + I_2 .$$

$$I_1 = \frac{(1-\delta)}{\lambda}\left[\Gamma(k+1)\, \lambda^{k+1} \right]$$

because $x^k e^{-x/\lambda} = x^{(k+1)-1} e^{-x/\lambda}$
= gamma kernel w/
shape = $k+1$
scale = λ

$$I_2 = \frac{2\delta}{\lambda}\left[\Gamma(k+1) \left(\frac{\lambda}{2}\right)^{k+1} \right]$$

because $x^k e^{-2x/\lambda} = x^{(k+1)-1} e^{-x/(\frac{\lambda}{2})}$
= gamma kernel w/
shape = $k+1$
scale = $\lambda/2$

Therefore

$$I_1 + I_2 = \frac{(1-\delta)}{\lambda}\left[\Gamma(k+1)\, \lambda^{k+1} \right] + \frac{2\delta}{\lambda}\left[\Gamma(k+1) \left(\frac{\lambda}{2}\right)^{k+1} \right]$$

$$= \lambda^k \Gamma(k+1)\left[(1-\delta) + 2\delta\left(\frac{1}{2}\right)^{k+1}\right]$$

$$= \lambda^k \Gamma(k+1)\left(1-\delta+\delta 2^{-k}\right),$$

as claimed.

(d) To find the MOM, we set up 2 equations:

$$(1) \qquad \bar{X} \overset{set}{=} E_{\underset{\sim}{\theta}}(X)$$

$$(2) \qquad \frac{1}{n}\sum_{i=1}^{n} X_i^2 \overset{set}{=} E_{\underset{\sim}{\theta}}(X^2)$$

Note that

$$E_{\underset{\sim}{\theta}}(X) = \lambda^1 \Gamma(1+1)\left(1-\delta+\delta 2^{-1}\right) = \lambda\left(1-\frac{\delta}{2}\right)$$

$$E_{\underset{\sim}{\theta}}(X^2) = \lambda^2 \Gamma(2+1)\left(1-\delta+\delta 2^{-2}\right) = 2\lambda^2\left(1-\frac{3\delta}{4}\right)$$

Therefore, set

$$\bar{X} \overset{set}{=} \lambda\left(1-\frac{\delta}{2}\right)$$

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \overset{set}{=} 2\lambda^2\left(1-\frac{3\delta}{4}\right)$$

and solve for $\underset{\sim}{\theta} = (\delta, \lambda)'$. Report the solution that falls in the parameter space (if any).

Problem 2

KYPHOSIS DATA

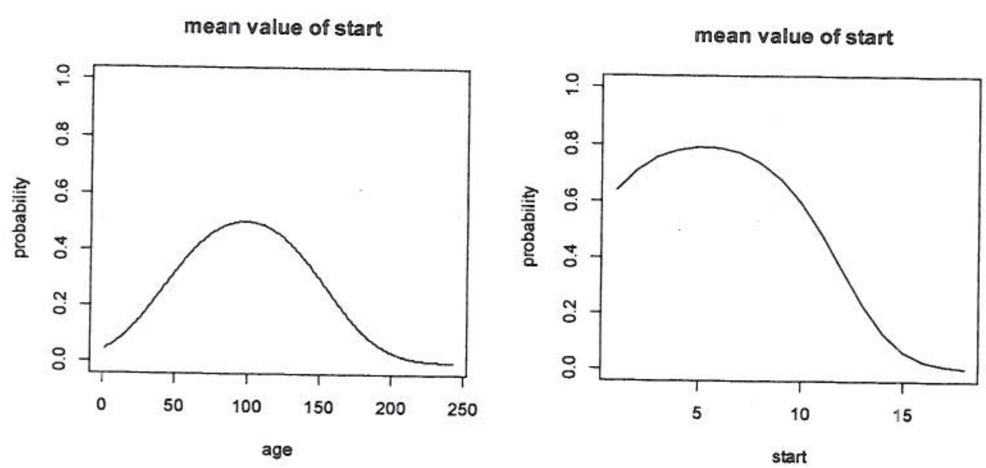
The response is Bernoulli, so logistic regression is the most obvious choice of model. If one includes quadratic and interaction terms and performs backwards elimination with a 0.05 cutoff, age, age^2, start, and start^2 are retained in the model; only the interaction is removed. The final model is:

Analysis of Maximum Likelihood Estimates

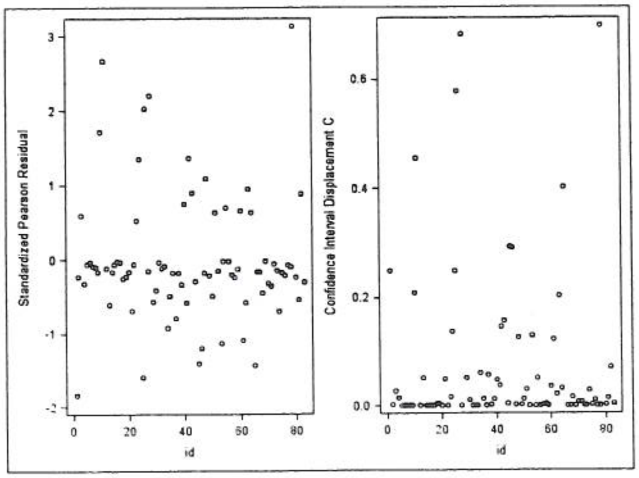
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.8692	1.4041	4.1755	0.0410
age	1	0.0630	0.0275	5.2576	0.0219
start	1	0.4425	0.3026	2.1387	0.1436
age*age	1	-0.00032	0.000161	3.9414	0.0471
start*start	1	-0.0421	0.0190	4.9009	0.0268

Since quadratic terms are in the model, we cannot interpret the output simply in terms of odds ratios; elimination of the interaction does allow for additive interpretation though. Both quadratic terms are negative, so probability increases, then decreases with age, ditto for start. The apex, i.e. greatest risk, occurs at 0.063-0.00032*2*age=0, or about age=98 months holding start constant. Similarly, greatest risk occurs at about start=5 holding age constant; this latter observation was not asked for, but may be of interest.

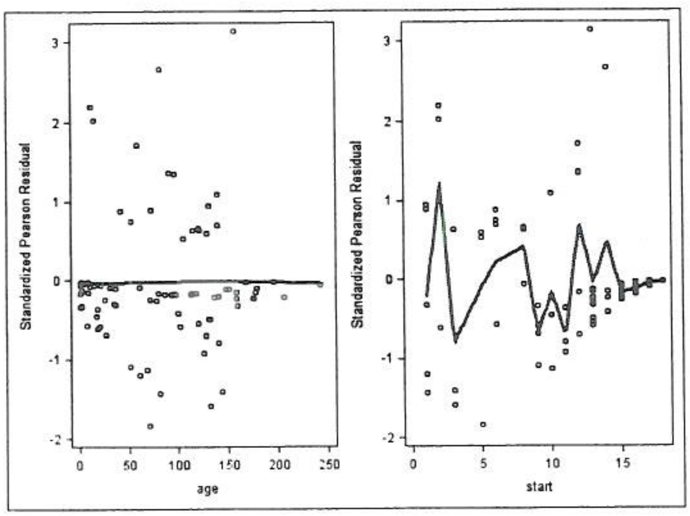
The mean age is about 85 months (range is 1 to 243), and the mean start is about 11 (range is 1 to 18). It is helpful (but not strictly necessary) to examine at the fitted probability of kyphosis for age over the observed range with start fixed at 11, and for start with age fixed at 85:



Studentized Pearson residuals and Cook's distance vs. index potentially shows one patient that is both influential and ill-fit (upper right). This is case 79, a child that had kyphosis at age=157 and start=13; the fitted probability of kyphosis for this child is 0.1.



The diagnostic plots show reasonable fit, although the default bandwidth for the loess smooth on start undersmooths the estimated mean:



Note that, if instead one INITIALLY fits the simpler model with age and start entered linearly into predictor, smoothed residual plots show lack-of-fit indicating a more flexible mean is required (like quadratic terms found necessary using backwards elimination).

SAS code:

```
proc logistic data=kyphosis plots=all;
model kyphosis(event='1')=age start age*age start*start age*start /
selection=backward;
run;

proc logistic data=kyphosis plots=all;
model kyphosis(event='1')=age start age*age start*start;
output out=out stdreschi=r c=c;
run;
```

```
proc sgscatter data=out;
plot (r c)*id;
run;

proc sgscatter data=out;
plot r*(age start) / loess;
run;

proc means data=kyphosis; run;

proc print data=out(where=(c>0.6 or r>3 or r<-3));
var kyphosis age start c r;
run;
```

R code:

```
f=function(a,s){exp(-2.8692+0.063*a+0.4425*s-0.00032*a^2-0.0421*s^2)/(1+exp(-
2.8692+0.063*a+0.4425*s-0.00032*a^2-0.0421*s^2))}
a=seq(1,243,1); s=seq(1,18,1); pa=a; ps=s
for(i in 1:243){pa[i]=f(a[i],11)}
for(i in 1:18){ps[i]=f(85,s[i])}
plot(a,pa,type="l",ylim=c(0,1),xlab="age",ylab="probability",main="mean value of
start")
plot(s,ps,type="l",ylim=c(0,1),xlab="start",ylab="probability",main="mean value of
start")
f(157,13)
```

Problem 3

(a) Use mgfs (easiest). The mgf of $T = X_1 + X_2$ is

$$M_T(t) = M_{X_1}(t)\, M_{X_2}(t) \qquad X_1 \perp X_2$$

$$= e^{\theta_1(e^t - 1)}\, e^{\theta_2(e^t - 1)}$$

$$= e^{(\theta_1+\theta_2)(e^t - 1)},$$

which is recognised as the mgf of a Poisson distribution with mean $\theta_1 + \theta_2$. Because mgfs are unique,

$$T = X_1 + X_2 \sim \text{Poisson}(\theta_1 + \theta_2).$$

(b) We want to calculate $f_{X_1|T}(x_1|t)$ for all $x_1 \in \mathbb{Z}$. Clearly $x_1 = 0, 1, 2, \ldots, t$ with positive probability; that is

$$f_{X_1|T}(x_1|t) \text{ is nonzero}$$

only if $x_1 = 0, 1, 2, \ldots, t$; $f_{X_1|T}(x_1|t) = 0$ o.w.

Now,

$$f_{X_1|T}(x_1|t) = \frac{f_{X_1,T}(x_1, t)}{f_T(t)} \qquad \begin{matrix}\text{joint}\\ \text{marg}\end{matrix}$$

$$= \frac{P(X_1 = x_1, T = t)}{P(T = t)}$$

$$= \frac{P(X_1 = x_1, X_2 = t - x_1)}{P(T = t)}$$

$$X_1 \perp X_2 \longrightarrow = \frac{P(X_1 = x_1)\, P(X_2 = t - x_1)}{P(T = t)} \qquad (*)$$

$X_1 \sim \text{Poisson}(\theta_1)$, $X_2 \sim \text{Poisson}(\theta_2)$, $T \sim \text{Poisson}(\theta_1 + \theta_2)$

Therefore,

$$\circledast = \frac{\dfrac{\theta_1^{x_1} \cancel{e^{-\theta_1}}}{x_1!} \dfrac{\theta_2^{t-x_1} \cancel{e^{-\theta_2}}}{(t-x_1)!}}{\dfrac{(\theta_1+\theta_2)^t \cancel{e^{-(\theta_1+\theta_2)}}}{t!}}$$

$x_1 = 0, 1, 2, \ldots, t$

$$= \frac{t!}{x_1!(t-x_1)!} \frac{\theta_1^{x_1} \theta_2^{t-x_1}}{(\theta_1+\theta_2)^t}$$

$$= \binom{t}{x_1} \frac{\theta_1^{x_1}}{(\theta_1+\theta_2)^{x_1}} \frac{\theta_2^{t-x_1}}{(\theta_1+\theta_2)^{t-x_1}}$$

$$= \binom{t}{x_1} \left(\frac{\theta_1}{\theta_1+\theta_2}\right)^{x_1} \left(\frac{\theta_2}{\theta_1+\theta_2}\right)^{t-x_1}$$

Note that

$$\frac{\theta_2}{\theta_1+\theta_2} = 1 - \frac{\theta_1}{\theta_1+\theta_2}$$

Therefore, X_1, conditional on $T = t$, is <u>binomial</u> with parameters t and $p = \dfrac{\theta_1}{\theta_1+\theta_2}$.

(c) Recall that

$$\lambda = \lambda(x_1, x_2) = \frac{\sup_{\underset{\sim}{\theta} \in \Theta_0} L(\underset{\sim}{\theta} \mid \underset{\sim}{x})}{\sup_{\underset{\sim}{\theta} \in \Theta} L(\underset{\sim}{\theta} \mid \underset{\sim}{x})}.$$

Here, $\underset{\sim}{\theta} = (\theta_1, \theta_2)'$, $\underset{\sim}{x} = (x_1, x_2)'$

$$\Theta_0 = \{(\theta_1, \theta_2)' : 0 < \theta_1 < \theta_2 < \infty\}$$

maximise L over Θ_0 and Θ.

$$\Theta = \{(\theta_1, \theta_2)' : 0 < \theta_1 < \infty,\ 0 < \theta_2 < \infty\}$$

First, the likelihood function.

$$\begin{aligned} L(\underset{\sim}{\theta} \mid \underset{\sim}{x}) &= L(\theta_1, \theta_2 \mid x_1, x_2) \\ &= f_{X_1}(x_1 \mid \theta_1)\, f_{X_2}(x_2 \mid \theta_2) \qquad \text{(indep)} \\ &= \frac{\theta_1^{x_1} e^{-\theta_1}}{x_1!} \; \frac{\theta_2^{x_2} e^{-\theta_2}}{x_2!} \end{aligned}$$

$$\ln L(\underset{\sim}{\theta} \mid \underset{\sim}{x}) = x_1 \ln \theta_1 - \theta_1 - c_1 + x_2 \ln \theta_2 - \theta_2 - c_2,$$

where $c_1 = \ln x_1!$ and $c_2 = \ln x_2!$, both free of $\underset{\sim}{\theta}$.

Now

$$\frac{\partial \ln L(\underset{\sim}{\theta} \mid \underset{\sim}{x})}{\partial \theta_1} = \frac{x_1}{\theta_1} - 1 \overset{\text{set}}{=} 0$$

$$\frac{\partial \ln L(\underset{\sim}{\theta} \mid \underset{\sim}{x})}{\partial \theta_2} = \frac{x_2}{\theta_2} - 1 \overset{\text{set}}{=} 0$$

The (unique) solution is

$$\hat{\underset{\sim}{\theta}} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

This solution is a maximiser because

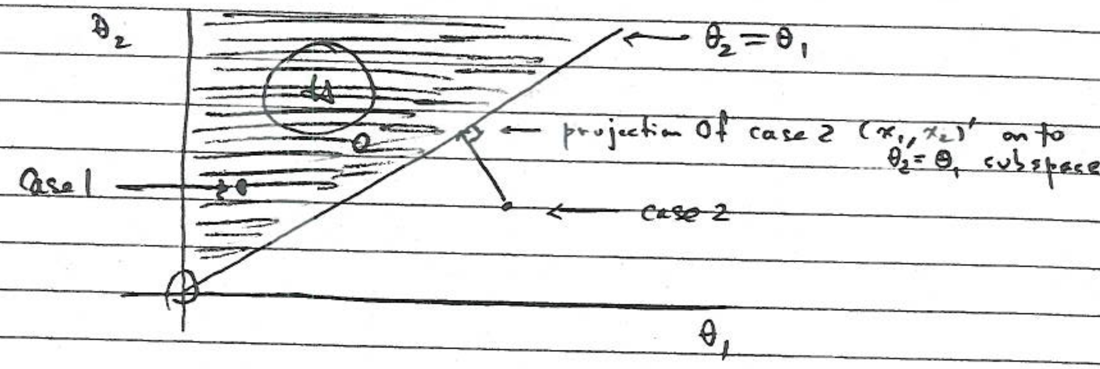
$$\underset{\sim}{a}' \begin{bmatrix} -\frac{x_1}{\theta_1^2} & 0 \\ 0 & -\frac{x_2}{\theta_2^2} \end{bmatrix} \underset{\sim}{a} \qquad \underset{\sim}{a} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$$

$$= -\frac{a_1^2 x_1}{\theta_1^2} - \frac{a_2^2 x_2}{\theta_2^2} < 0$$

$\forall \underset{\sim}{a} \in \mathbb{R}^2 - \{\underset{\sim}{0}\}$. That is, the Hessian is negative definite.

Remark: If $x_1 = 0$, then (because $e^{-\theta_1}$ is a decreasing function of θ_1 in L), the MLE is unchanged. Similarly if $x_2 = 0$ or if $x_1 = x_2 = 0$.

Now, to maximise L over Θ_0, we use a geometric argument. The parameter space Θ_0 is depicted below:



Case 1: If $x_1 \le x_2$, then the restricted estimate $\hat{\theta}_0$ matches the unrestricted one; i.e.,

$$\hat{\underset{\sim}{\theta}}_0 = \arg \sup_{\underset{\sim}{\theta} \in \Theta_0} L(\underset{\sim}{\theta} \mid \underset{\sim}{x})$$

$$= \arg \sup_{\underset{\sim}{\theta} \in \Theta} L(\underset{\sim}{\theta} \mid \underset{\sim}{x}) = \hat{\underset{\sim}{\theta}} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}.$$

(Clearly, in this case, $\lambda = 1$).

Case 2: If $x_1 > x_2$, then the unrestricted estimate $(x_1, x_2)'$ falls outside Θ_0. We therefore project $(x_1, x_2)'$ onto Θ_0; specifically, we choose the point in Θ_0 that is "closest" to (x_1, x_2); this point is

$$\hat{\underset{\sim}{\theta}}_0 = \begin{pmatrix} (x_1 + x_2)/2 \\ (x_1 + x_2)/2 \end{pmatrix}.$$

The restricted MLE is $\begin{pmatrix} (X_1 + X_2)/2 \\ (X_1 + X_2)/2 \end{pmatrix}$.

Therefore, if $x_1 \le x_2$ (Case 1), then

$$\lambda = \frac{L(x_1, x_2 \mid x_1, x_2)}{L(x_1, x_2 \mid x_1, x_2)} = 1.$$

If $x_1 > x_2$ (Case 2), then

$$\lambda = \frac{L\left(\frac{x_1+x_2}{2}, \frac{x_1+x_2}{2} \mid x_1, x_2\right)}{L(x_1, x_2 \mid x_1, x_2)}$$

$$= \frac{\dfrac{\left(\frac{x_1+x_2}{2}\right)^{x_1} \cancel{e^{-\left(\frac{x_1+x_2}{2}\right)}}}{\cancel{x_1!}} \; \dfrac{\left(\frac{x_1+x_2}{2}\right)^{x_2} \cancel{e^{-\left(\frac{x_1+x_2}{2}\right)}}}{\cancel{x_2!}}}{\dfrac{x_1^{x_1} \cancel{e^{-x_1}}}{\cancel{x_1!}} \; \dfrac{x_2^{x_2} \cancel{e^{-x_2}}}{\cancel{x_2!}}}$$

$$= \left(\frac{x_1+x_2}{2x_1}\right)^{x_1} \left(\frac{x_1+x_2}{2x_2}\right)^{x_2}$$

(d) Suppose $\theta_1 = 1$ and $\theta_2 = 2$. Then

$$X_3 \sim \text{Poisson}(3)$$
$$X_4 \sim \text{Poisson}(2)$$

$$f_{X_3,X_4}(x_3, x_4) = f_{X_3}(x_3)\, f_{X_4}(x_4)$$

$x_3 = 0, 1, 2, \dots$ $\longrightarrow$ $= \dfrac{3^{x_3} e^{-3}}{x_3!} \; \dfrac{2^{x_4} e^{-2}}{x_4!}$

$x_4 = 0, 1, 2, \dots$

If $\theta_1 = 2$ and $\theta_2 = 1$, then we get the same joint distribution. That is,

$$f_{X_3,X_4}(x_3, x_4 \mid \underset{\sim}{\theta}) = f_{X_3 X_4}(x_3, x_4 \mid \underset{\sim}{\theta}') \not\Rightarrow \underset{\sim}{\theta} = \underset{\sim}{\theta}'$$

$\Rightarrow$ θ_1 and θ_2 are not identifiable.

(c) Some members of the Qualifying Exam Committee believe that $E(U)$ and $\mathrm{var}(U)$ can be found in closed form (under the assumption that $\theta_1 = \theta_2 = 1$). Another perfectly acceptable solution is to use simulation.

1. Simulate

$$\left.\begin{array}{l} X_{11} \sim \text{Poisson}(1) \\ X_{12} \sim \text{Poisson}(1) \\ X_{13} \sim \text{Poisson}(2) \\ X_{14} \sim \text{Poisson}(1) \end{array}\right\} \text{independent.}$$

and calculate

$$U_1 = \frac{X_{11} + X_{12}}{1 + X_{13} + X_{14}}$$

2. Repeat Step 1 for $X_{b1}, X_{b2}, X_{b3}, X_{b4}$ and calculate U_b, $b = 2, 3, 4, \ldots, B$ (large).

3. After Step 2, we will have

$$U_1, U_2, \ldots, U_B \quad \text{iid} \quad f_U(u), \text{ say.}$$

Because

$$\overline{U} = \frac{1}{B}\sum_{b=1}^{B} U_b \xrightarrow{p} E(U), \quad \text{as } B \longrightarrow \infty \qquad \text{(WLLN)}$$

$$S_U^2 = \frac{1}{B-1}\sum_{b=1}^{B} (U_b - \overline{U})^2 \xrightarrow{p} \mathrm{var}(U), \quad \text{as } B \longrightarrow \infty \qquad \text{(WLLN)}$$

We can use $\overline{U}$ and S_U^2 as approximations of the true values $E(U)$ and $\mathrm{var}(U)$.

Simulations on the next page. $B = 1{,}000{,}000$ was used. The following approximations were obtained

$$E(U) \approx 0.634$$
$$\mathrm{var}(U) \approx 0.424.$$

7

```
# Monte Carlo Simulation
# Approximate E(U) and var(U)

# Number of MC samples (i.e., number of simulated values of U)
B = 1000000

# Sampling from Poisson distributions
MC.data = matrix(0, nrow = B, ncol = 4)
      MC.data[,1] <- rpois(B,1)
      MC.data[,2] <- rpois(B,1)
      MC.data[,3] <- rpois(B,2)
      MC.data[,4] <- rpois(B,1)

# Initialising step
U = MC.data[,1]*0
for (i in 1:B){
      U[i] <- (MC.data[i,1] + MC.data[i,2])/(1+MC.data[i,3] +
      MC.data[i,4])
      }

# Estimate of E(U)
mean(U)
# Estimate of var(U)
var(U)


Output:
> mean(U)
[1] 0.634259
> # Estimate of var(U)
> var(U)
[1] 0.4236196
```

Problem 4

(a) By Central Limit Theorem,

$$\sqrt{n}(\bar{X}_n - p) \xrightarrow{d} N(0, \ pq),$$

or, equivalently,

$$\frac{n\bar{X}_n - np}{\sqrt{npq}} \xrightarrow{d} N(0, 1),$$

i.e.,

$$\frac{S_n - np}{\sqrt{npq}} \xrightarrow{d} N(0, 1).$$

#

(b)

(i) Note that the event $\{\tilde{Y}_n \leq y\}$ is equivalent to $\{$no more than $m-1$ observations (among $Y_1, \cdots, Y_n$) exceed $y\}$. Define $X_i = I(Y_i > y)$, then $X_i \sim \text{Bernoulli}(r)$, where $r = 1 - G(y)$, for $i=1, \cdots$. Moreover, let $S_n = \sum_{i=1}^{n} X_i$, then

$$P(\tilde{Y}_n \leq y) = P(S_n \leq m-1),$$

where $S_n \sim \text{Binomial}(n, r)$.

(ii) From part (i), one has

$$P\{\sqrt{n}(\tilde{Y}_n - \theta) \leq w\}$$
$$= P(\tilde{Y}_n \leq w/\sqrt{n} + \theta)$$
$$= P(S_n \leq m-1),$$

where $S_n \sim \text{Binomial}(n, p_n)$, with

$$p_n = 1 - G\left(\frac{w}{\sqrt{n}} + \theta\right) = 1 - F\left(\frac{w}{\sqrt{n}}\right).$$

Note that $\lim_{n\to\infty} p_n = 1 - F(0) = 0.5$.

Now, envoking (1) above, one has, let $q_n = 1 - p_n$,

$$\lim_{n\to\infty} P\{\sqrt{n}(Y_n - \theta) \le w\}$$

$$= \lim_{n\to\infty} P(S_n \le m-1) \quad \text{, recall that } n = 2m-1,$$

$$= \lim_{n\to\infty} P\left(\frac{S_n - np_n}{\sqrt{np_n q_n}} \le \frac{\frac{n-1}{2} - np_n}{\sqrt{np_n q_n}}\right)$$

$$= \lim_{n\to\infty} \Phi\left(\frac{\frac{n-1}{2} - np_n}{\sqrt{np_n q_n}}\right),$$

where $\Phi(\cdot)$ is the cdf of $N(0,1)$.

Looking into the argument of $\Phi(\cdot)$ above, one has

$$\lim_{n\to\infty} \frac{\frac{n-1}{2} - np_n}{\sqrt{np_n q_n}}$$

$$= \lim_{n\to\infty} \frac{2\sqrt{n}\{F(\frac{w}{\sqrt{n}}) - \frac{1}{2}\} - \frac{1}{\sqrt{n}}}{2\sqrt{\{1-F(\frac{w}{\sqrt{n}})\}F(\frac{w}{\sqrt{n}})}}$$

$$= \lim_{n\to\infty} 2\sqrt{n}\left\{F\left(\frac{w}{\sqrt{n}}\right) - \frac{1}{2}\right\}$$

$$= \lim_{n\to\infty} \frac{2F(\frac{w}{\sqrt{n}}) - 1}{\frac{1}{\sqrt{n}}}$$

$$= \lim_{n\to\infty} \frac{2f(\frac{w}{\sqrt{n}})\left(-\frac{1}{2}\frac{w}{n^{3/2}}\right)}{-\frac{1}{2}\cdot\frac{1}{n^{3/2}}}$$

$$= 2f(0)w.$$

Hence,

$$\lim_{n\to\infty} P\{\sqrt{n}(\tilde{Y}_n-\theta)\le w\}=\Phi\{2f(0)w\},$$

that is,

$$\sqrt{n}(\tilde{Y}_n-\theta)\xrightarrow{d} N\left(0,\ \frac{1}{4f^2(0)}\right).$$

#

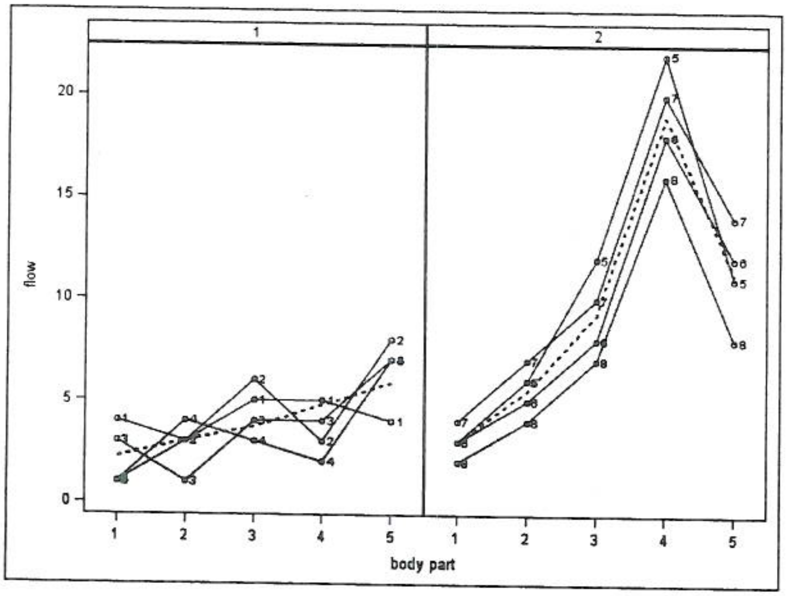
(c) One may use result in (2) to construct an asymptotically valid C.I. for θ. One just needs to find a way to estimate $f(0)$ consistently.

(We want to see what you can come up with.)

Problem 5

RAT EXERCISE BLOOD FLOW DATA

A picture is worth a thousand words:



Part (a): There is some variability in blood flow across body parts for the sedentary rats, but a great deal more variability for the exercising rats. There seems to be little difference between exercising and not exercising for body part 1 (bone), a small difference for 2 (brain), but a lot of difference for parts 3, 4, and 5 (skin, muscle, heart), especially 4. These superimposed averages (red dotted line) have different shapes across exercise/sedentary, so there is a body part by treatment interaction here.

Part (b): Based on the SAS output the correlation between repeated measurements is estimated to be 1.0833/(1.0833+1.95)=0.36.

Cov Parm	Estimate	Error
rat	1.0833	0.8580
Residual	1.9500	0.5629

Part (c) Here are the Type III tests

Effect	Num DF	Den DF	F Value	Pr > F
bodypart	4	24	49.94	<.0001
exercise	1	24	44.10	<.0001
bodypart*exercise	4	24	33.60	<.0001

The body part by interaction is significant. Differences in blood flow from sedentary to exercise differ across the body parts.

Part (d) SAS provides this test; we reject H0: sigma_rho=0 at the 5% level. Blocking on rat is beneficial.

Tests of Covariance Parameters
Based on the Restricted Likelihood

Label	DF	-2 Res Log Like	ChiSq	Pr > ChiSq	Note
No G-side effects	1	132.29	5.28	0.0108	MI

MI: P-value based on a mixture of chi-squares.

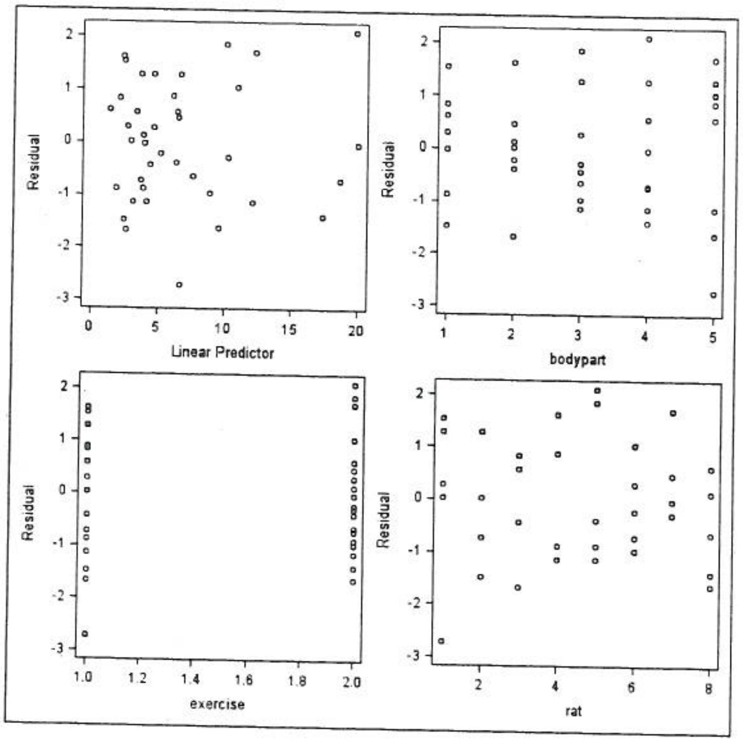
Part (e) SAS handles this request. Tukey is the most efficient (smallest intervals) from among Tukey, Scheffe, and Bonferroni.

Least Squares Means Estimates
Adjustment for Multiplicity: T

Effect	Label	Adj Lower	Adj Upper
bodypart*exercise	exercise vs. sedentary, bone	-1.7918	3.2918
bodypart*exercise	exercise vs. sedentary, brain	0.2082	5.2918
bodypart*exercise	exercise vs. sedentary, skin	2.2082	7.2918
bodypart*exercise	exercise vs. sedentary, muscle	12.9582	18.0418
bodypart*exercise	exercise vs. sedentary, heart	2.2082	7.2918

Echoing the profile plot, exercise significantly increases bloodflow in all regions except for bone. The greatest increase is seen in muscle where we are 95% confident that mean bloodflow increases between 13 and 18 units. All comparisons are carried out with a FER<0.05.

Part (f): There are no obvious patterns here. There does seem to be some slight differences in variability across bodypart.



Part (g): All tests for normality do not reject that the BLUPs are normal:

```
                        Tests for Normality

Test                  --Statistic---      -----p Value------
Shapiro-Wilk          W     0.926009      Pr < W      0.4805
Kolmogorov-Smirnov    D     0.203185      Pr > D     >0.1500
Cramer-von Mises      W-Sq  0.047203      Pr > W-Sq  >0.2500
Anderson-Darling      A-Sq  0.319327      Pr > A-Sq  >0.2500
```

SAS code:

```
proc sgpanel data=blood noautolegend;
 panelby exercise / rows=1 columns=2 novarname;
 series x=bodypart y=flow / group=rat lineattrs=(pattern=1 color=black
thickness=0.5);
 scatter x=bodypart y=flow / markerchar=rat;
 loess x=bodypart y=flow / lineattrs=(pattern=2 color=red);
 colaxis label="body part";
 rowaxis label="flow";
 run;

proc glimmix;
 class bodypart rat exercise;
 model flow=bodypart|exercise;
 random rat;
 lsmestimate bodypart*exercise
  "exercise vs. sedentary, bone" -1 1 0 0 0 0 0 0 0 0,
  "exercise vs. sedentary, brain" 0 0 -1 1 0 0 0 0 0 0,
  "exercise vs. sedentary, skin" 0 0 0 0 -1 1 0 0 0 0,
  "exercise vs. sedentary, muscle" 0 0 0 0 0 0 -1 1 0 0,
```

```
  "exercise vs. sedentary, heart" 0 0 0 0 0 0 0 0 -1 1 / adjust=t cl;
  output out=out residual=r pred=p;
 run;

proc sgscatter data=out;
 plot r*(p bodypart exercise rat);
run;

ods listing close; ods output SolutionR=rand;
proc glimmix;
 class bodypart rat exercise;
 model flow=bodypart|exercise;
 random rat /s;
 run;
ods output close; ods listing;

proc print data=rand; run;

proc univariate data=rand normal; var estimate;
run;
```

Problem 6

(a) Note that $f_X(x|\theta)$ can be written in exponential family form:

$$f_X(x|\theta) = 2x\,I(x>0)\,\frac{1}{\theta}\,e^{-\frac{1}{\theta}x^2}$$

$$= h(x)\,c(\theta)\exp[w_1(\theta)\,t_1(x)],$$

where $h(x) = 2x\,I(x>0)$

$c(\theta) = 1/\theta$

$w_1(\theta) = -1/\theta$

$t_1(x) = x^2$

We know $T = \sum_{i=1}^{n} X_i^2$ is a sufficient statistic (from exponential family theory).

Because this family is full rank (i.e., $d = k = 1$)

$d = \dim(\theta)$

$k = \#\ t_i$ statistics

we also know T is complete.

(b) Goal: Find function of T, say $\phi(T)$, so that

$$E_\theta(\phi(T)) = \tau(\theta) = \theta^2.$$

We know that the UMVUE estimator of $\tau(\theta)$
- is a function of T
- is an unbiased estimator of $\tau(\theta)$.

Start by calculating $E_\theta(T)$.

$$E_\theta(T) = E_\theta\left(\sum_{i=1}^{n} X_i^2\right) = \sum_{i=1}^{n} E_\theta(X_i^2)$$

$$= n\,E_\theta(X^2).$$

(a) Note that $f_X(x|\theta)$ can be written in exponential family form:

$$f_X(x|\theta) = 2x I(x>0) \frac{1}{\theta} e^{-\frac{1}{\theta}x^2}$$

$$= h(x)\, c(\theta) \exp[w_1(\theta) t_1(x)],$$

where $h(x) = 2x I(x>0)$
$c(\theta) = 1/\theta$
$w_1(\theta) = -1/\theta$
$t_1(x) = x^2$

We know $T = \sum_{i=1}^{n} X_i^2$ is a sufficient statistic (from exponential family theory).

Because this family is full rank (i.e., $d = k = 1$)

$d = \dim(\Theta)$
$k = \#\ t_i$ statistics

we also know T is complete.

(b) Goal: Find function of T, say $\phi(T)$, so that

$$E_\theta(\phi(T)) = \tau(\theta) = \theta^2.$$

We know that the UMVUE estimator of $\tau(\theta)$
- is a function of T
- is an unbiased estimator of $\tau(\theta)$.

Start by calculating $E_\theta(T)$.

$$E_\theta(T) = E_\theta\left(\sum_{i=1}^{n} X_i^2\right) = \sum_{i=1}^{n} E_\theta(X_i^2)$$

$$= n E_\theta(X^2).$$

Now,

$$E_\theta(X^2) = \int_0^\infty x^2 \, \frac{2x}{\theta} \, e^{-x^2/\theta} \, dx$$

$u = x^2$

$du = 2x\,dx$

$$= \int_0^\infty u \, \frac{2x}{\theta} \, e^{-u/\theta} \, \frac{du}{2x}$$

$$= \int_0^\infty u \, \frac{1}{\theta} \, e^{-u/\theta} \, du$$

$= E(U)$, where $U \sim$ exponential (θ)

Therefore $E_\theta(T) = n E_\theta(X^2) = n\theta.$

$$\Rightarrow E_\theta\left(\frac{T}{n}\right) = \theta$$

Candidate estimator: $\left(\frac{T}{n}\right)^2$.

$$E_\theta\left[\left(\frac{T}{n}\right)^2\right] = \frac{1}{n^2} E_\theta(T^2)$$

$$= \frac{1}{n^2}\left[\operatorname{var}_\theta(T) + [E_\theta(T)]^2\right]$$

$$= \frac{1}{n^2}\left[\operatorname{var}_\theta(T) + (n\theta)^2\right]$$

Find $\operatorname{var}_\theta(T)$.

$$\operatorname{var}_\theta(T) = \operatorname{var}_\theta\left(\sum_{i=1}^n X_i^2\right) = \sum_{i=1}^n \operatorname{var}_\theta(X_i^2)$$

$$= n \operatorname{var}_\theta(X^2)$$

Now,

$$\operatorname{var}_\theta(X^2) = E_\theta(X^4) - [E_\theta(X^2)]^2$$

$$= E_\theta(X^4) - \theta^2$$

Find $E_\theta(X^4)$.

$$E_\theta(X^4) = \int_0^\infty x^4 \frac{2x}{\theta} e^{-x^2/\theta}\, dx$$

$u = x^2$
$du = 2x\,dx$

$$= \int_0^\infty u^2 \frac{2x}{\theta} e^{-u/\theta} \frac{du}{2x}$$

$$= \int_0^\infty u^2 \frac{1}{\theta} e^{-u/\theta}\, du$$

$= E(U^2)$, where $U \sim$ exponential (θ)

$= 2\theta^2.$

Therefore,

$$\text{var}_\theta(X^2) = 2\theta^2 - \theta^2$$
$$= \theta^2$$

Therefore

$$\text{var}_\theta(T) = n\theta^2$$

Therefore,

$$E_\theta\left[\left(\frac{T}{n}\right)^2\right] = \frac{1}{n^2}\left(n\theta^2 + (n\theta)^2\right)$$

$$= \theta^2\left(\frac{1}{n} + 1\right)$$

$$= \left(\frac{n+1}{n}\right)\theta^2$$

$$\Rightarrow E_\theta\left[\left(\frac{n}{n+1}\right)\left(\frac{T}{n}\right)^2\right] = \theta^2$$

$$\Rightarrow \phi(T) = \frac{T^2}{n(n+1)}$$ <u>UMVUE</u>

(c) We know T is a sufficient statistic. Let $f_T(t|\theta)$ denote the pdf of T. Find $f_T(t|\theta)$ and show that

$$\{f_T(t|\theta);\ \theta > 0\}$$

has MLR.

First, find pdf of $Y = g(X) = X^2$. Note that $y = g(x) = x^2$ is a monotone function on $\mathbb{R}^+$

$$y = g(x) = x^2$$

$$\Rightarrow x = \sqrt{y} = g^{-1}(y)$$

Therefore, for $y > 0$

$$f_Y(y) = f_X[g^{-1}(y)|\theta] \left| \frac{d}{dy} g^{-1}(y) \right|$$

$$= \frac{2\sqrt{y}}{\theta} e^{-y/\theta} \left| \frac{1}{2\sqrt{y}} \right|$$

$$= \frac{1}{\theta} e^{-y/\theta} I(y>0).$$

Therefore $Y = g(X) = X^2 \sim$ exponential (θ). Therefore

$$T \sim \text{gamma}(n, \theta)$$

(easy mgf argument)

Show $f_T(t|\theta)$ has MLR.

Suppose $\theta_2 > \theta_1$ and form the ratio

$$\frac{f_T(t|\theta_2)}{f_T(t|\theta_1)} = \frac{\frac{1}{\cancel{\Gamma(n)}\theta_2^n} \cancel{t^{n-1}} e^{-t/\theta_2}}{\frac{1}{\cancel{\Gamma(n)}\theta_1^n} \cancel{t^{n-1}} e^{-t/\theta_1}} = \left(\frac{\theta_1}{\theta_2}\right)^n e^{-t\left(\frac{1}{\theta_2} - \frac{1}{\theta_1}\right)}$$

$$= \left(\frac{\theta_1}{\theta_2}\right)^n e^{t\left(\frac{1}{\theta_1} - \frac{1}{\theta_2}\right)}$$

Because $\theta_2 > \theta_1 \iff \frac{1}{\theta_2} < \frac{1}{\theta_1}$

$$\iff \frac{1}{\theta_1} - \frac{1}{\theta_2} > 0.$$

Therefore

$$\frac{f_T(t\mid\theta_2)}{f_T(t\mid\theta_1)} = \underbrace{\left(\frac{\theta_1}{\theta_2}\right)^n}_{>0} \underbrace{e^{t\left(\frac{1}{\theta_1} - \frac{1}{\theta_2}\right)}}_{\text{monotone increasing in } t.}$$

The Karlin Rubin Theorem therefore says that the UMP level α uses the test function

$$\delta(T) = I(T \leq c)$$

$$= \begin{cases} 1, & T \leq c \\ 0, & T > c \end{cases}$$

where c is chosen to satisfy

$$\alpha = E_{\theta_0}[\delta(T)] = P_{\theta_0}(T \leq c) \quad (*)$$

Because $T \sim \text{gamma}(n, \theta)$, clearly $(*)$ implies that we take

$$c = g_{n,\theta_0,\alpha},$$

the lower α quantile of the gamma(n, θ_0) distribution.

$$R = \{\underset{\sim}{x} \in \mathcal{X} : \textstyle\sum_{i=1}^n x_i^2 \leq g_{n,\theta_0,\alpha}\}$$

UMP level α rejection region

Remark: It is possible to write the rejection region in terms of a χ^2 quantile. Note that

$$\alpha = P_{\theta_0}(T \leq c) = P_{\theta_0}\left(\frac{2T}{\theta_0} \leq \frac{2c}{\theta_0}\right)$$

and $\frac{2T}{\theta_0} \sim \text{gamma}(n, 2) \stackrel{d}{=} \chi^2_{(2n)}$. Therefore,

$$R = \{\underset{\sim}{x} \in \mathcal{X} : \sum_{i=1}^n x_i^2 \leq g_{n,\theta_0,\alpha}\}$$
$$= \left\{\underset{\sim}{x} \in \mathcal{X} : \sum_{i=1}^n x_i^2 \leq \frac{\theta_0 \chi^2_{2n,\alpha}}{2}\right\}$$

same

(d) The power function is the probability of the rejection region

$$\beta(\theta) \equiv P_\theta(\underset{\sim}{X} \in R)$$
$$= P_\theta\left(\sum_{i=1}^n X_i^2 \leq g_{n,\theta_0,\alpha}\right)$$

Note: $T = \sum_{i=1}^n X_i^2 \sim \text{gamma}(n, \theta)$. Therefore,

CDF of $T \sim \text{gamma}(n, \theta)$

$$\beta(\theta) = F_T(g_{n,\theta_0,\alpha})$$
$$= \int_0^{g_{n,\theta_0,\alpha}} \frac{1}{\Gamma(n)\theta^n} t^{n-1} e^{-t/\theta}\, dt.$$

You could also write $\beta(\theta)$ in terms of a $\chi^2_{(2n)}$ CDF.