PhD Qualifying Examination–Part I

Department of Statistics
University of South Carolina
August 8, 2025 - 9:00AM–1:00PM

READ FIRST THESE INSTRUCTIONS

1. DO NOT write your name on any of your answer sheets. Instead, write your pre-assigned codename.
2. Use separate sheets of paper for each problem. Do not write on the back of any page.
3. There are four (4) problems on this examination.
4. You have four hours for this examination. All four problems will be graded and are of equal weight.
5. Formulas relating to distributions are provided on the last two pages.

The Problems

1. Suppose we have a random sample of size n, $\mathbf{X} = (X_1, \ldots, X_n)$, from a Poisson distribution with the probability mass function given by

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \text{ if } x = 0, 1, \ldots,$$

and $f(x) = 0$ otherwise, where $\lambda > 0$ is an unknown parameter.

   (a) Provide a uniformly minimum variance unbiased estimator (UMVUE) of λ based on $\mathbf{X}$. Is this estimator the unique UMVUE of λ based on $\mathbf{X}$? Explain.

   (b) What is the maximum likelihood estimator (MLE) of $\theta = e^{-\lambda}$? Denote this estimator by $\hat{\theta}_{\text{MLE}}$. What does $\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta)$ converge to in distribution as $n \to \infty$?

   (c) Define $Y = \mathbb{1}(X_1 = 0)$, where $\mathbb{1}(\cdot)$ is the indicator function.

      (i) Show that $E(Y) = \theta$.

      (ii) Show that a UMVUE of θ is given by

$$\hat{\theta}_{\text{UMVUE}} = \left(\frac{n-1}{n}\right)^{\sum_{i=1}^n X_i}.$$

   (d) Show that, as $n \to \infty$,

$$\sqrt{n}(\hat{\theta}_{\text{UMVUE}} - \theta) \xrightarrow{d} N(0,\, e^{-2\lambda}\lambda).$$

   The following result, which you do not need to prove, may be useful for this part,

$$\lim_{n\to\infty}\left(\frac{n-1}{n}\right)^n = e^{-1}.$$

2. Consider the following two-factor ANOVA model for the response data, with a levels of factor A and b levels of factor B,

$$Y_{ij} = \mu_{ij} + \epsilon_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}, \text{ for } i = 1, \ldots, a, \; j = 1, \ldots, b, \tag{1}$$

where μ is the (fixed) grand mean, α_i is the additive (fixed) main effect of level i of factor A, for $i = 1, \ldots, a$, β_j is the additive (fixed) main effect of level j of factor B, for $j = 1, \ldots, b$, and $\{\epsilon_{ij}, \, i = 1, \ldots, a, \; j = 1, \ldots, b\}$ are random errors assumed to be independent and identically distributed (i.i.d.) according to $N(0, \sigma^2)$. Moreover, the main effects satisfy the constraints that $\sum_{i=1}^{a} \alpha_i = 0$ and $\sum_{j=1}^{b} \beta_j = 0$. The observed data only include one observation per cell. Assume there is no interaction between the two factors.

(a) Consider the single observation in cell (i, j), Y_{ij}. Define the fitted value as

$$\hat{Y}_{ij} = \bar{Y}_{i\cdot} + \bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot} \tag{2}$$

for this cell's observation, where $\bar{Y}_{i\cdot} = b^{-1} \sum_{j=1}^{b} Y_{ij}$, $\bar{Y}_{\cdot j} = a^{-1} \sum_{i=1}^{a} Y_{ij}$, and $\bar{Y}_{\cdot\cdot} = (ab)^{-1} \sum_{i=1}^{a} \sum_{j=1}^{b} Y_{ij}$. Derive from first principles the variance of a fitted value, $\mathrm{Var}(\hat{Y}_{11})$, when $a = 2$ and $b = 3$. It can be helpful to write out the definitions of the sample means in (2).

(b) Suppose, for $a = 2$ and $b = 3$, the observed experimental data are $(Y_{11}, Y_{12}, Y_{13}) = (4.5, 3.6, 6.6)$ and $(Y_{21}, Y_{22}, Y_{23}) = (6.9, 5.2, 8.0)$. Derive a point estimate and a 95% interval estimate for μ_{11} based on the fitted value. The following t-distribution quantiles from `R` can be helpful:

```
> qt(.975,df=1:6)
[1] 12.706205  4.302653  3.182446  2.776445  2.570582  2.446912
```

(c) We initially assumed here that there was no interaction between factors A and B. If you observed the data in part (b), would that cause you to doubt the assumption of no interaction? Justify your answer in some way, either formally or informally.

(d) Discuss at least two other model assumptions that would be difficult to verify based on the sample data, and explain why checking these assumptions would be difficult.

(e) In contrast to the model assumed in parts (a)-(d), assume now that, instead of $N(0, \sigma^2)$, the random error ϵ_{ij} in (1) follows Uniform$(-\theta, \theta)$. Carefully write all the steps of an algorithm you could use to estimate θ and to get an approximately valid 95% confidence interval for μ_{11} in this case.

3. Generalized linear models (GLM) are a useful general family of models to analyze response values from the exponential family. Specifically, the response values $Y_1, \ldots, Y_n$ are independent and follow a distribution that is in the exponential family. Then, conditioning on covariates in $\mathbf{x}_i$ (that can contain a 1 for the intercept term as the first entry of $\mathbf{x}_i$), the mean response μ_i for subject i's response is connected to the linear predictor $\mathbf{x}_i^T\boldsymbol{\beta}$ through a link function $g(\cdot)$, that is, $g(\mu_i) = \mathbf{x}_i^T\boldsymbol{\beta}$, for $i = 1, \ldots, n$, where $\boldsymbol{\beta}$ is the vector of regression coefficients, and "T" refers to the transpose operator.

   (a) (i) A distribution belongs to the exponential family if the probability density/mass function can be written in the form

$$f(y|\theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\},$$

   where θ is the canonical parameter, ϕ is the dispersion parameter, $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ are known functions. Show that the mean of the distribution is $b'(\theta)$, that is, the derivative of $b(\theta)$.

   (ii) Verify that the binomial distribution Binomial(n, π) belongs to the exponential family by specifying the corresponding canonical parameter θ, dispersion parameter ϕ, and functions $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$.

   (iii) For the logistic regression, prove that the observed information (i.e., the negative of the Hessian matrix) and the expected information (i.e., the Fisher information) are the same.

   (b) In an experiment testing the effect of a toxic substance, 1500 experimental insects were divided at random into six groups of 250 each. The insects in each group were exposed to a fixed dose of the toxic substance. A day later, each insect was assessed. The results are shown below: X_i denotes the dose level (on a logarithmic scale) administered to the insects in group i, and Y_i denotes the number of insects that died out of the $n_i = 250$ insects in the group, for $i = 1, \ldots, 6$.

i:	1	2	3	4	5	6
X_i	1	2	3	4	5	6
n_i	250	250	250	250	250	250
Y_i	28	53	93	126	172	197
$\hat{\pi}_i$	0.122	0.215	0.349	0.513	0.674	0.802

   Fitting a logistic regression model to the data, and using Fisher's scoring method to find the maximum likelihood estimates of β_0 and β_1 yields the estimated regression coefficients $\hat{\beta}_0 = -2.644$ and $\hat{\beta}_1 = 0.674$. The corresponding fitted probabilities, $\{\hat{\pi}_i\}_{i=1}^6$, of dying are shown in the above table. The following distribution quantiles from `R` may be useful for part (b): (If additional quantiles are needed to draw a conclusion, you may guess the range of an appropriate quantile.)

```
> qnorm(c(.90,.95,.975,.99))
[1] 1.281552 1.644854 1.959964 2.326348
> qt(.975,df=1:6)
[1] 12.706205  4.302653  3.182446  2.776445  2.570582  2.446912
> qchisq(.95,df=1:6)
[1]  3.841459  5.991465  7.814728  9.487729 11.070498 12.591587
```

(i) Interpret $\exp(\hat{\beta}_1)$ and estimate the probability that an insect dies at the dose level of $X = 3.5$.

(ii) Use an appropriate test to assess the goodness of fit of the logistic regression model.

(iii) Estimate the asymptotic variance of the maximum likelihood estimate $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)^T$, and test if the effect of dose is significant.

4. (a) A random variable X follows a distribution supported on $[0, \pi/2]$, specified by an unknown probability density function $f(x)$. A random sample of size n, $\mathbf{X} = (X_1, \ldots, X_n)$, is drawn to test the null hypothesis

$$H_0 : f(x) = f_0(x)$$

against the alternative

$$H_1 : f(x) = f_1(x).$$

Show that, if

$$f_0(x) = c_0 \mathrm{e}^{\sin^2 x} \quad , \quad 0 \leq x \leq \pi/2,$$

and

$$f_1(x) = c_1 \mathrm{e}^{-2\cos x} \quad , \quad 0 \leq x \leq \pi/2,$$

where c_0 and c_1 are normalization constants, then the most powerful test for testing H_0 against H_1 is given by the decision function

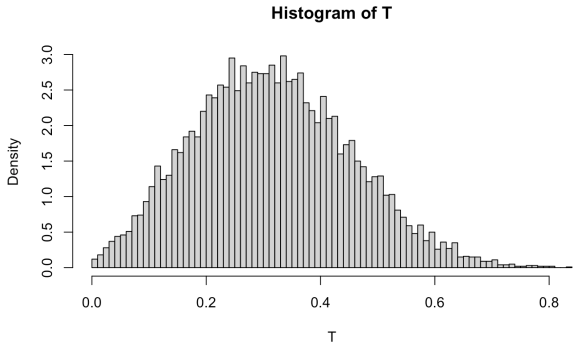
$$\delta(\mathbf{X}) = \mathbb{1}\left(\sum_{i=1}^{n} (\cos x_i - 1)^2 \geq K \right), \tag{3}$$

for some constant K, where $\mathbb{1}(\cdot)$ is the indicator function.

(b) A Monte Carlo simulation study was carried out for the test in part (a). A total of $10,000$ random samples of size $n = 5$ are drawn from H_0 and a histogram of the values of

$$T = \frac{1}{n} \sum_{i=1}^{n} (\cos x_i - 1)^2 \tag{4}$$

is given below,



What is an approximate value of K when $n = 5$ so that the test according to (3) is a level-0.05 test?

(c) Carry out the hypothesis test in part (b) for the following dataset (in radians)

$$(0.5,\ 1,\ 1.2,\ 1.5,\ 1.3)$$

and give your conclusion.

(d) Derive the limiting distribution of T in (4) as $n \to \infty$, without deriving the particular parameter(s) of the limiting distribution.

Table of Common Distributions

taken from *Statistical Inference* by Casella and Berger

Discrete Distrbutions

distribution	pmf	mean	variance	mgf/moment
Bernoulli(p)	$p^x(1-p)^{1-x};\ x=0,1;\ p\in(0,1)$	p	$p(1-p)$	$(1-p)+pe^t$
Beta-binomial(n,α,β)	$\binom{n}{x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(x+\alpha)\Gamma(n-x+\beta)}{\Gamma(\alpha+\beta+n)}$	$\frac{n\alpha}{\alpha+\beta}$	$\frac{n\alpha\beta}{(\alpha+\beta)^2}$	
Notes: If $X\|P$ is binomial (n,P) and P is beta(α,β), then X is beta-binomial(n,α,β).				
Binomial(n,p)	$\binom{n}{x}p^x(1-p)^{n-x};\ x=1,\ldots,n$	np	$np(1-p)$	$[(1-p)+pe^t]^n$
Discrete Uniform(N)	$\frac{1}{N};\ x=1,\ldots,N$	$\frac{N+1}{2}$	$\frac{(N+1)(N-1)}{12}$	$\frac{1}{N}\sum_{i=1}^{N}e^{it}$
Geometric(p)	$p(1-p)^{x-1};\ p\in(0,1)$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{pe^t}{1-(1-p)e^t}$
Note: $Y=X-1$ is negative binomial$(1,p)$. The distribution is *memoryless*: $P(X>s\|X>t)=P(X>s-t)$.				
Hypergeometric(N,M,K)	$\frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}};\ x=1,\ldots,K$	$\frac{KM}{N}$	$\frac{KM}{N}\frac{(N-M)(N-k)}{N(N-1)}$	?
	$M-(N-K)\le x\le M;\ N,M,K>0$			
Negative Binomial(r,p)	$\binom{r+x-1}{x}p^r(1-p)^x;\ p\in(0,1)$	$\frac{r(1-p)}{p}$	$\frac{r(1-p)}{p^2}$	$\left(\frac{p}{1-(1-p)e^t}\right)^r$
	$\binom{y-1}{r-1}p^r(1-p)^{y-r};\ Y=X+r$			
Poisson(λ)	$\frac{e^{-\lambda}\lambda^x}{x!};\ \lambda\ge 0$	λ	λ	$e^{\lambda(e^t-1)}$
Notes: If Y is gamma(α,β), X is Poisson$(\frac{x}{\beta})$, and α is an integer, then $P(X\ge\alpha)=P(Y\le y)$.				

Continuous Distributions

distribution	pdf	mean	variance	mgf/moment
Beta(α,β)	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1};\ x\in(0,1),\ \alpha,\beta>0$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$	$1+\sum_{k=1}^{\infty}\left(\prod_{r=0}^{k-1}\frac{\alpha+r}{\alpha+\beta+r}\right)\frac{t^k}{k!}$
Cauchy(θ,σ)	$\frac{1}{\pi\sigma}\frac{1}{1+(\frac{x-\theta}{\sigma})^2};\ \sigma>0$	does not exist	does not exist	does not exist
Notes: Special case of Students's t with 1 degree of freedom. Also, if X,Y are iid $N(0,1)$, $\frac{X}{Y}$ is Cauchy				
χ^2_p	$\frac{1}{\Gamma(\frac{p}{2})2^{\frac{p}{2}}}x^{\frac{p}{2}-1}e^{-\frac{x}{2}};\ x>0,\ p\in N$	p	$2p$	$\left(\frac{1}{1-2t}\right)^{\frac{p}{2}},\ t<\frac{1}{2}$
Notes: Gamma$(\frac{p}{2},2)$.				
Double Exponential(μ,σ)	$\frac{1}{2\sigma}e^{-\frac{\lvert x-\mu\rvert}{\sigma}};\ \sigma>0$	μ	$2\sigma^2$	$\frac{e^{\mu t}}{1-(\sigma t)^2}$
Exponential(θ)	$\frac{1}{\theta}e^{-\frac{x}{\theta}};\ x\geq 0,\ \theta>0$	θ	θ^2	$\frac{1}{1-\theta t},\ t<\frac{1}{\theta}$
Notes: Gamma$(1,\theta)$. Memoryless. $Y=X^{\frac{1}{\gamma}}$ is *Weibull*. $Y=\sqrt{\frac{2X}{\beta}}$ is *Rayleigh*. $Y=\alpha-\gamma\log\frac{X}{\beta}$ is *Gumbel*.				
F_{ν_1,ν_2}	$\frac{\Gamma(\frac{\nu_1+\nu_2}{2})}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})}\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}}\frac{x^{\frac{\nu_1-2}{2}}}{\left(1+(\frac{\nu_1}{\nu_2})x\right)^{\frac{\nu_1+\nu_2}{2}}};\ x>0$	$\frac{\nu_2}{\nu_2-2},\ \nu_2>2$	$2(\frac{\nu_2}{\nu_2-2})^2\frac{\nu_1+\nu_2-2}{\nu_1(\nu_2-4)},\ \nu_2>4$	$EX^n=\frac{\Gamma(\frac{\nu_1+2n}{2})\Gamma(\frac{\nu_2-2n}{2})}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})}\left(\frac{\nu_2}{\nu_1}\right)^n,\ n<\frac{\nu_2}{2}$
Notes: $F_{\nu_1,\nu_2}=\frac{\chi^2_{\nu_1}/\nu_1}{\chi^2_{\nu_2}/\nu_2}$, where the χ^2s are independent. $F_{1,\nu}=t^2_\nu$.				
Gamma(α,β)	$\frac{1}{\Gamma(\alpha)\beta^\alpha}x^{\alpha-1}e^{-\frac{x}{\beta}};\ x>0,\ \alpha,\beta>0$	$\alpha\beta$	$\alpha\beta^2$	$\left(\frac{1}{1-\beta t}\right)^\alpha,\ t<\frac{1}{\beta}$
Notes: Some special cases are exponential ($\alpha=1$) and χ^2 ($\alpha=\frac{p}{2},\beta=2$). If $\alpha=\frac{2}{3},\ Y=\sqrt{\frac{X}{\beta}}$ is *Maxwell*. $Y=\frac{1}{X}$ is *inverted gamma*.				
Logistic(μ,β)	$\frac{1}{\beta}\frac{e^{-\frac{x-\mu}{\beta}}}{\left[1+e^{-\frac{x-\mu}{\beta}}\right]^2};\ \beta>0$	μ	$\frac{\pi^2\beta^2}{3}$	$e^{\mu t}\Gamma(1+\beta t),\ \lvert t\rvert<\frac{1}{\beta}$
Notes: The cdf is $F(x\lvert\mu,\beta)=\frac{1}{1+e^{-\frac{x-\mu}{\beta}}}$.				
Lognormal(μ,σ^2)	$\frac{1}{\sqrt{2\pi}\sigma}\frac{1}{x}e^{-\frac{(\log x-\mu)^2}{2\sigma^2}};\ x>0,\sigma>0$	$e^{\mu+\frac{\sigma^2}{2}}$	$e^{2(\mu+\sigma^2)}-e^{2\mu+\sigma^2}$	$EX^n=e^{n\mu+\frac{n^2\sigma^2}{2}}$
Normal(μ,σ^2)	$\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}};\ \sigma>0$	μ	σ^2	$e^{\mu t+\frac{\sigma^2t^2}{2}}$
Pareto(α,β)	$\frac{\beta\alpha^\beta}{x^{\beta+1}};\ x>\alpha,\ \alpha,\beta>0$	$\frac{\beta\alpha}{\beta-1},\ \beta>1$	$\frac{\beta\alpha^2}{(\beta-1)^2(\beta-2)},\ \beta>2$	does not exist
t_ν	$\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})}\frac{1}{\sqrt{\nu\pi}}\frac{1}{(1+\frac{x^2}{\nu})^{\frac{\nu+1}{2}}}$	$0,\ \nu>1$	$\frac{\nu}{\nu-2},\nu>2$	$EX^n=\frac{\Gamma(\frac{\nu+1}{2})\Gamma(\nu-\frac{n}{2})}{\sqrt{\pi}\Gamma(\frac{\nu}{2})}\nu^{\frac{n}{2}},\ n$ even
Notes: $t^2_\nu=F_{1,\nu}$.				
Uniform(a,b)	$\frac{1}{b-a},\ a\leq x\leq b$	$\frac{b+a}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{bt}-e^{at}}{(b-a)t}$
Notes: If $a=0,\ b=1$, this is special case of beta ($\alpha=\beta=1$).				
Weibull(γ,β)	$\frac{\gamma}{\beta}x^{\gamma-1}e^{-\frac{x^\gamma}{\beta}};\ x>0,\ \gamma,\beta>0$	$\beta^{\frac{1}{\gamma}}\Gamma(1+\frac{1}{\gamma})$	$\beta^{\frac{2}{\gamma}}\left[\Gamma(1+\frac{2}{\gamma})-\Gamma^2(1+\frac{1}{\gamma})\right]$	$EX^n=\beta^{\frac{n}{\gamma}}\Gamma(1+\frac{n}{\gamma})$
Notes: The mgf only exists for $\gamma\geq 1$.				

PhD Qualifying Examination–Part II

Department of Statistics
University of South Carolina
August 9, 2025 - 9:00AM–1:00PM

READ FIRST THESE INSTRUCTIONS

1. DO NOT write your name on any of your answer sheets. Instead, write your pre-assigned codename.
2. Use separate sheets of paper for each problem. Do not write on the back of any page.
3. There are four (4) problems on this examination.
4. You have four hours for this examination. All four problems will be graded and are of equal weight.
5. Formulas relating to distributions are provided on the last two pages.

The Problems

1. Let $X_1, \ldots, X_n$ be independent and

$$X_i \sim \text{Poisson}\{(n+1-i)\theta\}\ , \quad i = 1, 2, \ldots, n,$$

with different means $(n+1-i)\theta$, where the probability mass function for the distribution Poisson(μ) is

$$f(x|\mu) = \frac{e^{-\mu}\mu^x}{x!}, \quad \text{if } x = 0, 1, 2, \ldots,$$

and $f(x|\mu) = 0$ otherwise, where $\mu > 0$ is the mean of Poisson(μ).

(a) Write down the likelihood given the sample $X_1 = x_1, \ldots, X_n = x_n$.

(b) Show that $\sum_{i=1}^n X_i$ is a sufficient statistic for θ.

(c) Derive the maximum likelihood estimator (MLE) of θ and show that it is unbiased. The following result, which you do not need to prove, may be useful for this part,

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}.$$

(d) Derive the Cramér-Rao lower bound for the variance of an unbiased estimator of θ and show that the MLE of θ achieves the lower bound.

(e) Now consider Bayesian inference for θ. Suggest an appropriate conjugate prior distribution for θ and derive the posterior distribution based on the n independent observations, $X_1, \ldots, X_n$.

(f) Obtain an appropriate Bayesian estimator of θ based on the posterior distribution found in part (e). Discuss how this Bayes estimator compares with the MLE of θ.

2. During exercise, blood flow increases in some parts of the body in response to metabolic demand. Using radioactive microspheres, an experiment was conducted to determine in which of the five parts of the body (factor B) this occurs. Microspheres are distributed in tissue as a function of blood flow. In particular, the greater the blood flow to a part of the body, the more microspheres (and radioactivity) it will contain. The experiment was designed to compare blood flow in five different parts of the body (factor B: bone, brain, skin, muscle, heart) between the resting control condition (factor A: level 1) and during exercise (factor A: level 2). A group of rats were injected intravenously with radioactive microspheres. After the microspheres were injected, some rats were exercised on a treadmill for 15 minutes (factor A: level 2), and the others were placed on the treadmill, but the treadmill was not turned on (factor A: level 1). At the end of a 15-minute period, the rats were sacrificed, and tissues in the five parts were harvested, and the radioactivity in the tissues was measured. The data for this blood flow experiment are presented in Table 1.

Table 1: Radioactivity data from the blood flow experiment

	Body Part				
Exercise Condition	$k=1$ (Bone)	$k=2$ (Brain)	$k=3$ (Skin)	$k=4$ (Muscle)	$k=5$ (Heart)
(No Exercise)	4	3	5	5	4
	1	3	6	3	8
$i=1$	3	1	4	4	7
	1	4	3	2	7
(Exercise)	3	6	12	22	11
	3	5	8	18	12
$i=2$	4	7	10	20	14
	2	4	7	16	8

Based on the corresponding R code and output (see "R code and output for Problem 2" following Problem 4), answer the following questions.

(a) A researcher applies a two-way ANOVA model to analyze the data:

$$Y_{ijk} = \mu_{...} + \alpha_i + \beta_k + (\alpha\beta)_{ik} + \epsilon_{ijk},$$

for $i = 1, 2$; $j = 1, \ldots, 4$; $k = 1, \ldots, 5$, where α_i and β_k denote the main effects of factor A and factor B, respectively, with constraints $\sum_{i=1}^{2} \alpha_i = 0$ and $\sum_{k=1}^{5} \beta_k = 0$, $(\alpha\beta)_{ik}$ denotes the AB interaction effect, and the random errors $\{\epsilon_{ijk},\ i = 1, 2, j = 1, \ldots, 4, k = 1, \ldots, 5\}$ are independent and identically distributed (i.i.d.) according to $N(0, \sigma^2)$, that is, $\epsilon_{ijk} \overset{i.i.d.}{\sim} N(0, \sigma^2)$, for $i = 1, 2, j = 1, \ldots, 4, k = 1, \ldots, 5$.

(i) Comment if the two-way ANOVA model is an appropriate model to analyze the data. Justify your answer clearly.

(ii) Write one or two paragraphs to summarize the analysis results. Justify your findings clearly.

(iii) Conduct a formal contrast comparison to see if the muscle has more blood flow than the other parts of the body during exercise.

(b) Now suppose each row of data in Table 1 is from the same rat. That means there are in total 8 rats (subjects) in the experiment, 4 of them were exercised and 4 of them were not. Given this additional information, the researcher proposes another model to analyze the data:

$$Y_{ijk} = \mu_{...} + \alpha_i + \rho_{j(i)} + \beta_k + (\alpha\beta)_{ik} + \epsilon_{ijk},$$

for $i = 1, \ldots, a$; $j = 1, \ldots, s$; $k = 1, \ldots, b$, where α_i and β_k denote the main effects of factor A and factor B, respectively, with constraints $\sum_{i=1}^{a} \alpha_i = 0$ and $\sum_{k=1}^{b} \beta_k = 0$, $(\alpha\beta)_{ik}$ denotes the AB interaction effect, and $\rho_{j(i)}$ the main random effect of subject j that is nested in level i of Factor A. Assume $\epsilon_{ijk} \overset{i.i.d.}{\sim} N(0, \sigma^2)$ and $\rho_{j(i)} \overset{i.i.d.}{\sim} N(0, \sigma_\rho^2)$, for $i = 1, \ldots, a$ and $j = 1, \cdots, s$, and ϵ_{ijk}'s and $\rho_{j(i)}$'s are independent. For this data set, $a = 2$, $s = 4$, and $b = 5$.

For this model, the analysis of variance and the expected mean squares are summarized as in Table 2.

Table 2: ANOVA and EMS for the blood flow experiment

Source of Variation	SS	E(MS)
Factor A	$bs\sum_{i=1}^{a}(\bar{Y}_{i..} - \bar{Y}_{...})^2$	$\sigma^2 + b\sigma_\rho^2 + bs\frac{\sum_{i=1}^{a}\alpha_i^2}{a-1}$
Factor B	$as\sum_{k=1}^{b}(\bar{Y}_{..k} - \bar{Y}_{...})^2$	$\sigma^2 + as\frac{\sum_{k=1}^{b}\beta_k^2}{b-1}$
AB interactions	$s\sum_{i=1}^{a}\sum_{k=1}^{b}(\bar{Y}_{i.k} - \bar{Y}_{i..} - \bar{Y}_{..k} + \bar{Y}_{...})^2$	$\sigma^2 + s\frac{\sum_{i=1}^{a}\sum_{k=1}^{b}(\alpha\beta)_{ik}^2}{(a-1)(b-1)}$
Subjects(within factor A)	$b\sum_{i=1}^{a}\sum_{j=1}^{s}(\bar{Y}_{ij.} - \bar{Y}_{i..})^2$	$\sigma^2 + b\sigma_\rho^2$
Error	$\sum_{i=1}^{a}\sum_{j=1}^{s}\sum_{k=1}^{b}(Y_{ijk} - \bar{Y}_{i.k} - \bar{Y}_{ij.} + \bar{Y}_{i..})^2$	σ^2

(i) Comment why the current model is a better model than the two-way ANOVA model to analyze the data after the additional information is provided. Provide as much support as possible.

(ii) Provide a point estimate of σ_ρ^2 and describe how to find an approximate confidence interval of σ_ρ^2.

(iii) In this model, homogeneous subject variances are assumed, that is, $\text{Var}(\rho_{j(i)}) = \sigma_\rho^2$, for $i = 1, 2$; $j = 1, 2, 3, 4$. Propose a method to test/check this homogeneous variance assumption against the heterogeneous variances, that is, $\text{Var}(\rho_{j(i)}) = \sigma_{\rho,i}^2$, for $i = 1, 2$; $j = 1, 2, 3, 4$.

(c) Consider the model in part (b), but now assume $\text{Var}(\rho_{j(i)}) = \sigma^2_{\rho,i}$, for $i = 1, 2$, and assume $j = 1, 2$ and $k = 1, 2, 3$ for simplicity.

(i) Derive the covariance matrix Σ_y of $\{Y_{ijk}, \text{for } i = 1, 2; j = 1, 2; k = 1, 2, 3\}$.

(ii) Find the joint distribution of $\{Y_{ijk}, \text{for } i = 1, 2; j = 1, 2; k = 1, 2, 3\}$, where the mean needs to be expressed in a linear regression model format with the design matrix and parameter vector clearly specified.

3. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random sample from a distribution specified by the probability density function $f(x) = (2x/\beta)\exp(-x^2/\beta)$, if $x > 0$, and $f(x) = 0$ otherwise, where $\beta > 0$ is the parameter of interest. Define the gamma function by $\Gamma(t) = \int_0^\infty u^{t-1}e^{-u}\,du$, for $t > 0$. Two useful properties of the gamma function are $\Gamma(t+1) = t\Gamma(t)$ and $\Gamma(1/2) = \sqrt{\pi}$.

   (a) Provide a method of moments estimator for β. Denote this estimator by $\hat{\beta}_1$. What does $\sqrt{n}(\hat{\beta}_1 - \beta)$ converge to in distribution as $n \to \infty$? Prove your claim.

   (b) Derive the maximum likelihood estimator of β, denoted by $\hat{\beta}_2$. What does $\hat{\beta}_2$ converge to in probability as $n \to \infty$? Prove your claim. Can you conclude a stronger mode of convergence for $\hat{\beta}_2$? Explain.

   (c) Is $\hat{\beta}_2$ is a uniformly minimum variance unbiased estimator (UMVUE) for β? Explain.

   (d) Provide a uniformly most powerful (UMP) level-α test for testing $H_0 : \beta \leq \beta_0$ versus $H_1 : \beta > \beta_0$, where β_0 is a pre-specified positive constant.

   (e) Propose two strategies for constructing 95% confidence intervals for β based on $\mathbf{X}$. Confidence intervals with a confidence coefficient approximately equal to 0.95 (when n is sufficiently large) are acceptable here.

4. For count response, Poisson log-linear regression is a widely used approach, i.e., modeling the log transformation of the (conditional) expectation of the response. Meanwhile, the Gaussian regression treats the response as following the Gaussian (normal) distribution. This problem will investigate the inferential results when we use the **identity link**, i.e., directly modeling the (conditional) expectation, for the Poisson regression, as a (misspecified) Gaussian regression, i.e. traditional linear regression.

   Suppose we have independent observations, $Y_1, \ldots, Y_n$, where

$$Y_i | x_i \sim \text{Poisson}(\alpha + \beta x_i), \text{ for } i = 1, \ldots, n,$$

   with $\alpha > 0$ and $\beta \geq 0$. Moreover, we assume that $0 < x_i < M_1 < \infty$ for all $i \in \{1, \ldots, n\}$, and $n^{-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2 = M_2 < \infty$, in which $\bar{x}_n = n^{-1} \sum_{i=1}^n x_i$.

   (a) Suppose we estimate α and β using the ordinary least squares (OLS) approach, where we minimize the following objective function with respect to (α, β),

$$L(\alpha, \beta) = \sum_{i=1}^n \{Y_i - (\alpha + \beta x_i)\}^2.$$

   Define

$$(\hat{\alpha}_{\text{OLS}}, \hat{\beta}_{\text{OLS}}) = \underset{(\alpha,\beta) \in \mathbb{R}^2}{\arg\min}\, L(\alpha, \beta).$$

   (i) State the Gauss-Markov theorem. Do the estimators of α and β obtained using the OLS approach enjoy its result?

   (ii) Are $\hat{\alpha}_{\text{OLS}}$ and $\hat{\beta}_{\text{OLS}}$ unbiased estimators for α and β?

   (iii) Show that $\hat{\beta}_{\text{OLS}}$ converges in probability to β as $n \to \infty$ under the conditions about $\{x_i,\, i = 1, \ldots, n\}$ stated above.

   (b) A randomized controlled trial can be viewed as a two-sample inference problem and handled by a (generalized) linear model.

   Suppose we are interested in comparing the length of stays in hospital for the treatment (new drug) and the control (traditional drug) group, each with n patients. For the i-th patient in the j-th group ($j = 0$ for the control group, and $j = 1$ for the treatment group), we have

$$Y_{ij} \sim \begin{cases} \text{Poisson}(\alpha), & \text{if } j = 0 \text{ (control group)}, \\ \text{Poisson}(\alpha + \beta), & \text{if } j = 1 \text{ (treatment group)}, \end{cases}$$

   for $i = 1, \ldots, n$, and $\{Y_{ij}, i = 1, \ldots, n, j = 0, 1\}$ are independent.

   Define the group mean $\bar{Y}_j = n^{-1} \sum_{i=1}^n Y_{ij}$.

(i) Derive the maximum likelihood estimators for α and β.

(ii) We are interested in testing the **reduction** of the length of stay in hospital. What is the corresponding statistical hypothesis? Express the likelihood ratio test statistic in terms of $\bar{Y}_0$ and $\bar{Y}_1$.

(c) In real applications, the two-sample t-test is widely used even when the data is obviously non-Gaussian. Suppose you "wrongly" run a two sample t-test assuming equal variances for the randomized controlled trial in (b) in a research paper, and the reviewer (probably with less training in mathematical statistics) of the paper wrote "*Since the data is non-Gaussian, I think running a two sample t-test does not make sense to me!*".

Argue your approach is "fine" by

(i) showing that the two-sample t-test statistic converges to the standard normal distribution under the null hypothesis of no treatment effect;

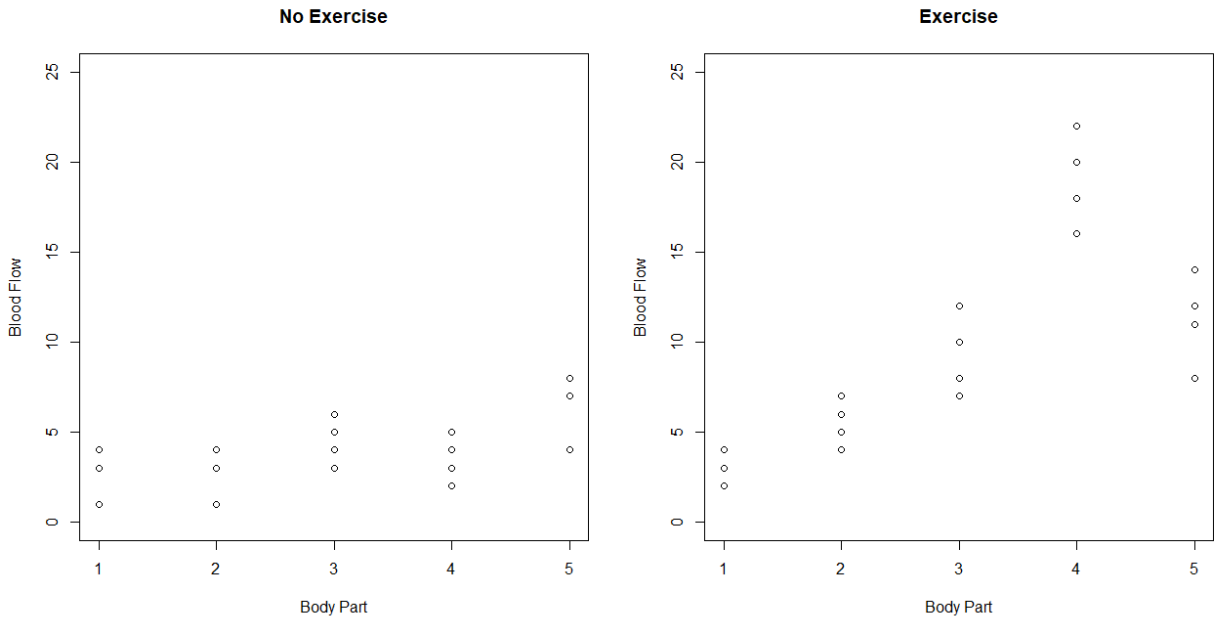
(ii) providing a **non-technical** response to the reviewer (of course, in a polite tone), assuming that you have several thousand patients in the treatment/control group.

R code and output for Problem 2

```
library(lme4); library(lmerTest); library(emmeans); library(car)

setwd( "C:/Users/lin9/Downloads")
da=read.table("BloodFlow.txt",header=T)
da
```

```
##    BF sub Exe part
## 1   4   1   1    1
## 2   3   1   1    2
## 3   5   1   1    3
## 4   5   1   1    4
## 5   4   1   1    5
## 6   1   2   1    1
## 7   3   2   1    2
## 8   6   2   1    3
## 9   3   2   1    4
## 10  8   2   1    5
## 11  3   3   1    1
## 12  1   3   1    2
## 13  4   3   1    3
## 14  4   3   1    4
## 15  7   3   1    5
## 16  1   4   1    1
## 17  4   4   1    2
## 18  3   4   1    3
## 19  2   4   1    4
## 20  7   4   1    5
## 21  3   1   2    1
## 22  6   1   2    2
## 23 12   1   2    3
## 24 22   1   2    4
## 25 11   1   2    5
## 26  3   2   2    1
## 27  5   2   2    2
## 28  8   2   2    3
## 29 18   2   2    4
## 30 12   2   2    5
## 31  4   3   2    1
## 32  7   3   2    2
## 33 10   3   2    3
## 34 20   3   2    4
## 35 14   3   2    5
## 36  2   4   2    1
## 37  4   4   2    2
## 38  7   4   2    3
## 39 16   4   2    4
## 40  8   4   2    5
```

```r
BF=da[,1]; sub=da[,2]; exe=da[,3]; part=da[,4]
sub=factor(sub);  exe=factor(exe);  part=factor(part)

par(mfrow=c(1,2))
plot(as.numeric(part[exe==1]),BF[exe==1],ylim=c(0,25),ylab='Blood Flow', xlab
='Body Part',main="No Exercise")
plot(as.numeric(part[exe==2]),BF[exe==2],ylim=c(0,25),ylab='Blood Flow', xlab
='Body Part',main="Exercise")
```




```r
fit=lm(BF~exe+part+exe:part)
summary(fit)
```

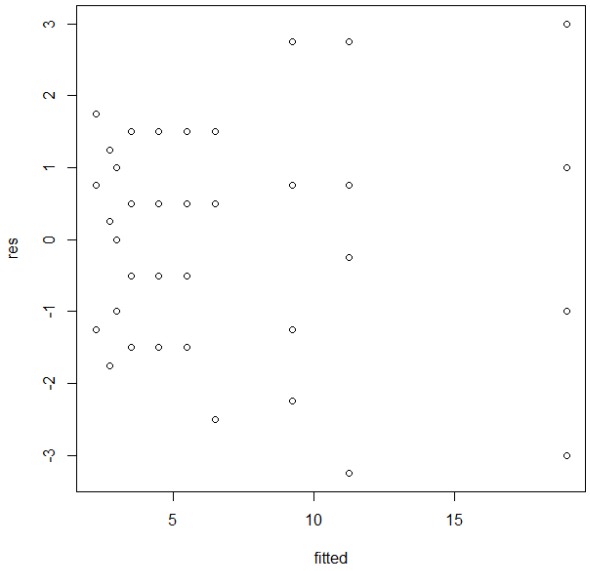
```
##
## Call:
## lm(formula = BF ~ exe + part + exe:part)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -3.25  -1.25   0.25   1.00   3.00
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2500     0.8708   2.584  0.01489 *
## exe2          0.7500     1.2315   0.609  0.54711
## part2         0.5000     1.2315   0.406  0.68762
## part3         2.2500     1.2315   1.827  0.07767 .
## part4         1.2500     1.2315   1.015  0.31822
## part5         4.2500     1.2315   3.451  0.00168 **
## exe2:part2    2.0000     1.7416   1.148  0.25990
## exe2:part3    4.0000     1.7416   2.297  0.02879 *
```

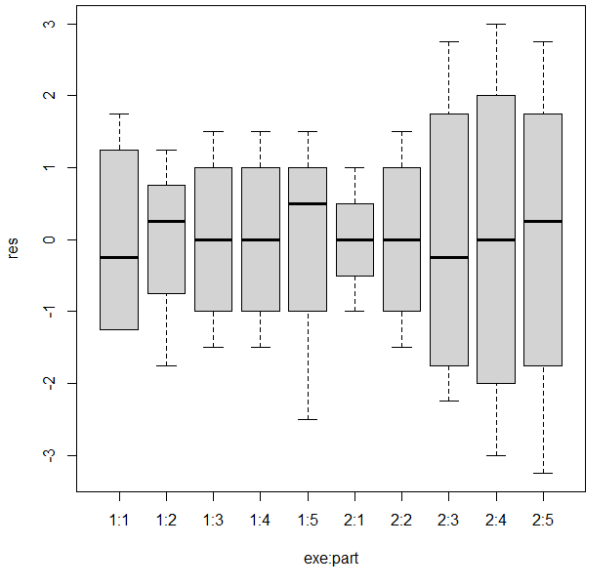
```
## exe2:part4   14.7500     1.7416   8.469 1.89e-09 ***
## exe2:part5    4.0000     1.7416   2.297  0.02879 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.742 on 30 degrees of freedom
## Multiple R-squared:  0.9148, Adjusted R-squared:  0.8892
## F-statistic: 35.77 on 9 and 30 DF,  p-value: 1.242e-13
```

```
anova(fit)
```

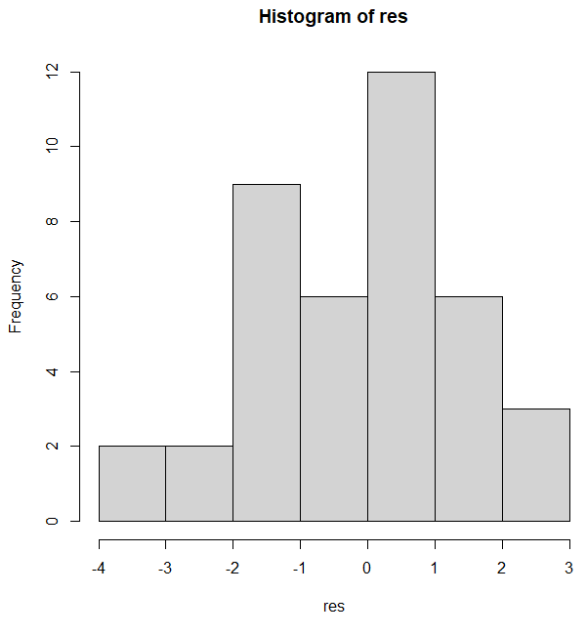
```
## Analysis of Variance Table
##
## Response: BF
##           Df Sum Sq Mean Sq F value    Pr(>F)
## exe        1  324.9  324.90 107.110 2.044e-11 ***
## part       4  389.5   97.38  32.102 1.903e-10 ***
## exe:part   4  262.1   65.52  21.602 1.783e-08 ***
## Residuals 30   91.0    3.03
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

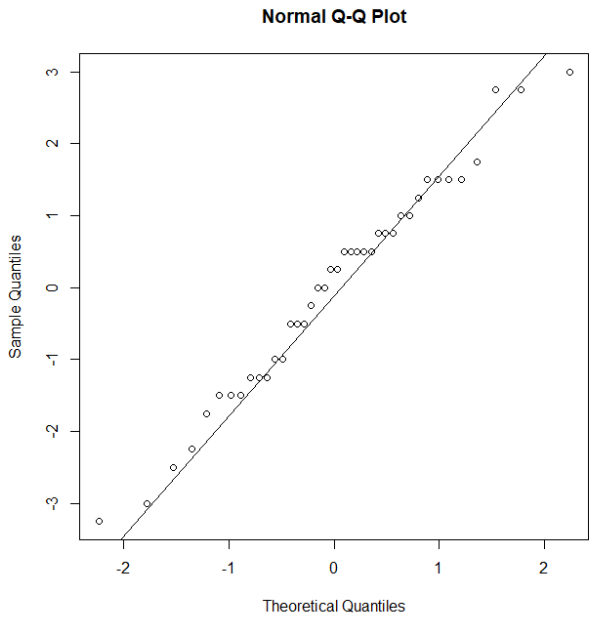
```
res=resid(fit);  fitted=fitted(fit);  group=exe:part

par(mfrow=c(1,2))
plot(fitted,res)
plot(group,res,xlab="exe:part",ylab="res")
```






```
hist(res);  qqnorm(res);  qqline(res)
```






```
leveneTest(res,group)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  9   1.037 0.4348
##       30
```

```
shapiro.test(res)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res
## W = 0.97925, p-value = 0.6618
```

```
em.exe=emmeans(fit,~exe)
```

```
## NOTE: Results may be misleading due to involvement in interactions
```

```
pairs(em.exe)
```

```
##  contrast    estimate    SE df t.ratio p.value
##  exe1 - exe2     -5.7 0.551 30 -10.349  <.0001
##
## Results are averaged over the levels of: part
```

```
em.part=emmeans(fit,~part)
```

```
## NOTE: Results may be misleading due to involvement in interactions
```

```
pairs(em.part)
```

```
##  contrast      estimate    SE df t.ratio p.value
##  part1 - part2    -1.50 0.871 30  -1.723  0.4360
##  part1 - part3    -4.25 0.871 30  -4.880  0.0003
##  part1 - part4    -8.62 0.871 30  -9.904  <.0001
##  part1 - part5    -6.25 0.871 30  -7.177  <.0001
##  part2 - part3    -2.75 0.871 30  -3.158  0.0275
##  part2 - part4    -7.12 0.871 30  -8.182  <.0001
##  part2 - part5    -4.75 0.871 30  -5.455  0.0001
##  part3 - part4    -4.38 0.871 30  -5.024  0.0002
##  part3 - part5    -2.00 0.871 30  -2.297  0.1739
##  part4 - part5     2.38 0.871 30   2.727  0.0732
##
## Results are averaged over the levels of: exe
## P value adjustment: tukey method for comparing a family of 5 estimates
```

```
em=emmeans(fit,~exe*part)
pairs(em)
```

```
##  contrast                estimate   SE df t.ratio p.value
##  exe1 part1 - exe2 part1    -0.75 1.23 30  -0.609  0.9998
##  exe1 part1 - exe1 part2    -0.50 1.23 30  -0.406  1.0000
##  exe1 part1 - exe2 part2    -3.25 1.23 30  -2.639  0.2438
##  exe1 part1 - exe1 part3    -2.25 1.23 30  -1.827  0.7140
##  exe1 part1 - exe2 part3    -7.00 1.23 30  -5.684  0.0001
##  exe1 part1 - exe1 part4    -1.25 1.23 30  -1.015  0.9889
##  exe1 part1 - exe2 part4   -16.75 1.23 30 -13.601  <.0001
##  exe1 part1 - exe1 part5    -4.25 1.23 30  -3.451  0.0456
##  exe1 part1 - exe2 part5    -9.00 1.23 30  -7.308  <.0001
##  exe2 part1 - exe1 part2     0.25 1.23 30   0.203  1.0000
##  exe2 part1 - exe2 part2    -2.50 1.23 30  -2.030  0.5862
##  exe2 part1 - exe1 part3    -1.50 1.23 30  -1.218  0.9635
##  exe2 part1 - exe2 part3    -6.25 1.23 30  -5.075  0.0007
##  exe2 part1 - exe1 part4    -0.50 1.23 30  -0.406  1.0000
##  exe2 part1 - exe2 part4   -16.00 1.23 30 -12.992  <.0001
##  exe2 part1 - exe1 part5    -3.50 1.23 30  -2.842  0.1677
##  exe2 part1 - exe2 part5    -8.25 1.23 30  -6.699  <.0001
##  exe1 part2 - exe2 part2    -2.75 1.23 30  -2.233  0.4578
##  exe1 part2 - exe1 part3    -1.75 1.23 30  -1.421  0.9109
##  exe1 part2 - exe2 part3    -6.50 1.23 30  -5.278  0.0004
##  exe1 part2 - exe1 part4    -0.75 1.23 30  -0.609  0.9998
##  exe1 part2 - exe2 part4   -16.25 1.23 30 -13.195  <.0001
##  exe1 part2 - exe1 part5    -3.75 1.23 30  -3.045  0.1116
##  exe1 part2 - exe2 part5    -8.50 1.23 30  -6.902  <.0001
##  exe2 part2 - exe1 part3     1.00 1.23 30   0.812  0.9978
##  exe2 part2 - exe2 part3    -3.75 1.23 30  -3.045  0.1116
##  exe2 part2 - exe1 part4     2.00 1.23 30   1.624  0.8262
##  exe2 part2 - exe2 part4   -13.50 1.23 30 -10.962  <.0001
##  exe2 part2 - exe1 part5    -1.00 1.23 30  -0.812  0.9978
##  exe2 part2 - exe2 part5    -5.75 1.23 30  -4.669  0.0021
##  exe1 part3 - exe2 part3    -4.75 1.23 30  -3.857  0.0171
```

```
##  exe1 part3 - exe1 part4     1.00 1.23 30   0.812  0.9978
##  exe1 part3 - exe2 part4   -14.50 1.23 30 -11.774  <.0001
##  exe1 part3 - exe1 part5    -2.00 1.23 30  -1.624  0.8262
##  exe1 part3 - exe2 part5    -6.75 1.23 30  -5.481  0.0002
##  exe2 part3 - exe1 part4     5.75 1.23 30   4.669  0.0021
##  exe2 part3 - exe2 part4    -9.75 1.23 30  -7.917  <.0001
##  exe2 part3 - exe1 part5     2.75 1.23 30   2.233  0.4578
##  exe2 part3 - exe2 part5    -2.00 1.23 30  -1.624  0.8262
##  exe1 part4 - exe2 part4   -15.50 1.23 30 -12.586  <.0001
##  exe1 part4 - exe1 part5    -3.00 1.23 30  -2.436  0.3412
##  exe1 part4 - exe2 part5    -7.75 1.23 30  -6.293  <.0001
##  exe2 part4 - exe1 part5    12.50 1.23 30  10.150  <.0001
##  exe2 part4 - exe2 part5     7.75 1.23 30   6.293  <.0001
##  exe1 part5 - exe2 part5    -4.75 1.23 30  -3.857  0.0171
##
## P value adjustment: tukey method for comparing a family of 10 estimates
```

```
em.exe1=emmeans(fit,~exe|part)
pairs(em.exe1)
```

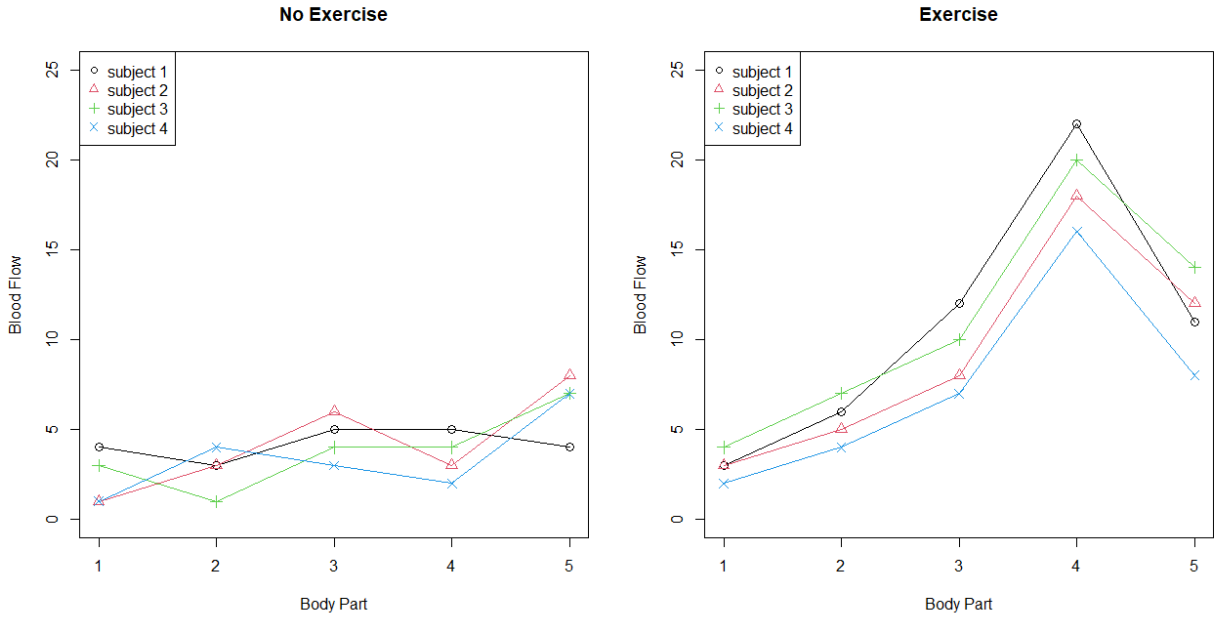
```
## part = 1:
##  contrast    estimate   SE df t.ratio p.value
##  exe1 - exe2    -0.75 1.23 30  -0.609  0.5471
##
## part = 2:
##  contrast    estimate   SE df t.ratio p.value
##  exe1 - exe2    -2.75 1.23 30  -2.233  0.0332
##
## part = 3:
##  contrast    estimate   SE df t.ratio p.value
##  exe1 - exe2    -4.75 1.23 30  -3.857  0.0006
##
## part = 4:
##  contrast    estimate   SE df t.ratio p.value
##  exe1 - exe2   -15.50 1.23 30 -12.586  <.0001
##
## part = 5:
##  contrast    estimate   SE df t.ratio p.value
##  exe1 - exe2    -4.75 1.23 30  -3.857  0.0006
```

```
em.part1=emmeans(fit,~part|exe)
pairs(em.part1)
```

```
## exe = 1:
##  contrast      estimate   SE df t.ratio p.value
##  part1 - part2    -0.50 1.23 30  -0.406  0.9940
##  part1 - part3    -2.25 1.23 30  -1.827  0.3775
##  part1 - part4    -1.25 1.23 30  -1.015  0.8465
##  part1 - part5    -4.25 1.23 30  -3.451  0.0135
##  part2 - part3    -1.75 1.23 30  -1.421  0.6194
##  part2 - part4    -0.75 1.23 30  -0.609  0.9726
```

```
##  part2 - part5    -3.75 1.23 30  -3.045  0.0359
##  part3 - part4     1.00 1.23 30   0.812  0.9249
##  part3 - part5    -2.00 1.23 30  -1.624  0.4943
##  part4 - part5    -3.00 1.23 30  -2.436  0.1334
##
## exe = 2:
##  contrast      estimate   SE df t.ratio p.value
##  part1 - part2    -2.50 1.23 30  -2.030  0.2767
##  part1 - part3    -6.25 1.23 30  -5.075  0.0002
##  part1 - part4   -16.00 1.23 30 -12.992  <.0001
##  part1 - part5    -8.25 1.23 30  -6.699  <.0001
##  part2 - part3    -3.75 1.23 30  -3.045  0.0359
##  part2 - part4   -13.50 1.23 30 -10.962  <.0001
##  part2 - part5    -5.75 1.23 30  -4.669  0.0005
##  part3 - part4    -9.75 1.23 30  -7.917  <.0001
##  part3 - part5    -2.00 1.23 30  -1.624  0.4943
##  part4 - part5     7.75 1.23 30   6.293  <.0001
##
## P value adjustment: tukey method for comparing a family of 5 estimates
```

```
BF1=BF[exe==1]; sub1=sub[exe==1]; part1=part[exe==1]
plot(c(1,5), c(0,25), type='n', ylab='Blood Flow', xlab='Body Part',main="No
Exercise")
for (i in 1:4){lines(part1[sub1==i], BF1[sub1==i], type='l', col=i, pch = i,
cex=1.2)}
for (i in 1:4){lines(part1[sub1==i], BF1[sub1==i], type='p', col=i, pch = i,
cex=1.2)}
legend("toplef",legend=c("subject 1","subject 2","subject 3","subject 4"),col
=1:4,pch=1:4)

BF2=BF[exe==2]; sub2=sub[exe==2]; part2=part[exe==2]
plot(c(1,5), c(0,25), type='n', ylab='Blood Flow', xlab='Body Part',,main="Ex
ercise")
for (i in 1:4){lines(part2[sub2==i], BF2[sub2==i], type='l', col=i, pch = i,
cex=1.2)}
for (i in 1:4){lines(part2[sub2==i], BF2[sub2==i], type='p', col=i, pch = i,
cex=1.2)}
legend("topleft",legend=c("subject 1","subject 2","subject 3","subject 4"),co
l=1:4,pch=1:4)
```




```
fit.new=lm(BF~exe+part+exe:part+exe/sub);  summary(fit.new)
```

```
## Call:
## lm(formula = BF ~ exe + part + exe:part + exe/sub)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.800 -0.800 -0.050  1.038  1.800
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.550e+00  8.832e-01   2.887 0.008098 **
## exe2         1.650e+00  1.249e+00   1.321 0.198940
## part2        5.000e-01  9.874e-01   0.506 0.617217
## part3        2.250e+00  9.874e-01   2.279 0.031876 *
## part4        1.250e+00  9.874e-01   1.266 0.217691
## part5        4.250e+00  9.874e-01   4.304 0.000244 ***
## exe2:part2   2.000e+00  1.396e+00   1.432 0.164975
## exe2:part3   4.000e+00  1.396e+00   2.864 0.008544 **
## exe2:part4   1.475e+01  1.396e+00  10.563 1.67e-10 ***
## exe2:part5   4.000e+00  1.396e+00   2.864 0.008544 **
## exe1:sub2    3.700e-15  8.832e-01   0.000 1.000000
## exe2:sub2   -1.600e+00  8.832e-01  -1.812 0.082575 .
## exe1:sub3   -4.000e-01  8.832e-01  -0.453 0.654681
## exe2:sub3    2.000e-01  8.832e-01   0.226 0.822765
## exe1:sub4   -8.000e-01  8.832e-01  -0.906 0.374035
## exe2:sub4   -3.400e+00  8.832e-01  -3.850 0.000770 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Residual standard error: 1.396 on 24 degrees of freedom
```

```
## Multiple R-squared:  0.9562, Adjusted R-squared:  0.9288
## F-statistic:  34.9 on 15 and 24 DF,  p-value: 1.192e-12
```

```
 anova(fit.new)
```

```
## Analysis of Variance Table
##
## Response: BF
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## exe        1  324.9  324.90 166.6154 2.721e-12 ***
## part       4  389.5   97.38  49.9359 2.718e-11 ***
## exe:part   4  262.1   65.52  33.6026 1.636e-09 ***
## exe:sub    6   44.2    7.37   3.7778  0.008637 **
## Residuals 24   46.8    1.95
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
 em.new.part=emmeans(fit.new,~part|exe)
```

```
## NOTE: A nesting structure was detected in the fitted model:
##     sub %in% exe
```

```
 pairs(em.new.part)
```

```
## exe = 1:
##  contrast      estimate    SE df t.ratio p.value
##  part1 - part2    -0.50 0.987 24  -0.506  0.9859
##  part1 - part3    -2.25 0.987 24  -2.279  0.1864
##  part1 - part4    -1.25 0.987 24  -1.266  0.7137
##  part1 - part5    -4.25 0.987 24  -4.304  0.0021
##  part2 - part3    -1.75 0.987 24  -1.772  0.4118
##  part2 - part4    -0.75 0.987 24  -0.760  0.9396
##  part2 - part5    -3.75 0.987 24  -3.798  0.0071
##  part3 - part4     1.00 0.987 24   1.013  0.8470
##  part3 - part5    -2.00 0.987 24  -2.025  0.2846
##  part4 - part5    -3.00 0.987 24  -3.038  0.0410
##
## exe = 2:
##  contrast      estimate    SE df t.ratio p.value
##  part1 - part2    -2.50 0.987 24  -2.532  0.1166
##  part1 - part3    -6.25 0.987 24  -6.330  <.0001
##  part1 - part4   -16.00 0.987 24 -16.204  <.0001
##  part1 - part5    -8.25 0.987 24  -8.355  <.0001
##  part2 - part3    -3.75 0.987 24  -3.798  0.0071
##  part2 - part4   -13.50 0.987 24 -13.672  <.0001
##  part2 - part5    -5.75 0.987 24  -5.823  <.0001
##  part3 - part4    -9.75 0.987 24  -9.874  <.0001
##  part3 - part5    -2.00 0.987 24  -2.025  0.2846
##  part4 - part5     7.75 0.987 24   7.849  <.0001
## Results are averaged over the levels of: sub
## P value adjustment: tukey method for comparing a family of 5 estimates
```

Table of Common Distributions

taken from *Statistical Inference* by Casella and Berger

Discrete Distrbutions

distribution	pmf	mean	variance	mgf/moment
Bernoulli(p)	$p^x(1-p)^{1-x};\ x=0,1;\ p\in(0,1)$	p	$p(1-p)$	$(1-p)+pe^t$
Beta-binomial(n,α,β)	$\binom{n}{x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(x+\alpha)\Gamma(n-x+\beta)}{\Gamma(\alpha+\beta+n)}$	$\frac{n\alpha}{\alpha+\beta}$	$\frac{n\alpha\beta}{(\alpha+\beta)^2}$	
Notes: If $X\|P$ is binomial (n,P) and P is beta(α,β), then X is beta-binomial(n,α,β).				
Binomial(n,p)	$\binom{n}{x}p^x(1-p)^{n-x};\ x=1,\ldots,n$	np	$np(1-p)$	$[(1-p)+pe^t]^n$
Discrete Uniform(N)	$\frac{1}{N};\ x=1,\ldots,N$	$\frac{N+1}{2}$	$\frac{(N+1)(N-1)}{12}$	$\frac{1}{N}\sum_{i=1}^{N}e^{it}$
Geometric(p)	$p(1-p)^{x-1};\ p\in(0,1)$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{pe^t}{1-(1-p)e^t}$
Note: $Y=X-1$ is negative binomial$(1,p)$. The distribution is *memoryless*: $P(X>s\|X>t)=P(X>s-t)$.				
Hypergeometric(N,M,K)	$\frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}};\ x=1,\ldots,K$	$\frac{KM}{N}$	$\frac{KM}{N}\frac{(N-M)(N-k)}{N(N-1)}$	?
	$M-(N-K)\le x\le M;\ N,M,K>0$			
Negative Binomial(r,p)	$\binom{r+x-1}{x}p^r(1-p)^x;\ p\in(0,1)$	$\frac{r(1-p)}{p}$	$\frac{r(1-p)}{p^2}$	$\left(\frac{p}{1-(1-p)e^t}\right)^r$
	$\binom{y-1}{r-1}p^r(1-p)^{y-r};\ Y=X+r$			
Poisson(λ)	$\frac{e^{-\lambda}\lambda^x}{x!};\ \lambda\ge 0$	λ	λ	$e^{\lambda(e^t-1)}$
Notes: If Y is gamma(α,β), X is Poisson$(\frac{x}{\beta})$, and α is an integer, then $P(X\ge\alpha)=P(Y\le y)$.				

Continuous Distributions

distribution	pdf	mean	variance	mgf/moment
Beta(α,β)	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1};\ x\in(0,1),\ \alpha,\beta>0$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$	$1+\sum_{k=1}^{\infty}\left(\prod_{r=0}^{k-1}\frac{\alpha+r}{\alpha+\beta+r}\right)\frac{t^k}{k!}$
Cauchy(θ,σ)	$\frac{1}{\pi\sigma}\frac{1}{1+(\frac{x-\theta}{\sigma})^2};\ \sigma>0$	does not exist	does not exist	does not exist
Notes: Special case of Students's t with 1 degree of freedom. Also, if X,Y are iid $N(0,1)$, $\frac{X}{Y}$ is Cauchy				
χ^2_p	$\frac{1}{\Gamma(\frac{p}{2})2^{\frac{p}{2}}}x^{\frac{p}{2}-1}e^{-\frac{x}{2}};\ x>0,\ p\in N$	p	$2p$	$\left(\frac{1}{1-2t}\right)^{\frac{p}{2}},\ t<\frac{1}{2}$
Notes: Gamma$(\frac{p}{2},2)$.				
Double Exponential(μ,σ)	$\frac{1}{2\sigma}e^{-\frac{\lvert x-\mu\rvert}{\sigma}};\ \sigma>0$	μ	$2\sigma^2$	$\frac{e^{\mu t}}{1-(\sigma t)^2}$
Exponential(θ)	$\frac{1}{\theta}e^{-\frac{x}{\theta}};\ x\geq 0,\ \theta>0$	θ	θ^2	$\frac{1}{1-\theta t},\ t<\frac{1}{\theta}$
Notes: Gamma$(1,\theta)$. Memoryless. $Y=X^{\frac{1}{\gamma}}$ is *Weibull*. $Y=\sqrt{\frac{2X}{\beta}}$ is *Rayleigh*. $Y=\alpha-\gamma\log\frac{X}{\beta}$ is *Gumbel*.				
F_{ν_1,ν_2}	$\frac{\Gamma(\frac{\nu_1+\nu_2}{2})}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})}\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}}\frac{x^{\frac{\nu_1-2}{2}}}{\left(1+(\frac{\nu_1}{\nu_2})x\right)^{\frac{\nu_1+\nu_2}{2}}};\ x>0$	$\frac{\nu_2}{\nu_2-2},\ \nu_2>2$	$2(\frac{\nu_2}{\nu_2-2})^2\frac{\nu_1+\nu_2-2}{\nu_1(\nu_2-4)},\ \nu_2>4$	$EX^n=\frac{\Gamma(\frac{\nu_1+2n}{2})\Gamma(\frac{\nu_2-2n}{2})}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})}\left(\frac{\nu_2}{\nu_1}\right)^n,\ n<\frac{\nu_2}{2}$
Notes: $F_{\nu_1,\nu_2}=\frac{\chi^2_{\nu_1}/\nu_1}{\chi^2_{\nu_2}/\nu_2}$, where the χ^2s are independent. $F_{1,\nu}=t^2_\nu$.				
Gamma(α,β)	$\frac{1}{\Gamma(\alpha)\beta^\alpha}x^{\alpha-1}e^{-\frac{x}{\beta}};\ x>0,\ \alpha,\beta>0$	$\alpha\beta$	$\alpha\beta^2$	$\left(\frac{1}{1-\beta t}\right)^\alpha,\ t<\frac{1}{\beta}$
Notes: Some special cases are exponential ($\alpha=1$) and χ^2 ($\alpha=\frac{p}{2},\beta=2$). If $\alpha=\frac{2}{3},\ Y=\sqrt{\frac{X}{\beta}}$ is *Maxwell*. $Y=\frac{1}{X}$ is *inverted gamma*.				
Logistic(μ,β)	$\frac{1}{\beta}\frac{e^{-\frac{x-\mu}{\beta}}}{\left[1+e^{-\frac{x-\mu}{\beta}}\right]^2};\ \beta>0$	μ	$\frac{\pi^2\beta^2}{3}$	$e^{\mu t}\Gamma(1+\beta t),\ \lvert t\rvert<\frac{1}{\beta}$
Notes: The cdf is $F(x\lvert\mu,\beta)=\frac{1}{1+e^{-\frac{x-\mu}{\beta}}}$.				
Lognormal(μ,σ^2)	$\frac{1}{\sqrt{2\pi}\sigma}\frac{1}{x}e^{-\frac{(\log x-\mu)^2}{2\sigma^2}};\ x>0,\sigma>0$	$e^{\mu+\frac{\sigma^2}{2}}$	$e^{2(\mu+\sigma^2)}-e^{2\mu+\sigma^2}$	$EX^n=e^{n\mu+\frac{n^2\sigma^2}{2}}$
Normal(μ,σ^2)	$\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}};\ \sigma>0$	μ	σ^2	$e^{\mu t+\frac{\sigma^2t^2}{2}}$
Pareto(α,β)	$\frac{\beta\alpha^\beta}{x^{\beta+1}};\ x>\alpha,\ \alpha,\beta>0$	$\frac{\beta\alpha}{\beta-1},\ \beta>1$	$\frac{\beta\alpha^2}{(\beta-1)^2(\beta-2)},\ \beta>2$	does not exist
t_ν	$\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})}\frac{1}{\sqrt{\nu\pi}}\frac{1}{(1+\frac{x^2}{\nu})^{\frac{\nu+1}{2}}}$	$0,\ \nu>1$	$\frac{\nu}{\nu-2},\nu>2$	$EX^n=\frac{\Gamma(\frac{\nu+1}{2})\Gamma(\nu-\frac{n}{2})}{\sqrt{\pi}\Gamma(\frac{\nu}{2})}\nu^{\frac{n}{2}},\ n$ even
Notes: $t^2_\nu=F_{1,\nu}$.				
Uniform(a,b)	$\frac{1}{b-a},\ a\leq x\leq b$	$\frac{b+a}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{bt}-e^{at}}{(b-a)t}$
Notes: If $a=0,\ b=1$, this is special case of beta ($\alpha=\beta=1$).				
Weibull(γ,β)	$\frac{\gamma}{\beta}x^{\gamma-1}e^{-\frac{x^\gamma}{\beta}};\ x>0,\ \gamma,\beta>0$	$\beta^{\frac{1}{\gamma}}\Gamma(1+\frac{1}{\gamma})$	$\beta^{\frac{2}{\gamma}}\left[\Gamma(1+\frac{2}{\gamma})-\Gamma^2(1+\frac{1}{\gamma})\right]$	$EX^n=\beta^{\frac{n}{\gamma}}\Gamma(1+\frac{n}{\gamma})$
Notes: The mgf only exists for $\gamma\geq 1$.				