Final Exam

Name: _____

This exam is worth 20% of your course grade.

Rules

- **Do** show your work.
- **Do** use prepared "cheatsheets" with definitions, formulae, but *no worked problems, working `R` code, or proofs.* **Don't** use other books, notes, or pieces of paper.

Procedure

1. Do **not** turn this page until directed. Read it thoroughly.
2. Turn in the written portion of your exam, *the cheatsheet(s)*, and any *scrap paper*.

Question	Points	Score
1	25	
2	20	
3	20	
4	25	
5	10	
Total:	100	

GOOD LUCK!!

Short Answer

1. USA Today/Gallup conducted a poll on Nov. 20-22 to assess female opinions about a government advisory panel's recommendation on mammograms. They asked 277 women (some did not give usable answers) who had previously had mammograms whether their experience had been very stressful at least once or never very stressful. They also asked these women whether they agreed with the advisory panel's recommendation to reduce the frequency of mammography screenings. The data are shown below in the Table.

***USA Today*/Gallup Poll on Mammograms.**

		Observed		Expected	
		Advisory Panel Guidelines		Advisory Panel Guidelines	
		Agree	Disagree	Agree	Disagree
Mammogram Very Stressful?	Yes	14	28	$\frac{(14+28)(14+46)}{14+28+46+189}$	$\frac{(14+28)(28+189)}{14+28+46+189}$
	No	46	189	$\frac{(46+189)(14+46)}{14+28+46+189}$	$\frac{(46+289)(28+189)}{14+28+46+189}$

[5 pts] (a) What test*s* could you perform to see if previous experience with mammograms influences female opinion on the guidelines?

> **Solution:** The chi square test of independence or the Fisher's exact test would both test the null hypothesis of no relation or association between experience and opinion.

[5 pts] (b) Fill in the expected values (leave unsimplified formulas, *e.g.* $27 + 95$) in the right side table while retaining row/column marginal sums and assuming that there is no relationship between previous experience and guideline opinion.

> **Solution:** Shown in table.

[5 pts] (c) The statistic $\sum_i \frac{(E_i - O_i)^2}{E_i}$, where O_i are the observed and E_i the expected counts in each cell, is 96.14. What is the asymptotic distribution of this statistic under the null hypothesis of no relationship between experience and opinion?

> **Solution:** Asymptotically chi squared χ^2_1 with 1 degree of freedom as sample size increases.

[5 pts] (d) Suppose it is discovered that more educated women tend to agree with the panel recommendations. A closer examination of the above dataset reveals that significantly more educated women were in the group who had experienced more stressful mammograms. How could you improve the design of the above study to remove the effect of education on the analysis. Write down what the data table would look like with hypothetical observed counts O_1, O_2, *etc.* as well as the test statistic and its asymptotic distribution.

Solution: Pairs of high-stress and no-stress women could be matched on education status (e.g. high school, college, post-graduate) and McNemar's test applied. Women with no match are discarded. A table of the data would look like

		High-Stress Mammogram	
		Agree	Disagree
No-Stress Mammogram	Agree	O_1	O_2
	Disagree	O_3	O_4

and the statistic is

$$\frac{(O_2 - O_3)^2}{O_2 + O_3} \sim \chi^2_1,$$

which asymptotically has a chi square distribution with 1 degree of freedom.

Alternatives. (1) Pairs of "agreeers" and "disagreers" could be matched on education status, with same test and statistic. (2) Logistic regression with panel agreement response and two predictors: education level and previous stressful mammogram experience. Specifically,

$$P(y_{ijk} = 1) = \frac{1}{1 + e^{-(\mu + \alpha_i + \beta_j)}},$$

where y_{ijk} indicates agreement status of kth woman with education status i and no stressful mammograms ($j = 0$) or at least one stressful mammogram ($j = 1$), and α_i and β_j are the education and stressful mammogram effects. Testing $H_0 : \beta_0 = \beta_1 = 0$ in the presence of education effect α_i can reveal if there is a stressful mammogram experience effect on the tendency to agree with the panel.

[5 pts] (e) Why might the cancer status of a woman's relatives be an important variable to consider? Write down a model (mathematical formulation) that would account for the presence x_{i1} of cancer in relatives and presence of stressful mammogram experience x_{i2} in predicting y_i, whether a woman agrees, or not, with the panel. (All three variables are indicator variables.)

Solution: Logistic regression is appropriate, and the mathematical formula is

$$P(y_i = 1) = E[y_i] = \frac{1}{1 + e^{-x_i^T \beta}},$$

where $x_i^T = (1, x_{i1}, x_{i2})$ and $\beta^T = (\beta_0, \beta_1, \beta_2)$.

Alternative.

$$\log\left(\frac{P(y_i = 1)}{1 - P(y_i = 1)}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$$

[5 pts] 2. (a) In a one-way ANOVA there are 10 treatments and 7 observations in each treatment. What is the ratio of the length of the simultaneous confidence intervals (CIs) by the Tukey method compared to the Bonferroni method. (Use notation like $t_{df}(\alpha)$ for the relevant critical values.)

Solution: There are $k = 10$ treatments, $n = 7$ observations per treatment. The error degrees of freedom are $k(n-1) = 60$. There are $\binom{10}{7} = 45$ possible pairwise comparisons to make, so the Bonferroni $\alpha' = \frac{\alpha}{45}$. The CI are

$$\text{Tukey} \quad : \quad \bar{Y}_{i\cdot} - \bar{Y}_{j\cdot} \pm q_{10,60}(1-\alpha)\frac{S_p}{\sqrt{7}}$$

$$\text{Bonferroni} \quad : \quad \bar{Y}_{i\cdot} - \bar{Y}_{j\cdot} \pm t_{60}\left(1 - \frac{\alpha}{90}\right)\frac{S_p\sqrt{2}}{\sqrt{7}},$$

where S_p is the pooled variance or square root of MSE. The lengths are twice the second term, so the ratio is

$$\frac{q_{10,60}\,(1-\alpha)}{t_{60}\left(1-\frac{\alpha}{90}\right)\sqrt{2}}.$$

Here, $q_{10,60}(1-\alpha)$ is the quantile of the studentized range distribution with parameters $k = 10$ and $k(n-1) = 60$ and $t_{60}\left(1-\frac{\alpha}{90}\right)$ is the quantile of a t distribution with 60 degrees of freedom. Because the studentized range is a maximum, a one tail quantile is used. The Bonferroni-corrected t-test may be unusually large or small, so a two tail quantile is used.

[5 pts] (b) What is the ratio of the Tukey method CI length and the CI length obtained without correcting for multiple testing.

Solution: The only change is in the α level of the t test, so the ratio is

$$\frac{q_{10,60}\,(1-\alpha)}{t_{60}\left(1-\frac{\alpha}{2}\right)\sqrt{2}}.$$

[5 pts] (c) Fill in (unsimplified expressions) the incomplete ANOVA table produced from the data described above.

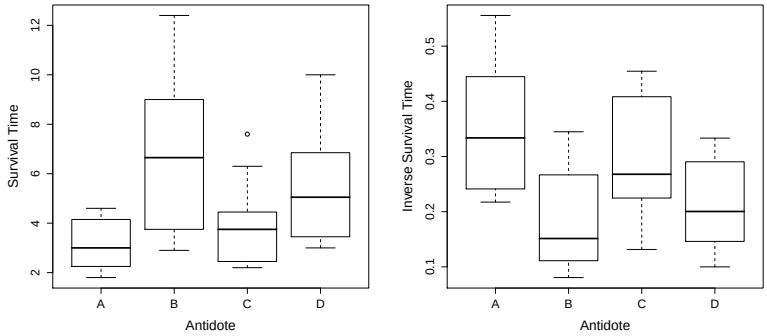
Source	*df*	SS	MS
Treatment	$10 - 1 = 9$	836131	$\frac{836131}{9} = 92903.44$
Error	$10 \times (7 - 1)$ $= 69 - 9 = 60$	$3410 \times 60 = 204600$ $\neq 1071421 - 836131 = 235290$	3410
Total	$9 + 60 = 10 \times 7 - 1$ $= 69$	1071421	

Solution: The multiple ways to compute each entry are shown. There was a typo in the total sum of squares. It should be 1040731, but either calculation was accepted.

[5 pts] (d) Given the above ANOVA table, what is the Tukey CI (unsimplified expression, but with appropriate numbers where available used)?

Solution: All you had to do is recognize that you now had an estimate of $S_p = \sqrt{3410}$, so the CI becomes

$$\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot} \pm q_{10,60}(1-\alpha)\sqrt{3410}/\sqrt{7}.$$



(a) Survival Time (b) Inverse Survival Time

Figure 1: Data Plots

[20 pts] 3. Shown are two analyses of survival time in animals exposed to three poisons, four antidotes, with four observations per treatment. The first ANOVA table is the result when the original times (Fig. 1(a)) were analyzed. The second ANOVA table is the result when the inverse of survival times (Fig. 1(b)) were analyzed. Fill in the degrees of freedom in the first table. What effects are significant? Defend your answer.

Source	*df*	SS	MS
Poison	I-1=2	-	51.52
Antidote	J-1=3	-	30.63
Interaction	(I-1)(J-1)=6	-	5.12
Error	(IJ(K-1)=36	-	2.19

Source	*df*	SS	MS
Poison		-	0.174
Antidote		-	0.068
Interaction		-	0.003
Error		-	0.002

Solution: The degrees of freedom are shown above in red for $I = 3$ poisons and $J = 4$ antidotes. Because critical values decrease with degrees of freedom, we can use the quantiles for 30 denominator degrees of freedom in the table. They are

$$\begin{aligned} F_{2,36}(0.05) &< 3.32 \\ F_{3,36}(0.05) &< 2.92 \\ F_{6,36}(0.05) &< 2.42, \end{aligned}$$

for testing poison effect, antidote effect, and interaction effect respectively. Clearly, both analyses show a significant poison ($\frac{MS_P}{MS_E} = \frac{51.52}{2.19} \approx 25$ or $= \frac{0.174}{0.002} \approx 100$) and antidote ($\frac{MS_A}{MS_E} = \frac{30.63}{2.19} \approx 15$ or $\frac{0.068}{0.002} \approx 34$) effect. The interaction effect is borderline for the first analysis ($\frac{MS_I}{MS_E} = \frac{5.12}{2.19} \approx 2.5$) and insignificant for the second ($\frac{MS_I}{MS_E} = \frac{0.003}{0.002} = 1.5$). Since the variance is clearly not constant for the survival data, but appears relatively constant for the inverse survival data, we trust the second set of results more and conclude that there is probably no interaction effect.

Biologically we conclude that both poisons and antidotes affect survival, but it doesn't seem to matter which antidote is applied to which poison. Apparently, these are generic antidotes.

[5 pts] 4. (a) Convert the following nonlinear model involving unknown parameters a and b to a linear model involving unknown parameters β_0 and β_1. What are β_0 and β_1 in terms of a and b? What are the new independent x and dependent y variables in terms of z and w?

$$z_i = ae^{-bw_i}$$

Solution:

$$\log z_i = \log\left(ae^{-bw_i}\right) = \log a + \log\left(e^{-bw_i}\right) = \log a - bw_i$$

so log the dependent variable $y_i = \log z_i$, leave the independent untouched $x_i = w_i$, and set $\beta_0 = \log a$ and $\beta_1 = -b$.

[5 pts] (b) What kind of error in the measurement of z_i can be accommodated in the linear model?

Solution: Multiplicative error, like $z_i = ae^{-bw_i}e_i$, is OK because it turns into the traditional

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

additive error with $\epsilon_i = \log e_i$ in the transformed model.

To give you an example, suppose measurement error is known to be *proportional* to the true measurement, like the manufacturer or data providers says "the error is within 3%." The log plays such a large role in transforming data because so many kinds of error tend to be this kind.

[10 pts] (c) After obtaining estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, you wish to predict the mean

$$\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

at an unobserved x_0. Use the following covariance matrix to derive the expression for $\text{Var}(\hat{\mu}_0)$.

$$\text{Cov}\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \sigma^2 \sum_i x_i^2 & -\sigma^2 \sum_i x_i \\ - & n\sigma^2 \end{pmatrix} \left[n \sum_i x_i^2 - \left(\sum_i x_i \right)^2 \right]^{-1}$$

Solution:

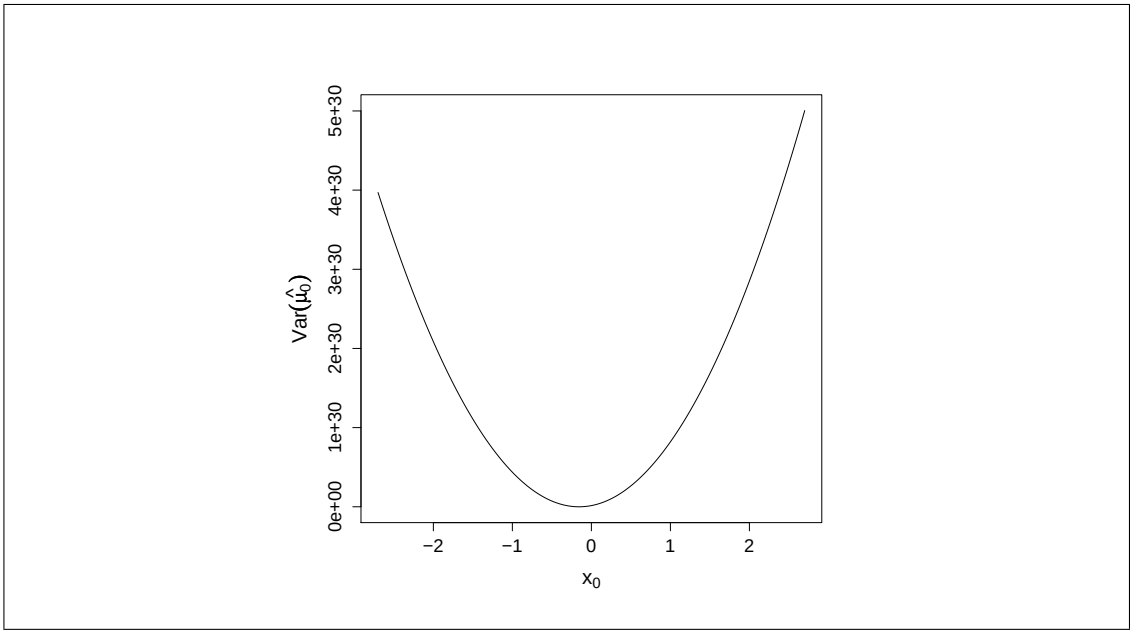
$$\begin{aligned}
\text{Var}(\hat{\mu}_0) &= \text{Var}(\hat{\beta}_0) + x_0^2 \text{Var}(\hat{\beta}_1) + 2x_0 \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) \\
&= \left(\sigma^2 \sum_i x_i^2 + x_0^2 n\sigma^2 - 2x_0\sigma^2 \sum_i x_i \right) / \left[n \sum_i x_i^2 - \left(\sum_i x_i \right)^2 \right] \\
&\quad \text{rest of simplification is optional} \\
&= \sigma^2 \sum_i (x_i - x_0)^2 / \left[n \sum_i x_i^2 - n^2 \bar{x}^2 \right] \\
&= \sigma^2 \sum_i (x_i - \bar{x} + \bar{x} - x_0)^2 / \left[n \sum_i (x_i^2 - 2\bar{x}^2 + \bar{x}^2) \right] \\
&= \sigma^2 \frac{\sum_i (x_i - \bar{x})^2 + \sum_i (\bar{x} - x_0)^2}{[n \sum_i (x_i - \bar{x})^2]} = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_i (x_i - \bar{x})^2} \right) \qquad (1)
\end{aligned}$$

[5 pts] (d) Sketch a plot of $\text{Var}(\hat{\mu}_0)$ as a function of x_0. Where does $\text{Var}(\hat{\mu}_0)$ reach its minimum?

Solution: Notice that the unsimplified (and simplified) expression above is a formula for a parabola in x_0. Whether it opens up or down is determined by the sign of x_0^2, which is

$$\frac{\sigma^2}{\sum x_i^2 - n\bar{x}^2} = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2} > 0,$$

indicating an upward-opening bowl shape. If you didn't see the denominator simplification, then think. The variance should be smallest where we have observed data x_i and high where we don't, so the parabola must open upward with a minimum where the data is located. In fact, you probably had formula (1) for $\text{Var}(\hat{\mu}_0)$ on your cheatsheet, and perhaps even a note about how the minimum is at $x_0 = \bar{x}$. In short, any sketch that looks like an upward-looking parabola would suffice. An example for $\sigma^2 = 1$ and $n = 100$ data points $x_i \overset{\text{iid}}{\sim} N(0, 1)$ is given below.


0e+00
1e+30
2e+30
3e+30
4e+30
5e+30
Var($\hat{\mu}_0$)
−2
−1
0
1
2
x_0

[10 pts] 5. In the squid example used in lecture, the following variables were measured on $n = 22$ squid specimens.

x_1	rostral length
x_2	wing length
x_3	rostral to notch length
x_4	notch to wing length
x_5	width
y	weight

A series of sequential sums of squares are given below

$$\begin{aligned} R(\beta_1 \mid \beta_0) &= 199.1 \\ R(\beta_2 \mid \beta_0, \beta_1) &= 0.1267 \\ R(\beta_3 \mid \beta_0, \beta_1, \beta_2) &= 4.120 \\ R(\beta_4 \mid \beta_0, \beta_1, \beta_2, \beta_3) &= 0.2635 \\ R(\beta_5 \mid \beta_0, \beta_1, \beta_2, \beta_3, \beta_4) &= 4.352, \end{aligned}$$

and the pooled sample variance was $s^2 = 0.4948$. Write an expression in terms of these sequential sums of squares that could be used to test $H_0 : \beta_2 = \beta_4 = 0$. Does it look like this null hypothesis should be rejected? Would you get the same result if the sequence of addition had been $\beta_0, \beta_2, \beta_4, \beta_3, \beta_5$?

Solution: Notice that the hypothesis is stated generally, but is really contingent on the order of addition $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5$. Thus, we can use statistic

$$\frac{R(\beta_2 \mid \beta_0, \beta_1) + R(\beta_4 \mid \beta_0, \beta_1, \beta_2, \beta_3)}{s^2} = \frac{(0.1267 + 0.2635)/2}{0.4948} < 1 \sim F(2, 16)$$

to reject the null hypothesis, but it could (not necessarily, but likely) give completely different results with a different ordering, e.g. $\frac{R(\beta_2|\beta_0)+R(\beta_4|\beta_0,\beta_2)}{s^2}$ under the second ordering. In particular, either β_2 or β_4 may be significant in the absence of β_1.

Alternatives. You could mention our usual test

$$\frac{R(\beta_2, \beta_4 \mid \beta_0, \beta_1, \beta_3, \beta_5)/2}{s^2} \sim F(2, 16)$$

but the numerator cannot be computed from the sequential sums of squares given in the problem. This test would be insensitive to order, but it may fail to notice significant relationships between x_2 or x_4 and the response because all the other variables are included.

You could also mention two tests

$$\frac{R(\beta_2 \mid \beta_0, \beta_1)}{s^2} \sim F(1, 16) \qquad \frac{R(\beta_4 \mid \beta_0, \beta_1, \beta_2, \beta_3)}{s^2} \sim F(1, 16)$$

but they test to separate hypotheses $H_0 : \beta_2 = 0$ and $H_0 : \beta_4 = 0$. These tests do depend on the order of addition.