

Directions: This test consists of 5 problems (5 pages, front and back). Read each problem carefully; some have multiple parts. Write all of your answers in the space provided. You may use the textbook and any notes you wish. You have 2 hours. Partial credit will be assigned where appropriate, so it is in your best interest to **SHOW ALL OF YOUR WORK**. This is a 30-point test. The point value of each problem is shown in square brackets. (Note: The datasets cited here may not be real.)

Problem 1. [7 points] In a study of the effect of the chemical dioxin on reproduction in fish, dioxin levels (in parts per billion) were measured for 18 different ponds in regions of Vietnam that had been exposed to agent orange during the Vietnam War. Researchers fertilized a sample of fish eggs in a sample of water from each of the ponds, then counted how many of the fertilized eggs eventually hatched.

Logistic regression output is provided below.

Logistic Regression Table

| Predictor | Coef | StDev | Z | P |
|-----------|-----------|----------|-------|-------|
| Constant | -0.3049 | 0.2382 | -1.28 | 0.200 |
| Dioxin | -0.029710 | 0.005485 | -5.42 | 0.000 |

Log-Likelihood = -272.368

Test that all slopes are zero: $G = 33.277$, $DF = 1$, $P\text{-Value} = 0.000$

Goodness-of-Fit Tests

| Method | Chi-Square | DF | P |
|----------|------------|----|-------|
| Deviance | 9.234 | 15 | 0.865 |

(a) [1] Explain how you can tell from the above output that **increased** dioxin level appears to be associated with a **decrease** in the probability of fish eggs hatching.

The coefficient of dioxin, $-.0297$, is negative.

Problem 1 cont'd

(b) [2] If π denotes the probability that a fish egg hatches in pond water with a dioxin level of 50 parts per billion, what is the estimated value of π according to the logistic regression model? Show work.

First, find the logit or log-odds: $-.3049 - .0297(50) = -1.79$

Next, exponentiate and convert from odds to a probability:

$$\frac{\exp(-1.79)}{1 + \exp(-1.79)} = \frac{.167}{1.167} = .143$$

(c) [1] The output says that $G = 33.277$. What does this statistic (and its associated p-value) tell us?

The small p-value gives evidence that dioxin is a meaningful predictor of the probability that a fish egg hatched.

(d) [1] The output says that $\text{Chi-square} = 9.234$. What does this statistic (and its associated p-value) tell us?

The large p-value indicates no evidence that this model is deficient in explaining the probability that a fish egg hatched.

(e) [2] Can you infer, based on what you know about this experiment, that increased dioxin levels **cause** a smaller proportion of fish to hatch? Explain.

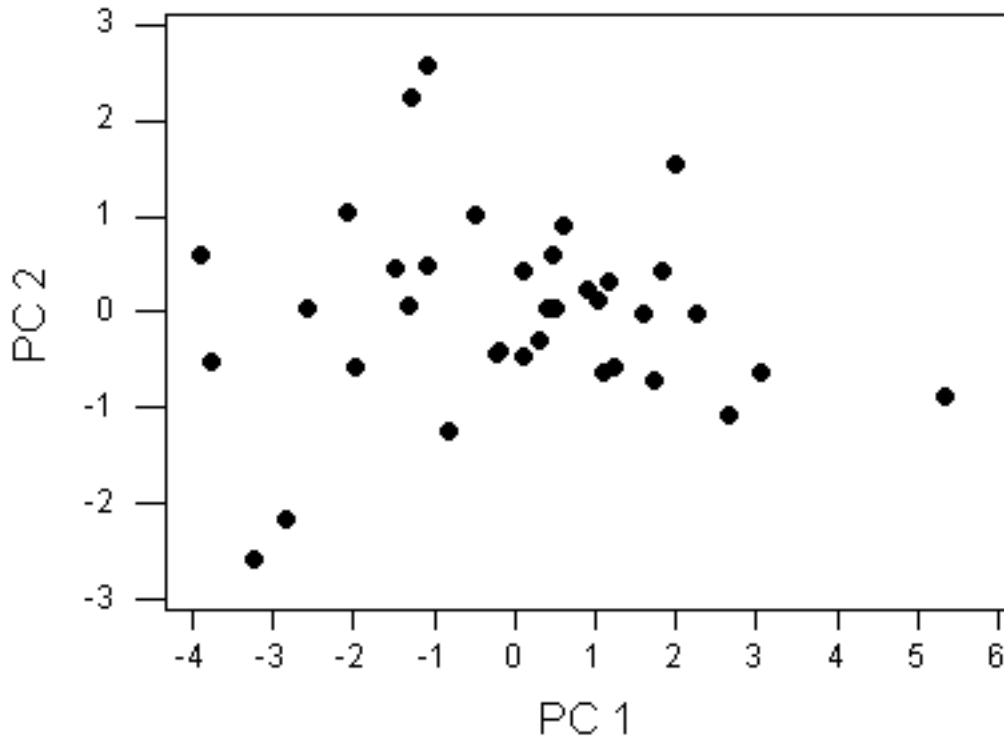
No. This is an observational study, so causal inference is not possible. The different levels of dioxin pollution were not assigned randomly to ponds.

Problem 2. [4 points] In a sample of 36 flea beetles, 6 different physical characteristics are measured for each beetle: head length, body length, length of one joint, length of a second joint, total weight, and body temperature. The first 4 are measured in micrometers, the 5th is measured in milligrams, and the 6th is measured in degrees Celsius. The outcome of a principal components analysis and a related plot are given below.

Eigenanalysis of the Correlation Matrix

| | | | | | | |
|------------|--------|--------|--------|--------|--------|--------|
| Eigenvalue | 4.0424 | 1.0419 | 0.8711 | 0.0372 | 0.0064 | 0.0010 |
| Proportion | 0.674 | 0.174 | 0.145 | 0.006 | 0.001 | 0.000 |
| Cumulative | 0.674 | 0.847 | 0.993 | 0.999 | 1.000 | 1.000 |

| Variable | PC1 | PC2 | PC3 |
|----------|-------|--------|--------|
| head | 0.489 | -0.001 | -0.077 |
| body | 0.495 | 0.065 | 0.028 |
| jnt1 | 0.437 | -0.084 | 0.393 |
| jnt2 | 0.372 | 0.083 | -0.723 |
| weight | 0.496 | -0.057 | 0.023 |
| temp | 0.035 | 0.824 | 0.562 |



Problem 2 cont'd

(a) [1] As you can see, the correlation matrix is used here instead of the covariance matrix. Explain why the correlation matrix is a more sensible choice than the covariance matrix for this analysis.

The variables are on completely different scales, so it makes sense to standardize them all before analyzing their contributions to the variation. This is what the correlation matrix does, not the covariance matrix.

(b) [1] How much of the variation in this dataset is explained by the first principal component? How much is explained by the first and the second principal components together?

The first PC explains 67.4%. The first two together explain 84.7%.

(c) [2 points] A plot of the principal component scores for the 36 beetles is shown. Imagine that the x and y axes are drawn onto the plot, dividing it into 4 quadrants: UR (upper right), UL (upper left), LR (lower right), and LL (lower left).

Suppose two new beetles, Bert and Ernie, are measured. Bert is much larger than any of the other beetles and is also slightly warmer. Ernie, on the other hand, is very small compared with the other beetles but like Bert, Ernie is warmer than the others.

For both Bert and Ernie, tell which quadrant of the above graph each would lie in. Explain briefly how you know.

Bert: UR

Ernie: UL

From the loadings, we can tell that the first PC is essentially nothing but the sum of the size measurements and the second PC is essentially nothing but the temperature. Thus, Bert will have a high PC 1 score and a high PC 2 score, whereas Ernie will have a low PC 1 score and a high PC 2 score.

Problem 3. [9 points] An observational study to contrast cholesterol levels in rural and urban Guatemalan Indians came up with measurements of serum total cholesterol levels (mg/l) for two samples of Indians: One of them is urban (n=45), while one of them is rural (n=49). The samples are not random samples. Researchers are interested in whether cholesterol levels among urban Indians tend to be higher than among rural Indians.

A particular kind of graphical presentation of the data is given below.

Leaf Unit = 10

| URBAN | 0 9 | RURAL |
|-------------|----------------------|-------|
| ----- | 1 0011 | ----- |
| | 33 1 222333333 | |
| | 5 1 4444444444555555 | |
| | 777 1 66677777 | |
| | 99998888 1 888999 | |
| 11000000000 | 2 0 | |
| 333322222 | 2 2223 | |
| 54444 | 2 | |
| | 77 2 | |
| | 888 2 | |
| | 3 | |
| | 3 3 | |

(a) [2] The plot allows us to assess visually two of the assumptions made by a 2-sample t test. Name these two assumptions. For each assumption, tell what this plot indicates about the validity of that assumption for this dataset.

The normality and equal variance assumptions both appear to be justified for this dataset.

(b)[2] State, using appropriate mathematical symbols, the correct null and alternative hypotheses for this experiment. Then express these hypotheses in words.

$$H_0 : \mu_U = \mu_R \qquad H_a : \mu_U > \mu_R$$

The null hypothesis says that the mean cholesterol for the urban population is equal to the mean cholesterol for the rural population. The alternative says that the urban mean is greater than the rural mean.

Problem 3 cont'd

(c)[1] Below is some Minitab output. Find the missing T statistic. Show work.

Two sample T for choles

| code | N | Mean | StDev | SE Mean |
|------|----|-------|-------|---------|
| 1 | 45 | 216.9 | 39.9 | 6.0 |
| 2 | 49 | 157.0 | 31.8 | 4.5 |

95% CI for mu (1) - mu (2): (45.1, 74.6)

T-Test mu (1) = mu (2) (vs >): T = ??? P = 0.0000 DF = 92

Both use Pooled StDev = 35.9

$$T = \frac{\bar{x}_U - \bar{x}_R}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{216.9 - 157.0}{35.9 \sqrt{\frac{1}{45} + \frac{1}{49}}} = 8.081$$

(d) [2] It is stated that the samples are not random samples. What does this mean about the types of inferences we can or cannot make from the results of the study?

We cannot infer the results of this study to the populations in question. In other words, we cannot assume that μ_U is truly larger than μ_R .

(e) [2] A nonparametric test (a test that doesn't make the assumptions in part (a)) is run on this dataset in Minitab. The test statistic found by Minitab is:

W = 2983.0

Give the name of this test (either the name given in the textbook or the name used by Minitab). Without actually attempting to do it, explain how to go about finding the W statistic for this dataset.

This is the rank-sum, or Mann-Whitney, test. The W statistic is found by ranking all 94 observations, then adding the ranks found in one of the groups.

Problem 4. [5 points] Previous studies suggest that vegetarians may not receive enough zinc in their diets. The zinc requirement is particularly important during pregnancy. Researchers conducted a study to determine whether vegetarian pregnant women are at greater risk from low zinc levels than nonvegetarian pregnant women. 23 women were monitored: twelve vegetarians who were pregnant, six nonvegetarians who were pregnant, and five vegetarians who were not pregnant. The zinc content was measured in a hair sample from each woman (Data are in micrograms per gram).

(a) [2] Fill in the blanks in the ANOVA table below.

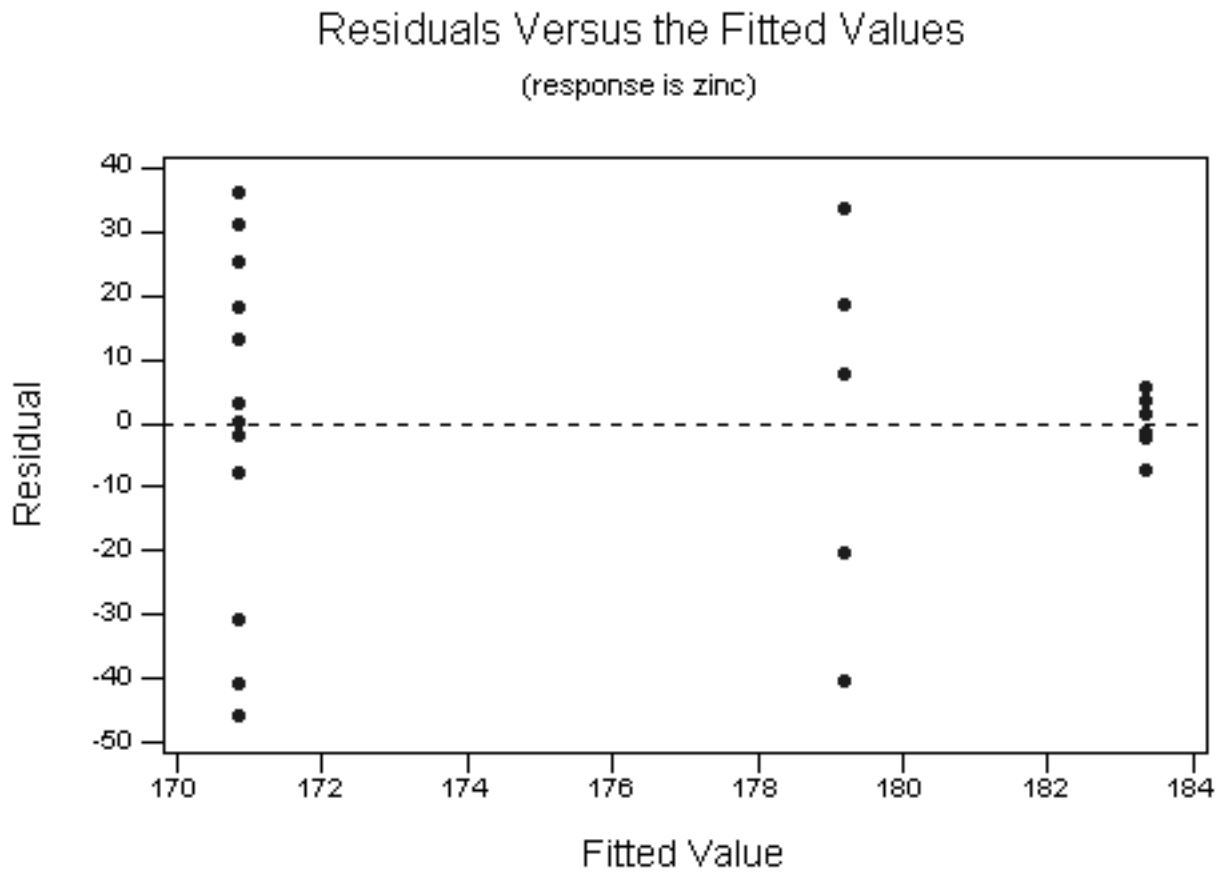
Analysis of Variance for zinc

| Source | DF | SS | MS | F | P |
|--------|-------|-------|-------|-------|-------|
| Group | 2 | 694 | 347 | 0.58 | 0.567 |
| | ===== | | ===== | ===== | |
| Error | 20 | 11900 | 595 | | |
| | ===== | ===== | | | |
| Total | 22 | 12594 | | | |
| | ===== | | | | |

(b) [1] Find the pooled estimate of variance based on the ANOVA table above.

The pooled sample variance is the same as the mean square error (595 in this case).

Problem 4 cont'd



(c) [2] Does the above plot indicate anything about the appropriateness of the model assumed here? Explain.

The assumption of equality of variances among the three groups is clearly inappropriate since the rightmost group has visibly a much smaller variance than the other two.

Problem 5. [5 points] It is often argued that victims of violence exhibit more violent behavior toward others. To study this hypothesis a researcher searched court records to find 905 individuals who had been victims of abuse as children (11 years or younger). She then found 660 individuals, with similar demographic characteristics, who had not been abused as children. Based on a search through subsequent years of court records, she was able to determine how many in each of these groups became involved in violent crimes, as shown in the following table:

| | Involved in violent crime? | |
|--------------|----------------------------|-----|
| | Yes | No |
| Abuse Victim | 104 | 801 |
| Control | 52 | 608 |

(a) [2] If ω denotes the odds that an individual will be involved in a violent crime, give a 95% confidence interval for the odds ratio $\omega_{\text{Abused}}/\omega_{\text{Controls}}$. Show all work.

The estimated odds ratio $\hat{\omega}_A/\hat{\omega}_C$ is $(104 \times 608/52 \times 801) = 1.518$. We use the fact that its log is approximately normally distributed with mean $\log(\omega_A/\omega_C)$ and standard error

$$\sqrt{\frac{1}{n_A \hat{\pi}_A (1 - \hat{\pi}_A)} + \frac{1}{n_B \hat{\pi}_B (1 - \hat{\pi}_B)}} = \sqrt{\frac{905}{104 \times 801} + \frac{660}{52 \times 608}} = .178$$

Thus, the interval for $\log(\omega_A/\omega_C)$ is $\log(1.518) \pm 1.96(.178) = (.0685, .7663)$. Exponentiating gives the confidence interval for ω_A/ω_C :

$$(1.071, 2.152)$$

(b) [2] This is an observational study, not a randomized experiment. Explain what this means about the types of inferences we can make, and then explain why a randomized experiment would not be feasible in this case.

We may not make causal inferences. In other words, we may not claim that abuse as children CAUSES them to become involved in violent crime later.

A randomized experiment would involve randomly assigning certain children to abusive homes and others to non-abusive homes, which is clearly unethical.

Problem 5 cont'd

(c) [1] The Minitab output below begins to compute a statistic, X. Fill in all four blanks below to tell what the statistic is called, finish computing it, and give its associated degrees of freedom.

X is a chi-squared statistic on 1 df, calculated by adding for each of the 4 cells $(\text{observed} - \text{expected})^2/\text{expected}$. The missing summand, corresponding to the upper left cell, is $(104 - 90.21)^2/90.21 = 2.108$.

Degrees of freedom equal $(\text{rows} - 1) \times (\text{columns} - 1)$.

Expected counts are printed below observed counts

| | Yes | No | Total |
|---------|--------------|---------------|-------|
| Abuse | 104 90.21 | 801 814.79 | 905 |
| Control | 52 65.79 | 608 594.21 | 660 |
| Total | 156 | 1409 | 1565 |

$$\begin{aligned}
 X = & 2.108 + 0.233 + \\
 & \text{=====} \\
 & 2.890 + 0.320 = 5.551 \\
 & \text{=====}
 \end{aligned}$$

$$\begin{aligned}
 \text{DF} = & 1, \text{ P-Value} = 0.018 \\
 & \text{=====}
 \end{aligned}$$

The statistic above is called chi-squared statistic
 =====