Ann Marie Weideman

MS Applied 2017

*Note:* This test was reviewed in a session held by a professor who had the correct test answers. All answers have been adapted to match the soln key reviewed in class (updated from my previous post, as some of my answers were incorrect).

1. **a) Determine the mean number of previous pregnancies per woman. (Report the value rounded to 2 decimal places.)**

$$mean = \frac{sum(ppnum)}{\# \ ID's} = \frac{679}{453} \approx 1.50 \ pregnancies/woman$$

**b) In other studies, the distribution of the number of previous pregnancies is approximated reasonably well by the Poisson distribution. Assuming a Poisson distribution having the mean calculated in part (a), how many women in this dataset would one expect to have had at least 4 previous pregnancies? Test whether the observed number differs significantly from the expected number.**

Note: Must assume data is not subject to overdispersion due to a data mean that is very different than the data variance. If it was, would be better to model with a negative binomial instead of a Poisson (which has identical mean and variance).

Using the pmf of the Poisson distribution, we can calculate the proportion of women with at least 4 pregnancies as $1 - \sum_{x=0}^{3} \frac{\exp(-1.5)(1.5)^x}{x!} = 1 - 0.93436 = 0.06564$ or at least 30 (0.06564*453) women. **Note:** Can also do using `1-ppois(3,1.5)` and multiply this by the number of rows in the dataset (nrow(dataset)) to get at least 30 women.

Use z-test for proportions to determine whether the observed proportion differs significantly from the expected proportion:

$H_0: p_1 - p_2 = 0$

$$Z = \frac{(p_1 - p_2) - 0}{\sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}}} = \frac{(0.06564 - 0) - 0}{\sqrt{\frac{0.06564(1 - 0.06564)}{453} + \frac{0(1 - 0)}{453}}} \approx 5.64$$

Assuming a significance level of $\alpha = 0.05$ and two-sided testing, have $Z = 5.64 > 1.96$. Thus, we reject the null hypothesis that the proportions are equivalent. We conclude that the observed number of pregnancies is significantly lower than expected.

Note, you can also do the above problem using a chi-squared test. If you square my answer, you should get your chi-squared value. I am just lazy and think z-scores are easier to calculate (and I know the default critical value, so why not use it).

**c) Test whether the normal distribution is a good model for the distribution of the ultrasound version of GA.**

Will use the Shapiro-Wilk test (solution used a KS test, but you get same result here either way) and q-q plot to determine if the data is approximately normally distributed.

Implementing this test in R using the shapiro.test function, we produce the test statistic W = 0.74326 with p-value < 2.2e-16. Assuming a significance level of $\alpha = 0.05$, we reject the null hypothesis that the data comes from a normal distribution and conclude that the normal distribution is NOT a good model for the ultrasound version of GA.

In addition to the Shapiro-Wilk test, we can visualize the data by producing a quantile-quantile (q-q) plot, as in the below figure (Note: According to professor, this was not necessary here). Note the heavy tails since the outer parts of the curve are steeper than the middle part. This plot is in agreement with our earlier conclusion, that this data would not be modeled well with a normal distribution.

**Normal Q-Q Plot**
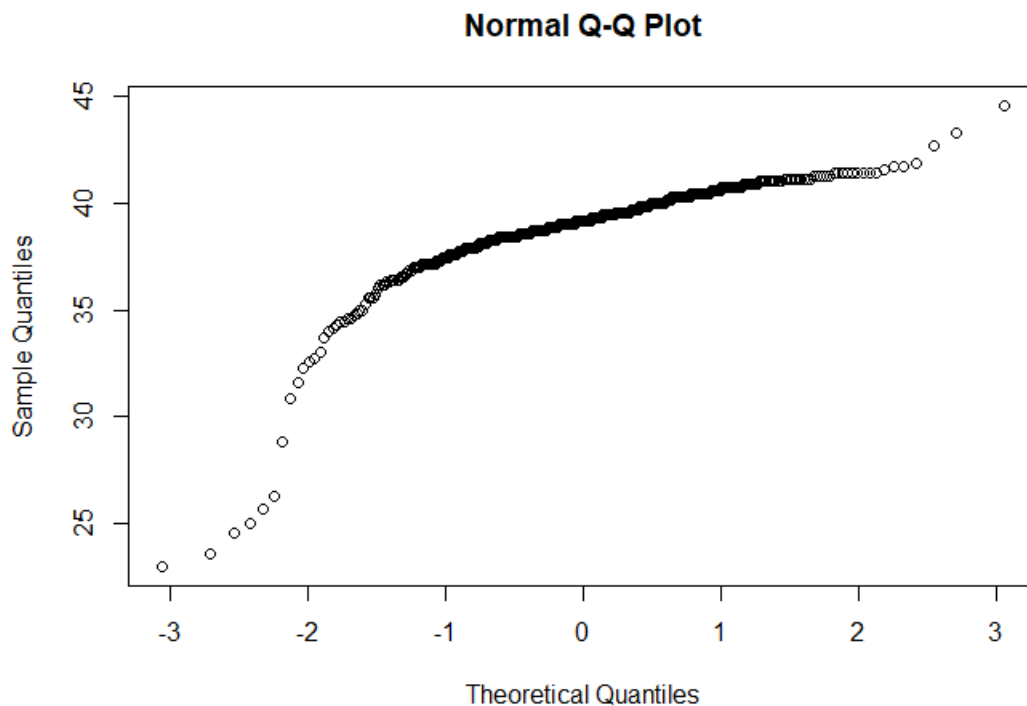


*d) Classify both versions of gestational age into 3 intervals, (0,37), [37,40), and [40,∞).*
*Determine how well the two versions agree and provide a 95% confidence interval for the true*
*agreement.*

We will estimate the Cohen's kappa statistic to test the null hypothesis that the extent of agreement is the same as random agreement (i.e., kappa=0).

Utilized the Kappa.test() function in the fmsb package in R to produce both the statistic and accompanying 95% CI.

Result:

$\kappa \approx 0.69$ with $p < 2.2e - 16$

95% $CI = (0.628, 0.752)$

Using a significance level of $\alpha = 0.05$, since the p-value is well below 0.05, and since we are 95% confident that the true value for the kappa statistic falls somewhere in the interval from 0.628 to 0.752 (not close to including 0), we can conclude that there is moderate to substantial agreement between the two versions of gestational age (after classifying into three intervals).

**e) Does the risk of preterm delivery vary monotonically with the number of previous pregnancies?**

Use Chi-squared test for trend.

$H_0: \pi_0 = \pi_1 = \pi_2 = \pi_3$ vs. $H_1: \pi_0 \leq \pi_1 \leq \pi_2 \leq \pi_3$ or $H_1: \pi_0 \geq \pi_1 \geq \pi_2 \geq \pi_3$ with at least one strict inequality and where 3 is the maximum number of observed pregnancies in the dataset.

(See R code to obtain below table)

| Pregnancies | Zero | One | Two | Three |
|---|---|---|---|---|
| Preterm | 12 | 12 | 12 | 12 |
| Full term | | | | |
| Totals | 90 | 154 | 102 | 107 |

(I didn't bother computing full term, but you can fill them in by subtracting if you'd like).

Calculation of Test Statistic:

$X^2_{trend} = 5.8 \times 10^{-5}$ (see R code, using prop.trend.test() function)

Critical Region: $C_{0.05} = \left\{ X^2_{trend} : X^2_{trend} > \chi^2_{3,0.95} = 7.82 \right\}$ (via Chi-Square table)

P-value: p=0.9939 (see R code)

Conclusion: Assuming a significance level of $\alpha = 0.05$, we fail to reject the null hypothesis since our statistic of $5.8 \times 10^{-5}$ falls well outside of the critical region and our p-value >> 0.05. We conclude that the risk of preterm delivery does NOT appear to vary monotonically with the number of previous pregnancies.

**f) Based on this study, is treating periodontal disease in pregnant women effective in terms of reducing the risk of prematurity?**

Use a chi-squared test to determine if the proportions of women with preterm delivery is the same in women treated for periodontal disease while pregnant versus those treated for periodontal disease post-delivery.

Assumptions for chi-squared test:

- Categories are mutually exclusive and each participant contributes to one, and only one, cell.
- The value of the expecteds should be 5 or more in at least 80% of the cells and no cell should have an expected of less than 1.

$H_0: \pi_1 = \pi_2$ vs $H_1: \pi_1 \neq \pi_2$

where $\pi_1$ is the proportion of women with preterm delivery in the prenatal group and $\pi_2$ is the proportion of women with preterm delivery in the postnatal group.

$X^2 = 0.426$ (see R code)

df = (nrow-1)*(ncol-1) = (2-1)*(2-1) = 1

Critical Region: $C_{0.05} = \{X^2 : X^2 > \chi^2_{1,0.95} = 3.84\}$

P-value: p = 0.514 (see R code)

Conclusion: Assuming a significance level of $\alpha = 0.05$, we fail to reject the null hypothesis. We conclude that the intervention does not appear to be successful in reducing the risk of prematurity in pregnant women.

### (g) For the prenatal treatment group, did the mean average pocket depth change from baseline to after delivery?

Want to compare mean change within a participant, so use either a paired t-test (parametric) or a Wilcoxon signed-rank test (non-parametric) depending on the data distribution. Since the sample size is large (n=176, excluding missingness), we can rely on CLT and Slutsky's and it would be acceptable to use a paired t-test.

$H_0: \mu_{diff} = 0$ vs $H_1: \mu_{diff} \neq 0$

Where $\mu_{diff}$ is the difference between the average pocket depth from baseline to post-delivery

Test statistic: $t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{0.00337733 - 0}{0.4993049/\sqrt{176}} \approx 0.08974$ (verified in R)

Critical Region: $C_{0.05} = \{t : |t| > t_{175,0.05} = 1.97\}$

P-value: p=0.929 (from R)

Conclusion: Assuming a significance level of $\alpha = 0.05$, since the p-value of 0.929 is well above 0.05, we fail to reject the null hypothesis. We conclude that there is no evidence of a significant increase in the mean average pocket depth from baseline to after delivery.

### (h) Did the mean change in average pocket depth differ between the two treatment groups?

Want to compare a mean change between groups. For this, we would use either an unpaired t-test (parametric) or a Wilcoxon rank sum test (non-parametric).

Since the sample size is large can rely on CLT and Slutsky's, so it is OK to use two-sample t-test.

$H_0: \mu_{diff1} = \mu_{diff2}$ vs $H_1: \mu_{diff1} \neq \mu_{diff2}$

Where $\mu_{diff1}$ is the difference between the average pocket depth from baseline to post-delivery in the prenatal group and $\mu_{diff2}$ is the difference between the average pocket depth from baseline to post-delivery in the postpartum group. Will assume unequal variances.

Test statistic: $t = \frac{(\bar{x}_1 - \bar{x}_2) - (0-0)}{\sqrt{\frac{var(\bar{x}_1)}{n_1} + \frac{var(\bar{x}_2)}{n_2}}} \approx 3.63$ (verified in R)

(Note: Due to unequal variances, the formula for calculating df is wonky. You will get df=345.94. You can Google the formula if you like. Either way, the sample size is large enough to simply assume that $t_{crit} = 1.96$)

P-value: p=0.0003 (from R)

Conclusion: Assuming a significance level of $\alpha = 0.05$, since the p-value of 0.0003 is well below 0.05, we reject the null hypothesis. We conclude that there is a significant difference in the mean change in average pocket depth between the two treatment groups. The prenatal treatment group saw a slight improvement in periodontal disease (decrease in pocket depth of -0.003 mm) from pre- to post- delivery. However, the postnatal treatment group saw a worsening in periodontal disease (increase in pocket depth of 0.17 mm) from pre- to post-delivery.

### (i) Based on the data on average pocket depth, discuss the effectiveness of the periodontal therapy and the consequences for the potential to affect birthweight.

The data suggests that periodontal therapy is not effective in reducing the risk of prematurity in women who are provided dental intervention while pregnant. However, while the intervention did not seem to help the child, it did, however, appear to have helped the mother. Those who received periodontal treatment while pregnant had improved periodontal outcomes after childbirth compared to those who received periodontal treatment postpartum.

2. **(a) The researchers would like to first verify whether the mean response in patients receiving chemotherapy differs across clinical subtypes. Please use cell means coding for this section.**

*i. Write a linear model in matrix notation (using cell means coding) to evaluate this question. Specify the dimension of each component in your model.*

No intercept in cell means coding scheme:

$$y_{168 \times 1} = X_{168 \times 2}\beta_{2 \times 1} + \varepsilon_{168 \times 1}$$

$$\begin{bmatrix} y_1 \\ \vdots \\ n_1 \\ \vdots \\ y_{168} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_{168} \end{bmatrix}$$

where $X$ is full rank 2, $n_1$ is the sample size in the HR+/HER2+ group (level=1), $\beta_1$ is the coefficient corresponding to the HR+/HER2+ group, and $\beta_2$ is the coefficient corresponding to the HR-/HER2+ group.

*ii. Fit this linear model and interpret the coefficients. Then, evaluate whether mean response in patients receiving chemotherapy differs across clinical subtypes.*

**Fit this linear model and interpret the coefficients:**

Model fit as in the above matrix notation using R code (see uploaded code). Since a cell means model does not include an intercept, the coefficients correspond to the two levels of the $\beta$ vector.

The estimate of 63.3 for the HR+/HER2+ subtype indicates that 63.3% of the tumor remains in chemo patients with this subtype. The estimate of 90.1 for the HR-/HER2+ subtype indicates that 90.1% of the tumor remains in chemo patients with this subtype.

**Then, evaluate whether mean response in patients receiving chemo differs across clinical subtypes.**

We can use two methods to compare these models in R. These two methods were reviewed during our review session using R code.

**Method 1:** Use a contrast matrix. Since testing $H_0: \beta_1 = \beta_2 \rightarrow \beta_1 - \beta_2 = 0 \rightarrow C = [1, -1]$

**Method 2:** Fit the full and reduced models and compare with anova().

Method 1 will return a t-value (or F value, depending on which package you use – see R code). Square this, and you will get the F value returned in Method 2. Honestly, I think Method 2 is easier because it requires less code (and you avoid having to create a contrast), but it's all personal preference. If you have a hypothesis that is more complicated, then supplying contrasts will make your life easier at that point.

Using Method 1 results returns $t = -5.547 \sim t_{(full\ df - red\ df)\ =\ rank(C)} \equiv t_1$ with $p = 1.13e - 07$

Assuming a significance level of $\alpha = 0.05$, since $p = 1.13e - 07 < \alpha$, we reject the null hypothesis and conclude that the mean response in patients receiving chemo differs significantly between the HR+/HER2+ and HR-/HER2+ subtypes.

*iii. The researchers hypothesize that the variance in chemotherapy response may be larger in the HR+/HER2+ group compared to the HR-/HER2+ group. Evaluate whether the variance in chemotherapy response is homogeneous across clinical subtypes, reporting the test, test statistic, df, p-value for the test, and your interpretation of the result. What additional steps should be taken given your test result, if any?*

Use Levene's test to test for homogeneity of variance using the leveneTest() function in R.

$F = 0.8171 \sim F_{1,166}$

$p = 0.37$

Assuming a significance level of $\alpha = 0.05$, since p=0.37 (>0.05), we fail to reject the null hypothesis that there is homogeneity between the groups. Thus, we can safely use hypothesis tests that require the groups have relatively equal variance.

*iv. Traditionally, the mean response to chemotherapy among HR+/HER2+ patients has been shown to be twice that of HR-/HER2+ patients. Evaluate whether this relation holds in this*

*study. Write the H0, HA, C, β and ϑ0 for carrying out this test in the GLH testing framework. Carry out this test and report the test statistic, its df and p-value. Does this result agree with or contradict previous results?*

$H_0: 2\beta_1 = \beta_2$ vs. $H_0: 2\beta_0 \neq \beta_1$

$C = (2, -1), \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \quad \theta_0 = 0$

Here $\hat{\theta} = C\hat{\beta} = (2, -1) \begin{pmatrix} 63.319 \\ 90.071 \end{pmatrix} = 36.567$

$$F = \frac{\{(\hat{\theta} - \theta_0)'[C(X'X)^{-1}C']^{-1}(\hat{\theta} - \theta_0)/Rank(C)\}}{SSE/dfE}$$

$$= \frac{\left\{(36.567)'\left[(2, -1)\begin{pmatrix} 1/121 & 0 \\ 0 & 1/47 \end{pmatrix}(2, -1)'\right]^{-1}(36.567)/1\right\}}{130712/166}$$

$$\approx 31.25 \sim F_{1,166}$$

P-value computed using `pf(31.25331,1,166,lower.tail=T)` in R gives p=9.13e-08, thus we reject the null hypothesis.

The above by-hand solution was also verified in R using the linearHypothesis function in the car package.

This result agrees with the previous results, which found the mean response to chemotherapy among HR+/HER2+ patients to be only 1.42x (not quite 2x) that of HR-/HER2+ patients.

***v. Calculate the estimated covariance between ˆβ1 and ˆβ2. Explain how one may arrive at this result without access to computer software.***

Can use the vcov() function in R to get the variance-covariance matrix,

$$\Sigma = \begin{pmatrix} 6.51 & 0 \\ 0 & 16.75 \end{pmatrix}$$

Then, grab the estimated covariance (upper off-diagonal entry), $Cov(\hat{\beta}_1, \hat{\beta}_2) = 0$.

The covariance between the two coefficients can by computed, by hand, using the below formula:

$$\Sigma = \hat{\sigma}^2(X'X)^{-1}$$

where $\hat{\sigma}^2$ is the MSE and $(X'X)^{-1}$ is the inverse SSCP. Again, grab the upper off-diagonal entry.

***(b) Next, the researchers hypothesize that the relative effectiveness of chemotherapy versus hormone therapy may vary across subtypes. That is, the difference in mean response between patients receiving chemotherapy and those receiving hormone therapy may not be the same across subtypes.***

***i. With respect to clinical subtypes, evaluate this claim by first choosing an appropriate model. Then write out the H0, HA, C, β and ϑ0 for carrying out this test in the GLH testing framework***

with your chosen model. Carry out this test and interpret the result. Based on the fitted model, suggest which therapy may be better for each clinical subtype and why.*

$$Response = \beta_0 + \beta_1 I(subtype = HR-) + \beta_2 treatment + \beta_3 I(subtype = HR-) * treatment + \varepsilon$$

Then, $H_0: \beta_3 = 0$ vs $H_A: \beta_3 \neq 0$

$$C = [0\ 0\ 0\ 1], \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}, \theta_0 = 0$$

(See three methods in R). We see that the difference in mean response between patients receiving chemotherapy and those receiving hormone therapy significantly differs across the clinical subtypes ($t = 5.147$ (on $1\ df$), $p = 4.82e - 07$).

Based on the fitted model, we can perform a Tukey's HSD (honest significant differences) test to determine which type of treatment is beneficial for each subtype. This test assumes there are equal variance between the groups and incorporates an adjustment for sample size in unbalanced designs. Note that, in the following, a negative difference implies the first type is better at reducing tumor volume and a positive difference implies the second type is better at reducing tumor volume.

If a patient is HR+/HER2+, it appears that s/he would benefit more from chemotherapy than hormone therapy (HR+:Chemo- HR+:Hormone = -19.01, p-adj=1.21e-05).

If a patient is HR-/HER2+, it appears that s/he would benefit more from hormone therapy than chemotherapy (Chemo:HR- - Hormone:HR- = 10.38, p-adj=0.06). However, this difference was not statistically significant at a significance level of $\alpha = 0.05$.

***ii) Next, the researchers would like to evaluate whether molecular subtype is better at explaining response to treatment than clinical subtype. Fit the same model as the prior sub-part, except this time using molecular subtype instead of clinical subtype. Report the R2 from each model. Can we use R2 to help choose between the two models? Why or why not? If no, propose an alternative measure and use it to help suggest which model may be more appropriate.***

R-squared for model in i): $R^2 = 0.17$

R-squared for model in ii): $R^2 = 0.58$

We cannot use R-squared to help us choose between the two models because the models are not nested. If they were nested, we would want to compared adjusted R-squared (which corrects for number of predictors). However, we can use AIC (or BIC) to compare the two non-nested models. The lower the AIC, the better.

AIC for model in i): $AIC = 2734$

AIC for model in ii): $AIC = 2541$

Thus, it may be more appropriate to choose the second model over the first, since its AIC is lower.

***iii) Researchers would like to directly test whether the fit of the previous model using clinical subtype is much better than the fit of the model using molecular subtype. Carry out this test. If you cannot carry out this test, explain why.***

Cannot carry out this test because the models are not nested. The AIC from the first is based on df=5 (covariates) and df=9 (covariates).

```
        df      AIC
 lm_bi_full    5 2734.061
 lm_bii_full   9 2541.049
```

***3.***

***(a) Test the null hypothesis that, in non-hypertensive women, low birth weight of the first child is not associated with low birth weight of the second child. Apply the same test to the hypertensive group.***

The non-hypertensive group:

|                      | Second born not LBW | Second born LBW |
|----------------------|---------------------|-----------------|
| First born not LBW   | 60                  | 10              |
| First born LBW       | 20                  | 10              |

The hypertensive group:

|                      | Second born not LBW | Second born LBW |
|----------------------|---------------------|-----------------|
| First born not LBW   | 48                  | 12              |
| First born LBW       | 22                  | 18              |

Chi-squared statistic formula for a 2x2 contingency table:

$$\chi^2 = \frac{(ad - bc)^2(a + b + c + d)}{(a + b)(c + d)(a + c)(b + d)}$$

Non-hypertensive: $\chi^2 = \frac{(ad-bc)^2(a+b+c+d)}{(a+b)(c+d)(a+c)(b+d)} = \frac{(60*10-20*10)^2(60+10+20+10)}{(60+10)(20+10)(60+20)(10+10)} \approx 4.76 \sim \chi_1^2$

Hypertensive: $\chi^2 = \frac{(ad-bc)^2(a+b+c+d)}{(a+b)(c+d)(a+c)(b+d)} = \frac{(48*18-22*12)^2(48+12+22+18)}{(48+12)(22+18)(48+22)(12+18)} \approx 7.14 \sim \chi_1^2$

When a comparison is made between one sample and another, a simple rule is that the degrees of freedom equal (number of columns minus one) x (number of rows minus one) not counting the totals for rows or columns. For our data this gives df = (2-1) x (2-1) = 1.

From the chi-squared table, $\chi^2 = 4.76$ on 1 df and $\alpha = 0.05$ has a p-value in the range 0.025<p<0.05.

From the chi-squared table, $\chi^2 = 7.14$ on 1 df and $\alpha = 0.05$ has a p-value in the range 0.001<p<0.01.

Thus, in both the non-hypertensive and hypertensive groups, low birth weight in the first child is associated with low birth weight in the second child.

***(b) Suggest a measure of association between low birth weight of the first child and low birth weight of the second child. Compute a point estimate and a 95% confidence interval for this measure, separately for each group of women.***

According to soln. from professor, can use OR, Kappa, or Pearson correlation. I think this may be a case-control study, for which we could not use RR. Professor wasn't sure though, when I asked if it was case-control. It appears to be so to me at least. Note: I forgot to write the correct answer here, so beware that my soln. for this part may have errors.

Here going to use odds ratio (OR).

Non-Hypertensive: $OR = \dfrac{ad}{bc} = \dfrac{60(10)}{10(20)} = 3.00$

$$SE(\ln(OR)) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} = \sqrt{\frac{1}{48} + \frac{1}{12} + \frac{1}{22} + \frac{1}{18}} = 0.52$$

$95\% \ CI(OR) = \exp\big(ln(OR) \pm 1.96 * SE(ln(OR))\big) = \exp(ln(3.00) \pm 1.96 * 0.52) = (1.08, 8.31)$

Hypertensive: $OR = \dfrac{ad}{bc} = \dfrac{48(18)}{12(22)} = 3.27$

$$SE(\ln(OR)) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} = \sqrt{\frac{1}{60} + \frac{1}{10} + \frac{1}{20} + \frac{1}{10}} = 0.45$$

$95\% \ CI(OR) = \exp\big(ln(OR) \pm 1.96 * SE(ln(OR))\big) = \exp(ln(3.27) \pm 1.96 * 0.45) = (1.35, 7.90)$

Thus, in the non-hypertensive group, for women whose first pregnancy was low birthweight, the odds of a second child with low birthweight are 3.00 times the odds for women whose first pregnancy was not low birthweight.

Thus, in the hypertensive group, for women whose first pregnancy was low birthweight, the odds of a second child with low birthweight are 3.27 times the odds for women whose first pregnancy was not low birthweight.

***(c) Test the null hypothesis that, in non-hypertensive women, the risk of low birth weight for the first born child is the same as the risk of low birth weight for the second born child. Apply the same test to the hypertensive group.***

***Note!!! A bunch of us used an independent proportions test here. However, the solution key said to give zero credit for this problem if students assumed the groups are independent. So, have to use McNemar's test, which is appropriate for correlated data. Think carefully about why this data might be correlated.***

$H_0: \pi_1 = \pi_2$ vs $H_1: \pi_1 \neq \pi_2$

where $\pi_1$ is the proportion of women with low birthweight for the first born child and $\pi_2$ is the proportion of women with low birthweight for the second born child.

Non-hypertensive: $X^2 = \frac{(b-c)^2}{b+c} = \frac{(10-20)^2}{10+20} = 3.33 \sim \chi_1^2$

Hypertensive: $X^2 = \frac{(b-c)^2}{b+c} = \frac{(12-22)^2}{12+22} = 2.94 \sim \chi_1^2$

From table of critical values for chi-square distribution, $\chi_{1,0.05}^2 = 3.84$. Since both chi-square statistics fall below this critical value, we fail to reject the null hypothesis and conclude that the risk of low birth weight for the first born child is no different from the risk of low birth weight for the second born child, both in the non-hypertensive and hypertensive groups.

**(d) Test the null hypothesis that the joint distribution of the first and second low birth weight outcomes is the same in the two groups of women.**

**Null:** The two populations follow the same distribution.

**Alternative:** The two populations have different distributions.

**Test:** Chi-squared test for homogeneity.

**Df:** (nrow-1)*(ncol-1) = (4-1)*(2-1) = 3

Note: (don't include totals in your row or columns counts in the above formula)

**Observed:**

| | Non-HTN | HTN | Totals |
|---|---|---|---|
| 1st normal, 2nd normal | 60 | 48 | 108 |
| 1st normal, 2nd low | 10 | 12 | 22 |
| 1st low, 2nd normal | 20 | 22 | 42 |
| 1st low, 2nd low | 10 | 18 | 28 |
| Totals | 100 | 100 | 200 |

**Expected:**

| | Non-HTN | HTN |
|---|---|---|
| 1st normal, 2nd normal | (100/200)*108 = 54 | (100/200)*108 = 54 |
| 1st normal, 2nd low | (100/200)*22 =11 | (100/200)*22 = 11 |
| 1st low, 2nd normal | (100/200)*42 = 21 | (100/200)*42 = 21 |
| 1st low, 2nd low | (100/200)*28 = 14 | (100/200)*28 = 14 |

$X^2 = \frac{(60-54)^2}{54} + \frac{(48-54)^2}{54} + \frac{(10-11)^2}{11} + \frac{(12-11)^2}{11} + \frac{(20-21)^2}{21} + \frac{(22-21)^2}{21} + \frac{(10-14)^2}{14} + \frac{(18-14)^2}{14} \approx 3.896$

(checked in R as well)

P-value: (for $\alpha = 0.05$ on 3 df from chi-squared table): $0.25 < p < 0.50$

(checked in R as well, exact p-value was 0.2729)

**Conclusion:** Assuming a significance level of $\alpha = 0.05$ (chi-squared test is one-sided), since $0.25<p<0.50$ we fail to reject the null hypothesis. We conclude that the non-hypertensive and hypertensive groups are likely to follow the same distribution.

*e) The data from only the first child (200 births) were used to fit a logistic regression model (using maximum-likelihood) with low birth weight as response and an indicator of hypertension (x = 0 if non-hypertensive, x = 1 if hypertensive) as a covariate,*

*logitP(low birth weight) = β1 + β2x.*

*To the extent possible, present a table of parameter and standard error estimates of β1 and β2. (Remember: only a calculator is allowed).*

Construct new contingency table for only first child births:

|  | *Non-HTN* | *HTN* | *Total* |
|---|---|---|---|
| *First born not LBW* | 60+10 = 70 | 48+12 = 60 | 130 |
| *First born LBW* | 20+10 = 30 | 22+18 = 40 | 70 |
| *Total* | 100 | 100 | 200 |

**For $\widehat{\beta}_1$:**

$$logit[P(\text{LBW in Non-HTN})] = \beta_1 + \beta_2(0) = logit\left(\frac{30}{100}\right) = \log\left(\frac{\frac{30}{100}}{1 - \frac{30}{100}}\right) \approx -0.847$$

$$\boldsymbol{SE}(\widehat{\boldsymbol{\beta}}_1) = \sqrt{\frac{1}{70} + \frac{1}{30}} \approx 0.218$$

(formula for SE of a log(OR))

**For $\widehat{\beta}_2$ (*change in log odds when $x = 1$*):**

$$logit[P(\text{LBW in HTN})] - logit[P(\text{LBW in Non-HTN})] = (\beta_1 + \beta_2(1)) - (\beta_1 + \beta_2(0)) =$$

$$logit\left(\frac{40}{100}\right) - logit\left(\frac{30}{100}\right) = \log\left(\frac{\frac{0.4}{(1-0.4)}}{\frac{0.30}{1-0.30}}\right) \approx 0.442$$

Or can do using:

$$logit[P(\text{LBW in HTN})] = \beta_1 + \beta_2 => \beta_2 = logit[P(\text{LBW in HTN})] - \beta_1 = \log\left(\frac{0.4}{1 - 0.4}\right) + 0.847$$
$$\approx 0.442$$

$$SE(\widehat{\beta}_2) = \sqrt{\frac{1}{70} + \frac{1}{60} + \frac{1}{30} + \frac{1}{40}} \approx 0.299$$

(formula for SE of a log(OR))

| | Estimate | Standard Error |
|---|---|---|
| $\widehat{\beta}_1$ | $-0.847$ | 0.218 |
| $\widehat{\beta}_2$ | 0.442 | 0.299 |

*(f) Consider the following model:*

*logitP(low birth weight) = γ1 −γ2x + γ3z − γ4z∗x,*

*where z is an indicator for the first child (z = 1 for the first child, z = 0 for the second child). Interpret γ3 and γ4. (Notice the negative signs in the model equation.)*

$\gamma_3$: Log ratio of odds of LBW in 1$^{st}$ child to odds of LBW in 2$^{nd}$ child in non-hypertensive women.

$\gamma_4$: How much the log odds ratio in non-hypertensive women is higher than the comparable log odds ratio in hypertensive women. I.e., it is the contrast between a non-hypertensive and hypertensive woman (not the other way around).

The above was the soln key's wording, not mine…

*(g) An investigator suggested using all the data (400 births) in PROC LOGISTIC in SAS to fit the above model to obtain estimates, confidence intervals, p-values, etc. Comment on this suggestion.*

It would be problematic to use all of the data because each mother would have two measures (first and second birth). This would violate the model independence assumption because these data would be more similar (correlated) than first/second births for other women. Thus, I would suggest that the investigator control account for this by adding a random effect to the model.