STAT 6200 Homework #7 - Solution

Ch. 15

6. a) $P(\chi^2(2) > 9.21) = 1 - \underbrace{P(\chi^2(2) \le 9.21)}_{= .989998 \text{ from Minitab}} = .010002 \approx .01$

b) $P(\chi^2(2) > 7.38) = 1 - \underbrace{P(\chi^2(2) \le 7.38)}_{= .975028} = .024972 \approx .025$

c) $\chi^2_{1-.10}(2) = \chi^2_{.90}(2) = 4.60517$ (from Minitab)

8. These data are not paired.

a) $$\chi^2 = \frac{n(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)} = \frac{5294\left[(1250)(1666) - (991)(1387)\right]^2}{(2241)(3053)(2637)(2657)}$$

$$= 55.355$$

$p = .000$

b) There ~~does seem~~ is sufficient evidence to conclude that there is an association between drunk driving and year of the survey. An examination of the expected cell counts reveals that a lower proportion of students drive after drinking in 1987 than expected.

$\hat{p}_1$ = sample proportion driving while drinking in '83 = $\frac{1250}{2637} = .4740$

$\hat{p}_2$ = " " " " " " '87 = $\frac{991}{2657} = .3730$

$\hat{p}_2 < \hat{p}_1$ prevalence of drunk driving seems to be going down over time.

c) $$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{.4740 - .3730}{\sqrt{.4233(1-.4233)\left(\frac{1}{2637} + \frac{1}{2657}\right)}}$$

$$\hat{p} = \frac{2241}{5294} = .4233 \qquad = 7.44$$

$$p = .000$$

Tests are equivalent. Reach the same conclusion either way.

d.) $$\hat{p}_1 - \hat{p}_2 \pm 1.96\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = (.0746, .1275)$$

e) It does not contain 0, which is what we would expect because we rejected $H_0: p_1 - p_2 = 0$

10. The Pearson chi-square test is appropriate for this problem, provided that the expected cell counts are adequate. See Minitab project hwk7-10.mpj for this analysis.

From the expected cell counts, we find that none are < 1 and $3/15 = 20\%$ are ≤ 5, so the expected cell counts are (barely) adequate for chi-square analysis.

a) The test statistic & p-value are $X^2 = 28.986$, $p = .000$

$$p = P\left(\chi^2\underbrace{((r-1)(c-1))}_{=8} > 28.986\right) = 1 - P(\chi^2(8) \leq 28.986) = 1 - .9997$$
$$= .0003$$

b.) We conclude that there is an association between specialty + surgical recommendation. Comparing observed + expected cell counts, one of the most striking features is that radiotherapists recommend option C much more frequently and option CR much less frequently than expected.

13. These are ~~pairs~~ matched pairs, and McNemar's test is appropriate here.

a) $$X^2 = \frac{(|O_{12} - O_{21}| - 1)^2}{O_{12} + O_{21}} = \frac{(|12-20| - 1)^2}{12+20} = 1.5313$$

At $\alpha = .05$, the critical value is $\chi^2_{.95}(1) = 3.84$

The p-value is $p = P(\chi^2(1) > 1.5313) = 1 - P(\chi^2(1) \leq 1.5313)$
$= 1 - .7841 = .2159$

b) So, we do not reject H_0. There is insufficient evidence to conclude that there is an association between retirement status and heart disease (cardiac arrest).

c) [illegible]

$$\hat{OR} = \frac{O_{12}}{O_{21}} = \frac{12}{20} = .6$$

d) 95% CI for OR given by (e^L, e^U)

where $L = \ln(\hat{OR}) - 1.96\sqrt{\frac{O_{12}+O_{21}}{O_{12}O_{21}}}$

$$U = \ln(\hat{OR}) + 1.96\sqrt{(O_{12}+O_{21})/(O_{12}O_{21})}$$

d) $L = \ln(.6) - 1.96 \underbrace{\sqrt{\frac{12+20}{12(20)}}}_{=.3651} = -1.2265$

$U = \ln(.6) + 1.96(.3651) = .2049$

$\Rightarrow$ 95% CI for OR $= \left(e^{-1.2265}, e^{.2049}\right) = (.29, 1.23)$

This interval does not contain 1, so we would not reject $H_0: OR = 1$. That is, this result agrees w/ parts a + b: There is insufficient evidence to conclude that there's association between retirement status + heart disease.

15. Again, these data are paired, so McNemar's test is appropriate

a) $X^2 = \frac{(|O_{12} - O_{21}| - 1)^2}{O_{12} + O_{21}} = \frac{(|2-8| - 1)^2}{2+8} = 2.5$

$p = P(\chi^2(1) > 2.5) = 1 - P(\chi^2(1) \leq 2.5) = 1 - .886154$
$= .1138$

b) So, we fail to reject H_0

There is insufficient evidence to conclude that there is an association between occurrence of headaches + level of exposure.
A problem w/ this analysis however, is the small sample size (over).

15(b) The rule of thumb for McNemar's test is $O_{12} + O_{21} \geq 20$. In this case $O_{12} + O_{21} = 2 + 8 = 10 < 20$. Therefore, an exact analysis would be preferred here. I don't expect you to perform the exact analysis. However, I did & do it, and the exact p-value for McNemar's test turns out to be .1094, which is fairly close to the approximate p-value obtained above, & which results in the same conclusion: fail to reject the null hypothesis of no association.

17.

a)

		Pelvic inflammatory disease: Yes	Pelvic inflammatory disease: No	
ectopic preg	Yes	28	251	279
	No	6	273	279
		34	524	558

There's nothing in the description to suggest that these are matched pair data. The equal sample sizes in the case group (ectopic pregnancy = yes) and control group suggests that maybe(!) it was a paired case-control study, but we have no way of knowing this, so we'll assume independent samples.

b) $$\hat{OR} = \frac{(28)(273)}{6(251)} = 5.076$$

The odds of pelvic disease are estimated to be 5.076 times higher for the cases than for the controls.

c) 99% CI for $\ln(OR)$ is

$$\ln(\hat{OR}) \pm z_{1-\frac{.01}{2}} \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$$

$$= \ln(5.076) \pm 2.576 \sqrt{\frac{1}{28}+\frac{1}{251}+\frac{1}{6}+\frac{1}{273}}$$

$$= (.4439, 2.8050)$$

99% CI for OR $= \left(e^{.4439}, e^{2.8050}\right) = (1.56, 16.53)$

Ch. 17

4. No. The Pearson correlation coefficient measures the strength of the linear association between two variables. Two variable may be highly dependent in a non-linear way and still have a population correlation coefficient of 0. See figure 17.2 d on p.401 for an example.

5. See the Minitab project file hwk7-5.mpj

a) See attached plot, Figure1 (in this pdf file)

b) Figure 1 does not show clear evidence of a linear relationship. It does appear that cholesterol level tends to increase as triglycerides increase, but this tendency seems to be driven by just two points. If the last two points in the data set are removed, then there doesn't seem to be much of a linear relationship.

5 c) The sample Pearson correlation coefficient can be computed and d) in Minitab by selecting

Stat → Basic Statistics → Correlation...

and then selecting the variables cholest + triglyc + then hitting OK. The result is $r = .650$. Minitab also gives the p-value for the test of $H_0: \rho = 0$. This p-value is $p = .042$ so at level $\alpha = .05$, we would reject H_0 and conclude that there is a signif. (positive) linear relationship between cholesterol level & triglyceride level.

More details on the computations:

One formula that can be used to compute r is

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

We implement this formula in hwk7-5.mpj. First the original variables are converted to z-scores z-chol and z-triglyc by using Calc → Standardize These results are stored in columns C4 + C5 of the worksheet. Then in column C6, C4 and C5 are multiplied together. This gives $\left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$.

Then the sum of column C6 is taken, which gives

$$\sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) = 5.84689$$

Finally, $r = \frac{1}{n-1}(5.84689) = \frac{1}{9}(5.84689) = .6497$

The test of $H_0: \rho = 0$ has test statistic

$$t = r\sqrt{\frac{n-2}{1-r^2}} = .6497\sqrt{\frac{10-2}{1-(.6497)^2}} = 2.4170$$

which has p-value $p = 2P(t(n-2) > |t|) = 2P(t(8) > 2.4170)$

$$= 2\left(P(t(8) < -2.4170)\right) = 2(.02102) = .042$$

e) r_s can be calculated in the same manner as r, but by operating on the ranks of cholest and triglyc. (f, g) These ranks can be computed in Minitab by selecting

Data → Rank...

These ranks are stored in C8 + C9.

r_s can now be computed by selecting

Stat → Basic Statistics → Correlation...

Then select variables rank-chol and rank-tri (columns C8, C9) and hit O.k. This yields

$r_s = .418$ w/ a p-value for $H_0: \rho_s = 0$ of $p = .229$

Therefore, r_s is smaller than r and we fail to reject $H_0: \rho_s = 0$

In this example, I would be more inclined to use the Spearman correlation because both cholesterol and triglyceride have somewhat skewed distributions, and there is at

least one outlier among the cholesterol values (observation #9). See Figure 2, which displays boxplots of cholest and triglyc.

7. See Minitab project file hwk7-7.mpj.
Since Apgar score is an ordinal variable, it is more appropriate to use Spearman's correlation ~~rather~~ rather than the Pearson correlation here.

a) $r_s = .108$, $p = .283$ for testing $H_0: \rho_s = 0$

(b, c)

See Figure 3. There is a very slight tendency for apgar5 scores to increase w/ sbp, but that positive relationship is not significant according to our test of H_0.

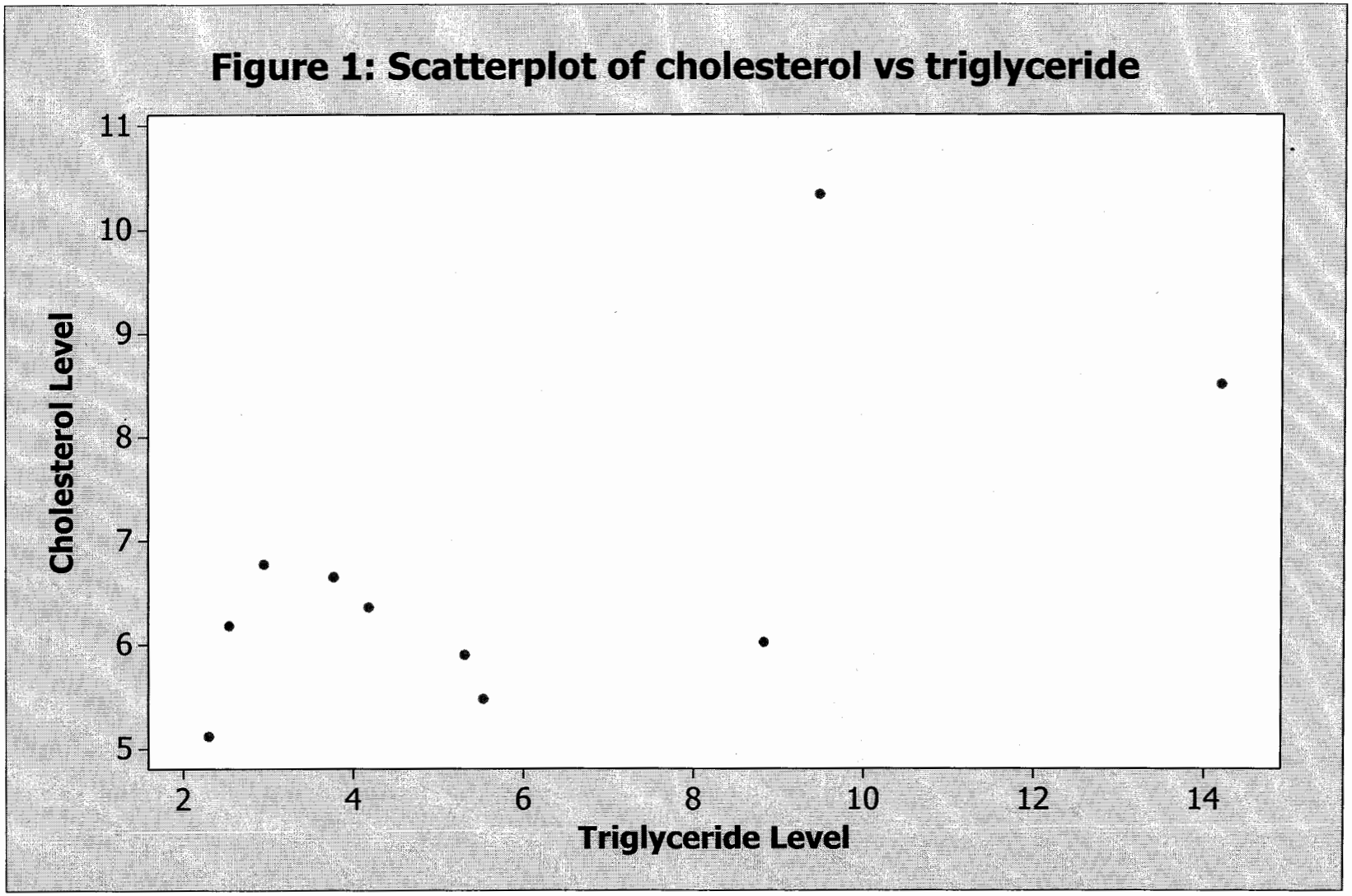
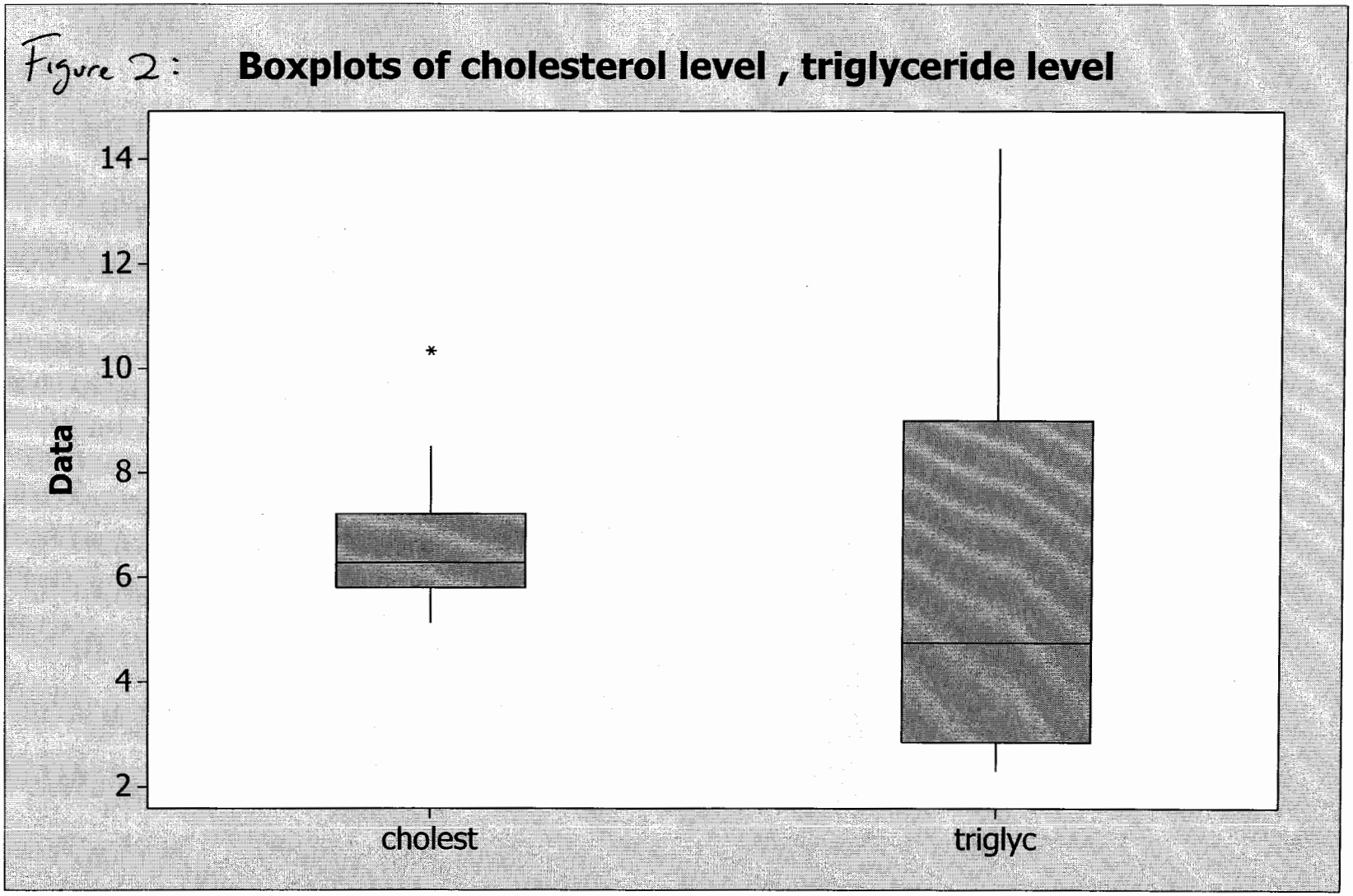


Figure 1: Scatterplot of cholesterol vs triglyceride
Cholesterol Level
11
10
9
8
7
6
5
2
4
6
8
10
12
14
Triglyceride Level

Figure 2: **Boxplots of cholesterol level , triglyceride level**



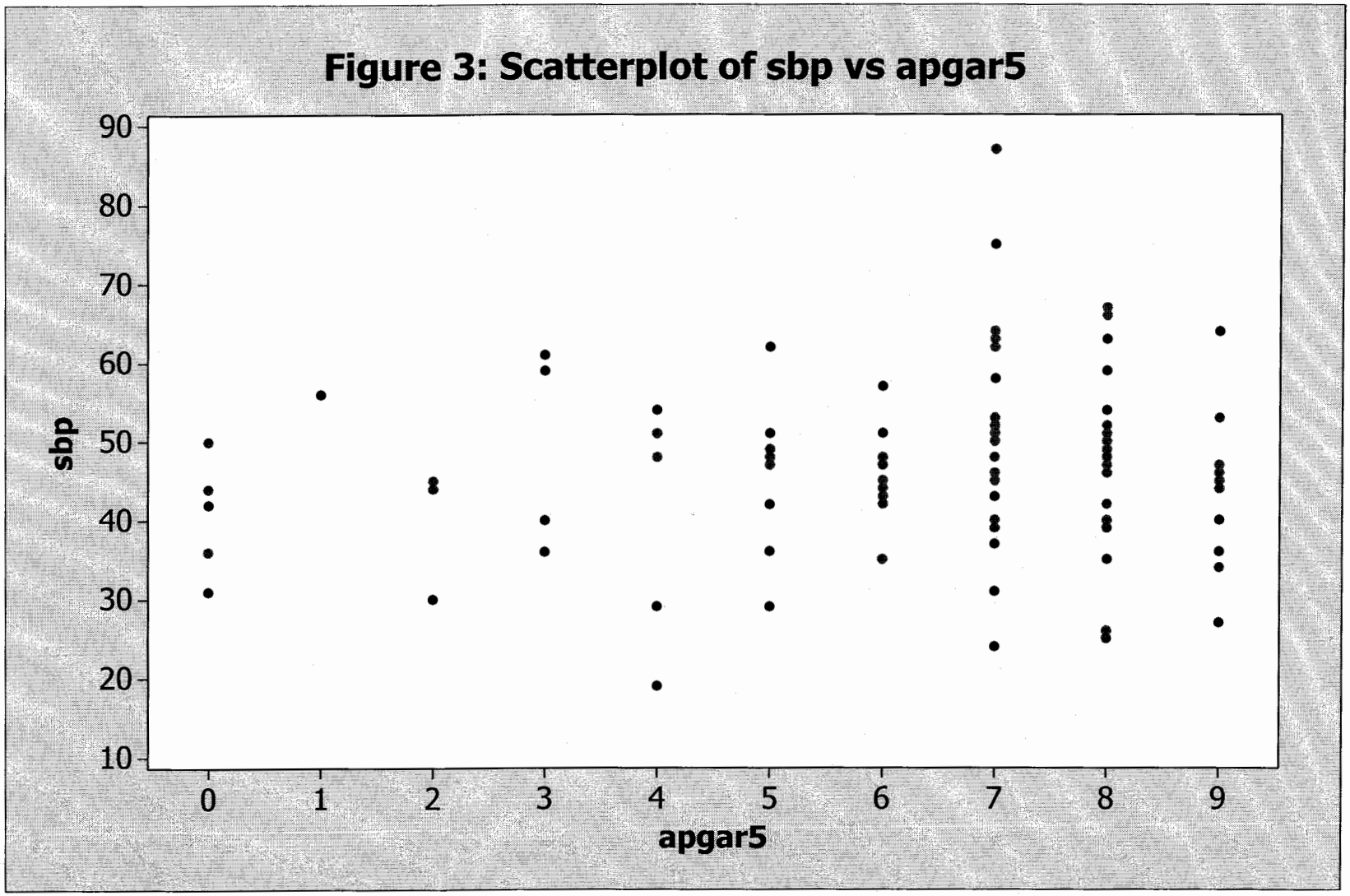


Figure 3: Scatterplot of sbp vs apgar5
sbp
90
80
70
60
50
40
30
20
10
0
1
2
3
4
5
6
7
8
9
apgar5