UCLA STAT 13 Statistical Methods - Final Exam Review Solutions

Chapter 7 – Sampling Distributions of Estimates

1. **(a)** **(i)** $\mu_{\overline{X}} = \mu$ **(ii)** $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}}$

(b) $\overline{X}$ is **exactly** Normally distributed.

(c) **(i)** $\overline{X}$ is **approximately** Normally distributed.

(ii) Central limit theorem.

2. **(a)** **(i)** $\mu_{\hat{P}} = p$ **(ii)** $\sigma_{\hat{P}} = \sqrt{\dfrac{p(1-p)}{n}}$

(b) For large samples $\hat{P}$ is **approximately** Normally distributed.

3. A **parameter** is a numerical characteristic of a population.

4. An **estimate** is a **known** quantity calculated from data in order to estimate an unknown **parameter**.

5. (4)

6. **(a)** $\mu_{\overline{X}}$ = 280 seconds $\sigma_{\overline{X}} = \dfrac{60}{\sqrt{16}}$ = 15 seconds

$\overline{X}$ ~ approximately Normal (μ = 280s, σ = 15s)

(b) pr($\overline{X}$ > 240) = 1 – pr($\overline{X}$ < 240)

= 1 – 0.0038

= 0.9962

7. **(a)** **(i)** $\mu_{\overline{X}} = 7.15$ litres $\sigma_{\overline{X}} = \dfrac{\sigma_X}{\sqrt{n}} = \dfrac{1.2}{\sqrt{1}} = 1.2$ litres

(ii) $\mu_{\overline{X}} = 7.15$ litres $\sigma_{\overline{X}} = \dfrac{\sigma_X}{\sqrt{n}} = \dfrac{1.2}{\sqrt{4}} = 0.6$ litres

(iii) $\mu_{\overline{X}} = 7.15$ litres $\sigma_{\overline{X}} = \dfrac{\sigma_X}{\sqrt{n}} = \dfrac{1.2}{\sqrt{16}} = 0.3$ litres

(b) The standard deviation differs. This is because as the sample size increases there is a decrease in the variability of the sample mean.

8. **(a)** The proportion of university students who belong to the student loan scheme.

(b) $\mu_{\hat{P}} = 0.65$ $\quad\quad$ $\sigma_{\hat{P}} = \sqrt{\dfrac{0.65(1-0.65)}{50}} = 0.0675$

$\hat{P}$ ~ approx Normal (μ = 0.65, σ = 0.0675)

(c) pr($\hat{P}$ > 0.7) = 1 – 0.7707 = 0.2293

(d) pr(0.45 < $\hat{P}$ < 0.55) = pr($\hat{P}$ < 0.55) - pr($\hat{P}$ < 0.45)

= 0.0691 – 0.0015

= 0.0676

9. **(a)** $\bar{x} = 10.125$, $s = 1.9477$

(b) $\bar{x} \pm 2 \times \dfrac{s}{\sqrt{n}}$ = 10.125 ± 2 x $\dfrac{1.9477}{\sqrt{8}}$

= (8.75, 11.50)

(c) **(i)** wider $\quad$ **(ii)** nothing $\quad$ **(iii)** narrower

10. $\hat{p} = \dfrac{36}{120} = 0.3$

$\hat{p} \pm 2 \times \sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$ = 0.3 ± $2 \times \sqrt{\dfrac{0.3 \times 0.7}{120}}$

= (0.216, 0.384)

11. (2)

UCLA STAT 13 Statistical Methods - Final Exam Review Solutions
Chapter 8 – Confidence Intervals

Section A: Confidence intervals for a mean, proportion and difference between means

1. **(a)** $\theta = \mu$, the population mean score for the old Stat 13 exam.

(b) $\hat{\theta} = \bar{x}$ = 38.20, the mean mark of the sample of 30 marks.

(c) se($\hat{\theta}$) = se($\bar{x}$) = $\dfrac{s}{\sqrt{n}} = \dfrac{10.85}{\sqrt{30}} = 1.9809$

(d) df = 30 – 1 = 29

(e) t-multiplier = 2.045

(f) 95% c.i. is: $\bar{x} \pm t \times se(\bar{x})$ = 38.20 ± 2.045 x 1.9809 = 38.20 ± 4.0509 = (34.15, 42.25)

(g) There are many ways of interpreting a confidence interval. Three different ways follow.

(1) With 95% confidence, we estimate that the population mean mark is somewhere between 34.15 and 42.25 marks.

(2) We estimate that the population mean mark is somewhere between 34.15 and 42.25 marks. A statement such as this is correct, on average, 19 times out of every 20 times we take such a sample.

(3) We estimate the population mean mark to be 38.20 with a margin or error of 4.05. A statement such as this is correct, on average, 19 times out of every 20 times such a sample is taken.

(h) We don't know. The population mean mark is not known so we don't know whether this particular 95% confidence interval contains the population mean. However, in the long run, the population mean will be contained in 95% of the 95% confidence intervals calculated from such samples.

2. **(a)** $\theta = p$, the proportion of female Spanish prisoners in 1995 who had tuberculosis.

(b) $\hat{\theta} = \hat{p} = \dfrac{36}{90} = 0.4$, the proportion in the sample of female Spanish prisoners who had tuberculosis.

(c) se($\hat{\theta}$) = se($\hat{p}$) = $\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\dfrac{0.4 \times 0.6}{90}}$ = 0.051640

(d) z-multiplier = 1.96

(e) 95% c.i. is: $\hat{p} \pm t \times se(\hat{p})$ = 0.4 ± 1.96 x 0.051640 = 0.4 ± 0.1012 = (0.299, 0.501)

(f) We estimate that the proportion of female Spanish prisoners in 1995 with tuberculosis is somewhere between 29.9% and 50.1%. A statement such as this is correct, on average, 19 times out of every 20 times we take such a sample.

(g) We don't know. The population proportion is not known so we don't know whether this particular 95% confidence interval contains the population proportion. However, in the long run, the population proportion will be contained in 95% of the 95% confidence intervals calculated from such samples.

3. (a) We estimate that the mean thiol level for people suffering from rheumatoid arthritis is somewhere between 1.08 and 2.01 greater than the mean thiol level for non-sufferers. A statement such as this is correct, on average, 19 times out of every 20 times we take such a sample.

(b) We don't know. The true difference in the thiol levels between the two populations is not known so we don't know whether this particular 95% confidence interval contains the true difference. However, in the long run, the true difference will be contained in 19 out of each batch of 20 confidence intervals calculated from such samples.

4. (4)

5. (2)

6. (3)

7. (3)

8. (5)

9. (4)

Section B: Confidence interval for a difference in proportions

1. (a) Situation (b): *Single sample, several response categories*

(b) Situation (a): *Two independent samples*

(c) Situation (c): *Single sample, two or more Yes/No items*

(d) Situation (a): *Two independent samples*

2. (a) Situation (b): *Single sample, several response categories*

(b) Situation (a): *Two independent samples*

(c) Situation (c): *Single sample, two or more Yes/No items*

(d) Situation (a): *Two independent samples*

(e) Situation (c): *Single sample, two or more Yes/No items*

(f) Situation (b): *Single sample, several response categories*

3. (2) Note: This was Situation (c): *Single sample, two or more Yes/No items* **.**

4. (a) Let p_1 represent the proportion of white prisoners who were infected with TB and p_2 represent the proportion of Gypsy prisoners who were infected with TB.
$\theta = p_1 - p_2$, the true difference in the above proportions.

(b) $\hat{\theta} = \hat{p}_1 - \hat{p}_2 = \frac{496}{886} - \frac{74}{152} = 0.5598 - 0.4868 = 0.0730$, the estimated difference in the above proportions.

(c) $\text{se}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{0.5598(1-0.5598)}{886} + \frac{0.4868(1-0.4868)}{152}} = 0.043837$

(d) $z = 1.96$

(e) 95% c.i. is: $(\hat{p}_1 - \hat{p}_2) \pm z \times \text{se}(\hat{p}_1 - \hat{p}_2)$ = 0.0730 ± 1.96 x 0.043837 = 0.0730 ± 0.08592 = (–0.0129, 0.1589) = (–1.3%, 15.9%)

(f) With 95% confidence, we estimate the proportion of white prisoners who were infected with TB to be somewhere between 1.3% lower and 15.9% higher than the proportion of Gypsy prisoners who were infected with TB.

5. (a) $\theta = p_W - p_G$, the difference in the proportion of prisoners infected with TB who were white and the proportion of prisoners infected with TB who were Gypsy.

(b) $\hat{\theta} = \hat{p}_W - \hat{p}_G = \frac{496}{592} - \frac{74}{592} = 0.8378 - 0.1250 = 0.7128$, the estimated difference in the above proportions.

(c) $\text{se}(\hat{p}_W - \hat{p}_G) = \sqrt{\frac{0.8378 + 0.1250 - (0.8378 - 0.1250)^2}{592}} = 0.027715$

(d) $z = 1.96$

(e) 95% c.i. is: $(\hat{p}_W - \hat{p}_G) \pm z \times se(\hat{p}_W - \hat{p}_G)$ = 0.7128 ± 1.96 x 0.027715 = 0.7128 ± 0.0543 = (0.6585, 0.7671) = (66%, 77%)

(f) We estimate that proportion of prisoners infected with TB who were white is somewhere between 66% and 77% greater than the proportion of prisoners infected with TB who were Gypsy. A statement such as this is correct, on average, 19 times out of every 20 times such a sample is taken.

6. (a) Situation (a). There are two independent samples – a sample of intravenous drug user prisoners and a sample of non-intravenous drug user prisoners.

(b) With 95% confidence, we estimate that proportion of intravenous drug user prisoners who were infected with TB is somewhere between 0.6% less than and 11.5% greater than the proportion of non-intravenous drug user prisoners who were infected with TB.

(c) Yes. Since zero is contained within the 95% confidence interval, zero is a plausible value for true difference between the population proportions.

7. (3)

8. (4)

9. (1)

10. (5)

UCLA STAT 13 Statistical Methods - Final Exam Review Solutions
Chapter 9 – Significance Testing: Using Data to Test Hypotheses

Section A:

1. The null hypothesis is the hypothesis tested by the statistical test. The alternative hypothesis specifies the type of departure from the null hypothesis we expect to detect.

2. **(a)** $H_0: \ \neq \theta_0$ **(b)** $H_0: \ > \theta_0$ **(c)** $H_0: \ < \theta_0$

3. A one-tailed test is used when the investigators have good grounds for believing the true value of θ was on one particular side of θ_0 before the study began. Otherwise, or if in doubt, a two-tailed test is used. Good grounds mean that there is prior information or there is a theory to tell the investigators which way the study will go.

θ θ θ

4. **(a)** $\dfrac{estimate - hypothsised\ value}{std\ \ error}$ **(b)** $\dfrac{\hat{\ } - \theta_0}{\mathrm{se}(\hat{\theta})}$

5. The *P-value* is the **probability** that, if the **null hypothesis** was true, **sampling variation** would produce an estimate that is **at least as far away** from the **hypothesised value** than our **data estimate**.

θ

6. The *P-value* measures the strength of evidence against the null hypothesis.

7.

P-value	**Evidence against H_0**
> 0.12	**none**
≈ 0.10	**weak**
≈ 0.05	**some**
≈ 0.01	**strong**
< 0.01	**very strong**

8. Nothing.

9. A confidence interval.

10. One possible value for the parameter, called the hypothesised value, is tested. The test determines the strength of evidence provided by the data against the proposition that the hypothesised value is the true value.

11. If the *P-value* is less than or equal to a specified value (usually 5% or 1%), the effect that was tested is said to be significant at that specified level (usually 5% or 1%). Therefore a significant test reveals that there is sufficient evidence against the null hypothesis.

Section B:

1. (a) Let p_W be the true proportion of white prisoners who were infected with TB and p_G be the true proportion of Gypsy prisoners who were infected with TB. Thus $\theta = p_W - p_G$.

(b) H_0: p_W - p_G = 0 vs H_1: p_W - $p_G \neq 0$

(c) $\hat{p}_W - \hat{p}_G = \frac{496}{886} - \frac{74}{152} = 0.5598 - 0.4868 = 0.0730$

(d) $\text{se}(\hat{p}_W - \hat{p}_G) = \sqrt{\frac{0.5598(1-0.5598)}{886} + \frac{0.4868(1-0.4868)}{152}} = 0.04384$

$z_0 = \frac{0.0730 - 0}{0.04384} = 1.665$

(e) *P-value* = 2 x pr(Z > 1.665)

= between 0.05 and 0.1 (in fact it is just less than 0.1)

(f) We have weak evidence against H_0.

(g) There is weak evidence that there is a difference between the proportion of White prisoners who had TB and the proportion of Gypsy prisoners who had TB.

(h) 95% confidence interval for $p_W - p_G$:

0.0730 ± 1.96 x 0.04384 = (-0.013, 0.159)

(i) With 95% confidence, we estimate that the proportion of White prisoners who had TB is somewhere between 1.3% lower than and 15.9% higher than the proportion of Gypsy prisoners who had TB.

2. (a) p, the population proportion (ie the proportion of residents of New York City who would have said that they would move somewhere else).

(b) H_0: p = 0.5 vs H_1: $p \neq 0.5$

(c) $\hat{p} = \frac{595}{1009} = 0.589693$

(d) z_0 = 5.70 *P-value* = 0.000

(e) There is very strong evidence that the true proportion of New York City residents who would have said that they would move somewhere else is greater than 50%.

(f) We estimate that the proportion of New York City residents who would have said that they would move somewhere else is somewhere between 55.9% and 62.0%. A statement such as this is correct, on average, 19 out of every 20 times such a sample is taken.

3. (a) Let μ_1 be the true mean daily revenue for laundry 1 and μ_2 be the true mean daily revenue for laundry 2. Thus the parameter used is $\mu_1 - \mu_2$, the difference in the mean daily revenue for the two laundries.

(b) H_0: μ_1 - μ_2 = 0 vs H_1: μ_1 - $\mu_2 \neq 0$

(c) $\bar{x}_1 - \bar{x}_2 = 635.4 - 601.6 = 33.8$

(d) $t_0 = 1.94$

(e) *P-value* = 0.057. We have some evidence that the mean daily revenue of the first laundry is greater than the mean daily revenue of the second laundry.

(f) With 95% confidence, we estimate that the mean daily revenue of the first laundry is somewhere between \$1 less than and \$69 more than the mean daily revenue of the second laundry.

(g) There are no reasons for doubting the validity of the results of this analysis because neither stem-and-leaf plot shows any non-Normal features.

(h) The computer uses a different formula for calculating *df*. This formula gives a larger value of *df* than the hand calculation based on the minimum of one less than each sample size.

Section C:

1. (1)

2. (1)

3. (4)

UCLA STAT 13 Statistical Methods - Final Exam Review Solutions
Chapter 10 – Data on a Continuous Variable

1. **(a)** A *t*-test on the differences. A pair of observations is made on the same subject so this is paired comparison data. A *t*-test on the differences is more appropriate.

(b) H_0: $\mu_{\text{Diff}} = 0$ vs H_1: $\mu_{\text{Diff}} > 0$.
P-value = 0.016
We have strong evidence against the null hypothesis of no difference between current and previous spending. It appears that access to the cable network increases spending in viewers. With 95% confidence, we estimate that viewers spent, on average, between \$3.10 and \$62.50 more when they had access to the cable network.
Note: A one-tailed test was used because if departure from H_0 does occur we expect an increase in sales.

(c) Since a *t*-test on the differences is used we look at the dot plot, Normal probability plot and *W*-test on the differences. The dot plot shows moderate positive skewness but the *t*-test is robust to such departures from Normality. The *W*-test for Normality has a *P-value* greater than 10% so it is believable that the data has an underlying Normal distribution. The results of the *t*-test should be valid in this situation.

(d) **Hypotheses:** $H_0 : \tilde{\mu}_{Diff} = 0$ vs $H_1 : \tilde{\mu}_{Diff} > 0$
Signs of the differences: + - + + - 0 + - + + - + - + + +
10 positive signs, 5 negative signs, 1 zero.
Since this is a one-tailed (>) test, evidence against the null hypothesis will be provided by a large number of observations above the hypothesised median of zero (and therefore a small number of observations below zero).
P-value = pr($Y \geq 10$) by considering positive signs, or
pr($Y \leq 5$) by considering negative signs, where Y ~ Binomial ($n = 15$, $p = 0.5$)
= 0.151
Interpretation: We have no evidence against the null hypothesis. We have no evidence to suggest that the median spending increases when the viewers have access to the cable network.

(e) The *t*-test is more appropriate. When the assumptions of the *t*-test are met reasonably well choose a *t*-test in preference to a nonparametric test. This is because a *t*-test will more readily detect departures from the null hypothesis when these departures do exist (That is, the *t*-test is more powerful than the Sign test).
Note that the *t*-test gave a *P-value* of 1.6% while the Sign test gave a *P-value* of 15.1%.

2. **(a)** The dot plot shows the presence of a possible outlier. The t-test is sensitive to the presence of outliers, especially with a small sample size of 11. Because the Normality assumption is not satisfied a nonparametric test should be used.

(b) $H_0 : \tilde{\mu} = 1.8$ vs $H_1 : \tilde{\mu} < 1.8$.

(c) *P-value* = 0.1719

(d) We have no evidence against the null hypothesis. We have no evidence that the median operating time for hedge trimmers was less than 1.8 hours.

(e) A high number of observations below the hypothesised median (and therefore a small number of observations above the hypothesised median) provide evidence against the null hypothesis.
P-value = pr($Y \leq 3$) by considering positive signs, or
pr($Y \geq 7$) by considering negative signs, where Y ~ Binomial ($n = 10$, $p = 0.5$).

Section B: More Than Two Independent Samples

1. **(a)** Analysis of variance (ANOVA).

(b) When we want to investigate differences between the underlying means of more than two groups.

(c) H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ (The population means are all equal.)
H_1: At least one of the population means is different from the other three.

(d) **(i)** Assumptions:
(I) The samples are independent of each other.
(II) The underlying distribution of each group is Normally distributed.
(III) The population standard deviations of each group are equal.

(ii) Checking assumptions:
(I) Ensure in the design of the experiment or study.
(II) By plotting the data. The choice of plot will depend on the sample sizes.
(III) By plotting the data and/or looking at the sample standard deviations.

2. **(a)** $f_0 = \dfrac{s_B^2}{s_W^2}$

(b) s_B^2 is called the between mean sum of squares and it measures the variability between the sample means.
s_W^2 is called the within mean sum of squares and it measures the average variability within the samples (that is, the internal variability within the samples themselves).

3. **(a)** $k = 6$, $n_{tot} = 16$

(b) $k = 9$, $n_{tot} = 153$

(c) $k = 5$, $n_{tot} = 30$

4. (a) There are three independent random samples.
There are signs of moderate positive skewness in all three of the samples. The Neither group also shows signs of three clusters, indicating possible multi-modality. However with the equal sample sizes and the moderate size of the three samples this should not cause any concern with the validity of the *F*-test.
The assumption of equality of the standard deviations of each underlying distribution is suspect because the sample standard deviation of the Neither group is larger than the sample standard deviations of the Drug and Placebo groups. However this difference is not large enough to cause concern with the assumption because the *F*-test is quite robust with respect to this assumption. Note that the ratio of the largest sample standard deviation to the smallest standard deviation is less than 2 ($\frac{23.14}{15.22} = 1.52$).

(b) ANOVA table:

	DF	SS	MS	F	P
Group	**2**	3330.427	**1665.21**	**4.8213**	0.011
Error	**72**	**24867.936**	345.388		
Total	**74**	**28198.363**			

(c) s_B^2 = 1665.21 s_W^2 = 345.888

(d) H_0: $\mu_{\text{Drug}} = \mu_{\text{Neither}} = \mu_{\text{Placebo}}$, where μ_{Drug}, μ_{Neither} and μ_{Placebo} are the mean number of minutes for patients to fall asleep in the Drug, Neither and Placebo groups respectively. That is, the mean number of minutes to fall asleep is the same for each of the three groups.
H_1: At least one of the three underlying means is different from the other two.
The *P-value* of 0.011 means that we have strong evidence against the null hypothesis. We have strong evidence that at least one of the groups has an different underlying mean number of minutes for patients to fall asleep.

(e) The confidence interval for the difference between the means of the **Drug** group and the **Neither** group does not contain zero. At the **95%** confidence level, it is believable that there is **a** real difference between the means of these two groups.

(f) (i) The family error rate of 5% means that in 5% of the times the test is done, at least one of the difference confidence intervals **will not** contain the true mean difference.
(ii) The family error rate of 5% means that in 95% of the times the test is done, each of the difference confidence intervals **will** contain its true mean difference.

(g) The individual error rate of 1.95% means that in **98.05**% of the times the test is done, the confidence interval for μ_{Drug} - μ_{Placebo} will contain the true mean difference.

Section C: Experimental Design

1. We should not use a paired design. A paired design would require measuring the number of heart attacks over two different time periods. In this study the time periods are long (10 years), which would make a paired comparison design impractical.

The most appropriate design for this experiment would be a two independent sample design, with one of the samples acting as a control group by receiving aspirin. A sample of 9500 subjects should be randomly selected from the 19,000 available and these should be given a regular dosage of aspirin. The remaining subjects should be given a regular dosage of the new drug. The subjects are chosen randomly, in an attempt to make the two groups differ only by the type of drug given to the subjects. The experiment should also be double blind to prevent any possible biases by the subjects or the people administering the treatments.

The hypotheses we would be interested in testing would be:
$H_0 : p_A - p_N = 0$ versus $H_1 : p_A - p_N \neq 0$ where p_A is the proportion of deaths due to heart attacks for the subjects given a regular dosage of aspirin and p_N is the proportion of deaths due to heart attacks for the subjects given a regular dosage of the new drug.

With such large samples there will be no concern about the Normality assumption for this *t*-test for two proportions.

2. The most appropriate design would be to use five samples, one for each spray and a control group, and use one-way analysis of variance. The trees should be randomly assigned into five groups of 16. Each of the four sprays should be randomly allocated to one of these groups, with the remaining group being left untreated.

The hypotheses we would be interested in testing would be:
H_0: The underlying mean number of bug infested leaves is the same for each group.
H_1: At least one group of trees has a different underlying mean number of bug infested leaves.

The assumptions are:
The five samples are independent.
The underlying distributions are Normally distributed.
The population standard deviations of each group are all equal.

3. The most appropriate design for this experiment would be a two independent sample design, with one sample filling in the first questionnaire and the other filling in the second questionnaire. The 50 subjects should be randomly allocated to the samples in an attempt to make the two samples differ only by the type of questionnaire given to the subjects. Interviewers should each interview 5 subjects from each group to allow for any possible bias due to interviewer effect. Also, subjects should be randomly allocated to interviewers.

   The hypotheses we would be interested in testing would be:
   $H_0: \mu_1 - \mu_2 = 0$ versus $H_1: \mu_1 - \mu_2 \neq 0$, where μ_1 is the mean time taken to fill in questionnaire 1 and μ_2 is the mean time taken to fill in questionnaire 2.

   The assumption that needs to be satisfied is that the underlying distributions are Normally distributed.

4. The most appropriate design for this experiment would be a paired design in order to attempt to reduce the variability between rat hearts due to age and condition. If a two independent sample design was used the variability within in each sample might overshadow any difference between the means. We have 20 hearts available (each sliced in two). For each heart, half of the heart can be used for one dosage of the digitalis treatment, and the remaining half used for the other dosage of the digitalis treatment.

   The hypotheses we would be interested in testing would be:
   $H_0: \mu_{Diff} = 0$ versus $H_1: \mu_{Diff} \neq 0$, where μ_{Diff} is the mean difference in the strength of muscle contractions between the pieces of heart treated with differing dosages of digitalis.

   The assumption that needs to be satisfied is that the distribution of the differences is Normally distributed.

Section D: Quiz

1. **(a)** Plot the data.
   (b) It is the quickest way to see what the data say.
   It often reveals interesting features that were not expected.
   It helps prevent inappropriate analyses and unfounded conclusions.
   Plots also have a central role in checking the assumptions made by formal methods.
2. All observations in the sample are independent and are from the same distribution.
3. **(a)** (i) and (iv)
   (b) (iv)
   (c) (i) [in that the underlying distribution of the differences is assumed to be Normal] and (iii)
   (d) (i), (ii) and (iv)
   (e) (i), (ii), (iv) and (v).
4. Try to check the original sources of the data. Any observations that you know to be mistakes should be corrected or removed. If in doubt, do the analyses with and without the outliers to see whether you come to the same conclusions.
5. They are less sensitive to outliers and do not assume any particular underlying distribution for the observations.
6. Nonparametric tests tend to be less effective at detecting departures from a null hypothesis and tend to give wider confidence intervals than parametric tests.

UCLA STAT 13 Statistical Methods - Final Exam Review Solutions
Chapter 11 – Tables of Counts

Section A: One-way Tables

1. **(a)** H_0: Relative market shares are the same as they were prior to Channel A altering its programming, ie $p_A = 0.1$, $p_B = 0.4$, $p_C = 0.5$.
   H_1: Relative market shares differ since Channel A altered its programming, ie the proportions are not $p_A = 0.1$, $p_B = 0.4$, $p_C = 0.5$.

   (b) Expected count for Channel A = 0.1 x 300 = 30
   Expected count for Channel B = 0.4 x 300 = 120
   Expected count for Channel C = 0.5 x 300 = 150

   (c) Cell contribution = $\dfrac{(125-120)^2}{120}$ = 0.2083

   (d) **(i)** *P-value* = pr($X^2 \geq 5.0417$) = 1 – 0.9196 = 0.0804
   (ii) There is weak evidence that the altered programming for Channel A has affected relative market shares.

2. **(a)** H_0: The proportions of the types are $p_{BC} = \frac{9}{16}$, $p_{Bc} = \frac{3}{16}$, $p_{bC} = \frac{3}{16}$, $p_{bc} = \frac{1}{16}$.
   H_1: The proportions of the types are not $p_{BC} = \frac{9}{16}$, $p_{Bc} = \frac{3}{16}$, $p_{bC} = \frac{3}{16}$, $p_{bc} = \frac{1}{16}$.

   (b) Expected count for type BC = $\frac{9}{16}$ x 160 = 90

   (c) Cell contribution = $\dfrac{(16-30)^2}{30}$ = 6.5333

   (d) **(i)** *P-value* = pr($X^2 \geq 9.8667$) = 1 – 0.9803 = 0.0197
   (ii) There is strong evidence that the observed frequencies are different from the expected frequencies.

Section B: Two-way Tables

1. **(a)** H_0: A person's primary source of news is independent of their age.
H_1: There is a relationship between a person's primary source of news and their age.

(b) **(i)** Expected count for the (Under 30, Radio) cell = $\frac{225 \times 250}{1000}$ = 56.25

(ii) Expected count for the (30-49, TV) cell = $\frac{575 \times 480}{1000}$ = 276

(c) Cell contribution = $\frac{(100 - 51.625)^2}{51.625}$ = 45.330

(d) The *P-value* = 0.000 to 3 decimal places.
We have very strong evidence to suggest that there is a relationship between a person's age and their primary news source.

(e) Yes. We could consider the samples of people under 30, people in the 30-49 age group and the people in the 50 and over age group as three independent sub-samples and carry out a Chi-square test of homogeneity with the primary news source as the response factor.

2. **(a)** H_0: The distribution of the opinions is the same for each group.
H_1: The distribution of the opinions is different for at least one group.

(b) **(i)** Expected count for the (Increase government expenditure, Economists) cell
= $\frac{64 \times 100}{300}$ = 21.33

(ii) Expected count for the (Other business incentives, Business executives) cell
= $\frac{56 \times 100}{300}$ = 18.67

(c) **(i)** Cell contribution = 0.005

(ii) (Increase government expenditure, Government officials) – more than expected
(Increase government expenditure, Business executives) – fewer than expected
(Other business incentives, Business executives) – more than expected

(iii) The *P-value* = 0.000 to 3 decimal places.
There is extremely strong evidence that the opinions of the three groups differ.

(d) No. This data has been collected as three independent samples and a Chi-square test for independence requires that the data is collected as one random sample.

3. (2)

4. (4)

5. (3)

UCLA STAT 13 Statistical Methods - Final Exam Review Solutions

Chapter 12 – Relationships between Quantitative Variables: Regression and Correlation

Section A: The Straight Line Graph

1. **(a)** $\beta_0 = 5$, $\beta_1 = 3$

(b) $\beta_0 = 10$, $\beta_1 = -14$

2. **(a)** $y = -3 + 2x$

(b) **(i)** 2

(ii) 12

Section B: Regression

1. **(a)** $\hat{y} = 11.238 + 1.309x$

(b) Predicted lung capacity = 11.238 + 1.309 x 30 = 50.5

(c) Predicted lung capacity = 11.238 + 1.309 x 25 = 44.0
Residual = Observed value – predicted value = 55 – 44.0 = 11

(d) 'Years smoking' is used to predict lung capacity.
'Years smoking' is a quantitative variable and 'Lung capacity' is continuous and random.
There is a possible linear trend but the observations (28, 30) and (33, 35) are possible outliers which cause concern with the appropriateness of the model.
The residuals versus 'Years smoking' plot along with the *P-value* for the *W*-test for Normality indicates some concern with the assumption that the errors are Normally distributed.

(e) H_0: $\beta_1 = 0$
H_1: $\beta_1 \neq 0$
P-value = 0.0086
There is strong evidence of a linear relationship between years of smoking and lung capacity.
With 95% confidence, we estimate that for every additional year of smoking an emphysema patient's lung capacity increases by between 0.44 and 2.18 units.

(f) **(i)** $r = 0.774$

(ii) *Excel* calls it Multiple R.

2. **(a)** For $x = 1.46$, $\hat{y}$ = -29.86 + 37.72 x 1.46 = 25.2
Residual = Observed value – predicted value = 11.6 – 25.2 = –13.6

(b) The lactic acid concentration is used to predict the taste score.
The lactic acid concentration is quantitative, and the taste score is continuous and random.
The scatter plot shows a linear trend with scatter about that trend.
From the plot of residuals versus lactic acid concentration there is no concern with the assumption that the errors are Normally distributed with mean 0 and with the same standard deviation for each value of *X*.

(c) H_0: $\beta_1 = 0$
H_1: $\beta_1 \neq 0$
P-value = 0.000
There is strong evidence of a linear relationship between lactic acid concentration and taste score.
95% confidence interval for β_1 is:
$37.720 \pm 2.048 \times 7.186 = (23.0, 52.4)$
With 95% confidence, we estimate that for every increase of one unit in the lactic acid concentration the taste score increases by between 23.0 and 52.4 units.

(d) **(i)** We predict that, on average, cheddar cheese with a lactic acid concentration of 1.8 will have a taste score of 38.04.

(ii) With 95% confidence, we estimate that the mean taste score for cheddar cheese with a lactic acid concentration of 1.8 will be somewhere between 31.2 and 44.9.

(iii) With 95% confidence, we predict that the next piece of cheddar cheese with a lactic acid concentration of 1.8 will be somewhere between 13.0 and 63.1.

(e) Estimated slope = 37.72
Estimated increase in taste score for a 1 unit change in lactic acid concentration is 37.72.
Estimated increase in taste score for a 0.05 unit change in lactic acid concentration is $0.05 \times 37.72 = 1.886$.

(1) is the correct response.

(f) (2)

Section C:

1. (4)
2. (2)
3. (5)
4. (5)
5. (2)
6. (3)
7. (1)
8. (3)
9. (5)
10. (3)
11. (3)
12. (5)