UCLA STAT 110A - Applied Statistics - Review Solutions
Sampling Distributions of Estimates, CLT

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

$\hat{P}$ ~ approx Normal ($\mu = 0.65$, $\sigma = 0.0675$)

(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) $\bar{x} \pm 2 \times \frac{s}{\sqrt{n}} = 10.125 \pm 2 \text{ x } \frac{1.9477}{\sqrt{8}}$

= (8.75, 11.50)

(c) **(i)** wider **(ii)** nothing **(iii)** narrower

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

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

= (0.216, 0.384)

11. (2)

UCLA STAT 110A Applied Statistics - Review Solutions
Confidence Intervals

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

1. **(a)** $\theta = \mu$, the population mean mark for the 1995 528.188 exam.

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

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

(d) $df = 30 - 1 = 29$

(e) t-multiplier = 2.045

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

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

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

(d) z-multiplier = 1.96

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

(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 110A Applied Statistics - Review Solutions
Hypothesis Testing

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)** $\dfrac{\hat{} - \theta_0}{\text{se}(\hat{\theta})}$

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: $\mu_1 - \mu_2 = 0$ vs H_1: $\mu_1 - \mu_2 \neq 0$

(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 110A - Applied Statistics - Review Solutions
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)