Questions 39 to **41** refer to the following information.

Death Penalty Survey Results

"Should convicted murderers be put to death?"

	Australia	N.Z.
Yes	46%	42%
No	39%	41%
Can't Say	15%	17%

est = $\hat{p} = 0.42$

[Polls of 1307 Australians & 1010 New Zealanders]

39. Based on previous studies, a researcher believes that the proportion of New Zealanders who agree that convicted murderers should be put to death would be more than forty percent. The hypotheses for this test would be:

Step type 2!

(1) H_0: $p = 0.40$; H_1: $p \neq 0.40$

(2) H_0: $p < 0.40$; H_1: $p = 0.40$

(3) H_0: $p = 0.40$; H_1: $p > 0.40$

(4) H_0: $p < 0.40$; H_1: $p \neq 0.40$

(5) H_0: $p < 0.40$; H_1: $p > 0.40$

$H_0: p = 0.40$

$H_1: p > 0.40$

1-sided test!

step 1 step 2

40. Let's assume instead that the researcher tested H_0: $p = 0.40$ versus H_1: $p \neq 0.40$ where p = the proportion of New Zealanders who agree that convicted murderers should be put to death. If the standard error, se($\hat{p}$) = 0.0155, then the value of the *t*-test statistic, t_0, and the degrees of freedom, *df*, to be used to calculate the *P-value* are given by:

step 3

2-sided test!

(1) $t_0 = 1.290$, $df = \infty$

(2) $t_0 = 1.290$, $df = \cancel{1009}$

(3) $t_0 = -1.290$, $df = \infty$

~~(4)~~ $t_0 = 1.290$, $df = \cancel{1.96}$

~~(5)~~ $t_0 = -1.290$, $df = \cancel{1009}$

$t_0 = \dfrac{\text{est} - \text{hyp val}}{\text{std err}} = \dfrac{\hat{p} - p_0}{se(\hat{p})} = \dfrac{.42 - .4}{.0155} = 1.290$ (3dp)

A difference considered by the researcher was between the proportion of Australians (p_{oz}) supporting the *death penalty for convicted murderers* and the proportion of New Zealanders supporting the *death penalty for convicted murderers*. (p_{NZ})

41. To test for a difference in the two proportions given above the hypotheses would be:

(1) $H_0 : \cancel{\mu_1 - \mu_2} = 0$ vs $H_1 : \cancel{\mu_1 - \mu_2} \cancel{>} 0$

(2) $H_0 : \hat{p}_1 - \hat{p}_2 = 0$ vs $H_1 : \hat{p}_1 - \hat{p}_2 \neq 0$

(3) $H_0 : p_1 - p_2 = 0$ vs $H_1 : p_1 - p_2 \neq 0$

(4) $H_0 : \hat{p}_1 - \hat{p}_2 \cancel{>} 0$ vs $H_1 : \hat{p}_1 - \hat{p}_2 \cancel{>} 0$

(5) $H_0 : p_1 - p_2 \cancel{>} 0$ vs $H_1 : p_1 - p_2 \cancel{=} 0$

$H_0: p_{oz} - p_{NZ} = 0$

$H_1: p_{oz} - p_{NZ} \neq 0$

42. The general formula for a Normality-based confidence interval for the difference between two proportions is:

$$\hat{p}_1 - \hat{p}_2 \pm t \times se(\hat{p}_1 - \hat{p}_2)$$

Which **one** of the following statements about Normality-based confidence intervals for the difference between two proportions is **false**?

T (1) The value of the t-multiplier depends on the confidence level.
T (2) In the long run, if we repeatedly take samples and calculate a 95% confidence interval from each sample, we expect that 95% of the intervals will contain the true value of $p_1 - p_2$.
T (3) The confidence interval is centred on $\hat{p}_1 - \hat{p}_2$.
F (4) The value of the t-multiplier depends on the sample size. df = ∞ so t = 1.96 for 95% CI!
T (5) The size of the standard error depends on the sampling situation.

a, b or c

Questions 43 to 47 refer to the following information.

A survey of 2171 men and 2412 women in Auckland in the early 1990s found that 10% of men abstained from drinking alcohol compared with 16% of women.

difference!

We wish to compare the proportion of female abstainers, p_{female}, with the proportion of male abstainers, p_{male}.

Story type 4. sit (a)!

43. The sampling situation is **best** described as:

(1) two independent samples. a
~~(2)~~ one sample, several response categories. b
~~(3)~~ one sample, many yes/no items. c
(4)? two samples, several response categories. a/b?
(5)? two samples, many yes/no items. a/c?

rubbish!

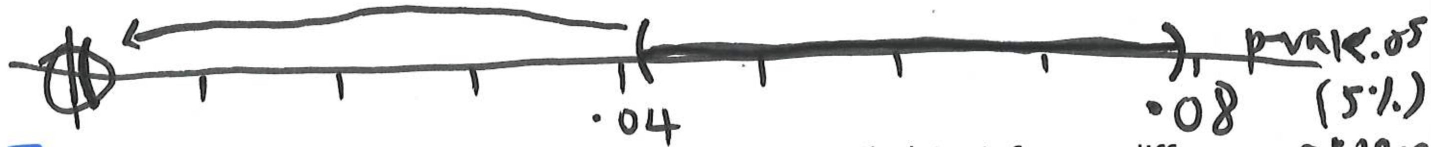
$n_f = 2412$

$n_m = 2171$

zero out of CI ∴ $p_f > p_m$

44. Based on the data, a 95% confidence interval for $p_{female} - p_{male}$ is (0.041, 0.079). Which **one** of the following statements is **false**?

T (1) Based on the data, a 99% confidence interval would be wider than 0.038. width of CI = .079 − .041 = .038
T (2) The point estimate of $p_{female} - p_{male}$ is 0.06. $\hat{p}_f - \hat{p}_m$ = .16 − .10 = .06
T (3) We are confident that the proportion of female abstainers is larger than the proportion of male abstainers.
F (4) Zero is not a plausible value for $p_{female} - p_{male}$.
T (5) Based on the data, a 95% confidence interval for $p_{male} - p_{female}$ is (−0.079, −0.041).



45. Consider the *P-value* associated with a two-tailed test for no difference between p_{female} and p_{male}. Based on the confidence interval in Question **13**, which **one** of the following statements is **true**? as zero is out of CI, by a long way.

(.041, .079)

(1) The *P-value* is much less than 5%.
~~(2)~~ The *P-value* is around 10%.
~~(3)~~ We do not have enough information to determine the approximate *P-value*.
~~(4)~~ The *P-value* is greater than 5%. ✗ zero just out of CI.
(5) The *P-value* is just below 5%. → (.0041, .079)

Questions 46 and 47 refer to the following **additional** information.

Overall, 13% of the 4583 people surveyed abstained from alcohol. We are interested in p_{abstain}, the proportion of people who abstain from alcohol. Step 1!

Story type 2...

A *t*-test of the hypotheses:

$$H_0 : p_{\text{abstain}} = 0.1$$

$$H_1 : p_{\text{abstain}} \neq 0.1$$

gives a test statistic of 6.04 and a *P-value* of 0.000. → could be .0001 or .00001 or .0001592... or...

46. Which **one** of the following statements is **false**?

T (1) The test is significant at the 1% level of significance. → p-val < .01
T (2) If the null hypothesis is true, it is extremely unlikely that sampling variability would give values further away from the hypothesised value, 0.1, than our sample estimate.
T (3) The sample estimate, $\hat{p}_{abstain}$, is approximately 6 standard errors above the hypothesised value, 0.1.
F (4) If the null hypothesis is true, sampling variability ~~could never~~ give values further away from the hypothesised value, 0.1, than our sample estimate.
T (5) The hypothesised value, 0.1, would be outside a 99% confidence interval for p_{abstain}. p-val < .01

47. Which **one** of the following statements gives the **best** interpretation of the hypothesis test result? p-val = .000 (3dp)

~~(1)~~ There is ~~some~~ evidence that the true population proportion, p_{abstain}, is **not** 0.1.
(2) There is very strong evidence that the true population proportion, p_{abstain}, is **not** 0.1.
~~(3)~~ There is ~~no~~ evidence that the true population proportion, p_{abstain}, is **not** 0.1.
~~(4)~~ There is very strong evidence that the true ~~sample~~ proportion, ~~$\hat{p}_{abstain}$~~, is **not** 0.1.
(5) There is very strong evidence that the true population proportion, p_{abstain}, is 0.1.
not

Questions 48 to 50 refer to the following information.

In 2008, as part of the first World Internet Project New Zealand survey, the Institute of Culture, Discourse and Communication published *The Internet in New Zealand 2007 Final Report*.

A random sample of 1430 people aged 12 and over were surveyed via telephone and were asked many questions in order to ascertain New Zealanders' usage of, and attitudes towards, the Internet in 2007.

One of the questions asked:

> *'How important is the Internet in your daily life — important, neutral, not important?'*

The respondents were also categorised by ethnicity. Table 2 shows the response to this question.

	Importance of the Internet			
Ethnicity	Important	Neutral	Not important	**Total**
Pakeha	476	137	302	915
Maori	46	25	44	115
Pasifika	50	26	10	86
Asian	130	16	11	157
Other	77	33	47	157
Total	779	237	414	1430

Table 2: Importance of the Internet in daily life

Assume these 1430 respondents form a random sample from the population of all New Zealanders aged 12 and over.

Let:

$p_{Pasifika}$ p1 be the true proportion of Pasifika New Zealanders aged 12 and over who think that the Internet is important in their daily life

and

p_{Pakeha} p2 be the true proportion of Pakeha New Zealanders aged 12 and over who think that the Internet is important in their daily life.

Step 4 !

48. An estimate of $p_{Pasifika}$ - p_{Pakeha}, rounded to two decimal places, is:

(1) 0.30	(4) −0.58
(2) 0.06	(5) −0.06
(3) −0.30	

$\hat{p}_1 - \hat{p}_2 = 50/86 - 476/915$

49. The sampling situation associated with $se(\hat{p}_{Pasifika} - \hat{p}_{Pakeha})$ is best described as:

sit (a)!

(1) one single sample of size 779, several response categories.

(2) two independent samples, of sizes 86 and 915. $n_1 = 86$

(3) one single sample of size 779, many yes/no items. $n_2 = 915$

(4) one single sample of size 1001, several response categories.

(5) two independent samples, of sizes 50 and 476.

zero in CI

50. A 95% confidence interval for $p_{Pasifika}$ - p_{Pakeha} is (−0.0480, 0.1704). Which one of the following statements is **true**? so $p_1 \approx p_2$

F (1) It would not be surprising to see a different sample of 1430 New Zealanders aged 12 and over produce a poll result with $\hat{p}_{Pasifika}$ smaller than $\hat{p}_{Pakeha}$.

F (2) The difference between $\hat{p}_{Pasifika}$ - $\hat{p}_{Pakeha}$ is not significant at the 5% level of significance. 0 in CI ∴ p-val > .05

T (3) The estimated difference between $p_{Pasifika}$ - p_{Pakeha} could just be sampling error.

F (4) The margin of error for this 95% confidence interval for $p_{Pasifika}$ - p_{Pakeha} is approximately ~~22%~~ 11%.

F (5) ~~Since~~ although $\hat{p}_{Pasifika}$ is bigger than $\hat{p}_{Pakeha}$, we can't claim that $p_{Pasifika}$ is bigger than p_{Pakeha}.

51. Which **one** of the following statements about a *P-value* is **false**?

F (1) The ~~larger~~ smaller a *P-value*, the stronger the evidence against the null hypothesis.

T (2) A *P-value* measures the strength of evidence against the null hypothesis.

T (3) A relatively large test statistic results in a relatively small *P-value*.

T (4) A *P-value* is the conditional probability of observing a test statistic as extreme as that observed or even more so, if the null hypothesis were true.

T (5) A *P-value* says nothing about the size of an effect or difference. CI does!

width of CI $= .1704 - {}^{-}.0480 = .2184$

moe $= \dfrac{\text{width}}{2} = \dfrac{.2184}{2} = .1092 \approx 11\%$

Questions 52 to **55** refer to the following information in this appendix.

Researchers (Ridker et al., 2008) conducted a study to see whether taking a particular drug reduces the rate of a first major cardiovascular event (FMCE).

17 802 apparently healthy men and women were randomly assigned to receive 20 mg of this drug daily or to receive a placebo. The researchers recorded whether or not an FMCE occurred within five years for each participant and found that 142 of the 8901 subjects in the drug group had an FMCE within five years compared to 251 of the 8901 subjects in the placebo group.

Let:

p_{Ros} be the proportion who would have had an FMCE within five years if all 17 802 subjects received 20mg of the drug daily → $\hat{p}_{Ros} = \frac{142}{8901}$

and

p_{Pla} be the proportion who would have had an FMCE within five years if all 17 802 subjects received a placebo daily. → $\hat{p}_{Pla} = \frac{251}{8901}$

Using the t-procedure tool, a 95% confidence interval for $p_{Ros} - p_{Pla}$ is calculated to be $(-0.0166, -0.0079)$. ← zero out of CI ∴ not a plausible value → $p_{Ros} < p_{Pla}$ by between .79 & 1.66 %age points.

52. The sampling situation associated with se($\hat{p}_{Ros} - \hat{p}_{Pla}$) can be described as:

b (1) one sample of size 17 802, several response categories.

a (2) two independent samples, of sizes 142 and 251. Sit (a)!

a (3) two independent samples, both of size 8901.

c (4) one sample of size 8901, many yes/no items. $n_{Ros} = 8901$

b (5) one sample of size 8901, several response categories. $n_{pla} =$ 〃

53. When considering a two-tailed t-test for no difference between p_{Ros} and p_{Pla}, which one of the following statements is **false**? p-val < .05 (5%.)

T (1) At the 5% level of significance, it may be claimed that p_{Ros} is less than p_{Pla}. stat sig @ 5% level & 10% level!

T (2) If a similar (second) study was conducted using the same number of participants, it would be surprising if the second study produced a result with p_{Ros} greater than p_{Pla}.

T (3) The observed difference between p_{Ros} and p_{Pla} is significant at the 5% level of significance.

T (4) At the 10% level of significance, it is not plausible that the observed difference between p_{Ros} and p_{Pla} could be due to chance alone.

F (5) The observed difference between p_{Ros} and p_{Pla} is ~~not~~ significant at the 10% level of significance.

54. Which **one** of the following statements is **true**?

F (1) This study is an experiment and we can conclude that when an apparently healthy adult is treated with the drug it will lower the rate of having a first major cardiovascular event within five years.

T (2) This study is an experiment and, for the 17 802 participants, the observed reduction in the proportion in the drug (treatment) group who had a first major cardiovascular event within five years compared to the proportion in the placebo group can be attributed to the drug.

F (3) This study is an observational study but, for the 17 802 participants, the observed reduction in the proportion in the drug (treatment) group who had a first major cardiovascular event within five years compared to the proportion in the placebo group cannot be attributed to the drug.

F (4) This study is an experiment but, for the 17 802 participants, the observed reduction in the proportion in the drug (treatment) group who had a first major cardiovascular event within five years compared to the proportion in the placebo group cannot be attributed to the drug.

F (5) This study is an observational study and, for the 17 802 participants, the observed reduction in the proportion in the drug (treatment) group who had a first major cardiovascular event within five years compared to the proportion in the placebo group can be attributed to the drug.

55. Suppose the study had used 5000 participants (i.e. 2500 in each group), instead of the 17 802 participants it actually used, and that $\hat{p}_{Ros}$ and $\hat{p}_{Pla}$ (i.e. the observed group proportions) remained unchanged.

Which **one** of the following statements is **true**?

If 5000 participants took part in the study instead of 17 802 participants, then the resulting 95% confidence interval for $p_{\text{Ros}} - p_{\text{Pla}}$ would be:

(1) wider and the estimation statement would be less precise.

(2) narrower and the estimation statement would be less precise.

(3) narrower and the estimation statement would be more precise.

(4) wider and the estimation statement would have the same precision.

(5) unchanged and the estimation statement would have the same precision.

$8901 \xrightarrow{\div 4 \text{ (approx)}} 2500$

Question 56 refers to the following information.

The Auckland Marathon is an annual running event which involves thousands of runners. Within the event, there are five different races: the full marathon (42 km), the half marathon (21 km), the traverse (12 km), the challenge (5 km), and the kids marathon (2.2 km). Information about each runner who enters the Auckland Marathon is made available to the public each year on the website www.aucklandmarathon.co.nz. Data about a random sample of 1000 runners who entered the Auckland Marathon in 2015 was scraped (with permission) from this website to create a data set.

cate, nom   cate, nom

An excerpt of this sample data set is shown in Figure 3.

bib number	name	gender	division	age division	distance in km	completed event	time in hours	place	mean pace km per hr
11055	KATIE CARROLL	F	F0034	Up to 34	21	No			
755	JULIE CAHILL	F	F4549	45 to 49	21	Yes	3.4	4,793	6.2
10610	VICKY FOSTER	F	F0034	Up to 34	21	No			
9998	GEORJARNA MACCORMACK	F	F0034	Up to 34	21	No			
4528	DEON STOLTZ	M	M6064	60 to 64	42	Yes	7	1,507	6
4648	GARRY DONOGHUE	M	M7074	70 to 74	42	Yes	4.7	1,069	8.9
25459	JOHANN RLYNON	F	F0034	Up to 34	12	No			
21940	FERNANDA STEWART	F	F0034	Up to 34	12	No			
21872	GABOR PERJESSY	M	M0034	Up to 34	12	Yes	1.1	235	10.9
3804	CHRISTOPHER ABESAMIS	M	M4044	40 to 44	42	Yes	4.2	694	10
20910	LIZA CLARK	F	F5054	50 to 54	12	Yes	1.3	566	9.2
25236	HELEN MARINOVICH	F	F5054	50 to 54	12	Yes	2.1	1,771	5.7
14887	ZHENG DONG	M	M0034	Up to 34	21	Yes	1.9	937	11.1
11737	ELLA STENSNESS	F	F0034	Up to 34	21	No			
20282	HEATHER IRVINE	F	F4044	40 to 44	12	Yes	2	1,703	6
11773	ALEXANDRA BARNETT	F	F0034	Up to 34	21	Yes	2.2	2,365	9.5
4364	SAMUEL EASTON	M	M0034	Up to 34	42	Yes	4.7	1,044	8.9
388	SHARON RANDELL	F	F4549	45 to 49	42	Yes	4	594	10.5
2660	TREVOR THEYS	M	M4549	45 to 49	21	Yes	2.3	2,806	9.1

Figure 3: Excerpt of the sample data set created

56. Assuming the conditions for the underlying assumptions are satisfied, which **one** of the following types of analysis would be **most appropriate** to investigate the relationship between **gender** and **completed event**?

(1) Correlation. → 2 numeric vars
(2) Simple linear regression. → 2 numeric vars } e.g. time vs mean pace
(3) *t*-test on a difference between two proportions. → 2 categorical vars
(4) *F*-test for one-way analysis of variance. → 1 numeric & 1 cate
(5) One-sample *t*-test on a proportion. → 1 categorical var

For "Spot the Analysis" practice across Chapters 7 to 10, use Anna Fergusson's app here: www.tinyURL.com/stats-spot

var with 3 or more levels
e.g. time vs age division

Questions 57 to **59** refer to the following information.

Students enrolled in stage one statistics courses at the University of Auckland were surveyed regarding their access to, and experience with, computers. The survey was included as a question in an assignment, and students were given marks for completing it (irrespective of the answers they gave). Staff administering the courses wished to use the results of this survey to draw conclusions about future stage one statistics students.

One question asked: 'At the start of the course, how would you describe your Excel experience?'. A total of 918 students answered this question. Each of the 918 answers were classified according to the response given by the student, and the stream the student attended. The results are given in the table below, where 101G, 108 and 101 refer to the various streams.

	Stream			
Response	101G	101	108	Total
None	15	36	102	153
Very Little	44	89	119	252
Some	74	150	200	424
Lots	9	29	51	89
Total	142	304	472	918

$\hat{p}_{101} = \frac{36}{304} = .1184$ (4dp)

$\hat{p}_{108} = \frac{102}{472} = .2161$ (4dp)

57. A hypothesis test is performed on the data for no difference between the proportion of **101** students who responded **None** and the proportion of **108** students who responded **None**. Which one of the following statements about this hypothesis test **is true**?

F (1) The degrees of freedom used depends on the number of 101 and 108 students in the sample. → ∞!

F (2) The one sample *t*-test should be used. → diff in 2 proportions!

T (3) The test should be two-tailed.

F (4) The test could ~~only~~ be used to show a difference existed in the ~~sample~~ pop. proportions.

F (5) An appropriate null hypothesis is that the difference between the proportion of 101 students who responded None and the proportion of 108 students who responded None, is ~~not~~ zero. $H_0: p_{101} - p_{108} = 0$

58. The standard error for the difference in the proportions tested in Question 57 would be calculated using the sampling situation described as:

sit (a)!

$n_{101} = 304$

$n_{108} = 472$

b (1) ~~one~~ sample of size 304, several response categories.

C (2) one sample of size 472, many yes/no items.

c (3) one sample of size 304, many yes/no items.

a (4) two independent samples of sizes 304 and 472.

b (5) one sample of size 472, several response categories.

St.ev. against

No!

59. The *P-value* for the statistical test mentioned in Question 57 is 0.004. Which one of the following statements gives the **best** interpretation of this *P–value*?

F (1) There is some evidence that the underlying proportions of 101 students and 108 students that would respond None are different.

F (2) There is strong evidence that the sample proportions of 101 students and 108 students that responded None are different.

T (3) There is strong evidence that the underlying proportions of 101 students and 108 students that would respond None are different.

F (4) There is no evidence that the underlying proportions of 101 students and 108 students that would respond None are different.

F (5) There is weak evidence that the underlying proportions of 101 students and 108 students that would respond None are different.

ANSWERS

1. **(2)**	2. **(4)**	3. **(2)**	4. **(4)**	5. **(2)**	6. **(1)**
7. **(3)**	8. **(2)**	9. **(5)**	10. **(2)**	11. **(3)**	12. **(1)**
13. **(1)**	14. **(4)**	15. **(3)**	16. **(3)**	17. **(2)**	18. **(4)**
19. **(1)**	20. **(1)**	21. **(2)**	22. **(5)**	23. **(1)**	24. **(3)**
25. **(2)**	26. **(5)**	27. **(5)**	28. **(4)**	29. **(1)**	30. **(2)**
31. **(4)**	32. **(4)**	33. **(4)**	34. **(4)**	35. **(5)**	36. **(1)**
37. **(2)**	38. **(1)**	39. **(3)**	40. **(1)**	41. **(3)**	42. **(4)**
43. **(1)**	44. **(4)**	45. **(1)**	46. **(4)**	47. **(2)**	48. **(2)**
49. **(2)**	50. **(3)**	51. **(1)**	52. **(3)**	53. **(5)**	54. **(2)**
55. **(1)**	56. **(3)**	57. **(3)**	58. **(4)**	59. **(3)**	

WHAT SHOULD I DO NEXT?

- **Do Question 2 of Assignment 3!** & Q5 a,b
- **Go through the <u>proportions material</u> in Chapter 6 and 7 of your Lecture Workbook. (Chapter 6: pages 7 & 8, 12-18, 20, 23, 25 & 26; Chapter 7: Pages 5 & 6, 11-16, 18-20, 22).**
- **Try Chapter 6 & 7 questions from three of the past five exams that are relevant to this workshop (i.e. on proportions not means!)**

→ S118 Exam Q2-4

→ S218 Exam Q25

→ S119 Q21-23 & S219 Q31&32

FORMULAE

Confidence intervals and *t*-tests

Confidence interval: $\textit{estimate} \pm t \times \text{se}(\textit{estimate})$

Ch 6,7,8,10

t-test statistic: $$t_0 = \frac{\textit{estimate} - \textit{hypothesised value}}{\textit{standard error}}$$

Ch 7,8,10

Applications:

~~1. Single mean μ: $\textit{estimate} = \overline{x}$; $\textit{df} = n - 1$~~

2. Single proportion p: $\textit{estimate} = \hat{p}$; $\textit{df} = \infty$

~~3. Difference between two means $\mu_1 - \mu_2$: (independent samples)~~

~~$\textit{estimate} = \overline{x}_1 - \overline{x}_2$; $\textit{df} = \min(n_1 - 1, n_2 - 1)$~~

Ch 6/7

HTP

4. Difference between two proportions $p_1 - p_2$:

   $\textit{estimate} = \hat{p}_1 - \hat{p}_2$; $\textit{df} = \infty$

   Situation (a): *Proportions from two independent samples*

   Situation (b): *One sample of size n, several response categories*

   Situation (c): *One sample of size n, many yes/no items*

The *F*-test (ANOVA)

F-test statistic: $f_0 = \dfrac{s_B^2}{s_W^2}$; $\quad df_1 = k - 1,\ df_2 = n_{\text{tot}} - k$

Ch 8 HTM2

The Chi-square test

Chi-square test statistic: $$\chi_0^2 = \sum_{\text{all cells in the table}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

Ch 9

Expected count in cell $(i, j) = \dfrac{R_i C_j}{n}$

CJT

$\textit{df} = (I - 1)(J - 1)$

Regression

Ch 10

Fitted least-squares regression line: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Inference about the intercept, β_0, and the slope, β_1: $\textit{df} = n - 2$

RC