Q1(a) Let μ_T be the mean percentage of body fat in the treatment group and let μ_P be the mean percentage of body fat in the placebo group.

Test $H_0: \mu_T = \mu_P$ against $H_1: \mu_T \neq \mu_P$

pooled sample variance $S_p^2 = \dfrac{(26-1)\times 0.078^2 + (13-1)\times 0.043^2}{26+13-2}$

$= 0.00471$

Test statistic $t = \dfrac{(0.368 - 0.327) - 0}{\sqrt{0.00471}\sqrt{1/26 + 1/13}}$

$= 1.759$

p-value $= 2\times \min\{\mathbb{P}(T_{37} > 1.759), \mathbb{P}(T_{37} < 1.759)\}$

$= 2\times 0.0434 = 0.0869$ (tcdf(1.759, 37, 'upper') in Matlab)

or using T-table with 30 degrees of freedom gives a p-value of $2\times\mathbb{P}(T_{30} > 1.759)$ which is between 0.05 and 0.1.

There is weak evidence against the null hypothesis, suggesting there is a difference in mean percentage of body fat between treatment + placebo group.

(b) Let μ be the mean decrease in fasting insulin.

Test $H_0: \mu = 0$ against $H_1: \mu > 0$

Test statistic $t = \dfrac{5.7 - 0}{26.7/\sqrt{26}} = 1.0886$

p-value $= \mathbb{P}(T_{25} \geq 1.0886)$

$\mathbb{P}(T_{25} \geq 0.684) = 0.25$ and $\mathbb{P}(T_{25} \geq 1.316) = 0.10$

So the p-value is between 0.1 and 0.25

This is inconclusive evidence against H_0, suggesting there is no decrease in mean fasting insulin.

Q2. (a) $\dfrac{496 + 406}{1248 + 1057} \approx 0.391$

(b) $\dfrac{496}{1248} - \dfrac{406}{1057} \approx 0.0133$

(c) 95% CI for the difference of proportions

$$\hat{p}_1 - \hat{p}_2 \pm 1.96 \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$n_1 = 1248$ $\quad \hat{p}_1 = \dfrac{496}{1248} \approx 0.397$ $\quad n_2 = 1057$ $\quad \hat{p}_1 = \dfrac{406}{1057} = 0.384$

$$0.0133 \pm 1.96 \sqrt{\frac{0.397 \times (1-0.397)}{1248} + \frac{0.384 \times (1-0.384)}{1057}}$$

$= 0.133 \pm 0.03996$

As the 95% CI covers 0, there does not appear to be a difference in the accidental poisoning rates due to household cleaning products between the two hospitals at the 5% significance level.

(d) Test H_0: 'Hospital' and 'Discharge category' are independent.
against H_1: some association between 'Hospital' and 'Discharge category'.

Expected Counts

	Usual Residence	Admitted	Transfer	
Clayton	$\frac{678 \times 1248}{2305} = 367.1$	$\frac{1350 \times 1248}{2305} = 730.9$	$\frac{277 \times 1248}{2305} = 150$	1248
Casey	$\frac{678 \times 1057}{2305} = 310.9$	$\frac{1350 \times 1057}{2305} = 619.1$	$\frac{277 \times 1057}{2305} = 127$	1057
	678	1350	277	2305

$$X^2 = \sum_i \frac{(e_i - o_i)^2}{e_i} = \frac{(367.1-317)^2}{367.1} + \frac{(310.9-361)^2}{310.9} + \frac{(730.9-596)^2}{730.9}$$

$$+ \frac{(619.1-754)^2}{619.1} + \frac{(150-144)^2}{150} + \frac{(127-133)^2}{127} \simeq 5.99$$

Under H_0 X^2 has approximately χ^2_{df} distribution where

$$df = (\#\text{rows} - 1) \times (\#\text{columns} - 1) = (2-1) \times (3-1) = 2$$

$$\text{p-value} = \mathbb{P}(\chi^2_2 \geq 5.99) \simeq 0.05$$

There is mild/weak evidence against the null hypothesis, suggesting an association between Hospital + Discharge Category.

Q3 (a) $H_0: \mu_1 = \mu_2$ against $H_1: \mu_2 > \mu_1$.

(b) $$S_p = \sqrt{\frac{(10-1)\times 2.56^2 + (15-1)\times 2.81^2}{10+15-2}} = 2.715$$

(c) Test statistic $$t = \frac{(74.2-70.3)-0}{2.715\sqrt{1/10+1/15}} = 3.518$$

$$\text{p-value} = \mathbb{P}(T_{23} \geqslant 3.518)$$

$$\mathbb{P}(T_{23} \geqslant 3.485) = 0.001 \quad \text{and} \quad \mathbb{P}(T_{23} \geqslant 3.768) = 0.0005$$

The p-value is between 0.0005 and 0.001 (0.000921 using MATLAB).

This is strong evidence against the null hypothesis, suggesting a higher mean weight gain with the new diet.

Q4 (a) $H_0: \mu = 0$ against $H_1: \mu > 0$ (there is a reduction)

(b) Test statistic $$t = \frac{6.1-0}{12.36/\sqrt{10}} = 1.56$$

$$\text{p-value} = \mathbb{P}(T_9 \geqslant 1.56)$$

$$\mathbb{P}(T_9 > 1.383) = 0.10 \quad \text{and} \quad \mathbb{P}(T_9 \geqslant 1.833) = 0.05$$

The p-value is between 0.05 and 0.1.

There is weak evidence against H_0, suggesting a learning effect is present.

(c) 95% CI $\bar{x} \pm t_{9;0.975}\ s.e.(\bar{x})$

$$6.1 \pm 2.262 \times \frac{12.36}{\sqrt{10}}$$

$$\Rightarrow 6.1 \pm 8.84 \text{ cm}$$

i.e. we are 95% confident that the reduction is between -2.74 cm and 14.94 cm.

(d) We want the margin of error to be 3cm

$$3 \simeq 1.96 \times \frac{12.36}{\sqrt{n}}$$ (assuming standard dev is 12.36)

$$\Rightarrow n = \left(\frac{1.96 \times 12.36}{3}\right)^2 \simeq 65$$

Around 65 subjects are needed.

Q5. Test H_0: 'Fish intake' and 'Asthma status' are independent.
against H_1: some association between 'Fish intake' and 'Asthma status'.

We could test this as for two proportions, ~~are~~ but here I use the χ^2 test of independence.

Expected counts

Fish intake	Yes	No	
Asthma	$\frac{2448 \times 659}{3535} = 456.4$	$\frac{1087 \times 659}{3535} = 202.6$	659
No Asthma	$\frac{2448 \times 2876}{3535} = 1991.6$	$\frac{1087 \times 2876}{3535} = 844.4$	2876
	2448	1087 1087	3535

$$X^2 = \frac{(456.4 - 360)^2}{456.4} + \frac{(1991.6 - 2088)^2}{1991.6} + \frac{(202.6 - 229)^2}{202.6} + \frac{(844.4 - 788)^2}{844.4}$$

$$= \cancel{31.07} \; 81.329$$

$$\text{p-value} = \mathbb{P}(\chi^2_1 \geq 81.329) < 0.001$$

$$(\text{As } \mathbb{P}(\chi^2_1 \geq 10.83) = 0.001)$$

There is (very) strong evidence against the null hypothesis, suggestion there is some association between Fish intake and asthma status.

Q6 (a) $$s_p^2 = \frac{(51-1) \times 20^2 + (67-1) \times 19^2}{51 + 67 - 2} = 377.81$$

(b) Let μ_{2013} be the mean length of bluefin tuna in 2013, and
" μ_{2014} " " " " " 2014.

Test $H_0: \mu_{2013} = \mu_{2014}$ against $H_1: \mu_{2013} > \mu_{2014}$

Test statistic $$t = \frac{(180 - 175) - 0}{\sqrt{377.81}\sqrt{1/51 + 1/67}} = 1.384$$

p-value $= \mathbb{P}(T_{116} \geq 1.384)$ is between 0.05 and 0.1

as

$$\mathbb{P}(T_{116} \geq 1.384) \simeq \mathbb{P}(T_\infty \geq 1.384) \text{ and}$$

$$\mathbb{P}(T_\infty \geq 1.282) = 0.10 \text{ and } \mathbb{P}(T_\infty \geq 1.645) = 0.05$$

There is weak evidence against the null hypothesis, suggesting a decline in the mean length of Bluefin tuna.

(c) $t_{116;0.975} \simeq 1.96$ so

95% CI is $(180-175) \pm 1.96 \times \sqrt{377.81}\sqrt{1/51 + 1/67}$

$= 5 \pm 7.08$ (cm).

Q7 Let p_1 = proportion of smokers who get pregnant in first cycle
p_2 = proportion of non-smokers who get pregnant in first cycle.

Test $H_0: p_1 = p_2$ against $H_1: p_1 < p_2$

$\hat{p}_1 = \frac{29}{100} = 0.29$ $\hat{p}_2 = \frac{198}{198+288} \approx 0.407$

(pooled) $\hat{p} = \frac{29+198}{29+71+198+288} = \frac{227}{586} = 0.387$

test statistic $= \frac{(0.407 - 0.29) - 0}{\sqrt{0.387 \times (1-0.387)}\sqrt{\frac{1}{486} + \frac{1}{100}}}$

$= 2.195$

p-value $= \mathbb{P}(Z \geqslant 2.195) = 0.014$

This is moderate evidence against the null hypothesis, suggesting that there is a higher proportion of non-smokers who fall pregnant in first cycle.

Q8 a) CI estimate $\pm$ (critical value) $\times$ s.e.(estimate)

$$19.6 \pm t_{89;0.995} \times \frac{8.2}{\sqrt{90}}$$

Either use $t_{80;0.995} = 2.639$ from tables
or $t_{89;0.955} = 2.6322$ using MATLAB. So, using
the $t_{89;0.955}$ value, the CI is

$$19.6 \pm 2.6322 \times 8.2 / \sqrt{90}$$
$$= 19.6 \pm 2.275$$

We are 99% confident the true mean number of polyester particles per 250ml of beach sediment is between 17.32 and 21.87

b) Let μ_V be the mean number of polyester microparticles per 250ml of beach sediment in Victoria and let μ_Q be the corresponding mean for North Queensland.

Test $H_0: \mu_Q = \mu_V$ against $H_1: \mu_Q \neq \mu_V$

test statistic $= \dfrac{\text{estimate} - \text{hypothesis}}{\text{s.e.(estimate)}}$

We need the pooled variance estimate to get the s.e.

$$S_p^2 = \frac{(45-1)\times 8.54^2 + (45-1)\times 9.82^2}{45+45-2} = 84.682$$

test statistic $t = \dfrac{(16.67 - 21.22) - 0}{\sqrt{84.682}\sqrt{1/45 + 1/45}} = \dfrac{-4.55}{9.2023 \times 0.2108} = -2.345$

compare to a t_{88}-distribution to get the p-value.

p-value $= 2 \times \min\{\mathbb{P}(T_{88} \geqslant -2.345), \mathbb{P}(T_{88} \leqslant -2.345)\}$

By symmetry of the t-distribution

$$\mathbb{P}(T_{88} \leqslant -2.345) = \mathbb{P}(T_{88} \geqslant 2.345)$$

Using MATLAB, p-value $= 2 \times 0.0106 = 0.0213$

Approximation using tables:

$\mathbb{P}(T_{80} \geqslant 1.990) = 0.025$ and $\mathbb{P}(T_{80} \geqslant 2.374) = 0.01$

So the p-value is between 0.02 and 0.05.

There is moderate evidence against the null hypothesis, suggesting a difference in the mean number of microparticles in 250ml of beach sediment between North Queensland and Victoria.

Q9(a) CI estimate $\pm$ (critical value) $\times$ s.e. (estimate)

$\hat{p} = 0.43$ $\quad$ s.e.$(\hat{p}) = \sqrt{0.43 \times (1-0.43)/1000} = 0.01566$

critical value from normal distribution $Z_{0.95} = 1.645$

$$0.43 \pm 1.645 \times 0.01566$$
$$= 0.43 \pm 0.0257$$

We are 90% confident that the proportion of adult australians who didn't pay with cash in the past 7 days is between 0.404 and 0.456.

(b) let p_{over} and p_{under} be the respective proportion of over 35 and under 35 who made no cash transactions in the past seven days.

Test $H_0 : p_{over} = p_{under}$ against $H_1 : p_{over} \neq p_{under}$

$\hat{p}_{over} = 198/566 = 0.3498$ $\hat{p}_{under} = 232/434 = 0.5346$

pooled estimate $\hat{p} = 430/1000 = 0.43$

Test statistic $$t = \frac{(\hat{p}_{over} - \hat{p}_{under}) - 0}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{1/n_{over} + 1/n_{under}}}$$

$$= \frac{(0.3498 - 0.5346) - 0}{\sqrt{0.43(1-0.43)}\sqrt{1/566 + 1/434}}$$

$$= -5.85$$

$$\text{p-value} = 2 \times \min\{\mathbb{P}(Z \leq -5.85), \mathbb{P}(Z \geq -5.85)\}$$

$$\leq 2 \times 0.0001$$

The p-value is less than 0.0002. This is very strong evidence against the null hypothesis, suggesting a difference in the proportion of people who made no cash transaction between the two age groups.

Q10 We test

H_0: 'Relapse' and 'Treatment drug' are independent.

against H_1: There is some association between 'Relapse' and 'Treatment drug'.

Observed Counts

	Relapse	Did not relapse	
Placebo	20	4	24
Lithium	18	6	24
Desipramine	10	14	24
	48	24	72

Expected counts

	Relapse	Did not relapse
Placebo	16	8
Lithium	16	8
Desipramine	16	8

$$E_{11} = \frac{48 \times 24}{72} = 16$$

$$E_{12} = \frac{24 \times 24}{72} = 8$$

(all other expected counts similar)

$$X^2 = \sum_i \frac{(E_i - O_i)^2}{E_i}$$

$$= \frac{(16-20)^2}{16} + \frac{(8-4)^2}{8} + \frac{(16-18)^2}{16} + \frac{(8-6)^2}{8} + \frac{(16-10)^2}{16} + \frac{(8-14)^2}{8}$$

$$= 10.5$$

degrees of freedom

$= (\text{rows} - 1) \times (\text{columns} - 1)$

$= (3-1) \times (2-1) = 2$

p-value $= \mathbb{P}(\chi^2_2 \geq 10.5)$

As $\mathbb{P}(\chi^2_2 \geq 9.210) = 0.01$ and $\mathbb{P}(\chi^2_2 \geq 10.60) = 0.005$, the p-value is between 0.005 and 0.01.

This is strong evidence against the null hypothesis, suggesting some association between treatment drug and whether or not the patient relapses.

Q11 (a) Let μ_A and μ_B be the mean PCB concentrations in fish caught in Lakes A and B, respectively.

Test $H_0 : \mu_A = \mu_B$ against $H_1 : \mu_A \neq \mu_B$.

test statistic $t = \dfrac{\text{estimate} - \text{hypothesis}}{\text{s.e.(estimate)}}$

$$= \frac{(11.170 - 11.988) - 0}{s_p \sqrt{1/10 + 1/8}}$$

pooled variance estimate $s_p^2 = \dfrac{(10-1)\times 0.862^2 + (8-1)\times 0.738^2}{10+8-2}$

$$= 0.656244$$

$$s_p = 0.810$$

$$t = \frac{-0.818}{0.810\sqrt{1/10 + 1/8}} = -2.1288$$

p-value $= 2 \times \min\{\mathbb{P}(T_{16} \geq -2.1288), \mathbb{P}(T_{16} \leq -2.1288)\}$

Note that $\mathbb{P}(T_{16} \leq -2.1288) = \mathbb{P}(T_{16} \geq 2.1288)$ by symmetry of the t-distribution

As $\mathbb{P}(T_{16} \geq 2.120) = 0.025$ and $\mathbb{P}(T_{16} \geq 2.583) = 0.01$, the p-value is between 0.02 and 0.05

This is moderate evidence against the null hypothesis, suggesting a difference in the mean PCB concentration in fish between lakes A and B.

(b) CI estimate $\pm$ (critical value) $\times$ s.e.(estimate)

Critical value is $t_{16;0.975} = 2.120$. We have the s.e.(estimate) from part (a)

$$-0.818 \pm 2.120 \times 0.810 \times \sqrt{1/10 + 1/8}$$
$$-0.818 \pm 0.8146$$

We are 95% confident the true difference in mean PCB concentration in fish between lakes A and B is between -1.633 and -0.0034. This suggest a difference in the true means at the 5% significance level.

Q12 (a) Let μ be the mean difference in profit between row-planted crop and broadcast planted crop.

Test $H_0: \mu = 0$ against $H_1: \mu \neq 0$

Test statistic $t = \dfrac{\text{estimate} - \text{hypothesis}}{\text{s.e.(estimate)}}$

$$= \frac{-1.0143 - 0}{1.6067/\sqrt{7}} = -1.6702$$

p-value $= 2 \times \min\{\mathbb{P}(T_6 \geq -1.6702), \mathbb{P}(T_6 \leq -1.6702)\}$

As the t-distribution is symmetric $\mathbb{P}(T_6 \leq -1.6702) = \mathbb{P}(T_6 \geq 1.6702)$

$\mathbb{P}(T_6 \geq 1.440) = 0.1$ and $\mathbb{P}(T_6 \geq 1.943) = 0.05$. So the p-value is between 0.1 and 0.2.

There is no evidence against the null hypothesis, suggesting no difference in the mean profit from the two methods of sowing.

b) CI estimate $\pm$ (critical value) $\times$ s.e.(estimate)

critical value is $t_{6;0.975} = 2.447$. We have s.e.(estimate) from part (a) so

$$-1.014 \pm 2.447 \times \frac{1.6067}{\sqrt{7}}$$

$$= -1.014 \pm 1.486$$

We are 95% confident the difference in mean profit from the two sowing methods is between -2.500 (units) and 0.4717 (units).

Q13 a) Let p_G and p_P be the proportion of the population who would have heart-attacks being treated with Gemfibrozil and the placebo, respectively.

Test $H_0: p_G = p_P$ against $H_1: p_G < p_P$

$\hat{p}_G = \frac{56}{1995+56} = 0.0273$ $\hat{p}_P = \frac{84}{1946+84} = 0.04138$

(pooled) $\hat{p} = \frac{56+84}{1995+56+1946+84} = 0.03431$

test statistic $t = \frac{\text{estimate} - \text{hypothesis}}{\text{s.e.(estimate)}}$

$$t = \frac{(0.0273 - 0.04138) - 0}{\sqrt{0.03431 \times (1 - 0.03431)}\sqrt{\frac{1}{2051} + \frac{1}{2030}}}$$

$$= -2.47009$$

p-value $= \mathbb{P}(Z \leq -2.47009)$ (As alternative is $p_G < p_P$)

The p-value is between 0.005 and 0.01. This is strong evidence against the null hypothesis, suggesting the proportion of the population that would have a heart attack with the Gemfibrozil treatement is smaller than the proportion with placebo.

b) CI estimate $\pm$ (critical value) $\times$ s.e.(estimate)

critical value is $z_{0.95} = 1.645$. So the CI is

$$(0.0273 - 0.04138) \pm 1.645 \times \sqrt{\frac{0.0273(1-0.0273)}{2051} + \frac{0.04138(1-0.04)}{2030}}$$

$$= -0.01408 \pm 1.645 \times 0.004989$$

$$= -0.01408 \pm 0.00821$$

We are 90% confident the true reduction in the proportion of heart attacks is between 0.005867 and 0.02228 under the Gemfibrozil treatment relative to the placebo.

Q14 Let p be the proportion of students who sit at the front of a lecture who are shorter than 150cm

$\hat{p} = \frac{10}{50} = 0.2$ CI estimate ± (critical value) × s.e.(estimate)

Critical value is 1.960

$$CI = 0.2 \pm 1.960 \times \sqrt{\frac{0.2 \times (1-0.2)}{50}}$$

$$= 0.2 \pm 0.1109$$

We are 95% confident the proportion of students who sit at the front of a lecture who are shorter than 150cm is between 0.089 and 0.311.