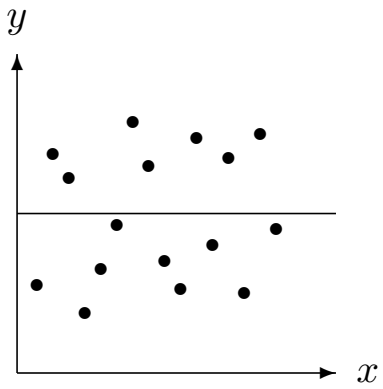
13 Regression model - Part II

13.1 Testing the significance of regression model

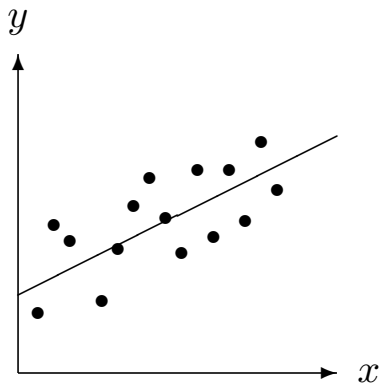
We have learnt how to fit a regression line and check its goodness-of-fit using r^2 and residual plot.

However, we do not know whether the fitted regression line, in particular the slope of the regression line, is significantly different from zero. If the slope is just zero, the fitted line is essentially horizontal, indicating no relationship between X and Y.

A random scatter plot indicates the model is insignificant with zero slope.



A random scatter plot indicates the model is significant.



To construct a hypothesis test to test whether $\beta = 0$ in the regression model

$$\hat{Y}_i = \alpha + \beta x_i,$$

we need a distribution for its estimate b. It can be proved that

$$b \sim N\left(\beta, \frac{\sigma^2}{L_{xx}}\right)$$

where σ^2 is the variance of the residuals r_i. An estimate of σ^2 is

$$s^2 = \frac{\sum_{i=1}^{n} r_i^2}{n-2} = \frac{L_{yy} - bL_{xy}}{n-2}$$

which is the ratio of the *residual sum of squares* (RSS) to df. The df is $n-2$ due to the two parameter estimates a and b. Then the test statistic is

$$t_{\text{obs}} = \frac{b}{s/\sqrt{L_{xx}}} \sim t_{n-2}$$

under $H_0 : \beta = 0$. In summary, the four steps of the test are:

1. Hypotheses: $H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$.
2. Test statistic: $t_{\text{obs}} = \dfrac{b}{s/\sqrt{L_{xx}}}$ where $s^2 = \dfrac{L_{yy} - bL_{xy}}{n-2}$.
3. P-value: $2P(T_{n-2} > |t_{\text{obs}}|)$ for 2-sided test.
4. Conclusion: If P-value < 0.05, the data are against H_0 and hence the regression line is significant.

Example: Test if the regression line for the dose and concentration data discussed before is significant.

Solution: We have $L_{xx} = 844.25$, $L_{yy} = 74$, $L_{xy} = 230$ and $b = 0.272$. The hypothesis test on the slope of a regression line is:

1. Hypotheses: _____.

2. Test statistic: $s^2 = \dfrac{L_{yy} - bL_{xy}}{n-2} =$ _____.

   Hence $t_{\text{obs}} = \dfrac{b}{s/\sqrt{L_{xx}}} =$ _____.

3. P-value: _____ (2-sided test).

4. Conclusion: _____. Hence the regression line is significant.

Attempt Question 2(e) of June 2011 exam. for a second example.

13.2 Summary

Sum. stat.	location: $\bar{x}=\frac{\sum_i x_i}{n}$, Q_1, Q_2, Q_3, LT=Q_1-1.5×IQR, UT=Q_3+1.5×IQR, mode spread: $s^2=\frac{\sum_i x_i^2-\frac{(\sum_i x_i)^2}{n}}{n-1}$, range=max-min, IQR=$Q_3$-$Q_1$, outliers$\notin$(LT,UT) plot: stem-and-leaf plot, boxplot, histogram
Prob.	$P(A\cup B)=P(A)+P(B)-P(A\cap B)$ If $P(A\cap B)=0$, A & B mutually exclusive & $P(A\cup B)=P(A)+P(B)$ $P(A\cap B)=P(A)P(B\|A)$ If $P(B\|A)=P(B)$, A & B independent & $P(A\cap B)=P(A)P(B)$

	Normal (continuous)	Binomial (discrete)
Dist.	$X\sim N(\mu,\sigma^2)$ $X_1\pm X_2\sim N(\mu_1\pm\mu_2,\sigma_1^2+\sigma_2^2)$ $P(X<x)=P(Z<\frac{x-\mu}{\sigma})$	$X\sim B(n,p)$ $E(X)=np$; $Var(X)=np(1-p)$ $P(X=r)=\binom{n}{r}p^r(1-p)^{n-r}$
Sam. dist.	$\bar{X}\sim N(\mu,\frac{\sigma^2}{n})$	$\hat{p}=\frac{X}{n}\overset{CLT}{\sim}N(p,\frac{p(1-p)}{n})$ if n large
One sam.	One sample or matched pairs t-test $H_0:\mu=\mu_0$ (normal, σ^2 unknown)	One sample Z-test $H_0:p=p_0$ (n large)
CI: HT:	$\bar{x}\mp t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$ t-test: $t_o=\frac{\bar{x}-\mu_0}{s/\sqrt{n}}\sim t_{n-1}$	$\hat{p}\mp z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ Z-test: $z_o=\frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/\sqrt{n}}}\sim N(0,1)$
Two sam.	Two sample t-test H_0: $\mu_1=\mu_2$ (normals, σ_i^2 unk. & equal)	Two sample Z-test H_0: $p_1=p_2$ (n_1, n_2 large)
CI: HT:	$\bar{x}_1-\bar{x}_2\mp t_{n_1+n_2-2,\alpha/2}s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ t-test: $t_o=\frac{\bar{x}_1-\bar{x}_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim t_{n_1+n_2-2}$ $s_p^2=\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$	$\hat{p}_1-\hat{p}_2\mp z_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ Z-test: $z_o=\frac{\hat{p}_1-\hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim N(0,1)$ $\hat{p}=\frac{X_1+X_2}{n_1+n_2}$
Cate.	One var.: H_0: $p_i=p_{i0}$ ($E_i=np_{i0}\geq 5$) χ^2-test: $\chi^2_o=\sum_i\frac{(x_i-np_{i0})^2}{np_{i0}}\sim\chi^2_{g-1}$	Two var.: H_0: $p_{ij}=p_ip_j$ ($E_{ij}=\frac{r_ic_j}{n}\geq 5$) χ^2-test: $\chi^2_o=\sum_{i,j}\frac{(x_{ij}-\frac{r_ic_j}{n})^2}{\frac{r_ic_j}{n}}\sim\chi^2_{(r-1)(c-1)}$
Reg.	$L_{xy}=\sum_i x_iy_i-\frac{1}{n}(\sum_i x_i)(\sum_i y_i)$ $L_{xx}=\sum_i x_i^2-\frac{1}{n}(\sum_i x_i)^2$ $L_{yy}=\sum_i y_i^2-\frac{1}{n}(\sum_i y_i)^2$ Cor.: $r=\frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}}$, r^2=% of fit	Reg. line: $y=a+bx$ $b=\frac{L_{xy}}{L_{xx}}$, $a=\bar{y}-b\bar{x}$ Test H_0:$\beta=0$: $t_o=\frac{b}{s/\sqrt{L_{xx}}}\sim t_{n-2}$ $s^2=\frac{L_{yy}-bL_{xy}}{n-2}$

Note:

1. In HT, the alternative hypothesis H_1 can be:
   upper-sided (larger), lower-sided (less) or
   two-sided (different) where P-value is *2 times* the 1-sided.

2. In HT, H_0 is rejected if
   (1) P-value< 0.05 or
   (2) μ_0, p_0 (1-sample) or 0 (2-sample) lies *outside* $(1-\alpha)\%$ CI *except* the CI for $\hat{p}$ when s.e. is calculated using $\hat{p}$, not p_0.

3. The df in the t-test is $n-1$ for 1-sample,
   n_1+n_2-2 for 2-sample and $n-2$ for regression.
   The df in the χ^2-test is $g-1$ for χ^2 GOF test (1 categorical var.) and $(r-1)(c-1)$ for χ^2 test for independence (2 categorical var.).

Revision Questions

We will discuss 2012 exam paper in detail. Attempt the problems before lecture.

1. A study is conducted to compare female adolescents who suffer from bulimia with healthy females with similar body compositions and levels of physical activities. Listed below are measures of daily caloric intake, recorded in kilocalories per kilogram of body weight, for 24 female adolescents who suffer from bulimia and 15 healthy female adolescents. Measurements are sorted from the smallest to the largest. Values a, b, and c are unknown.

Daily Caloric Intake

Bulimia X			**Healthy** Y	
17	20	27	a	31
18	20	29	b	33
19	21	30	20	34
19	21	31	22	37
19	22	33	25	39
20	23	34	26	40
20	24	38	28	41
20	26	41	c	

   (a) *[4 marks]* For the healthy group, the sample mean, median and range are known to be 29.6, 31 and 23, respectively. Find a, b and c

   (b) *[3 marks]* For the healthy group, find the lower and upper quartiles and hence the interquartile range.

(c) *[2 marks]* For the Bulimia group, draw a stem-and-leaf plot using 10, 20, 30 and 40 as stems.

(d) A researcher classifies the daily caloric intake as `low` if it is 20 or less, `high` if it is at least 35 and `normal` otherwise. He finds that the ratio of `low: normal: high` levels of daily caloric intake is 1:2:1 among healthy females adolescents. Perform a suitable test at *5% significance level* to confirm if this ratio applies to female adolescents who *suffer from bulimia* using the following steps.

(i) *[1 mark]* State the hypotheses associated with this test.

(ii) *[1 mark]* Find the observed counts O_1, O_2, O_3 respectively for `low`, `normal` and `high` levels of daily caloric intake among the 24 female adolescents who suffer from bulimia.

(iii) *[2 marks]* Find the expected counts E_1, E_2, E_3 respectively for `low`, `normal` and `high` levels of daily caloric intake among the 24 female adolescents who suffer from bulimia.

(iv) *[4 marks]* Complete the table below to calculate a suitable test statistic.

Class	Observed count O_i	Expected count E_i	$\frac{(O_i-E_i)^2}{E_i}$
low			
normal			
high			
Total	24	24	

(v) *[2 marks]* Write down an expression of the p-value for the test in (iv) and calculate the p-value.

(vi) *[1 mark]* Draw your conclusion based on the p-value of (v).

Solution:

(a) Range = __ and hence a = _____.
Median = __ and hence c = __.
Mean = ____,
b = _____.

(b) The lower group: _____,
Q_1 = median of lower group = _____.
The upper group: _____,
Q_3 = median of upper group = _____.
IQR = _____.

(c) The stem and leaf represent ten and unit values respectively.

(d) (i) _____.

(ii) _____.

(iii) _____

_____.

(iv)

Class i	Obs O_i	Exp E_i	$\frac{(O_i-E_i)^2}{E_i}$
low	__	_	_____
normal	__	__	_____
high	_	_	_____
Total	__	__	_____

(v) P-value: _____

(vi) Conclusion: _____.

2. The birth weight y (in kg) and gestation age x (in weeks) for a random sample of 10 infants are given in the table below:

Birth weight (y)	2.94	3.13	2.42	2.45	2.76	2.44	3.226	3.301	2.729	3.41
gestation age (x)	38	38	36	34	39	35	40	42	37	40

$$\sum_{i=1}^{10} x_i = 379,\ \sum_{i=1}^{10} x_i^2 = 14419,\ \sum_{i=1}^{10} y_i = 28.806,\ \sum_{i=1}^{10} y_i^2 = 84.2498,\ \sum_{i=1}^{10} x_i y_i = 1099.175$$

(a) (i) *[4 marks]* Calculate the correlation coefficient between x and y.

(ii) *[1 mark]* Comment on the correlation coefficient in (i).

(iii) *[2 marks]* What proportion of the variability in birth weight y is explained by gestation age x using a linear regression model of y on x?

(b) *[4 marks]* Calculate and state the regression model for the data.

(c) *[7 marks]* Test whether the linear relationship between birth weight y and gestation age x in (b) is significant. Justify your answer using an appropriate t-test and perform all the steps.

(d) *[2 marks]* Predict the birth weight for an infant born prematurely at 28 weeks gestation. Comment on the suitability of the prediction.

Solution:

(a) (i)

$$L_{xx} = \sum_i x_i^2 - \frac{(\sum_i x_i)^2}{n} = \underline{\qquad\qquad}$$

$$L_{xy} = \sum_i x_i y_i - \frac{(\sum_i x_i)(\sum_i y_i)}{n} = \underline{\qquad\qquad}$$

$$L_{yy} = \sum_i y_i^2 - \frac{(\sum_i y_i)^2}{n} = \underline{\qquad\qquad}$$

$$r = \frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}} = \underline{\qquad\qquad}$$

(ii) The correlation between y and x is _____.

(iii) The proportion of variation explained by the regression model is

$$R^2 = r^2 = \_\_\_\_\_$$

(b) $b = \dfrac{L_{xy}}{L_{xx}} = \_\_\_\_\_$

$a = \bar{y} - b\bar{x} = \_\_\_\_\_$

The regression line is:

_____.

(c) The t-test for the significance of the slope parameter:

1. Hypotheses: _____
2. Test statistic:
$T_0 = \dfrac{b}{s/\sqrt{L_{xx}}} = \_\_\_\_\_$ where
$s^2 = \dfrac{L_{yy} - bL_{xy}}{n-2} = \_\_\_\_\_$
3. P-value: _____
4. Conclusion: Since the P-value < 0.05, the data are against H_0. The regression model is significant.

(d) Predicted value:
$\hat{y} = a + bx = \_\_\_\_\_$.

_____.

3. Carbon monoxide in cigarettes is thought to be hazardous to the fetus of a pregnant woman who smokes. A researcher investigates whether low-tar cigarettes help pregnant women to reduce the amount of carbon monoxide

in their blood. Blood samples were drawn from 10 pregnant women after smoking ordinary cigarettes. The measurements taken were the percentages of blood hemoglobin bound to carbon monoxide as carboxyhemoglobin (COHb). After a wide enough time interval, the same procedure was applied to the same 10 women after they have shifted to smoke low-tar cigarettes. The results for the 10 women are shown in the following table.

Ordinary, o_i	7.6	4.0	5.0	6.3	5.8	6.0	6.4	5.0	4.2	5.2
Low-tar, l_i	5.3	4.2	3.5	4.9	5.1	3.8	4.6	5.6	3.7	5.0
Difference, $d_i = o_i - l_i$	2.3	-0.2	1.5	1.4	0.7	2.2	1.8	-0.6	0.5	0.2

(a) (i) *[3 marks]* Calculate the sample mean and standard deviation for the *difference* d_i.

(ii) *[6 marks]* Construct a 95% confidence interval for the mean *difference* in percentages of COHb in the blood from smoking an ordinary cigarettes to low-tar cigarettes. State clearly your model assumptions. Does the confidence interval suggest a difference in percentages of COHb at 5% significance level?

(b) A suitable test is performed to infer at the *5% significance level* that there is a *reduction* in the percentage of COHb in the blood of pregnant women who shift from smoking ordinary cigarettes to low-tar cigarettes using the following steps.

(i) *[1 marks]* State the null and alternative hypotheses.

(ii) *[5 marks]* Calculate a suitable test statistic to test your hypothesis in (i).

(iii) *[3 marks]* Calculate the p-value of the test based on the statistic in (ii).

(iv) *[2 marks]* Using the p-value in part (iii), state the conclusion for the test in (b).

Solution:

(a) (i) Sample means: _____ and sample sd: _____.

(ii) The 95% confidence interval for the mean reduction is:

$$(\bar{d} - t_{1-\alpha/2}\,\frac{s_d}{\sqrt{n}},\ \bar{d} - t_{1-\alpha/2}\,\frac{s_d}{\sqrt{n}})$$

$=$ _____

$=$ _____

Assumption: _____.

_____

_____

_____.

(b) (i) _____

(ii) Test statistics:

$t = \dfrac{\bar{d}}{s/\sqrt{n}} =$ _____.

(iii) P-value: _____

(iv) Conclusion: _____
_____.

2011 June Exam - Extended Answer Section

1. The ages of a sample of 25 customers in a department store are presented in the following stem-and-leaf plot:

```
2 | 0378
3 | 1355689
4 | 002344789
5 | 236
6 | 7
7 | 3
```

The stem and leaf represent ten and unit values respectively.

(a) *[1 mark]* Find the range of the data.

(b) *[2 marks]* Find the median of the data.

(c) *[3 marks]* Find the lower quartile, upper quartile and the interquartile range of the data.

(d) *[5 marks]* Find the lower and upper thresholds and hence identify any outliers.

(e) *[2 marks]* If the mean of the data is 41.72, find the new mean based on 24 observations excluding the maximum.

(f) *[3 marks]* Complete the frequency distribution of the data given below:

Class	Midpoint	Frequency
$(18, 28]$		
$(28, 38]$		
$(38, 48]$		
$(48, 58]$		
$(58, 68]$		
$(68, 78]$		

(g) *[4 marks]* Calculate the *grouped* variance of the data, using only your answer from part (f).

Solution:

(a) 1 mark: The range =max-min=_____.

(b) 2 marks: Median = _____.

(c) 3 marks:
The lower group: _____,
Q_1=median of lower group=_____.
The upper group: _____,
Q_3=median of upper group=_____.
IQR=_____

(d) 5 marks:
Lower threshold=$Q_1 - 1.5 \times IQR$=_____
Upper threshold=$Q_3 + 1.5 \times IQR$=_____

_____

(e) 2 marks:
The new mean = _____

(f) 3 marks: The frequency distribution table is

Class i	Midpoint u_i	Frequency f_i	$u_i f_i$	$u_i^2 f_i$
(18,28]				
(28,38]				
(38,48]				
(48,58]				
(58,68]				
(68,78]				
Total				

(g) 4 marks:

$$\text{group mean } \bar{x} = \frac{1}{n}\sum_i f_i u_i = \underline{\qquad\qquad}$$

$$\text{group var. } s^2 = \frac{1}{n-1}\left(\sum_i f_i u_i^2 - n\bar{x}^2\right)$$

$$= \underline{\qquad\qquad\qquad\qquad}$$

2. In the modern Olympic era, performances in track and field have been steadily improving. The table below gives the winning distance (in inches) for the Olympic men's long jump from 1952 to 1984. Below are some summary statistics for a simple regression of distance (y) on year (x).

Year (x_i)	1952	1956	1960	1964	1968	1972	1976	1980	1984
Distance (y_i)	298	308.25	319.75	317.75	350.5	324.5	328.5	336.25	336.25

$$\sum_{i=1}^{9} x_i = 17712, \sum_{i=1}^{9} x_i^2 = 34858176, \sum_{i=1}^{9} y_i = 2919.75, \sum_{i=1}^{9} y_i^2 = 949218, \sum_{i=1}^{9} x_i y_i = 5747113$$

(a) *[3 marks]* Using the information above, calculate the correlation coefficient between x and y.

(b) *[4 marks]* Calculate the estimated regression line.

(c) *[2 marks]* What proportion of the variability in distance is explained by year using the simple linear regression model?

(d) *[3 marks]* In the next Olympics, Carl Lewis (USA) won the gold medal after jumping 8.72 meters. Calculate the residual if you use the regression model to predict the next year's Olympic value. (Hint: 1 inch = 2.54 cm.)

(e) *[8 marks]* Test whether there is a significant linear relationship between years and distance. Justify your answer using an appropriate t-test and perform all the steps.

Solution:

(a) 3 marks:

$L_{xx} = \sum_i x_i^2 - \frac{(\sum_i x_i)^2}{n} =$ _____

$L_{xy} = \sum_i x_i y_i - \frac{(\sum_i x_i)(\sum_i y_i)}{n} =$ _____

$L_{yy} = \sum_i y_i^2 - \frac{(\sum_i y_i)^2}{n} =$ _____

$r = \frac{L_{xy}}{\sqrt{L_{xx} L_{yy}}} =$ _____

(b) 4 marks

$b = \frac{L_{xy}}{L_{xx}} =$ _____

$a = \bar{y} - b\bar{x} =$ _____

The regression line is: distance = -1818.74+1.089 X year.

(c) 2 marks: The proportion of variation explained by the regression model is

$$R^2 = r^2 = 0.754^2 = 0.568516$$

(d) 3 marks:
$\hat{y}_{10}=a+bx_{10}=$ _____
Since $y_{10} = 8.72(\text{m}) =$ _____
The residual$=y_{10}-\hat{y}_{10} =$ _____

(e) 8 marks: The t-test for the significance of the slope parameter:
1. Hypotheses: _____
2. Test statistic: $T_0 = \dfrac{b}{s/\sqrt{L_{xx}}} =$ _____
where $s^2=\dfrac{L_{yy} - bL_{xy}}{n-2}=$ _____
3. P-value = _____
4. Conclusion: _____
_____

3. A dentist investigates if among his patients, the number problem teeth that smokers have are more than the non-smokers. He takes a random sample of 10 smoker patients and 8 non-smoker patients and counts their number of

problem teeth. The result is summarised in the following table.

Patient	**Observation**
Smoker	6 3 3 3 4 3 6 5 5 4
Non-smoker	4 2 1 2 3 1 2 3

Perform a suitable test to infer at the 1% significance level that smoker patients have more problem teeth on average, than non-smoker patients using the following steps.

(a) *[2 marks]* State the null and alternative hypotheses.

(b) *[4 marks]* Calculate the sample mean and variance of problem teeth for smokers and non-smokers.

(c) *[6 marks]* Calculate a suitable test statistic to test your hypothesis in part (a).

(d) *[3 marks]* Calculate the P-value of the test based on

(e) *[3 marks]* Using the P-value in part (d), state the conclusion for the test in part (a).

(f) *[2 marks]* Do you think any assumption(s) needed for the test you just performed might have been violated? Explain your answer.

Solution: Two independent samples t-test:

(a) 2 marks: Hypotheses: _____

(b) 4 marks:
Sample means: _____ and _____ for smokers and non-smokers respectively.
Sample variance: _____ and _____.

(c) 6 marks: Test statistics:

$$s_p^2 = \frac{s_1^2 \times (n_1 - 1) + s_2^2 \times (n_2 - 1)}{n_1 + n_2 - 2}$$

$$= \underline{\qquad\qquad\qquad\qquad\qquad\qquad}$$

$$T_{16} = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \underline{\qquad\qquad\qquad\qquad\qquad\qquad}$$

(d) 3 marks: P-value = _____.

(e) 3 marks: Conclusion: _____

_____

_____

(f) 2 marks: _____

_____

_____

2010 June Exam - Extended Answer Section

1. A new material, that is applied to wounds to stop bleeding, has been investigated and compared to traditional material. In a random sample of six volunteers, on each person two similar wounds were inflicted. One of the wounds was treated with the new material and the other with the traditional material. The following information about time until bleeding stops was recorded (in seconds):

Person	1	2	3	4	5	6
Traditional material	86	77	78	85	75	73
New material	80	72	69	76	88	71
Difference	6	5	9	9	-13	2

   (a) *[2 marks]* Calculate the sample mean and the sample standard deviation of the data corresponding to wounds treated with the new material.

   (b) *[6 marks]* Assuming a normal population distribution, construct a 98% confidence interval for the mean time until bleeding stops for all patients on which the new material is applied.

   (c) *[12 marks]* The manufacturer of the new material claims that on average their material stops bleeding in a shorter period of time than the traditional material. Check this claim by following the steps below. (i) Set up the null and the alternative hypotheses specifying all parameters. (ii) Identify a suitable test statistic and calculate its value using the given information.

(iii) Calculate the corresponding P-value. (iv) Draw your conclusion based on the P-value in (iv). (v) Specify any assumptions you may need to validate the above test (no need to verify them).

Solution:

(a)

$\bar{x}_{NM} =$ _____

$s^2_{NM} =$ _____

$=$ _____

$s_{NM} =$ _____

(b) 98% CI for μ_{NM} is

_____

_____

(c) (i) Hypotheses: _____
where

μ_{NM} is the average time until bleeding stops with new material

μ_{TM} is the average time until bleeding stops with traditional material

(ii) Test statistic:

_____

where _____

_____

_____

(iii) P-value: _____

(iv) Conclusion: _____

_____

(v) Assumption: _____

_____

2. In a recent study of chronic fatigue syndrome, $\frac{8}{9}$ of the patients received the treatment A and the remainder received the treatment B. It is known that the treatment A is successful with probability $\frac{7}{10}$ and that the treatment B is successful with probability $\frac{17}{20}$.

   (a) *[5 marks]* Suppose that a patient is randomly selected from this study. Using a tree diagram (or otherwise) calculate the probability that the patient was treated successfully.

   (b) *[8 marks]* In a similar study with 9 patients, the probability of successful treatment is 0.64. Assuming the patients' results are independent of each other, find the probability that at least 2 were treated successfully. What is the expected value and the variance of the number of successfully treated patients?

   (c) *[7 marks]* Calculate the probability in part (b) using the normal approximation.

Solution:

(a) The tree diagram:

(b) Let X = no. of successfully treated patients out of 9 studied patients. Then $X \sim$ _____.

$$
\begin{aligned}
P(X \geq 2) &= \\
&= \\
&= \\
&= \\
&= \\
E(X) &= \\
\mathrm{Var}(X) &=
\end{aligned}
$$

(c) You haven't learned normal approximation to a binomial random variable.

3. A 1996 medical survey examined the relationship between estrogen replacement therapy (ERT) and premature death

rates among post-menopausal women. Researchers searched the medical records of the Kaiser Permanente Medical Care Program in Oakland, California for women born between 1900 and 1915 who had taken estrogen supplements for at least one year starting in 1969. There were 232 such women and 53 of them had died prematurely from different causes. The researchers also selected a sample of records of women born between 1900 and 1915 who had not used ERT at all. There were 222 women in this sample and 87 of them had died prematurely. Assume that these samples of women are representative samples from the populations of women born between 1900 and 1915 who did and did not take estrogen supplements.

(a) *[9 marks]* Test the hypotheses $H_0 : p_{ERT} = 0.4$ vs. $H_1 : p_{ERT} < 0.4$, where p_{ERT} is the premature death rate in the population of all women born between 1900 and 1915 who did take estrogen supplements.

(b) *[11 marks]* Fill in the following two-way table and test whether varaibles death and ERT are independent.

Death	ERT		
	Use	Not use	Total
Premature			
Non-premature			
Total			

Solution:

(a) The sample proportion is _____.

The test statistic:

$Z_{\text{obs}} =$ _____

P-value: _____ Conclusion: _____

_____

_____

(b) The two-way table is

Death	ERT		
	Use	Not use	Total
Premature Non-premature			
Total			

1. Hypotheses:

_____ vs

_____

2. Test statistic:

$$E_{11} = \_\_\_\_\_$$

$$E_{12} = \_\_\_\_\_$$

$$E_{21} = \_\_\_\_\_$$

$$E_{22} = \_\_\_\_\_$$

$$X^2_{\text{obs}} = \_\_\_\_\_$$

$$= \_\_\_\_\_$$

$$= \_\_\_\_\_$$

3. P-value: _____

4. Conclusion: _____

_____

_____

2009 June Exam - Some MC questions

1. Let $Z \sim N(0,1)$. Find k such that $P(|Z| > k) = 0.05$.

   Solution: $P(|Z| > k) = 0.05$ refers to a two-sided area being 0.05. Hence for the upper sided,

   $$P(Z > k) = \underline{\qquad\qquad}$$
   $$\Rightarrow \quad P(Z < k) = \underline{\qquad\qquad}$$

   From the normal table, $k =$ ____.

2. Let $\bar{X}$ be the mean of a sample of size 25 drawn from $N(10, 64)$. State the sampling distribution of $\bar{X}$. Find $P(\bar{X} > 12)$.

   Solution: Since $\bar{X} \sim$ _____ $\Rightarrow \sigma =$ _____,

   $P(\bar{X} > 12) =$ _____.

3. If $n = 10$, $\bar{x} = 12$ and $s^2 = 8$ in Q3, find a 98% CI for μ.

   Solution: The 98% CI for μ is

   _____

2009 June Exam - Extended Answer Section

1. (a) An experiment looked at the effect of position on the level of blood pressure. In the experiment 32 patients had their blood pressures measured while lying down with their arms at their sides (R) and again standing with their arms supported at heart level (S). The differences, $x = R - S$ in their systolic blood pressures (Bernard Rosner, Fundamental of Biostatistics (2006), Thomson p39) are:

-8 -6 -2 0 2 2 4 4 4 6 6 6 8 8 8 8

5 10 10 10 12 12 12 14 14 14 14 14 16 18 26 28

Comment on the effect of position on the levels of systolic blood pressures. Obtain a suitable double stem and leaf plot for these data. Given that $\sum x = 282$ and $\sum x^2 = 4340$, find the proportion of observations in $(\bar{x} - 2s, \bar{x} + 2s)$, where $\bar{x}$ is the sample mean and s is the sample standard deviation.

Solutions: Comment: _____

_____

_____.

A double stem-and-leaf plot is the stem-and-leaf plot with two lines for each stem unit. In the following plot, the stem unit = 10 and leaf unit =1.

$n = 12$

Now the interval

and hence there are _____ of the observations are in this interval.

(b) A surgical procedure can be classified as unsuccessful (U) or successful (S) with probabilities P(U)=0.15 and P(S)=0.85. If four patients receive this procedure in turn and their outcomes are independent, compute the probability that the first three are successful and the last one is unsuccessful. What is the probability of having any three of the operations successful?

Solution: The probability of 3 successes and then a failure is

$P(SSSU) =$ _____

Let X be the number of successful operations out of 4. Therefore, $X \sim$ _____ and hence

$P(X = 3) =$ _____

$=$ _____

2. (a) A consumer group claims that the cure rate of a new drug for skin cancer is less than that of the previous rate of 0.90. To test this a medical team took a random sample of 280 patients and found that only 240 said they obtained relief using this medication. Based on a suitable statistical argument verify the consumers claim.

Solution: Let p be the true cure rate using the new drug. The one-sample z-test on the true proportion p is:

1. Hypotheses: _____
2. Test statistic: _____
3. P-value: _____

4. Conclusion: _____
_____
_____
_____.

(b) As a part of an experiment in a community counselling center, 42 new clients are assigned at random to be interviewed by either a student or a staff clinician. The variable of interest is the number of days, x between the initial visit and the follow-up visit. The resulting data are summarized as follows:

	Student	Staff
n	23	19
$\bar{x}$	51	46
s	12.8	13.4

Do these data suggest that the mean interval between visits is significantly greater when the interviewer is a student than when the interviewer is a staff member? *Hint: Conduct an appropriate t test assuming normality and draw your conclusion based on the associated P-value.*

Solution: Let μ_1 be the mean interval for student interviewers and μ_2 be the mean interval for staff interviewers. The two-sample $t-$test for the difference in means $\mu_1 - \mu_2$ is

1. Hypotheses: _____
2. Test statistic:

_____

3. P-value: _____
4. Conclusion: _____
_____
_____
_____
_____.

3. (a) The clinic of a health plan selects a random sample of records for 20 male members and summarizes the following information on age (x) and the number of visits to the clinic (y) in the past year: $\sum x = 850$, $\sum y = 34$, $\sum x^2 = 41128$, $\sum y^2 = 88$ and $\sum xy = 1796$. Find the correlation between x and y and interpret this value. Find the regression line relating number of visits to the age of the plan member. Given that a male of 39 years visited the clinic twice, find the residual of y when $x = 39$.

Solution:

Correlation coefficient:

Hence the relationship between X and Y is _____ _____.

Regression line:

Fitted line: _____

When $x = 39$: _____

The residual: _____

(b) A biologist wants to test the claim that the number of red, white, yellow and pink flowered plants resulting after cross-pollination follow the ratio 1:2:2:5. After cross- pollination of 300 seeds, he obtains the results below:

Colour	Red	White	Yellow	Pink	Total
Number	26	55	61	158	300

Test whether the given model fits the data well.

Solution We first complete the following table:

	Red	White	Yellow	Pink	Total
O_i	26	55	61	158	300
E_i					300

The χ^2 GOF test is

1. Hypotheses: _____ vs
   _____
2. Test statistic:

   _____
3. P-value: _____
4. Conclusion: _____
   _____
   _____
   _____

2007 June Exam - Extended Answer Section

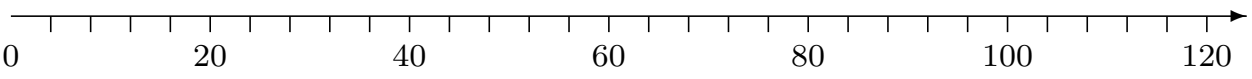
1. *[10 marks]*
The following set of data gives the number of passengers on 52 runs of a ferryboat:

32	35	37	40	41	43	45	49	50	50	51	52	53
53	54	56	57	58	58	59	60	61	62	62	64	64
65	65	66	67	68	70	71	71	73	75	75	77	78
78	80	80	82	84	86	89	91	92	95	97	104	107

(a) *Find* the first quartile, Q_1 and the third quartile, Q_3 of these data. *Produce* a boxplot for this data set indicating any outliers. *Find* the proportion of observations that lie in the interval (LT,UT)? (Recall that $LT = Q_1 - 1.5 * IQR$ and $UT = Q_3 + 1.5 * IQR$ refer to the lower and upper threshold values, and IQR is the interquartile range.)

Solution: Half of data is the first 52/2=26 data in the first two rows. Hence

The boxplot is



The proportion of observations in (LT,UT) is _____ .

(b) Given that $\sum x_i = 3432$ and $\sum x_i^2 = 242726$ for the above data, *calculate* the mean and standard deviation of these data.

Solution: The mean and standard deviation are

$$\text{mean} = \_\_\_\_\_$$

$$s^2 = \_\_\_\_\_$$

$$= \_\_\_\_\_$$

$$s = \_\_\_\_\_$$

2. *[10 marks]*
In a study of 10 patients with hypertriglyceridemia the cholesterol (x) and triglyceride (y) levels were measured after a special diet. The following summary is available for x and y:

s_{xx}=22.00, s_{yy}=131.22, s_{xy}=34.90, $\sum x$=67.3 and $\sum y$=59.1.

Note that in past exams (prior to 2008), L_{xx}, L_{yy}, L_{xy} have been replaced by s_{xx}, s_{yy}, s_{xy} respectively.

(a) Assuming a normal distribution for y, *compute* a 95% confidence interval for the mean triglyceride levels.

Solution: We have

_____

_____

Now a 95% CI for μ_y is:

_____

$=$ _____

$=$ _____

(b) (i) *Compute* r, the correlation coefficient between the two variables, x and y. *Find* the proportion of variability in y explained by a linear regression of y on x.

(ii) *Find* the least squares regression line in order to predict the amount of triglyceride. *Predict* the amount of triglyceride when the cholesterol level is 5.

Solution:

(i)

_____

The proportion of variability explained by the regression line is $r^2 =$ _____.

(ii) The slope is

_____

The y-intercept is

_____

The regression line is _____

When $x = 5$, _____

3. *[10 marks]*

(a) Researchers studied the entry and distribution of the drugs chloropromazine (CPZ) and vinblastine (VBL) into human red blood cells during endocytosis (absorption). The following data are the percentages of CPZ found in membrane fractions of red blood cells suspended in Hanks' solution for samples from eight donors.

Donor	1	2	3	4	5	6	7	8
CPZ percentages	5	7	10	15	17	14	14	27

Assuming the percentages have a symmetric distribution, test the claim, $H_0 : \mu = 10$ against $H_1 : \mu > 10$ using the sign test.

Solution: Not covered in 2011.

(b) It is known that a toss of a pair of fair dice gives a sum of the two digits from 2 to 12 with $P(sum = 7) = 1/6$ and $P(sum = 11) = 1/18$.
In 360 tosses of a pair of dice, 74 "sevens" and 24 "elevens" are observed. By considering the three outcome classes, "7", "11" and "Other", use a goodness-of-fit test to *test* the hypothesis that the dice are fair.

Solution: The outcomes are "sum=7", "sum=11" and "sum=neither 7 nor 11" which correspond to a categorical variable. Hence the χ^2 GOF test is

1. Hypotheses: _____

   _____

2. Test statistic: Complete the following table to calculate the expected frequencies E_i:

Outcome	**Obs. freq.** O_i	**Expected freq.** $E_i = np_i$	$\frac{(O_i - E_i)^2}{E_i}$
7	74	_____	_____
11	24	_____	_____
Other	262	_____	_____
Total	360	360	_____

$\chi^2_{\text{obs}} =$ _____

3. P-value: _____
4. Conclusion: _____

   _____

   _____

   _____

4. *[10 marks]*
   A researcher was given a project to compare the abrasive wear of two different laminated materials, X and Y. Twelve pieces of randomly selected material X were tested, by exposing each piece to a machine measuring wear. Similarly ten randomly selected pieces of material Y were tested.

The following summary gives the average wear and the standard deviation of each material:
For material X: $\bar{x} = 85,\ s_x = 4$
For material Y: $\bar{y} = 82,\ s_y = 5$
It is believed that the material X is easier to wear than the material Y. The researcher plans to apply the two sample t test to verify this claim.

(a) *State* the assumptions required for the test.

(b) *Set up* the hypotheses carefully defining any parameters to be used.

(c) *Provide* an expression for the P-value associated with the test in (ii) after evaluating the test statistic.

(d) Draw your conclusion using a suitable statistical argument based on (iii).

Solution:

(a) _____

_____

_____

(b) Let μ_x be the population mean wear of x.
Let μ_y be the population mean wear of y.
Hypotheses: _____

(c) The pooled variance estimate is:

_____

and the test statistic is:

(d) Conclusion:

5. *[10 marks]*

(a) It is known that the probability of a randomly chosen individual from a certain county with an IQ less than 106 is 0.7. Let X be the number of people in a sample of 50 with an IQ less than 106. *Write down* an expression for the exact probability, $P(20 \leq X < 30)$. *Find* an approximate value for this probability. *Hint: Use a suitable normal approximation with correction for continuity.*

Solution: We have $X \sim$ _____.

The exact probability is

The normal approximation with continuity correction is not covered in 2011.

(b) A hot dog manufacturer asserts that his new product has an average fat content of 18g with a standard deviation of 1g. To test this claim (ie. $H_0 : \mu = 18$ against $H_1 : \mu > 18$), a random sample of 36 hot dogs was examined by an agency. The sample average fat content was 18.4g. Assuming the population standard deviation is known and $\sigma = 1$, use the Central Limit Theorem (CLT) to *find* the P-value associated with the above test given by $P(\bar{X} \geq 18.4)$.

Solution: Since the population $\sigma^2 = 1$ is known, $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ i.e. _____ by CLT under $H_0 : \mu = 18$. Hence the P-value is

2008 June Exam - Extended Answer Section

1. *[10 marks]*
   The following are the hemoglobin readings for each of 48 healthy adult women in an agricultural town of Tomba.

   12.08 12.53 12.57 12.63 12.79 13.03 13.03 13.09 13.28 13.32 13.39 13.42
   13.42 13.51 13.51 13.54 13.80 13.82 13.87 13.96 14.06 14.10 14.10 14.16
   14.32 14.44 14.55 14.60 14.70 14.87 14.90 14.92 14.94 15.05 15.07 15.11
   15.19 15.27 15.49 15.49 16.13 16.17 16.28 16.40 16.48 17.03 17.48 18.70

   (i) Obtain a stem and leaf plot for this data. *(Hint: Take the two digits to the left of the decimal place as stems and the two digits to the right as leaves*

   (ii) Find the five number summary for the data.

   (iii) Identify any outliers.

   (iv) Sketch the boxplot of the data.

   (v) Describe the shape of the data.

2. *[10 marks]*
   (a) The probabilities that a husband and wife will be alive 25 years from now are given by 0.70 and 0.80 respectively. Find the probability that in 25 years at least one will be alive. *(Assume that the events that the husband and wife respectively, will be alive in 25 years are independent.)*

   (b) Mendellian inheritance predicts that the ratio of red, white and pink should be 1:1:2 in cross-pollination. A biologist wanted to test this claim and counted the

number of red, white and pink flowered plants resulting after cross-pollination of 260 white and red sweet peas. The results were:

Colour	Red	White	Pink	Total
Number	72	63	125	260

Test the null hypothesis that the model fits well for the given data.

3. *[10 marks]*
The following table shows the ages x (in months) and yield y (in kilograms per year) of 12 orange trees grown under special conditions:

Age, x	56	42	72	36	63	47	55	49	38	42	68	60
Yield, y	147	125	160	118	149	128	150	145	115	140	152	155

($\sum x = 628$, $\sum y = 1684$, $\sum x^2 = 34414$, $\sum y^2 = 238822$, and $\sum xy = 89894$).

(a) Explain why it is reasonable to use the linear least-squares regression of y on x for these data. *[Hint: Find the correlation between x and y and interpret this value].*

(b) Estimate the yield of a tree whose age is 45 months.

4. *[10 marks]*
Angioplasty is a surgical procedure that uses a catheter and a balloon to open up blocked arteries, usually those of the heart. A cardiologist wanted to verify that a second supplier of angioplasty ballons was producing balloons of larger strength than the current supplier. 10 balloons from

the current supplier A and 12 from the new supplier B were randomly selected and tested for their bursting pressures, measured in bars (1 bar=1 kg/cm^2). The data obtained were:

Supplier	Mean	SD	Size
A	17.42	2.04	10
B	18.72	1.82	12

(a) State the null and alternative hypotheses of the required test.

(b) Conduct an appropriate t test, and draw your conclusion based on the associated P-value.

Numerical answers to 2008 Exam.:

1. (ii) min=12.08; Q_1=13.42; Q_2=14.24; Q_3=15.15; max=18.70
   (iii) IQR=1.73; LT=10.82; UT=17.75; one outlier at 18.70
   (v) slightly right skewed

2. (a) 0.94 (b) $\chi^2_{\text{obs}} = 1.008$, consistent with H_0.

3. (a) $r = 0.8967$; r^2=80.4% (b) $y = 80.701 + 1.139x$; 132kg

4. (a) $H_0 : \mu_B = \mu_A$ vs $H_1 : \mu_B > \mu_A$
   (b) $t_{\text{o}} = 1.580$, consistent with H_0.

GOOD LUCK!