Problem 1

a) F
b) F
c) T
d) T
e) F
f) F
g) T
h) T
i) T
j) F
k) F
l) T
m) F
n) F
o) F
p) T
q) T
r) T
s) F
t) F
u) F
v) F

Problem 2

a) $\frac{\text{\# with gene}}{\text{total people}} = \frac{39+32}{39+31+32+17} = 59.66\%$

b) $\frac{\text{\# not respond}}{\text{total people}} = \frac{32+17}{39+31+32+17} = 41.18\%$

c) $\frac{\text{\# gene \& responded}}{\text{\# responded}} = \frac{39}{39+31} = 55.71\%$

d)

	Gene	No Gene	Total
Yes	39	31	70
No	32	17	49
Total	71	48	119

→

	Gene	No Gene
Yes	41.76	28.24
No	29.24	19.76

} Expected Counts

$* \text{ Expected} = \frac{\text{Row} \times \text{Col}}{\text{Total}}$

$= \frac{71 \times 49}{119} = 29.23529$

e) "Chi-Squared Residuals"

	Gene	No Gene
Yes	0.1824	0.2697
No	0.2605	0.3855

$= \frac{(31-28.24)^2}{28.24}$

* 'Chi-Squared Residual' ('Chi-value')

$\frac{(obs-exp)^2}{exp}$

* The 'ordinary' residual is just obs − pred. So for 31, the residual is...

$31 - 28.24 = 2.76$

f) $df = (\#row - 1)(\#col - 1)$

$= (2-1)(2-1)$

$= 1$

g) $\begin{cases} H_0: \text{Rows/Col are independent} \\ H_a: \text{There is an association between rows/col.} \end{cases}$

h) $X^2 = 0.1824 + 0.2697 + 0.2605 + 0.3855 = 1.0981$

$\underset{\text{dof 1}}{P(X^2 \geq 1.0981)} > 0.25$ (exact value 0.2938)

So fail to reject null. No evidence to suggest relationship between row/col.

Problem 3

a) 0.50

b) $\frac{\#\text{heads}}{\text{flips}} = \frac{5103}{10{,}000} = 0.5103$

c)

H	T
5103	4897

But flipped 10,000 times & expected 50% of each...

Expected

H	T
5000	5000

'Residuals'

H	T
2.1218	2.1664

$\frac{(\text{obs} - \text{pred})^2}{\text{pred}} = \frac{(5103 - 5000)^2}{5000}$

d) $df = \#\text{col} - 1 = 2 - 1 = 1$

$\chi^2 = 2.1218 + 2.1664 = 4.2882$

$P(\chi^2 \geq 4.2882) = 0.05$ (really between 0.05 & 0.025)
dof 1
↳ Note: exact value 0.0404

Use $\alpha = 0.05$. So there is enough evidence to suggest that the coin is not a fair coin.

* Note: This is Chi-Squared Goodness of Fit.

Problem 4

a) $\frac{\#\text{VE with emp.}}{\#\text{VE}} = \frac{52}{52+44} = 54.17\%$

b) $\frac{\#\text{avg. not emp.}}{\#\text{not emp.}} = \frac{60}{60+47+36+44} = 32.09\%$

c) $\frac{\#\text{SE}}{\text{total}} = \frac{50+47}{35+60+50+47+61+36+52+44} = 25.19\%$

d)

	A	SE	E	VE
Y	48.85	49.89	49.89	49.37
N	46.14	47.11	47.11	46.63

$= \frac{95 \cdot 198}{385}$

$= \frac{96 \cdot 198}{385}$

Note: Totals

	A	SE	E	VE	Total
Y	35	50	61	52	198
N	60	47	36	44	187
Total	95	97	97	96	385

e)

	A	SE	E	VE
Y	-13.85	0.114	11.11	2.629
N	13.857	-0.114	-11.114	-2.63

* Note: This is an 'ordinary' residual table.

$= \text{obs} - \underbrace{\text{pred}}_{\text{expected}} = 44 - 46.63$

$= \text{obs} - \text{pred}$
$= 35 - 48.85$

f) 'Chi-Squared Res.' Table

	A	SE	E	VE
Y	3.93	≈ 0	2.47	0.140
N	4.16	≈ 0	2.62	0.149

$= \frac{(47-47.11)^2}{47.11} = 0.00025$

$= \frac{(44-46.63)^2}{46.63} = 0.148336$

$X^2 = 3.93 + 4.16 + 0 + 0 + 2.47 + 2.62 + 0.140 + 0.149 = 13.469$

$df = (4-1)(2-1)$
$= 3 \cdot 1$
$= 3$

$\begin{cases} H_0: \text{the row/col are indep} \\ H_a: \text{the row/col are dep.} \end{cases}$

$P(\underbrace{\chi^2 \geq 13.469}_{\text{doF } 3}) = 0.005$

exact value 0.003708

So there is sufficient evidence to suggest (at $\alpha = 0.05$ & 0.01) there is a relation between row/col.

Problem 5: Researchers try to see if there is a connection between age and the type of alcohol one regularly drinks. Their data is summarized below:

TABLE 5. 'Regular' alcohol usage among different age groups.

	15–22	22–29	30–35	35–40	40–60	Total
Liquor	1271	1205	607	345	401	3829
Beer	1497	2112	2115	2456	3019	I
Wine	262	IV	II	1772	6998	12,034
None	2063	1206	2097	2419	10,042	17,827
Total	5093	III	6333	6992	20460	44,889

TABLE 6. Expected values of 'regular' alcohol usage among different age groups.

	15–22	22–29	30–35	35–40	40–60
Liquor	434	513	540	596	1745
Beer	1271	VI	1580	1744	5104
Wine	1365	1611	1698	VII	5485
None	V	2387	2515	2777	8125

TABLE 7. 'Chi–Squared Residual' values of 'regular' alcohol usage among different age groups.

	15–22	22–29	30–35	35–40	40–60
Liquor	1,610.97	934.66	8.26	X	1,035.36
Beer	40.336	250.053	IX	290.307	851.991
Wine	891.625	VIII	19.892	5.598	417.358
None	0.807	584.449	69.49	46.096	452.091

(a) Fill in the missing values from the table of counts.
(b) Fill in the missing values from the expected count table.
(c) Fill in the missing values from the 'chi–residual' table.
(d) What is the degree of freedom?
(e) State H_0 and H_a.
(f) What is the p–value? At $\alpha = 0.05$, what is the conclusion?
(g) Should the conclusion have been expected by looking at the original table?

Problem 5

a)

$i = 44889 - 17827 - 12034 - 3829 = 11199$

$ii = 6333 - 2097 - 2115 - 607 = 1514$

$iii = 44889 - 5093 - 6333 - 6992 - 20460 = 6011$

$iv = 6011 - 1206 - 2112 - 1205 = 1488$

b)

$$v = \frac{5093 \cdot 17827}{44889} = 2022$$

$$vi = \frac{6011 \cdot 11199}{44889} = 1500$$

$$vii = \frac{6992 \cdot 12034}{44889} = 1875$$

c)

$$viii = \frac{(1488 - 1611)^2}{1611} = 9.39$$

$$ix = \frac{(2115 - 1580)^2}{1580} = 181.155$$

$$x = \frac{(345 - 596)^2}{596} = 105.706$$

d) $dF = (\# col - 1)(\# row - 1) = (5-1)(4-1) = 4 \cdot 3 = 12$

e) $\begin{cases} H_0: \text{Row/Col. indep.} \\ H_a: \text{Row/Col dep.} \end{cases}$

f) X^2 = sum 'Chi Squared Res' Table = 7805.9578 (may be diff. due to rounding)

$P(X^2 \geq 7805.9) < 0.00001$. Reject null. There is relation between row/col.

$\underset{dof\ 12}{}$

g) Yes, 'easy' to get association with large N chi-squared tables.

Problem 6

Income	0-20	20-40	40-60	60-80	80-100
Count	1215	1232	1205	921	552

Total: 5125

Expected Count Table:

Income	0-20	20-40	40-60	60-80	80-100
Count	1204.38	1250.50	1142.88	937.875	589.375

= expected proportion × total = 0.235 × 5125

= 0.244 × 5125

'Chi-Squared Residual' Table

Income	0-20	20-40	40-60	60-80	80-100
Res.	0.093	0.274	3.376	0.304	2.37

$= \frac{(1215-1204.38)^2}{1204.38}$

$\frac{(1205-1142.88)^2}{1142.88}$

$dF = \#\,col - 1 = 5 - 1 = 4$

$X^2 = 0.093 + 0.274 + 3.376 + 0.304 + 2.37 = 6.417$

$P(\underset{\text{dof } 4}{X^2 \geq 6.417}) = 0.1699$ (exact value)

$\alpha = 0.10$ (why not?)

There is not enough evidence to suggest that their study consists of a different dist. of Hispanic-Amer. than the rest of the country.

Problem 6: Researchers are trying to analyze various demographics of Hispanic Americans. The factors they are interested in are influenced by one's income level. The researchers want to check that their study 'fits' the national distribution of Hispanic American income levels. On average 23.5% of Hispanic Americans are in the bottom 20% of income levels, 24.4% are in the next 20% of incomes, 22.3% are in the next income bracket, 18.3% are in the next, and 11.5% are in the top 20% of income levels. Their study consists of the following distribution of incomes among the Hispanic Americans surveyed: According to their data, do their surveyed

TABLE 8. Surveyed Hispanic Americans broken down by income level.

Income Level	0–20%	20–40%	40–60%	60–80%	80–100%
Count	1215	1232	1205	921	552

Hispanic Americans 'fit' the average distribution of Hispanic American income levels?

Problem 7: A company is trying to predict the average production costs for the coming fiscal year. They have a statistician create a linear model predicting the cost (in thousands) for production given the amount of items they produce (in thousands). The computer output for the model is found below.

```
Analysis of Variance

          Source      DF   Adj SS   Adj MS  F-Value  P-Value
          Regression  I    93657    V       VII      0.000
          Items       II   93657    VI      VIII     0.000
          Error       III  IV       472.9
          Total       16   100750

Model Summary

                  S    R-sq   R-sq (adj)  R-sq (pred)
                 IX   92.96%      92.49%       90.47%

Coefficients

          Term        Coef  SE Coef  T-Value  P-Value   VIF
          Constant     2.0     10.1     0.20    0.846
          Items     10.101    0.718     X       0.000  1.00

The regression equation is

production cost = ______________________________________
```

Problem 7

a) $I = 1$

$II = 1$

$III = 15$

IV: $MSE = \frac{SSE}{DFE} \rightarrow SSE = MSE \cdot DFE$

$= 472.9 \cdot 15$

$= 7093.5$

V: $MSR = \frac{SSR}{DFR} = \frac{93657}{1} = 93657$

VI: $MSI = \frac{SSI}{DFI} = \frac{93657}{1} = 93657$

VII: $F = \frac{MSM}{MSE} = \frac{93657}{472.9} = 198.048$

$VIII$: Same as VII

IX: $S^2 = \frac{SSE}{DFE} = MSE = 472.9 \rightsquigarrow S = \sqrt{472.9} = 21.7463$

X: $t = \frac{\text{coef}}{\text{SE coef}} = \frac{10.101}{0.718} = 14.0682$

Reg. Equation:

production cost = 2.0 + 10.101 Items

b) Total = 16 + 1 = 17

c) 2.0 + 10.101 (18) = 183.818 (or \$183,818)

d) $R = \sqrt{R^2} = \sqrt{0.9296} = 0.9642$

e) $R^2 = 92.96\%$

f) 0.9296

g) 2.0 : This is the fixed cost for the company.

h) est ± t^* SE

Use $t(15)$, want 95%

$t^* = 2.131$

$10.101 \pm (2.131) 0.718$

$(8.578, 11.624)$

There is a 95% chance that on average for inc. the number of items produced by 1000 increases costs between 8.57 & 11.62 thousand dollars.

i) $\begin{cases} H_0: \beta_1 = 0 \\ H_a: \beta_1 \neq 0 \end{cases}$

$p = 2 \cdot 0.000 = 0.000 \xrightarrow{\alpha = 0.05}$ Reject null.

$df = 15$ (DFE)

There is a relationship between items & prod. cost.

(Can't say for sure lin. just that there is one.)

(a) Fill in the missing items in the tables above.
(b) What was the total amount of data points used to create the model?
(c) Predict the average production cost if the company produces 18,000 items.
(d) What is the correlation coefficient?
(e) What is the coefficient of determination?
(f) What proportion of the variation in y is explained by the variation in x for this model?
(g) What is the constant in the model? Interpret the constant in the context of the problem.
(h) Find a 95% confidence interval for β_1. Interpret the result.
(i) Conduct a hypothesis test for $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$. State the p-value, degrees of freedom, and interpret the result carefully. [Use $\alpha = 0.05$.]

Problem 8: A company is trying to determine if more experienced workers are more productive than 'fresh' employees. The company hires a statistician to create a linear model predicting the average number of new customers an employee processes each year based on the number of years the employee has been at the company. The model is summarized below.

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	727.9	727.88	8.75	V
Years	1	727.9	727.88	8.75	VI
Error	9	II	IV		
Total	I	III			

Model Summary

S	R-sq	R-sq (adj)	R-sq (pred)
VII	VIII	43.65%	14.17%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	99.99	IX	X	XI	
Years	2.572	0.870	2.96	0.016	1.00

The regression equation is

Number Customers = ______________________________

(a) Fill in the missing entries in the tables above. [Note: $\sum(x_i - \bar{x})^2 = 110$ and the average employee used to create the model had worked at the company for 5 years.]
(b) Use the model to predict the number of customers an employee would add on average that year if the employee had worked at the company for 9 years.
(c) What is the correlation coefficient?

Problem 8

a) $I = 1 + 9 = 10$

$IV: \quad F = \frac{MS}{MSE} \rightarrow MSE = \frac{MS}{F} = \frac{727.88}{8.75} = 83.22$

$II: \quad SS = MS \cdot DF = 83.22 \cdot 9 = 749.0$

$III = 727.9 + 749.0 = 1476.9$

$V: \quad dF = 9 \quad ; \quad F(1,9) \rightarrow 0.016$

$VI: \quad dF = 9 \; ; \; F(1,9) \rightarrow 0.016$

$VII: \quad S = \sqrt{MSE} = \sqrt{83.22} = 9.12249$

$VIII: \quad R^2 = \frac{SSM}{SST} = \frac{727.9}{1476.9} = 49.29\%$

$IX: \quad SE_{b_0} = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}} = 9.12249\sqrt{\frac{1}{11} + \frac{5}{110}} = 5.15$

$X: \quad t = \frac{coef}{SE} = \frac{99.99}{5.15} = 19.43$

$XI: \quad$ p-value = 0.000

Number Customers = 99.99 + 2.572 Years

b) $99.99 + 2.572(9) = 123.138 \rightarrow 123$

c) $R = \sqrt{R^2} = 0.702068$

d) $R^2 = 49.29\%$

e) 0.4929

f) = $10 + 1 = 11$

g) df = 9

$t^* = 1.833$

Coef ± t^* SE = 99.99 ± (1.833)5.15 → (90.5485, 109.431)

h) β_0 = number new cust. after 0 years ... but then they don't even work at the company yet...

i) $\begin{cases} H_0: \beta_1 = 0 \\ H_a: \beta_1 > 0 \end{cases}$ $t = 2.96$, $p = 0.016$, $df = 9$

Fail to reject null. Not enough evidence to suggest relation between avg. new cust. and # years at company.

j) $p = 0.016$ Enough evidence to suggest relationship but $R^2 = 49.29\%$ so 'probably' not linear

K) Exact same as j.

(d) What is the coefficient of determination?
(e) What proportion of the variation in y is explained by the variation in x for this model?
(f) What was the total number of data points used to create the model?
(g) Construct a 90% confidence interval for β_0.
(h) Does β_0 have any meaning in this problem? Explain.
(i) Conduct a hypothesis test $H_0 : \beta_1 = 0$ versus $H_0 : \beta_1 > 0$ using $\alpha = 0.10$. Interpret your results. Be sure to give the t–value, p–value, and degrees of freedom.
(j) Can we say that the data is linear?
(k) Conduct an F–test for the regression. State H_0 and H_a. How does this test differ from the hypothesis test you performed above?

Problem 9: A statistics student is preparing for an exam. They have to fill in the missing values from the model given below. However, they are nervous and want to be able to check their answers. Fill in the values so that the student will have a solution manual to which compare their answers. Be sure to indicate for the student whether this was a simple linear regression or a multiple linear regression and how many data values were used to create the model.

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	I	0.29477	IV	4.00	0.067
Error	13	0.95920	0.07378		
Lack-of-Fit	11	0.90620	0.08238	3.11	0.268
Pure Error	2	III	0.02650		
Total	II				

Model Summary

S	R-sq	R-sq (adj)	R-sq (pred)
V	VI	17.62%	0.19%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	0.185	0.126	VII	0.166	
Regression	0.541	0.270	2.00	0.067	1.00

The regression equation is

y = ____

Problem 9

I = 1 (notice under coefficients, the only variable is 'Regression')

II = 1 + 13 = 14

III: $SS = MS \cdot DF = 0.02650 \cdot 2 = 0.05300$

IV: $MS = \frac{SS}{DF} = \frac{0.29477}{1} = 0.29477$

V: $S = \sqrt{MSE} = \sqrt{0.07378} = 0.27163$

VI: $R^2 = \frac{SSM}{SST} = \frac{0.29477}{0.95920} = 23.51\%$

VII: $t = \frac{coef}{SE} = \frac{0.185}{0.126} = 1.47$

Total Data Values = 14 + 1 = 15

This is a simple linear regression.

Reg. Eq:

$y = 0.185 + 0.541 \cdot \text{Regression}$

Problem 10: The same student returns to you for more help–you'd better start charging! Fill in the values so that the student will have a solution manual to which compare their answers. Be sure to indicate for the student whether this was a simple linear regression or a multiple linear regression and how many data values were used to create the model. Furthermore, explain to the student whether one can predict y using the input variable(s). Explain whether the model is linear or not. Then help the student construct a confidence interval (a 95% confidence interval) and hypothesis test (of $H_0 : \beta_i = 0$ versus $H_a : \beta_i \neq 0$) on the variable(s).

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	I	3317.23	1658.62	VII	0.000
Var1	II	2653.19	VI	3151.70	0.000
Var2	III	V	664.04	788.81	0.000
Error	IV	27.78	0.84		
Total	35				

Model Summary

S	R-sq	R-sq (adj)	R-sq (pred)
VIII	IX	99.12%	99.01%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	15.081	0.352	42.90	0.000	
Var1	2.5134	0.0448	56.14	0.000	1.00
Var2	-1.2574	0.0448	-28.09	0.000	1.00

The regression equation is

y = ____

Problem 10

I = 2

II = 1

III = 1

IV = 35 - 2 = 33

V: $MS = \frac{SS}{DF} \rightarrow SS = MS \cdot DF$
$= 664.04 \cdot 1$
$= 664.04$

VI: $MS = \frac{SS}{DF} = \frac{2653.19}{1} = 2653.19$

VII: $F = \frac{MS}{MSE} = \frac{1658.62}{0.84} = 1974.55$

VIII: $S = \sqrt{MSE} = \sqrt{0.84} = 0.9165$

IX: $R^2 = \frac{SSR}{SST} = \frac{3317.23}{3345.01} = 99.17\%$

$SST = 3317.23 + 27.78$

Reg. Eq:

$y = 15.081 + 2.5134\, Var1 - 1.2574\, Var2$

* Multiple lin. reg. (used 2 variables)

* Data values = 35 + 1 = 36

* p-value (model) = 0.000 & $R^2 = 99.17\%$. So evidence to suggest relationship between variables & y. Cannot say for sure if lin. (but probably)

* We will do this for i = 2, ie variable 1

$\begin{cases} H_0: \beta_2 = 0 \\ H_a: \beta_2 \neq 0 \end{cases}$ $t = -28.09$, p-value = 0.000 → Reject Ho. There is a relationship

C.I.: $-1.2574 \pm \underbrace{(2.032)}_{t_{95\%,\ dof\ 34}} 0.0448 \rightarrow (-1.348, -1.166)$

Problem 11: A finance company is investigating the effects of different variables on total average pre–tax income (measured in US dollars). The variables used in the model were education (in number of years), age (in years), residence (number of years living at the current residence), savings (in US dollars), debt (in US dollars), and the number of credit cards one has. Mathematicians at the company use a statistics program to create a multivariate linear regression. The output is given below.

```
Analysis of Variance
```

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	I	5673515090	810502156	4.19	XIV
Education	II	2021974124	2021974124	10.45	0.004
Age	III	603358164	603358164	XIII	0.091
Residence	IV	6336044	6336044	0.03	0.858
Employment	V	45377529	XII	0.23	0.633
Savings	VI	727084527	727084527	3.76	0.066
Debt	VII	253851061	253851061	1.31	0.264
Credit Cards	VIII	XI	51710387	0.27	0.610
Error	IX	4257582430	X		
Total	29	9931097520			

```
Model Summary
```

S	R-sq	R-sq (adj)	R-sq (pred)
XV	57.13%	43.49%	8.85%

```
Coefficients
```

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-90642	37172	-2.44	XVIII	
Education	4827	1493	3.23	0.004	1.52
Age	2098	XVII	1.77	0.091	4.53
Residence	XVI	2313	-0.18	0.858	12.71
Employment	1291	2666	0.48	0.633	14.31
Savings	-1.015	0.524	-1.94	0.066	3.00
Debt	-0.854	0.746	-1.15	0.264	2.12
Credit Cards	1348	2607	0.52	0.610	1.60

The regression equation is

income = ____

____

Problem 11

a) $I = 7$

$II = III = IV = V = VI = VII = VIII = 1$

$IX = 29 - 7 = 22$

$X = \frac{SSE}{DFE} = \frac{4257582430}{22} = 193526474$

$XI = SS = MS \cdot DF = 51710387 \cdot 1 = 51710387$

$XII = MS = \frac{SS}{DF} = \frac{45377529}{1} = 45377529$

$XIII = F = \frac{MS}{MSE} = \frac{603358164}{193526474} = 3.12$

XIV : use $F(7, 22)$. $F = 4.19 \rightarrow p = 0.005$ (exact value)

XV : $S = \sqrt{MSE} = \sqrt{193526474} = 13911.4$

XVI : $coef = t \cdot SE = -0.18 \cdot 2313 = -416.34$

$XVII$: $SE = \frac{2098}{1.77} = 1185.31$

$XVIII$: same idea as above: 0.023 (exact)

$income = -90642 + 4827\,Ed + 2098\,Age - 416.34\,Res$

$+ 1291\,Emp - 1.015\,Sav - 0.854\,Debt + 1348\,CC.$

b) At $\alpha = 0.05$: Education

c) At $\alpha = 0.05$: Age, Res., Emp., Savings, Debt, Cred. Cards

d) Plug into the model :

Income = \$26,191.20

e) dof = 22

$t^* = 2.074$

$$\underbrace{-0.854}_{est} \pm \underbrace{(2.074)}_{t^*} \underbrace{0.746}_{SE} \rightarrow (-2.4012, 0.693)$$

f) 5^{th} variable Savings

$$\begin{cases} H_0: \beta_5 = 0 \\ H_a: \beta_5 < 0 \end{cases} \qquad p = 0.066 \qquad \alpha = \begin{cases} 0.10 & \text{Reject} \\ 0.05 & \text{Don't Rej.} \\ 0.01 & \text{Don't Rej.} \end{cases}$$

g) p-value : 0.005. So model sig. Can't say for sure that it is lin.

But likely not lin ($36.05 = R^2$)

h) Uses less variables (it is a 'sub' model)

i) Org: $p = 0.091 \rightarrow$ No

New : $p = 0.002 \rightarrow$ Yes

p-values can change. It is a 'new' model. With | Without other variables then predicts better | worse