



10 Statistics

10.1 Correlation and Regression

1. A new CEO of a renewable energy company is about to hire new employees as part of the company's expansion plan. The company has no set pay structure in their policy documents. The CEO decides to examine the existing relationship between years of experience in the renewable energy sector and salary. The results are shown in the table below. The salary is in thousands of euro.

Years of Exp.	3	5	10	8	12	20	17	2
Salary (€1000's)	36	40	53	56	65	105	88	36

- Draw a scatter diagram of the data
 - Calculate the correlation coefficient; and comment on the correlation.
 - Add the line of best fit to your scatter plot.
 - Find the equation of the line of best fit.
 - In order to stay in line with existing company salaries, what salary should the CEO offer an new employee who has 15 years experience in renewable energy?
2. A math lecturer in a local business university wishes to examine the relationship between how many classes his students miss during the semester, and their grades in the final exam. There were 25 classes in total during the semester, and the grades are marked out of 100. The results are shown in the table below.

No. of Absences	3	0	7	2	1	11	5	16
Final Grade (%)	79	93	62	85	95	32	68	41

- Plot the results on a scatter graph.
- Calculate the correlation coefficient.
- Add the line of best fit to your scatter plot.
- Find the equation of the line of best fit.
- What grade would you expect from someone who missed 4 classes?
- What grade would you expect from someone who missed 20 classes?





3. An ice-cream salesman from Ballybunion wishes to predict his future sales, based on the weather forecast, more importantly the temperature. Over the course of four weekends he tracks the high temperature for the day and his sales for that day. Results are shown in the table below.

Temp($^{\circ}C$)	17	19	22	15	23	25	18	19
Sales (€)	853.50	920	970.50	515	1155	1254	815.50	884

- Represent the data on a scatter graph.
- Calculate the correlation coefficient.
- Does it appear that ice cream sales are affected by temperature? Explain your answer.
- Add the line of best fit to your scatter plot.
- Find the equation of the line of best fit.
- How much should the salesman expect to earn if the the temperature reached $30^{\circ}C$

10.2 Confidence Intervals

- A sample of the results of 200 Leaving Cert students has a mean of 375 points. The standard deviation of the sample is 90 points. Find the 95% confidence interval for the true mean in relation to Leaving Cert points.
- A sample of 40 different days in Limerick hospital found on average there was 25 patients waiting on trolleys. The standard deviation of the sample was 3.5. Find the 95% confidence interval of the mean number of patients waiting on trolleys in Limerick hospital.
- A sample of 350 college students found that on average they spent €40 on alcohol per week. Find the 95% confidence interval for the true mean, if the standard deviation of the sample was €16.
- A study of 50 people found that the average time it took a person to find a new job was 4 months. If the standard deviation of the sample was 2.9 months, find the 95% confidence interval of the true mean.
- Find the 95% confidence interval for the mean attendance at Thomond Park for Munster Rugby Games. A random sample of the attendance is shown.

7862	6982	15824
9865	22145	21987
17569	9513	14562
6598	15354	13974
24865	18954	25630
19541	13254	8751
25630	25630	22159
16789	9512	17498
23541	25630	15971
7875	25630	14654





6. A survey of 1500 third level students found that 33% will go abroad this Summer. Find the 95% confidence interval of the true proportion of students who will go abroad this Summer.
7. An survey done by the Department of Social Protection found that in a sample of 1000 unemployed workers, 25% were not interested in retuning to work. Find the 95% confidence interval of the true proportion of unemployed workers who do not wish to return to work.
8. A survey of 190 families showed that 111 owned at least one pet. Find the 95% confidence interval of the true proportion of families who own at least one pet.
9. A survey found that out of 2000 workers, 750 said they spend at least an hour per day on social networks. Find the 95% confidence interval of the population proportion of workers spend at least an hour a day on social networks.
10. A survey of 500 young teachers (under 30) showed that 35% belonged to a union. Find the 95% confidence interval of the true proportion of young teachers who belong to a union.
11. A survey of 1000 leaving cert students found that 290 sit higher level maths. Find the 95% confidence interval of the true proportion of leaving cert students who sit higher level maths.

10.3 Hypothesis Testing

1. A radio station claimed that only 30% of leaving cert students have a part time job. A sample of 200 leaving cert students found that 39% of them had a job. Is there enough evidence to reject the radio stations claim at the 5% significance level.
2. A drug company produces a new drug to help significantly lower cholesterol. The company claims that the drug works 78% of the time. A group of doctors doubted the company's claim. They prescribed the drug for a group of 2500 patients. After one year 1850 of these patients' cholesterol had been significantly lowered by the drug. Test the company's claim that the drug had an 78% success rate, at the 5% significance level.
3. A national newspaper is investigating a claim made by the CEO of a large airline. The CEO claims that 90% of the company's customers are satisfied with the service they received. The newspaper surveyed 500 customers. Among the sampled customers, 441 said they were satisfied with the company's service. Test the CEO's claim, at the 5% significance level.
4. A magazine claims that more than 40% of leaving cert students smoke. A random sample of 2500 leaving cert students found that 1200 students smoked. Test the magazines claim at the 5% significance level.
5. John claims a die is biased towards the number 6. Out of 100 rolls, the number 6 came up 25 times. Is there enough evidence to support John's claim, at the 5% significance level.





6. A casino manager thinks a player is cheating somehow in a game of dice, perhaps by using biased dice. The game involves rolling two dice and adding the scores. If a player rolls a total of 7 or 11 they will win. Out of 50 games, a certain player won 28 times. Test the claim that the player is cheating, at the 5% significance level.
7. A travel company claims that the average cost of a hotel room in Limerick is €65. A researcher selects a sample of 50 hotel rooms and finds that the average cost is €62 with a standard deviation of €12.53. At $\alpha = 0.05$, is there enough evidence to reject the travel company's claim?
8. The director of a multinational company claims that the average hourly wage of the company's factory workers is €12.50 per hour. A sample of 50 employees had a mean hourly wage of €10.50 with a standard deviation of €2.26. At $\alpha = 0.05$, is there enough evidence to reject the directors claim?
9. A reporter suggests that the average attendance at Thomond Park for Munster Rugby games is 20000. Using the data for attendance from the table below, is there enough evidence to reject the reporters claim?

7862	6982	15824
9865	22145	21987
17569	9513	14562
6598	15354	13974
24865	18954	25630
19541	13254	8751
25630	25630	22159
16789	9512	17498
23541	25630	15971
7875	25630	14654

10. A tyre manufacturer claims that their tyres last for an average of 70000 miles. A random sample of 152 tyres found that they lasted for an average of 63000 miles, with a standard deviation of 15000 miles. Test the manufacturers claim at the 5% significance level.
11. A casino owner claims that the average age of people at the weekly BINGO sessions is 65 years . A sample of 30 BINGO players is selected, and their ages are recorded. Test the owner's claim at the 5% significance level.

59	80	24	61	69	68
53	72	46	65	76	81
80	66	70	61	71	67
52	82	74	39	49	69
32	56	70	74	62	45

12. A politician claims that the average student spends €60 a week on alcohol. A sample of 350 college students found that on average they spent €40 on alcohol per week, with a standard deviation of €26. Test the politician's claim at the 5% significance level.
13. From past records in Limerick maternity hospital, it was found that in the years 2000 to 2010, the average weight of new born babies was 7.1 pounds. A doctor claims that





in recent years, newborns are getting bigger and heavier. In 2015, a sample of 80 newborn babies, had a mean weight of 7.5 pounds with a standard deviation of 1.64 pounds. Is there enough evidence to support the doctors claim, at the 5% significance level?

14. A psychologist suggests that teenagers have too many Facebook friends. He claims that the average teenager has 800 Facebook friends. A sample of 250 teenagers, showed an average of 770 friends, with a standard deviation of 293.3. At the 5% significance level, is there enough evidence to reject the psychologists claim?
15. The average IQ is 100 with a standard deviation of 15. A school principal claims that students in his school have a higher than average IQ. A random sample of 30 students from his school had an average IQ of 110 with a standard deviation of 18.5. Is there enough evidence to support the principal's claim, at the 5% significance level?

10.4 Exam Questions

1. **2016 Paper 2** Data on earnings were published for a particular country. The data showed that the annual income of people in full-time employment was normally distributed with a mean of €39 400 and a standard deviation of €12 920.
 - (a)
 - i. The government intends to impose a new tax on incomes over €60 000. Find the percentage of full-time workers who will be liable for this tax, correct to one decimal place.
 - ii. The government will also provide a subsidy to the lowest 10 % of income earners. Find the level of income at which the government will stop paying the subsidy, correct to the nearest euro.
 - iii. Some time later a research institute surveyed a sample of 1000 full-time workers, randomly selected, and found that the mean annual income of the sample was €38 280. Test the hypothesis, at the 5 % level of significance, that the mean annual income of full-time workers has changed since the national data were published. State the null hypothesis and the alternative hypothesis. Give your conclusion in the context of the question.
 - (b) The research institute surveyed 400 full-time farmers, randomly selected from all the full-time farmers in the country, and found that the mean income for the sample was €26 974 and the standard deviation was €5120. Assuming that annual farm income is normally distributed in this country, create a 95 % confidence interval for the mean income of full-time farmers.
 - (c) It is known that data on farm size are not normally distributed. The research institute could take many large random samples of farm size and create a sampling distribution of the means of all these samples. Give one reason why they might do this.
 - (d) The research institute also carried out a survey into the use of agricultural land. n farmers were surveyed. If the margin of error of the survey was 4.5 %, find the value of n .





2. **2015 Paper 2** A survey of 100 shoppers, randomly selected from a large number of Saturday supermarket shoppers, showed that the mean shopping spend was €90.45. The standard deviation of this sample was €20.73.
- Find a 95% confidence interval for the mean amount spent in a supermarket on that Saturday.
 - A supermarket has claimed that the mean amount spent by shoppers on a Saturday is €94. Based on the survey, test the supermarket's claim using a 5% significance level. Clearly state your null hypothesis, your alternative hypothesis, and your conclusion.
 - Find the p -value of the test you performed in part (b) above and explain what this value represents in the context of the question.
3. **2014 Paper 2** Table 1 below gives details of the number of males (M) and females (F) aged 15 years and over at work, unemployed, or not in the labour force for each year in the period 2004 to 2013.

Table 1										
Labour Force Statistics 2004 to 2013 - Persons aged 15 years and over (000's)										
Year	At work			Unemployed			Not in labour force			Total
	M	F	Total	M	F	Total	M	F	Total	
2004	1045.9	738.9	1784.8	79.6	31.6	111.2	457.1	854.2	1311.3	3207.3
2005	1087.3	779.7	1867.0	81.3	33.5	114.8	459.5	846.6	1306.1	3287.9
2006	1139.8	815.1	1954.9	80.6	38.1	118.7	457.6	844.9	1302.5	3376.1
2007	1184.0	865.6	2049.6	84.3	39.2	123.5	472.4	852.7	1325.1	3498.2
2008	1170.9	889.5	2060.4	106.3	41.0	147.3	494.8	872.5	1367.3	3575.0
2009	1039.8	863.5	1903.3	234.0	82.4	316.4	505.6	874.9	1380.5	3600.2
2010	985.1	843.5	1828.6	257.6	98.2	355.8	529.2	884.6	1413.8	3598.2
2011	970.2	843.2	1813.4	260.7	103.4	364.1	540.1	881.5	1421.6	3599.1
2012	949.6	823.8	1773.4	265.2	108.0	373.2	546.5	896.9	1443.4	3590.0
2013	974.4	829.0	1803.4	227.7	102.3	330.0	557.8	895.0	1452.8	3586.2

- Suggest two categories of people, aged 15 years or over, who might not be in the labour force.
- Find the median and the interquartile range of the total persons at work over the period.
- The following data was obtained from Table 1. The percentages of persons aged 15 years and over at work, unemployed, or not in the labour force for the year 2006 are given below.

		At work	Unemployed	Not in the labour force
Persons aged 15 years and over	2006	57.9%	3.5%	38.6%
	2011			

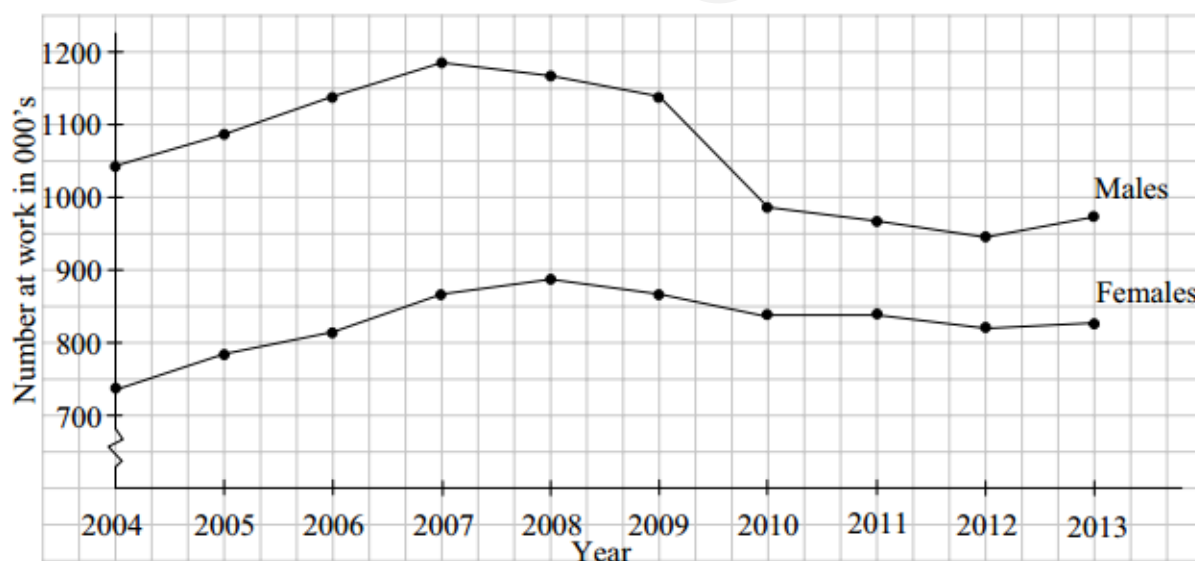




- i. Complete the table for the year 2011. Give your answer correct to one decimal place.
- ii. A census in 2006 showed that there were 864449 persons in the population aged under 15 years of age. The corresponding number in the 2011 census was 979590. Assuming that none of these persons are in the labour force, complete the table below to give the percentages of the *total population* at work, unemployed, or not in the labour force for the year 2011.

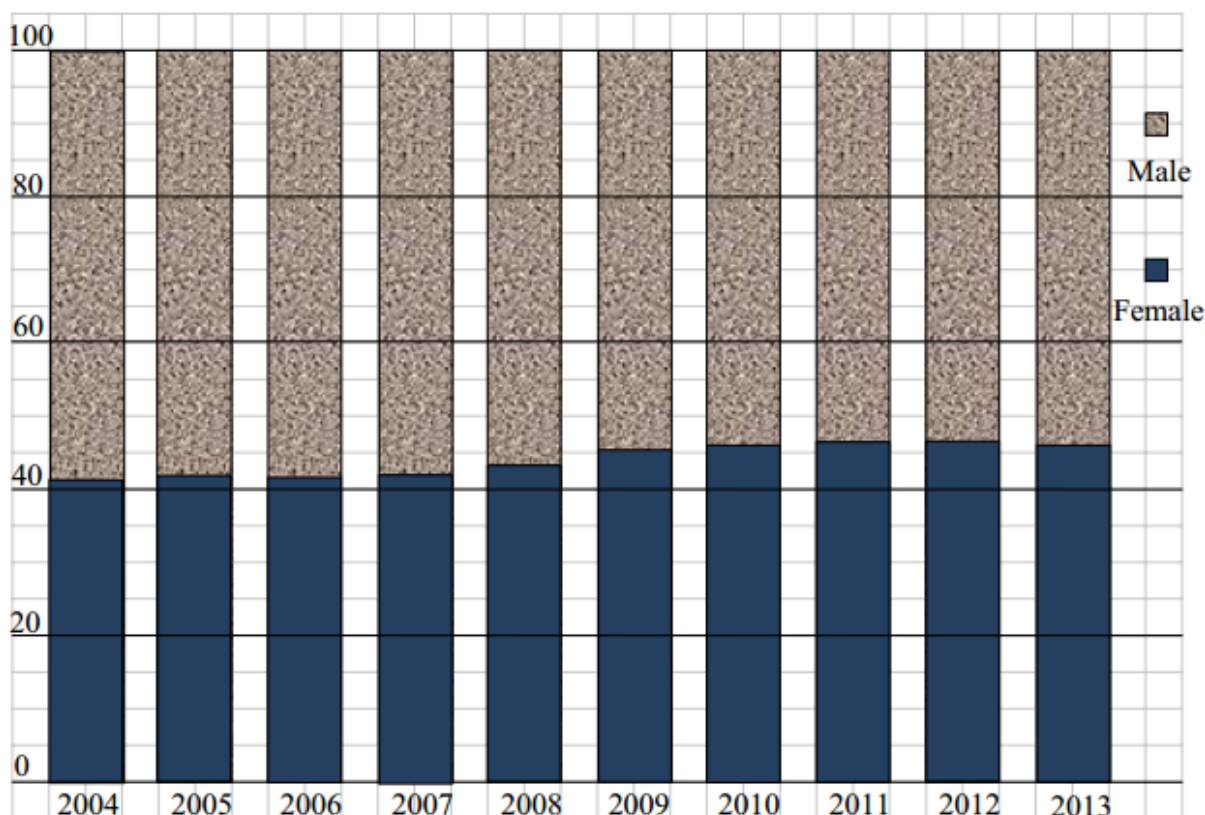
		At work	Unemployed	Not in the labour force
Total population	2006	46.1%	2.8%	51.1%
	2011			

- iii. A commentator states that "The changes reflected in the data from 2006 to 2011 make it more difficult to balance the Government's income and expenditure". Do you agree with this statement? Give two reasons for your answer based on your calculations above.
- (d) Liam and Niamh are analysing the number of males and the number of females at work over the period 2004 to 2013. Liam draws the following chart, using data from Table 1.





Niamh uses the same data and calculates the number of females at work as a percentage of the total number of persons at work and then draws the following chart.



- i. Having examined both charts, a commentator states "females were affected just as much as males by the downturn in employment". Do you agree or disagree with this statement? Give reasons for your conclusion.
 - ii. Which, if any, of the two charts did you find most useful in reaching your conclusion above? Give a reason for your answer.
 - iii. Use the data in Table 1, for the years 2012 and 2013 only, to predict the percentage of persons, aged 15 years and over, who will be at work in 2014.
4. **2014 Paper 2** A generic drug used to treat a particular condition has a success rate of 51%. A company is developing two drugs, A and B, to treat the condition. They carried out clinical trials on two groups of 500 patients suffering from the condition. The results showed that Drug A was successful in the case of 296 patients. The company claims that Drug A is more successful in treating the condition than the generic drug.
 - (a) Use a hypothesis test at the 5% level of significance to decide whether there is sufficient evidence to justify the company's claim. State the null hypothesis and state your conclusion clearly.
 - (b) The null hypothesis was accepted for Drug B. estimate the greatest number of patients in the trial who could have been successfully treated with Drug B.





5. 2013 Paper 2

Go Fast Airlines provides internal flights in Ireland, short haul flights to Europe and long haul flights to America and Asia. On long haul flights the company sells economy class, business class and executive class tickets. All passengers have a baggage allowance of 20 kg and must pay a cost per kg for any weight over the 20 kg allowance. Each month the company carries out a survey among 1000 passengers. Some of the results of the survey for May are shown below.

Gender	Male: 479	Female: 521
--------	-----------	-------------

Previously flown with <i>Go Fast Airlines</i>	Yes: 682	No: 318
---	----------	---------

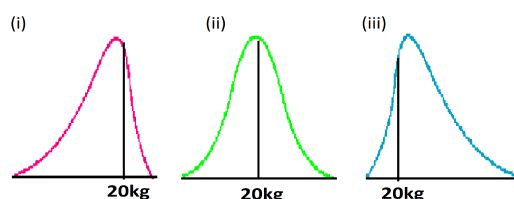
Would fly again with <i>Go Fast Airlines</i>	Yes: 913	No: 87
--	----------	--------

Passenger Age	Mean age: 42	Median age: 31
Spend on in-flight facilities	Mean spend: €18.65	Median spend: €32.18

Was flight delayed?	Yes	No	Don't know
	231	748	21

Passenger satisfaction with overall service	Satisfied	Not satisfied	Don't know
	664	238	98

- (a) *Go Fast Airlines* used a **stratified random sample** to conduct the survey.
- Explain what is meant by a **stratified random sample**.
 - Write down 4 different passenger groups that the company might have included in their sample.
- (b)
- What is the probability that a passenger selected at random from this sample
 - had his/her flight delayed Answer:
 - was not satisfied with the overall service Answer:
 - An employee suggests that the probability of selecting a passenger whose flight was delayed and who was not satisfied with the overall service should be equal to the product of the two probabilities in (i) above. Do you agree with the employee?
- (c) Which of the graphs below do you think is most likely to represent the distribution of the weight of passengers' baggage?



Give a reason for your answer.





- (d)
 - i. Draw a sketch of the possible distribution of the ages of the passengers based on the data in the survey.
 - ii. Explain your answer
- (e)
 - i. The company repeatedly asserts that 70% of their customers are satisfied with their overall service. Use a hypothesis test at the 5% level of significance to decide whether there is sufficient evidence to conclude that their claim is valid in May. Write the null hypothesis and state your conclusion clearly.
 - ii. A manager of the airline says: "If we survey 2000 passengers from June on, we will halve the margin of error in our surveys." Is the manager correct? Explain your answer.

- (f) The response of ten individual passengers to the question on age and in-flight spend are given below.

Age(years)	46	29	37	18	25	75	52	35	40	31
In-flight spend (euro)	30	15	20	0	10	45	25	20	20	30

- i. Draw a scatter plot of the data.
 - ii. Calculate the correlation coefficient between passengers age and in-flight spend.
 - iii. What can you conclude from the completed scatter plot and the correlation coefficient?
 - iv. Sketch the line of best fit in the completed scatter plot above.
6. **2012 Paper 2** In a 2011 study it was found that 48% of mortgages were in negative equity. In early 2012, researchers want to know whether the proportion of properties in negative equity has changed. They analyse 2000 randomly selected properties with mortgages. They discover that 552 of them are in negative equity. Use a hypothesis test at the 5% level of significance to decide whether there is sufficient evidence to conclude that the situation has changed since December 2010.
Be sure to state the null hypothesis clearly, and to state the conclusion clearly.
7. **2014 Sample paper 2** A survey is being conducted of voters' opinions on several different issues.
- (a) What is the overall margin of error of the survey, at 95% confidence, if it is based on a simple random sample of 1111 voters?
 - (b) A political party had claimed that it has the support of 23% of the electorate. Of the voters in the sample above, 234 stated that they support the party. Is this sufficient evidence to reject the party's claim, at the 5% level of significance?
8. **2014 Sample Paper 2** The *King of the Hill* triathlon race in Kinsale consists of a 750 metre swim, followed by a 20 kilometre cycle followed by a 5 kilometre run. The questions below are based on data from 224 athletes who completed this triathlon in 2010.
Maire is analysing data from the race, using statistical software. She has a data file with each competitor's time for each part of the race, along with various other details of the competitors.
Maire gets the software to produce some *summary statistics* and it produces the

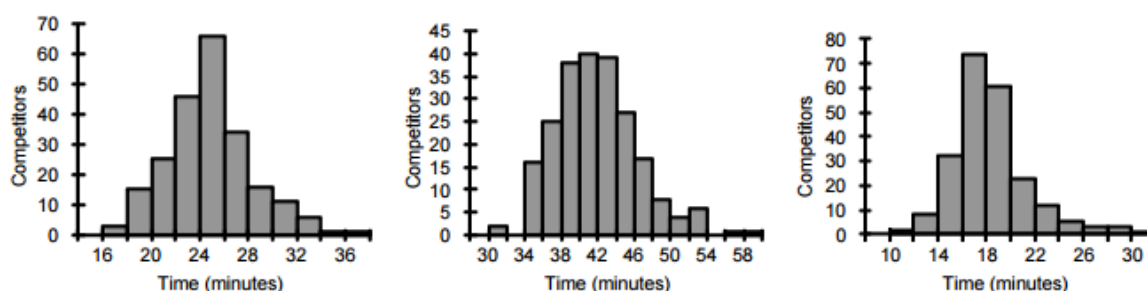




	<i>Swim</i>	<i>Cycle</i>	<i>Run</i>
Mean	18.329	41.927	?
Median	17.900	41.306	?
Mode	#N/A	#N/A	#N/A
Standard Deviation	?	4.553	3.409
Sample Variance	10.017	20.729	11.622
Skewness	1.094	0.717	0.463
Range	19.226	27.282	20.870
Minimum	11.350	31.566	16.466
Maximum	30.576	58.847	37.336
Count	224	224	224

following table. Three of the entries in the table have been removed and replaced with question marks (?).

Maire produces histograms of the times for the three events. Here are the three histograms, without their titles.

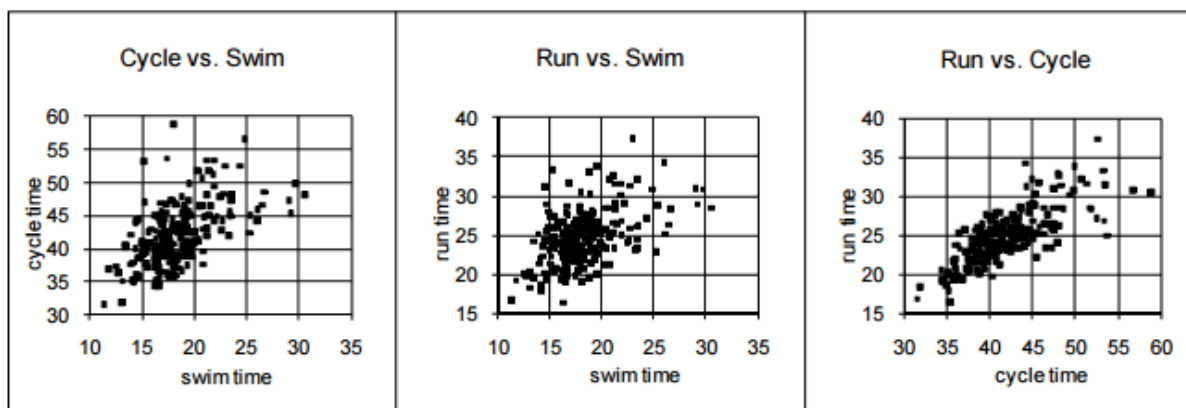


- (a)
- Use the summary statistics in the table to decide which histogram responds to each event. Write the answers above the histograms.
 - The mean and the median time for the run are approximately equal. Estimate this value from the corresponding histogram.
 $mean \approx median \approx \dots\dots\dots$
 - Estimate from the relevant histogram the standard deviation of the times for the swim.
 $Standard\ deviation \approx \dots\dots\dots$
 - When calculating the summary statistics, the software failed to find a *mode* for the data sets. Why do you think this is?





Maire is interested in the relationship between the athletes' performance in the three different events. She produces the following three scatter diagrams



- (b) Give a brief summary of the relationship between performance in the different events, based on the scatter diagrams.
- (c) The best-fit line for run-time based on swim-time is $y = 0.53x + 15.2$. The best-fit line for run-time based on cycle-time is $y = 0.58x + 0.71$. Brian did the swim in 17.6 minutes and cycle in 35.7 minutes. Give your best estimate of Brian's time for the run, and justify your answer.
- The mean finishing time for the overall event was 88.1 minutes and the standard deviation was 10.3 minutes.
- (d) Based on a assumption that the distribution of overall finishing times is approximately normal, us the empirical rule to complete the following sentence : "95% of the athletes took between..... and minutes to complete the race."
- (e) Using normal distribution tables, estimate the number of athletes who completed the race in less than 100 minutes.
- (f) After the event, a reporter wants to interview two people who took more than 100 minutes to complete the race. She approaches athletes at random and asks them their finishing time. She keeps asking until she finds someone who took over 100 minutes, interviews that person, and continues until she finds a second such person. Assuming the athletes are cooperative and truthful, what is the probability that the second person she interviews will be the sixth person she approaches?

