

|  |
| --- |
| ­Seat Number: \_\_\_\_\_\_\_\_\_\_ Room: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Student Number: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Surname: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Given Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Questions that are covered in the week 6 test are highlighted in yellow  The last 3 section in question B are the most important  Lecturer/Tutor:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |

**This paper and all materials issued MUST be returned at the end of the examination.**

**They are NOT to be removed from the examination room.**

**Examination Conditions:**

King’s Own Institute examinations are run in accordance with KOI examination procedures. Any student breaching these procedures will be dealt with in accordance with KOI’s Student Academic Misconduct Policy.

It is your responsibility to ensure that you are in the correct examination.

It is your responsibility to fill out and complete your details in the space provided on all the examination material provided. Use the time before your examination to complete this task as you will not be granted any additional time to do so.

You are **NOT** permitted to have at your desk or on your person any unauthorized material. This includes but is not limited to:

* Mobile Phones
* Smart watches
* Electronic Devices
* Notes

You are not permitted to obtain assistance by improper means or to ask for or give help to any other person.

You may not leave the room, including using the bathroom, for the first 60 minutes or during the last 15 minutes.

During the examination you must seek permission (by raising your hand) from a supervisor if you wish to:

* Leave early
* Use the bathroom
* Access your bag
* Require assistance

**BUS105 INTRODUCTION TO BUSINESS STATISTICS**

**Final Examination T318**

**Time Allowed**: 2 *hours plus 10 minutes reading time*

Reading time is for reading only. You are not permitted to write, calculate or mark your paper in any way during reading time.

**This is a CLOSED BOOK exam**

**Permitted materials for this exam:**

* Non-programmable calculators
* Dictionaries.

**Materials provided for the examination:**

* Formula sheet (at the end of this exam paper)
* Statistical tables (at the end of this exam paper)

**Instructions for Students: (example below)**

* This exam consists of two (2) sections: A and B.
* Section A consists of 15 multiple choice questions each worth 1 mark. Multiple choice answer sheet is provided.
* Sections B consists of eight (8) short answer questions to be answered in the space provided. Questions in Section B is worth 15 marks in total.
* This exam will be marked out of 30 in total, and then will be scaled to a mark out of 50.

**DO NOT open your examination paper until instructed**

**SECTION A – Multiple Choice Questions (MCQ)**

**Question 1**

What is the 97.72th percentile of the Binomial distribution X when n=10,000 and p=0.5?

Hint: X is approximately normally distributed with mean 5,000 and standard deviation 50.

1. 4,900
2. 5,000
3. 5,100
4. None of the Above

Answer C

Answer 97.72th percentile has zscore 2   
x= mean+zscore\*stdev   
=5000+2\*50=5100  
**Question 2**

Suppose you obtain a sample of 100 Australians and only 20 Support the Prime Minister. The formula for the 90% confidence interval for *p* is

Answer d, use row z (bottom row of the table) because you have a sample proportion.

**Question 3**

What is the standard deviation of the sample 10, 8 ,12?

1. 5
2. 2
3. 0
4. None of the above

**Answer b**

**Question 4**

Consider the 2 cases below. In both cases the null and alternative hypothesis being tested  
are   
H0: µ=100   
H1: µ>100 (is the population mean more than 100).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Case | n |  | *s* |  | Standard error | Test stat | p-value |
| 1 | 400 | 100.4 | 4 | 0.4 | 0.2 | 2 | 0.023 |
| 2 | 100 | 100.4 | 4 | 0.4 | 0.4 | 1 | 0.16 |

Which pair of diagrams gives a plausible representation of test stat and p-value for Case 1 and Case 2?

|  |  |
| --- | --- |
| Case 1 | Case 2 |



|  |  |
| --- | --- |
| Case 1 | Case 2 |

|  |  |
| --- | --- |
| Case 1 | Case 2 |

|  |  |
| --- | --- |
| Case 1 | Case 2 |

Answer D shade to the right because the question says more than >

First cas has a smaller pvalue so it has a smaller shaded area

**Question 7**

If the mean of a quantitative variable X is not close to the median

1. The variable X **IS** skewed, so it **CANNOT** be normally distributed
2. The variable X **IS** skewed, so it **CAN** be normally distributed
3. The variable X **IS NOT** skewed, so it **CANNOT** be normally distributed
4. The variable X **IS NOT** skewed, so it **CAN** be normally distributed

**Answer A , if the mean is not close the median then variable is skewed and not normal**

**Question 8**

Consider the following two-way table based on a survey of 200 gym customers.

|  |  |  |
| --- | --- | --- |
|  | Males | Females |
| Want the gym to be Unisex | 10 | 80 |
| Do Not want the gym to be Unisex | 90 | 20 |
| Total | 100 | 100 |

What is the probability that the customer wants the gym to be Unisex if the hypothesis the cutomer is female is true?

1. 0
2. 0.8
3. 2
4. None of the above

Answer B

Answer 80/100=0.8

**Question 9**

Consider the following two data sets:

|  |  |
| --- | --- |
| Dataset 1 | Data set 2 |

A linear regression model would be appropriate for:

1. Data set 1
2. Data set 2
3. Data set 1 and Data set 2
4. None of the above

Answer C both lines are straight

**Question 10**

Suppose you are going to get a sample of 100 customers to test the claim that proportion of males is more than 50%. What is an appropriate test stat and critical value for testing the claim at the 5% level of significance?

1. Test stat =, critical value -*t*99,0.025 = -1.984
2. Test stat =, critical value -*t*99,0.025 = -0.5
3. Test stat =, critical value -*z*0.025 = -1.96
4. None of the above

Answer

D, question says more than so crical value should be positive

Correct critical value is Z0.05 =1.645

**Question 11**

In a sample of 100 customers, 80 would buy the product.

1. The estimate of *p* is 0.8 with standard error 0.04
2. The estimate of *p* is 0.8 with standard error 0.2
3. The estimate of *µ* is 0.8 with standard error 0.04
4. The estimate of *µ* is 0.8 with standard error 0.2

**Answer A**

**Estimate of p** is =80/100 = 0.8, n=100

**Standard error = =0.04**

**Question 12**

Compare the two cases below. If you are investigating the claim that there is a significant difference between means

Case 1

|  |  |  |  |
| --- | --- | --- | --- |
|  | Average amount spent | Standard deviation of amount spent | Sample size |
| Males | $100 | $30 | 100 |
| Females | $120 | $40 | 100 |

Case 2

|  |  |  |  |
| --- | --- | --- | --- |
|  | Average amount spent | Standard deviation of amount spent | Sample size |
| Males | $100 | $30 | 10 |
| Females | $120 | $40 | 10 |

1. Case 1 has **higher** standard error and **higher** p-value
2. Case 1 has **lower** standard error and **higher** p-value
3. Case 1 has **higher** standard error and **lower** p-value
4. Case 1 has **lower** standard error and **lower** p-value

Answer D, Bigger samples are more accurate.

**Use the following information to answer questions 13,14 and 15**

Consider the following scatterplot and computer output based on a sample size of 62 houses that lets you investigate the relationship between the variable

X house size in square meters, and

Y house price in thousands $

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | 353.9 |  |  |  |
| Size | 3 | 0.05 | 60 | 3.26E-79 |

**Question 13**

What is a 95% confidence interval for the slope? You may use the fact that in this case 95% of sample slopes are within 2 standard errors of the population slope.

1. Between 2.9 and 3.1
2. Between 0.05 and 0.25
3. Between 2.95 and 3.05
4. None of the above

Answer A

**Answer between 3-2\*0.05=2.9 and 3+2\*0.05=3.1**

**Question 14**

The regression line is *y*=353.9+3*x*. Predict price (*y*) when size (*x*) is 100 square meters.

1. 35690
2. 653.9
3. 100
4. None of the above

**Answer B**

**Y=353.9+3\*100=653.9**

**Question 15**

The p-value for testing the claim the slope is different to 0 is 3.2E-79. What is the appropriate interpretation of the p-value?

1. The p-value is **more** than 0.05 so **reject H0** because there is strong evidence the slope is different to 0
2. The p-value is **more** than 0.05 so **do not reject H0** because there is strong evidence the slope is different to 0
3. The p-value is **less** than 0.05 so **reject H0** because there is strong evidence the slope is different to 0
4. The p-value is **less** than 0.05 so **do not reject H0** because there is strong evidence the slope is different to 0

**Answer C**

**SECTION B**

***Short answer section. Write your answers in the space provided.***

**Question 1**

Suppose an estimate X is normally distributed with mean 0.09 and standard deviation 0.01. What is the 97.5th percentile of the estimate? (2 marks) Add shading and labels on the Histogram of many estimates and the Ogive of many estimates given below to explain what is meant by the 97.5th percentile

|  |  |
| --- | --- |
| Histogram of many estimates | Ogive of many estimates |

Answer : the zscore of the 97.5th percentile is 1.96 so the  
 97.5th percentile = mean+zscore\*stdev=0.09+1.96\*0.01=0.11

|  |  |
| --- | --- |
| 97.5%  Z=2  Estimate =0.11 | 2.5%  97.5%  Z=2  Estimate =0.11 |

**Question 2**

1. What would be an appropriate null and alternative Hypothesis for testing the claim there is a linear relationship between the variables “age of customer measured in years” and “the amount spent by the customer”? (1 mark)

answer

H0 :population slope = 0 , H1 :population slope ≠0

1. What would be an appropriate null and alternative Hypothesis for testing the claim “The average amount people spend is below $100.”? (1 mark)

**answer**

H0 :µ= 100 , H1 : µ<100

**Question 3**

1. What is 4% of $700? (0.5 marks)

Answer 0.04\*700=28

1. If 700 students did an exam, then exam mark X is a quantitative variable. If 50 is the 4th percentile of exam mark X (so P(X<50) = 0.04), how many students have a mark less than 50? (0.5 marks)

Answer 0.04\*700=28

**Question 4**

Consider the two cases below.

Case 1

A two-way table which is a summary of a sample from the assignment data set.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Manager wants to keep selling the product** | **Manager does not want to keep selling the product** | **Total** |
| region 1 | 80 | 20 | 100 |
| region 2 | 60 | 40 | 100 |

Two tailed p-value 0.002

To find the p-value H0: *p1=p2* was assumed to be true, evidence against H0. H1:: *p1 ≠ p2* because the test is two tailed.

Case 2

A two-way table which is a summary of a sample from the assignment data set.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Manager wants to keep selling the product** | **Manager does not want to keep selling the product** | **Total** |
| region 1 | 8 | 2 | 10 |
| region 2 | 6 | 4 | 10 |

Two tailed p-value 0.391

To find the p-value H0: *p1=p2* was assumed to be true, evidence against H0. H1:: *p1 ≠ p2* because the test is two tailed.

What does the p-value measure in the cases above and explain why the different cases have different p-values? (1 mark)

**Answer the p-value measures the amount of evidence for the claim that there is a difference between proportions, the lower the p-value the greater the evidence. The case with the larger sample size has the lower p-value   
  
Case 1 has a larger sample size than case 2 so it’s pvalue is smaller**

*A simple comment above is enough but it is worth understanding*

*In case 1 the*  difference in proprotions is - =80/100-60/100=0.2   
in case 2 *the*  difference in proprotions is - =8/10-6/10=0.2   
in both cases the difference in proporitons is the same and   
 measures the effect the variabale “gender?” has on the variable   
“would have on the product ” , however in case 2 the sample size is larger

**Question 5**

a) Consider just the first 4 elements (cases) of a dataset.

|  |  |  |
| --- | --- | --- |
|  | which version? | How much would you pay? |
| person 1 | New version | $28 |
| person 2 | Old version | $24 |
| person 3 | Old version | $26 |
| person 4 | New version | $32 |

What is the difference between the sample means? (0.5 marks)

New version average = (28+32)/2=30  
Old version average = (24+26)/2=25  
so difference is 30-25=5, students should note that it also acceptable to give the answer   
  
25-30=-5

b) Consider the summary of 200 elements of the same dataset used in part (a).

|  |  |  |  |
| --- | --- | --- | --- |
| Which version | Sample average of amount they would pay | Sample standard deviation | Sample size |
| New | $31 | $5 | 100 |
| Old | $30 | $5 | 100 |

Two tailed p-value 0.62

To find the p-value H0: *μ1=μ2* was assumed to be true, evidence against H0. *μ1 ≠ μ2* because the test is two tailed.

1. What are the variables in the dataset? For each variable, state if the variable is categorical or quantitative. (1 mark)
2. Make a simple comment about the relationship between the variables. (0.5 marks)
3. Interpret the p-value and explain what the p-value is measuring. (1 mark)

*You will not have to comment on p-value in the week 6 test*

Answer   
i) which version (old or new) is a categorical variable   
the amount they would pay is a quantitative variable

ii) people would pay more for the new version.

*A simple comment is enough but it is worth understanding* A difference in means -=31-30=1 measures the effect the variabale “which version?” has on the variable “amount they would pay ”

(iii) The p-value is measuring the amount of evidence for the difference between means   
since the p-value is not less than 0.05 there is not strong evidence that there is a difference between means.

**Question 6**

1. Consider the first 5 elements (cases) of a dataset

|  |  |  |
| --- | --- | --- |
|  | gender | would they buy the product? |
| person 1 | Male | No |
| person 2 | Male | No |
| person 3 | Female | Yes |
| person 4 | Male | No |
| person 5 | Female | No |

What is the difference between the proportion of females that would buy the product and the proportion of males that would to buy the product? (0.5 marks)

Answer 1/2 -0/3=1/2

Just the calculation is required however understand   
 A difference in proprotions is - =1/2-0/3=0.5 measures the effect the variabale “gender?” has on the variable “would have on the product ”

1. Consider the following two way table. Which is a summary of 200 elements of the same dataset used in part (a)?

|  |  |  |  |
| --- | --- | --- | --- |
|  | Would buy | Would not buy | Total |
| Female | 80 | 20 | 100 |
| Male | 35 | 65 | 100 |

Two tailed p-value 0.000

To find the p-value H0: *p1=p2* was assumed to be true, evidence against H0. H1:: *p1 ≠ p2* because the test is two tailed.

1. What are the variables in the dataset? For each variable, state if the variable is categorical or quantitative. (1mark)
2. Make a simple comment about the relationship between the variables. (0.5 marks)
3. Using simple terms, explain what the p-value is measuring and give a conclusion based on the p-value (1 mark)

Students do not have to comment on the pvalue in the week 6 test

Answer   
Ii) gender (male or female) is a categorical variable   
would they buy the product (would buy or would not buy) is a categoricalvariable

(ii) A higher proportion of females would buy the product

*A simple comment is enough but it is worth understanding* A difference in proprotions is - =80/100-35/100=0.45 measures the effect the variabale “gender?” has on the variable “would have on the product ”

(iii) The p-value is measuring the amount of evidence for the difference between proportions  
since the p-value is less than 0.05 there is strong evidence that there is a difference between proportions

**Question 7**

1. Suppose you ask a sample 20 people if they support a new law and their answers were

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| no | no | no | no | YES | no | no | no | no | no |
| no | YES | no | no | no | no | no | no | no | no |

1. What is the sample proportion of people that support the new law? (0.5 marks)

Answer 2/20=0.1 (2 out of 20 people said yes)

1. Is it reasonable to calculate a confidence interval for the population proportion above, given that calculating the confidence interval requires you to assume normality? (0.5 marks)

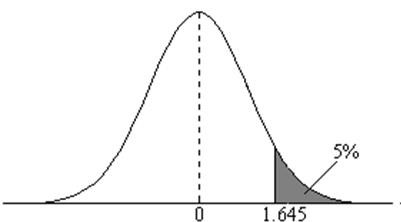
This question will not be in the week 6 test

No , *np* is not above 5

1. Suppose the proportion of people that support a new law is 50%.
2. Calculate the z score of the sample proportion if you have a sample where 60 out of 100 people support a new law. (1 mark)

Answer , n=100, p=0.5,

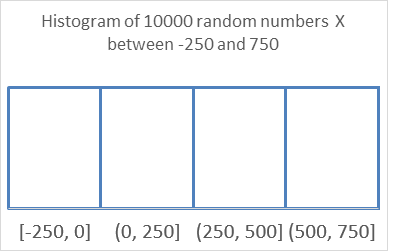
==2

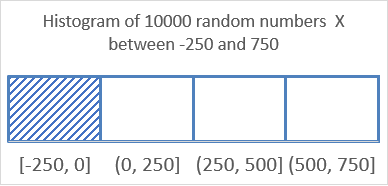
1. Test the claim the proportion of people that support a new law is more than 50% if you have a sample where 60 out of 100 people support a new law. Hint: the test stat is the answer to part (i). (1 mark)   
   answer   
   H0 p=0.5 H1 p>0.5   
   

Test stat of 2 is above the critical value 1.645 from the table so reject H0   
there is strong evidence the proportion is above 60%

**Question 8**

Add shading arrows and comments to the graphs below that explain that 0 is the 25th percentile, The graphs are graphical display of the distribution of 10,000 uniform random numbers between -250 and 750

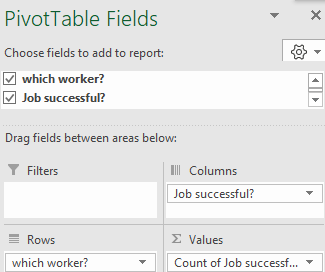
****

Answer.

Looking at the ogive you can see that 25% of values are below 0 and 75% are above

1 out of 4 of the equally sized rectangles are less than 0, so proportion is 1 /4=0.25=25% are less than 0

Question 9

1. Consider the following sample of dataset with just 5 elements (cases) and pivot table commands. What is the output? (1 mark)

|  |  |
| --- | --- |
| which worker? | Job successful? |
| worker A | Successful |
| worker A | Not Successful |
| worker B | Successful |
| worker B | Successful |
| worker B | Not Successful |

Answer the following is enough, it is OK to swap rows and columns

|  |  |  |
| --- | --- | --- |
|  | Not Successful | Successful |
| Worker A | 1 | 1 |
| Worker B | 1 | 2 |

*Note for students: note that the summary in part (b) helps you do part (a)*

1. The following is a summary of a much larger sample of the same dataset used in part (a)

|  |  |  |  |
| --- | --- | --- | --- |
|  | Not successful | Successful | Total |
| Worker A | 5 | 15 | 20 |
| Worker B | 10 | 10 | 20 |

p-value for test the claim there is a difference in proportions 0.10247

1. Comment on the relationship between the fields (variables) (1 mark)

Answer: worker A has a higher proportion of success ,

*A simple comment is enough but it is worth understanding* A difference in proprotions is - =15/20-10/20=0.25 This measures the effect the variabale “which worker?” has on the variable “Is the job successful”

1. Comment on the pvalue (1 mark)

Answer p-value is less 0.05 so there is not strong evidence that there is a difference between the proportions

*You will not discuss p-value in the week 6 test*

Question 10

1. Consider the following Scatter plot

mean of X: , Standard deviation of X: =$10

Mean of Y: , Sandard deviation of Y: =10

Correlation *r* =-0.8

Some predictions based on some possible future values of X are given below.   
Fill in the blanks and plot the future values of X and the predicted values on the scatterplot

|  |  |  |
| --- | --- | --- |
| Future value of X | Z score of X | Predicted y = (Z score of X)\* |
| $20 | (20-40)/10=-2 | 316 |
| $35 |  |  |
| $45 |  |  |
| $60 | 2 | 284 |

Answer

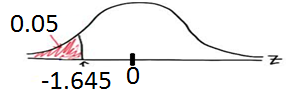
|  |  |  |
| --- | --- | --- |
| Future value of X | Z score of X | Predicted y = (Z score of X)\* |
| $20 | (20-40)/10=-2 | 300+-0.8\*-2\*10=316 |
| $35 | (35-40)/10=-0.5 | 300+-0.8\*-0.5=304 |
| $45 | (45-40)/10=0.5 | 300+-0.8\*0.5=296 |
| $60 | (60-40)/10=2 | 300+-0.8\*2=284 |

Some information about normally distributed variables

Lower tail case (one sided)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| z-score | -1 | -0.674 | | -1.282 | | -1.645 | | -1.960 | | -2 | | -3 | |
| Which Percentile | 15.87th | 25th | 10th | | 5th | | 2.5th | | 2.28th | | 0.135th | |
| P(Z<z-score) | 0.1587 | 0.25 | 0.1 | | 0.05 | | 0.025 | | 0.0228 | | 0.00135 | |

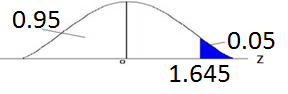
Example: the 5th percentile has a z-score of -1.645 so P(Z<-1.645)=0.05



Upper tail case (one sided)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| z-score | 1 | 0.674 | 1.282 | 1.645 | 1.96 | 2 | 3 | |
| Which Percentile | 84.13 | 75th | 90th | 95th | 97.5th | 97.72th | 99.865 | |
| P(Z>z-score) | 0.1587 | 0.25 | 0.1 | 0.05 | 0.025 | 0.0228 | 0.00135 | |
|  |  |  |  |  |  |  |  |  | |

Example: The 95th percentile has a z-score of 1.645 so P(Z>1.645) =0.05

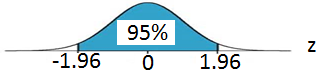
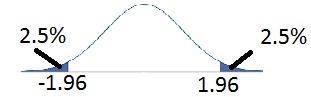


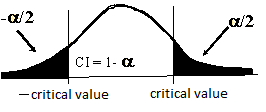
Two sided case

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| z-scores | ±1 | ±0.674 | ±1.282 | ±1.645 | ±1.960 | ±2 | ±3 |
| P(|Z|>|z-score|) | 0.3174 | 0.5 | 0.2 | 0.1 | 0.05 | 0.0456 | 0.0027 |
| proportion between  Both z-scores | 68.26% | 50% | 80% | 90% | 95% | 95.44% | 99.73% |

Example: The middle 95% of values of normally distributed variable have  
 z-scores between -1.96 and 1.96.

Example: middle 95% of values of normally distributed variable have z-scores between -1.96 and 1.96 and 5%,so 5% of values are not between   
(2.5% of values are below and 2.5% are above)  
so P(-1.96<Z<1.96)=0.95

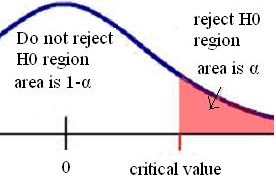


Two-sided t table 

the columns of the table below are the significance level α, if α is not given assumeα=0.05.  
The numbers in the table are the critical values which give the you the boundaries of the shaded region.

If the test stat is in the shaded region, then p-value< α so you should reject H0.

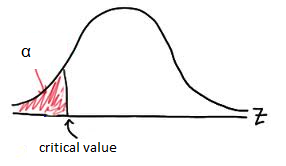
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Α | 0.5 | 0.3 | 0.2 | 0.1 | 0.05 | 0.02 | 0.01 |
| Df |  |  |  |  |  |  |  |
| 1 | ±1.000 | ±1.963 | ±3.078 | ±6.314 | ±12.706 | ±31.821 | ±63.657 |
| 2 | ±0.817 | ±1.386 | ±1.886 | ±2.920 | ±4.303 | ±6.965 | ±9.925 |
| 3 | ±0.765 | ±1.250 | ±1.638 | ±2.353 | ±3.182 | ±4.541 | ±5.841 |
| 4 | ±0.741 | ±1.190 | ±1.533 | ±2.132 | ±2.776 | ±3.747 | ±4.604 |
| 5 | ±0.727 | ±1.156 | ±1.476 | ±2.015 | ±2.571 | ±3.365 | ±4.032 |
| 6 | ±0.718 | ±1.134 | ±1.440 | ±1.943 | ±2.447 | ±3.143 | ±3.707 |
| 7 | ±0.711 | ±1.119 | ±1.415 | ±1.895 | ±2.365 | ±2.998 | ±3.499 |
| 8 | ±0.706 | ±1.108 | ±1.397 | ±1.860 | ±2.306 | ±2.896 | ±3.355 |
| 9 | ±0.703 | ±1.100 | ±1.383 | ±1.833 | ±2.262 | ±2.821 | ±3.250 |
| 10 | ±0.700 | ±1.093 | ±1.372 | ±1.812 | ±2.228 | ±2.764 | ±3.169 |
| 11 | ±0.697 | ±1.088 | ±1.363 | ±1.796 | ±2.201 | ±2.718 | ±3.106 |
| 12 | ±0.696 | ±1.083 | ±1.356 | ±1.782 | ±2.179 | ±2.681 | ±3.055 |
| 13 | ±0.694 | ±1.079 | ±1.350 | ±1.771 | ±2.160 | ±2.650 | ±3.012 |
| 14 | ±0.692 | ±1.076 | ±1.345 | ±1.761 | ±2.145 | ±2.624 | ±2.977 |
| 15 | ±0.691 | ±1.074 | ±1.341 | ±1.753 | ±2.131 | ±2.602 | ±2.947 |
| 16 | ±0.690 | ±1.071 | ±1.337 | ±1.746 | ±2.120 | ±2.583 | ±2.921 |
| 17 | ±0.689 | ±1.069 | ±1.333 | ±1.740 | ±2.110 | ±2.567 | ±2.898 |
| 18 | ±0.688 | ±1.067 | ±1.330 | ±1.734 | ±2.101 | ±2.552 | ±2.878 |
| 19 | ±0.688 | ±1.066 | ±1.328 | ±1.729 | ±2.093 | ±2.539 | ±2.861 |
| 20 | ±0.687 | ±1.064 | ±1.325 | ±1.725 | ±2.086 | ±2.528 | ±2.845 |
| 21 | ±0.686 | ±1.063 | ±1.323 | ±1.721 | ±2.080 | ±2.518 | ±2.831 |
| 22 | ±0.686 | ±1.061 | ±1.321 | ±1.717 | ±2.074 | ±2.508 | ±2.819 |
| 23 | ±0.685 | ±1.060 | ±1.319 | ±1.714 | ±2.069 | ±2.500 | ±2.807 |
| 24 | ±0.685 | ±1.059 | ±1.318 | ±1.711 | ±2.064 | ±2.492 | ±2.797 |
| 25 | ±0.684 | ±1.058 | ±1.316 | ±1.708 | ±2.060 | ±2.485 | ±2.787 |
| 26 | ±0.684 | ±1.058 | ±1.315 | ±1.706 | ±2.056 | ±2.479 | ±2.779 |
| 27 | ±0.684 | ±1.057 | ±1.314 | ±1.703 | ±2.052 | ±2.473 | ±2.771 |
| 28 | ±0.683 | ±1.056 | ±1.313 | ±1.701 | ±2.048 | ±2.467 | ±2.763 |
| 29 | ±0.683 | ±1.055 | ±1.311 | ±1.699 | ±2.045 | ±2.462 | ±2.756 |
| 99 | ±0.677 | ±1.042 | ±1.290 | ±1.660 | ±1.984 | ±2.365 | ±2.626 |
| 100 | ±0.677 | ±1.042 | ±1.290 | ±1.660 | ±1.984 | ±2.364 | ±2.626 |
| 1000 | ±0.675 | ±1.037 | ±1.282 | ±1.646 | ±1.962 | ±2.330 | ±2.581 |
| ∞(z\*) | ±0.674 | ±1.036 | ±1.282 | ±1.645 | ±1.960 | ±2.326 | ±2.576 |
|  |  |  |  |  |  |  |  |
|  | 50%CI | 70%CI | 80%CI | 90%CI | 95%CI | 98%CI | 99%CI |

Right tailed t table 

The columns of the table below are the significance level α, if α is not given assumeα=0.05.  
The numbers in the table are the critical value which give the you the boundary of the shaded region.

If the test stat is in the shaded region, then p-value< α so you should reject H0.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Α | 0.25 | 0.2 | 0.15 | 0.1 | 0.05 | 0.025 | 0.01 | 0.005 |
| **Df** |  |  |  |  |  |  |  |  |
| **1** | 1 | 1.376 | 1.963 | 3.078 | 6.314 | 12.706 | 31.821 | 63.657 |
| **2** | 0.817 | 1.061 | 1.386 | 1.886 | 2.92 | 4.303 | 6.965 | 9.925 |
| **3** | 0.765 | 0.979 | 1.25 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 |
| **4** | 0.741 | 0.941 | 1.19 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 |
| **5** | 0.727 | 0.92 | 1.156 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 |
| **6** | 0.718 | 0.906 | 1.134 | 1.44 | 1.943 | 2.447 | 3.143 | 3.707 |
| **7** | 0.711 | 0.896 | 1.119 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 |
| **8** | 0.706 | 0.889 | 1.108 | 1.397 | 1.86 | 2.306 | 2.896 | 3.355 |
| **9** | 0.703 | 0.883 | 1.1 | 1.383 | 1.833 | 2.262 | 2.821 | 3.25 |
| **10** | 0.7 | 0.879 | 1.093 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 |
| **11** | 0.697 | 0.876 | 1.088 | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 |
| **12** | 0.696 | 0.873 | 1.083 | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 |
| **13** | 0.694 | 0.87 | 1.079 | 1.35 | 1.771 | 2.16 | 2.65 | 3.012 |
| **14** | 0.692 | 0.868 | 1.076 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 |
| **15** | 0.691 | 0.866 | 1.074 | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 |
| **16** | 0.69 | 0.865 | 1.071 | 1.337 | 1.746 | 2.12 | 2.583 | 2.921 |
| **17** | 0.689 | 0.863 | 1.069 | 1.333 | 1.74 | 2.11 | 2.567 | 2.898 |
| **18** | 0.688 | 0.862 | 1.067 | 1.33 | 1.734 | 2.101 | 2.552 | 2.878 |
| **19** | 0.688 | 0.861 | 1.066 | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 |
| **20** | 0.687 | 0.86 | 1.064 | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 |
| **21** | 0.686 | 0.859 | 1.063 | 1.323 | 1.721 | 2.08 | 2.518 | 2.831 |
| **22** | 0.686 | 0.858 | 1.061 | 1.321 | 1.717 | 2.074 | 2.508 | 2.819 |
| **23** | 0.685 | 0.858 | 1.06 | 1.319 | 1.714 | 2.069 | 2.5 | 2.807 |
| **24** | 0.685 | 0.857 | 1.059 | 1.318 | 1.711 | 2.064 | 2.492 | 2.797 |
| **25** | 0.684 | 0.856 | 1.058 | 1.316 | 1.708 | 2.06 | 2.485 | 2.787 |
| **99** | 0.677 | 0.845 | 1.042 | 1.29 | 1.66 | 1.984 | 2.365 | 2.626 |
| **100** | 0.677 | 0.845 | 1.042 | 1.29 | 1.66 | 1.984 | 2.364 | 2.626 |
| **1000** | 0.675 | 0.842 | 1.037 | 1.282 | 1.646 | 1.962 | 2.33 | 2.581 |
| **∞(z\*)** | 0.674 | 0.841 | 1.036 | 1.282 | 1.645 | 1.96 | 2.326 | 2.576 |
|  |  |  |  |  |  |  |  |  |
|  | 50%CI | 60%CI | 70%CI | 80%CI | 90%CI | 95%CI | 98%CI | 99%CI |

left tailed t table 

The columns of the table below are the significance level α, if α is not given assumeα=0.05.  
The numbers in the table are the critical value which give the you the boundary of the shaded region.

If the test stat is in the shaded region, then p-value< α so you should reject H0.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Α | 0.25 | 0.2 | 0.15 | 0.1 | 0.05 | 0.025 | 0.01 | 0.005 |
| **df** |  |  |  |  |  |  |  |  |
| **1** | -1 | -1.376 | -1.963 | -3.078 | -6.314 | -12.706 | -31.821 | -63.657 |
| **2** | -0.817 | -1.061 | -1.386 | -1.886 | -2.92 | -4.303 | -6.965 | -9.925 |
| **3** | -0.765 | -0.979 | -1.25 | -1.638 | -2.353 | -3.182 | -4.541 | -5.841 |
| **4** | -0.741 | -0.941 | -1.19 | -1.533 | -2.132 | -2.776 | -3.747 | -4.604 |
| **5** | -0.727 | -0.92 | -1.156 | -1.476 | -2.015 | -2.571 | -3.365 | -4.032 |
| **6** | -0.718 | -0.906 | -1.134 | -1.44 | -1.943 | -2.447 | -3.143 | -3.707 |
| **7** | -0.711 | -0.896 | -1.119 | -1.415 | -1.895 | -2.365 | -2.998 | -3.499 |
| **8** | -0.706 | -0.889 | -1.108 | -1.397 | -1.86 | -2.306 | -2.896 | -3.355 |
| **9** | -0.703 | -0.883 | -1.1 | -1.383 | -1.833 | -2.262 | -2.821 | -3.25 |
| **10** | -0.7 | -0.879 | -1.093 | -1.372 | -1.812 | -2.228 | -2.764 | -3.169 |
| **11** | -0.697 | -0.876 | -1.088 | -1.363 | -1.796 | -2.201 | -2.718 | -3.106 |
| **12** | -0.696 | -0.873 | -1.083 | -1.356 | -1.782 | -2.179 | -2.681 | -3.055 |
| **13** | -0.694 | -0.87 | -1.079 | -1.35 | -1.771 | -2.16 | -2.65 | -3.012 |
| **14** | -0.692 | -0.868 | -1.076 | -1.345 | -1.761 | -2.145 | -2.624 | -2.977 |
| **15** | -0.691 | -0.866 | -1.074 | -1.341 | -1.753 | -2.131 | -2.602 | -2.947 |
| **16** | -0.69 | -0.865 | -1.071 | -1.337 | -1.746 | -2.12 | -2.583 | -2.921 |
| **17** | -0.689 | -0.863 | -1.069 | -1.333 | -1.74 | -2.11 | -2.567 | -2.898 |
| **18** | -0.688 | -0.862 | -1.067 | -1.33 | -1.734 | -2.101 | -2.552 | -2.878 |
| **19** | -0.688 | -0.861 | -1.066 | -1.328 | -1.729 | -2.093 | -2.539 | -2.861 |
| **20** | -0.687 | -0.86 | -1.064 | -1.325 | -1.725 | -2.086 | -2.528 | -2.845 |
| **21** | -0.686 | -0.859 | -1.063 | -1.323 | -1.721 | -2.08 | -2.518 | -2.831 |
| **22** | -0.686 | -0.858 | -1.061 | -1.321 | -1.717 | -2.074 | -2.508 | -2.819 |
| **23** | -0.685 | -0.858 | -1.06 | -1.319 | -1.714 | -2.069 | -2.5 | -2.807 |
| **24** | -0.685 | -0.857 | -1.059 | -1.318 | -1.711 | -2.064 | -2.492 | -2.797 |
| **25** | -0.684 | -0.856 | -1.058 | -1.316 | -1.708 | -2.06 | -2.485 | -2.787 |
| **99** | -0.677 | -0.845 | -1.042 | -1.29 | -1.66 | -1.984 | -2.365 | -2.626 |
| **100** | -0.677 | -0.845 | -1.042 | -1.29 | -1.66 | -1.984 | -2.364 | -2.626 |
| **1000** | -0.675 | -0.842 | -1.037 | -1.282 | -1.646 | -1.962 | -2.33 | -2.581 |
| **∞(z\*)** | -0.674 | -0.841 | -1.036 | -1.282 | -1.645 | -1.96 | -2.326 | -2.576 |
|  |  |  |  |  |  |  |  |  |
|  | 50%CI | 60%CI | 70%CI | 80%CI | 90%CI | 95%CI | 98%CI | 99%CI |

**BUS105 formula sheet**

# An example that explains the concepts “quantitative variable”, “percentile” and “tail probability” If a group of students did an exam the teacher would mark the papers and obtain a list of Marks, different students get different marks so the list of marks is a dataset with one quantitative variable. The symbol for a quantitative variable is X. If 20% of the students in the class got a mark X less than 50 then the mark x=50 is the 20th percentile. The lower tail probability P(X<50)=0.2

**Formulas for calculating sample statistics**

Sample mean**: **

Sample variance**:**

Sample standard deviation**: **

**Formulas for calculating population parameters given a discrete distribution**

Population mean 

Population variance ,   
Population standard deviation σ =

**Properties of the binomial distribution**

Mean = *np*  
Standard deviation=

The mean and standard deviation will be calculated for you, so you can find the z-score

the binomial is approximately normally distributed   
If np≥5 and n(1-p) ≥5

**z-score of a quantitative variable X**

Z-score = (X-mean)/stdev (Note that stdev is the standard deviation)  
 So x= mean+z-score\*stdev   
Note that if a question asks for a percentile you have to calculate x

**Some tips for interpreting summaries of the data sets that compare two groups**

\*The difference between sample means -(pronounced xbar1-xbar2) estimates the difference between the population means μ1-μ2

\*The difference between sample proportions - (pronounced phat1-phat2) estimates the difference between the population proportions p1-p2

If you get many samples from the same population the sample estimate is a variable   
the properties are given below.

**An that estimate summarizes a single quantitative variable**

the sample mean is the estimate for the population mean μ

standard error of =

Z-score for the sample mean

**An estimate that summarizes a single categorical variable**the sample proportion is the estimate for the population proportion *p*

standard error of =

Z-score for a sample proportion

Formula for the test statistic when you have an estimate and the standard error

In a hypothesis test you only get one sample but you imagine you get many samples

Test statistic =

Where the claimed value is given in the question, you assume the claimed value is the mean of the sample estimate.

NOTE: if you use sample standard deviation s to calculate the standard error the test stat has   
a t-distribution with df= n-1 this takes into account that the sample standard deviation is not accurate and can be too small.

**Confidence intervals**

For a confidence interval you get a single sample, but you rely on the fact there is a high chance the estimate is not many standard errors away from true value. (if you get many samples most of the samples would be close to the estimate)

For single categorical variable

Confidence interval for population proportion p is

For a single quantitative variable   
confidence interval for the population mean μ is

**Alternate formula page 1 of 2**

**USEFUL FORMULAE WHEN YOU HAVE A SINGLE SAMPLE MEAN**

**Standard error of sample mean:** 

**Estimated standard error of the mean**: ,   
note that if n is small this could be much too small, use the t-table to take this into account

## Confidence interval formula for a single mean

****  if**** is given **, **  *, =df =n-1*if **** unknown,

**Hypothesis testing formula for a single mean**

**The test stat is z= **  if**** is given, The test stat is t= ****  if**** is unknown

### If you use the sample standard deviation s then use df=n-1

### 

**USEFUL FORMULAE WHEN YOU HAVE A SINGLE SAMPLE PROPORTION**

Standard error of sample proportion  (if you do not have p use the estimate )  
Many textbooks use the symbol *q=1-p* instead of writing 1-p

The distribution of the sample proportion is close to the normal distribution, if *np*>5 and *n*(1-*p*)>5

## Confidence interval formula for a single proportion

 Confidence Interval for p

## Hypothesis testing for a single proportion

Test statistic

Use z-table to find the critical value

**Alternative test statistic,**

**Alternate formula page 2 of 2**

SIMPLE LINEAR REGRESSION This is regression with only one predictor

**The sample least square regression line : **

The standard errors (se) se() and se()

will always be given in the output next to the coefficient

## Confidence interval formula for the slope



**The test stat for testing a hypothesis about the slope is**

 the critical value comes from the t-table with df n-2

Some tips on explaining what effect one variable has on another variable

To measure the effect that a categorical variable has on another quantitative variable  
you can calculate difference between sample means - (xbar1-xbar2).

You can explain the effect by   
Describing is the mean of category 1

Describing is the mean of category 2  
Stating the mean of category 1 is - higher than the mean category 2

Example (the exam formula sheet probably won’t have an example)   
The mean weight of some males is 60 (=60)

The mean weight of some females is 40 (=40)  
So the mean weight of males is - =60-40=20 higher than the mean of females   
- =20 measures the effect that gender has on weight

To measure the effect that a categorical variable has on another categorical variable  
 calculate the difference between sample means sample proportions - (phat1-phat2)

You can explain the effect by   
Describing is the proportion of category 1

Describing is the proportion of category 2  
Stating the category 1 proportion is -higher than the category 2 proportion

Example (the exam formula sheet probably won’t have an example)   
The proportion of some males that like action movies is 0.6 ( =0.6)  
The proportion of some females that like action movies is 0.4 ( =0.4)

So the male proportion is - =0.6-0.4=0.2 higher than the female proportion   
-=0.2 measure the effect gender has on liking action movies

To measure the effect that a quantitative variable has on another quantitative variable calculate the correlation coefficient *r*

You can explain the effect by saying   
“increasing *x* by 1 standard deviation increases *y* by *r* standard deviations”

General discussion about making prediction not given on the exam formula sheet

Case 1 ,one categorical and quantitative variable

If

The mean weight of some males is 60 (=60)

The mean weight of some females is 40 (=40)  
So the mean weight of males is - =60-40=20 higher than the mean of females   
- =20 measures the effect that gender has on weight

Then you can make a prediction about the relationship in the future   
  
The predicted mean weight of a future group of males is 60 (=60)

The predicted mean weight of future group of females is 40 (=40)

Case 2, two categorical variables

If

The proportion of some males that like action movies is 0.6 ( =0.6)  
The proportion of some females that like action movies is 0.4 ( =0.4)

So the male proportion is - =0.6-0.4=0.2 higher than the female proportion   
-=0.2 measure the effect gender has on liking action movies

You can make a prediction for a future groups of ,say, 100,000 males and 100,000 females

You would predict that 100,000\*0.6=60,000 males like action movies   
You would predict that 100,000\*0.4=40,000 females like action movies

Case 3, two quantitative variables with a linear relationship

Where the mean of the variable X is   
Where the mean of the variable Y is   
and the coefficient correlation r

If a future value of x is the predicted value y is   
“increasing *x* the future value of x 1 standard deviation increases the predicted *y* by *r* standard deviations”