UCLAEX Stat XL 10, Statistical Methods - Final Exam Review

Probabilities and Proportions

1. Suppose there were 2011 students enrolled in either Stat 10 or Stat11 at UCLA. The numbers of female and male students are given in the following table.

	Females	Males	**Total**
Stat10	604	593	1197
Stat11	387	427	814
Total	991	1020	2011

(a) Convert the above table of counts into a probability table (to 4 dp).

	Females	Males	Total
Stat 10			
Stat11			
Total			

(b) One of the 2011 students is chosen at random. What is the probability that the student chosen is:

(i) a male taking Stat 10?

(ii) a female?

(iii) a female taking Stat 11?

(c) Given that a student is taking Stat 11, what is the probability that they are male?

(d) What is the probability that a randomly chosen male student is taking Stat 11?

2. Consider drivers stopped at random for breath testing. Below is a partially completed probability table providing information about such drivers, with regards to their age (40 or under, over 40) and whether they were (or were not) wearing seat belts.

	40 or under	Over 40	**Total**
Wearing a seat belt	0.484		0.853
Not wearing seat belt		0.081	
Total			1

(a) Complete the table.

(b) What is the probability that a driver stopped at random is not wearing a seat belt?

(c) If a driver stopped at random is not wearing a seat belt, then what is the probability the driver is over 40?

(d) What is the probability that a driver stopped at random is 40 or under?

3. The medical records of a group of diabetic patients presenting at a clinic showed that 50 presented as serious cases, while 36 presented as mild cases. Of the 31 patients aged under 40, 16 presented as mild cases.

(a) Present this information in the table below.

(b) A patient is chosen at random. Find the probabilities that:

(i) the patient is under 40 and has a mild case.

(ii) the patient is at least 40 years old or has a serious case.

(iii) the patient has a serious case and is at least 40 years old.

(c) Of those presenting with serious cases, what proportion are aged under 40?

(d) Of those aged at least 40, what proportion present with mild cases?

4. A bank classifies borrowers as high-risk or low-risk. Of all its loans, 5% are in default. Forty percent (40%) of those loans in default are to high-risk borrowers, while 77% of loans not in default are to low-risk borrowers.

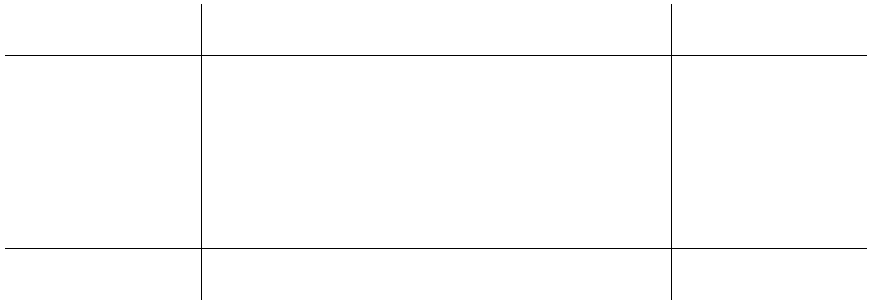
(a) Complete the table.

(b) What percentage of loans is made to borrowers in the high-risk category?

(c) What is the probability that a high-risk borrower will default on his or her loan?

5. According to recent figures from the National Centre of Educational Statistics (US), 17.5% of all bachelor's degrees are in business. 27% of bachelor's degrees in business are obtained by women and 48.75% of other degrees are obtained by men.

(a) Complete the table.



(b) What is the probability that a randomly selected recent bachelor's degree graduate will be a man?

(c) What is the probability that a randomly selected recent bachelor's degree graduate will be a man with a degree in business?

(d) What is the probability that a randomly selected female recent bachelor's degree graduate will have a degree in business?

6. .

A drinking pattern found by a survey is that 19% of male drinkers and 10% of female drinkers drink alcohol daily. Also, 51% of all drinkers are male (a 'drinker' was defined as someone who had consumed alcohol in the previous 12 months).

The probability that a randomly selected drinker from this survey who drinks alcohol daily is female is:

(1) 0.3448
(2) 0.3358
(3) 0.0490
(4) 0.1459
(5) 0.2041

UCLA Stat 10 Review Final Exam

Confidence Intervals

Section A: Confidence intervals for a mean, proportion and difference between means

1. An exam had a possible total of 64 points.
A random sample of 30 scores was selected from all of the exams. The data collected and its summary is as follows:

46	32	24	20	51	33	35	43	26	29	59	41	30	35	49
53	32	50	52	23	25	53	51	34	26	29	40	38	45	42

sample size	sample mean	sample standard deviation
30	38.20	10.85

You will use this sample to construct a 95% confidence interval for the mean.

(a) State the parameter θ (using a symbol and in words).

(b) State the estimate $\hat{\theta}$ (using a symbol, in words and as a number).

(c) Calculate se($\hat{\theta}$).

(d) State the value of *df*.

(e) Use the table for the Student's *t*-distribution to write down the value of the *t*-multiplier.

(f) Calculate the 95% confidence interval for the mean.

(g) Interpret the confidence interval.

(h) Does the confidence interval contain the true mean? Discuss briefly.

2. Tuberculosis (TB) is known to be a highly contagious disease. In 1995 a study was carried out on a random sample of 1074 Spanish prisoners. The study investigated factors that might be associated with the tuberculosis infection. Some of the results follow.

	Prisoners with tuberculosis	Total number of prisoners
Male	556	984
Female	36	90

You will use this sample to construct a 95% confidence interval for the proportion of female prisoners who had tuberculosis.

(a) State the parameter θ (using a symbol and in words).

(b) State the estimate $\hat{\theta}$ (using a symbol, in words and as a number).

(c) Calculate se($\hat{\theta}$).

(d) Use the table for the Student's *t*-distribution to write down the value of the *z*-multiplier.

(e) Calculate the 95% confidence interval for the proportion of female prisoners who had tuberculosis.

(f) Interpret the confidence interval.

(g) Does the confidence interval contain the true proportion? Discuss briefly.

3. .

Banford et al. [1982] noted that thiol concentrations within human blood cells are seldom determined in clinical studies, in spite of the fact that they are believed to play a key role in many vital processes. They reported a new reliable method for measuring thiol concentration and demonstrated that, in one disease at least (rheumatoid arthritis), the change in thiol status in the lysate from packed blood cells is substantial. There were two groups of volunteers, the first group being "normal" and the second suffering from rheumatoid arthritis. We shall treat the two groups as random samples from the normal and rheumatoid populations respectively (for the area in which the study was undertaken) and will estimate μ_R - μ_N, the difference in true mean thiol levels between the rheumatoid and normal populations.

Computer Output

```
Two sample T for Rheumatoid vs Normal

          N      Mean     StDev   SE Mean
Rheumato  6     3.465     0.440      0.18
Normal    7    1.9214    0.0756     0.029

95% CI for mu Rheumato - mu Normal: ( 1.08,  2.012)
T-Test mu Rheumato = mu Normal (vs not =): T = 8.48  P = 0.0004  DF = 5
```

(a) Interpret the confidence interval.

(b) Does the confidence interval contain the difference in true mean thiol levels between the rheumatoid and normal populations? Discuss briefly.

4. .

Consider constructing a confidence interval for the mean of a population. Which of the following would have an effect on the width of the confidence interval?

I: The size of the sample used to construct the interval.

II: The confidence level for the interval.

III: The amount of variability in the population.

---

(1) I only.

(2) I and II only.

(3) I and III only.

(4) I, II, and III.

(5) II and III only.

5. .

A 95% confidence interval for the difference between the mean haemoglobin levels of people with Type III and people with Type II sickle cell disease, $\mu_{\text{Type III}}$ - $\mu_{\text{Type II}}$, is [-0.80, 2.39]. A **correct** interpretation of this interval would be:

(1) Since zero is in the interval, there is a difference between the average haemoglobin levels for people with Type II sickle cell disease and people with Type III sickle cell disease.

(2) We estimate, with 95% confidence, the average haemoglobin level for people with Type III sickle cell disease to be somewhere between 0.80g/dL lower and 2.39g/dL higher than the average haemoglobin level for people with Type II sickle cell disease.

(3) We estimate, with 95% confidence, the average haemoglobin level for people with Type II sickle cell disease to be somewhere between 0.80g/dL lower and 2.39g/dL higher than the average haemoglobin level for people with Type III sickle cell disease.

(4) On average, people with Type II sickle cell disease have a lower haemoglobin level than people with Type III sickle cell disease.

(5) Since zero is in the interval, there is no difference between the average haemoglobin levels for people with Type II sickle cell disease and people with Type III sickle cell disease.

Questions 6 and **7** refer to the following information.

.

In 1990 *CNN/Time* sought information on how young American adults viewed their parents' marriage. In a telephone poll, one of the questions they asked of six hundred and two (602) 18-29 year old Americans was "Would you like to have a marriage like the one your parents have?" Forty-four percent (44%) responded "Yes".

6. *CNN/Time* were interested in determining what proportion of the 18-29 year old American population would answer "Yes" to this question. Which **one** of the following statements is **false**?

(1) The value of the parameter of interest is an unknown quantity.

(2) In this context, 0.44 is an estimate for the parameter of interest.

(3) The parameter of interest depends on the sample and hence is a random quantity.

(4) A confidence interval for the parameter of interest will give a range of possible values for this parameter.

(5) The parameter of interest is the proportion of 18-29 year old Americans who would have answered "Yes" in 1990.

7. An approximate 95% confidence interval for the proportion of the 18-29 year old American population who would have answered "Yes" to this question in 1990 is [0.400, 0.480]. If two thousand four hundred (2400) 18-29 year old Americans had been sampled instead of six hundred and two (602) 18-29 year old Americans, then the new 95% confidence interval would be approximately:

(1) twice as wide.

(2) one-quarter as wide.

(3) half as wide.

(4) four times as wide.

(5) equally as wide.

8. .

The *Listener/Heylen* poll from August 6, 1994 reported results on what New Zealanders think about the "Ten Commandments" from a sample of 1000 randomly chosen New Zealanders. A 99% confidence interval for the proportion of New Zealanders who believed that the commandment "I am the Lord your God; worship no god but me" fully applied to them, p_G, is given by (0.282, 0.358). Which **one** of the following statements is **true**?

(1) The interval (0.282, 0.358) will cover the true, but unknown parameter p_G for 99% of samples taken.

(2) Between 28.2 and 35.8 per cent of New Zealanders believe that this commandment fully applies to them 99% of the time.

(3) A 95% confidence interval for p_G would be wider than this interval.

(4) The probability that the interval (0.282, 0.358) covers the sample proportion is 0.99.

(5) The probability that another interval calculated in the same way from a new sample of 1000 New Zealanders covers p_G is 0.99.

9. .

The results of a survey of 1146 New Zealanders were published in the 23 March 1992 issue of Time magazine. In response to the question "Is it a good time to buy a major household item?" 585 respondents replied "yes", 332 replied "no" and 229 replied "don't know".

Let p represent the true proportion of New Zealanders who think it is a good time to buy a major household item. Using the results of this survey a 99% confidence interval for p and a 95% confidence interval for p were constructed. A two standard error interval for p was also constructed.

Which **one** of the following statements is **true**?

The 99% confidence interval would:

(1) be completely contained by the corresponding 95% confidence interval for p.

(2) be narrower than the corresponding two standard error interval for p.

(3) be wider if a much larger sample had been taken.

(4) be wider than the corresponding 95% confidence interval for p.

(5) have confidence limits which are twice as far apart as the confidence limits for the corresponding 95% confidence interval for p.

Section B: Confidence interval for a difference in proportions

1. In 1991 a random sample of New Zealand adults were surveyed about their working hours and the number of jobs they had. A similar survey was carried out in 1994.

Identify the sampling situation as:
Situation (a): *Two independent samples*,
Situation (b): *Single sample, several response categories*,
Situation (c): *Single sample, two or more Yes/No items*,
in the following cases.

(a) We want to compare the proportion of females working 1-39 hours in 1994 with the proportion of females working 40 hours or more in 1994.

(b) We want to compare the proportion of males working 40 hours or more in 1991 with the proportion of females working 40 hours or more in 1991.

(c) In the same survey people were also asked if they had 2 or more jobs. We want to compare the proportion of people who had 2 or more jobs in 1994 with the proportion of people who worked 40 hours or more per week in 1994.

(d) We want to compare the proportion of females working 40 hours or more in 1994 with the proportion of females working 40 hours or more in 1991.

Questions 2 to 6 refer to the following information.
Tuberculosis (TB) is known to be a highly contagious disease. In 1995 a study was carried out on a random sample of 1074 Spanish prisoners. The study investigated factors that might be associated with the tuberculosis infection. The results follow.

Variable		Prisoners with tuberculosis	Total number of prisoners
Gender	Male	556	984
	Female	36	90
Race	White	496	886
	Gypsy	74	152
	Other	22	36
Intravenous Drug Users	Yes	361	629
	No	231	445
HIV Positive	Yes	186	294
	No	406	780
Re-imprisonment	Yes	272	456
	No	320	618

2. Identify the sampling situation as:
Situation (a): *Two independent samples*,
Situation (b): *Single sample, several response categories*,
Situation (c): *Single sample, two or more Yes/No items*,
in the following cases:

(a) Of those prisoners who had TB, we want to compare the proportion of white prisoners with the proportion of Gypsy prisoners.

(b) We want to compare the proportion of male prisoners who had TB with the proportion of female prisoners who had TB.

(c) We want to compare the proportion of prisoners who were intravenous drug users with the proportion of prisoners who had been re-imprisoned.

(d) We want to compare the proportion of white prisoners who had TB with the proportion of Gypsy prisoners who had TB.

(e) Of those prisoners who had TB, we want to compare the proportion who were intravenous drug users with the proportion who were HIV-positive.

(f) We want to compare the proportion of Gypsy prisoners with the proportion of prisoners whose race was categorised as "other".

3. The standard error of the difference between the proportion of prisoners who have TB that are intravenous drug users and the proportion of prisoners who have TB that are HIV positive is:

(1) $\sqrt{\dfrac{0.6098(1-0.6098)+0.3142(1-0.3142)}{592}}$

(2) $\sqrt{\dfrac{0.6098+0.3142-(0.6098-0.3142)^2}{592}}$

(3) $\sqrt{\dfrac{0.6098+0.3142+(0.6098-0.3142)^2}{592}}$

(4) $\sqrt{\dfrac{0.6098(1-0.6098)}{629}+\dfrac{0.3142(1-03142)}{294}}$

(5) $\sqrt{\dfrac{0.6098^2+0.3142^2}{592}}$

4. Construct a 95% confidence interval for the difference between the proportion of White prisoners who were infected with TB and the proportion of Gypsy prisoners who were infected with TB. State what your interval tells you in plain English.

(a) State the parameter θ (using symbols and in words).

(b) State the estimate $\hat{\theta}$ (using symbols, in words and as a number).

(c) Calculate se($\hat{\theta}$).

(d) Use the table for the Student's *t*-distribution to write down the value of the *z*-multiplier.

(e) Calculate the confidence interval.

(f) Interpret the confidence interval.

5. Construct a 95% confidence interval for the difference in the proportion of prisoners infected with TB who were white and the proportion of prisoners infected with TB who were Gypsy.

(a) State the parameter θ (using symbols and in words).

(b) State the estimate $\hat{\theta}$ (using symbols, in words and as a number).

(c) Calculate se($\hat{\theta}$).

(d) Use the table for the Student's *t*-distribution to write down the value of the *z*-multiplier.

(e) Calculate the confidence interval.

(f) Interpret the confidence interval.

6. Let p_Y be the proportion of intravenous drug user prisoners who were infected with TB, and p_N be the proportion of non-intravenous drug user prisoners who were infected with TB. The *Excel* worksheet below shows the calculations for a 95% confidence interval based on the data shown on the first page.

Two population proportions

Input data	
X1_sample	361
X2_sample	231
n1_total	629
n2_total	445
p1_ratio	0.573926868
p2_ratio	0.519101124
pdiff	0.054825744
Alpha	0.05
Calculated value	
se	0.03081794
t-multiplier	1.959961082
Confidence Interval	
Lower limit	-0.00557622
Upper limit	0.115227708

(a) Which sampling situation applies here? Briefly explain why.

(b) Interpret the confidence interval.

(c) Is it plausible that p_Y is equal to p_N? Justify your answer.

7. .

A *Time/CNN* poll was based on a telephone survey of 800 adult Hong Kong residents conducted two weeks before the hand over of Hong Kong to China. p_c is the proportion of people in Hong Kong who think "Corruption" is the issue which worries them most, and p_f is the proportion of people in Hong Kong who think "Reduced personal freedoms" is the issue which worries them most.

A 95% confidence interval for $p_c - p_f$ is (0.012, 0.088). Which **one** of the following statements is **false**?

(1) In repeated sampling, we would expect that 95% of the 95% confidence intervals produced contain the true value of $p_c - p_f$.

(2) In light of the data, the interval (0.012, 0.088) contains the most plausible values for $p_c - p_f$.

(3) The true value of $p_c - p_f$ must be in the interval (0.012, 0.088).

(4) At this level of confidence, statements such as "p_c is bigger than p_f by somewhere between 0.012 and 0.088" are true, on average, 19 out of 20 times.

(5) With 95% confidence, the true value of $p_c - p_f$ is 0.05 with a margin of error of ±0.038.

8. .

In a Time Morgan poll (July 1994) 662 voters were interviewed by telephone and asked whether *developing the economy* or *protecting the environment* would be more important in the short term. There were 238 National and 162 Labour supporters in the poll.

Let p_N be the true proportion of National supporters and let p_L be the true proportion of Labour supporters who think that *protecting the environment* is more important in the short term. A 95% confidence interval for the difference between the proportions p_N - p_L is given by [-0.1526, 0.0326]. Which **one** of the following interpretations is **true**?

(1) With a probability of 0.95, the true difference of proportions p_N - p_L lies between –0.1526 and 0.0326.

(2) In repeated sampling the 95% confidence interval [-0.1526, 0.0326] will contain the true difference in proportions in 95% of the samples taken.

(3) In repeated sampling the true proportion p_N will be somewhere between 0.1526 larger and 0.0326 smaller than p_L.

(4) With 95% confidence the true proportion p_N is somewhere between 0.1526 smaller and 0.0326 larger than p_L.

(5) With 95% confidence the true proportion p_N is 0.1852 larger than p_L.

UCLA Stat 10 Final Exam Review

Hypothesis Testing

Section A:

1. What is the difference between the null hypothesis and the alternative hypothesis?

2. List the three alternative hypotheses that are possible when testing the null hypothesis $H_0: \quad = \theta_0$.

(a) **(b)** **(c)**

3. On what basis do we decide whether to do a one-tailed or a two-tailed test?

θ

4. Use the Formulae Sheet to fill in the formula for the *t*-test statistic:

(a) In words:

(b) In symbols:

5. Fill in the gaps in this description of the *P-value* from a *t*-test:

The *P-value* is the _____ that, if the _____ _____ was true, _____ _____ would produce an estimate that is _____ _____ _____ _____ _____ from the _____ _____ than our _____ _____.

6. What does the *P-value* measure?

7. In the table below, fill in the descriptions used for the given *P-values*:

P-value	**Evidence against H_0**
> 0.12	
≈ 0.10	
≈ 0.05	
≈ 0.01	
< 0.01	

8. What does the *P-value* tell us about the size of the effect?

9. What tool do we use to tell us about the size of an effect?

10. What is a significance test?

11. What does a significant test reveal?

Section B:

1. Tuberculosis (TB) is known to be a highly contagious disease. In 1995 a study was carried out on a random sample of 1074 Spanish prisoners. The study investigated factors that might be associated with the tuberculosis infection. The results follow.

Variable		Prisoners with tuberculosis	Total number of prisoners
Gender	Male	556	984
	Female	36	90
Race	White	496	886
	Gypsy	74	152
	Other	22	36
Intravenous Drug Users	Yes	361	629
	No	231	445
HIV Positive	Yes	186	294
	No	406	780
Re-imprisonment	Yes	272	456
	No	320	618

Is there any evidence to suggest that the race of the prisoner (White or Gypsy) makes any difference to whether they contracted tuberculosis? Carry out a significance test to answer this question and then calculate an appropriate 95% confidence interval.

Let p_W be the proportion of White prisoners infected with TB and p_G be the proportion of Gypsy prisoners infected with TB.

(a) Identify the parameter θ.

(b) State the hypotheses.

(c) Write down the estimate and its value.

(d) Calculate the value of the *t*-test statistic.

(e) Find the *P-value*.

(f) Interpret the *P-value*.

(g) Answer the original question.

(h) Calculate a 95% confidence interval for the parameter.

(i) Interpret the 95% confidence interval.

2. In a poll conducted for TIME and CNN (TIME 17 September 1990, page 51), 1009 residents of New York City were asked "If you could choose where you live, would you live in New York City or move somewhere else?" 595 of the residents said that they would move somewhere else. Could you conclude that this is the opinion of a majority of residents of New York City?

The information below is a Compuiter output for a significance test and a 95% confidence interval. Use this output to answer the questions below.

Test and Confidence Interval for One Proportion

```
Test of p = 0.5 vs p not = 0.5

Sample      X      N  Sample p        95.0 % CI       Z-Value  P-Value
1         595   1009  0.589693  (0.559342, 0.620044)     5.70    0.000
```

(a) State the parameter used in this analysis.

(b) State the hypotheses used in this *t*-test.

(c) Write down the estimate and its value.

(d) Write down the value of the test statistic and the *P-value*.

(e) Answer the original question. (I.e., could you conclude that this is the opinion of a majority of residents of New York City?)

(f) Interpret the confidence interval.

3. A businessperson is interested in buying a coin-operated laundry and has a choice of two different businesses. The businessperson is interested in comparing the average daily revenue of the two laundries, so she collects some data. A simple random sample for 50 days from the records for the past five years of the first laundry and a simple random sample for 30 days from the records for the past three years of the second laundry reveal the following summary statistics:

	Sample size	Sample mean	Sample standard deviation
Laundry 1	50	\$635.40	\$71.90
Laundry 2	30	\$601.60	\$77.70

Computer output

Two Sample T-Test and Confidence Interval

```
Two sample T for Laundry1 vs Laundry2

           N      Mean     StDev   SE Mean
Laundry1  50     635.4      71.9        10
Laundry2  30     601.6      77.7        14

95% CI for mu Laundry1 - mu Laundry2: ( -1,  69)
T-Test mu Laundry1 = mu Laundry2 (vs not =): T = 1.94  P = 0.057  DF = 57
```

```
Stem-and-leaf of Laundry1  N  = 50
Leaf Unit = 10

  1    4 7
  1    4
  2    5 0
  5    5 233
  6    5 5
 10    5 6777
 16    5 889999
 18    6 01
(10)   6 2222233333
 22    6 4444445555
 13    6 67
 11    6 889
  8    7 1
  7    7 23
  5    7 5
  4    7 77
  2    7 99
```

```
Stem-and-leaf of Laundry2  N  = 30
Leaf Unit = 10

  1    4 3
  2    4 6
  9    5 0122344
 15    5 558889
 15    6 113344444
  6    6 77
  4    7 023
  1    7 5
```

(a) State the parameter used in this analysis.

(b) State the hypotheses used in this *t*-test.

(c) Write down the estimate and its value.

(d) Write down the value of the test statistic.

(e) Interpret the test.

(f) Interpret the confidence interval.

(g) Do the stem-and-leaf plots give you any reasons for doubting the validity of the results of this analysis? Briefly explain.

(h) If this analysis were done by hand the value of *df* would have been 29. Why does the output show that *df* = 57?

Section C:

1. Which **one** of the following statements regarding significance testing is **false**?

(1) A highly significant test result means that the size of the difference between the estimated value of the parameter and the hypothesised value of the parameter is significant in a practical sense.

(2) A *P-value* of less than 0.01 is often referred to as a highly significant test result.

(3) A nonsignificant test result does not necessarily mean H_0 is true.

(4) A two-tail test of H_0: $\theta = \theta_0$ is significant at the 5% significance level if and only if θ_0 lies outside a 95% confidence interval for θ.

(5) Testing at the 5% level of significance means that the null hypothesis is rejected whenever a *P-value* smaller than 5% is obtained.

2. Which **one** of the following statements is **false**?

(1) In hypothesis testing, a nonsignificant result implies that H_0 is true.

(2) In hypothesis testing, a two-tail test should be used when the idea for doing the test has been triggered by the data.

(3) In surveys, the nonsampling error is often greater than the sampling error.

(4) Larger sample sizes lead to smaller standard errors.

(5) In hypothesis testing, statistical significance does not necessarily imply practical significance.

3. Which **one** of the following statements regarding significance testing is **false**?

(1) Formal tests can help determine whether effects we see in our data may just be due to sampling variation.

(2) The *P-value* associated with a two-sided alternative hypothesis is obtained by doubling the *P-value* associated with a one-sided alternative hypothesis.

(3) The *P-value* says nothing about the size of an effect.

(4) The data should be carefully examined in order to determine whether the alternative hypothesis needs to be one-sided or two-sided.

(5) The *P-value* describes the strength of evidence against the null hypothesis.