DAWSON COLLEGE
MATHEMATICS DEPARTMENT
Probability & Statistics

201-BZS-05 S01
Fall 2016
Final Examination
December 15th, 2016
Time Limit: 3 hours

Name: _____

ID#: _____

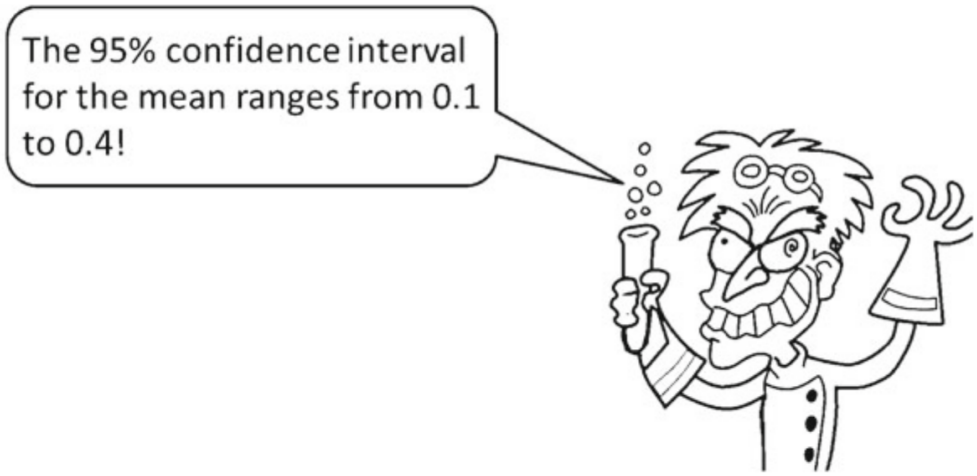
Instructor: Y. Lamontagne

- This exam contains 16 pages (including this cover page) and 17 problems. Check to see if any pages are missing.
- Answer the questions in the spaces provided on the question sheets. If you run out of room for an answer, continue on the back of the page, and please indicate that you have done so.
- Give the work in full; – unless otherwise stated, reduce each answer to its simplest, exact form; – and write and arrange your exercise in a legible and orderly manner.
- You are only permitted to use the **Sharp EL-531XG** or **Sharp EL-531X** calculator.
- This examination booklet must be returned intact.
- Good luck!

Question	Points	Score
1	3	
2	2	
3	2	
4	5	
5	5	
6	4	
7	5	
8	11	
9	5	
10	5	
11	10	
12	6	
13	5	
14	5	
15	5	
16	5	
17	11	
Total:	94	

1. [1]

Professor Bumbledorf conducts an experiment, analyzes the data, and reports:



Please mark each of the statements below as "true" or "false". False means that the statement does not follow logically from Bumbledorf's result. Also note that all, several, or none of the statements may be correct:

(a) (1 mark) The probability that the true mean is greater than 0 is at least 95 %. **True** or **False**

Answer:
False

(b) (1 mark) There is a 95 % probability that the true mean lies between 0.1 and 0.4. **True** or **False**

Answer:
False

(c) (1 mark) If we were to repeat the experiment over and over, then 95 % of the time the true mean falls between 0.1 and 0.4. **True** or **False**

Answer:
False

2. (2 marks) Briefly explain the meaning of a 95 % confidence interval.

Answer:
95 % of confidence intervals contructed from point estimates will contain the population parameter.

3. (2 marks) What is the probability of randomly answering Question 1. and getting at least 2 correct answers?

[1] modified from Hoekstra, R., Morey, R.D., Rouder, J.N. et al. Psychon Bull Rev (2014) 21: 1157.

Answer:

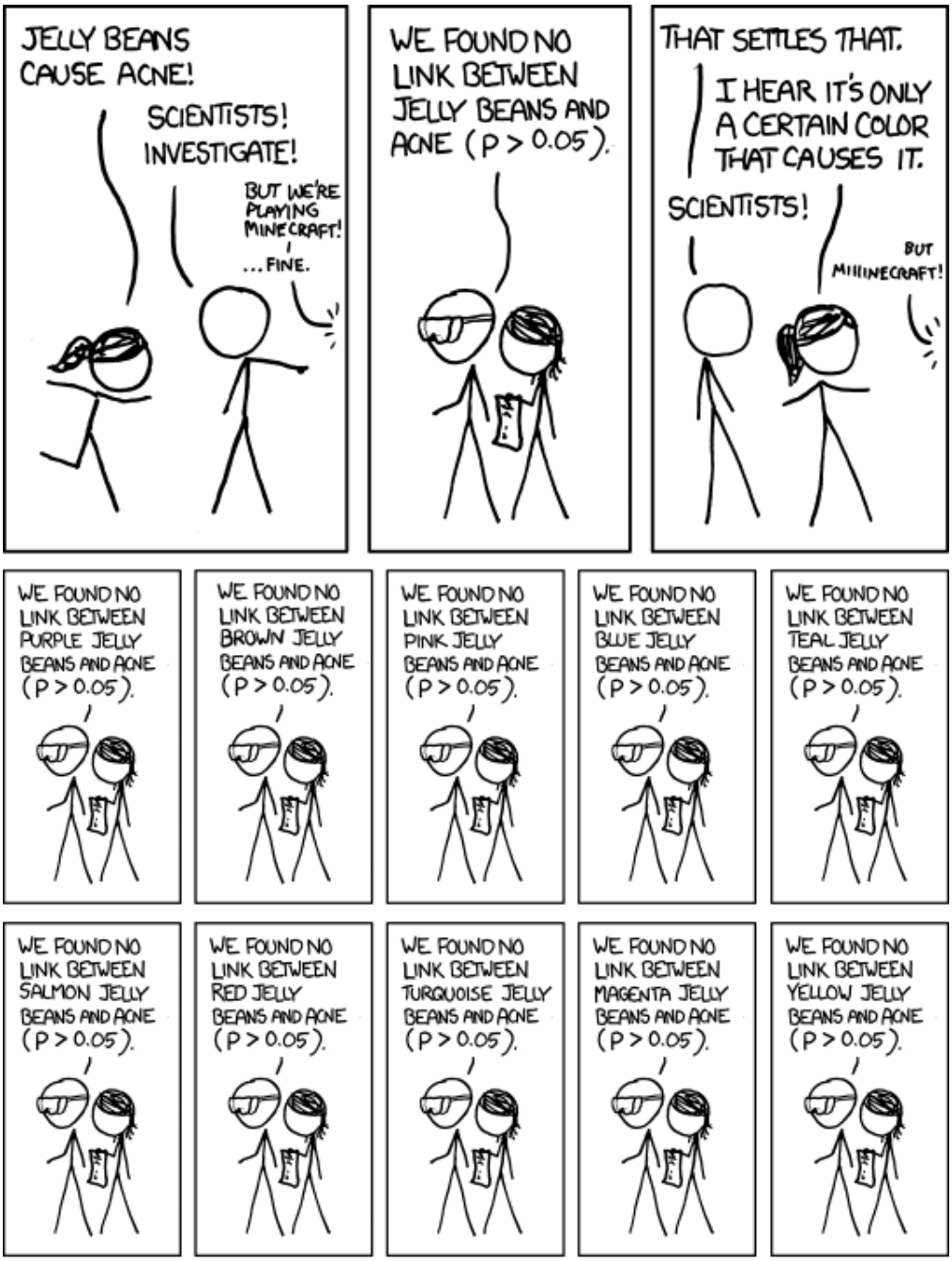
0.5

4. (a) (2 marks) Give the definition of *p-value*.

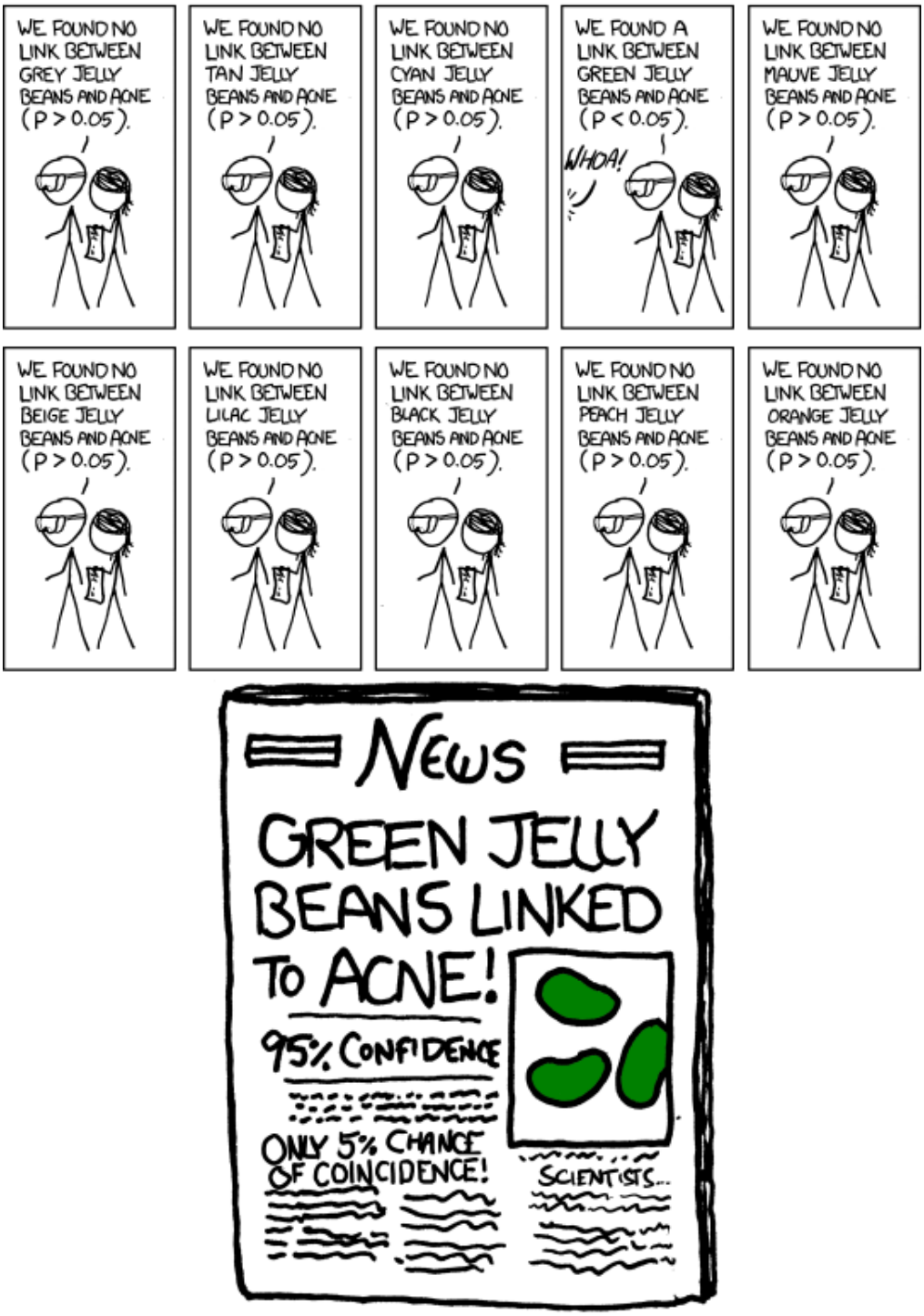
Answer:

The *p-value* of a given point estimate is the probability of observing a point estimate at least as extreme as the given point estimate in the direction of the alternative hypothesis when the null hypothesis is true.

(b) (3 marks) Read the following comic strip[2] *(there is part of the comic on the next page)* then explain why the "green jelly beans and acne study" featured in the comic should probably not have been published. Your explanation should be based on the p-value.



[2]Comic copied from https://xkcd.com/882/



Answer:

Suppose that there is no link between color and acne and suppose the conditions for a hypothesis test.

The probability that a point estimate has a p-value smaller than 0.05 is $0.05 = \frac{1}{20}$. If an experiment is performed a significant number of times where point estimates are obtained, then some point estimates will have p-values smaller than a given significance level.

Hence observing one point estimate (with p-value smaller than the significance level) out of many can provide little evidence for the alternate hypothesis.

In fact the probability of the occurance of the above scenario (that is, having one p-value out of 20 being smaller than the significance level) is about 0.38 which does not provide evidence for the alternate hypothesis (Green jelly beans linked to acne).

5. Given

$$f(x) = \begin{cases} \frac{1}{x} & \text{if } x \in [1,b] \\ 0 & \text{if } x \notin [1,b] \end{cases}.$$

(a) (2 marks) Determine the value of b such that $f(x)$ probability density function.

Answer:
$b = e$

(b) (3 marks) Find μ and $P(\frac{3}{2} < X < \mu)$ where X is continuous random variable with density function $f(x)$ and the value of b found in part a.

Answer:
$\mu = e-1, \quad P\left(\frac{3}{2} < X < \mu\right) = \ln\left(\frac{2e-2}{3}\right)$

6. The expression used in evaluating the *R score*[3] is:

$$\text{R score} = (Z + ISG + 5) \times 5$$

where Z and ISG are the numerical expressions for the Z score and the indicator of the strength of the group, respectively. The ISG is defined as

$$ISG = \frac{\text{average grade of the group at the secondary level} - 75}{14}$$

(a) (2 marks) Suppose that the average grade of the group at the secondary now taking Statistics and Probability at CEGEP is 87%. The Statistics and Probability class average is 75% and standard deviation is 15%. Compute the R score of a student who has an average of 85%.

Answer:
32.6

(b) (2 marks) For a different course, the student earns an R score of 38. Find the percentile in which the student ranks in the course given that the students' classmates are the same as in part a. and the grades are normally distributed.

Answer:
96 %

[3]CREPUQ. The R score : what it is, and what it does. September 3, 2004

7. (5 marks) It is believed that 2% of the population suffer from Lupus disease. The test to detect Lupus is 98% accurate if a person actually has the disease. The test is 74% accurate if a person does not have the disease. There is a line from a television show that is used after a patient tests positive for Lupus: "It's never Lupus." Do you think there is truth to this statement? Use appropriate probabilities to support your answer.[4]

Answer:
93 % of tests are false positives. Hence there is some truth to this statement.

8. The daily high temperature reading on January 1 was collected in 1968 and 2008 for 51 randomly selected locations in the continental US. Then the difference between the two readings (temperature in 2008 - temperature in 1968) was calculated for each of the 51 different locations. The average of these 51 values was 1.1 degrees with a standard deviation of 4.9 degrees. We are interested in determining whether this data provides strong evidence of temperature warming in the continental US between the years 1968 and 2008.[5]

   (a) (1 mark) Is there a relationship between the observations collected in 1968 and 2008? Or are the observations in the two groups independent? Explain.

   Answer:
   Dependent since temperature measurements are from the same locations for the two years.

   (b) (1 mark) Write hypotheses for this research in symbols and in words.

   Answer:
   $H_0 : \mu_d = \mu_{2008} - \mu_{1968} = 0$ No difference in temperature for the two years in the continental US.
   $H_a : \mu_d = \mu_{2008} - \mu_{1968} > 0$ Temperature greater in 2008 in the continental US.

   (c) (1 mark) Check the conditions required to complete this test.

   Answer:
   sample is independent, sample is large

   (d) (2 marks) Calculate the test statistic and an interval for the p-value.

   Answer:
   $T = 1.60, \quad df = 50, \quad 0.050 < \text{p-value} < 0.100$

   (e) (1 mark) What do you conclude at 5% significance? Interpret your conclusion in context.

   Answer:
   Fail to reject H_0. There is not strong evidence in this experiment to suggest temperature warming in the continental US between the years 1968 and 2008.

[4] modified from OpenIntro Statistics by D.M. Diez, C.D. Barr and M. Çetinkaya-Rundel, OpenIntro LaTeX, code, and PDFs are released under a Creative Commons BY-SA 3.0 license.

[5] modified from OpenIntro Statistics by D.M. Diez, C.D. Barr and M. Çetinkaya-Rundel, OpenIntro LaTeX, code, and PDFs are released under a Creative Commons BY-SA 3.0 license.

(f) (1 mark) What type of error might have been made? Explain what the error means in the context of this problem?

Answer:
A type II error. Failing to conclude that the experiment suggests temperature warming, given that there is in fact temperature warming.

(g) (4 marks) Find an approximation*("a good approximation")* of the probability of the above error given, the hypothesis test at 5% significance and $\mu_a = 1.1$.

Answer:
$\beta = 0.53$

9. A professor using an open source introductory statistics book predicts that 60% of the students will purchase a hard copy of the book, 25% will print it out from the web, and 15% will read it online. At the end of the semester he asks his students to complete a survey where they indicate what format of the book they used. Of the 126 students, 71 said they bought a hard copy of the book, 30 said they printed it out from the web, and 25 said they read it online.[6]

(a) (1 mark) State the hypotheses for testing whether the professor's predictions were inaccurate.

Answer:
H_0 : The professor's predictions were accurate.
H_a : The professor's predictions were inaccurate.

(b) (1 mark) How many of his 126 students did the professor expect to buy the book, print the book, and read the book exclusively online?

Answer:
$e_1 = 75.6, \quad e_2 = 31.5, \quad e_3 = 18.9$

(c) (1 mark) Is an appropriate setting for a chi-square test? List the conditions required for a test and verify they are satisfied.

Answer:
Assuming a random sample and $e_i > 5$.

(d) (1 mark) Calculate the chi-squared statistic.

Answer:
$\chi^2 = 2.32$

(e) (1 mark) Based on the above test statistic, what is the conclusion of the hypothesis test? Interpret your conclusion.

[6]modified from OpenIntro Statistics by D.M. Diez, C.D. Barr and M. Çetinkaya-Rundel, OpenIntro LaTeX, code, and PDFs are released under a Creative Commons BY-SA 3.0 license.

Answer:
$df = 2$, p-value > 0.3 hence fail to reject H_0. Therefore, fail to conclude that professor's predictions were inaccurate.

10. The local police decide to listen to journalists' phone calls but have limited resources and therefore can only listen to calls of the same 5 journalists in any given month. The local news agency A has 13 journalists, news agency B has 14 journalists.[7]

    (a) (1 mark) How many different groups of 5 journalists can be listened to by the police in any given month?

    Answer:
    $\binom{27}{5}$

    (b) (2 marks) If 5 journalists are selected by the police at random what is the probability that at least 2 are from the news agency A.

    Answer:
    0.814

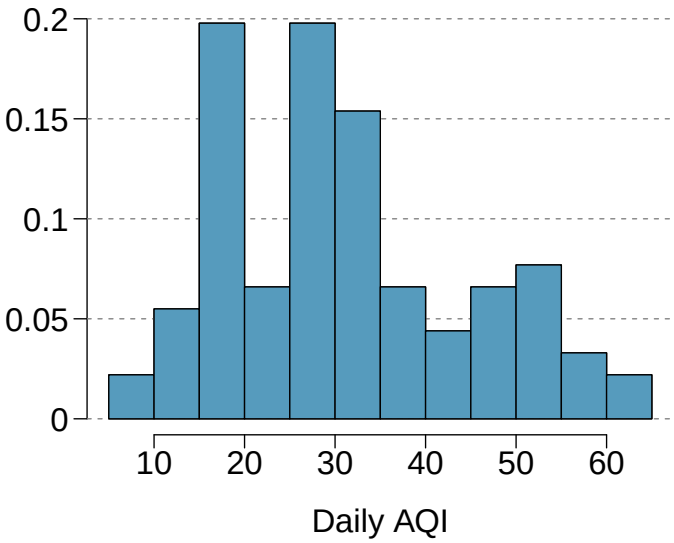
    (c) (2 marks) If 5 journalists are selected by the police at random what is the probability that at least 1 is from the news agency B given that at least 2 are from the news agency A.

    Answer:
    0.8716

11. Daily air quality is measured by the air quality index (AQI) reported by the Environmental Protection Agency. This index reports the pollution level and what associated health effects might be a concern. The index is calculated for five major air pollutants regulated by the Clean Air Act and takes values from 0 to 300, where a higher value indicates lower air quality. AQI was reported for a sample of 91 days in 2011 in Durham, NC. The relative frequency histogram below shows the distribution of the AQI values on these days.[8]

[7] Almost entirely fictional.

[8] modified from OpenIntro Statistics by D.M. Diez, C.D. Barr and M. Çetinkaya-Rundel, OpenIntro LaTeX, code, and PDFs are released under a Creative Commons BY-SA 3.0 license.



(a) (2 marks) Find an approximation of the mean of the sample using the mark of each class of the histogram.

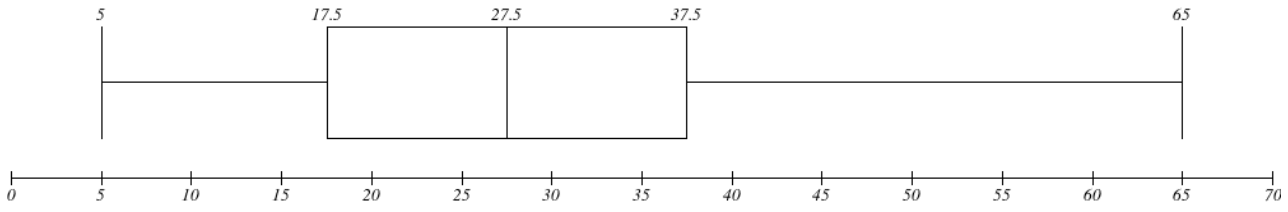
Answer:
≈ 32

(b) (2 marks) Estimate the median, Q1, Q3, and IQR for the distribution.

Answer:
Q1 ≈ 17.5, median ≈ 27.5, Q3 ≈ 37.5, $IQR \approx 20$

(c) (3 marks) Sketch an approximate box plot of the distribution.

Answer:



(d) (1 mark) Would any of the days in this sample be considered to have an unusually low or high AQI? Explain your reasoning.

Answer:
No, the data is contained within the wiskers.

(e) (1 mark) Describe the type of data and the level of measurement in this problem. (i.e. numerical data, categorical data, continuous data, discrete data, nominal data, etc).

Answer:
numerical data, ratio data

(f) (1 mark) If the above data was used for a study would it be an observational study or experimental study? Justify.

Answer:
Observational study, since there is no treatment group.

12. Given the events A, B and C where A and C are disjoint events, A and B are independent events, and the following assigned probabilities: $P(B) = 0.05, \quad P(C) = 0.25, \quad P(A \cup B) = 0.1, \quad P(B \cup C') = 0.2$

(a) (2 marks) Compute $P(A)$

Answer:
0.05263

(b) (2 marks) Compute $P(B \cap C')$

Answer:
0.6

(c) (2 marks) Compute $P(A \cup B \cup C \mid C)$

Answer:
1

13. (5 marks) Given a set of n bivariate data (x_i, y_i), its associated means $\bar{x}$ and $\bar{y}$, and the least square line $\hat{y} = b_0 + b_1 x$ for the bivariate data. Show that the sum of the residuals is equal to zero, that is, $\sum_{i=1}^{n} (\hat{y}_i - y_i) = 0$. *(Hint: you can use the fact $\bar{y} = b_1\bar{x} + b_0$)*

Answer:

$$
\begin{aligned}
\bar{y} &= b_1\bar{x} + b_0 \\
\frac{\sum y_i}{n} &= b_1\frac{\sum x_i}{n} + b_0 \\
\sum y_i &= b_1\sum x_i + nb_0 \\
\sum y_i &= \sum(b_1x_i + b_0) \\
\sum y_i &= \sum \hat{y}_i \\
0 &= \sum \hat{y}_i - \sum y_i \\
0 &= \sum(\hat{y}_i - y_i)
\end{aligned}
$$

14. (5 marks) Yann is testing braking efficiency of two brands of winter bicycle tyres. For each trial he maintains a specific speed and at a marker applies the rear brakes then measures the distance between the marker and the position where the bicycle came to a stop. The data is summarized in the table below. Assuming normal distribution of both populations conduct a hypothesis test to determine if this data provides evidence ($\alpha = 0.05$) that one brand is superior to the other. *(Write the hypotheses, check conditions, write a conclusion supported by the test statistic)*

	Brand A	Brand B
n	6	9
$\sum x$	27	46
$\sum x^2$	127	278

Answer:

$H_0: \mu_A - \mu_B = 0$
$H_a: \mu_A - \mu_B \neq 0$
Assuming that populations are normally distributed and the samples are independent.
$df = 11$, $t_{\alpha/2} = 2.20$, $T = -0.69$
Fail to reject H_0, the data does not provide evidence that one brand is superior to the other.

15. (5 marks) A 2010 survey asked 827 randomly sampled registered voters in California "Do you support? Or do you oppose? Drilling for oil and natural gas off the Coast of California? Or do you not know enough to say?" Below is the distribution of responses, separated based on whether or not the respondent graduated from college.[9]

Conduct a hypothesis test to determine if the data provides strong evidence that the proportion of college graduates who do not have an opinion on this issue is different than that of non-college graduates. *(Write the hypotheses, check conditions, write a conclusion supported by the p-value)*

	College Grad	
	Yes	No
Support	154	132
Oppose	180	126
Do not know	104	131
Total	438	389

Answer:

$H_0: p_g - p_{ng} = 0$
$H_a: p_g - p_{ng} \neq 0$
Samples are assumed to be independent since sample size is less than 10% of population.
Since $\hat{p}_g n_g$, $(1-\hat{p}_g)n_g$, $\hat{p}_{ng} n_{ng}$, $(1-\hat{p}_{ng})n_{ng} \geq 10$ we can use the normal distribution.
$\hat{p} = 0.2842$, SE $= 0.0314$, $Z = -3.16$, p-value $= 0.0014$
Reject H_0, since the p-value is very small. Therefore the data supports that there is a difference in the proportions.

16. The 2010 General Social Survey asked 1,259 US residents: "Do you think the use of marijuana should be made legal, or not?" 48% of the respondents said it should be made legal.[10]

[9] OpenIntro Statistics by D.M. Diez, C.D. Barr and M. Çetinkaya-Rundel, OpenIntro LaTeX, code, and PDFs are released under a Creative Commons BY-SA 3.0 license.

[10] OpenIntro Statistics by D.M. Diez, C.D. Barr and M. Çetinkaya-Rundel, OpenIntro LaTeX, code, and PDFs are released under a Creative Commons BY-SA 3.0 license.

(a) (1 mark) Is 48% a sample statistic or a population parameter? Explain.

Answer:
Sample statistic, since it is a survey conducted on a sample.

(b) (2 marks) Construct a 95% confidence interval for the proportion of US residents who think marijuana should be made legal.

Answer:
$(0.45, 0.51)$

(c) (1 mark) A critic points out that this 95% confidence interval is only accurate if the statistic follows a normal distribution, or if the normal model is a good approximation. Is this true for these data? Explain.

Answer:
True since $\hat{p}n,\ (1-\hat{p})n$ is large.

(d) (1 mark) A news piece on this survey's findings states, "Majority of Americans think marijuana should be legalized." Based on your confidence interval, is this news piece's statement justified?

Answer:
Not justified, the population parameter could be smaller than 0.50 since the confidence interval contains 0.50.

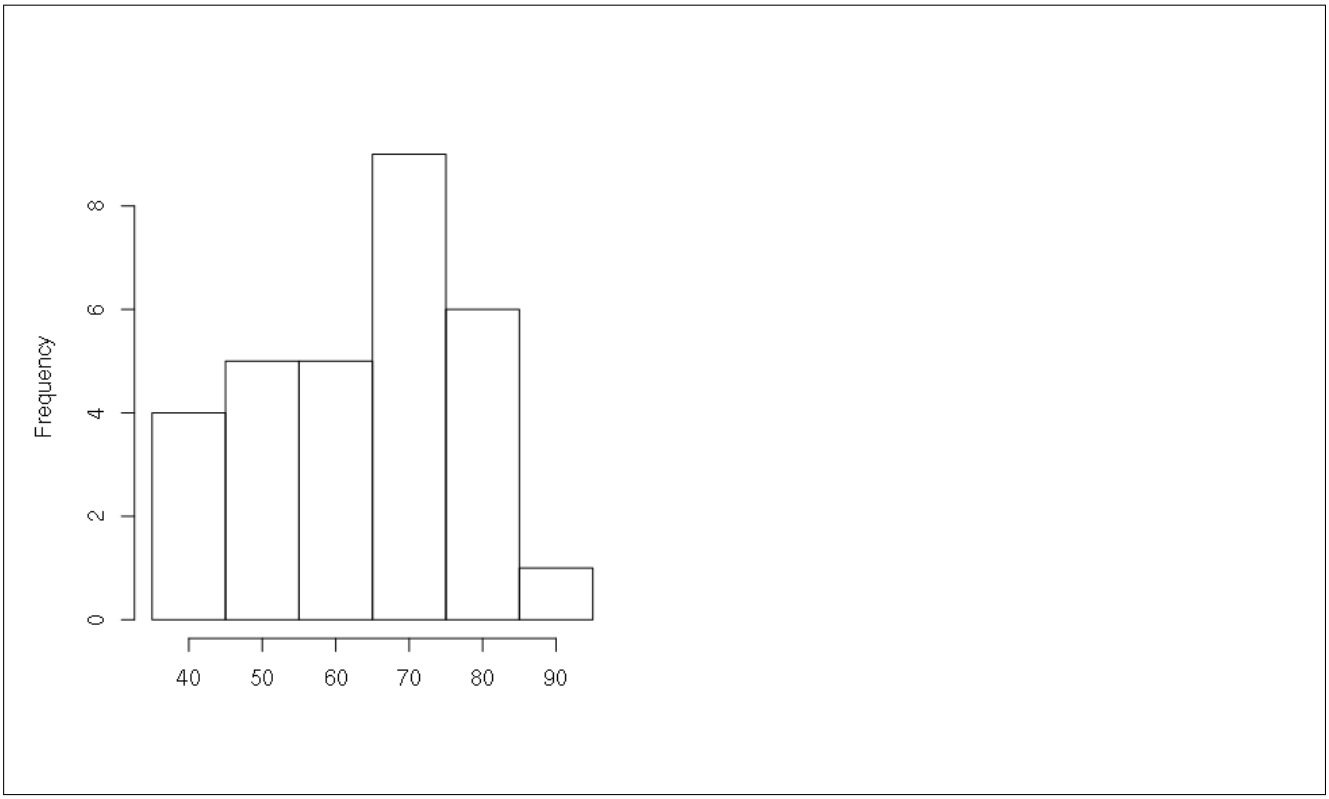
17. Given a sample of a multi-section CEGEP course*(a total of 564 students registered)*:
36 36 42 44 46 48 52 53 55 57 62 64 64 65 66 66 70 71 71 72 73 74 75 76 77 78 80 80 82 92

(a) (2 marks) Compute the sample mean and standard deviation *(use your calculator to save time.)*.

Answer:
$\bar{x} = 64.23,\ s_x = 14.46$

(b) (4 marks) Sketch a histogram representing the above data (use $\sqrt{n}$ to determine the number of classes).

Answer:



(c) (5 marks) Using part a. test for normality using χ^2 hypothesis test.

i. Complete the missing entries in the table.

$a_i \leq X < b_i$	Observed freq.	$z_i \leq Z < z_{i+1}$	$P(z_i \leq Z < z_{i+1})$	Expected freq.
$-\infty < X < 52.08$	7	$-\infty < Z < -0.84$	0.2	6
$52.08 \leq X <$		$-0.84 \leq Z <$	0.2	6
$\leq X <$		$\leq Z <$	0.2	6
$\leq X < 76.38$		$\leq Z < 0.84$	0.2	6
$76.38 \leq X < \infty$	6	$0.84 \leq Z < \infty$	0.2	6

Answer:

$a_i \leq X < b_i$	Observed freq.	$z_i \leq Z < z_{i+1}$	$P(z_i \leq Z < z_{i+1})$	Expected freq.
$-\infty < X < 52.08$	7	$-\infty < Z < -0.84$	0.2	6
$52.08 \leq X < 60.62$	3	$-0.84 \leq Z < -0.25$	0.2	6
$60.62 \leq X < 67.84$	6	$-0.25 \leq Z < 0.25$	0.2	6
$67.85 \leq X < 76.38$	8	$0.25 \leq Z < 0.84$	0.2	6
$76.38 \leq X < \infty$	6	$0.84 \leq Z < \infty$	0.2	6

ii. Calculate the chi-squared statistic.

Answer:
$\chi^2 = 2.33$

iii. Based on the above test statistic, what is the conclusion of the hypothesis test?

Answer:
H_0 : Data follows a normal distribution.
H_a : Data does not follow a normal distribution.

$df = 3,$ p-value > 0.3, Fail to reject H_0, hence data follows a normal distribution.

Formula Sheet

$$\begin{aligned}\bar{x} &= \frac{\sum x}{n} \\ &= \frac{\sum xf}{n} \\ SS(x) &= \sum (x-\bar{x})^2 \\ &= \sum x^2 - n(\bar{x})^2 \\ &= \sum x^2 - \frac{(\sum x)^2}{n} \\ &= \sum x^2 f - \frac{(\sum xf)^2}{n} \\ \sigma^2 &= \frac{SS(x)}{n} \\ s^2 &= \frac{SS(x)}{n-1}\end{aligned}$$

$$\begin{aligned}\mu = E(X) &= \sum xP(x) \\ \sigma^2 = Var(X) &= \sum (x-\mu)^2 P(x) \\ &= \sum x^2 P(x) - \mu^2\end{aligned}$$

$$\begin{aligned}\mu = E(X) &= \int_{-\infty}^{\infty} x f(x)\,dx \\ \sigma^2 = Var(X) &= \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx \\ &= \int_{-\infty}^{\infty} x^2 f(x)\,dx - \mu^2\end{aligned}$$

$$\begin{aligned}P(X=x) &= \binom{n}{x} p^x (1-p)^{n-x} \\ \mu &= np \\ \sigma^2 &= np(1-p)\end{aligned}$$

$$\begin{aligned}P(X=x) &= \frac{\mu^x e^{-\mu}}{x!} \\ \sigma^2 &= \mu\end{aligned}$$

$$\begin{aligned}P(X=x) &= \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \\ \mu &= n\frac{M}{N} \\ \sigma^2 &= \frac{N-n}{N-1} n \frac{M}{N} \frac{N-M}{N}\end{aligned}$$

$$P(a \le X \le b) = \int_a^b \frac{e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}{\sigma\sqrt{2\pi}}\,dx$$

$$\begin{aligned}\mu_{\bar{x}} &= \mu \\ \sigma_{\bar{x}} &= SE = \frac{\sigma_x}{\sqrt{n}}\end{aligned}$$

$$\begin{aligned}\mu_{\bar{d}} &= \mu_d = \mu_{x_1} - \mu_{x_2} \\ \sigma_{\bar{d}} &= SE = \frac{\sigma_d}{\sqrt{n}}\end{aligned}$$

$$\begin{aligned}\mu_{\overline{x_1}-\overline{x_2}} &= \mu_{x_1} - \mu_{x_2} \\ \sigma_{\overline{x_1}-\overline{x_2}} &= SE = \sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}} \\ df &= \frac{\left(\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}\right)^2}{\frac{{s_1}^4}{{n_1}^2(n_1-1)} + \frac{{s_2}^4}{{n_2}^2(n_2-1)}}\end{aligned}$$

$$\begin{aligned}\mu_{\hat{p}} &= p \\ \sigma_{\hat{p}} &= SE = \sqrt{\frac{p(1-p)}{n}}\end{aligned}$$

$$\begin{aligned}\mu_{\hat{p}_1-\hat{p}_2} &= p_1 - p_2 \\ \sigma_{\hat{p}_1-\hat{p}_2} &= SE \\ &= \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \\ \hat{p} &= \frac{x_1+x_2}{n_1+n_2} \\ \sigma_{\hat{p}_1-\hat{p}_2} &= SE \approx \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\end{aligned}$$

$$\chi^2 = \sum \frac{(f_i - e_i)^2}{e_i}$$

$$\begin{aligned}SS(xy) &= \sum (x-\bar{x})(y-\bar{y}) \\ &= \sum xy - \frac{(\sum x)(\sum y)}{n} \\ SSE &= \sum (y-\hat{y})^2 \\ b_1 &= \frac{SS(xy)}{SS(x)} \\ &= \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2} \\ b_0 &= \frac{\sum y - (b_1 \cdot \sum x)}{n} \\ &= \bar{y} - (b_1 \cdot \bar{x}) \\ r &= \frac{SS(xy)}{\sqrt{SS(x)SS(y)}}\end{aligned}$$