Stat 302 Winter 2016 Assignment 1 KEY

(This due date overrides the syllabus due date of Jan 26).

There are six multi-part questions and a set of questions based on 11 pages of reading.
Please see the Stats Workshop for help, or see me in office hours (Tue 1-2, Thur 3:30-4:30), or e-mail me at jackd@sfu.ca

Explanations and relevant computer output should be included, but plots do not need to be.
Necessary R code is included for every question, after the reading questions.
There are also four practice problems, mostly surrounding review material.

Total /65
Q1 /6
Q2 /7
Q3 /8
Q4 /4
Q5 /10
Q6 /12
Reading /18

Name _____

Student Number _____

1) Consider the different types of data. Specifically numeric and categorical. [6 points, 2 per question.]

1a) Name two discrete numeric variables, and give a possible value for each.

A discrete numeric variable is one we measure in countable numbers. Possible answers include:

Number of cars in a parking lots, e.g. 17 cars.
Number of new HIV infections in a month, e.g. 3 new infections.

1b) Name two continuous numeric variables, and give a possible value for each.

A continuous numeric variable is one we measure in continuous, 'gapless' numbers. Possible answers include:

Kilometers driven by a car, e.g. 3223.4 km
Dry weight of a plant, e.g. 423 grams
Lung capacity, e.g. 3.8 Litres of air

Note: Where there are millions of a countable thing, it's more appropriate to treat that

variable as continuous. e.g. Grains of sand, molecules of water, people in a large country.

1c) Name two categorical variables, and give a possible set of values for each.

Categorical variables are those that can't be described as an amount or count.

Possible answers include:
Stage of cancer, e.g. Stage 1, Stage 2, Stage 3a, Stage 3b, Stage 4
Movie genre, e.g. Action, Comedy

There are two kinds of categorical variables, nominal and ordinal. However, for the work we are doing in this class, there's little practical difference between them. With ordinal variables, there is a natural order to the categories (e.g. low, medium, high, very high). With nominal variables there is not.

---

2) Consider this dataset from the a collection of measurements of iris setosa flowers. (A1setosa.csv)
[7 points total. Do not include plots.]

2a) Describe the five variables of the dataset. Are they numeric or categorical? Are they discrete or continuous? [2 pts]

NOTE: The R code and output is just shown here for you. It is NOT required for full marks.

```
> head(Q2)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
```

The five variables are:
Sepal.Length, a continuous numeric variable
Sepal.Width , a continuous numeric variable
Petal.Length, a continuous numeric variable
Petal.Width , a continuous numeric variable

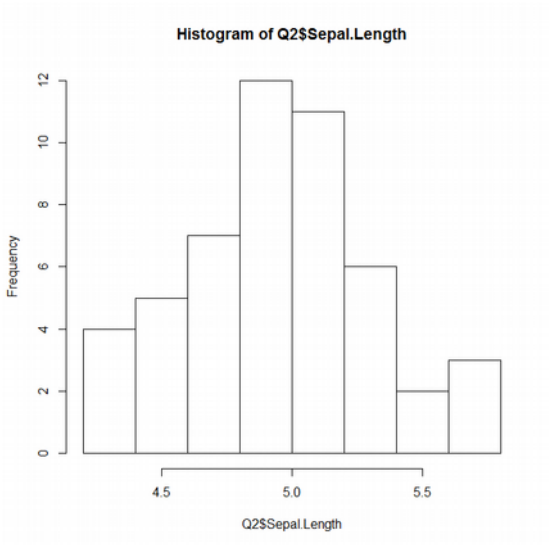
Species, a categorical variable

2b) Find the mean and median of the Sepal length. [1 pt]

The mean is 5.006, and the median is 5.

2c) Create a histogram of the distribution of the first variable (Sepal length). [4 pts, do not include histogram]
Use the histogram and the mean and median to describe this distribution.
Is it unimodal (one peak), or multimodal (multiple peaks)?

Is it left/negatively skewed (extreme values on left/lower), or right/positively skewed (extreme values on right/higher)?
Explain how you know this from the mean and the median.



The histogram shows a unimodal distribution.

The mean and median are very close, so this distribution is symmetrical.
(Also acceptable is slightly right skewed)

The mean is sensitive to extreme values and the median is not. If there were more extreme values on either side, the mean would be 'pulled' more to that side. In short, if the mean was larger than the median, we would say this distribution is right/positively skewed.

3) Consider this dataset of the spending habits of 100 families. Specifically, their food spending per month and their clothing spending per year. (A1spend.csv) [8 points, 2 per question. Do not include plots.]

3a) Find the Pearson correlation coefficient. Test the hypothesis that the parameter rho = 0 at the 0.01 level?

```
> cor.test(A1spend$food, A1spend$clothing)

        Pearson's product-moment correlation

data:  A1spend$food and A1spend$clothing
t = 9.3692, df = 98, p-value = 2.887e-15
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5676046 0.7786727
sample estimates:
      cor
0.6873874
```

The correlation is 0.687

Against the null hypothesis that the coefficient rho is 0,
our p-value is 2.887 x 10^{-15}, which is less than 0.01.

Therefore, there is sufficient evidence to reject this hypothesis. There IS a correlation.

3b) How much of the variation in clothing spending can be explained by food spending?

Proportion of variance explained is r-squared: 0.4719

47.19% of the variation in y is explained by x.

3c) Verify the statistical significance of the correlation with a two-sided t-test. Report the t-score, degrees of freedom and p-value.

The t-score we obtained from cor.test was 9.37, and the degrees of freedom was 98 (there are 100 observations, and we are estimating 2 parameters).

We can compare this to the critical t-score. If the obtained t-score is larger than the

critical value, we can reject the null. The significance level is 0.01, so that leaves 0.005 on both the negative and positive sides, so the quantile we use is .995.

qt(.995, df=98) gives us a critical value of 2.62

9.37 > 2.62,

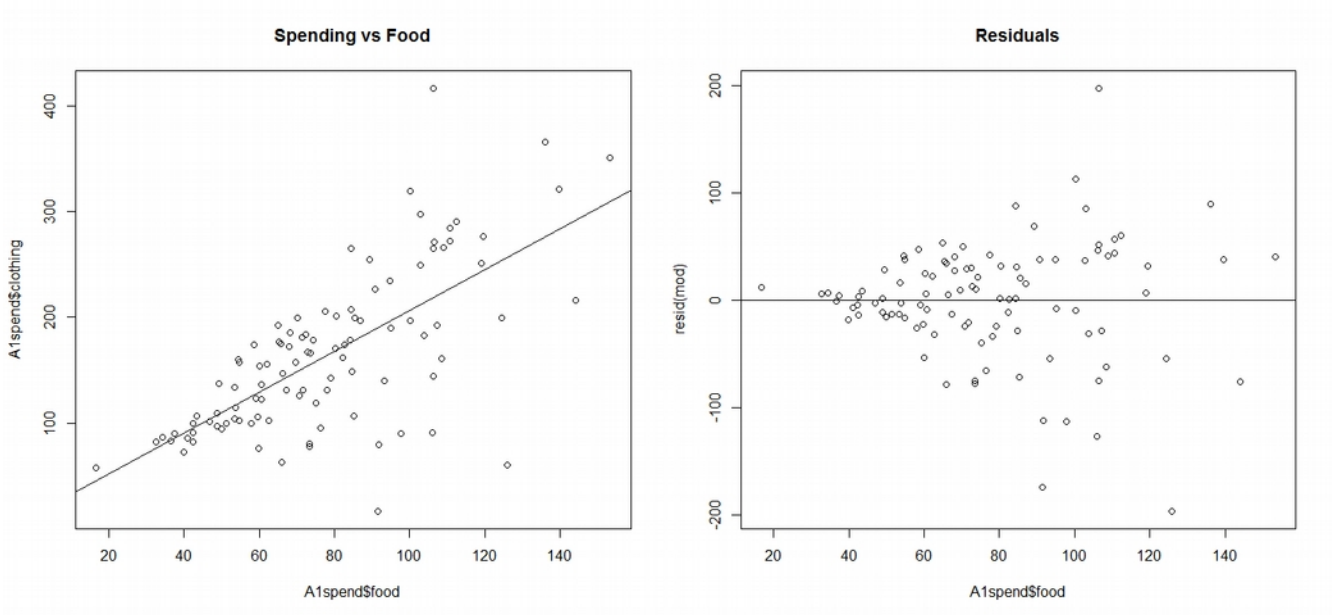
so we reject the null. There is indeed a correlation.

Note solving this with the p-value in 3a, and solving this with a critical will give the same answer. This is always true.

3d) Produce a scatterplot to better see the relation. Is there any trend in the plot that could be a problem?

The amount of variation around the regression line increases as x increases. In other words, the variance is unequal. In fancy words, we have heretoscedasticity.

Regression assumes that the variance is equal. In this case, however, the variance increases with the mean, so the higher values affect the regression line more than the lower values.



Notice that the residuals show an obvious fanning-out pattern.

4) The correlation between two abstract variables. X and Y, is r = +0.76
[4 points, 1 per question.]

4a) What proportion of the variance in Y is explained by X?

Variance is the regression coefficient squared. $0.76^2 = 0.5776$
57.76% of the variance in Y is explained by X.

4b) If X and Y were switched, what would the correlation become?

Correlations just measure the strength and relationship between two variables. It doesn't matter which variable is the X and which is the Y.

So the correlation would still be +0.76.

4c) If Y were multiplied by 10, what would the correlation become?

Correlation is invariant to scaling. If X or Y is multiplied by any positive number

So the correlation would STILL be +0.76.

4d) If Y were multiplied by -10, what would the correlation become?

When either X or Y is negated (but not both), the direction of the correlation is also negated. The strength of the correlation is preserved.

Therefore the correlation is -0.76

---

5) Consider the gross domestic product (GDP, a measure of the mean income of the citizens in a country), and life

expectancy (at birth) in the countries of the world in 2003. (A1countries).
[10 points, 2 per question. Do not include plots.]

5a) Find the Pearson correlation coefficient. Is it significant at the 0.01 level?

The correlation is +0.633.
The p-value against no correlation is less than 0.01, so yes it is significant at the 0.01 level.

5b) How strong is this correlation?
Does it fully describe the strength of the relationship between GDP and life expectancy?

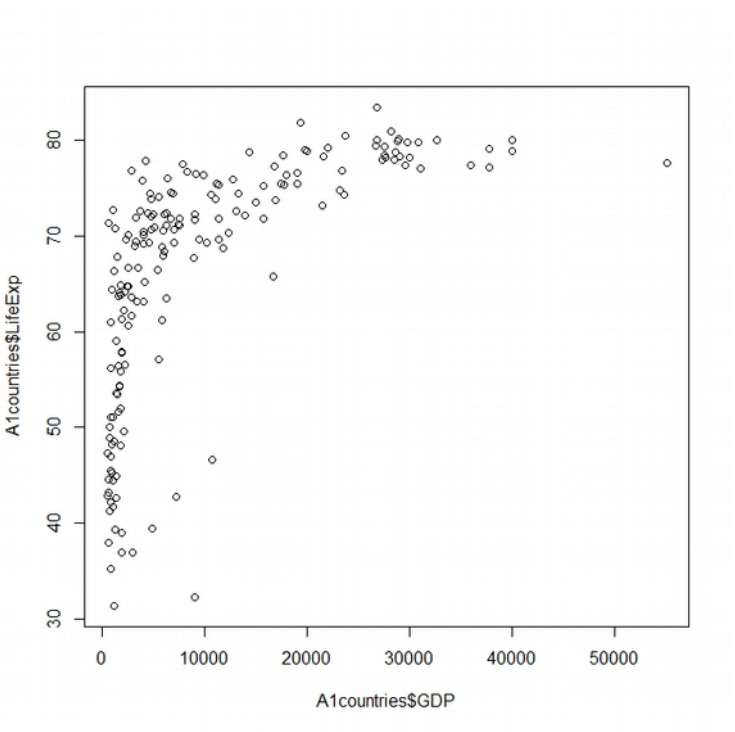
This correlation is moderate to very strong.
It does not fully describe the strength of this relationship though. Possible reasons why include:
** The non-linear nature of the relationship.*
** The natural limit to life expectancy.*
- Covariates like climate, inequality, and war status.
- That a single value for average lifetime does not cover specific aspects like child mortality.

Any one of these reasons is enough. The ones with stars were the intended answers, but I don't want to punish creative problem solving.

5c) Produce a scatterplot of Life expectancy(y) over GDP(x) and describe it to back up your answer in 5b.

There is a curved relationship between GDP and Life Expectancy. Life Expectancy climbs very quickly per dollar/year for low incomes, and then it levels out. It also seems like no country can push past 80 years. This scatterplot lends additional evidence to the following explanations.

** The non-linear nature of the relationship.*
** The natural limit to life expectancy.*



5d) What else might better capture the strength of this relationship? Why?

Something that handles the curved relationship. Possible answers include

** Using the transformed data Log10GDP.*
** Using a Spearman correlation instead of a Pearson correlation.*
- Using a regression fit that includes a curve, such as a polynomial fit.

5e) Confirm you answer in 5d by finding a different correlation between GDP and life expectancy, and by find the Pearson correlation between Log10GDP and life expectancy.

The Pearson correlation of Life Expectancy and Log10GDP is +0.770, which is a substantial improvement over the correlation +0.633 in 5a.

*The Spearman correlation of Life Expectancy against BOTH GDP and against Log10GDP is 0.840. Spearman correlation uses the ranks of the values, and those are not affected by transformations.**

** There are functions we could use, like square, that don't preserve rank, but they should not be used in cases where they change the rank.*

6) Consider the dataset of standings in the National Hockey for 2011-12 regular season, in "A1nhl.csv"
Create a linear regression of wins (W) in response to goals scored against (GA).
To clarify, wins is the 'y' variable, and goals against is the 'x' variable.
[12 points, 2 per question. Do not include plots.]

6a) Determine the least-squares estimates of the intercept and slope parameters.

```
> mod = lm(W ~ GA, data=A1nhl)
> summary(mod)

Call:
lm(formula = W ~ GA, data = A1nhl)

Residuals:
    Min      1Q  Median      3Q     Max
-9.6963 -3.4041 -0.3588  4.1214  9.4601

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 78.82954    8.41628   9.366    4e-10 ***
GA          -0.16873    0.03731  -4.522 0.000102 ***
```

The slope estimate is 78.83
The intercept estimate is -0.169

I didn't ask for an interpretation, but it would look like this:
- A team that allows 0 goals can expect to win 78.83 games (out of 82)
- For every goal that a team allows, they win 0.169 fewer games.

6b) Test the null hypothesis of zero slope. Interpret the result.

The p-value associated with slope is very small, 0.0001. This is very strong evidence against the null, which is that the slope is zero.

We reject this null, meaning that the number of wins changes as goals-against changes.

6c) Find a 95% confidence interval of the slope of wins (W) over goals against (GA).

Explain your choice for degrees of freedom.

Explain your choice for quantile (hint: 95% refers to the MIDDLE 95% of the distribution)

Interpret the confidence interval.

The formula for the confidence interval is

(estimate) +/- (critical value) X (standard error)

From the summary information, we have the estimate (-0.169) and standard error (0.037) of the slope. But we need the critical value.

Because there are 30 observations (hockey teams) and 2 parameters being estimated. That leaves 28 degrees of freedom for residuals.

As for quantile, we are looking for a 95% confidence interval, which leaves 5% of the distribution on the outside. That's 2.5% on either side. We should choose the quantile that cuts off the top 2.5%. That's 0.975.

Alternatively, we can look at the cutoff at the bottom 2.5%. That's 0.025.

So we use the code

qt(.975, df=28)

to find the critical value, 2.048, for this confidence interval.

Therefore the confidence interval is
(estimate) +/- (critical value) X (standard error)
-0.169 +/- (2.048) X (0.037)

-0.245 to -0.093

6d) The value -0.2 is inside the 95% confidence interval for the slope in 6c. Using only this information, can you tell whether -0.2 is inside the 99% confidence interval or not? Explain.

The 95% confidence interval is -0.245 to -0.093, meaning that any value between these two numbers is inside the confidence interval.

The 99% confidence interval has the same middle, the estimate. However, it has a larger critical value, which implies that the +/- portion will be larger. This means the bottom will be lower and the top will be higher.

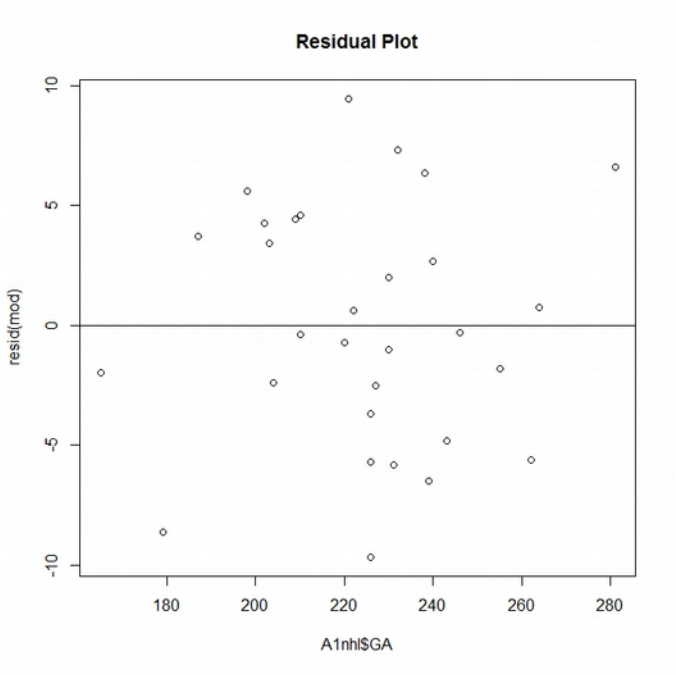
This is a general rule for confidence intervals. An interval at a higher confidence level (99%) includes every value that would be in a lower confidence level's interval (95%).

Since the value -0.2 is inside the 95% confidence interval, it must also be inside the 99% confidence interval. (and the 99.9% confidence interval, and so on).

6e) Plot the residuals of this regression against GA. Are there any potential problems visible? Explain.

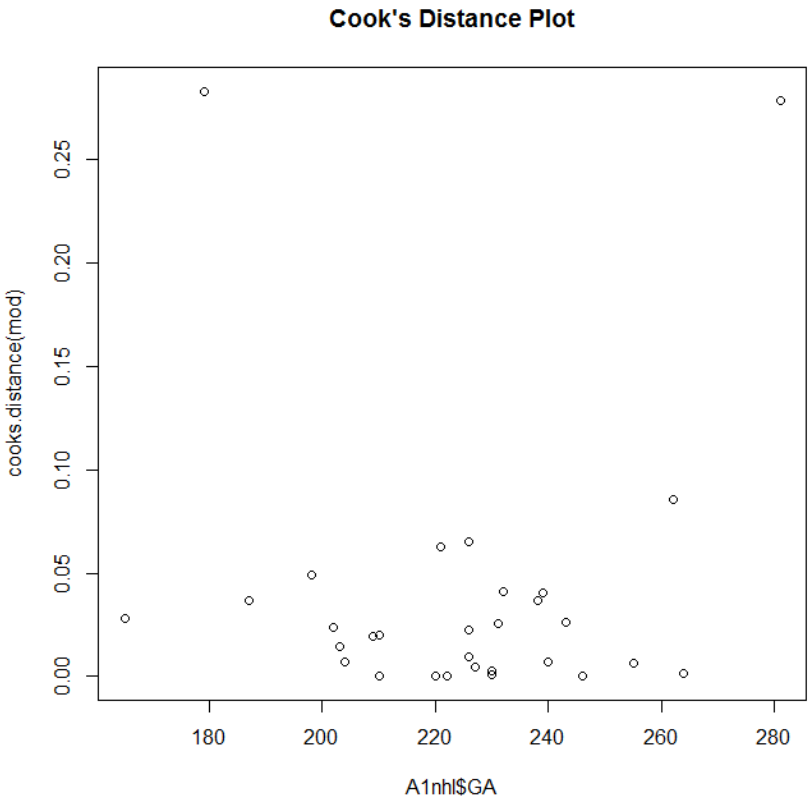
There are no obvious patterns in the residuals. Specifically, we are looking for curves, fan-shapes, and extremely positive or negative values. None of these appear, so the residuals don't reveal any problems.

Patterns in residuals should be obvious to be a potential problem, like in question 3d.



6f) Plot the Cook's Distance for this regression. Are there any potential problems visible? Explain.

Cook's Distance is used to measure the influence of specific cases. We are not looking for trends here, we are looking for values that are extremely high.



There are two cases that could be exerting a very strong influence on the regression line. This is a potential problem.

In order to check if this were a real problem, we would have to remove those two points and see the regression line changes much. However, we haven't covered how to do that yet.

Reading questions)
These questions pertain to the first 3 sections of "For Objective Causal Inference, Design Trumps Analysis", by Donald B. Rubin.
The Annals of Applied Statistics
2008, Vol. 2, No. 3, 808–840 (Stop at Page 818)

The answer to R1 appears first in the text, and so on. [18 points, 3 per question.]

R1) Randomized experiments make causal inference valid because we know the 'scores' "are known from the design of the experiment". What are these 'scores'? (Name only)

The 'scores' in question are "Propensity Scores". These are the probabilities that each subject will take the treatment in question. (i.e. be in the treatment group, see R2 for times when this isn't the case)

R2) Rubin is stating that randomized experiments are 'the gold standard' for causal inference. Does that mean causality can be inferred from all randomized experiments, or are some "poorly suited"? If so, give an example.

Experiments in which there is a high rate of non-compliance. That is, experiments where assigning a treatment does not necessary translate to taking the treatment. In the paper, an example is given where 80% of subjects refuse to take the drugs they have been assigned to take.

Quoting this example or one similar to it is acceptable for full marks.

R3) What are the two design steps that are "absolutely essential" for objective inferences?

Understanding your ***potential outcomes*** *and choosing an* ***assignment mechanism****.*

R4) In Section 2.1, how is a treatment defined?

'A treatment is an action or intervention that can be initiated or withheld from that unit at some per-specified time t.'

A verbatim quote of this is acceptable.

R5) What are covariates, "in contrast to" outcome/response variables?

"In contrast to outcome variables, covariates are variables, X, that for each unit take the same value no matter which treatment is applied to the unit"

In other words, covariates are variables that don't change whether the treatment or

the control is applied. In human trials, most commonly used covariates include demographic information like age, sex, race, and relevant medical history.

R6) What "self optimizing" behaviour happens when treatments are not assigned randomly by the experimenter?

"Self optimizing" behaviour is behaviour in which subjects take whatever treatment (or control) that they think will lead the best outcome for them. Most commonly, this means that people will seek out a treatment for an illness to fix that particular illness.

This is good for people in general, but it makes observational studies difficult because it is hard to separate the effect of a treatment from the effect of whatever led to a person seeking a treatment.

For example, blood pressure medication is typically only given to people that have high blood pressure to begin with. So if we blindly compare people on blood pressure medication to those that are not, we would find that people on the medication actually have WORSE blood pressure.

Practice Problem 1)

When nutrition guides were being made, a large group of healthy people's diets were surveyed. The intake of many nutrients was found to be normally distributed. The amounts that were two standard deviations below the mean were set as the recommended daily minimums.

a) What proportion of surveyed healthy people met or exceeded this minimum for calcium?
Anyone above two standard deviations below meets or exceeds this. The area above a z-score of -2 is 97.72% using the table, or 97.5% using the 95% rule. Both are acceptable answers.

b) Consider three surveyed healthy people. What is the probability that all three are getting enough calcium?
These three people are independent, so we use the multiplication rule and our value from part a.
(.9772) X (.9772) X (.9772) = .9331 or 93.31%
Or… (.975) X (.975) X (.975) = .9269 or 92.69%

c) Adult men in the survey took in 14 mg/day of zinc, with a standard deviation of 1.5 mg/day,

what is the daily recommended minimum of zinc for adult men?

Two standard deviations below the mean is the minimum. Using: $x = \mu + z\sigma$
X = 14 + (-2)(1.5) = 11

d) All adult men, healthy or otherwise, take a mean of 12mg/day of zinc and a standard deviation of 2 mg/day. What proportion of them get their daily recommended amount of zinc?

X = 11 from part c. mean = 12, std.dev. = 2. Using: $z = \frac{x-\mu}{\sigma}$
We find z = (11 – 12)/2 = -0.5. By the table we find this to correspond to 19.15% + 50% = 69.15% of people.

Practice Problem 2)

Which are statistics and which are parameters? Identify the population of interest for every question and identify the sample when applicable. (Sources are for interest only, everything you need is in the sentence.) ***The population in several cases is open to interpretation. Anything close that is a complete collection like all of (People in city) should do.***

a) "The federal agency said the census data showed that ***85.2 PARAMETER*** per cent of Canada's population was under the age of 65. "(Source: Canada.com, "Canada still youngest amongst G8" May 29,2012) ***Population: Canadians.***

b) 25 courses were ***randomly selected*** from all those available in the summer, and an average of ***2.4 STATISTIC*** required textbooks was found. (modified from your textbook). ***Sample: The 25 selected courses. Population: All courses available in summer.***

c) A research organization conducted a ***survey*** and found that ***45% STATISTIC*** of respondents felt they were better off economically than before. ***Sample: Those in the survey. Population: All people in the area.***

d) "Statistics Canada says the caseload in youth courts fell ***seven per cent PARAMETER*** in 2010-2011, a second straight annual drop. The agency says youth courts handled about ***52,900 PARAMETER*** cases last year." (source: Toronto Star: Youth Courts Caseload Falling: Statscan." May 28, 2012) ***Population: All Canadian Youth Courts. (Or: All of Canada)***

e) "A ***survey*** featured in [a recent] report reveals ***82 per cent STATISTIC*** of mothers cite safety concerns as reasons why they restrict outdoor play [of their children]." (Source: The Globe and Mail: "Canadian kids get failing grade in physical activity: report." May 29, 2012) ***Sample: Surveyed Mothers. Population: All Canadian Mothers***

e) “A survey featured in [a recent] report reveals 82 per cent of mothers cite safety concerns as reasons why they restrict outdoor play
[of their children].”
(Source: The Globe and Mail: “Canadian kids get failing grade in physical activity: report.” May 29, 2012)

Practice Problem 3)
Hummus, a food product, comes in packages that claim there is 227 grams of hummus inside the container. In reality not every container has the exactly the same amount of hummus. The weight of the product has a normal distribution with mean 232 grams, and standard deviation of 4 grams.

a) What is the probability that a randomly selected hummus package is underweight (contains less than the advertised amount)?
Z = (227 – 232) / 4 = -5 / 4 = -1.25. → Table → .1038 or 10.38%
We want the underweight amount, which is beyond the mean for all of these. (as in farther away from the mean than our cutoff, and in one direction. Seriously, draw it!!)

b) What is the probability that an average of 4 packages is less than 227 grams?
Z = (227 – 232) / (4 / sqrt(4)) = -5/2 = -2.5 → Table → .0062 or 0.62%

c) What is the probability that an average of 25 packages is less than 227 grams?
Z = (227 – 232) / (4 / sqrt(25)) = -5/0.8 = -6.25 → Table → Less than .00003 or 0.003%.
The table only goes to 4, where the area beyond is .003. Since 6.25 is beyond that the area beyond has to be smaller, we just don’t know how much smaller.

(for your interest, the area beyond is approx. 2.05×10^{-10}, or .000000000205.)

d) Give an upper bound to the chance that an average of 400 packages is less than 227 grams?
Z = (227 – 232) / (4 / sqrt(400)) = -5/0.2 = -25 → Table → Less than .00003 or 0.003%.
The upper bound is .00003.

(for your interest, the area beyond is approx. 3.06×10^{-138}, which as a decimal starts with 137 zeroes. At that level, even the computer’s answers become suspect.)

e) If you took the average of a larger and larger sample, what would expect to happen to the sample mean of the weights?
The sample mean of the weights should get closer and closer to the true mean of 232.
(for your interest, this is called the Law of Large Numbers)

Practice Problem 4)

1) In each situation, would a large value of significance level alpha (more than .05) or a small level of alpha be more appropriate (.05 or less)? **Only the small/large part is required. Justification is here just so you know why.**

   A) A trial for which the penalty is death (null hypothesis is innocence).
   A small alpha. Rejecting the null would mean making the decision that someone is guilty. Since making that decision wrongfully carries such a heavy penalty, we want to minimize the chance of that happening.

   B) Determining if you should take extra vitamin C to ward off a cold this morning (null hypothesis is you're not in danger of getting a cold).
   Large alpha. If we falsely reject the null that means assuming you could get a cold and eating a vitamin pill for no reason. There's almost no penalty for making this mistake, so we can accept a large chance of making it.

   C) A patient might have a disease that would require cutting off a leg. (Null hypothesis is that the leg doesn't need to be cut off).
   Small alpha. This is like the death penalty one. If we're going to cut someone's leg off, we should be very sure that it has to be done. We're only willing to accept a small chance of cutting the leg off without need. That acceptable chance level is alpha.

   D) You have to decide whether to look before crossing the street. (Null hypothesis is that it's safe to cross without looking).
   Large alpha. Like the vitamin one, we should be willing to accept a large chance of assuming it's dangerous to cross when it's safe. All it would cost is to look both ways.

   E) You have an exam in the morning and you're deciding whether to set the alarm or not. (Null hypothesis is that you'll wake up in time without an alarm)
   Large alpha. See B and D.