Statistical Analysis in R

1 Purpose

The purpose of this activity is to provide you with an understanding of statistical inference and to both develop and apply that knowledge to the use of the R statistical programming environment.

2 Overview

This lab can be completed during class time and at-home. You should allocate time to complete the relevant portions of the lab in line with the scheduled topics for each week. The final quantitative problem set in LT relates to, roughly, the first half and second half of the material covered in this lab (this document containing the first half) and builds on our previous work with data summaries and visualizations in R.

For this activity, we focus on the idea of "statistical significance." Statistical significance is a concept related to *statistical* hypothesis testing. If you recall from earlier in the course, we discussed two different "flavours" of hypothesis testing — one associated with Fisher and one associated with Neyman and Pearson. We will see how both kinds of hypothesis testing manifest and how current statistical practice is a blend of these two perspectives. That practice most closely approximates Fisher's ideas (the calculation of a *p*-value) but differs in other ways (the estimation of a "confidence interval").

3 Your Task

Using R as instructed, complete the following activities.

3.1 Statistical Significance

1. In your own words, attempt to explain what it means for a research conclusion to be *significant.* You may find it useful to distinguish *substantive significance* from *statistical significance.*

2. To develop an initial understanding of the idea of statistical hypothesis testing (and statistical significance), we will think about the idea of "outliers." Outliers are unusual values in a variable (or set of variables). For example, consider the following variable:

   ```
   v <- c(1,1,1,1,2,2,2,2,2,3,3,3,3,10)
   ```

   It should be clear that the value 10 here is an outlier. If we draw a boxplot of this variable, it becomes even more clear:

   ```
   library("ggplot2")
   ggplot(, aes(x = 1, y = v)) + geom_boxplot() + scale_y_continuous(limits = c(0,12))
   ```

3. If we expected all the values of `v` to be relatively similar, we could point out that the value 10 is more than 3 standard deviations away from the mean of the variable: `(10-mean(v))/sd(v)`. This idea that a value is an outlier relative to some assumed central tendency is the basic logic of statistical hypothesis testing.

4. In statistical hypothesis testing, however, we are instead interested to know how unusual a *statistic* is relative to our baseline expectation of what the value of the statistic should be. Commonly we are interested in statistics that describe a sample of cases, such as:

Example	**Statistic**	**Unit of Analysis**	**Concepts/Variables**
Mean years of education for working-age adults			
Proportion of university graduates			
The male–female mean difference in annual income			
The correlations between education, sex, and income			
The causal effect of a university degree on annual income			

For each of the above examples, what are the *statistic*, *unit of analysis*, and *variables/concepts* involved in the analysis?

5. Each of the above statistics can be calculated on a sample of cases in order to make descriptive (or causal) inferences about the population of cases as a whole. But, given what we know about sampling (from MT), these inferences carry with them a degree of uncertainty that is a function of the size of the sample of cases selected from that population.

   (a) A sample statistic provides an estimate of the population *parameter*. As sample size increases, what happens to the amount of uncertainty we have about the precise value of the population parameter?

   (b) Is this change in amount of uncertainty linear?

6. Carrying forward this basic logic of *sampling variability* (or *sampling error*, *sampling uncertainty*), we may want to characterize a sample statistic as *statistically significant*. That is, we may want to say that a sample estimate appears unusual in light of some prior expectation about the population parameter value ought to or is thought to be. To figure that out, we have to describe — in R.A. Fisher's language — a "null hypothesis." This is a value that we think the parameter might have in the population.[1] We are therefore trying to see whether the statistic in our sample data differs from this (hypothetical) null population parameter.

   Based upon what you know or your own intuition, how might we determine whether a statistic is unusual? What process would we use to assert that?

[1] Often, the null hypothesis is set to something that would be theoretically uninteresting: a proportion is 0.5, a difference between two means is 0, a correlation between two variables is 0, the mean of a variable is 0, etc. We can pick any value, but these theoretically uninteresting values are conventional.

7. At its core, statistical significance testing attempts to wrestle with the fact that an observed sample statistic that *appears* unusual may still be possible (or even highly probable because of sampling variability) to see from a population whose parameter is equal to the null hypothesis value. In other words, an apparently large difference between our sample statistic and the null hypothesis parameter value might be erroneous. Statistical hypothesis testing therefore asserts that we only consider a *sample* statistic to indicate that its popular parameter differs from the null hypothesis value — and is thus "statistically significant" — when it is quite far from the null hypothesis value. But how far exactly is "quite far"?

   To decide that, we return to the idea of repeated sampling that we examined in MT when we discussed notions of *representativeness*. We consider a statistic "statistically significant" when it differs more from our null expectation of the population parameter (i.e., the null hypothesis) than the vast majority of the variation in sample statistics we could observe across repeated samples from a population where that null hypothesis was true.

   In your own words, try restating that definition of statistical significance.

8. For a mean (like the one we looked at earlier), this would be stated as a null expectation that the mean level of education in a country is 12 years. If we collect a sample of data and find that the mean is different from 12, we would consider that mean to be statistically significantly different from our null expectation of 12 years if the sample mean was further from 12 (i.e., much larger or much smaller) than the variation in sample means we could expect to observe from same-sized samples drawn from a population where the true mean was 12.[2] Revisit the exercises from Lab 2 if you remain unclear about the underlying dynamics of sampling.

9. Now let's put this into practice using R. First, create a vector of education values, `x`, which contains a population of education levels (think of these numbers as years of education for individuals from a population of 1,000,000 people):

```
set.seed(1)      # this makes sure we all get the same answer
x <- sample(0:20, 1e7, TRUE, c(1,2,3,4,5,6,7,8,9,10,11,12,13,12,11,10,9,8,7,6,5))
# ^^ don't worry about the details of what's happening in this line of code
```

10. Use the `length()` function to verify how many people there are in our population.

11. Use `mean()` to calculate the "true" population mean of this population. How many years of education, on average, does this population have?

12. Now, we will draw a small sample from this population using the `sample()` function. (Note: we used this above to generate some fake population data; now we will use in a different way.). Start by drawing a small sample of just ten observations:

```
s1 <- sample(x, 5, FALSE)
```

13. Use **ggplot2** to draw a histogram of these data:

[2] If we were talking about a different statistic, the logic would be the same. For a proportion, we might have a null expectation that the proportion is 0.5. An observed proportion in our data would be considered statistically significantly different from 0.5 if the proportion were larger than the variation in estimated proportions we would see across multiple same-sized samples from a population where the parameter was 0.5.

```
library("ggplot2")
ggplot(, aes(x = s1)) + geom_histogram(bins = 21)
```

14. What is the sample mean of this sample? How variable are the data as shown in the histogram?

15. Now, repeat this but drawing a larger sample size of size 100 (call this vector `s2`). How variable are these data? What is the sample means of `s2`?

16. Recall that the *standard error* is meant to capture the idea that if we repeated our sampling process and calculated our statistic of interest (in this case, the mean) on each sample, the standard deviation of those estimates around the true mean would be equal to our standard error. To get a better grasp on this idea, we are going to simulate the process of drawing random samples from our population and then compare the standard deviation of our estimates from each sample to the standard errors we calculate above.

17. To do this, we are going to use the `replicate()` function. This function allows us to repeat a calculate multiple times and return the results in a convenient form. To understand how it works, try generating a single sum of two random numbers: `rnorm(1) + rnorm(1)`. Then, use `replicate()` to do this five times:

    ```
    replicate(5, rnorm (1) + rnorm (1))
    ```

    Note how the result is simply a vector.

18. Now, we want to apply this function to the calculation of the sample mean, as above. To do so, we simply write:

    ```
    # five samples of size n = 5
    replicate(5, mean(sample(x, 5, FALSE)))

    # 1,000 samples of size n = 5
    dist5 <- replicate(1000, mean(sample(x, 5, FALSE)))

    # 100 samples of size n = 100
    dist100 <- replicate(100, mean(sample(x, 100, FALSE)))
    ```

    This vector of sample means is often called the "sampling distribution" of the mean. This term refers to the distribution of a given statistic across repeated samples of the same size from a population.[3] Note that these operations may take some time. When they are done, examine the results:

    - What does the histogram look like: `ggplot(, aes(x = dist5)) + geom_histogram(bins = 21)` ?
    - Are the sample means "unbiased" (meaning the mean of the sample means is close to the population mean): `mean(dist5) mean(dist100)`?
    - How does the standard deviation of the sample means correspond to the standard errors you calculated above: `sd(dist5)` and `sd(dist100)` ?

[3] So hear we focus on the sampling distribution of the mean, but we could also create a sampling distribution for the maximum of each sample, for the count of observations with 10 years of education, the proportion with more than 12 years, etc. The process works the same for any sample statistic, we just focus here on the mean.

19. The margin of error is a form of "interval estimation" in which we express our uncertainty about the value of a parameter by stating the range of values that the parameter is expected to be in based upon our sample estimate. The interval that is equivalent to estimate +/- 2 times the standard error is also called a 95% confidence interval (for reasons we will return to below). You can compare the confidence interval from your data (mean +/- 2 SE) to the distribution of estimated sample means (`quantile(dist100, c(0.025, 0.975))`) to see how closely that interval compares to the interval of estimated values from repeated sampling.

20. To ensure you are comfortable with these ideas, try repeating all of the above but for a different kind of variable. Rather than using an ordinal or interval measure (as above), try using a binary variable. You can create one using `rbinom()`:

```
y <- rbinom(1000000, 1, prob = .5)
# ^^ the ‘prob‘ argument controls the ratio of 1s and 0s
```

21. Now, let's focus in on our samples of size $n = 100$. What is the standard deviation of the sampling distribution `dist100` (i.e., what is the standard error of the sample mean)?

22. In statistical hypothesis testing, we will say that an estimate is statistically significantly different from the null hypothesis value when the sample statistic is more than a certain number of standard errors away from the null hypothesis value. But how far is far enough to be considered *significant*? Take a look at how many observations in our `dist100` vector are more than 3 standard errors, more than 2 standard errors, and more than 1 standard error above or below the true mean of `x`:

```
dist100[dist100 > (mean(x) + 3*sd(dist100))]
dist100[dist100 < (mean(x) - 3*sd(dist100))]
dist100[dist100 > (mean(x) + 2*sd(dist100))]
dist100[dist100 < (mean(x) - 2*sd(dist100))]
dist100[dist100 > (mean(x) + 1*sd(dist100))]
dist100[dist100 < (mean(x) - 1*sd(dist100))]
```

How many of the sample means are within 1, 2, and 3 standard errors of the mean?

23. Now here is where a short moment of consideration is required before we proceed. In statistical hypothesis testing, we are declaring a null hypothesis value as a *known* population and seeing how likely different sample estimates are when we draw repeated random samples from that population. We declare a sample statistic "statistically significant" when the sample statistic is unusually far from the null hypothesis value in that *sampling distribution* because it would seem to suggest that the sample is drawn from a population with a different population parameter value than one with a population parameter equal to the null hypothesis value. In other words, the sample statistic is so unusual for the population from which we — under the null hypothesis — believe that the data are drawn from, that we decide that the data are actually drawn from a population with a different population mean. That's the essence of statistical hypothesis testing. Restate these ideas in your own words:

24. Hopefully, you've caught on that there's a leap being made here: in practice we don't generally have population data in front of us (indeed, that's the whole point of why we're analyzing a sample in the first place — to learn something about the population that we haven't fully observed). We don't actually know the population mean (or any population parameter), so we can't draw repeated random samples from the population to create a sampling distribution. We just have the data in front of us. So, if we set a null expectation of the population parameter value and find that our sample statistic differs from that considerably, we are inclined to "reject the null hypothesis" and believe the population parameter value is different from our null expectation. But, this judgment is based upon seeing a sample statistic that is apparently *unusual*, not *impossible*, under the null hypothesis. We might therefore reject the null hypothesis sometimes when the population parameter actually is equal to the null hypothesis value but where this particular sample (due to the inherent limitations of sampling) just happens to be unusual.

25. To avoid doing this too often, we have to define an "error rate" or "significance level" that only allows us to make these kind of "false positive" judgments quite rarely. We'll call this significance level α in order to explore different possible values and what consequences that has for the probability of obtaining a "false positive." A commonly used value of α is 0.05. This means that we will only declare a sample statistic to be "statistically significant" when it is as far or farther from the null hypothesis value as 5% of the possible sample statistics generated from random samples from a population with a parameter equal to the null hypothesis value. In essence, we're looking for an outlier statistic that is farther than the 2.5% percentile or 97.5% percentile of the sampling distribution. If we look at our sampling distribution, we can see how far a sample statistic would have to be in order for us declare it statistically significantly different from the population mean:

    ```
    quantile(dist100, c(0.025, 0.975))
    ```

    Check your understanding by assessing why these particular quantiles are used when $\alpha = 0.05$. Those quantiles reflect a "two-tailed" hypothesis test in which we look to see whether a sample statistic is an outlier in either direction. What quantiles would we consider if we were doing a "one-tailed" hypothesis to see if the sample statistic was an outlier only on the upper end of the distribution? only on the lower end of the distribution?

26. Different values of α are possible. Small values mean we want a lower chance of a "false positive" judgment, at the expense of making many more "false negatives" (judgments where we say a sample statistic is consistent with the null hypothesis when in fact it is drawn from a different population). Try some different values. Why are these two rates (false positives and false negatives) trade-offs?

27. But recall, we rarely have access to the population mean and we rarely have the ability to repeatedly sample from a population in order to create a sampling distribution. We therefore rely on "distributional" (or "parametric") assumptions. Rather than generating a sampling distribution from repeatedly sampling (because we only have one sample), we instead rely upon a mathematical theorem — the central limit theorem — that shows that the sampling distribution of the mean is normally distributed. You can get a sense of this by repeating our repeated sampling exercise above by collecting more samples.

    The more samples we draw from the population (even if they are all size $n = 10$), the more and more the shape of the sampling distribution will resemble the normal distribution's bell curve:

```
ggplot(, aes(x = replicate(100, mean(sample(x, 10, FALSE))))) +
    geom_histogram(bins = 21) + xlim(0,20)
ggplot(, aes(x = replicate(500, mean(sample(x, 10, FALSE))))) +
    geom_histogram(bins = 21) + xlim(0,20)
ggplot(, aes(x = replicate(2500, mean(sample(x, 10, FALSE))))) +
    geom_histogram(bins = 21) + xlim(0,20)
```

This property enables us to calculate how unusual a sample estimate is against a null hypothesis by simply calculating the probability of seeing different statistic values given the normal distribution (which has a well-defined formula). In R we can calculate these probabilities using the `pnorm()` function. If we are thinking about sample means, we can calculate the probability of different sample means against any particular null hypothesis parameter value. To do so, however, requires that we express the mean as a difference from the null hypothesis value and rescale to the scale of standard errors. In our example, we have to convert the sample mean to number of years different from the population mean and then rescale to units in number of standard errors: i.e., from `mean(s2)` to

```
zstat <- (mean(s2) - 12) / ( sd(s2)/sqrt(length(s2)) )
```

This value is called a z-statistic. If we had a different null hypothesis value (e.g., that the population mean was 3), we would use that instead of `12` in the above calculation.

28. To see how unusual this z-statistic is, we plug it into the `pnorm()` function: `pnorm(zstat)`. The output is called a p-value and can be understood as the probability of seeing a test statistic this far or farther from the null hypothesis value. Formally, it is “the probability of a t-statistic as extreme as the one we observed, if the null hypothesis was true.” When this p-value is smaller than α level, we judge the sample mean to be “statistically significantly different from zero.” When the p-value is larger than α, we declare that the test statistic is not statistically significantly different from the null hypothesis value. However, it is important to remember when reading this that a $p - value$ is not:

    - the probability that a hypothesis is true or false
    - a reflection of our confidence or certainty about the result
    - the probability that the true mean is in any particular range of values
    - a statement about the importance or substantive size of the effect

    Those are all common misconceptions of how to interpret a p-value.

29. To see how large a z-statistic has to be to cross different α levels, you can explore the `qnorm()` function, which takes a p-value as input and returns the corresponding z-statistic.

```
qnorm(0.0251)  # 5% significance level
qnorm(0.050)  # 10% significance level
qnorm(0.330)  # 33% significance level
qnorm(0.250)  # 50% significance level
```

30. We now return to the idea from MT of a “confidence interval,” which is an interval estimate like the interval created by calculating a margin of error back when we discussed sampling. A confidence interval (or “CI”) is simply a range, centered on our sample estimate that tells us about the likely location of the population parameter within a stated range of uncertainty.[4] To formalize this, a confidence interval tells us:

[4] Because it is just a transformation of the margin of error, it is based on the variability of the data, sampling procedures, and — most importantly — sample size.

> Were we to repeat our procedure of sampling, analyzing the sample to produce a sample estimate and standard error, and transforming those estimates into a confidence interval *repeatedly* from the population, a fixed percentage of the resulting intervals would include the true population-level parameter.

This does not say for sure whether the particular estimated confidence interval *this time* actually includes that true population parameter. We never know that. Why?

31. To get at the notion of the confidence interval, we are going back to our population data `x` and repeatedly drawing new samples. This time, however, we are going to store our results in a data frame and we are going to save not only how far the sample mean deviates from the population mean, but we are also going to save the standard error calculated from each sample:

```
alpha <- qnorm(0.025) # 5% significance level; 95% confidence interval

set.seed(123)
k <- 100  # number of samples to draw and estimate statistic on
ci <- data.frame(i = seq_len(k),
                 means = numeric(k),
                 se = numeric(k),
                 off = logical(k))
for (i in seq_len(k)){                       # Take 100 samples from our distribution
  tmp <- sample(x, 100, replace=FALSE)       # Store samples in 'temp'
  ci$means[i] <- mean(tmp)-mean(x)           # calculate and store mean
  ci$se[i] <- (sd(tmp)/sqrt(length(tmp)))    # calculate and store upper CI limit
}
```

    You can explore this new object, `ci`, perhaps using `summary(ci)`.

32. What we have generated here is a data.frame of "centered" sample means: we are expressing how far our sample mean is from the population mean. This allows us to calculate a so-called "confidence" interval. A confidence interval is an interval — recall how we have already encountered it in the form of a margin of error — that attempts to provide information about the location of the true population value. The confidence interval is sized based upon the variability of our data, the size of the sample we draw, and α. The choice of α, as in the rest of statistical hypothesis testing, is important because it dictates our error rate.

    When we set the width of the confidence level as $1 - \alpha$, we are saying we will allow α proportion of our confidence intervals (were we to sample an infinite number of times) to not include the true population parameter. If $\alpha = 0.05$, then we are drawing 95% confidence intervals. Thus only 5% of the intervals we draw from this population are likely to "miss" the true population parameter. If we therefore find in our particular sample that the interval differs from our null expectation (e.g., the sample mean does not equal zero and the 95% confidence interval based upon our sample data does not include 0), then we would say the sample mean difference is statistically significantly different from zero. (The p-value in this case would be less than 0.05.) So this either indicates that the population parameter is truly not equal to zero or that our particular sample happens to have produced one of the 5% of confidence intervals that are expected (given our sampling procedure, sample size, and α level) to not cover the true population parameter, simply due to chance.

- Can we know with certainty which interval — either (a) an interval from a population different from the one characterized by our null hypothesis value, or (b) a false positive — we observe in a given situation? Why or why not?

33. How many of our confidence intervals do not cover the population mean? (Recall we have rescaled the mean to be a different from the true value.):

```
ci$off <- ((ci$means-(abs(alpha)*ci$se)) > 0 & (ci$means+(abs(alpha)*ci$se)) > 0) |
          ((ci$means-(abs(alpha)*ci$se)) < 0 & (ci$means+(abs(alpha)*ci$se)) < 0)
table(ci$off)
```

34. We can graph all of our confidence intervals to see which include the true mean and which do not (colored by the `off` variable we just generated):

```
ggplot(ci, aes(x = i, y = means, colour = off)) +
    geom_errorbar(aes(ymin = (means - alpha*se),
                      ymax = (means + alpha*se)), width=.1) +
    geom_point() + coord_flip()
```

This exercise shows that if a particular parameter value is true (in this case, the population mean is zero), we can draw confidence intervals for that mean to try to estimate where the mean is located. Most of these intervals will "cover" the true population parameter value, but not all of them. The number that cover the true population parameter value depends on the width of the confidence interval we draw. If we draw a wider confidence interval (say 95%), then 95% of the confidence intervals drawn from samples of this size from the population will cover the true population parameter value. If we draw a narrower confidence interval (say 50%), then only 50% of the confidence intervals drawn from samples of this size from the population will cover the true population value.

Check your intuition this discussion:

- How wide is a 100% confidence interval?
- How wide is a 0% confidence interval?
- If we increase the sample size from $n = 100$ to $n = 500$, what happens to the width of a 95% confidence interval drawn on each sample? Which is wider?
- Consider, again, increasing the sample size from $n = 100$ to $n = 500$ and drawing a 95% confidence interval for each. For which sample size is the 95% confidence interval more likely to include the true population value?

35. Try to repeat the above sampling and graphing procedure but tweak the values of α and sample size. As α increases, fewer of the confidence intervals drawn from the population with cover the true population parameter. That means that we are more likely to make "false positive" judgments and to have incorrect beliefs about the location of the population parameter.

36. We can also see, in the above data, the equivalence of a confidence interval, a z-statistic, and a p-value. When our confidence interval does not overlap the null hypothesis value, we describe it as statistically significant. In such cases, the z-statistic is larger than the cut-off threshold for our chosen α level and the p-value will be less than α. For a sample mean, the z statistic is simply the sample mean divided by the standard error; when this test statistic exceeds the critical value, α, then the statistic is deemed statistically significantly different from the null hypothesis value. If these — test statistic, p-value, confidence interval — are all equivalent, why would we display one versus another?

3.2 Application to "Real" Data

37. Returning to the QoG data we used for Problem Set 3, load the data into R:

```
library("rio")
d <- import("qog_std_cs_jan16.dta")
t.test(d$bl_asy15f, mu = 12)
```

How many cases are in this dataset? What's the unit of analysis?

38. Then apply what we've just learned to assess whether the mean years of female educational attainment differ from a null hypothesis value. You can do this quickly in R using the `t.test()` function, where `mu` is the value of your null hypothesis:

```
# country-level mean
mean(d$bl_asy15f, na.rm = TRUE)

# t-test
t.test(d$bl_asy15f, mu = 12)
```

Try out different possible null hypothesis values, and specify them as `mu`.

39. A further test we might be interested in is whether countries' average educational attainment for men differs from the educational for women. This is what is known as a two-sample t-test. Like exercise we just performed to generate confidence intervals, this is based on a null hypothesis about the *mean-difference* (the difference between two variables' means). Conventionally, and in this example, we might have a null hypothesis that the mean-difference is zero (i.e., that the two variables — educational attainment for men and women in each country — do not differ). Try it out:

```
t.test(x = d$bl_asy15f, y = d$bl_asy15m)
```

Does the mean country-level educational attainment for men and women differ?

40. Another further test we might be interested in is whether two subsets of cases differ from one another with respect to a given statistic. To capture this, we can also use `t.test()`, this time with a *formula* notation. For example we might test whether the countries with high and low levels of democracy differ in female educational attainment:

```
dem_high <- factor(d$fh_polity2 > mean(d$fh_polity2, na.rm = TRUE))
t.test(d$bl_asy15f ~ dem_high)
```

41. Try to interpret all of the above results. What do they mean? Are they substantively and/or statistically significant?

To be clear, statistical significance tells us whether an estimate, a relationship, or an effect is large relative to a hypothetical distribution of test statistics corresponding to null expectation. This says nothing about whether that effect is large or important in substantive terms.[5] If we find that democracy and non-democracies differ by $50 in per capita GDP that this difference is statistically different from zero, that is a statistically significant difference. Whether that difference is large or important depends upon the state of broader scientific understanding, the amount of dispersion in the data (is the difference large if measured in number of standard deviations), the size of other differences (do regions of the world vary more than one another on this variable), the research context, and our own judgment. $50 may be a substantively small effect when talking about GDP but it may be a large effect when talking about the cost of tonight's dinner. This is something for you as a researcher to consider.

[5] If we have enough data (i.e., our sample is large enough), almost any test statistic will "statistically significant" but that does not mean that the estimated parameter is large or important.