MEDI 504A - Lab 3

Matthew Manson

MEDI 504 Lab 3

Basic biostatistics

By the end of this lab, students should be able to:

- Identify the different types of data analysis questions and categorize a question into the correct type
- Identify a suitable analysis type to answer an inferential question, given the data set at hand
- Use the R programming language to carry out analysis to answer inferential question
- Interpret and communicate the results of the analysis from an inferential question

Exercise 1: types of data analysis questions

rubric={10 points}

In the reading **Types of data analytic questions**, you were introduced to different types of statistical questions. Let us refresh our knowledge of these here and play name that statistical question! For each question below, assign the answer to one of the following types of statistical question being asked:

- **Descriptive.**
- **Exploratory.**
- **Inferential.**
- **Predictive.**
- **Causal.**
- **Mechanistic.**

Exercise 1.1:

Is wearing sunscreen associated with a decreased probability of developing skin cancer in Canada?

> Inferential

Exercise 1.2:

Is there a relationship between alcohol consumption and socioeconomic status in the 2018 City of Vancouver survey data set?

> Exploratory

Exercise 1.3:

Does performing strength training 3 times a week lead to an increase in bone density in the elderly?

Causal

Exercise 1.4:

How do changes in human behaviour lead to a reduction in the number of COVID-19 confirmed cases?

Mechanistic

Exercise 1.5:

Does reduced caloric intake cause weight-loss in adults?

Causal

Exercise 1.6:

Do countries with lower COVID-19 vaccination rates have higher levels of hospitalizations compared to countries with higher COVID-19 vaccination rates?

Inferential

Exercise 1.7:

Is vaccination against COVID-19 negatively associated with the presence long-COVID symptoms?

Inferential

Exercise 1.8:

How many patients will go to the emergency department at Vancouver General Hospital tomorrow?

Predictive

Exercise 1.9:

How many COVID-19 patients are in BC hospitals today?

Descriptive

Exercise 1.10:

Are high contrast images associated with better visual discrimination by the visually impaired?

Inferential

Exercise 2: identifying a suitable analysis method for a given question and data set

rubric={10 points}

Given the statistical question below and the data set description and snippet, name the type of statistical question and a suitable analysis method. Justify your choices for both the question type and analysis method. *Note: this case is fictional, but based on available medical methods.*

Statistical question: Is there a difference in the proportion of miscarriages in in-vitro fertilization (IVF) patients whose embryos undergoe preimplantation genetic testing for aneuploidy (PGT) compared to those whose embryos do not?

Question Type: Inferential

Justification: This question suggests that the results of the analysis of data from a sample of those who undergo IVF will be generalized to all individuals who undergo IVF (i.e the broader population). For this reason this question can be considered an inferential statistical question - it aims to make statements beyond the sample of data on which statistical tests were conducted.

Data set: Data from 457 patients was collected from the a local fertility clinic. Patients had the choice of opting for PGT or not. There is an added financial cost for PGT and therefore not all patents chose to opt for this added treatment. 196 patients opted to undergo PGT screening of their embryos, and 261 opted to forgo this screening. Miscarriage proportion for each patient was calculated as the number of unsuccessful embryo transfers divided by the total number of embryo transfers (successful + unsuccessful). A snippet of the data is shown below:

`patient_id`	`miscarriage_proportion`	`pgt`
2361344	0.25	yes
2361932	0.33	no
2397563	0	no
...	...	...
2595244	1	yes

Analysis Method: A method to compare two proportions - z-test/proportion-test (in R - `prop.test()`), or Chi-square test.

Justification: Here, the researchers are trying to compare the *proportion* of miscarriages experienced by two groups (specifically, the task is to compare the average proportion of miscarriages between the two groups - those who underwent PGT and those who did not). When comparing two *proportions*, we can use one of the two methods above, however their assumptions differ. I would use the R function prop.test(), as it is specifically designed for this case (i.e. the documentation specifies it is used for testing the null hypothesis that proportions are the same). prop.test() assumes that sample size is large (i.e. rule of thumb = the number of successes in each group is more than 5), and that the proportions are independent and a result of random sampling and treatment randomization. In this case, all assumptions are likely met (although more information would be needed to confirm), supporting its use.

Exercise 3: using R to visualize uncertainty of point estimates

rubric={20 points}

In a recent study by Jiang et al (2019), they investigated the effects of intramuscular and vaginal progesterone supplementation on frozen-thawed embryo transfer during in-vitro fertilization (IVF). This is an important question because progesterone supplementation is critical during IVF frozen-thawed embryo transfer, and intramuscular supplementation has many negative side effects (e.g., inconvenience, local pain and inflammation at the injection site). Patients were assigned to one of two groups: - group A with progesterone

intramuscular injection (60 mg/d) - group B with progesterone vaginal sustained-release gel of progesterone (90 mg/d)

The response variable of interest was whether a pregnancy resulted in a live birth (coded as 1) or not (coded as 0).

Your task here is to load the `data/jiang-live-birth.csv` file and create an effective data visualization which communicates the estimates for each group (proportion of live births) as well as the uncertainty of those point estimates at a 95% confidence interval.

```
# load the data
jiang_live_birth <- read.csv("data/jiang-live-birth.csv")

# look at the data (first 10 rows)
head(jiang_live_birth, 10)
```

```
##            group live_birth
## 1  intramuscular          0
## 2        vaginal          0
## 3  intramuscular          1
## 4  intramuscular          1
## 5  intramuscular          0
## 6  intramuscular          1
## 7  intramuscular          0
## 8  intramuscular          0
## 9  intramuscular          0
## 10 intramuscular          0
```

```
# creating new data frame displaying the number of live birth
# successes, total attempts, and proportion of successes for each group
(jiang_live_birth1 <- jiang_live_birth %>%
  group_by(group) %>%
  summarise(live_births = sum(live_birth), total = n()) %>%
  mutate(prop = live_births / total))
```

```
## # A tibble: 2 x 4
##   group         live_births total  prop
##   <chr>               <int> <int> <dbl>
## 1 intramuscular         811  1988 0.408
## 2 vaginal               461  1025 0.450
```

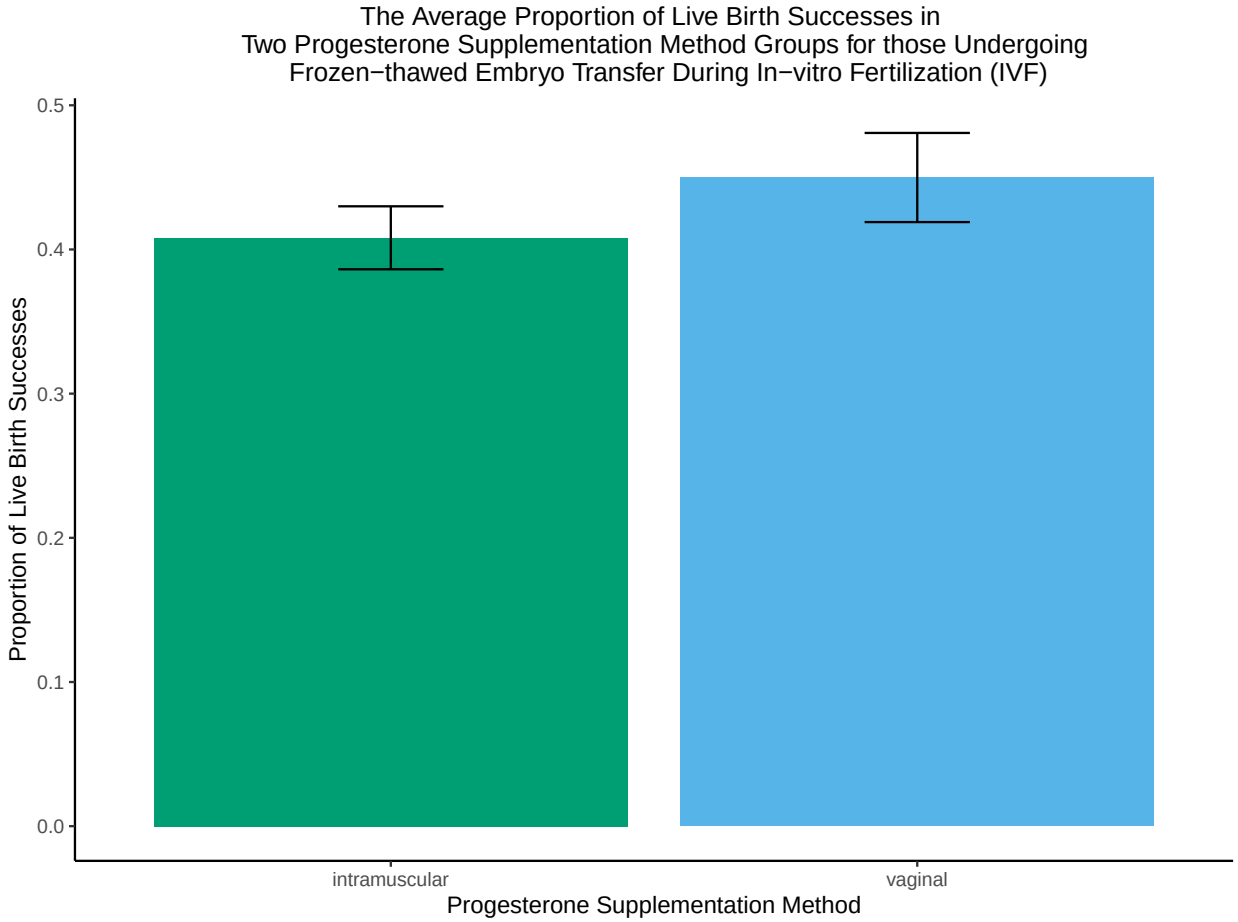
```
# 95% confidence interval for each group, using exact method
(confidence <- binom.confint(jiang_live_birth1$live_births,
  jiang_live_birth1$total,
  conf.level = 0.95,
  methods = "exact"))
```

```
##   method   x    n      mean     lower     upper
## 1  exact 811 1988 0.4079477 0.3862420 0.4299253
## 2  exact 461 1025 0.4497561 0.4189955 0.4808060
```

```r
# binding confidence interval upper and lowers to jiang_live_birth_1
# and creating final data frame for plotting, jiang_live_birth_final
(jiang_live_birth_final <-
  cbind(jiang_live_birth1, confidence$lower, confidence$upper) %>%
  # rename columns
  rename(conf_lower = "confidence$lower", conf_upper = "confidence$upper"))
```

```
##           group live_births total      prop conf_lower conf_upper
## 1 intramuscular         811  1988 0.4079477  0.3862420  0.4299253
## 2       vaginal         461  1025 0.4497561  0.4189955  0.4808060
```

```r
# plotting proportions of live births in each group
jiang_live_birth_final %>%
  ggplot(aes(x = group, y = prop)) +
  geom_col(aes(fill = group), show.legend = FALSE) +
  # adding 95% CI error bars, custom width
  geom_errorbar(aes(ymin = conf_lower, ymax = conf_upper), width = 0.2) +
  labs(
    title = "The Average Proportion of Live Birth Successes in
    Two Progesterone Supplementation Method Groups for those Undergoing
    Frozen-thawed Embryo Transfer During In-vitro Fertilization (IVF)",
    x = "Progesterone Supplementation Method",
    y = "Proportion of Live Birth Successes") +
  scale_fill_manual(values = c("#009E73", "#56B4E9")) +
  theme_classic() +
  theme(plot.title = element_text(size = 12, hjust = 0.5))
```



The above figure displays the proportion of live birth successes in each progesterone supplementation group, and the uncertainty of these point estimates using 95% confidence interval error bars.

Exercise 4: using R to infer group differences

rubric={20 points}

The error bars representing the uncertainty of our estimates in the visualization overlap! From this visualization alone, it is not yet clear as to whether the observed difference in the estimates is statistically significant. Perform a suitable analysis to answer this question. If any hypotheses or assumptions are made in your analysis, state them. Clearly communicate your results.

To summarise, so far, I have:

1. imported the data on live births in two groups (those receiving IM and Intravaginal Progesterone Supplements),
2. looked at the data set and calculated the proportion of live births within each group,
3. calculated the 95% confidence interval for the two point estimates (i.e. the proportion of live births in each group), and
4. plotted the proportion of live births in each supplementation group with error bars corresponding to the 95% confidence intervals of the point estimates.

Now I need to perform an analysis to determine whether the difference in point estimates is statistically clear. Lets do this.

Question: Is there a difference between the proportion of live birth successes among those who recieve intramuscular progesterone supplementation (group A, proportion A, pA) and the those who receive progesterone vaginally (group B, proportion B - pB) while undergoing frozen-thawed embryo transfer during in-vitro fertilization (IVF)?

Null Hypothesis: The observed proportion of live birth successes in group A (intramuscular progesterone supplementation) is equal to the observed proportion of live birth successes in group B (vaginal progesterone supplementation). (pA = pB)

Alternative Hypothesis: The observed proportion of live birth successes in group B (intramuscular progesterone supplementation) is different than the observed proportion of live birth successes in group B (vaginal progesterone supplementation). (pA != pB)

This is a two tailed test (not looking for greater than or less than, just looking for difference).

I can use `prop.test()`, an R function that is designed to test the null hypothesis that the proportions of success are equal across groups, to test the null hypothesis outlined above. Note that this method assumes that sample size is large (i.e. rule of thumb = the number of successes in each group is more than 5), and additionally that the proportions are independent and a result of random sampling and treatment randomization. The data we are working with reports 811 successes in one group and 461 in the other, and we will assume that these data are independent and random allocation took place. First, i'll run the proportion test and assign it to prop_test_births:

```
# proportion test - produces p-value and confidence interval
# for difference in proportions
(prop_test_births <- prop.test(jiang_live_birth_final$live_births,
  jiang_live_birth_final$total,
  alternative = "two.sided",
  conf.level = 0.95))
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  jiang_live_birth_final$live_births out of jiang_live_birth_final$total
## X-squared = 4.6761, df = 1, p-value = 0.03059
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.079886541 -0.003730282
## sample estimates:
##    prop 1    prop 2
## 0.4079477 0.4497561
```

Lets have a look at these results in a tidier format, using broom package function `tidy()`:

```
tidy(prop_test_births)
```

```
## # A tibble: 1 x 9
##   estimate1 estimate2 statistic p.value parame~1 conf.~2 conf.h~3 method alter~4
##       <dbl>     <dbl>     <dbl>   <dbl>    <dbl>   <dbl>    <dbl> <chr>  <chr>
## 1     0.408     0.450      4.68  0.0306        1 -0.0799 -0.00373 2-sam~ two.si~
## # ... with abbreviated variable names 1: parameter, 2: conf.low, 3: conf.high,
## #   4: alternative
```

The results of the proportion test suggest that there is a statistically clear difference in the proportion of live birth successes between the two observed groups receiving different progesterone supplementation methods ($p = 0.031$, (smaller than standard practice alpha level of 5%)). The 95% confidence interval for the difference in proportions is (-0.0799, -0.0037), suggesting a very minor but statistically clear higher proportion of birth successes for those receiving vaginal supplementation (does not cross zero - we can be 95% confident that the true difference in proportions falls between these values).

The question remains, just how large is this observed difference? The confidence interval suggests that it is quite small, and a small difference is likely to have little meaning in clinical practice. Effect size calculations can also help to quantify this. I'll calculate Cohen's h, a method of quantifying the difference between proportions. Cohen's h can be used to classify the difference between proportions as small, moderate, or large, and ultimately, decide whether such difference is likely to be meaningful. To do so, i'll use the function `ES.h()` from the package pwr.

```
# effect size
ES.h(jiang_live_birth_final[1, 4], jiang_live_birth_final[2, 4])
```

```
## [1] -0.08450315
```

The calculated Cohen's h value is ~ -0.085. Jacob Cohen provides criteria for interpreting this number in his book, 'Statistical Power Analysis for the Behavioral Sciences' (Cohen, 1988). An effect size of $h = 0.2$ is "small", $h = 0.5$ is "medium", and $h = 0.8$ is "large". Note that interpreting this value with respect to Cohen's criteria requires consideration of the absolute value (i.e. ignore the sign), although the sign does indicate the direction of the effect. In this case, the Cohen's h value of -0.085 suggests that IM injections tend to have less live birth success, however, the Cohen's h value is also extremely small (smaller than 0.2), suggesting that the effect of this difference would likely not be clinically meaningful.

For Jiang et al., the results of this analysis suggest that, although there was a statistically clear difference in the proportion of live birth successes across each group, this difference is extremely small and unlikely to be meaningful in clinical practice. For patients undergoing IVF, this analysis suggests that there is likely to be minimal clinically meaningful difference in the live birth success based on the method of progesterone supplementation, and because the vaginal method places less burden on patients, it may be the better approach.

(Optional) Exercise 5: using R to handle multiple comparisons

rubric={3 bonus points}

We will be working with the results from a Genome-wide analysis-like study found in `data/GWAS_results` from Timbers et al. (2016). The dataset contains two columns: a list of gene names (`gene`) and a list of unadjusted p-values (`pval`) generated from the analysis (the particular statistical test used is the Sequence Kernel Analysis test). These p-values were created by repeating the analysis on many variables from the same dataset. Thus we have a multiple testing problem to deal with. Each p-value corresponds to a gene and tests whether that gene is associated with a phenotype.

Note: before you get started on this question we recommend you read the following: - Types of errors section of the Modern Dive statistics textbook - Why is multiple testing a problem and what do I need to do about it? slides

```
# import data, select relevant rows, print to screen
GWAS_results <- read_csv("data/GWAS_results.csv", show_col_types = FALSE) %>%
  select(gene = public_gene_name, pval = `p-value`)
GWAS_results
```

```
## # A tibble: 1,150 x 2
##    gene          pval
##    <chr>        <dbl>
##  1 osm-1   0.00000102
##  2 che-3   0.0000161
##  3 F01D4.9 0.0000556
##  4 mdf-1   0.0001
##  5 cnt-1   0.000124
##  6 lars-1  0.00153
##  7 F43D9.1 0.00169
##  8 hlb-1   0.00180
##  9 jac-1   0.00255
## 10 col-135 0.00287
## # ... with 1,140 more rows
```

Using in a sample of 480 mutant C. elegans (nematode worms), the question in the analysis was: **are there any genes, which when mutated, associated with a phenotype defined as a decrease in the ability to uptake a fluorescent dye into their sensory neurons?** This would indicate there might be a problem with their sensory neurons and that this gene might be important for sensory neuron development or function. The study can be accessed here: http://dx.doi.org/10.1371/journal.pgen.1006235

(Optional) Exercise 5.1:

Answer the following questions

- How many genes are present in the total dataset? For this dataset, this corresponds to the number of multiple comparisons that were performed.
- How many genes are associated with the phenotype (a decrease in the ability to uptake a fluorescent dye into their sensory neurons) at the unadjusted $\alpha = 0.05$?

```
# number of genes in the dataset:
n_distinct(GWAS_results$gene)
```

```
## [1] 1150
```

```
# number of genes associated with the phenotype:
GWAS_results %>%
  filter(pval <= 0.05) %>%
  count()
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1   100
```

There are 1150 genes in the dataset. 100 genes are associated with the phenotype, if alpha = 0.05 is considered the threshold.

(Optional) Exercise 5.2:

Briefly describe (in one or two sentences) why it would be misleading to report only one of the "significant" tests (and ignoring the fact that others were done too).

Description: In this case, multiple hypothesis tests were done simultaneously, and because of this, based on chance alone, there is some probability that some of these significant results are false positives (and this probability increases as more hypothesis tests are done!). 'Cherry picking' a single positive result and reporting it alone from a pool of significant tests provides no context to the viewer/reader, and there is relatively high probability that this significant result is just due to chance.

(Optional) Exercise 5.3:

Use the function `p.adjust()` to calculate adjusted p-values using `method = "bonferroni"`. How many and which genes are associated with the treatment after the adjustment, at the $\alpha = 0.05$ significance level?

```
# calculating adjusted p-values, adding them to the data set,
# and selecting relevant columns
(GWAS_results_adjpval <- GWAS_results %>%
  mutate(adjpval = p.adjust(pval, method = "bonferroni")) %>%
  select(gene, pval, adjpval))
```

```
## # A tibble: 1,150 x 3
##    gene          pval adjpval
##    <chr>        <dbl>   <dbl>
##  1 osm-1   0.00000102 0.00117
##  2 che-3   0.0000161  0.0185
##  3 F01D4.9 0.0000556  0.0639
##  4 mdf-1   0.0001     0.115
##  5 cnt-1   0.000124   0.143
##  6 lars-1  0.00153    1
##  7 F43D9.1 0.00169    1
##  8 hlb-1   0.00180    1
##  9 jac-1   0.00255    1
## 10 col-135 0.00287    1
## # ... with 1,140 more rows
```

```
# finding the genes associated with the treatment, after adjustment
GWAS_results_adjpval %>%
  filter(adjpval <= 0.05)
```

```
## # A tibble: 2 x 3
##   gene        pval adjpval
##   <chr>      <dbl>   <dbl>
## 1 osm-1 0.00000102 0.00117
## 2 che-3 0.0000161  0.0185
```

After using the Bonferroni method to adjust the p-values and account for multiple testing, only two genes are associated with the treatment. These include osm-1 and che-3.

References

Cohen, Jacob (1988). Statistical Power Analysis for the Behavioral Sciences (2nd ed.).

Jiang, L., Luo, ZY., Hao, GM. et al. Effects of intramuscular and vaginal progesterone supplementation on frozen-thawed embryo transfer. Sci Rep 9, 15264 (2019). https://doi.org/10.1038/s41598-019-51717-5

Timbers TA, Garland SJ, Mohan S, Flibotte S, Edgley M, et al. (2016) Accelerating Gene Discovery by Phenotyping Whole-Genome Sequenced Multi-mutation Strains and Using the Sequence Kernel Association Test (SKAT). PLOS Genetics 12(8): e1006235. https://doi.org/10.1371/journal.pgen.1006235