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?

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

This would be an inferential question. This research question is looking at relationships/patterns, and trying to generate a hypothesis from it. The wording of the question does not give a direction of the association, but rather evaluating whether there is a difference at all, meaning it cannot be causal. There are also other factors that were not controlled for, which would influence the results of the analysis. Like a randomized experiment for example To analyse this data set, I want to look at the variables, where the explanatory variabel is the 'pgt': the patients who undergo genetic testing, and the response variable is the miscarriage proportion. The response variable is numerical continuous and the explanatory variable is categorical. Assuming the data is normally distributed, we are comparing two groups, the pgt testing positive vs negative, therefore we can use a z test, because there is a proportion variable here and we need to create a standard normal distribution note from class: premutation test

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 live birth data
live_births <- read.csv('data/jiang-live-birth.csv')
head(live_births)
```

```
##           group live_birth
## 1 intramuscular          0
## 2       vaginal          0
## 3 intramuscular          1
```

```
## 4 intramuscular          1
## 5 intramuscular          0
## 6 intramuscular          1
```

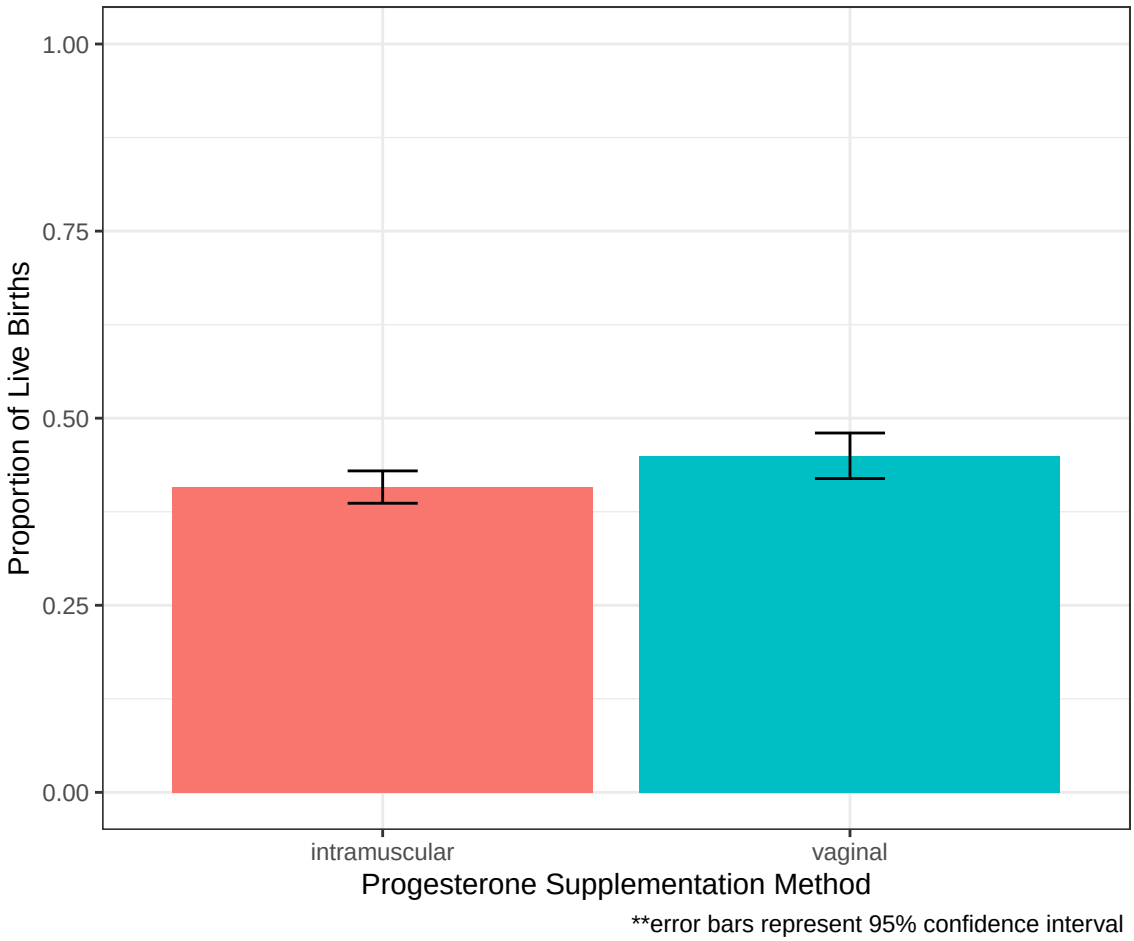
```
#group and count the number of '1' or successes (live births) in this scenario
births_success <- live_births %>%
  group_by(group) %>%
  count(live_birth) %>%
  group_by(live_birth)
births_success
```

```
## # A tibble: 4 x 3
## # Groups:   live_birth [2]
##   group         live_birth     n
##   <chr>              <int> <int>
## 1 intramuscular          0  1177
## 2 intramuscular          1   811
## 3 vaginal                0   564
## 4 vaginal                1   461
```

```
#create the confidence intervals using asymptotic method
im_ci <- binom.confint(811, 1988, conf.level = .95, methods = "asymptotic")
vg_ci <- binom.confint(461, 1025, conf.level = .95, methods = "asymptotic")

#add new column before merging the two data frames
im_ci$type <- "intramuscular"
vg_ci$type <- "vaginal"

births_calc <- merge(x = im_ci, y = vg_ci, all = TRUE)
births_calc
```

```
##       method   x    n      mean     lower     upper          type
## 1 asymptotic 461 1025 0.4497561 0.4193015 0.4802107       vaginal
## 2 asymptotic 811 1988 0.4079477 0.3863443 0.4295511 intramuscular
```

```
#plot the visualization using the means and confidence intervals
birth_plot <- births_calc %>%
  ggplot(aes(type, mean, fill = type)) +
  geom_bar(stat = "identity") +
  geom_errorbar(aes(ymin = lower, ymax = upper),
                width = .15)+
  #visuals
  xlab("Progesterone Supplementation Method") +
  ylab("Proportion of Live Births") +
  labs(caption = "**error bars represent 95% confidence interval ")+
  #scale 0-1 because this is a proportion measure, and when I don't scale it, it looks almost misleadin
  ylim(0,1)+
  theme_bw() +
  theme(
    legend.position = 'none'
  )

birth_plot
```



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.

```
# Your code goes here
#x is successes, with two entries, with intramuscular and vaginal respectively, n represents the total
test <- prop.test(x=c(811, 461), n=c(1988, 1025), conf.level = .95, correct = FALSE )
test
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(811, 461) out of c(1988, 1025)
## X-squared = 4.846, df = 1, p-value = 0.02771
## alternative hypothesis: two.sided
## 95 percent confidence interval:
```

```
##  -0.079147227 -0.004469595
## sample estimates:
##    prop 1    prop 2
## 0.4079477 0.4497561
```

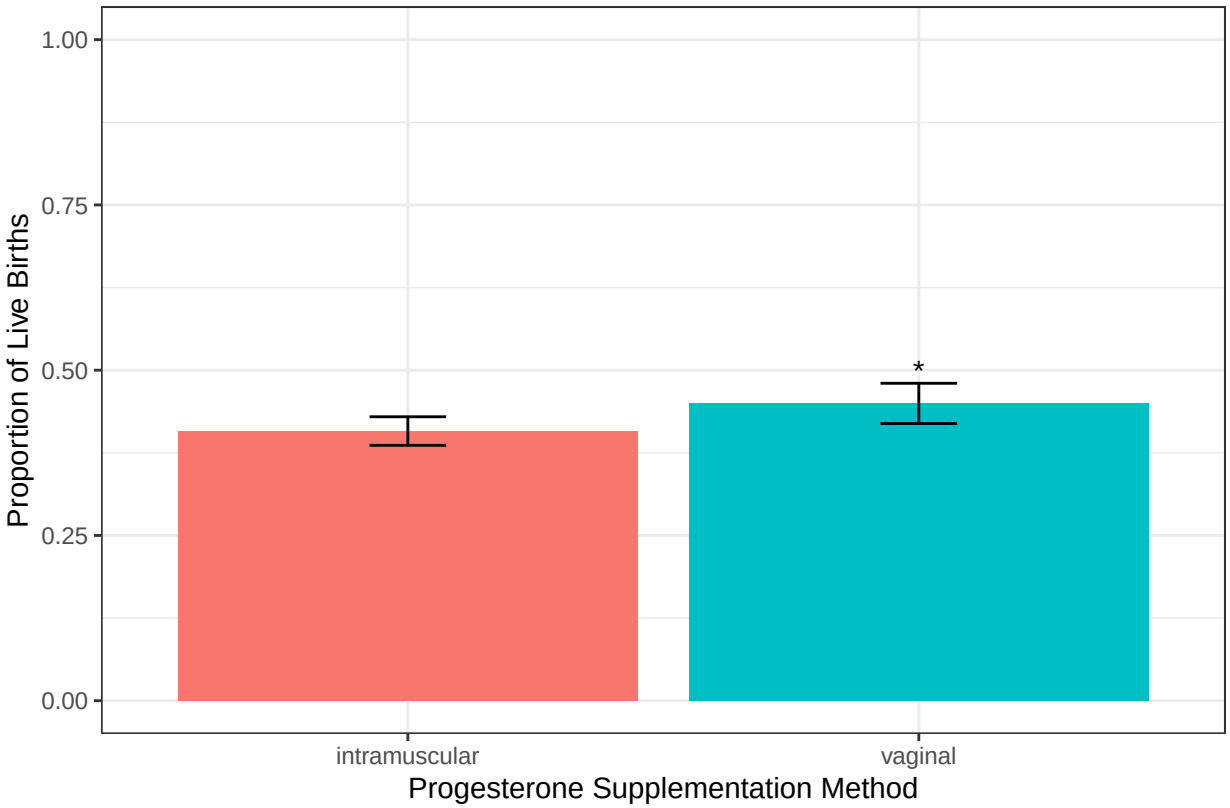
```
#this gave out a non-tidy version of data, so I will tidy it
tidy_test <- tidy(test)
tidy_test
```

```
## # A tibble: 1 x 9
##   estimate1 estimate2 statistic p.value parameter conf.low conf.high method
##       <dbl>     <dbl>     <dbl>   <dbl>     <dbl>    <dbl>     <dbl> <chr>
## 1     0.408     0.450      4.85  0.0277         1  -0.0791  -0.00447 2-sample t~
## # ... with 1 more variable: alternative <chr>
```

```
#I am most interested in the p value, to answer the question
tidy_test$p.value
```

```
## [1] 0.02771078
```

```
#which gives 0.0277

# I will now add this p value to the previous visualization, just to visualize this, in addition to the
birth_plot_2 <- births_calc %>%
  ggplot(aes(type, mean, fill = type)) +
  geom_bar(stat = "identity") +
  geom_errorbar(aes(ymin = lower, ymax = upper),
                width = .15)+
  #visuals
  xlab("Progesterone Supplementation Method") +
  ylab("Proportion of Live Births") +
  labs(caption = "error bars represent 95% confidence interval, * denotes statistical difference, p = .
  #scale 0-1 because this is a proportion measure, and when I don't scale it, it looks almost misleadin
  ylim(0,1)+
  theme_bw() +
  theme(
    legend.position = 'none'
  )+
   annotate("text", x = "vaginal", y = 0.5, label = "*")

birth_plot_2
```



> I will perform a z test, using prop.test and setting the correct to false - this will allow me to do the z test. The reason I will use this is because it is better suties for looking at the proportions or probability of successes, which in this case is the live births. Because the visualization was not clear before, we need to run a statistical test which will test whether the null hypothesis (that the proportions in the two groups are the same). > null hypothesis: there is no difference in proportion between intramuscular and vaginal method of progesterone administration on live births > alternative hypothesis: there is a difference (in either way b/c this is two sided) in proportion between intramuscular and vaginal method of progesterone administration on live births > Results: the p-value is reported as 0.028. For a statistical difference, we would need the p value to be <0.05, under the 95% confidence level. 0.028<0.05, meaning that we are lending support to the alternative hypothesis. Because we defined the null hypothesis that there is no difference in proportion between intramuscular and vaginal methods, it means that there is a difference between the two groups. Using the visualization to help us, these findings suggest that the vaginal group is the more effective progesterone administration method, although we will need a follow up experiment with considerably more controls (causal question) to better address this. > the results of this analysis have thus generated a more specific hypothesis for future studies.

(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

```
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

Exercise 6.1:

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 $\alpha = 0.05$?

```
# Your code goes here
```

> Your answer goes here.

(Optional) Exercise 6.2:

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).

> This is an issue because those other gene mutations for which this comparison was performed for could have an effect on other genes, which are reported. Basically, this is ignoring that these gene mutations and significant tests are independent of each other, when they might not be.

(Optional) Exercise 6.3:

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?

```
# Your code goes here
```

Your answer goes here.

References

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