MEDI504 Lab 3

Basic Biostatistics

ALLISON JUNG - 55018618

Contents

Learning Goals

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.
- Given the data set at hand, identify a suitable analysis type to answer an inferential question.
- Use the `R` programming language to carry out analysis to answer an inferential question.
- Interpret and communicate the results of the analysis from an inferential question.

Setup

You will need the below packages in subsequent questions. If you fail to load any package, you can install them and try loading the library again.

Run the cell below before proceeding.

Exercise 1: Types of data analysis questions

rubric={10 points}

In the reading **Types of data analytic questions**, you are 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.**

> **Heads-up:** No need to justify your answer.

Q1.1

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

ANSWER:

Inferential

Q1.2

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

ANSWER:

Exploratory

Q1.3

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

ANSWER:

Causal

Q1.4

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

ANSWER:

Mechanistic

Q1.5

Does reduced caloric intake cause weight-loss in adults?

ANSWER:

Causal

Q1.6

Suppose you have a comprehensive worldwide dataset of COVID-19 vaccination rates on a country level. That said, do countries with lower COVID-19 vaccination rates have higher levels of hospitalizations compared to countries with higher COVID-19 vaccination rates?

ANSWER:

Inferential

Q1.7

Is vaccination against COVID-19 negatively associated with the presence of long-term COVID-19 symptoms in adults living in B.C.?

ANSWER:

Inferential

Q1.8

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

ANSWER:

Predictive

Q1.9

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

ANSWER:

Descriptive

Q1.10

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

ANSWER:

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 dataset 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.

> **Heads-up:** 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?*

Dataset

Data from $n = 457$ patients was collected from a local fertility clinic. Patients had the choice of opting for PGT or not. There is an added financial cost for PGT, so not all patients choose to opt for this added treatment. 196 patients opted to undergo PGT screening of their embryos, and 261 opted to forgo this screening. The 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

ANSWER:

The statistical question is inferentialbecause it seeks to determine if there is a statistically significant difference in miscarriage proportions between two groups of IVF patients (those who underwent PGT screening and those who did not). Inferential questions aim to draw conclusions or make generalizations about a population based on data from a sample. In this case, the goal is to infer whether the observed difference in miscarriage proportions in the sample can be generalized to the broader population of IVF patients.

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`).

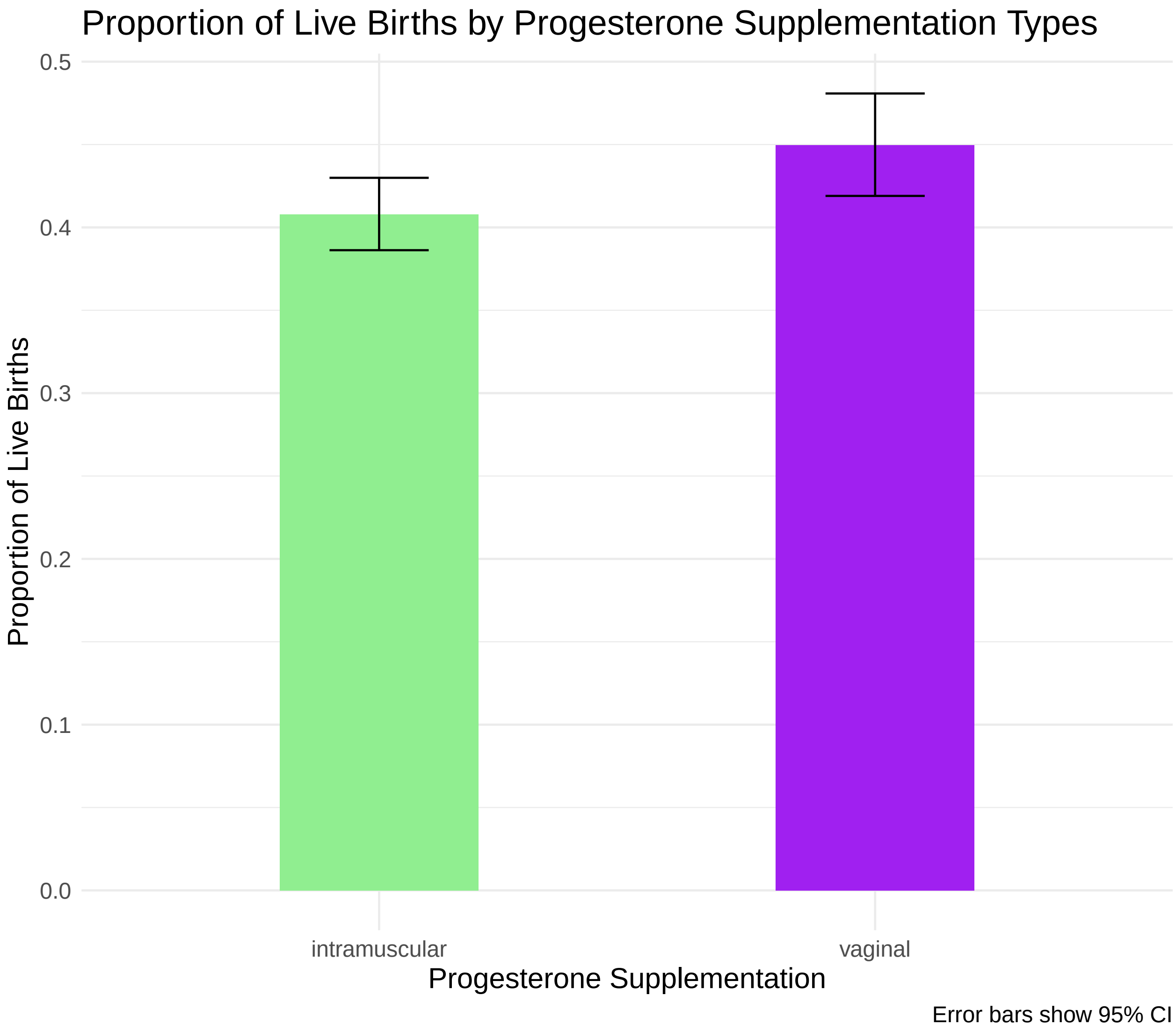
Your task here is to load the `data/jiang-live-birth.csv` file and create an effective data visualization which communicates the estimates for each group (proportion of live births) as well as the uncertainty of those point estimates at a 95% confidence interval.

> **Heads-up:** You will need to use specific functions from **{tidyverse}** for data wrangling and plotting along with a function from **{binom}** to compute the uncertainty of your proportion estimates.

```r
library(tidyverse)
library(binom)

data <- read.csv("~/Desktop/ubc/medi 504/MEDI504-basic-biostats-student-handout/data/jiang-live-birth.cs

# Inspect the first few rows of the data
head(data)
```

```
##           group live_birth
## 1 intramuscular          0
## 2       vaginal          0
## 3 intramuscular          1
## 4 intramuscular          1
## 5 intramuscular          0
## 6 intramuscular          1
```

```r
# Summarize the data by group
summary_data <- data %>%
  group_by(group) %>%
  summarize(n = n(),
            proportion = mean(live_birth),
            ci_lower = binom.confint(sum(live_birth), n(), method = "exact")$lower,
            ci_upper = binom.confint(sum(live_birth), n(), method = "exact")$upper
  )

# Create plot
ggplot(summary_data, aes(x = group, y = proportion, fill = group)) +
  geom_bar(stat = "identity", width = 0.4, show.legend = FALSE) +
  geom_errorbar(aes(ymin = ci_lower, ymax = ci_upper), width = 0.2) +
  scale_fill_manual(values = c("intramuscular" = "lightgreen", "vaginal" = "purple"))+
  labs(
    title = "Proportion of Live Births by Progesterone Supplementation Types",
    x = "Progesterone Supplementation",
    y = "Proportion of Live Births",
    caption = "Error bars show 95% CI"
  ) +
  theme_minimal() +
  theme(text = element_text(size = 14))
```



Exercise 4: Using `R` to infer group differences

rubric={20 points}

The error bars representing the uncertainty of our estimates in the previous visualization overlap in **Exercise 3**! From the above visualization alone, **it is not yet clear as to whether the observed difference in the proportion estimates of live births (between group A and group B) is statistically significant**.

Then, perform a suitable analysis to answer this question. If any hypotheses or assumptions are made in your analysis, state them. Finally, clearly communicate your results.

> **Hint:** We are dealing with proportions in this case. Therefore, a proper test for them must be found. Use a significance level $\alpha = 0.05$.

> **Heads-up:** You will need to use specific functions from **{broom}** to properly display your inferential results. Also, based on the initial inquiry in this exercise, indicate the corresponding hypotheses.

ANSWER:

*Steps for Analysis:

State the hypotheses:

1. Null Hypothesis (H): There is no difference in the proportions of live births between the two groups.
2. Alternative Hypothesis (H): There is a significant difference in the proportions of live births between the two groups.

Test for the difference in proportions: I will perform a two-proportion z-test to assess if the proportions of live births between the two groups differ significantly.

Perform the z-test:

- The test compares the observed difference in proportions with the expected difference under the null hypothesis, using the standard error of the difference in proportions.
- The p-value will be calculated to assess the significance of the difference.

Interpret the results: I will reject the null hypothesis if the p-value is less than the significance level = 0.05, indicating a significant difference between the two groups.*

```r
# Load necessary libraries
library(tidyverse)
library(broom)

# Summarize data by group
group_summary <- data %>%
  group_by(group) %>%
  summarise(live_births = sum(live_birth == 1),
            total = n()) %>%
  ungroup()

# Separate group intramuscular and group vaginal counts
group_A <- group_summary %>% filter(group == "intramuscular")
group_B <- group_summary %>% filter(group == "vaginal")

print(group_summary)
```

```
## # A tibble: 2 x 3
##   group         live_births total
##   <chr>               <int> <int>
## 1 intramuscular         811  1988
## 2 vaginal               461  1025
```

```
# Two-proportion z-test
prop_test <- prop.test(x = c(group_A$live_births, group_B$live_births),
                       n = c(group_A$total, group_B$total),
                       alternative = "two.sided",
                       conf.level = 0.95)

tidy(prop_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      4.68  0.0306         1  -0.0799  -0.00373 2-sample t~
## # i 1 more variable: alternative <chr>
```

```
# Extract p-value and confidence interval from prop_test
result <- tidy(prop_test)

p_value <- result$p.value
conf_low <- result$conf.low
conf_high <- result$conf.high

# Interpret result
if (p_value < 0.05) {
  message("There is a statistically significant difference in live birth proportions between Group A anc
} else {
  message("There is no statistically significant difference in live birth proportions between Group A ar
}
```

```
## There is a statistically significant difference in live birth proportions between Group A and Group I
```

Interpretation: I reject the null hypothesis as the p-value is 0.031, which is less than the significance level = 0.05. This indicates a significant difference between the two groups.*

(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 (i.e., column `gene`) and a list of **unadjusted** p-values (i.e., column `pval`) generated from the analysis (the particular statistical test used is the **sequence kernel association 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.

> **Heads-up:** 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.**

Before proceeding, run the below code to load the dataset.

```
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
## # i 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**.

(Optional) Q5.1

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. **Answer in one sentence.**
- 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 $\alpha = 0.05$? **Answer in one sentence.**

ANSWER:

1. The total number of genes in the dataset is 1150 total_genes. This corresponds to the number of multiple comparisons that were performed.

2. The number of genes associated with the phenotype at the unadjusted = 0.05 is 100 significant_genes.

Provide the necessary code to support both answers.

```r
# Count the total number of genes
total_genes <- nrow(GWAS_results)
total_genes
```

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

```r
# Count how many genes have p-values less than 0.05
significant_genes <- GWAS_results %>%
  filter(pval < 0.05) %>%
  nrow()
significant_genes
```

```
## [1] 100
```

(Optional) Q5.2

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).

ANSWER:

Reporting only one "significant" test without acknowledging the other tests can be misleading because it increases the risk of Type I errors, otherwise known as false positives. In multiple testing, some associations may appear significant by chance, so failing to adjust for the number of tests conducted can lead to overestimating the evidence for an effect.

(Optional) Q5.3

Use the function `p.adjust()` to calculate adjusted p-values via the Bonferroni correction. How many and which genes are associated with the treatment after the adjustment, at the $\alpha = 0.05$ significance level? **Answer in one or two sentences.**

ANSWER:

After applying the Bonferroni correction, there are 2 genes associated with the treatment at the =0.05 significance level. The significant genes are: osm-1 & che-3.

Provide the necessary code to support your answer.

```r
# Apply Bonferroni correction
GWAS_results <- GWAS_results %>%
  mutate(pval_adjusted = p.adjust(pval, method = "bonferroni"))

# Filter for significant genes with adjusted p-value < 0.05
significant_bonferroni <- GWAS_results %>%
  filter(pval_adjusted < 0.05)
```

```
# Count the number of significant genes after Bonferroni adjustment
num_significant_genes <- nrow(significant_bonferroni)

significant_bonferroni
```

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

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