Medical Statistics – Lab 3

R version

Welcome to lab 3 in the medical statistics course. For today's exercises, we will continue exploring the `lowbwt.sav` dataset, which you can downloaded from the Dataset section of the menu.

```
library(haven)
library(dplyr)
library(ggplot2)
library(car)

# Load the dataset
lowbwt <- read_sav("lowbwt.sav")

# Convert all variables to factors where needed
lowbwt <- lowbwt %>% mutate(across(where(is.labelled), as_factor))
```

As a reminder, the dataset includes the following variables (see the previous lab for more details):

Variable	Abbreviation
Identification Code	ID
Low Birth Weight (0 = Birth Weight 2500g, 1 = Birth Weight < 2500g)	low
Age of the Mother in Years	age
Weight in Pounds at the Last Menstrual Period	lwt
Ethnicity of the mother (1 = Caucasian, 2 = Afro-American, 3 = Asian)	ethnicity
Smoking Status During Pregnancy (1 = Yes, 0 = No)	smoke

Variable	Abbreviation
History of Premature Labor (0 = None, 1 = One, etc.)	ptl
History of Hypertension (1 = Yes, 0 = No)	ht
Presence of Uterine Irritability (1 = Yes, 0 = No)	urirr
Number of Physician Visits During the First Trimester (0 = None, 1 = One, 2 = Two, etc.)	pvft
Birth Weight in Grams	bwt

Part 1: Independent Samples t-test and Mann-Whitney U Test

In this part of the lab, we will examine the effect of smoking during pregnancy on birth weight.

Exploratory data analysis

We will start by creating a boxplot to visualize the distribution of birth weights for mothers who smoked and those who did not.

```
# Create boxplot comparing birth weights for mothers who smoked and those who did not
ggplot(lowbwt, aes(x = smoke, y = bwt)) +
  geom_boxplot() +
  labs(x = "Smoking Status", y = "Birth Weight (grams)", title = "Birth Weight by Smoking Sta
```

> Question 1
>
> Based on the boxplot, do you expect the smoking status to have an effect on birth weight?

Independent samples t-test

Next, we will perform an independent samples t-test to compare the mean birth weights between mothers who smoked and those who did not. In R, this can be done using the `t.test()` function:

```r
# Perform independent samples t-test
independent_t_test <- t.test(bwt ~ smoke, data = lowbwt, var.equal = TRUE)

# Print results
print(independent_t_test)
```

Explanation:

- The first argument `bwt ~ smoke` specifies the formula for the test, indicating that we are comparing the birth weights (`bwt`) between the two groups defined by the `smoke` variable.
- The second argument `data = lowbwt` specifies the dataset.
- The argument `var.equal = TRUE` indicates that we are assuming equal variances in the two groups, meaning that the classical independent samples t-test is performed. If you suspect unequal variances, Welch's t-test can be conducted by setting `var.equal = FALSE`.

> Question 2
>
> Based on the results of the independent samples t-test, is there a statistically significant difference in birth weight between mothers who smoked and those who did not?

Checking of assumptions

To assess whether the assumption of normality holds for the outcome variable in both groups, we create the following plot:

```r
# Create histograms of birth weight by smoking status
ggplot(lowbwt, aes(x = bwt)) +
  geom_histogram(binwidth = 200, fill = "blue", colour="black", alpha = 0.7) +
  facet_wrap(~smoke) +
  labs(x = "Birth Weight (grams)", title = "Histograms of Birth Weight by Smoking Status")
```

> Question 3
>
> Do the histograms indicate that the birth weight data are approximately normally distributed for both groups?

We also need to check whether the assumption of a common population standard deviation holds. This can be done by performing the Levene test:

```r
# Perform Levene's test for homogeneity of variances
levene_test <- leveneTest(bwt ~ smoke, data = lowbwt)
levene_test
```

> Question 4
>
> Based on the Levene test, does the assumption of equal variances hold?

95% Confidence Interval for the mean difference

In addition to performing hypothesis tests, it is often informative to estimate the effect size and its uncertainty. One way to do this is by calculating a confidence interval for the mean difference in birth weight between the two groups. The 95% confidence interval is included in the default output of the `t.test()` function, so in principle we could extract it from there. As an exercise, we are also going to calculate it manually based on the formulas provided in the lecture/course syllabus.

To calculate the summary statistics required for the manual calculation, we use the following code:

```r
# Calculate summary statistics by smoking status
summary_stats <- lowbwt %>%
  group_by(smoke) %>%
  summarise(mean = mean(bwt), sd = sd(bwt), n = n())
summary_stats
```

> Exercise
>
> Based on these summary statistics, calculate the pooled standard deviation and the standard error of the mean difference. Then compute the 95% confidence interval for the mean difference in birth weight between mothers who smoked and those who did not. You may simplify the calculation by using the 97.5th percentile from the standard normal distribution (1.96) rather than the corresponding percentile from the t-distribution.

> Question 5
>
> Does your manually calculated 95% confidence interval for the mean difference in birth weight between the two groups agree with the one provided in the output of the `t.test()` function?

Mann-Whitney U Test

In case the assumptions of the independent samples t-test are violated, we can use the Mann-Whitney U test as a non-parametric alternative. This test can be performed in R using the `wilcox.test()` function:

```
# Perform Mann-Whitney U test
wilcoxon_test <- wilcox.test(bwt ~ smoke, data = lowbwt)
wilcoxon_test
```

> Question 6
>
> What are the null and alternative hypotheses for the Mann-Whitney U test, and what does the p-value indicate about the difference in birth weight between mothers who smoked and those who did not?

Part 2: One-Way ANOVA and Kruskal-Wallis Test

In this part of the lab, we are going to examine the effect of ethnicity on birth weight.

Explaratory data analysis

```
# Create boxplot comparing birth weights across ethnic groups
ggplot(lowbwt, aes(x = ethnicity, y = bwt)) +
  geom_boxplot() +
  labs(x = "Ethnicity", y = "Birth Weight (grams)", title = "Birth Weight by Ethnicity")
```

> Question 7
>
> What does the boxplot suggest about the distribution of birth weights across different ethnic groups?

One-way ANOVA

To test the null hypothesis that the mean birth weights are equal across all ethnic groups, we can perform a one-way ANOVA using the `aov()` function:

```r
# Perform One-Way ANOVA
anova_result <- aov(bwt ~ ethnicity, data = lowbwt)

# Print summary of ANOVA
summary(anova_result)
```

> Question 8
>
> What conclusions can be drawn from the results of the one-way ANOVA?

Post-hoc tests

If the one-way ANOVA indicates a statistically significant difference in birth weight across ethnic groups, we can perform Bonferroni-corrected post-hoc tests to determine which specific groups differ from each other. In R, this can be done using the `pairwise.t.test()` function:

```r
# Perform pairwise comparisons with Bonferroni correction
posthoc <- pairwise.t.test(lowbwt$bwt, lowbwt$ethnicity, p.adjust.method = "bonferroni")
posthoc
```

Explanation:

- The first argument of the `pairwise.t.test()` function specifies the outcome variable, which is the column of the lowbwt dataset that contains the birth weight values
- The second argument specifies the grouping variable, which is the column of the lowbwt dataset that contains the ethnicity values
- The `p.adjust.method = "bonferroni"` argument specifies that the p-values should be adjusted using the Bonferroni correction

> Question 9
>
> What conclusions can be drawn from the post-hoc comparisons?

Checking of assumptions

To assess whether the results of the one-way ANOVA are valid, we need to check the assumptions of normality and homogeneity of variances. This step is analogous to the previous examples, and is left as an exercise.

Kruskal-Wallis Test

If the assumptions of the one-way ANOVA are violated, we can use the Kruskal-Wallis test as a non-parametric alternative. The test can be performed in R using the `kruskal.test()` function:

```
# Perform Kruskal-Wallis test
kruskal_test <- kruskal.test(bwt ~ ethnicity, data = lowbwt)
kruskal_test
```

> Question 10
>
> Are the results of the Kruskal-Wallis test consistent with the one-way ANOVA results?

Part 3: Unguided exercises

Effect of hypertension on birth weight

Examine the effect of history of hypertension on birth weight by performing the following steps:

- Create a boxplot to visualize the distribution of birth weights by history of hypertension
- Perform an independent samples t-test to compare the mean birth weights between mothers with and without a history of hypertension
- Check the assumptions of the t-test, including normality and homogeneity of variances
- If the assumptions of the t-test are violated, perform a Mann-Whitney U test as a non-parametric alternative

Comparing red cell folate levels across ventilation strategies in cardiac bypass patients

Twenty-two patients undergoing cardiac bypass surgery were randomized to one of three ventilation groups:

- **Group I**: Received a 50% nitrous oxide and 50% oxygen mixture continuously for 24 hours
- **Group II**: Received a 50% nitrous oxide and 50% oxygen mixture only during the operation
- **Group III**: Received no nitrous oxide and a 35-50% oxygen mixture continuously for 24 hours

The data file `ex5_6.sav` contains the red cell folate levels for the three groups after 24 hours of ventilation. The aim of this study is to compare the three groups and test whether they have the same red cell folate levels.

Tasks

1. **Exploratory data analysis** Create a boxplot to visualize the distribution of red cell folate levels by ventilation group. Based on this plot:
    - What are your first conclusions regarding the means and variances of the different groups?
2. **Perform a one-way ANOVA**:
    - Interpret the results
    - Are the assumptions satisfied?
3. **Try a log transformation on the data**:
    - Perform another one-way ANOVA
    - Are the assumptions satisfied after the transformation?
4. **Determine which means differ**:
    - Which means do you think differ?
    - Explain your reasoning.
5. **Try a non-parametric approach**:
    - What are your conclusions from this method?