Estimation, confidence intervals and statistical testing

1 Incidence of pancreatitis after ERCP

We consider again the data from "Data_indo.csv".

1.1 Content of the dataset

These data come from a randomized trial designed to study the effect of an anti-inflammatory drug (indomethacin) on the incidence of pancreatitis after a medical procedure (ERCP) aiming at diagnosing problems with the liver, gallbladder, bile ducts and pancreas. The variables collected on each patient included in this study are as follows:

- **id**: subject id,
- **site**: study site (center): **1** = University of Michigan, **2** = Indiana University, **3** = University of Kentucky, **4** = Case Western,
- **age**: age in years,
- **risk**: risk score for post-ERCP pancreatitis,
- **gender**: male or female,
- **sod**: presence of sphincter of oddi dysfunction,
- **pep**: previous post-ERCP pancreatitis (PEP),
- **recpanc**: recurrent pancreatitis,
- **outcome**: outcome of post-ERCP pancreatitis,
- **status**: outpatient status,
- **type**: sphincter of Oddi dysfunction type/level - higher numbers are more severe,
- **rx**: treatment arm,
- **bleed**: a gastrointestinal bleed occurred.

Variables **sod**, **pep**, **recpanc** are known risk factors for pancreatitis and **bleed** is an adverse event of the intervention.

1.2 Some preliminary calculations

Assume we are interested in estimating the incidence of pancreatitis in untreated patients.

1. Specify the statistical model associated with this problem.
2. What is the parameter of interest?
3. Derive an estimator of this parameter by the maximum likelihood method.
4. What are the expected value and the variance of this estimator?
5. Use the central limit theorem to obtain a pivotal statistics and derive an approximated confidence interval for the parameter of interest at confidence level $1 - \alpha$.

1.3 Application

1. Import the data into **Rstudio** using the **tidyverse** library.
2. Use the following function to extract the observations corresponding to untreated patients

```
filter(dataframe.name,variable.name==condition)
```

3. Compute the estimation of the incidence of pancreatitis in untreated patients based on the previously derived estimator.
4. Compute the bounds of the confidence interval for the incidence of pancreatitis at confidence level 0.95.
5. Discuss the influence of the sample size on the estimator and on the confidence interval.
6. The disadvantage of the previous confidence interval is that it is calculated from an approximation that is only valid under certain conditions. Other methods exist to calculate confidence intervals for proportions. One of them is implemented in the R function **prop.test**.

    a. For that purpose, we need first to specify the reference modality (on which the proportion of interest is based) with the following instruction:

```
reordered_factor <- relevel(factor_name, ref='reference_level')
```

If this operation is not performed, the proportion will be calculated on the default reference modality which is not necessarily the one we are interested in. Here, this could lead to estimating a proportion of patients without pancreatitis rather than with.

    b. Then, we need to compute the contingency table for the factor of interest and pass it as an argument to the **prop.test** function:

```
prop.test(table(dataframe.name$factor.name),conf.level = ...)
```

7. Compare the estimates and confidence intervals obtained by hand and with the **prop.test** function.
8. Change the value of the confidence level. What are the effects of changing the confidence level on the parameter estimate and on the confidence interval?
9. By repeating the previous steps, estimate the incidence of pancreatitis in treated patients and provide a confidence interval at confidence level 0.95. Compare with the ones obtained on untreated patients and comment on the results.

1.4 Comparison between groups

We wish to compare the average risk (variable **risk**) between the two treatment arms (variable **rx**).

1. Specify the statistical model associated with this problem.
2. What are the parameters of interest?
3. Formulate the hypotheses H_0 and H_1 of the statistical test that would answer the question.

The appropriate test statistic here depends on whether the variance is the same in the two groups being compared or not. Therefore, a test is first performed to compare variances in both groups with the following command:

```
var.test(indo_rct$risk~indo_rct$rx)
```

4. Formulate the hypotheses H_0 and H_1 of this preliminary test.
5. State your conclusion based on the outputs obtained.

Means comparison is performed using the command below:

```
t.test(indo_rct$risk~indo_rct$rx,var.equal=T)
```

6. What is the purpose of the option **var.equal=T**?
7. What is the conclusion of the test?

2 Poisson distribution : maximum likelihood estimation

We are interested in the **epil** dataset from the **MASS** library in which seizure counts are reported in patients suffering from epilepsy. Among the patients followed, some receive a treatment and the others receive a placebo (variable **trt**). To assess the benefit of treatment, the total number of seizures per patient over an 2-week period is counted (variable **y**).

In this exercise, we will analyze variable **y** by assuming that the number of seizures per patient follows a Poisson distribution. The Poisson distribution is a discrete probability distribution taking all non-negative integer values with the following probabilities:

$$P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!},\ k \in \mathbf{N}$$

Here, $\lambda > 0$ is the parameter of the distribution. The expected value and the variance of such distribution are both given by λ.

1. Create a function **LV** that computes the log-likelihood in λ of a sample from a Poisson distribution $\mathcal{P}(\lambda)$.

Indications:

i. Function **dpois** can be used to evaluate the probabilities of each count value for any given λ.

ii. To create a function, use the following command types:

```
name <- function(arg_1, arg_2, ...) expression
```

where **name** is the name of the created function **arg__1**, **arg__2**, ... of the function and **expression** is an **R** expression, that uses the arguments, **arg__i**, to calculate the value to be returned by the function.

2. Calculate $LV(\lambda)$ for $\lambda = 1, 5, 8$ in treated patients. Which one to choose?
3. Calculate the maximum likelihood of the mean seizure count in treated patients. The **optimize** function can be used (see the help for precisions).
4. Compare the result with the mean seizure count computed on the the treated patients subsample.