## Project C: Post-weaning diarrhea I

Post-weaning diarrhea (PWD) is a worldwide economically important disease in pigs in piggeries. The disease is characterised by increased mortality, weight loss, retarded growth, increased treatment costs and higher use of antibiotics. Enterotoxigenic Escherichia coli is considered to be the most important cause of the disease.

Currently the disease is often controlled by using antimicrobials, but the emergence of antimicrobial resistence in E. coli urges the need for alternative control strategies. For example, inclusion of additional dietary fiber and reduction of crude protein levels, but also the addition of zinc oxide (ZnO) has been demonstrated to have beneficial effects. However, by 2022 this zinc may no longer be used (EU legislation).

Another strategy is to vaccinate the piglets. In this study we are interested in the effect of vaccination as compared to the addidition of ZnO and nutraceuticals (e.g. fibers) to the feed. In particular the following treatments are considered (in the dataset this variables is names Treatment):

A: normal feed + ZnO

B: normal feed + nutraceuticals

C: vaccination + high energy/protein in phases 2 and 3 (time periods)

D: vaccination + high energy/protein in phases 1, 2 and 3

E: vaccination + high energy/protein in phases 1, 2 and 3 + nutraceutics

We are interested in the following outcomes:

ADWG0021, ADWG2150, ADWG0050: average daily weight gain (g/day) in the period between 0 and 21 days post-weaning, between day 21 and day 50 post-weaning and in the period between 0 and 50 days post-weaning, repectively.

A few details about the design of the study:

* Piglets live in pens (≈ cages) (16 piglets in one pen).
* We only have the total weights of the piglets living together in a pen.
* In this study each treatment group consists of 128 piglets (thus 8 pens of 16 piglets).
* The five treatments were randomised over the the 5×8=40 pens

. The data can be read as shown in the next chunck of R code: load("PWD.RData")

## Data reading

Read in the dataset and analyze the skim through the data variables.

## Descriptive statistics

* Summarize your data and calculate the following: mean, median, minimum, maximum, first and third quartile (for each variable).
* For the categorical variable existing, calculate a frequency table
* Calculate the correlation coefficient (ADWG0021 and ADWG2150*)* and (ADWG0021and ADWG0050)

## Graphics

* Generate a bar chart of a categorical variable for the gender (*Sex* parameter).
* Generate a bar chart graph with mean ADWG0021in males and females
* Make a histogram of a continuous variable: “ADWG2150” as well as “ADWG0021”.
* Make a scatterplot of 2 continuous variables ADWG0050 and ADWG0021, and add the regression lines for each gender
* Make a boxplot of ADWG0021in and a separate boxplots per *Treatment* (as.factors).

## Outlier detection

* Explore the data for any existing outliers, identify them (do NOT remove them if found).
* What do you think?

## Testing for normality/ homoscedasticity (for all features mentioned above)

* Check the normality using two methods
* Check the homoscedasticity using two methods.
* What do you think?

1. Statistical Inference

* Calculate the 90%, 95%, 99% confidence interval for the means of ADWG0021per each *gender*.
* How would you describe those inferences and what do you observe in terms of the interval width when request higher confidence (i.e. 99% C.I.)?

1. Hypothesis testing

* We hypothesis that ADWG0021is different between male vs female. Assuming normality and homoscedasticity, can you test this hypothesis using statistical hypothesis framework
* Assess whether the previous test assumptions have been meet for the test.
* We hypothesis that ADWG0021is “different” in the group receiving *Treatment A (*normal feed + ZnO)compared to the Treatment B (normal feed + nutraceuticals). Can you test this hypothesis assuming heteroscedasiticy
* Assess the previous test assumption
* We hypothesis that ADWG0021is different between the different Treatments . Can you perform comparison between the different groups, after assessing the assumptions and performing post-hoc testing (assuming normality and homoscedasticity).

1. Linear model

* Fit a linear regression to the data and interpret the regression coefficient (for the one of the hypotheses mentioned above)
* Calculate and interpret a 95% confidence interval of the regression slope , (bonus)
* Estimating the average ADWG0021 change for with changing the gender from 1 to 2 (bonus).