Prelim-Data Analysis Semester Project

AUTHOR
Olaitan Comfort Shekoni

Data Description

The dataset used for this analysis is from the second study of my MSc. thesis titled "Evaluation of the Efficacy of Bovine Adenovirus-Vectored Avian Influenza Vaccine in Poultry,". This study investigates how different mucosal vaccine routes influence protection against avian influenza. For this preliminary analysis, I used qPCR viral-load data **(Ct values and log10 genomic equivalents)** from tracheal swabs collected at three time points—2, 4, and 6 days post-challenge (DPC2, DPC4, DPC6)—in chickens that received a single dose of the BAdV-H5HA+H7NP vaccine ($1x10^8$pfu) in a prime-booster dose vaccine administration via two routes —**intraocular (IO)** and **intramuscular (IM)**—and challenged with two avian influenza virus strains (**H5N1** and **H7N2**).

Each Excel sheet corresponds to one **Virus × Timepoint** combination (e.g., DPC2-H5N1).
Key variables:

- **Vaccine_Group**: Mock, Mock Challenge, Empty-Vector, BAds-AIV
- **Route**: IM (intramuscular) or IO (intraocular)
- **Bird_ID**: unique sample identifier
- **Ct**: qPCR cycle threshold (continuous)
- **log10GE**: $\log_{10}$ genome equivalents/mL (continuous)

Since "Mock" birds were not challenged, they are excluded.
Data are nested: **Bird_IDs** are nested within **Vaccine_Group × Route** combinations, and each sheet (timepoint) is nested within each **Virus**.

```
#LOADING LIBRARIES
library(tidyverse)
```

```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
```

```
✔ dplyr     1.1.4     ✔ readr     2.1.5
✔ forcats   1.0.0     ✔ stringr   1.5.1
✔ ggplot2   4.0.0     ✔ tibble    3.3.0
✔ lubridate 1.9.4     ✔ tidyr     1.3.1
✔ purrr     1.1.0
── Conflicts ─────────────────────────────────────────
tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>)
to force all conflicts to become errors
```

```
library(stringr)
library(readxl)
library(dplyr)
library(tidyr)
library(ggplot2)
library(readr)   # for parse_number()
library(purrr)
library(broom)
library(multcomp) # for Tukey
```

```
Loading required package: mvtnorm
Loading required package: survival
Loading required package: TH.data
Loading required package: MASS

Attaching package: 'MASS'

The following object is masked from 'package:dplyr':

    select


Attaching package: 'TH.data'

The following object is masked from 'package:MASS':

    geyser
```

```
library(car)
```

```
Loading required package: carData
```

```
Attaching package: 'car'

The following object is masked from 'package:dplyr':

    recode

The following object is masked from 'package:purrr':

    some
```

```
library(emmeans)
```

```
Welcome to emmeans.
Caution: You lose important information if you filter this
package's results.
See '? untidy'
```

```
library(multcompView)
library(ggpubr)
library(glm2)
```

```
Attaching package: 'glm2'

The following object is masked from 'package:MASS':

    crabs

The following object is masked from 'package:survival':

    heart
```

```
library(glmertree)
```

```
Loading required package: lme4
Loading required package: Matrix

Attaching package: 'Matrix'

The following objects are masked from 'package:tidyr':

    expand, pack, unpack
```

```
Loading required package: partykit
Loading required package: grid
Loading required package: libcoin
```

```
library(glmmTMB)
```

```
Warning in check_dep_version(dep_pkg = "TMB"): package version
mismatch:
glmmTMB was built with TMB package version 1.9.17
Current TMB package version is 1.9.18
Please re-install glmmTMB from source or restore original
'TMB' package (see '?reinstalling' for more information)
```

```
library(lme4)
library(pscl)
```

```
Classes and Methods for R originally developed in the
Political Science Computational Laboratory
Department of Political Science
Stanford University (2002-2015),
by and under the direction of Simon Jackman.
hurdle and zeroinfl functions by Achim Zeileis.
```

```
library(ZIM)
library(TMB)
library(bbmle)
```

```
Loading required package: stats4

Attaching package: 'bbmle'

The following object is masked from 'package:dplyr':

    slice
```

```
library(DHARMa)
```

```
This is DHARMa 0.4.7. For overview type '?DHARMa'. For recent
changes, type news(package = 'DHARMa')
```

```
library(patchwork)
```

```
Attaching package: 'patchwork'

The following object is masked from 'package:MASS':

    area
```

```
# Loading dataset
# Importing my qPCR excel data file
excel_path <- path.expand("~/Desktop/Entomology tech-Fall 2025/
excel_path <- "ENT_Project_Tracheal_qPCR_Clean.xlsx .xlsx"

ENT_Project_Tracheal_qPCR_Clean_xlsx_ <- read_excel("ENT_Projec

#qPCR data wrangling process
excel_path
```

```
[1] "ENT_Project_Tracheal_qPCR_Clean.xlsx .xlsx"
```

```
#GETTING SHEET NAMES FROM THE CHOSEN FILE
sheets <- readxl::excel_sheets(excel_path)
#READING ALL SHEETS AND BINDING INTO ONE DATA FRAME
qpcr <- purrr::map_dfr(sheets, ~ readxl::read_excel(excel_path,
dplyr::glimpse(qpcr)        #quick peek into the selected data
```

```
Rows: 450
Columns: 8
$ Timepoint     <chr> "DPC2", "DPC2", "DPC2", "DPC2", "DPC2",
"DPC2", "DPC2", …
$ Virus         <chr> "H5N1", "H5N1", "H5N1", "H5N1", "H5N1",
"H5N1", "H5N1", …
$ Vaccine_Group <chr> "Mock", "Mock", "Mock", "Mock", "Mock",
"Mock", "Mock", …
$ Route         <chr> "NA", "NA", "NA", "NA", "NA", "NA",
"NA", "NA", "NA", "N…
$ Challenge     <chr> "No", "No", "No", "No", "No", "No",
"No", "No", "No", "N…
$ Ct            <dbl> 38.867, 0.000, 0.000, 0.000, 0.000,
0.000, 0.000, 0.000,…
$ log10GE       <dbl> 3.807920, 3.250908, 3.250908, 3.250908,
3.250908, 3.2509…
$ Bird_ID       <dbl> 712, 713, 714, 715, 716, 717, 718, 719,
```

```
720, 721, 722, 7…
```

```
# Checking counts per original sheet
qpcr %>% count(Timepoint, Virus)
```

```
# A tibble: 6 × 3
  Timepoint Virus     n
  <chr>     <chr> <int>
1 DPC2      H5N1     77
2 DPC2      H7N2     73
3 DPC4      H5N1     77
4 DPC4      H7N2     73
5 DPC6      H5N1     77
6 DPC6      H7N2     73
```

```
qpcr <- lapply(sheets, function(s) {
  df <- read_excel(excel_path, sheet = s)
  df$Sheet <- s
  return(df)
}) |> bind_rows()
```

```
# Clean and set factors
qpcr <- qpcr %>%
  filter(Vaccine_Group != "Mock") %>%
  mutate(
    Timepoint = factor(Timepoint, levels = c("DPC2","DPC4","DPC
    Virus = factor(Virus, levels = c("H5N1","H7N2")),
    Vaccine_Group = factor(Vaccine_Group, levels = c("Mock Chal
    Route = factor(Route, levels = c("IM","IO")),
    GroupRoute = interaction(Vaccine_Group, Route, sep = ":")
  )

glimpse(qpcr)
```

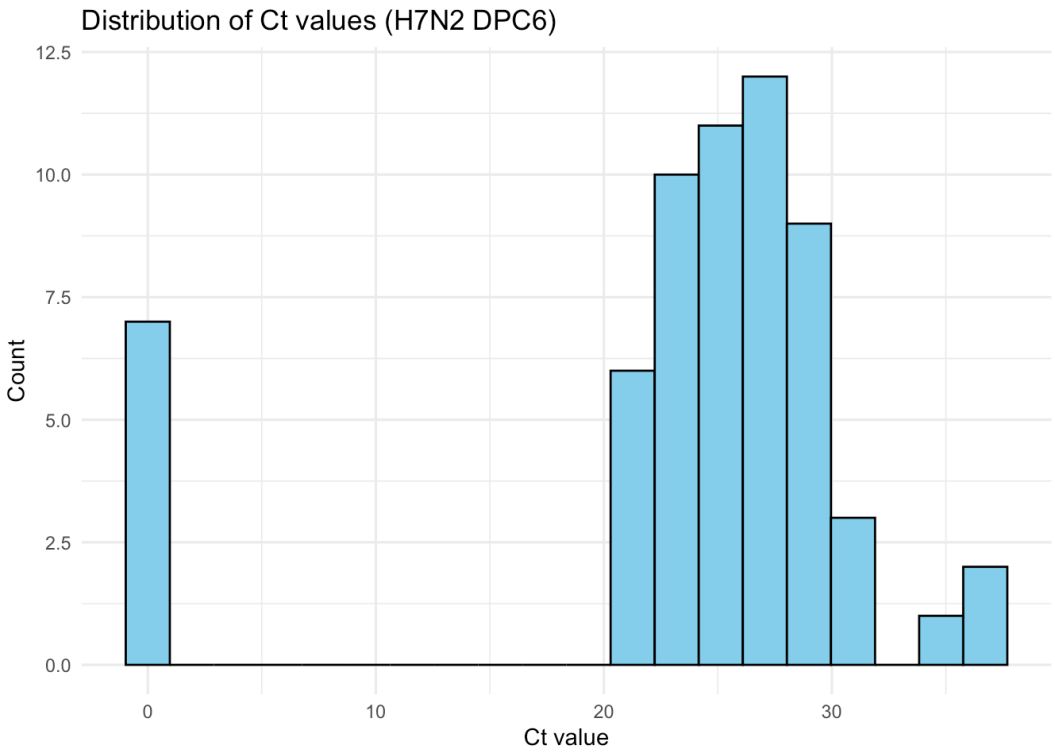
```
Rows: 378
Columns: 10
$ Timepoint     <fct> DPC2, DPC2, DPC2, DPC2, DPC2, DPC2,
DPC2, DPC2, DPC2, DP…
$ Virus         <fct> H5N1, H5N1, H5N1, H5N1, H5N1, H5N1,
H5N1, H5N1, H5N1, H5…
$ Vaccine_Group <fct> Mock Challenge, Mock Challenge, Mock
Challenge, Mock Cha…
```

```
$ Route         <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA, NA, NA, NA, …
$ Challenge     <chr> "Yes", "Yes", "Yes", "Yes", "Yes",
"Yes", "Yes", "Yes", …
$ Ct            <dbl> 28.294, 29.838, 30.844, 29.953, 29.932,
31.468, 28.799, …
$ log10GE       <dbl> 6.923950, 6.468909, 6.172425, 6.435017,
6.441206, 5.9885…
$ Bird_ID       <dbl> 775, 776, 777, 778, 779, 780, 781, 782,
783, 784, 785, 7…
$ Sheet         <chr> "DPC2-H5N1", "DPC2-H5N1", "DPC2-H5N1",
"DPC2-H5N1", "DPC…
$ GroupRoute    <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA, NA, NA, NA, …
```

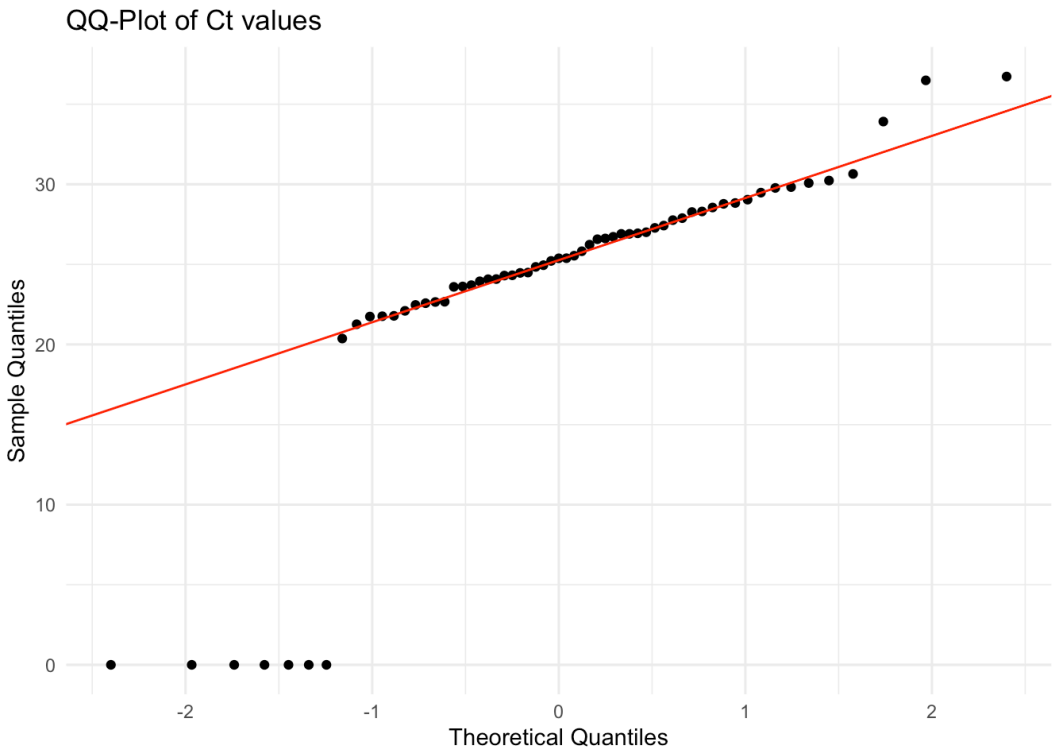
Checking Normality (Distribution)

Before model fitting, I checked whether **Ct** and **log10GE** are approximately normally distributed.

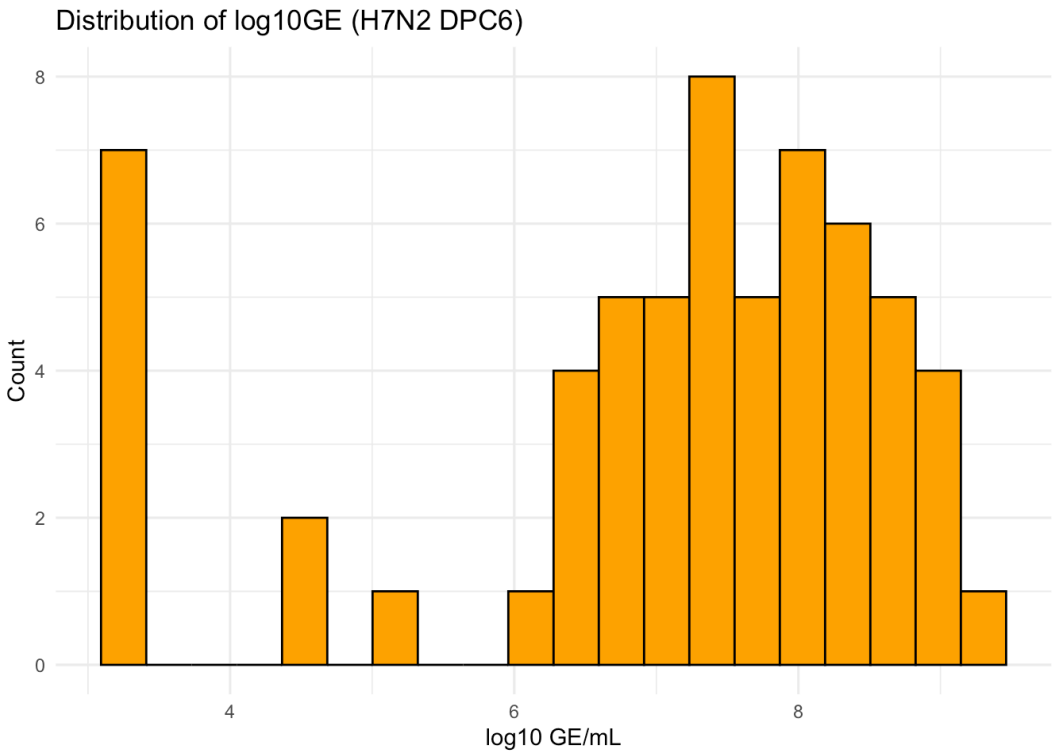
```
# Example: visualize H7N2 DPC6 data only

panel <- qpcr %>%
  filter(Virus == "H7N2", Timepoint == "DPC6")

# Histogram and QQ plot for Ct

ggplot(panel, aes(x = Ct)) +
  geom_histogram(bins = 20, fill = "skyblue", color = "black")
  labs(title = "Distribution of Ct values (H7N2 DPC6)", x = "Ct
  theme_minimal()
```



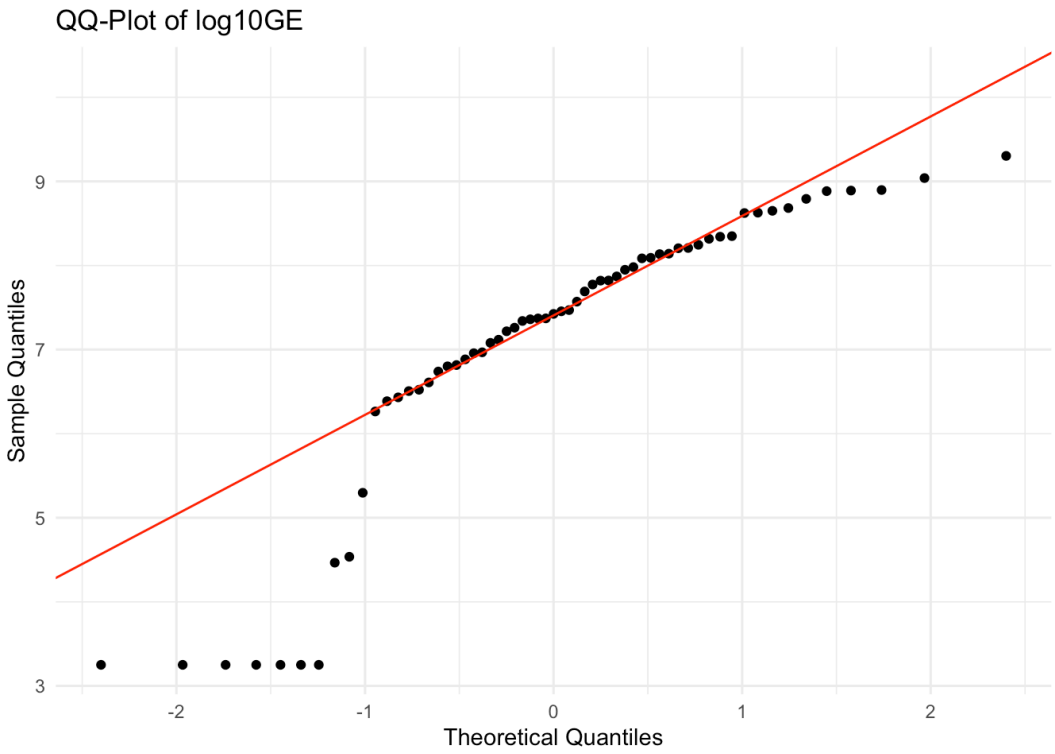

```
ggplot(panel, aes(sample = Ct)) +
  stat_qq() + stat_qq_line(color = "red") +
  labs(title = "QQ-Plot of Ct values", x = "Theoretical Quantil
  theme_minimal()
```




```
# Histogram and QQ plot for log10GE

ggplot(panel, aes(x = log10GE)) +
  geom_histogram(bins = 20, fill = "orange", color = "black") +
  labs(title = "Distribution of log10GE (H7N2 DPC6)", x = "log1
  theme_minimal()
```




```
ggplot(panel, aes(sample = log10GE)) +
  stat_qq() + stat_qq_line(color = "red") +
  labs(title = "QQ-Plot of log10GE", x = "Theoretical Quantiles
  theme_minimal()
```



Interpretation:

The histograms show approximately continuous, right-skewed distributions—typical of qPCR data. log10GE values are closer to normal than raw Ct values but still slightly skewed.

Model Structure and Selection Rationale

Data Nesting: Each measurement is **nested** as follows:

Nesting structure: Bird_ID ⊂ (Vaccine_Group × Route) ⊂ Timepoint ⊂ Virus (i.e., multiple birds belong to each treatment group (combination of vaccine and route), within each virus and timepoint.

Example model formula (for next stage)

Continuous response (log10GE) ~ fixed effects (Vaccine_Group, Route) + random effects (Bird_ID nested within Timepoint). Not yet running this model; just specifying for rationale

model_example <- “lmer(log10GE ~ Vaccine_Group * Route + (1 | Timepoint/Bird_ID), data = qpcr)” model_example

Chosen Model

Since my data contain repeated measurements of viral load (Ct and log10GE) across multiple timepoints, birds, and routes, a **nested or mixed-effects model** is appropriate.

The hierarchical structure (*Bird_ID nested within Timepoint within Virus*) requires random effects to account for correlation. Residuals for log10GE appear approximately normal, so a **Linear Mixed Model (LMM)** with Gaussian error is suitable. So, If future analysis shows strong skewness or heteroscedasticity, a **Tweedie GLMM** will be considered to model zero-inflated or right-skewed data. Thus, I will proceed using a **Gaussian LMM** framework with fixed effects for *Vaccine_Group*, *Route*, and *Virus*, and random effects for *Timepoint* and *Bird_ID*.