20074027 STAT2003 TASK 2

Bruce Odek Omondi

# QUESTION A

1. Using an effective experimental design, this experiment aimed to study and evaluate the antiviral effect of different combinations of drugs (the 5 drugs in the study) against a virus (HSV-1), AND to detect any interactions among the different drug components.
2. The experimental factors are each of the 5 drugs(A,B,C,D,E): A - Alpha, B - beta, C - gamma (Interferon drugs) D - Ribavirin, E - Acyclovir (Chemical drugs)
3. There are 3 levels for each experimental factor, i.e., each drug. Each one has a different purpose: High (+1) - maximum dosage from pilot study, Middle (0) - 32 times diluted from high dosage Low (-1) - no drug

* Having these 3 levels allows the researchers to capture and estimate not only the linear and quadratic effects, but also the interaction effect among the drug combinations better than a 2-level design, and this is crucial for the purpose of understanding how the drugs work together.

1. The treatments are the combinations of the different drugs at different dosage levels as shown in each run. The researchers did this using a composite design where:

* 16 treatments have 3 levels encoded as -1 and 1 to match their dosage level, allowing for estimation of linear and interaction effects.
* 18 treatments have 3 levels encoded as 0,-1, 1 to match their dosage level. This allowed estimation of linear, quadratic, and interaction effects.
* So, there were 34 total treatments (runs) applied to the infected cells.

1. The response is the readout, i.e., the percentage of cells infected with the virus. A lower response value indicates better outcome in the drug trial, i.e., the drugs are more effective at reducing the viral infection.
2. The basic principles of experiment design used in this study are:
3. Blocking - although implicit, the use of a single batch for all experiments to reduce variation from different batches indicates the implementation of blocking
4. Replication - each treatment was carried out in 2 separate runs to account for variability
5. Randomization - the 2 independent researchers used random orders AND the 34 runs were randomly assigned to the wells in the plates to avoid bias
6. Control - the study mentioned a run where there was viral infection but no drug application, which can be used for comparison, indicating the implementation of a control in the experiment

# QUESTION B

## Part 1: Analysis for 16-run fractional factorial design only [main response]

### Design and its properties

The design used is a **folded-over version of a two-level fractional factorial design**. I started with 16 runs and Resolution V half-fraction with generator E=ABCD. The factors are each of the 5 drugs (A–E) each tested at two dosage levels: **–1 (no drug)** and **+1 (high dose)** allowing for estimation of all main effects and two-factor interactions.

The design was **saturated**, as it left no degrees of freedom for error estimation, nor p-values to check the estimate significance, so I used a fold-over design that added 16 additional runs, reversing the **signs of all five factorial levels** to create a geometric mirror image of the original 16-run design. This resulted in a **32-run design** that now resolved aliasing of two-factor interactions, provided **16 residual degrees of freedom**, output the estimate p-values, and allowed for improved precision.

### Analysis

I analyzed the response variable as the average of the two replicates i.e., **y = avg\_readout**. The model fitted was a **full second-order interaction model**, which gave me the estimates for the 16-runs but as mentioned, the design was **saturated**, so not much else.

I decided to use **fold-over** where I got an additional 16 runs from flipping all of the factorial levels and simulating the avg\_readouts for these additional runs making sure the simulated data was in line with the mean and std.dev of the original data. I then defined a function to plot the **half-normal plots of effects** to visualize the absolute size of the effects and identify significant **factors**.

### Interpretation of Results

1. **Significant Main Effects**: **D** had a large negative coefficient, indicating good antiviral effect; **B** and **C** showed moderate main effects, suggesting moderate antiviral effect.
2. **Important Interactions**: **DE** - strong positive interaction indicating synergy and good antiviral effect; **BC** - suggested potential interactions among the Interferon drugs.

**\*\*R2 value** (74.7) also showed that the model was a good fit for the data.

**\*\* Replicate effect** was negligible, confirming reproducibility across researchers.

1. **Half-Normal Plot**: Clearly separated effects from noise. Effects like D, DE, and BC stood out from the fitted line, confirming their significance.
2. **Diagnostic plots**: showed good residual normality and no major heteroscedasticity.

### Limitations

The 2-level design is restricted to only the main and interaction effects, so **no quadratic effects can be estimated**. It also **does not explore mid-dose or extreme concentration combination**s. Lastly, although more robust than the 16-run model, it still **assumes that 3-factor interactions are negligible**.

## Part 2: Analysis to replicate estimates a, b, and c in Table III using all 34 runs (16-run & 18-run design) [bonus marks]

### Design and its properties

The study implemented a **34-run composite experimental design** composed of:

1. A **16-run two-level fractional factorial design (Resolution V)** defined by generator E= ABCD, enabling the estimation of all main effects and two-factor interactions.
2. An **18-run three-level orthogonal array**, which allows the estimation of linear, quadratic, and some interaction effects.

Each of the five drugs (A to E) was tested at **three levels**: -1 **(no drug),** 0 **(mid-dose),** and +1 **(high dose).** This combined design had enough degrees of freedom for estimating a full **second-order model** with 25 parameters (5 linear, 5 quadratic, 10 interactions, 1 intercept, and 1 replicate effect), while still being small enough to run within one batch of biological samples. Having two **independent replicates** also enhanced reliability and enabled modelling of any variation between researchers.

### Analysis & Interpretation of Results

I ran the same **second-order** model used in Section 3.1 of the study on different versions of the readout: 1) y= avg\_readout values, 2) y=sqrt of avg\_readout values, and 3) y=sqrt of avg\_readout values **and** outlier removed. In all 3 instances, the model gave insight into the main, interaction, and quadratic effects, as well as the output of the p-values to determine the significant effect estimates.

From the outputs:

1. **Strongest linear effects**: **D, E -** highly significant; **B -** moderately significant; **C -** marginally significant
2. **Quadratic effects**: **D2 -** suggests diminishing returns or overuse toxicity; **E2 -** suggests increased efficacy with higher doses up to a certain point
3. **Key interaction effects**: **DE -** synergistic effect between Ribavirin and Acyclovir (chemical drugs); **AB**, **AC:** moderate interactions within Interferon group

**\*\* Replicate effect** was insignificant, suggesting no major difference between researchers.

**\*\* Outlier (Run 14, Rep 1)** was identified due to high influence and discrepancy and was removed for modelling the estimates in c. Removing the outlier improved model assumptions and precision.

These results aligned with the findings in the study - chemical antiviral drugs (D and E) are highly effective, while Interferons have moderate or synergistic effects. The observed quadratic and interaction terms highlighted the complex, nonlinear nature of antiviral drug combinations. It also showed that **the 2 groups did not have much interaction** between them but **had intra-group interactions** to **moderate effect for the Interferon** drugs, and to **a good positive effect for the chemical** drugs.

### Limitations

Firstly, **model A had persistent residual skewness**, which required Box-Cox (sqrt) transformation. Additionally, model C improved fit but **there was still variability in the residuals** and the square root transformation may not always clear that variability. Lastly, the **composite design used is not fully orthogonal**, hence some aliasing may remain.

# QUESTION C

## Alternative design to 16-run fold-over

An alternative design to the fold-over 16-run design is a **Central Composite Design (CCD)** which is ideal for modelling **curvature**, estimating **quadratic effects** (e.g., diminishing or accelerating effects at high doses) unlike 2-level fold-overs, and detecting **optimum response regions,** hence better identifies optimal drug concentrations for minimum infection.

The CCD builds on the existing factorial setup and supports **second-order polynomial** modelling, adds **star points** to estimate quadratic effects and **center points** to detect non-linearity, and also **independently estimates experimental error**. It can also be **constructed to maintain** **rotatability**, such that variance of prediction is uniformly spread. Lastly, CCD can be made with less total runs, making it **more efficient** and at the same time **more powerful**.

### Assumptions and Considerations

CCD **assumes more than two levels** for drug trial doses, which is more in line with most controlled lab experiments. It also **assumes that higher-order (3+ factor) interactions are negligible**. Lastly, **the center points enable a robust check for lack-of-fit** and give **more replication for estimation of pure error**.

### Structure and Generation of the CCD

1. **Factorial core**: Use the same 16-run fractional factorial as the core with Resolution V
2. **Star points**: Add 10 star runs, where one factor is set to ±α (with others at 0) to estimate curvature
3. **Center points**: Include 5–6 runs with all five drugs at mid-dose level (coded 0) to estimate pure error and detect curvature.

**\* Total runs**: 16 (factorial), 10 (star), 5–6 (center) = **~31–32 total runs.** (matches the 16-run fold-over but estimates a lot more useful information)

**\*\*** It is possible to generate CCD designs directly in R using rsm library **(package used for Response Surface Methodologies)** to define the design.