Bacteria Drying

NCSU ST 542 Consulting Project

Bruce Campbell

Bacteria Drying

Scientists study bacteria for a variety of reasons. Bacteria have been the source of great human suffering and more recently it has become known that bacteria are also an important part of human health. In addition to matters of human health and disease, scientists study bacteria as forms of alternative energy and bio-remediation. The ability to store bacteria for long periods of time is an important research objective. Scientists need to be able to keep libraries of bacteria to compare bacterial populations over time and to share bacteria among colleagues. For alternative energy purposes, if bacteria can reliably be dried and re-hydrated then growth plants do not have to be collocated with productions plants. This has the potential to effectively make bacteria a stored energy source. For bio-remediation if bacteria can be reliably re-hydrated on site then deployment becomes easier as storing and transporting the bacteria in a dehydrated state is more cost effective.

This study aims to evaluate a variety of bacterial drying methodologies to determine optimal conditions for re-hydration. The specific aim is to have mechanism capable of producing ethanol/acetate at rates higher than 20% of undried bacteria. Variables used in the drying process include the addition of chemicals, different mechanisms for removal of oxygen, and different storage conditions. Samples from the same bacterial population are exposed to different drying mechanisms, are re-hydrated, and then measured for anaerobic respiration. The population surviving the drying process is measured through the amount of carbon monoxide produced after re-hydration. Increasing numbers of bacteria that survive the drying process lead to decreasing carbon monoxide levels. Since this measurement is the sum of many small contributions we expect the response to be normally distributed. The overall goals are to find the conditions which minimize the carbon monoxide levels - thus providing optimal conditions for bacteria re-hydration.

Methods

Method introduction

The bacteria drying experiments are carried out carefully in a traditional biological laboratory. Bacteria are grown in culture media prior to the experiments and samples from the same media are used in the different experiments for drying and storage. After storage bacteria are re-hydrated in individual tubes which are first evacuated with argon and then infused with

hydrogen and carbon dioxide to enable respiration. Treatments are grouped into experiments and carried out by different individuals. There are two phases to the experiment - a drying phase and a re-hydration phase. The treatment could be present in the drying phase or the re-hydration phase. The design is both qualitative and quantitative in the treatments applied.

Vacuum exposure is known to cause damage to bacteria cell membranes and that there is a repair mechanism for eliminating the damage(Frankenberg-Schwager et al. 1975). This study noted that the damage can be repaired after storage when the temperature is raised. It's also know that freeze-drying can cause genetic injuries in bacteria with the number of mutations increasing with the length of the storage time (Ashwood-Smith 1980). These are just some of the underlying factors that the experiments will be implicitly measuring.

Detailed notes are included with the experiments indicating any anomalies in them, and describing the lab procedures and materials used. Bacteria are grown from 7 cultures which are used in the experiments. Each sample is a unit of a cell formulation which is dried, stored, and then re-hydrated. Media is spun in a Beckman, and then rinsed with 20 mL from 25 mL bottle of 1YCM PCG with 1 mL Cys and 0.2 M NaOH. After that the sample is tube centrifuged in rotana for 15min at 4C and 6000rcf. After drying tubes are tubes flushed with argon at 20 psi for 25 minutes and then stored.

An experiment generally contains 10-25 formulations. In all there are 43 experiments with 775 samples. 73 samples have missing responses which are redacted from the analysis. The measures for this study are the experiment ID's which encode the conditions for the experiment and response conditions of the outcome. The responses are measures of cellular respiration; H2,O2,N2,CO CO2.

Statistical Analysis

We will be investigating a several approaches to analyzing this data. The freest step will be to implement a 1 way ANOVA on the experiment and treatment. This will determine if the mean responses of the experiment are different. The model we use is that the repose y_{ij} of replicate i in the j is expressed as $y_{i,j} = \mu + \tau_j + \varepsilon_{i,j}$ where $i = 1, \ldots, I$ an index over the replicates $j = 1, \ldots, J$ is the index over experiments, μ_j is the mean of the CO measurements for experiment j,μ is the overall mean CO, and τ_j is the treatment effect for experiment j - a deviation from the overall mean. The errors $\epsilon_{i\,j} \sim N(0, \sigma^2)$ are normally distributed with equal variance.

The response is assumed to be normal but we perform a non parametric Kruskal-Wallis (Kruskal and Wallis 1952) one-way analysis of variance by ranks for validation of the one way ANOVA, and as an precursor to selecting the optimal experimental conditions. The Kruskall-Wallis analysis of variance is an extension of the Mann-Whitney U test to multiple groups. We expect the Kruskal-Wallis result will be consistent with the usual 1 way ANOVA. The assumption of the Kruskal-Wallis is that all the groups come from identical distributions with common variance. The null hypothesis is the medians are all the same for the groups. Using the notation above, the hypothesis are

$$H_0 : P(\mu_j > \mu_k) = \frac{1}{2} \ \forall \ j, k \in 1, \ldots, J$$

$$H_a : \exists \ j, k \ \ \in 1, \ldots, J : P(\mu_j > \mu_k) \neq \frac{1}{2}$$

A significant Kruskal-Wallis test indicates that at least one experimental median stochastically dominates the others. We do not know which one. The common options for determining stochastic order are to perform pairwise Mann-Whitney tests or to perform Dunn's Multiple Comparison Test (Dunn 1961). Dunn's test has several technical advantages over running pairwise Mann-Whitney tests. The main advantages are that it uses the pooled variance estimate from the Kruskal-Wallis test, and that it reuses the average rank scores from the Kruskal-Wallis test. Appendix A [REVISIT] contains a detailed description of the Kruskal-Wallis and Dunn tests.

The input to the Dunn test is the per group average rank $\bar{R}_j = \frac{R_j}{n_j}$ and the z-test statistic for groups j, k is

$$z_{j,k} = \frac{\bar{R}_j - \bar{R}_k}{\sigma_{j,k}}$$

$\sigma_{j,k}$ is a funtion of the overall, group, and rank tie counts (see appendix A). The null hypothesis of the group pairwise comparison is that the probability of observing a randomly selected value from the first group that is larger than a randomly selected value from the second groups is one half - $H_0 : P(X_j > X_k) = \frac{1}{2}$.

The null hypothesis for each pairwise comparison is that the probability of observing a randomly selected value from the first group that is larger than a randomly selected value from the second group equals one half; this null hypothesis corresponds to that of the Wilcoxon-Mann-Whitney rank-sum test.

The Kruskal-Wallis is implemented in R in the `kruskal.test` function in the core `stats` package. The Dunn test is implemented in the `dunn.test` package. R version 3.4.2 (2017-09-28) is used to perform the analysis.

Data Description and Preparation

An excel workbook is kept with the experimental data and other lab notes. This consists of 63 worksheets which are pre-processed from the lab workbook into a series of files suitable for ingest into an R data frame object. An Excel VBA script is run to extract the individual experiments into csv files and the resulting files were investigated for inclusion in the data analysis. Not all of the worksheet in the data workbook contain experimental data. After looking at each file - there were 44 experiments eligible for inclusion in the data analysis. Of these 44 one was determined to not have any response measurements and was redacted from the statistical analysis. The VBA code for extracing the worksheets is listed in Appendix A

Create master data frame from experiment files.

```
setwd("c:/e/brucebcampbell-git/ncsu-consulting-project/")
wd <- getwd()
nwd <- paste(wd, "data/r-load-data/", sep = "/")
setwd(nwd)
file.list <- list.files(nwd)

header.list <- list()
# Inspect the Headers
for (i in 1:length(file.list)) {
    file.name <- file.list[[i]]
    experiment.name <- tools::file_path_sans_ext(file.name)
    df <- read.csv(paste(nwd, file.name, sep = "/"))
    header.list[[i]] <- names(df)
}

str.list <- list()
for (i in 1:length(header.list)) {
    file.name <- file.list[[i]]
    str.wr <- paste(as.character(c(file.name, header.list[[i]])), collapse = ", ")
    str.list[[i]] <- str.wr
}

dfmaster <- data.frame()
for (i in 1:length(file.list)) {
    file.name <- file.list[[i]]
    # experiment.name <- strsplit(as.character(file.name),split = '.')[[2]]
    experiment.name <- tools::file_path_sans_ext(file.name)
    df <- read.csv(paste(nwd, file.name, sep = "/"))
    df$experiment <- rep(experiment.name, nrow(df))
    df$experiment.count <- rep(nrow(df), nrow(df))
    df$experiment.mean <- rep(mean(df$CO, na.rm = TRUE), nrow(df))
    names(df)

    if ("Formulation.pH" %in% colnames(df)) {
        cat(file.name)
    } else {
        df$Formulation.pH <- NULL
    }


    if ("Formulation" %in% colnames(df)) {
        cat(file.name)
    } else {
```

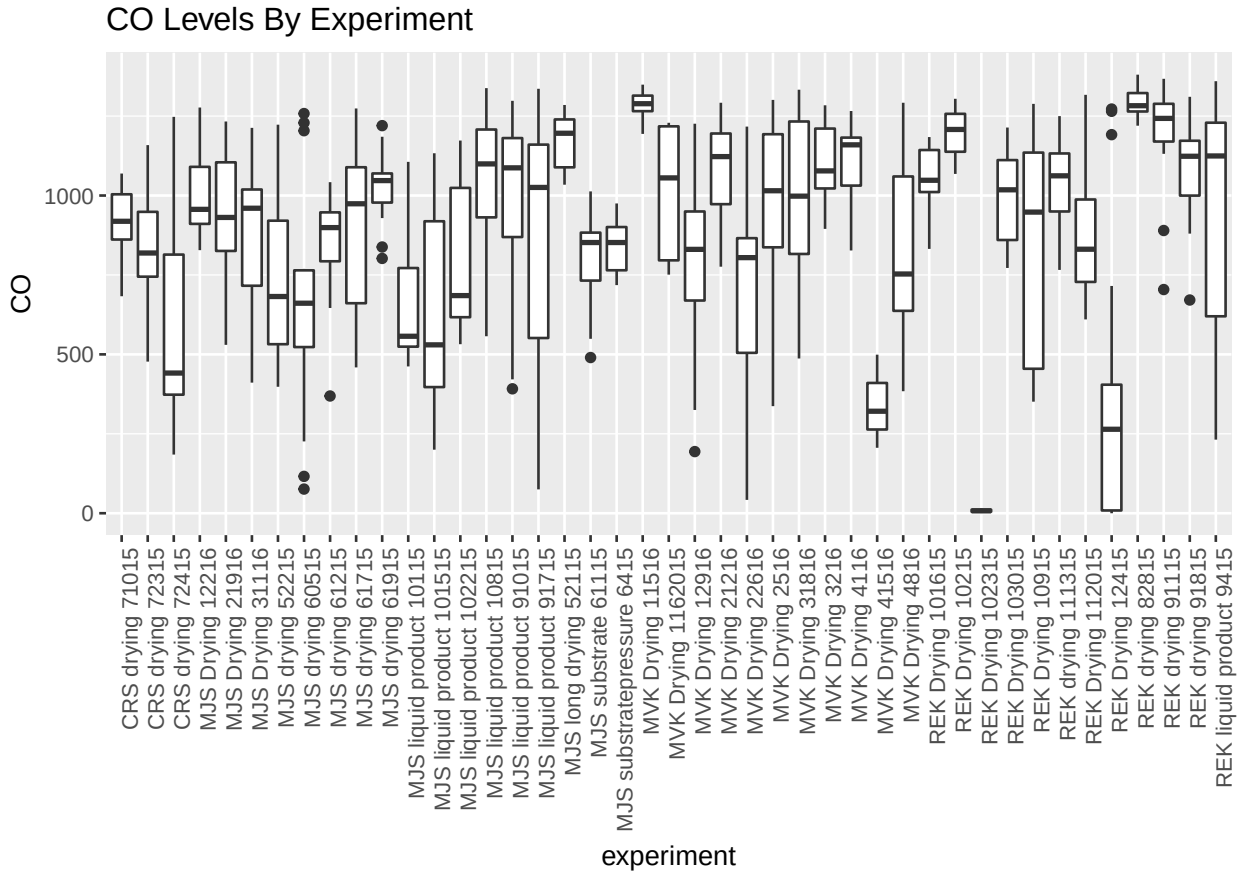
```
        df$Formulation <- "NA"
    }


    names.to.keep <- c("experiment", "Formulation", "CO", "experiment.count",
        "experiment.mean")

    df.keep <- df[, names.to.keep]
    df.keep$CO <- as.numeric(df.keep$CO)
    # plot(df.keep$CO, main=experiment.name)

    sum(is.na(df.keep$CO))
    dfmaster <- rbind(df.keep, dfmaster)
}
```

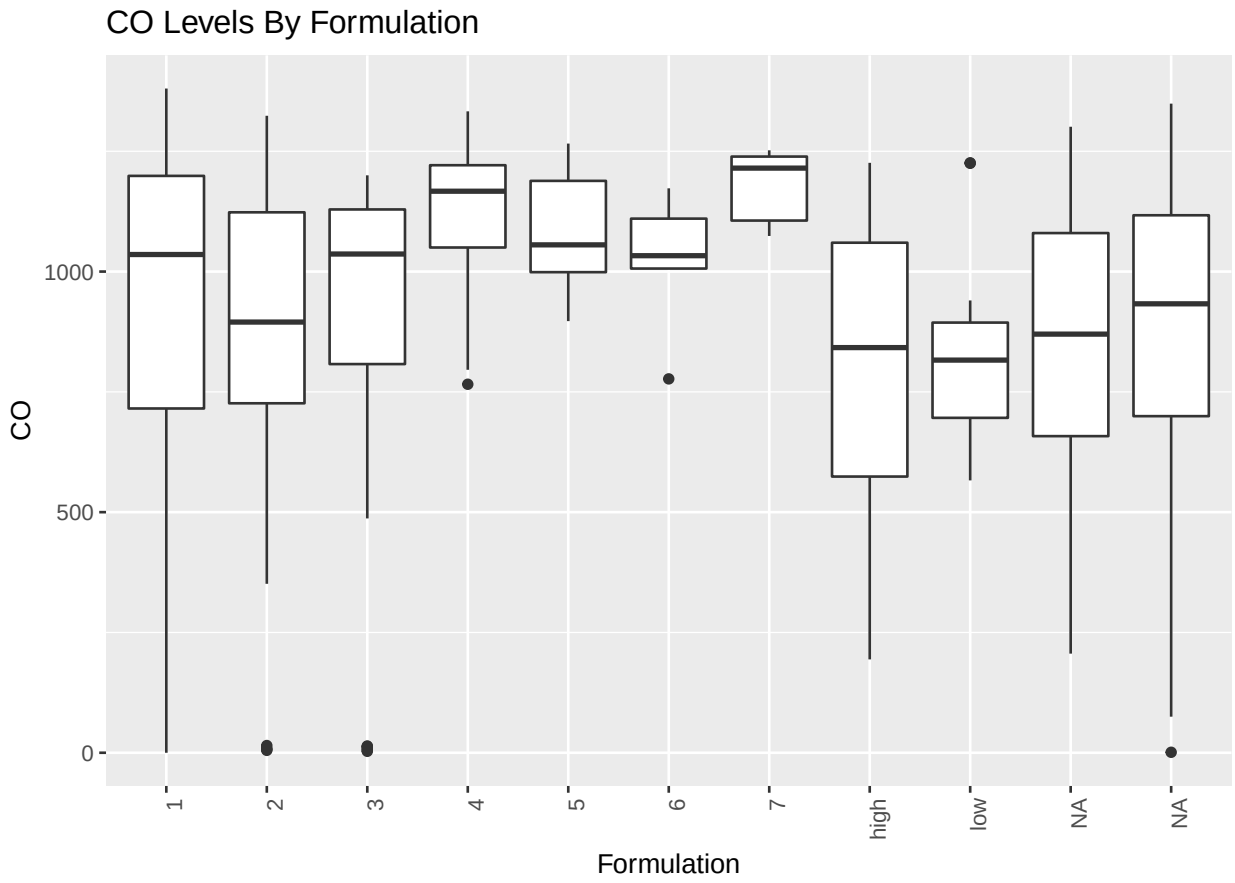
```
## CRS drying 71015.csvCRS drying 72315.csvCRS drying 72415.csvMJS Drying 12216.csvMJS D
```

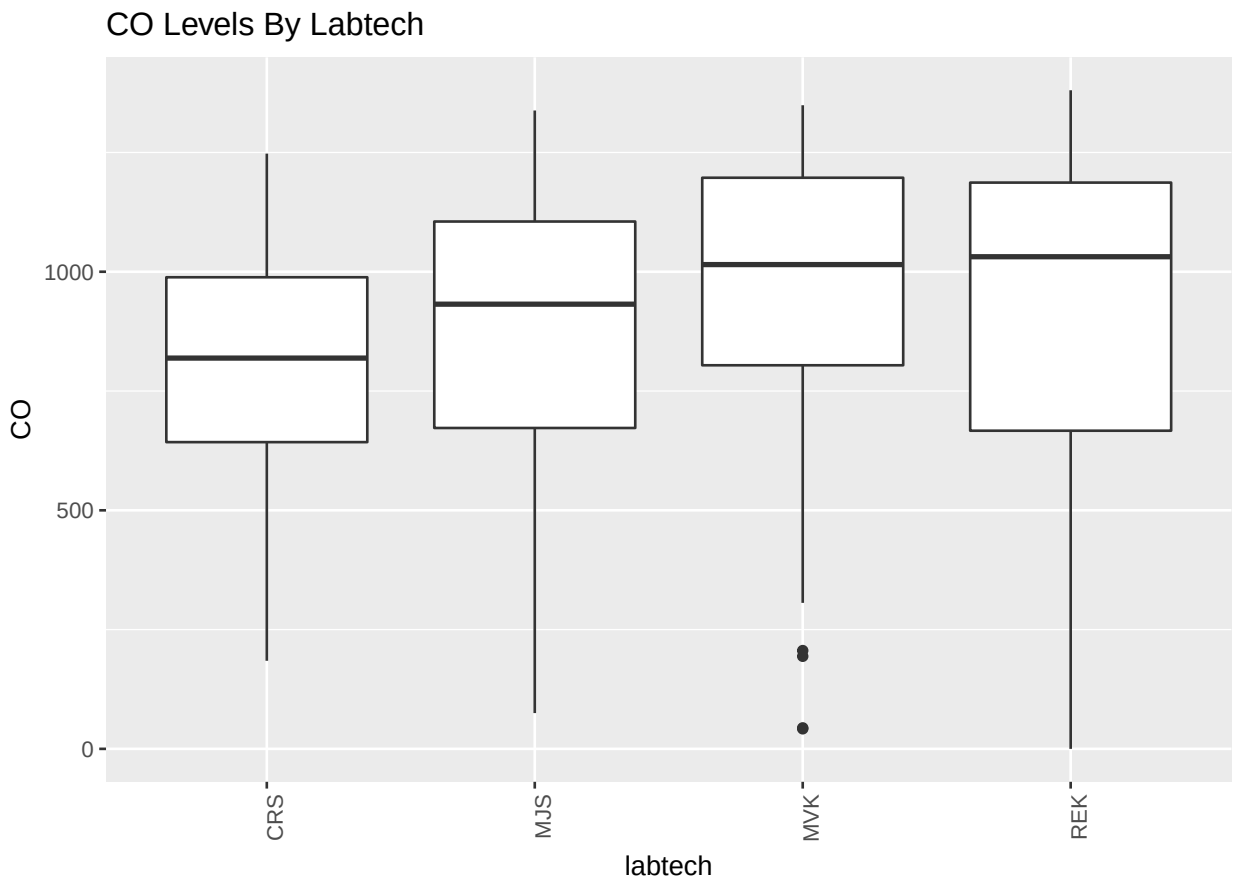
```
dfmaster <- dfmaster[order(dfmaster$experiment.mean, decreasing = TRUE), ]

dfmaster$labtech <- rep(NA, nrow(dfmaster))
for (i in 1:nrow(dfmaster)) {
    dfmaster[i, ]$labtech <- substr(as.character(dfmaster[i, ]$experiment),
        1, 3)
}
```

Table 1: Count of data elements redacted due to missing response.

missing.response
73




CO Levels By Formulation
CO
1000
500
0
1
2
3
4
5
6
7
high
low
NA
NA
Formulation


CO Levels By Labtech
CO
1000
500
0
CRS
MJS
MVK
REK
labtech


Replicates Per Experiment
df.counts$count
30
20
10
CRS drying 71015
CRS drying 72315
CRS drying 72415
MJS Drying 12216
MJS Drying 21916
MJS Drying 31116
MJS drying 52215
MJS drying 60515
MJS drying 61215
MJS drying 61715
MJS drying 61915
MJS liquid product 10115
MJS liquid product 101515
MJS liquid product 102215
MJS liquid product 10815
MJS liquid product 91015
MJS liquid product 91715
MJS long drying 52115
MJS substrate 61115
MJS substratepressure 6415
MVK Drying 11516
MVK Drying 1162015
MVK Drying 12916
MVK Drying 21216
MVK Drying 22616
MVK Drying 2516
MVK Drying 31816
MVK Drying 3216
MVK Drying 4116
MVK Drying 41516
MVK Drying 4816
REK Drying 101615
REK Drying 10215
REK Drying 102315
REK Drying 103015
REK Drying 10915
REK drying 111315
REK Drying 112015
REK Drying 12415
REK drying 82815
REK drying 91115
REK drying 91815
REK liquid product 9415
experiment

Analysis Section

Levene Test for Equality of Variances

Test for Homogeneity of Variances Levene's test (Levene 1960) is used to test if k samples have equal variances. Equal variances across samples is called homogeneity of variance. Some statistical tests, for example the analysis of variance, assume that variances are equal across groups or samples. The Levene test can be used to verify that assumption. Levene's test is an alternative to the Bartlett test. The Levene test is less sensitive than the Bartlett test to departures from normality. If you have strong evidence that your data do in fact come from a normal, or nearly normal, distribution, then Bartlett's test has better performance.

```
library(car)
# leveneTest(as.vector(scale(CO.n)), experiment.r)

leveneTest(CO.n, experiment.r)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value    Pr(>F)
## group  42  1.9467 0.0004332 ***
##       659
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The Levene test is significant so we now look to the Brown and Forsyth.

Brown and Forsythe Test for equality of variances.

The Levene test uses spread from the mean to check for variances while the Brown-Forsythe uses deviation from the median.

```
library(onewaytests)
bf.test(CO.n ~ experiment.r, data = d)
```

```
##
##   Brown-Forsythe Test
## -------------------------------------------------------
##   data : CO.n and experiment.r
##
##   statistic  : 3.217861
##   num df     : 42
##   denom df   : 487.9038
##   p.value    : 4.918218e-10
##
##   Result     : Difference is statistically significant.
## -------------------------------------------------------
```

ANOVA. First for comparison we will perform standard ANOVA. ANOVA is fairly roubust to violations of the contant variance assuption.

ANOVA Test

```
d.anova <- d[, c("experiment.r", "CO.r")]
lm <- lm(CO.r ~ experiment.r, data = d.anova)
lm.null <- lm(CO.r ~ 1, data = d.anova)
anova(lm.null, lm)
```

```
## Analysis of Variance Table
##
## Model 1: CO.r ~ 1
## Model 2: CO.r ~ experiment.r
##   Res.Df RSS Df Sum of Sq F Pr(>F)
## 1    701   0
## 2    659   0 42         0
```

The p-value is very small so the null hypothesis that there is no effect of experiment on CO level is rejected.

Welch Test

The R package `oneswaytest` is used to perform Welch's heteroscedastic F test see (Welch 1951) for more details. The test is a generalization to the Welch-t test to more than 2 groups. For large samples it amounts to a χ^2 test.

```
welch.test(CO.n ~ experiment.r, data = d, rate = 0, alpha = 0.05, na.rm = TRUE,
    verbose = TRUE)
```

```
##
##   Welch's Heteroscedastic F Test
## ------------------------------------------------------------
##   data : CO.n and experiment.r
##
##   statistic  : 8.155479
##   num df     : 42
##   denom df   : 184.6413
##   p.value    : 1.219526e-24
##
##   Result     : Difference is statistically significant.
## ------------------------------------------------------------
```

Kruskal-Wallis Rank Sum Test

```
kruskal.test(CO.r ~ experiment.r, data = d)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  CO.r by experiment.r
## Kruskal-Wallis chi-squared = 109.28, df = 42, p-value = 6.686e-08
```

Dunn Test

The most promising experiment is `MJS substratepressure 6415`

Apendix A VBA Data Extract

This is the listing of code for the extract of worksheets from excel to csv format.

```
Private Sub SaveAllSheetsAsCSV()
On Error GoTo Heaven

' each sheet reference
Dim Sheet As Worksheet
' path to output to
Dim OutputPath As String
' name of each csv
Dim OutputFile As String

Application.ScreenUpdating = False
Application.DisplayAlerts = False
Application.EnableEvents = False

' ask the user where to save
OutputPath = InputBox("Enter a directory to save to", "Save to directory", Path)

If OutputPath <> "" Then

    ' save for each sheet
    For Each Sheet In Sheets

        OutputFile = OutputPath & "\" & Sheet.Name & ".csv"

        ' make a copy to create a new book with this sheet
        ' otherwise you will always only get the first sheet
        Sheet.Copy
        ' this copy will now become active
        ActiveWorkbook.SaveAs Filename:=OutputFile, FileFormat:=xlCSV, CreateBackup:=Fal
        ActiveWorkbook.Close
    Next

End If

Finally:
Application.ScreenUpdating = True
Application.DisplayAlerts = True
Application.EnableEvents = True

Exit Sub
```

```
Heaven:
MsgBox "Couldn't save all sheets to CSV." & vbCrLf & _
        "Source: " & Err.Source & " " & vbCrLf & _
        "Number: " & Err.Number & " " & vbCrLf & _
        "Description: " & Err.Description & " " & vbCrLf

GoTo Finally
End Sub
```

Bibliography

Ashwood-Smith, M.J. 1980. "Mechanisms of Dehydration Injury in Bacteria." *International Journal of Refrigeration* 3 (4): 205–12. doi:https://doi.org/10.1016/0140-7007(80)90047-X.

Dunn, Olive Jean. 1961. "Multiple Comparisons Among Means." *Journal of the American Statistical Association* 56 (293). Taylor & Francis: 52–64. doi:10.1080/01621459.1961.10482090.

Frankenberg-Schwager, M, G Turcu, C Thomas, H Wollenhaupt, and H Bucker. 1975. "Membrane Damage in Dehydrated Bacteria and Its Repair." *Life Sciences and Space Research* 13: 83–88. http://europepmc.org/abstract/MED/12180476.

Kruskal, William H., and W. Allen Wallis. 1952. "Use of Ranks in One-Criterion Variance Analysis." *Journal of the American Statistical Association* 47 (260). Taylor & Francis: 583–621. doi:10.1080/01621459.1952.10483441.

Welch, B. L. 1951. "On the Comparison of Several Mean Values: An Alternative Approach." *Biometrika* 38 (3/4). [Oxford University Press, Biometrika Trust]: 330–36. http://www.jstor.org/stable/2332579.