IST772 Summary Template: Chapter 6 – Comparing Groups and Analyzing Experiments

Originality Assertion: By submitting this file you affirm that this writing is your own.

Name: Hendi Kushta
Date: Today's Date

****Important** Copying and/or pasting anything from the textbook will not be acceptable for your chapter notes submissions. You must write your notes in your own words and generate your own code, results, and graphs in R. This is what forces your brain to process the material that you read.**

INTRODUCTION

In the previous chapters of the book, the focus was on understanding the logic of statistical inference and how much trust can be placed in sample statistics to estimate population parameters. The chapter discusses how statistical inference considers and quantifies the role of sampling error in such estimates. In the last chapter, a t-test and Bayesian algorithm were used to compare the means of two groups, and in this chapter, a more general case is considered where the differences in means among any number of groups can be analyzed simultaneously. This technique is called analysis of variance (ANOVA), and it is a powerful way to compare multiple groups without having to perform individual t-tests repeatedly. ANOVA can be used to examine combinations of different factors, such as comparing various colors on a web page while varying font sizes.

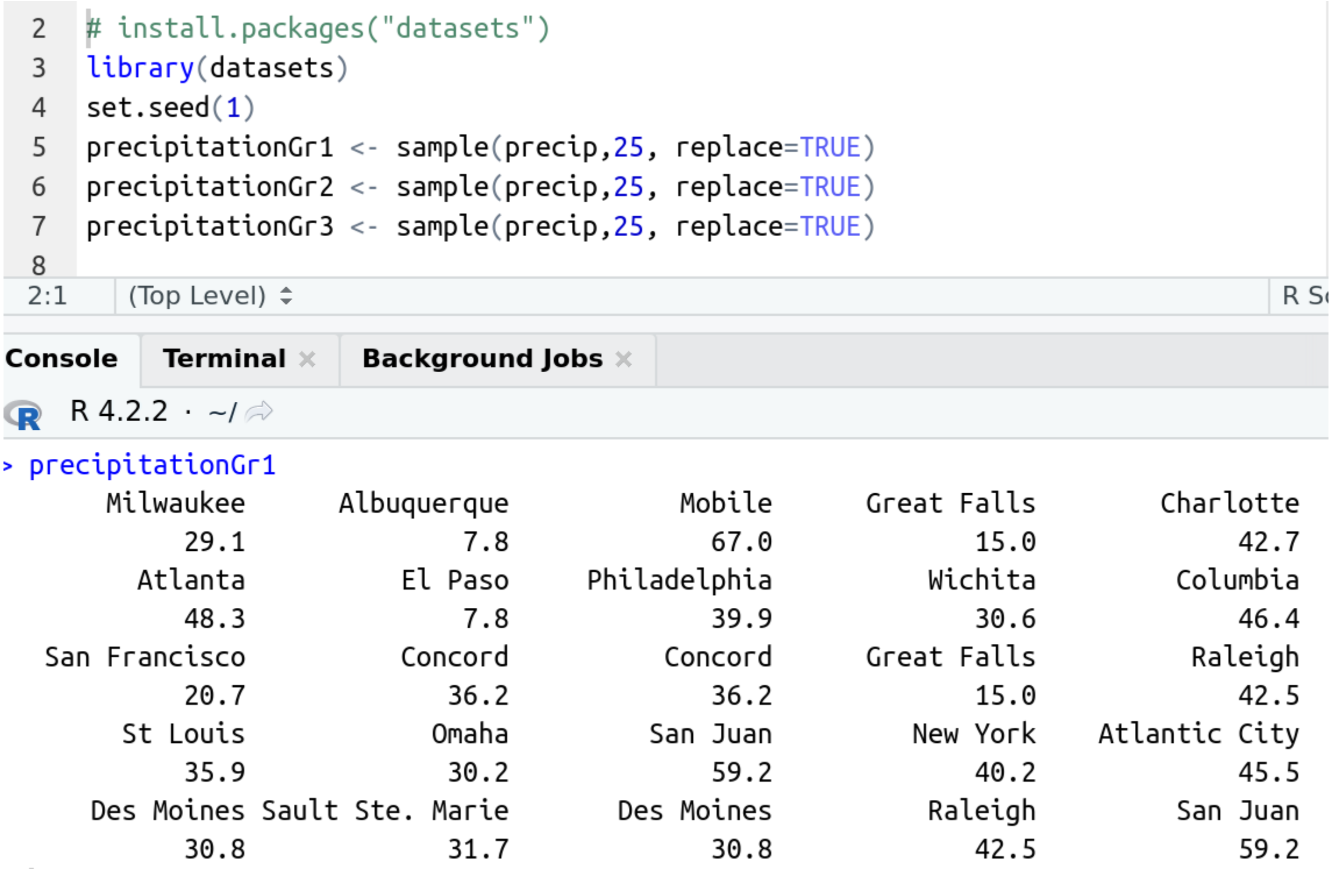
ANOVA is a member of the general linear model (GLM) family, which also includes other techniques that use a set of independent variables to predict a dependent variable. The GLM family of techniques can provide several useful outputs, such as an overall effect size value that represents the quality of the statistical result, inferential statistical tests to test hypotheses about the model quality or coefficient values, coefficients for each independent variable used to predict the dependent variable, and diagnostic information to ensure necessary assumptions about the data are satisfied. A practical example is given, where the cost of a tablet computer can be modeled based on the operating system, screen size, and memory size using a GLM analysis on a sample containing these variables.

ANOVA focuses on comparing variances between groups rather than means, and it partitions the overall variance into between-groups and within-groups variance to assess whether samples come from the same underlying population. The independent variables in ANOVA tend to be categorical, while the GLM family also includes techniques with metric independent variables.

Below is an example it R for a prebuilt precipitation data in R. Start with getting a sample of 25 cities each state in USA. I have shown the sample from first group of samples.

IST772 Summary Template: Chapter 6 – Comparing Groups and Analyzing Experiments

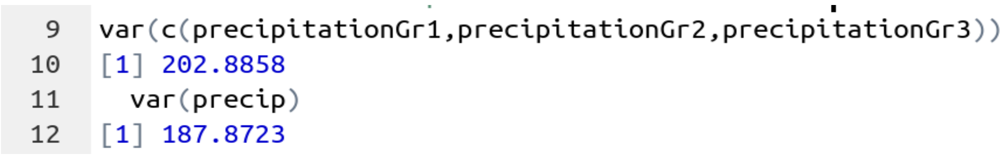
Originality Assertion: By submitting this file you affirm that this writing is your own.



```
# install.packages("datasets")
library(datasets)
set.seed(1)
precipitationGr1 <- sample(precip,25, replace=TRUE)
precipitationGr2 <- sample(precip,25, replace=TRUE)
precipitationGr3 <- sample(precip,25, replace=TRUE)
```

```
> precipitationGr1
      Milwaukee     Albuquerque          Mobile     Great Falls       Charlotte 
           29.1             7.8            67.0            15.0            42.7 
        Atlanta         El Paso    Philadelphia         Wichita        Columbia 
           48.3             7.8            39.9            30.6            46.4 
  San Francisco         Concord         Concord     Great Falls         Raleigh 
           20.7            36.2            36.2            15.0            42.5 
       St Louis           Omaha        San Juan        New York   Atlantic City 
           35.9            30.2            59.2            40.2            45.5 
     Des Moines Sault Ste. Marie     Des Moines         Raleigh        San Juan 
           30.8            31.7            30.8            42.5            59.2 
```

After sampling the population, In 3 samples, I have mixed them in one to find the variance and compare it to the population's variance.

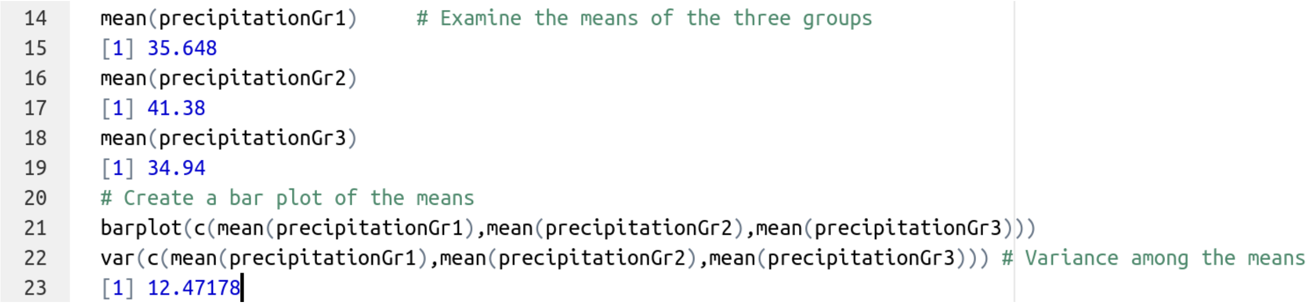


```
var(c(precipitationGr1,precipitationGr2,precipitationGr3))
[1] 202.8858
  var(precip)
[1] 187.8723
```

Another technique to examine variance is to compare the variance in the means of the three separate groups, as shown in the following example:

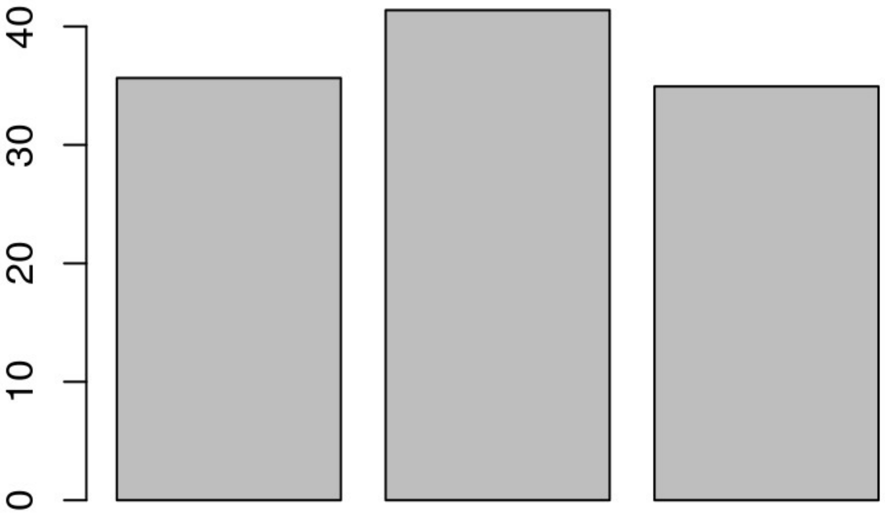
IST772 Summary Template: Chapter 6 – Comparing Groups and Analyzing Experiments

Originality Assertion: By submitting this file you affirm that this writing is your own.



```
mean(precipitationGr1)     # Examine the means of the three groups
[1] 35.648
mean(precipitationGr2)
[1] 41.38
mean(precipitationGr3)
[1] 34.94
# Create a bar plot of the means
barplot(c(mean(precipitationGr1),mean(precipitationGr2),mean(precipitationGr3)))
var(c(mean(precipitationGr1),mean(precipitationGr2),mean(precipitationGr3))) # Variance among the means
[1] 12.47178
```

The code describes an experiment where the means of three different groups are examined to calculate the variance among them. The variance of the means is found to be 12.47, which is relatively small compared to the variances of the raw data for each group.

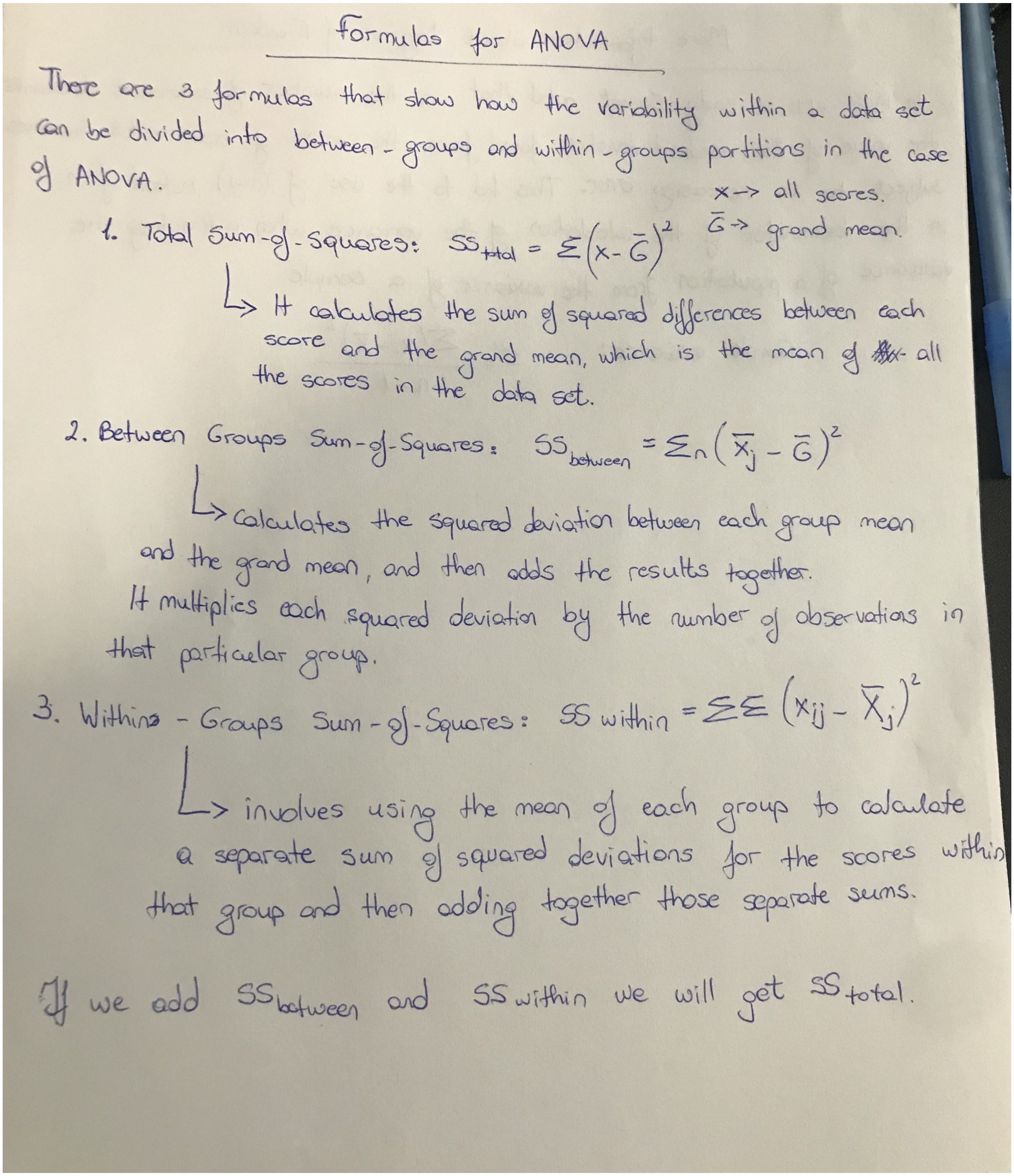


The F-ratio is a statistical measure that compares the between-groups variance to the within-groups variance in a sample. The numerator of the F-ratio is the scaled variance among the means of the groups (between-groups variance), and the denominator is the variance within the groups (within-groups variance). If the F-ratio is substantially larger than 1.0, it suggests that at least one of the groups is from a population with a different mean. The F-ratio can vary each time we sample because the data and variances among group means are different, but under the assumption that all groups are from the same population (null hypothesis), most F-ratios should be very close to 1.0.

IST772 Summary Template: Chapter 6 – Comparing Groups and Analyzing Experiments

Originality Assertion: By submitting this file you affirm that this writing is your own.

BOX ON P.93: FORMULAS FOR ANOVA



formulas for ANOVA

There are 3 formulas that show how the variability within a data set can be divided into between - groups and within - groups portitions in the case of ANOVA.

x -> all scores.

$\bar{G}$ -> grand mean.

1. Total Sum-of-Squares: $SS_{total} = \Sigma(x-\bar{G})^2$

-> It calculates the sum of squared differences between each score and the grand mean, which is the mean of all the scores in the data set.

2. Between Groups Sum-of-Squares: $SS_{between} = \Sigma n(\bar{X}_j - \bar{G})^2$

-> Calculates the squared deviation between each group mean and the grand mean, and then adds the results together. It multiplies each squared deviation by the number of observations in that particular group.

3. Withins - Groups Sum-of-Squares: $SS\ within = \Sigma\Sigma(x_{ij} - \bar{X}_j)^2$

-> involves using the mean of each group to calculate a separate sum of squared deviations for the scores within that group and then adding together those separate sums.

If we add $SS_{between}$ and SS_{within} we will get SS_{total}.

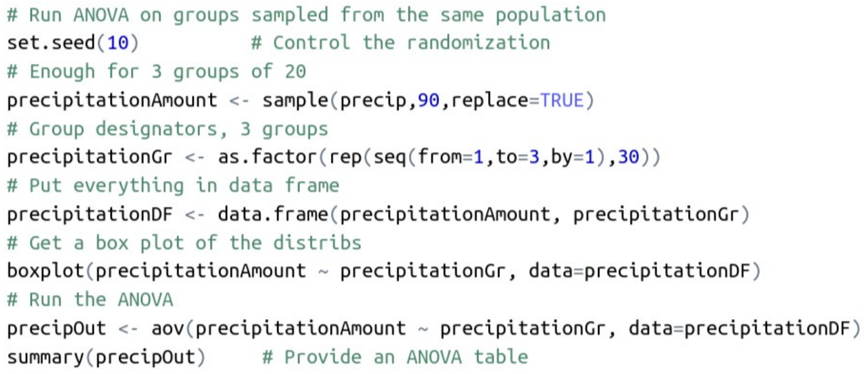
IST772 Summary Template: Chapter 6 – Comparing Groups and Analyzing Experiments

Originality Assertion: By submitting this file you affirm that this writing is your own.

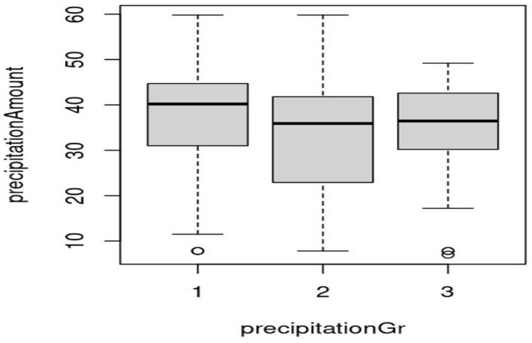
FREQUENTIST APPROACH TO ANOVA

In this part, I will use statistical procedures in R to test variance using the Null Hypothesis Significance Test (NHST). The aov() function, which is a wrapper around the lm() function and uses least-squares fitting to implement the general linear model. The passage provides the code for sampling a new set of observations from the "precip" data set, creating three groups, checking their distributions, running an ANOVA, and checking the output.

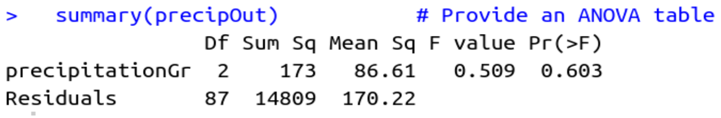
Below is all the necessary code to generate a new set of 90 observations from the "precip" dataset, divide them into three groups of 30, examine their distributions, conduct an ANOVA, and review the results.



```
# Run ANOVA on groups sampled from the same population
set.seed(10)          # Control the randomization
# Enough for 3 groups of 20
precipitationAmount <- sample(precip,90,replace=TRUE)
# Group designators, 3 groups
precipitationGr <- as.factor(rep(seq(from=1,to=3,by=1),30))
# Put everything in data frame
precipitationDF <- data.frame(precipitationAmount, precipitationGr)
# Get a box plot of the distribs
boxplot(precipitationAmount ~ precipitationGr, data=precipitationDF)
# Run the ANOVA
precipOut <- aov(precipitationAmount ~ precipitationGr, data=precipitationDF)
summary(precipOut)     # Provide an ANOVA table
```



The code block uses the set.seed() function to control the random number sequence and draw 90 data points at random from the built-in "precip" data set. The rep() and seq() commands generate a repeating sequence of 1, 2, 3 to assign group designators. The two variables, precipitation data and group designators, are combined using data.frame() to create a data frame. The boxplot is used to check the result and shows that the medians of the three distributions are similar. The aov() command runs the ANOVA, which tests the dependent variable, "precipitationAmount," as a function of the independent variable, "precipitationGr." The output is stored in a variable called precipOut, and the summary() command is used to produce a standard ANOVA table which is shown below.



```
>   summary(precipOut)           # Provide an ANOVA table
                Df Sum Sq Mean Sq F value Pr(>F)
precipitationGr  2    173   86.61   0.509  0.603
Residuals       87  14809  170.22
```

IST772 Summary Template: Chapter 6 – Comparing Groups and Analyzing Experiments

Originality Assertion: By submitting this file you affirm that this writing is your own.

The ANOVA table contains important statistical quantities for the analysis, such as degrees of freedom (df), sum of squares (Sum Sq), mean squares (Mean Sq), F-ratio (F-value), and the probability of a larger F-ratio (Pr(>F)). The table includes results for the grouping variable (independent variable) specified in the call to aov(), as well as the residuals, which account for all of the within-group variability. Residuals represent the variance that remains after all of the systematic variance has been removed, in this case, the variance that is left over after the "precipitationGr" variable has been taken into account. Understanding how to interpret an ANOVA table is an essential skill.
In interpreting the ANOVA table, the F-value and the probability of finding a larger F are the most important values. For the ANOVA result to be statistically significant, the F-value must substantially exceed one, and the Pr(>F) must be less than the chosen alpha level (typically 0.05, 0.01, or 0.001) before conducting the test. In this example, the p-value is 0.603, which is much larger than any of these conventional alpha values, so we have failed to reject the null hypothesis that all three groups were sampled from the same population. Sampling all three groups from the same population is by and large the very definition of the null hypothesis. It is important to note that sampling errors can cause some results to look significant even when they are not, resulting in a false positive or a Type 1 error.

BOX ON P.99: MORE INFORMATION ABOUT DEGREES OF FREEDOM

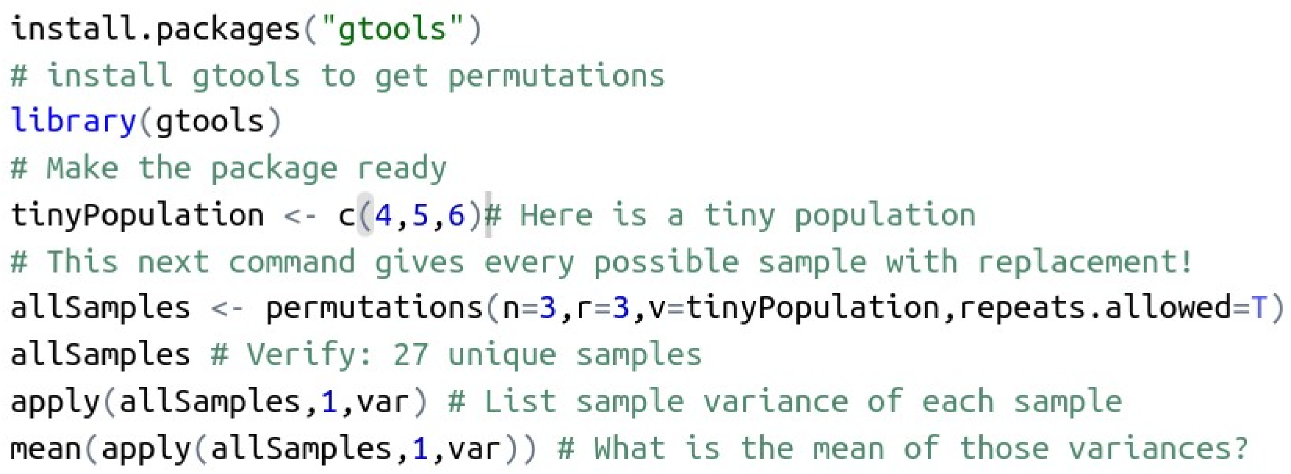
In his 1908 paper "The Probable Error of a Mean", William Sealy Gosset noted that as the number of experiments decreases, the value of the standard deviation found from the sample of experiments becomes subject to an increasing error. This led to the use of (n-1) instead of n in the denominator of the calculation of the variance, when estimating the variance of a population from the variance of a sample.

$$s^2 = \frac{\Sigma\left(x_i - \overline{x}\right)^2}{\left(n-1\right)}$$

The formula for sample variance uses (n-1) in the denominator, making the variance slightly larger than it would be if n were used. This suggests that using n instead of (n-1) might lead to an underestimation of population variance. The cause of this underestimation can be explored using a code snippet.

IST772 Summary Template: Chapter 6 – Comparing Groups and Analyzing Experiments

Originality Assertion: By submitting this file you affirm that this writing is your own.

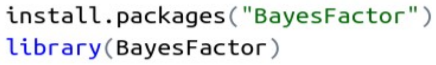


```
install.packages("gtools")
# install gtools to get permutations
library(gtools)
# Make the package ready
tinyPopulation <- c(4,5,6)# Here is a tiny population
# This next command gives every possible sample with replacement!
allSamples <- permutations(n=3,r=3,v=tinyPopulation,repeats.allowed=T)
allSamples # Verify: 27 unique samples
apply(allSamples,1,var) # List sample variance of each sample
mean(apply(allSamples,1,var)) # What is the mean of those variances?
```

It is explained why statisticians use (n-1) instead of n in the denominator of the sample variance formula when estimating the variance of a population from a sample. The reason is that when we calculate the sample variance with n in the denominator, it leads to an underestimation of population variance due to the sample mean bias.
To demonstrate this, I have provided an example with a three-element population, 3,4,5, where the population variance is easy to calculate as 0.6666667. By using the (n-1) denominator in the sample variance calculation and averaging across all possible samples, the result is also 0.6666667. This unbiased estimator corrects for the uncertainty in the sample mean and borrows one degree of freedom from the sample.

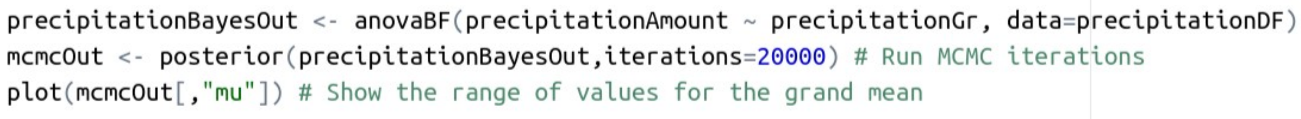
THE BAYESIAN APPROACH TO ANOVA

The passage discusses two different approaches for analyzing data: the frequentist approach and the Bayesian approach. In the frequentist approach, the data is used to reject or fail to reject the null hypothesis by estimating the population variance and creating a test statistic called the F-ratio, which is compared to a theoretical distribution to obtain a p-value. If the p-value is below a preselected alpha threshold, the null hypothesis is rejected. However, this approach does not provide information about the probability that the null or alternative hypothesis is true. In contrast, the Bayesian approach uses data to calculate the probability of the null or alternative hypothesis being true.



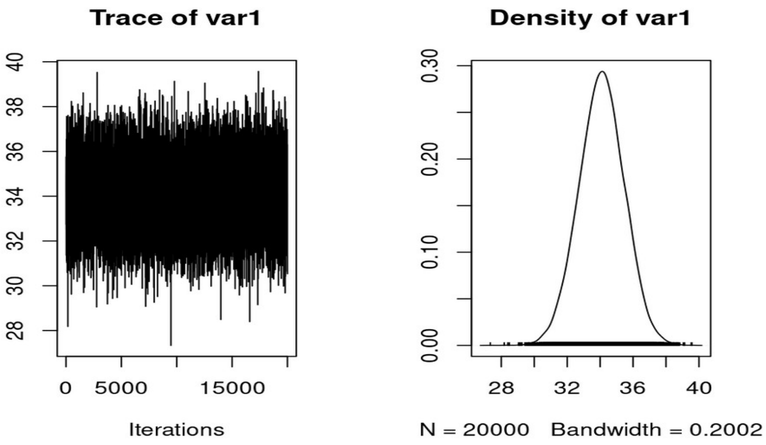
```
install.packages("BayesFactor")
library(BayesFactor)
```

The BayesFactor package has other packages it relies on, such as the "coda" package. These dependencies will usually be automatically downloaded and installed, but it's important to be aware of any warnings or error messages that may appear in the console. The anovaBF function in the package uses the same formula notation as the aov() function.

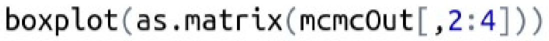


```
precipitationBayesOut <- anovaBF(precipitationAmount ~ precipitationGr, data=precipitationDF)
mcmcOut <- posterior(precipitationBayesOut,iterations=20000) # Run MCMC iterations
plot(mcmcOut[,"mu"]) # Show the range of values for the grand mean
```

In this section, the author discusses the use of the BayesFactor package to perform Bayesian ANOVA analysis. The anovaBF function is used to estimate the posterior distributions for the population parameters and generate estimates of deviations from the grand mean for each group. The resulting trace plot and density histogram provide information on the range of population values obtained from the MCMC analysis of the posterior distribution of the grand mean, and can help confirm that the MCMC run has converged on a stable result. Ultimately, the group means' deviation from the grand mean is the most important factor to consider when interpreting the ANOVA analysis results.



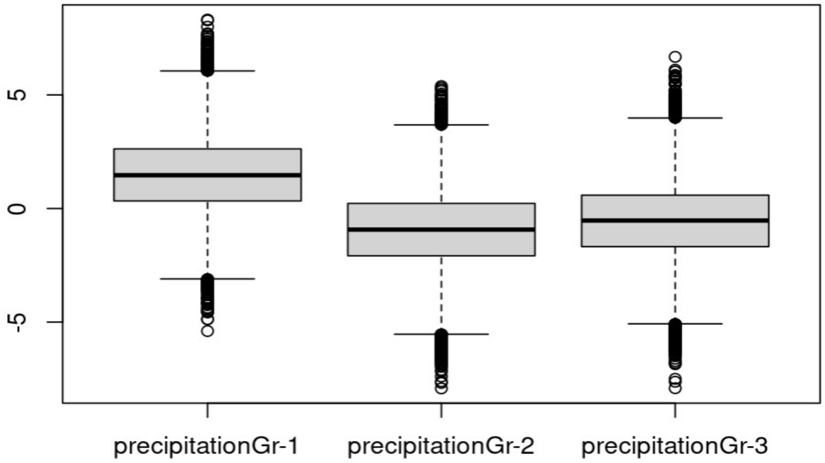
The section discusses summarizing the results of an ANOVA analysis with a boxplot, which is used to visualize the posterior distributions of all three groups together. The boxplot shows the deviations of each group mean from the grand mean, and the black bar in the middle of each box represents the median. The boxes contain the central 50% of all observations, and the whiskered areas for each group overlap each other, suggesting no differences among the group means. The section also provides a command for reviewing the numeric results of the MCMC sampling.



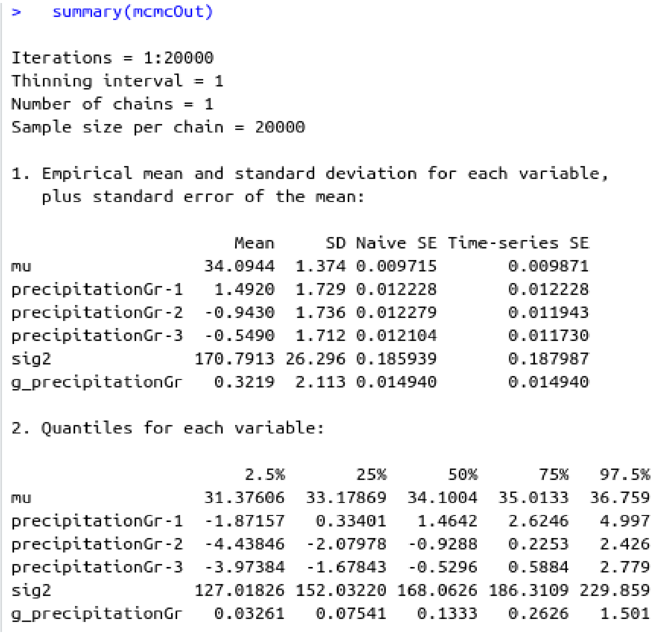
```
boxplot(as.matrix(mcmcOut[,2:4]))
```

IST772 Summary Template: Chapter 6 – Comparing Groups and Analyzing Experiments

Originality Assertion: By submitting this file you affirm that this writing is your own.



The output of the MCMC sampling for the Bayesian analysis is organized into two sections. The first section provides the empirical mean and standard deviation for each variable, plus the standard error of the mean. The second section provides quantiles for each variable, containing the lower and upper bounds of the 95% Highest Density Interval (HDI) for each parameter. The HDI is used to infer whether there are credible differences among the groups. The output shows that the 95% HDIs for all three groups overlap substantially with each other, suggesting no credible differences among the group means. The output also confirms that none of the group means is credibly different from the grand mean, supporting the inference that these three groups were all drawn from the same population. In addition, the BayesFactor package calculates the odds ratio for comparing two statistical models and provides a clear way of talking about the strength of evidence in favor of one model or another.



```
>   summary(mcmcOut)

Iterations = 1:20000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 20000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

                      Mean     SD Naive SE Time-series SE
mu                 34.0944  1.374 0.009715       0.009871
precipitationGr-1   1.4920  1.729 0.012228       0.012228
precipitationGr-2  -0.9430  1.736 0.012279       0.011943
precipitationGr-3  -0.5490  1.712 0.012104       0.011730
sig2              170.7913 26.296 0.185939       0.187987
g_precipitationGr   0.3219  2.113 0.014940       0.014940

2. Quantiles for each variable:

                       2.5%       25%      50%      75%   97.5%
mu                 31.37606  33.17869  34.1004  35.0133  36.759
precipitationGr-1  -1.87157   0.33401   1.4642   2.6246   4.997
precipitationGr-2  -4.43846  -2.07978  -0.9288   0.2253   2.426
precipitationGr-3  -3.97384  -1.67843  -0.5296   0.5884   2.779
sig2              127.01826 152.03220 168.0626 186.3109 229.859
g_precipitationGr   0.03261   0.07541   0.1333   0.2626   1.501
```

IST772 Summary Template: Chapter 6 – Comparing Groups and Analyzing Experiments

Originality Assertion: By submitting this file you affirm that this writing is your own.

BOX ON P.103: GIVING SOME THOUGHT TO PRIORS

The article discusses the role of prior probabilities in Bayesian thinking and how they are used in statistical analysis. Bayesian thinking starts with some prior probabilities about the situation that are modified using data to develop some posterior probabilities. However, in the Bayesian t-tests and ANOVA conducted in previous chapters, it appears that no priors were set.

Some statisticians argue that as long as you have a sizable dataset, the prior probabilities are not especially influential on the results. Uninformative priors may suffice except when working with small samples of data. In these cases, it may be more meaningful to sample priors from normal distributions, which are known as noncommittal priors.

The article also describes how Bayesian ANOVA procedures use priors from the Cauchy distribution. The Cauchy distribution is a heavy-tailed distribution that assigns more weight to extreme values than a normal distribution. By using these priors, the procedure is able to detect differences between groups that may not be detected using frequentist methods.

Overall, the article emphasizes the importance of priors in Bayesian analysis and the various approaches that can be taken to specify them. It also highlights the advantages of Bayesian methods over frequentist methods, particularly in cases where the data is limited or the effect size is small.

BOX ON P.110: INTERPRETING BAYES FACTOR

Here are some essential principles for interpreting Bayes factors:

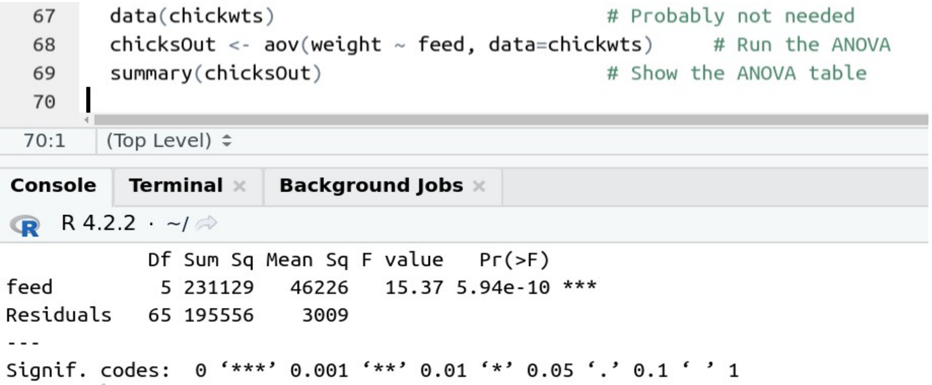
- Bayes factors are one way to look at statistical models and data, and should be considered alongside other kinds of results.
- Every Bayes factor represents a comparison between two models, and the favored model is the one with the higher Bayes factor. However, both models being compared could be poor choices.
- Choose appropriate and meaningful hypotheses to compare before collecting data, instead of only considering how to reject a null hypothesis as in frequentist thinking.
- The strength of the Bayes factor should be interpreted. An odds ratio weaker than 3:1 is not worth mentioning, odds ratios from 3:1 up to 20:1 are positive evidence, odds ratios from 20:1 up to 150:1 are strong evidence, and odds ratios of more than 150:1 are very strong evidence for the favored hypothesis.
- The strength of evidence needed depends on the research situation and consequences. For high-stakes situations, stronger evidence is needed.

IST772 Summary Template: Chapter 6 – Comparing Groups and Analyzing Experiments

Originality Assertion: By submitting this file you affirm that this writing is your own.

FINDING AN EFFECT

The previous section demonstrated the concept of null hypothesis and the importance of having a baseline. Real data is now used to explore the statistics and observe what they tell us. The "chickwts" dataset is used, which consists of 71 observations of six-week-old chicks assigned randomly to one of six feed groups. The essential research question is whether the type of feed affects the growth of chicks and which types promote more or less growth. The ANOVA null hypothesis test is conducted to determine whether the variation among means exceeds the variance within groups. The F-value is calculated, and the p-value is determined by positioning the observed value of F on the appropriate F distribution.



```
data(chickwts)                                   # Probably not needed
chicksOut <- aov(weight ~ feed, data=chickwts)     # Run the ANOVA
summary(chicksOut)                               # Show the ANOVA table
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)    
feed         5 231129   46226   15.37 5.94e-10 ***
Residuals   65 195556    3009                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The passage discusses the analysis of real data using the "chickwts" dataset, which contains observations of 6-week-old chicks assigned to one of six feed groups. The question of whether the type of feed affects the growth of chicks is explored using ANOVA. The results show that the p-value is statistically significant, indicating that we must reject the null hypothesis that all six groups were sampled from the same population. The eta-squared effect size is calculated to be 0.54, suggesting that feed type explained about 54% of the variance in weight. Follow-up tests, such as post hoc testing, may be necessary to examine the mean differences between different pairs of groups in more detail.