Professional Masters in
Applied Statistics and Data Science
(PM-ASDS)

A Report on The Statistical Inference
Procedures of Population Variances and Proportions

Submitted to:

Dr. Tapati Basak
Associate Professor
Department of Statistics, JU

Submitted by:
Mohammad Saiduzzaman Sayed
ID: 20215063
Batch: 5th
Sec: A

Course Name: Statistical Inference
Course Code: PM-ASDS03

Inference Procedures of Population Variances and Proportions

CONTENTS

LIST OF FIGURES

Abstract

We assessed the inference procedures of Population Variances and Proportions, described what steps to follow for solving problems, and computed mathematical descriptions.

We are using the CO2 dataset which is built-in dataset in R. Using conc and uptake variable for Variance's testing. And Proportion testing, we are using Type and Treatment variable.

We are going to solve our problems in R programming language. We will discuss how we can make a decision through hypothesis testing.

This procedure helps us to gather knowledge about a dataset and test variances and proportions.

Chapter 1

Introduction

Statistical Inference:

With the help of Statistical inference, we can conclude about a population from a sample that is drawn from that population by following some process.

Motivating example:

If we consider a problem, how much earning in a month of a freelancer in Chittagong City? We randomly select a sample of freelancers in Chittagong City and collect data on their monthly income and other characteristics.

We can use inferential statistics to conclude about freelancer monthly income in Chittagong City from that sample which we have drawn from our population.

The two general areas of Statistical inference:

1. Estimation
2. Hypothesis Testing

Sampling:

Sampling is a statistical procedure that is concerned with the selection of the individual observation; it helps us to make statistical inferences about the population.

Simple random sampling is a method of selecting n elements from a population of size N elements in such as a way that each combination of n elements has the equal chance or probability of being selected as every other combination.

Probability and Probability Distribution:

A numerical measure of uncertainty of an event of an experiment is probability.

A probability distribution shows the possible outcomes of an experiment and the probability of each of these outcomes.

Different types of probability distribution:

i) Bernoulli distribution

ii) Binomial distribution

iii) Poisson distribution

iv) Normal distribution

The Central Limit Theorem:

The Central Limit Theorem states that the distribution of averages of iid variables becomes that of a standard normal as the sample size increases.

Two rules of CLT:

i. We observed one average.

ii. We don’t know what the population distribution is.

For large n the result has a standard normal distribution.

The result is that,

$$\frac{\overline{\mathrm{X}}_{\mathrm{n}} - \mu}{\sigma \,/\, \sqrt{n}} = \frac{\sqrt{\mathrm{n}}(\overline{\mathrm{X}}_{\mathrm{n}} - \mu)}{\sigma} = \frac{\text{Estimate} - \text{Mean of estimate}}{\text{Standard error of estimate}}$$

Estimation:

The process of calculating a statistic from a sample's data serves as an estimate of the relevant parameter of the population from which the sample was drawn. Estimation is the value of the parameter.

There are two types of estimates we can compute:

1. A Point Estimate
2. An Interval Estimate

A Point Estimate:

A point estimate is a single numerical value used to estimate the corresponding population parameter. For Quantitative data, we can estimate mean and for categorical data, we can estimate proportion. In a normal distribution, the mean is considered more efficient that the median.

Example: We want to estimate the average income of a freelancer in a month in Chittagong City. Let, our sample size is 50 and their total income in a month is 89000 dollars. So, we can say that the average income of a freelancer in a month in Chittagong City is 1780 dollar.

Interval Estimate:

An interval estimate consists of two numerical values which are defining a range of values that most likely includes the parameter being estimated. We define one value as a lower confidence limit and another value as a higher confidence limit. Their difference is the width of the confidence interval.

Confidence intervals:

A confidence interval refers to the probability that the parameter of a population contains between two numerical values

Example:

With random sampling, a 95% confidence interval of [1700 1800] means we are 95% confident that the average monthly income of a freelancer is between 1700 and 1800 dollars.

Hypothesis Testing:

A Hypothesis is a claim (assumption) about a population parameter. Hypothesis Testing can determine whether or not such statements are compatible with the available data. Hypothesis helps us to make decisions about population formally.

Example:

An experimenter is interested in the average monthly income of a freelancer. A simple random sample of 50 individuals was drawn from the population of interest. From this sample, a mean of $\bar{x}=1780$ has been computed. From the central limit theorem, it is assumed that the sample is normally distributed because our sample size is large enough. Can he/she conclude that the mean monthly income of this population is different from the 1780 dollars?

There are two types of Hypotheses:

1. Research Hypothesis
2. Statistical Hypothesis

Research Hypothesis: Research hypothesis refers to the research being motivated by a hypothesis or a presumption.

Statistical Hypothesis: Statistical hypotheses are considered to be adequate that allows assumptions to be tested using statistical procedures.

Procedures of Testing Hypothesis:

- Specify the Null Hypothesis and the Alternative Hypothesis.
- Set the Significance Level, α
- Specify the Test Statistic.
- Use the level of significance and alternative hypothesis to develop the rejection rule.
- Draw conclusion based on sample.

Types of Statistical Hypothesis:

Null Hypothesis H_0:

A maintained hypothesis is held to be true unless sufficient evidence to the contrary is obtained. Null Hypothesis always contains the "=", "≤", "≥" sign. It may or may not be rejected.

Alternative Hypothesis H_1:

A hypothesis against which the null hypothesis is tested, and which will be held to be true if the null is held false. It never contains the "=", "≤", "≥" sign. It is generally the hypothesis that the researcher is trying to support.

One-Sided or Two-sided Alternative: An alternative hypothesis involving all possible values of a population parameter on either one side or the other (that is, either greater than or less than) or both sides of the value specified by a simple null hypothesis.

Hypothesis Test Decisions: A decision rule is formulated, leading the investigator to either reject or fail to reject the null hypothesis based on sample evidence.

Significance Level: The probability of rejecting a null hypothesis that is true. This probability is sometimes expressed as a percentage, so a test of significance level a is referred to as an $\alpha 100\%$ level test.

Critical value: The boundary points between the acceptance region and critical region.

p-value: Probability of obtaining a test statistic more extreme ($\leq$ or $\geq$) than the observed sample value given H_0 is true.

Type I Error: The rejection of a true null hypothesis.

Type II Error: The failure to reject a false null hypothesis.

		Condition of Null Hypothesis	
		True	False
Possible Action	Fail to Reject H_0	Correct Action	Type II Error
	Reject H_0	Type I Error	Correct Action

Draw conclusion:

Reject H_0 if $p - value \ \leq \ \alpha$

Dataset:

The CO2 data.frame is a dataset built into R showing the results of an experiment on the cold tolerance of grass. Grass samples from two regions (Quebec and Mississippi) were grown in either a chilled or nonchilled environment, and their CO2 uptake rate was tested.

R-Programming Language:

R is a programming language. R is often used for statistical computing and graphical presentation to analyze and visualize data.

In the next chapter, we discuss the mathematical descriptions of inference procedures for population variances and proportions.

Chapter 2

Methodologies

Inference Procedures of Population Variances:

Sampling and Sample Size determination:

With the help of simple random sampling without replacement, we can take samples from our population using the sample size determination formula.

Sample size determination for continuous data,

Formula 1:

$$n_0 = \frac{z^2 \sigma^2}{d^2}$$

Formula 2:

$$n = \frac{n_0}{1 + \frac{n_0}{N}}$$

Here,

n_0 = estimated sample size.

z = statistical certainty chosen (1.96 for 5% level of significance).

σ = population standard deviation, obtained from prior knowledge of the population.

d = width of the interval desired (precision).

n = desired sample size.

A Point Estimate:

For Quantitative data, we can estimate mean and for categorical data. In a normal distribution, the mean is considered more efficient than the median.

For mean,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{\mathrm{n}} x_i$$

Here, $\bar{x}$ = Sample mean

n = Sample size

$\frac{1}{n}\sum_{i=1}^{\mathrm{n}} x_i$ = Estimator

Estimation and Tests for a Population Variance:

Properties of Chi-Squared Distribution:

~ Parameter = k degrees of freedom

~ Mean = k

~ Not symmetrical about the mean.

~ It is a component of the t-distribution and the F-distribution used in t-tests, ANOVA.

~ Chi-Squared Distribution tends to normal distribution if sample size increases.

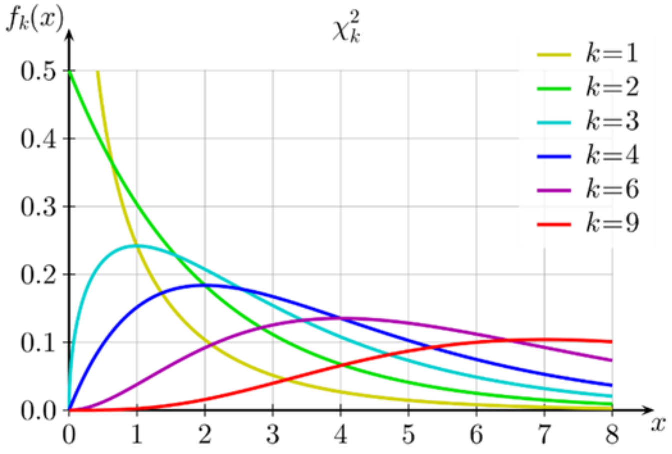

Fig. 1 Chi-Squared Distribution

Confidence Interval for the Variance of a Normally Distributed Population:

~ Population Variance, $\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$

~ Sample Variance, $s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$

~ It's a measure of how much dispersion there is in the data. The sample variance, s^2 is a point estimator of population variance, σ^2.

The assumptions of the one-sample inference procedures for a variance:

~ We have a simple random sample from the population of interest.

~ The population is normally distributed.

Under these conditions:

~ s^2 is an unbiased estimator of σ^2.

~ $\frac{(n-1)s^2}{\sigma^2}$ has a Chi-squared distribution, χ^2 with n-1 degrees of freedom.

A $(1-\alpha)100\%$ confidence interval for σ^2 is given by:

$$\frac{(n-1)s^2}{\chi^2_{1-(\alpha/2)}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{\alpha/2}}$$

Hypothesis Testing for a single population variance:

$H_0: \sigma^2 = \sigma_0^2$

$H_0: \sigma^2 \geq \sigma_0^2$

$H_0: \sigma^2 \leq \sigma_0^2$

Here, the test statistic is,

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

In terms of χ^2, $H_a: \sigma^2 \neq \sigma_0^2$

$H_a: \sigma^2 > \sigma_0^2$

$H_a: \sigma^2 < \sigma_0^2$

Estimation and Tests for the Ratio of Two Population Variances:

Properties of F Distribution:

~ Parameter = $d_1 = (n_1 - 1)$ and $d_2 = (n_2 - 1)$ degrees of freedom

~ Not symmetrical about the mean.

~ The F distribution is a Continuous probability Distribution that frequently appears as the null distribution of a test statistic, particularly in analysis of variance.

~ F Distribution tends to normal distribution if sample size increases.

~ F-distribution is an instance of ratio distributions.

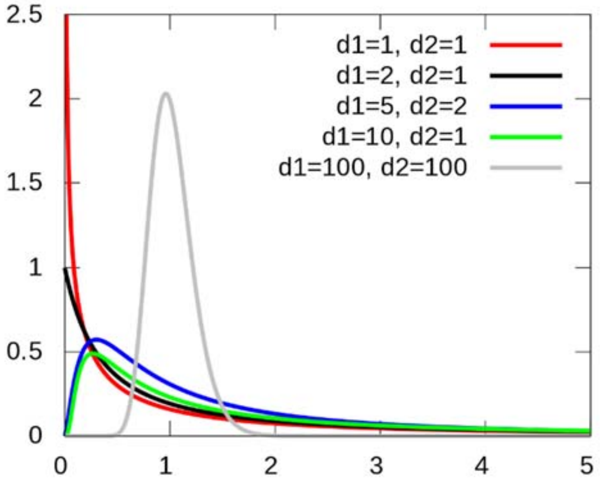

Fig. 2 F Distribution

Confidence Interval for the Ratio of the Variances of two Normally Distributed Populations:

The assumptions of the two-sample inference procedures for variances:

~ We have two simple random samples from the population of interest.

~ The population is normally distributed.

Under these conditions:

~ s_1^2 is an unbiased estimator of σ_1^2. And s_2^2 is an unbiased estimator of σ_2^2.

~ $\frac{s_1^2/s_2^2}{\sigma_1^2/\sigma_2^2}$ has a F distribution with $(n_1 - 1)$ which is numerator and $(n_2 - 1)$ which is denominator degrees of freedom.

A $(1-\alpha)100\%$ confidence interval for $\frac{\sigma_1^2}{\sigma_2^2}$ is given by:

$$\frac{s_1^2/s_2^2}{F_{1-(\alpha/2)}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2/s_2^2}{F_{\alpha/2}}$$

Hypothesis Testing for the Ratio of Two population variances:

$H_0: \sigma_1^2 = \sigma_2^2$

$H_0: \sigma_1^2 \geq \sigma_2^2$

$H_0: \sigma_1^2 \leq \sigma_2^2$

Here, the test statistic is,

$$F = \frac{s_1^2}{s_2^2}$$

In terms of F statistics,

$H_a: \sigma_1^2 \neq \sigma_2^2$

$H_a: \sigma_1^2 > \sigma_2^2$

$H_a: \sigma_1^2 < \sigma_2^2$

Inference Procedures of a Population Proportions:

Probability Distribution:

Important Properties of Normal Distribution:

- The distribution is symmetric about μ.
- The mean, median and mode of the distribution is equal.
- The mean of the distribution is μ and variance is σ^2 .
- The curve has a single peak for unimodal.
- $\mu \pm \sigma, \mu \pm 2\sigma, \mu \pm 3\sigma$ covers 68.27%, 95.45%, 99.73% area respectively.
- All odd central moments of the distribution are zero.
- Most of the distributions occurring in practice can be approximated by the normal distribution. Moreover, many of the sampling distributions e., g., Student's t, Snedecor's F, Chi-square distributions, etc. tend to normal for large samples.
- Normal distribution finds large applications in Statistical Quality Control in industry for setting control limits.
- Skewness is zero that is $\beta_1 = 0$ and kurtosis is 3 that is $\beta_2 = 3$.

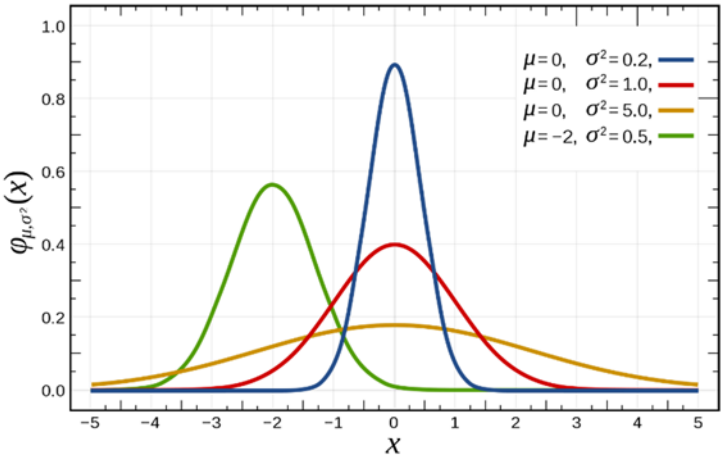

Fig. 3 Normal Distribution

Sampling and Sample Size:

With the help of simple random sampling without replacement, we can take samples from our population using the sample size determination formula.

Sample size determination for population proportion,

Formula 1:

$$n_0 = \frac{z^2pq}{d^2}$$

Formula 2:

$$n = \frac{n_0}{1 + \frac{n_0}{N}}$$

Here,

n_0 = estimated sample size.

z = statistical certainty chosen (1.96 for 5% level of significance).

p = estimated prevalence; (0.5 if unknown); q = 1-p.

d = width of the interval desired (precision).

n = desired sample size.

A Point Estimate:

For categorical data, we can estimate proportion.

For proportion,

$$\hat{p} = \frac{\mathrm{r}}{n}$$

Here, $\hat{p}$ = sample proportion

r = number of the attribute of interest

n = sample size

The Sampling Distribution of $\hat{p}$:

~ As the sample size increases, the sampling distribution of $\hat{p}$ becomes approximately normal.

~ The mean of the sampling distribution is p.

~ The standard deviation of the sampling distribution is $\sqrt{\frac{p(1-p)}{n}}$

Confidence Interval and Hypothesis Testing for a Population Proportion:

Confidence Interval:

~ A sample is drawn from the population of interest, and the sample proportion is computed.

~ Sample proportion, p̂ is an unbiased estimator of Population proportion, P.

Under these conditions:

~ If $\mu = np$ and *n(1-p)* are greater than 5, we assumed that the sampling distribution of p̂ comes from the normal distribution.

A $(1-\alpha)100\%$ confidence interval for P is given by:

$$\hat{p} \pm z_{(1-\alpha/2)}\sigma_{\hat{p}}$$

here, $\sigma_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Hypothesis Testing:

$H_0 : \hat{p} = p_0$

Here, the z statistic is

$$\mathrm{z} = \frac{\hat{\mathrm{p}} - \mathrm{p}_0}{\sqrt{\frac{\mathrm{p}_0(1-\mathrm{p}_0)}{\mathrm{n}}}}$$

In terms of Z statistics,

H_a: $\hat{\mathrm{p}} \neq \mathrm{p}_0$

H_a: $\hat{\mathrm{p}} \geq \mathrm{p}_0$

H_a: $\hat{\mathrm{p}} \leq \mathrm{p}_0$

Inference Procedures of two Population Proportions:

Confidence Interval and Hypothesis Testing for the Difference Between Two Population Proportions:

Confidence Interval:

- ~ Two samples are drawn from the population of interest, and the sample proportions are computed.
- ~ Sample proportions, $\hat{p}_1$and $\hat{p}_2$ are unbiased estimator of Population proportions, P_1 and P_2 .

Under these conditions:

- ~ When sample sizes, $n_1$and n_2 are large
- ~ The population proportions are not too close to 0 or 1.
- ~ The central limit theorem applies, and confidence intervals can be calculated using normal distribution theory.

A $(1-\alpha)100\%$ confidence interval for P_1 and P_2 is given by:

$$(\hat{p}_1 - \hat{p}_2) \pm z_{(1-\alpha/2)}\sigma_{(\hat{p}_1 - \hat{p}_2)}$$

here,

$$\sigma_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{n}}$$

Test Statistics:

Properties of Z test:

- ~ If data is normally distributed or sample size is large enough based on dataset, we can use Z-test.
- ~ Z-tests are related to t-tests, but we use t-tests when we have a small sample size.
- ~ t-test tends to z-test when sample size increases.

Hypothesis Testing:

Null Hypothesis, H_0: $\mathrm{p}_1 = \mathrm{p}_2$

Here, the z statistic is

$$\mathrm{z} = \frac{(\hat{\mathrm{p}}_1 - \hat{\mathrm{p}}_2) - (\mathrm{p}_1 - \mathrm{p}_2)_0}{\hat{\sigma}_{(\hat{\mathrm{p}}_1 - \hat{\mathrm{p}}_2)}}$$

here,

$$\hat{\sigma}_{(\hat{\mathrm{p}}_1 - \hat{\mathrm{p}}_2)} = \sqrt{\frac{\bar{\mathrm{p}}(1-\bar{\mathrm{p}})}{\mathrm{n}_1} + \frac{\bar{\mathrm{p}}(1-\bar{\mathrm{p}})}{\mathrm{n}_2}}$$

$$\bar{\mathrm{p}} = \frac{\mathrm{x}_1 + \mathrm{x}_2}{\mathrm{n}_1 + \mathrm{n}_2}$$

$$\sigma_{(\hat{\mathrm{p}}_1 - \hat{\mathrm{p}}_2)} = \sqrt{\frac{\hat{\mathrm{p}}_1(1-\hat{\mathrm{p}}_1)}{n} + \frac{\hat{\mathrm{p}}_2(1-\hat{\mathrm{p}}_2)}{n}}$$

In terms of Z,

H_a: $\mathrm{p}_1 \neq \mathrm{p}_2$

H_a: $\mathrm{p}_1 \geq \mathrm{p}_2$

H_a: $\mathrm{p}_1 \leq \mathrm{p}_2$

Dataset Description:

Carbon Dioxide Uptake in Grass Plants

Description:

The CO2 data frame has 84 rows and 5 columns of data from an experiment on the cold tolerance of the grass species $Echinochloa\ crus - galli$.

Usage:

CO2

Format:

An object of class

$c("nfnGroupedData", "nfGroupedData", "groupedData", "data.frame")$

containing the following columns:

Plant: an ordered factor with levels $Qn1 < Qn2 < Qn3 < Qc1 < Qc3 < Qc2 < Mn3 < Mn2 < Mn1 < Mc2 < Mc3 < Mc1$ giving a unique identifier for each plant.

Type: a factor with levels $Quebec\ and\ Mississippi$ giving the origin of the plant

Treatment: a factor with levels $nonchilled\ and\ chilled$

conc: a numeric vector of ambient carbon dioxide concentrations (mL/L).

uptake: a numeric vector of carbon dioxide uptake rates ($umol/m^2\ sec$).

Details:

The CO2 uptake of six plants from Quebec and six plants from Mississippi was measured at several levels of ambient CO2 concentration. Half the plants of each type were chilled overnight before the experiment was conducted.

This dataset was originally part of package $nlme$, and that has methods (including for $[as.data.frame, plot\ and\ print$) for its grouped-data classes.

Source:

Potvin, C., Lechowicz, M. J. and Tardif, S. (1990) "The statistical analysis of ecophysiological response curves obtained from experiments involving repeated measures", *Ecology,* **71**, 1389–1400.

Pinheiro, J. C. and Bates, D. M. (2000) *Mixed-effects Models in S and S-PLUS,* Springer.

Descriptive statistics:

Plant	Type	Treatment	conc	uptake
Qn1:7 Qn2:7 Qn3:7 Qc1:7 Qc3:7 Qc2:7 Mn3:7 Mn2:7 Mn1:7 Mc2:7 Mc3:7 Mc1:7	Quebec:42 Mississippi:42	nonchilled:42 chilled :42	Min.: 95 1st Qu.: 175 Median: 350 Mean: 435 Mean: 435 Max.:1000	Min.: 7.70 1st Qu.:17.90 Median:28.30 Mean:27.21 3rd Qu.:37.12 Max.:45.50

Conditioning Plot:

This plot refers to produces two variants of the conditioning plots. We can check pairwise relationship between conc and uptake conditional on a Plant.

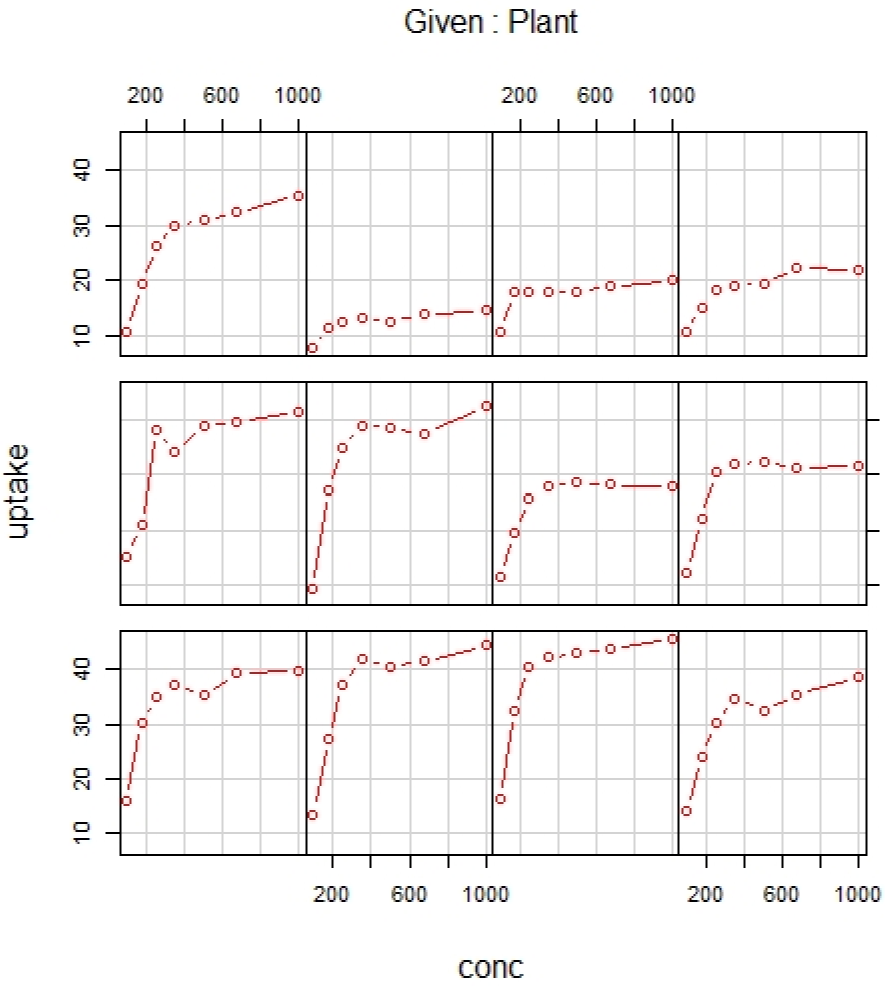

Fig. 4 Conditioning Plot

Scatter Plot:

We can see the relationship between our two quantitative variables from the scatter plot. And correlation between conc and uptake.

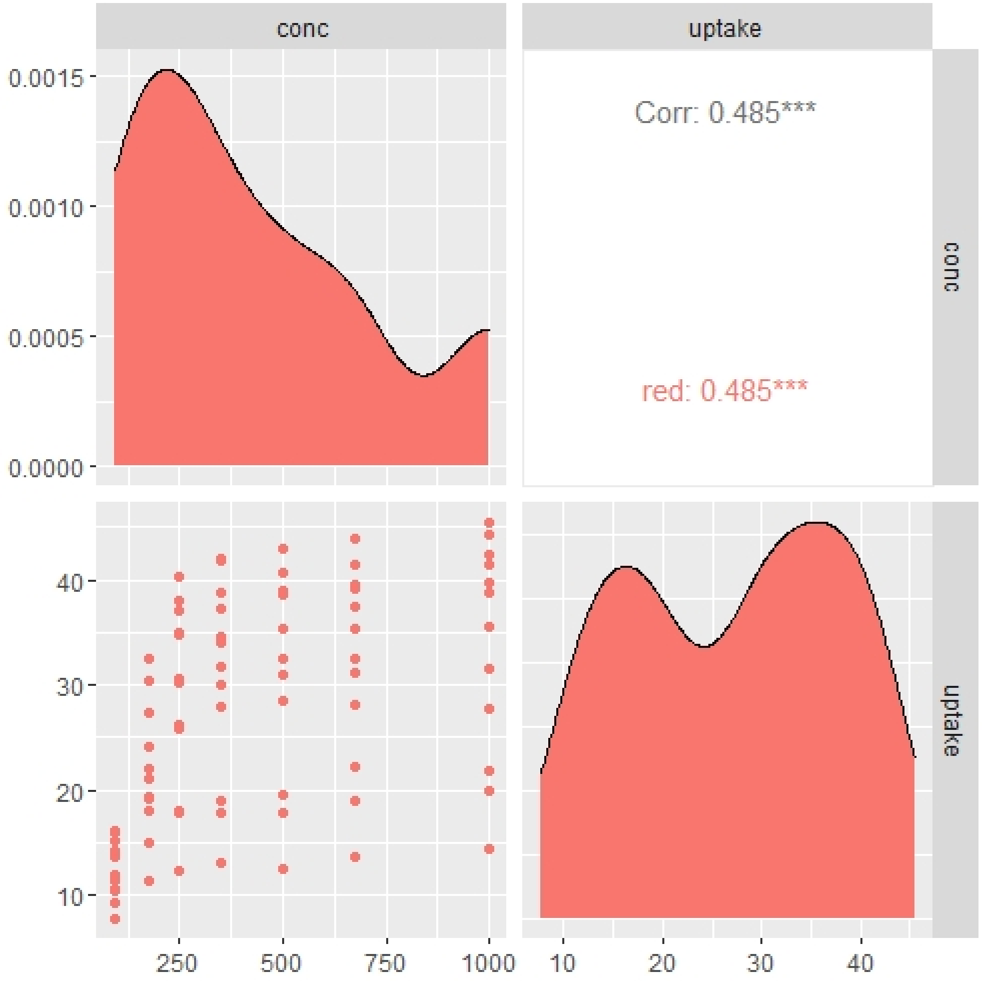

Fig. 5 Scatter Plot

Normality Checking:

For univariate normality checking, we can do a Normal Q-Q plot.

The steps leading to a Q–Q plot are as follows:

i. Order the original observations to get $x_{(1)}, x_{(2)}, \ldots, x_{(\mathrm{n})}$ and their corresponding probability values $\frac{(1-0.5)}{\mathrm{n}}, \frac{(2-0.5)}{\mathrm{n}}, \ldots, \frac{(\mathrm{n}-0.5)}{\mathrm{n}}$.

ii. Calculate the standard normal quantiles $\mathrm{q}_{(1)}, \mathrm{q}_{(2)}, \ldots, \mathrm{q}_{(\mathrm{n})}$.

iii. Plot the pairs of observations $(\mathrm{q}_{(1)}, x_{(1)}), (\mathrm{q}_{(2)}, x_{(2)}), \ldots, (\mathrm{q}_{(\mathrm{n})}, x_{(\mathrm{n})})$, and examine the "straightness" of the outcome.

The straightness of the Q−Q plot can be measured by calculating the correlation coefficient of the points in the plot. The pair of points $(\mathrm{q}_{(\mathrm{i})}, x_{(\mathrm{i})})$ lie very nearly along a 45-degree straight line, we can say the data is approximately normally distributed.

Here we check normality for conc and uptake variable.

For conc variable,

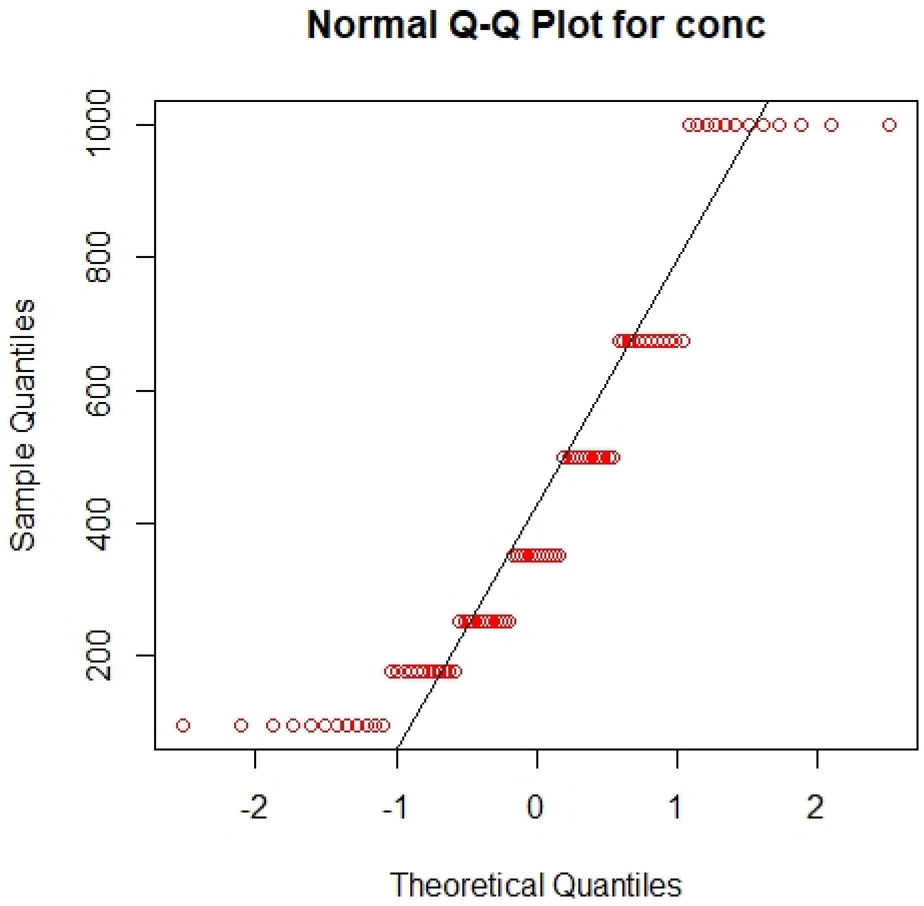

Fig. 6 Normal Q-Q Plot for conc

In this case, the straight line is far from the 45-degree, which means our conc variable is not normally distributed.

For uptake variable,

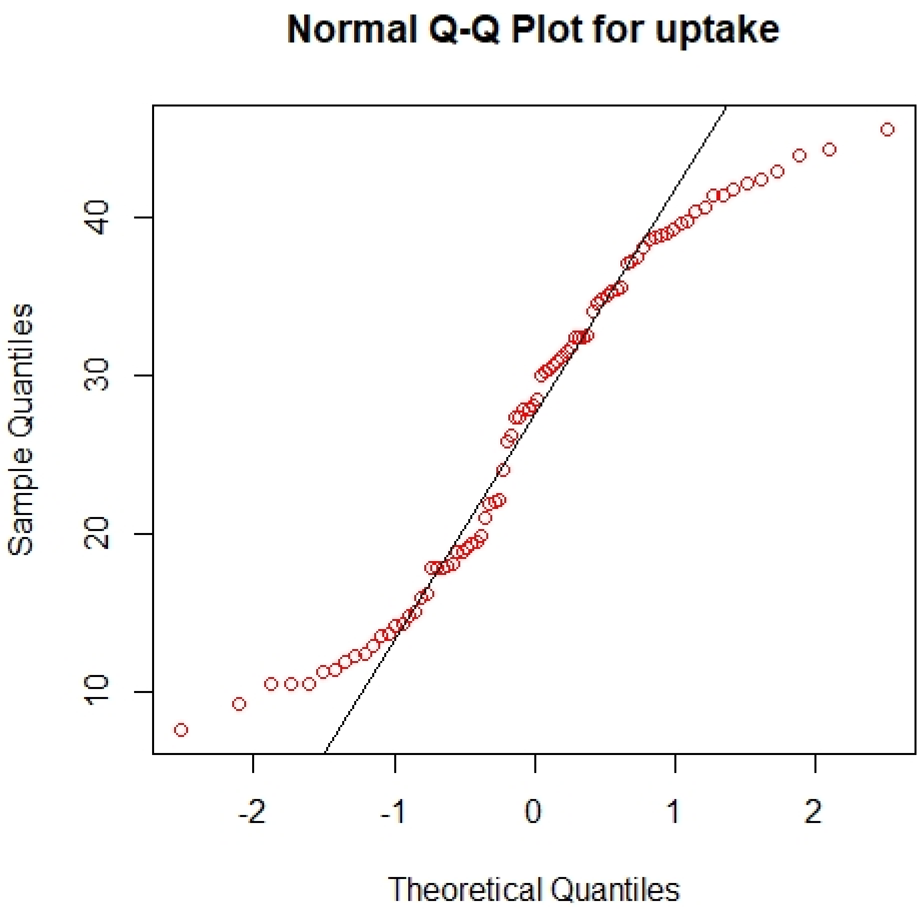

Fig. 7 Normal Q-Q Plot for conc

In this case, the straight line is far from the 45-degree, which means our uptake variable is not normally distributed.

In the next chapter, we discuss how we compute our inference procedure for population variances and proportions in R and interpretations.

Chapter 3

Practical Applications

Practical Applications with the data and interpretations:

Here, we import our built-in dataset “CO2” and we already discussed dataset and descriptive statistics with some plots as conditioning plot, Scatter Plot, Normal Q-Q plot in the previous chapter.

```
# Import built-in dataset
data("CO2")

# Descriptive statistics
summary(CO2)

# Conditioning Plots
library(stats)
library(graphics)
coplot(uptake ~ conc | Plant, data = CO2, show.given = FALSE, type = "b", col="red")

# Scatter plot
x<- CO2[,4:5]
x
library(ggplot2)
library(GGally)
ggpairs(data = x, mapping = aes(color="red"))

#normal Q-Q plot
qqnorm(CO2[,4], main = "Normal Q-Q Plot for conc", col="red")
qqline(CO2[,4])

qqnorm(CO2[,5], main = "Normal Q-Q Plot for uptake", col="red")
```

Sampling and Sample size determination:

```
#Estimated sample size

library(plotrix)
N<-length(CO2[,1])
z<- 1.96                          # (1.96 for 5% level of significance)
std<- sd(CO2$conc)           #population standard deviation
d<- std.error(CO2$conc)    #width of the interval desired (precision)
n1 <- z^2 * std^2/ d^2
n<- n1 / (1 + n1/N)

set.seed(12345)

# Sample for variances
s1<- sample(CO2$conc, size = n, replace = FALSE)
s2<- sample(CO2$uptake, size = n, replace = FALSE)

# Sample for Proportions
s3<- sample(CO2$Type, size = n, replace = FALSE)
```

After calculating sample size determination, our sample size is n=66 for all variables. And with the help of simple random sampling without replacement, we obtained our samples. The first sample is s1 drawn from the conc variable and the second sample is s2 drawn from the uptake variable which are quantitative variables.

So, the s1 sample,

1000, 175, 250, 250, 175, 500, 175, 175, 250, 250, 350, 500, 350, 675, 175, 95, 500, 250, 175, 1000, 675, 675, 1000, 675, 175, 350, 350, 675, 250, 675, 500, 350, 95, 250, 350, 95, 250, 500, 95, 500, 250, 175, 175, 350, 250, 95, 250, 500, 1000, 675, 95, 95, 675, 175, 1000, 1000, 500, 95, 1000, 350, 500, 175, 500, 675, 350, 95

And s2 sample,

19.5, 41.8, 35.5, 18.9, 13.0, 16.2, 39.7, 41.4, 31.1, 37.2, 12.5, 38.8, 11.3, 14.9, 32.4, 13.7, 40.3, 34.8, 42.1, 10.6, 35.0, 11.4, 37.1, 43.9, 40.6, 37.5, 17.9, 42.4, 39.2, 24.1, 30.9, 19.4, 27.3, 12.0, 10.5, 27.8, 38.9, 9.3, 34.6, 30.0, 38.7, 27.9, 19.2, 15.1, 13.6, 44.3, 18.1, 28.5, 45.5, 39.6, 41.4, 18.9, 42.9, 7.7, 10.6, 30.3, 35.3, 21.0, 27.3, 34.0, 21.9, 35.4, 14.4, 38.6, 32.5, 31.8

And s3 and s4 samples are drawn from Type and Treatment variables respectively, which are categorical variables.

So, the s3 sample,

Mississippi, Mississippi, Quebec, Quebec, Quebec, Quebec, Quebec, Quebec, Mississippi, Quebec, Quebec, Quebec, Mississippi, Mississippi, Mississippi, Quebec, Quebec, Mississippi, Mississippi, Mississippi, Quebec, Mississippi, Quebec, Quebec, Mississippi, Quebec, Mississippi, Mississippi, Quebec, Mississippi, Mississippi, Mississippi, Mississippi, Mississippi, Mississippi, Mississippi, Mississippi, Mississippi, Quebec, Mississippi, Quebec, Quebec, Quebec, Mississippi, Mississippi, Quebec, Quebec, Quebec, Mississippi, Mississippi, Mississippi, Mississippi, Quebec, Quebec, Mississippi, Quebec, Quebec, Mississippi, Quebec, Mississippi, Quebec, Quebec, Quebec, Mississippi, Quebec, Mississippi

Levels: Quebec Mississippi

And s4 sample,

Nonchilled, nonchilled, nonchilled, chilled, chilled, chilled, chilled, chilled, nonchilled, chilled, nonchilled, nonchilled, nonchilled, nonchilled, chilled, chilled, nonchilled, chilled,

chilled, chilled, chilled, nonchilled, nonchilled, nonchilled, nonchilled, chilled, chilled, nonchilled, nonchilled, nonchilled, chilled, nonchilled, chilled, nonchilled, chilled, chilled, chilled, chilled, chilled, chilled, nonchilled, chilled, chilled, nonchilled, nonchilled, nonchilled, chilled, nonchilled, nonchilled, nonchilled, chilled, nonchilled, chilled, chilled, nonchilled, nonchilled, nonchilled, chilled, chilled, chilled, nonchilled, chilled, nonchilled, chilled, chilled, chilled

Levels: nonchilled chilled

Inference Procedures for Population Variances:

Estimation:

```
#point estimate
mean(s1)
mean(s2)
```

From Central limit theorem, we assumed that our sample is large enough and normally distributed. In normal distribution, the mean is more efficient than other estimation process.

The mean of s1 sample is 405.3788

And the mean of s2 sample is 28.0303

Confidence Interval for the Variance of a Normally Distributed Population:

```
#CI for a population variance
library(EnvStats)
varTest(s1, conf.level = 0.95)
```

For one sample s1 variance, we can use EnvStats package to Chi-Squared Test on Variance.

Here,

Chi-Squared = 4962816

df = 65

p-value = 2.2e-16

95 percent confidence interval:

55651.2	111266.4

sample estimates: variance = 76351.01

alternative hypothesis: true variance is not equal to 1

Confidence Interval for the Ratio of the Variances of two Normally Distributed Populations

```
#CI for two population variances
library(dplyr)
var.test(s1, s2, alternative = "two.sided")
```

For two sample s1 and s2 variances, we can use dplyr package to F Test to compare two Variances. Here,

F = 579.42

df = 65

p-value = 2.2e-16

95 percent confidence interval:

354.7917	946.2675

sample estimates: Ratio of variances = 579.4203

alternative hypothesis: true ratio of variances is not equal to 1

Inference Procedures for Population Proportions:

Estimation:

```
#point estimate
#s3
table1<-table(s3)
table1
prop1 <- prop.table(table1)
cat_M = prop1["Mississippi"]
cat_M
x_value= cat_M*N
x_value

#s4
table2<-table(s4)
table2
prop2 <- prop.table(table2)
cat_c = prop2["chilled"]
cat_c
xValue= cat_c*N
xValue
```

From Central limit theorem, we assumed that our sample is large enough and normally distributed. Proportion is used for an estimate in the categorical variable.

The proportion of s3 sample:

Quebec	Mississippi
0.4848485	0.5151515

The x value of Mississippi category is 43.27273

And the proportion of s4 sample:

nonchilled	chilled
0.469697	0.530303

The x value of chilled category is 44.54545

Confidence Interval for a Population Proportion:

```
#CI for a population proportion
#s3
prop.test(x=x_value,n=N,conf.level = 0.95)
```

For one sample s3 proportion, we calculate confidence interval for Mississippi category.

Here,

X-squared = 0.028434

p-value = 0.8661

95 percent confidence interval:

0.4042531	0.6246525

sample estimates: p = 0.5151515

alternative hypothesis: true p is not equal to 0.5

Confidence Interval for the Difference Between Two Population Proportions:

```
#CI for two population proportions
two_x<-c(x_value,xValue)
total_length<-c(N,N)
prop.test(two_x,n=total_length,conf.level = 0.95)
```

For two sample s3 and s4 proportion, we calculate confidence interval for Mississippi and chilled category respectively.

Here, X-squared = 0.0017746

p-value = 0.9664

95 percent confidence interval:

-0.1780971	0.1477941

sample estimates:

Prop 1	Prop 2
0.5151515	0.5303030

The two proportion can be equal because 0 lies in confidence interval.

Chapter 4
Conclusions

Conclusions:

In previous chapters, we have seen the inference procedures of population variances and proportions. We talked about basic Statistical Inference. After that, we discussed mathematical approaches to how we compute our results. In addition, we calculated some procedures in the R programming language with results for our dataset.

Also, we have seen probability distribution, sampling distribution, Central Limit Theorem, Normality checking, sampling, and sample size determination.

Based on our results, we can conclude our population and make decisions. Thus, we describe our population parameter on the basis of sample statistic.

We can solve real-life problems related to the Statistical Inference of Variances and Proportions by following discussed procedures in previous chapters.

References

Caffo,B. (2016). Statistical inference for data science. Retrieved from https://leanpub.com/LittleInferenceBook/read

Igual,L. Segui,S. (2017). Introduction to Data Science. Retrieved from https://link.springer.com/book/10.1007/978-3-319-50017-1

Rohatgi,V.K., Saleh,A.K.E. (1939). An Introduction to Probability and Statistics. Retrieved from https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwiB7IGe7LD2AhVgyzgGHbJ2C_EQFnoECAQQAQ&url=https%3A%2F%2Fwww.usb.ac.ir%2FFileStaff%2F7344_2018-12-2-10-44-10.pdf&usg=AOvVaw1dhSpOECUDNcBksj3q0P2E

Acknowledgements

First and foremost, I would like to express my gratitude to my professor, Dr. Tapati Basak, whose expertise is invaluable in formulating the teaching procedure and methodology. Your informative remarks encouraged me to improve my thoughts and raise the quality of my work.

I would like to acknowledge my supervisor, Professor Dr. Md. Abdus Salam and Chairman, Professor Dr. Mohammad Alamgir Kabir, Professor Dr. Ajit Kumar Majumder, Associate Professor Md. Habibur Rahman and Assistant Professor Nasrin Khatun, for their valuable guidance throughout my studies.