Statistical Inference

Jerry Kiely

2023-11-25

Contents

1 Introduction **2**
- 1.1 Basic terms . . . 2
- 1.2 What is Statistical Inference? . . . 2
- 1.3 Three branches . . . 2
- 1.4 Parameter, Estimator, Estimate . . . 2
- 1.5 Sampling distribution / error . . . 2
- 1.6 Hypothesis Testing . . . 2
- 1.7 Types of Error . . . 3
- 1.8 One-tailed and two-tailed tests . . . 3

2 Parametric Tests **4**
- 2.1 Normality Tests . . . 4
- 2.2 t-distribution . . . 6
- 2.3 Degrees of freedom . . . 6
- 2.4 One sample t-test . . . 6
- 2.5 Independent samples t-test . . . 7
- 2.6 Paired samples t-test . . . 8
- 2.7 t-test for correlation . . . 9

3 Tests for Equality of Variances **10**
- 3.1 F-test . . . 10

4 Analysis of Variance **11**
- 4.1 What is Analysis of Variance . . . 11
- 4.2 One Way ANOVA . . . 11
- 4.3 Two Way ANOVA . . . 12
- 4.4 Three Way ANOVA . . . 12

5 Non-Parametric Tests **13**
- 5.1 Wilcoxon Signed Rank Test . . . 13
- 5.2 Mann-Whitney Test . . . 13
- 5.3 Wilcoxon Signed Rank Test For Paired Data . . . 14
- 5.4 Kruskal Wallis Test . . . 14
- 5.5 Chi-Square Test Of Association . . . 14

6 Summary **16**
- 6.1 Table . . . 16

1 Introduction

1.1 Basic terms

- variable
- population
- sample
- distribution
- factor
- descriptive statistics

1.2 What is Statistical Inference?

- drawing conclusions about unknown population properties
- based on samples drawn from the population
- such as mean, proportion, variance, etc.
- unknown population properties - known as parameters

1.3 Three branches

- point estimation
- interval estimation
- testing of hypothesis

1.4 Parameter, Estimator, Estimate

- parameter - unknown property or characteristic of population
- estimator - rule / function based on sample observations used to estimate parameter
- estimate - value computed from estimator

1.5 Sampling distribution / error

- sampling distribution of sample means
- standard error is the standard deviation of the sample means

1.6 Hypothesis Testing

- hypothesis - assertion of distribution / parameter of one or more random variables
- null hypothesis (H_0) - assertion believed to be true until rejected
- alternative hypothesis (H_1) - claim that contradicts H_0

You test a hypothesis to decide if a statement / hypothesis about a population parameter is true based on sample data.

- test statistic - the statistic on which the decision to reject the null hypothesis is defined
- critical / rejection region - the region within which, if the value of the test statistic falls, the null hypothesis is rejected

1.7 Types of Error

- Type I Error
- Type II Error

	H_0 is true	H_0 is false
reject H_0	Type I Error	Correct
fail to reject H_0	Correct	Type II Error

- level of significance - probability of Type I Error (α)
- generally set at 5% or 0.05
- p-value - smallest level of significance that would lead to rejection of H_0
- H_0 rejected if observed risk (or p-value) is less than level of significance
- α = Probability [Type I Error] = Probability [Reject H_0 | H_0 is True]
- β = Probability [Type II Error] = Probability [Do not reject H_0 | H_0 is not True]
- power of the test - $1 - \beta$

1.8 One-tailed and two-tailed tests

One-tailed test:

$H_0 : \mu = \mu_0$

$H_0 : \mu > \mu_0$ (right-tailed) or $H_0 : \mu < \mu_0$ (left-tailed)

Two-tailed test:

$H_0 : \mu = \mu_0$

$H_0 : \mu \neq \mu_0$

2 Parametric Tests

2.1 Normality Tests

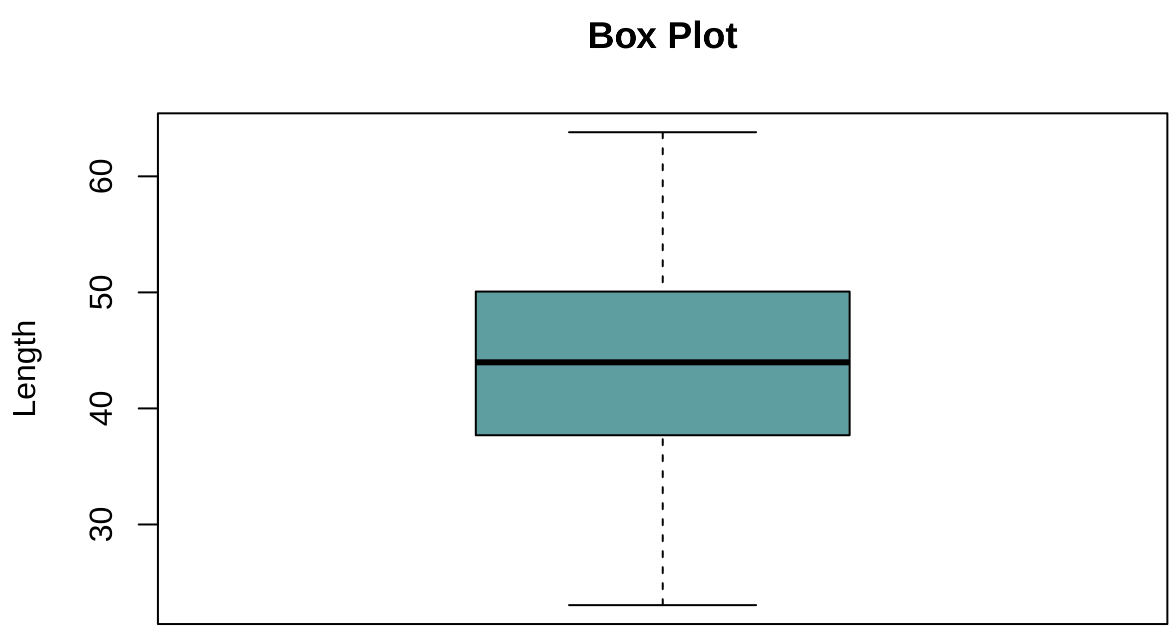
A prerequisite for many statistical tests - normal data is an underlying assumption in parametric tests. Normality can be assessed using two approaches:

- graphical
  - Box-Whisker plot
  - Q-Q plot
- numerical
  - Shapiro-Wilk test (small samples)
  - Kolmogorov-Smirnov test (large samples)

2.1.1 Box-Whisker Plot

Powerful visual method for assessing symmetry.

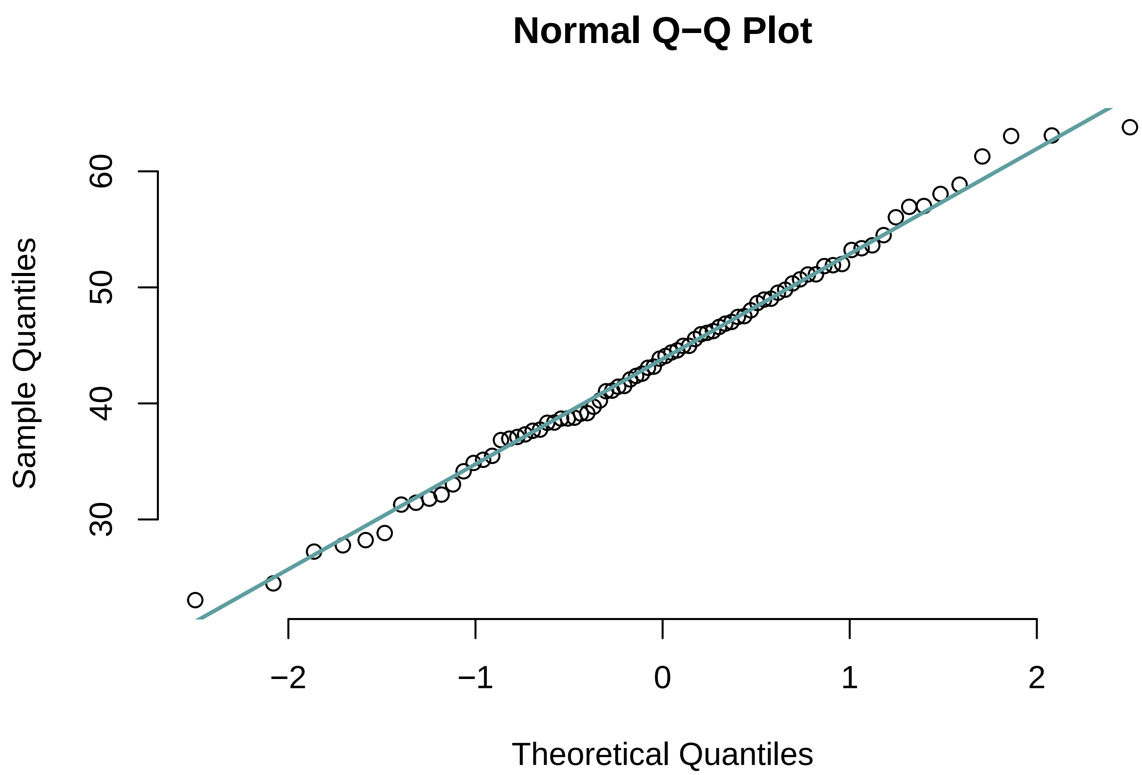
```r
boxplot(data$csi, main = "Box Plot", ylab = "Length", col = "cadetblue")
```



2.1.2 Q-Q Plot

Powerful visual method for assessing normality.

```r
qqnorm(data$csi, pch = 1, frame = FALSE)
qqline(data$csi, col = "cadetblue", lwd = 2)
```



2.1.3 Shapiro-Wilk test

A widely used test for assessing normality.

```r
shapiro.test(data$csi)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$csi
## W = 0.99196, p-value = 0.9038
```

2.1.4 Kolmogorov-Smirnov test

Another widely used test for assessing normality.

```
lillie.test(data$csi)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data$csi
## D = 0.042387, p-value = 0.9764
```

2.2 t-distribution

- symmetric
- resembles bell shape of the normal distribution
- as the sample size increases, as the degrees of freedom increases, it approaches the normal distribution with mean 0 and variance 1

2.3 Degrees of freedom

- the number of independent terms
- n values would have $n-1$ degrees of freedom
- $S = x_1 + x_2 + x_3 + x_4 + x_5 \implies x_1 = S - (x_2 + x_3 + x_4 + x_5)$

2.4 One sample t-test

- test a hypothesis about a single population mean
- a single sample drawn from a defined population
- compare sample statistic to hypothesized value of a population parameter

The assumptions of the one sample t-tests:

- random sampling from a defined population
- population is normally distributed
- variable under study is continuous

Normality tests can be performed using any of the methods described previously. The validity of the test is not significantly affected by moderate deviations from the normality assumption.

```r
t.test(data$Time, alternative = "greater", mu = 90)
```

```
##
##  One Sample t-test
##
## data:  data$Time
## t = 1.9176, df = 11, p-value = 0.04074
## alternative hypothesis: true mean is greater than 90
## 95 percent confidence interval:
##  90.22748      Inf
## sample estimates:
## mean of x
##  93.58333
```

2.5 Independent samples t-test

- compares means of two independent groups on the same continuous variable
- hypothesis tested
  - H_0: $\mu_1 = \mu_2$
  - H_1: $\mu_1 \neq \mu_2$

The assumptions of the independent samples t-test:

- samples drawn are random samples
- populations from which samples are drawn have equal and unknown variances
- populations follow normal distribution

Normality tests can be performed using any of the methods described previously.

```r
t.test(data$time_g1, data$time_g2, alternative = "two.sided", var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  data$time_g1 and data$time_g2
## t = 0.22346, df = 24, p-value = 0.8251
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.216185  5.239994
## sample estimates:
## mean of x mean of y
##  93.58333  93.07143
```

```r
t.test(data$time_g1, data$time_g2, alternative = "two.sided", var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  data$time_g1 and data$time_g2
## t = 0.21966, df = 21.103, p-value = 0.8282
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.333074  5.356884
## sample estimates:
## mean of x mean of y
##  93.58333  93.07143
```

2.6 Paired samples t-test

- used to determine if the mean difference between two sets of observations is 0
- each subject is measured twice - paired observations
- typically before / after
- hypothesis tested
  - H_0: $\mu_1 - \mu_2 = 0$
  - H_1: $\mu_1 - \mu_2 \neq 0$

The assumptions of the paired sample t-test:

- random sampling from a defined population
- population is normally distributed

A normality test can be performed using any of the methods described previously. The validity of the test is not significantly affected by moderate deviations from the normality assumption.

```r
t.test(data$time_before, data$time_after, alternative = "greater", paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  data$time_before and data$time_after
## t = 8.2295, df = 14, p-value = 4.919e-07
## alternative hypothesis: true mean difference is greater than 0
## 95 percent confidence interval:
##  6.549798      Inf
## sample estimates:
## mean difference
##        8.333333
```

2.7 t-test for correlation

- correlation coefficient summarizes the strength of a linear relationship between two variables
- a t-test is used to test if there is a significant correlation between two variables
- sample correlation coefficient is calculated using bivariate data
- hypothesis tested
  - H_0: there is no significant correlation between two variables under study ($\rho = 0$)
  - H_1: there is correlation between two variables under study ($\rho \neq 0$)

```
cor.test(data$aptitude, data$job_prof, alternative = "two.sided", method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  data$aptitude and data$job_prof
## t = 2.8769, df = 23, p-value = 0.008517
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1497097 0.7558981
## sample estimates:
##       cor
## 0.5144107
```

3 Tests for Equality of Variances

3.1 F-test

- used to test the equality of two population variances
- a prerequisite for many statistical tests
- hypothesis tested
  - H_0: $\sigma_1^2 = \sigma_2^2$
  - H_1: $\sigma_1^2 \neq \sigma_2^2$

The assumptions of the F-test:

- random sampling from a defined population
- population is normally distributed

The F-test is used to validate the assumption of the equality of variances. The parent population is assumed to follow a normal distribution.

```r
var.test(data$time_g1, data$time_g2, alternative = "two.sided")
```

```
##
##  F test to compare two variances
##
## data:  data$time_g1 and data$time_g2
## F = 1.5434, num df = 11, denom df = 13, p-value = 0.4524
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.4826988 5.2348866
## sample estimates:
## ratio of variances
##          1.543428
```

4 Analysis of Variance

4.1 What is Analysis of Variance

- a collection of statistical models used to analyze the difference among more than two group means
- developed by Ronald Fisher
- variance due to
  - assignable causes
  - chance causes
- ANOVA is the separation of variance ascribable to one group of causes from the variance ascribable to another

The assumptions of ANOVA:

- samples drawn are random samples
- populations from which samples are drawn have equal and unknown variances
- populations follow normal distribution

A normality test can be performed using any of the methods described previously.

4.2 One Way ANOVA

- an extension of the t-test for independent samples
- used to test equality of K population means
  - when $K = 2$ t-Test can be used
  - when $K = 2$ t-Test and one way ANOVA provide identical results
- hypothesis tested
  - H_0: $\mu_1 = \mu_2 = ... = \mu_K = \mu$
  - H_1: $\mu_i \neq \mu_j, i \neq j$

```
anova <- aov(satindex ~ dept, data = data)
summary(anova)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## dept         2  220.1  110.03   2.308  0.115
## Residuals   34 1620.9   47.67
```

4.3 Two Way ANOVA

- used when there are 2 factors under study
- each factor can have 2 or more levels
- three hypothesis tested
  - Factor A
    - H_{01}: all group means are equal
    - H_{11}: at least one mean is different from other means
  - Factor B
    - H_{02}: all group means are equal
    - H_{12}: at least one mean is different from other means
  - Interaction
    - H_{03}: the interaction is not significant
    - H_{13}: the interaction is significant

```
anova <- aov(satindex ~ dept*exp, data = data)
summary(anova)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## dept         2  164.2   82.11   1.679  0.204
## exp          1   78.0   78.03   1.595  0.216
## dept:exp     2   20.2   10.11   0.207  0.814
## Residuals   30 1467.2   48.91
```

4.4 Three Way ANOVA

- two way ANOVA can be extended to assess the effects of three or more factors
- with three factors A, B, and C we look at
  - the effects of A, B, and C
  - two way interactions - A*B, A*C, and B*C
  - three way interaction - A*B*C

```
anova <- aov(growth ~ campaign*region*size, data = data)
summary(anova)
```

```
##                      Df Sum Sq Mean Sq F value   Pr(>F)
## campaign              2  1.817   0.909  24.475 2.71e-08 ***
## region                2 24.656  12.328 332.024  < 2e-16 ***
## size                  1  0.009   0.009   0.239   0.6266
## campaign:region       4  1.102   0.275   7.418 7.75e-05 ***
## campaign:size         2  0.370   0.185   4.986   0.0103 *
## region:size           2  0.175   0.088   2.360   0.1041
## campaign:region:size  4  0.221   0.055   1.485   0.2196
## Residuals            54  2.005   0.037
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5 Non-Parametric Tests

- tests based on t and F distributions assume populations are normally distributed
- large body of statistical methods which do not make assumptions about the population distribution
- non-parametric or distribution-free tests
- if underlying normality assumption is met then parametric test should be chosen

5.1 Wilcoxon Signed Rank Test

- considered a non-parametric alternative to one sample t-test
- used to determine if the mean / median of a sample is equal to a known value when the variable is ordinal or continuous, but not normally distributed
- hypothesis tested
  - H_0: the mean / median of the population is m_0
  - H_1: the mean / median of the population is not m_0

```
wilcox.test(data$Score, mu = 50, alternative = "less")
```

```
##
##  Wilcoxon signed rank exact test
##
## data:  data$Score
## V = 54, p-value = 0.8833
## alternative hypothesis: true location is less than 50
```

5.2 Mann-Whitney Test

- considered a non-parametric alternative to independent samples t-test
- used to compare differences between two independent groups when the dependent variable is either ordinal or continuous, but not normally distributed
- equivalent to Wilcoxon rank-sum test (WRS)
- hypothesis tested
  - H_0: the distributions of both groups are identical
  - H_1: the distributions of both groups are not identical

```
wilcox.test(aptscore ~ Group, data = data)
```

```
##
##  Wilcoxon rank sum exact test
##
## data:  aptscore by Group
## W = 18, p-value = 0.7308
## alternative hypothesis: true location shift is not equal to 0
```

5.3 Wilcoxon Signed Rank Test For Paired Data

- considered a non-parametric alternative to paired t-test
- used to compare differences between two related or paired groups when the dependent variable is either ordinal or continuous, but not normally distributed
- hypothesis tested
  - H_0: the median of difference in the population is 0
  - H_1: not H_0

```r
wilcox.test(data$Before, data$After, paired = TRUE, alternative = "less")
```

```
##
##  Wilcoxon signed rank exact test
##
## data:  data$Before and data$After
## V = 4, p-value = 0.001709
## alternative hypothesis: true location shift is less than 0
```

5.4 Kruskal Wallis Test

- considered a non-parametric alternative to one way ANOVA
- used to compare differences between more than two independent groups when the dependent variable is either ordinal or continuous, but not normally distributed
- hypothesis tested
  - H_0: K samples come from the same population
  - H_1: not H_0

```r
kruskal.test(aptscore ~ Group, data = data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  aptscore by Group
## Kruskal-Wallis chi-squared = 2.2309, df = 2, p-value = 0.3278
```

5.5 Chi-Square Test Of Association

- also known as Pearson's Chi-Square Test
- used to test if there is a relationship between two categorical variables
- the two categorical variables can be nominal or ordinal
- hypothesis tested
  - H_0: the two attributes are independent
  - H_1: not H_0

```r
CrossTable(data$performance, data$source, chisq = TRUE)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  870
##
##
##                  | data$source
## data$performance |    Campus |  Internal | Jobportal | Row Total |
## -----------------|-----------|-----------|-----------|-----------|
##        Excellent |       150 |       100 |        40 |       290 |
##                  |    14.545 |     3.333 |    33.218 |           |
##                  |     0.517 |     0.345 |     0.138 |     0.333 |
##                  |     0.455 |     0.400 |     0.138 |           |
##                  |     0.172 |     0.115 |     0.046 |           |
## -----------------|-----------|-----------|-----------|-----------|
##             Good |       100 |       100 |       100 |       300 |
##                  |     1.672 |     2.207 |     0.000 |           |
##                  |     0.333 |     0.333 |     0.333 |     0.345 |
##                  |     0.303 |     0.400 |     0.345 |           |
##                  |     0.115 |     0.115 |     0.115 |           |
## -----------------|-----------|-----------|-----------|-----------|
##             Poor |        80 |        50 |       150 |       280 |
##                  |     6.467 |    11.531 |    34.405 |           |
##                  |     0.286 |     0.179 |     0.536 |     0.322 |
##                  |     0.242 |     0.200 |     0.517 |           |
##                  |     0.092 |     0.057 |     0.172 |           |
## -----------------|-----------|-----------|-----------|-----------|
##     Column Total |       330 |       250 |       290 |       870 |
##                  |     0.379 |     0.287 |     0.333 |           |
## -----------------|-----------|-----------|-----------|-----------|
##
##
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 =  107.3786     d.f. =  4     p =  2.635987e-22
##
##
##
```

6 Summary

6.1 Table

Parametric	Non-Parametric
one sample t-test	Wilcoxon signed rank test
independent samples t-test	Mann-Whitney test
paired samples t-test	Wilcoxon signed rank test for paired data
one way ANOVA	Kruskal Wallis test
two way ANOVA	
three way ANOVA	