Inductive Statistics
Formulary

1. Sampling Distribution

Sampling distribution of the population mean:

$$\mu_{\bar{x}} = \mu$$

μ: Mean of the population

$\mu_{\bar{x}}$: Mean of the sampling distribution of the mean.

$$\sigma^2_{\bar{x}} = \begin{cases} \dfrac{\sigma^2}{n} & \text{with replacement} \\ \dfrac{\sigma^2}{n} \cdot \left(\dfrac{N-n}{N-1} \right) & \text{without replacement} \end{cases}$$

Shape of the sampling distribution of the mean

1.
If a sample selected from a *normally distributed population* then the distribution of the sample mean will also be *normal* with

$$\mu_{\bar{x}} = \mu, \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

2.
If a "large" sample ($n \geq 30$) is selected from "*any*" *population*, then the distribution of the mean will be *approximately normal* with

$$\mu_{\bar{x}} = \mu, \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

Sampling Distribution of Proportion

The probability distribution of the sample distribution, $\bar{p}$, is called its *sampling distribution of proportion.*

$$\mu_{\bar{P}} = P,$$

$$\sigma_{\bar{p}} = \begin{cases} \sqrt{\dfrac{P(1-P)}{n}} & \text{,when} \quad \dfrac{n}{N} \le 0.05 \\ \sqrt{\dfrac{P(1-P)}{n}} \cdot \sqrt{\dfrac{N-n}{N-1}} & \text{,when} \quad \dfrac{n}{N} > 0.05 \end{cases}$$

<u>Shape of the Sampling Distribution of the Sample Proportion</u>

The sampling distribution of the sample proportion, $\bar{p}$, is approximately normally distributed for a sufficiently large sample size.

<u>Rule of thumb 1</u>: The sample size is considered to be sufficiently large if

$$n \cdot P > 5 \ \wedge \ n \cdot (1-P) > 5.$$

<u>Rule of thumb</u> 2:

$$n \ge 30.$$

The $z-$value for a value of $\bar{p}$ is calculated as

$$z = \frac{\bar{p} - \mu_{\bar{p}}}{\sigma_{\bar{p}}}.$$

2. Estimation Theory

Point Estimates, Interval Estimates

An estimate of a population parameter given by a single number is called a *point estimate* of the parameter.
An estimate of a population parameter given by two numbers between which the parameter may be considered to lie is called an *interval estimate* of the parameter.

Interval Estimation of the population Mean

Population	Sample size	σ known	σ unknown
Normally distributed	Large $(n \geq 30)$	$\bar{x} \pm z\sigma_{\bar{x}}$	$\bar{x} \pm ts_{\bar{x}}$ or $\bar{x} \pm zs_{\bar{x}}^{**}$
	Small $(n < 30)$	$\bar{x} \pm z\sigma_{\bar{x}}$	$\bar{x} \pm ts_i$
Not normally distributed	Large $(n \geq 30)$	$\bar{x} \pm z\sigma_{\bar{x}}^{*}$	$\bar{x} \pm ts_{\bar{x}}^{*}$ or $\bar{x} \pm zs_{\bar{x}}^{***}$
	Small $(n < 30)$	Nonparametric procedures directed toward the median generally would be used.	

* Central limit theorem is invoked.
** z is used as an approximation of t.
*** Central limit theorem is invoked, and z is used as an approximation of t.

The minimum required sample size:

$$n = \left(\frac{z \cdot \sigma}{E} \right)^2 .$$

E: maximum error of estimate.

Estimation of Population Proportion: Large Samples
A sample is *large* if

$$n \cdot p > 5 \ \wedge \ n \cdot (1 - p) > 5.):$$

1. The sampling distribution of the sample proportion, p, is (approximately) normal.
2. The mean, μ_p, of the sampling distribution of p is equal to the population proportion, P:

$$\mu_p = P$$

3. The standard deviation of the sample proportion, p, is

$$\sigma_p = \sqrt{\frac{P(1-P)}{n}}.$$

When estimating the value of a population proportion, we do not know the value of P. Consequently, we cannot compute σ_p. Therefore, in the estimation of a population proportion, we use the value of s_p as an estimate of σ_p.

<u>Confidence Interval for the Proportion, P</u>

The $(1-\alpha)100\%$ confidence interval for the population proportion, P, is

$$p \pm zs_p.$$

The value of z used here is obtained from the standard normal distribution table for the given confidence level, and $s_p \coloneqq \sqrt{\frac{p \cdot (1-p)}{n}}$.

(The term zs_p is called the *margin of error, E*.)

3.Testing Hypotheses

Null Hypothesis:
A *null hypothesis* is a statement about a population parameter that is assumed to be true until it is declared false.
The null hypothesis will be denoted by H_0.

Alternative Hypothesis:
An *alternative hypothesis* is a statement about the population parameter that will be true if the null hypothesis is false.
The alternative hypothesis will be denoted by H_1.

Tapes of error:

Type I Error
A *type I error* occurs when a true null hypothesis is rejected. The value of α represents the probability of committing this type of error, that is,

$$\alpha = P(H_0 \text{ is rejected}/ H_0 \text{ is true})$$

The value of α represents *significance level* of the test.

Type II Error
A *type II error* occurs when a false null hypothesis is not rejected. The value of β represents the probability of committing this type of error, that is,

$$\beta = P(H_0 \text{ is not rejected}/ H_0 \text{ is false})$$

The value of $1-\beta$ is called *power of the test*. It represents the probability of not making a type II error.

Steps to Perform a Test of Hypothesis with the Critical-Value Approach:

1. State the null and alternative hypothesis.
2. Select
    i. *The normal distribution if σ is known*
    ii. *t distribution if σ is unknown*

3. Determine the rejection and non-rejection regions by calculating by calculating

   z_{crit} if σ is known and t_{crit} if σ is unknown

4. Calculate the value of the test statistic.

$$z_{stat} = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \quad \text{where} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}, \quad \text{if } \sigma \text{ is known}$$

$$t_{stat} = \frac{\bar{x} - \mu}{s_{\bar{x}}} \quad \text{where} \quad s_{\bar{x}} = \frac{s}{\sqrt{n}}, \quad \text{if } \sigma \text{ is unknown}$$

5. Compare z_{stat} with z_{crit} or t_{stat} with t_{crit} and make a decision.

<u>Steps to Perform a Test of Hypothesis with the p-Value Approach:</u>

1. Select the null and alternative hypothesis.
2. Select

   i. The normal distribution if σ is known

   with $z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$ where $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

   ii. t distribution if σ is unknown

   with $t = \frac{\bar{x} - \mu}{s_{\bar{x}}}$ where $s_{\bar{x}} = \frac{s}{\sqrt{n}}$.

3. Calculate the *p-value*.
4. Reject the null hypothesis if $p-value < \alpha$. Do not reject the null hypothesis if $p-value \geq \alpha$

Tests about a Population Proportion

Steps to Perform a Test of Hypothesis with the Critical-Value Approach:

1. State the null and alternative hypothesis.
2. Use the normal distribution if

$$n.P > 5 \quad \wedge \quad n \cdot (1 - P) > 5 .$$

3. Find $z_{critical}$.

4. Compute

$$z_{statistic} = \frac{p - P}{\sigma_p} \quad \text{wehere} \quad \sigma_p = \sqrt{\frac{P \cdot (1 - P)}{n}}.$$

5. Make a decision by comparing $z_{critical}$ with $z_{statistic}$

Steps to Perform a Test of Hypothesis with the p-Value Approach:

1. State the null and alternative hypothesis.
2. Use the normal distribution if

$$n.P > 5 \quad \wedge \quad n \cdot (1 - P) > 5 .$$

3. Find the *p-value*

4. Reject the null hypothesis if $p - value < \alpha$. Do not reject the null hypothesis if $p - value \geq \alpha$

4. Simple Linear Regression

Simple Linear Regression Model

$$y = \beta_0 + \beta_1 x + \varepsilon .$$

Simple Linear Regression Equation

$$E(y) = \beta_0 + \beta_1 x.$$

Estimated linear Regression Equation

$$y^* = b_0 + b_1 x .$$

Normal Equations:

$$\begin{cases} n \cdot b_0 + b_1 \cdot \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ b_0 \cdot \sum_{i=1}^{n} x_i + b_1 \cdot \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i \cdot y_i \end{cases}$$

Important Sums:

$$SST = SSR + SSE$$

$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$: Total sum of squares

$SSR = \sum_{i=1}^{n} (y_i^* - \bar{y})^2$: Sum of square due to regression

$SSE = \sum_{i=1}^{n} (y_i - y_i^*)^2$: Sum of squares due to error

Coefficient of Determination

$$r^2 := \frac{SSR}{SST}$$

Simple Correlation Coefficient

$$r := (\text{sign of } b_1)\sqrt{r^2}, \quad -1 \le r \le 1.$$

or

$$r = \frac{n \cdot \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left(n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2 \right) . \left(n \cdot \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i \right)^2 \right)}}$$

Test for Significance in Simple Linear Regression

Step 1:

Formulate the test hypotheses:

$$H_0 : \beta_1 = 0, \qquad H_1 : \beta_1 \neq 0$$

Step 2:

Calculate the test statistic:

$$t = \frac{b_1}{s_{b_1}}$$

where

$$s_{b_1} = \frac{s}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}},$$

$$s = \sqrt{\frac{SSE}{n-2}}$$

Step 3:

Conclusion:

$$\text{Reject } H_0 \text{ if } \begin{cases} p - value \leq \alpha & p - value \text{ approach} \\ t \leq -t_{\alpha/2} \text{ or } t \geq t_{\alpha/2} & \text{critical value approach} \end{cases}$$

F Test for Significance in Simple Linear Regression

Step 1:

Formulate the test hypotheses:

$$H_0 : \beta_1 = 0, \qquad H_1 : \beta_1 \neq 0$$

Step 2:
Calculate the test statistic:

$$F = \frac{MSR}{MSE}$$

where

$$MSR = \frac{SSR}{\text{Number of independent variables}}$$

$$MSE = s^2$$

Step 3:
Conclusion:

$$\text{Reject } H_0 \text{ if } \begin{cases} p - value \le \alpha & p - value \text{ approach} \\ F \ge F_\alpha & \text{critical value approach} \end{cases}$$

(F_α is based on an F distribution with 1 degree of freedom in the numerator and $n-2$ degree of freedom in the denominator.)

Confidence Interval for β_1

The form of *confidence interval* for β_1 is as follows:

$$\beta_1 \in \left[b_1 - t_{\alpha/2} s_{b_1},\ b_1 + t_{\alpha/2} s_{b_1} \right]$$

Confidence Interval for the Mean of y

The form of *confidence interval* for the *mean value of y* is as follows:

$$E(y_p) = \left[y_p^* - t_{\alpha/2} \cdot s_{y_p^*},\ y_p^* + t_{\alpha/2} \cdot s_{y_p^*} \right]$$

where

x_p : the particular given value for x

y_p : the value of y corresponding to the given x_p

$E(y_p)$: the expected value of y

$y_p^* = b_0 + b_1 x_p$: the point estimate of $E(y_p)$, when $x = x_p$

$$s_{y_p^*} = s \cdot \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

<u>Confidence Interval for an Individual Value of y</u>

The form of *confidence interval* for *an individual value of y* is as follows:

$$y_p \in \left[y_p^* - t_{\alpha/2} \cdot s_{ind},\ y_p^* + t_{\alpha/2} \cdot s_{ind} \right]$$

where

$$s_{ind} = s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\ .$$

5. One-Way ANOVA

Algorithm (***One-Way ANOVA***)

Step 1 (Formulation of the Hypotheses):

$H_0: \mu_1 = \mu_2 = ... = \mu_k$; $H_1: \mu_i \neq \mu_j$, for at least one $i \neq j$, $i, j = 1,2,...,k$, $k \geq 3$.

Step 2 (Calculation of the Test Statistic)

We define:

x_{ij}: Value of the observation i for treatment j

n_j: Number of observations for treatment j

$\bar{x}_j$: Sample mean for treatment j

s_j^2: Sample variance for treatment j

s_j: Sample standard deviation for treatment j .

Formulas:

$$\bar{x}_j = \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j}$$

$$s_j^2 = \frac{\sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)}{n_j - 1}$$

$$\bar{\bar{x}} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij}}{n_T} \quad ; \quad n_T = n_1 + n_2 + ... + n_k. \quad (\textit{Overall sample mean})$$

(If the size of each sample is n, then $n_T = k \cdot n$ und

$$\bar{\bar{x}} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij}}{kn} = \frac{\frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij}}{n}}{k} = \frac{\sum_{j=1}^{k} \bar{x}_j}{k}.)$$

$$SSTR := \sum_{j=1}^{k} n_j (\bar{x}_j - \bar{\bar{x}})^2 \quad (\textit{Sum of Squares due to TreaTments})$$

$$MSTR := \frac{SSTR}{k-1} \quad (\textit{Mean Square due to TreaTment})$$

$$SSE = \sum_{j=1}^{k} (n_j - 1)s_j^2 \qquad (\underline{S}\textit{um of } \underline{S}\textit{quares due to } \underline{E}\textit{rror})$$

$$MSE := \frac{\sum_{j=1}^{k} (n_j - 1)s_j^2}{n_T - k} = \frac{SSE}{n_T - k} \quad (\underline{M}\textit{ean } \underline{S}\textit{quare due to } \underline{E}\textit{rror})$$

$$\boxed{F = \frac{MSTR}{MSE}}$$

Step 3 (Decision)

$p-value-$approach : Reject H_0 if $p-Value \le \alpha$

Critical-Value-approach: Reject H_0 if $F \ge F_\alpha$

(F_α is based on an F distribution with $k-1$ numerator degree of freedom and $n_T - k$ denominator degree of freedom.)

Confidence Intervals for Each of the Population Means

$$\mu_j \in \left[\bar{x}_j - t_{\alpha/2} \cdot \frac{\sqrt{MSE}}{n_j},\ \bar{x}_j + t_{\alpha/2} \cdot \frac{\sqrt{MSE}}{n_j} \right], \quad j = 1, 2, ..., k,$$

The degrees of freedom for the *t* value are the degrees of freedom associated with the within-treatments of σ^2, namely $n_T - k$.

Multiple Comparison Procedures:

Fisher's Least Significance Difference, LSD

Step 1:
Formulate the hypotheses:

$$H_0 : \mu_i = \mu_j \qquad H_1 : \mu_i \ne \mu_j \quad `$$

Step 2:
Calculate the test statistic:

$$t = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{MSE \cdot \left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}$$

Step 3:
Decision:

$p-value-$approach : Reject H_0 if $p-Value \le \alpha$
Critical-Value-approach: Reject H_0 if $t \le -t_{\alpha/2}$ or $t \ge t_{\alpha/2}$

(The value of $t_{\alpha/2}$ is based on t distribution with $n_T - k$ degrees of freedom.)

Fisher's Procedure Based on the Test Statistic $\bar{x}_i - \bar{x}_j$

Step 1:
Formulate the hypotheses:

$$H_0 : \mu_i = \mu_j \qquad H_1 : \mu_i \neq \mu_j$$

Step 2:
Calculate the test statistic:

$$\bar{x}_i - \bar{x}_j$$

Step 3:
Decision:

Reject H_0 if $|\bar{x}_i - \bar{x}_j| > LSD$, where $LSD = t_{\alpha/2} \cdot \sqrt{MSE \cdot \left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$

Confidence Interval Estimate of the Difference between Two Population Means

$$\left(\mu_i - \mu_j\right) \in \left[(\bar{x}_i - \bar{x}_j) - LSD,\ (\bar{x}_i - \bar{x}_j) + LSD \right],$$

where $LSD = t_{\alpha/2} \cdot \sqrt{MSE \cdot \left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$.

(Revised: 05.12.2019)