Chapter 4

In any statistical problem, we are interested in certain characteristics of the members of the population. These are called variables and are denoted by the capital letters X, Y, Z, etc. The population is too large to study. So, we must draw conclusions by studying only a portion of the population called a sample. The number of objects in the sample is called the sample size and is denoted by n.

Data: We will denote the n observations of the simple random variable X as $x_1, x_2, \ldots, x_n$. We will model each observation from this sample as a random variable, that is X_i is the i^{th} observation. We will assume that the trials are independent or equivalently that there is a simple random sampling plan (SRS) being used. If we repeat the same experiment or we draw samples using an SRS, then the random variables X_1,…, X_n are said to be identically distributed.

Consider a random sample $X_1, X_2, \cdots, X_n$. A function of the random sample is called a statistic. Often we use a statistic

$$\widehat{\Theta} = h(X_1, X_2, \ldots, X_n)$$

to estimate a population parameter θ. We say that $\widehat{\Theta}$ is a point estimator of θ.

Now we must note that $\widehat{\Theta}$ is a random variable. Its observed value of the random variable $\widehat{\Theta}$ which is $\hat{\theta} = h(x_1, x_2, \ldots, x_n)$ is called a point estimate of θ.

Point Estimates are a single value given as an estimate of a parameter of a population.

- The sample standard deviation (s) is a **point estimate** of the population standard deviation, σ
- The sample mean ($\bar{x}$) is a **point estimate** of the population mean, μ
- The sample proportion $(\widehat{p})$ is a **point estimate** of the population proportion, p

Statistics

Suppose we have a quantitative variable X whose observed values are $x_1, \ldots, x_n$. A statistic is a function of these observed value. Descriptive statistics give a useful summary of the data. The most commonly used descriptive statistics are:

a) the mean, median and mode
b) the standard deviation and variance
c) minimum and maximum values, and range
d) the quartiles Q1, Q3, and the interquartile range

We will describe the center of the distribution with either the mean, median or mode:

The sample **mean** $\bar{x}$ is the average of the n observations. If we put a weight (mass) of 1/n on each observation x_i, then $\bar{x}$ is the center of the mass. It is calculated as:

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

When the variable x takes on only two possibilities the statistic we use is the
The sample proportion $\hat{P}$ is defined as follows: $\hat{P} = X/n$, where X is the number of selected items that satisfy the attribute of interest among n items.

The sample **median** of n observations is the sample value which divides the sample into two approximately equal-sized data subsets. To obtain the median, arrange the sample values $x_1, x_2, \ldots, x_n$ in ascending order $y_1 \le y_2 \le \ldots \le y_n$. The latter are called order statistics.

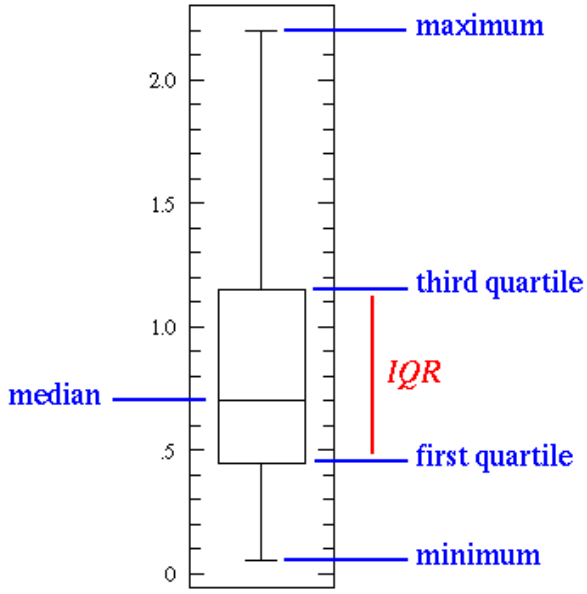
$$\text{Median} = \begin{cases} x_{(n+1)/2} & \text{If n is odd} \\ 1/2(x_{(n/2)} + x_{n/2+1)} & \text{If n is even} \end{cases}$$

The sample **mode** of n observations is the most common value.

We will describe the spread of the distribution with either the standard deviation, the variance, the range, and/or the interquartile range.

The sample **minimum** is the smallest ordered value.
The sample **maximum** is the largest ordered value.
The **range** is the maximum – minimum.
The **quartiles** split the distribution into 4 equal parts, Q1, Q2 (the median), Q3 and Q4

The **interquartile range** is the difference between Q3-Q1.

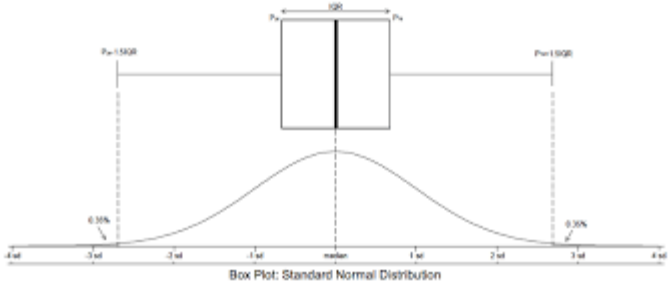


A **boxplot** is a graphical display of the 5-number summary, which is constructed as follows:

- Extend the box from the first quartile to the third quartile. (The box displays the interquartile range.)
- Within the box, display a line at the median.
- Imaginary fences are placed at a distance of 1.5 IQR above the third quartile and below the first quartile.
- Whiskers extend from the ends of the box to the smallest and the largest value in the sample, within the imaginary fences.
- Values outside the fences are called outliers. Each outlier is displayed as a point.

Sometimes a different symbol is used for extreme outliers that are at least 3 IQR above q3 or below q1.

Example: Normal Data



The sample **standard deviation** is denoted by s and is the most commonly used measure for dispersion. Its square is called the sample **variance** and is calculated by:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{1}{n(n-1)}\left[n\sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2\right]$$

The standard deviation of a point estimator is called the **standard error** of the estimate. So the standard error of the mean and of the sample proportion are respectively:

$\sigma_X = \sigma / \sqrt{n}$ and $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$.

Note: The standard error usually involves an unknown parameter, so in practice when giving data we replace this unknown parameter by its point estimate. So, the estimated standard error of the mean and of the sample proportion become in this case:

$\hat{\sigma}_X = s / \sqrt{n}$ and $\hat{\sigma}_{\hat{p}} = \sqrt{\hat{p}(1-\hat{p})/n}$.

Unbiased Estimators:

Let $\hat{\theta}$ be an estimator for the parameter θ. We say that $\hat{\theta}$ is an unbiased estimator for θ if

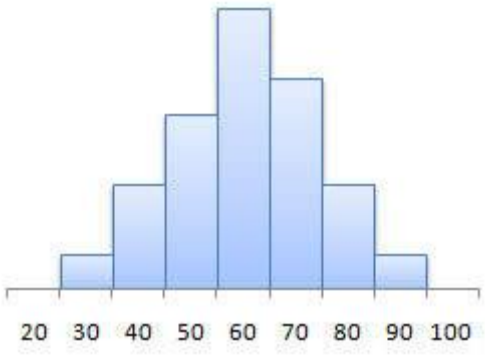
$$E(\hat{\theta}) = \theta$$

We say that an estimator is accurate if it has a low variability.

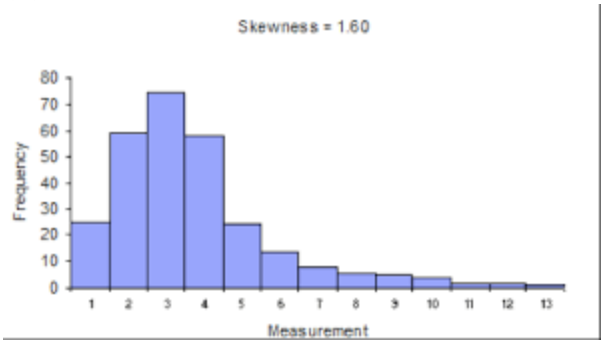
Histogram:

A histogram is used to describe the shape of the distribution of a numerical variable. The scale of the variable determines the placement of the bars. Usually, the vertical bars have the same width. Unlike the bar charts, the bars are not separated by some arbitrary space. Some examples of histograms that are respectively approximately symmetric, skewed to the right, skewed to the left. The asymmetry is in the direction of the atypical values.

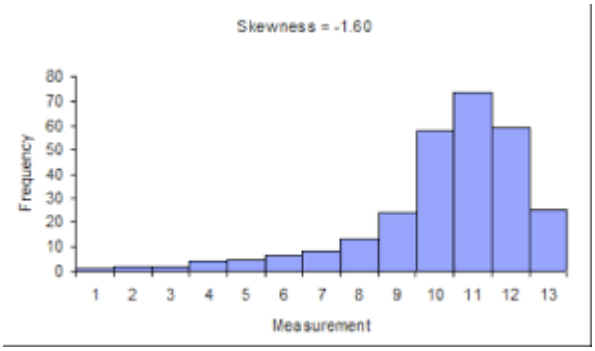
(a) approximately symmetric



(b) skewed to the right



(c) skewed to the left



Central Limit Theorem (CLT):

Definition: Let $X_1, \ldots, X_n$ be a random sample and $h(X_1, \ldots, X_n)$ be a statistic of this sample. The statistic is a random variable. We call its distribution a sampling distribution. Let $X_1, \ldots, X_n$ be a random sample from a population $N(\mu, \sigma^2)$. The sample mean X has the following sampling distribution $N(\mu, \sigma^2/n)$. In other words, $X \sim N$ $(\mu, \sigma^2/n)$.

The following theorem will allow us to approximate the sample distribution of the sample mean.

CLT: Let $X_1, \ldots, X_n$ be a random sample from a population with mean μ and variance σ^2. The random variable

$$Z = \frac{(X-\mu)}{\sigma/\sqrt{n}}$$

has approximately a standard normal distribution for sufficiently large n ($n \geq 30$).

Example when to use CLT:
The examination scores in an university course have mean 56 and sd 11. In a class of 49 students, what is the probability that the average mark is less than 50? What is the probability that the average mark lies between 50 and 60?

Independent Samples:
Let $X_1, \ldots, X_{n1}$ a random sample from a population with mean μ1 and variance σ^2_1 and let $Y_1, \ldots, Y_{n2}$ be a random sample from a population with mean μ2 and variance σ^2_2. We will assume that the two random samples are independent, i.e. $X_1, \ldots, X_{n1}$, $Y_1, \ldots, Y_{n2}$ are independent random variables.
1. If the two population are normal, i.e. $N(\mu_1, \sigma^2_1)$ and $N(\mu_2, \sigma^2_2)$, respectively, then, by the reproductive property of the normal distribution,

$$X \sim N(\mu_1, \sigma_1^2) \text{ and } Y \sim N(\mu_2, \sigma_2^2)$$

then

$$X - Y \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$$

2. If we weaken the assumption of normality, be we assume that both sample sizes are large, i.e. $n1 \geq 30$ and $n2 \geq 30$, then, using the CLT and the reproductive property of the normal distribution, we can show that has approximately a standard normal distribution

$$Z_0 = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

Example of this type of question:

The effective life of a component in a jet-turbine aircraft is normally distributed with mean 5000 hours and standard deviation 40 hours. Suppose that when using an improved manufacturing process, the effective is normally distributed with mean 5050 and standard deviation 30. Let X_1 be the average life of n1 = 16 components produced under the old manufacturing process and let X_2 be the average life of n2 = 25 components produced with the new manufacturing process. Find the following probability:
P $(\overline{X}_2 - \overline{X}_1) > 25$

Interval Estimates are an interval within which the value of a parameter of a population has a stated probability of occurring.

- A Confidence Interval (CI) is an **interval estimate**, with a specified coverage or probability that the population estimate falls under repeated sampling within the interval.
    - For example, if **repeated samples** were taken and the 95% **confidence interval** was computed for each **sample**, 95% of the **intervals** would contain the population mean.
    - This method produces an interval of possible values that includes the unknown parameter with a high probability.
- A Prediction Interval (PI) is an **interval estimate** associated with a random variable yet to be observed, with a specified probability of the random variable lying within the **interval**.

The statistic $\overline{X}$ is called a point estimator because it gives a single numeric value. This value is not exactly the value of the parameter μ of interest. We could use the standard error of the estimate as a measure of the error. It measures approximately the error that we will observe on average. The method of estimation by confidence intervals gives an interval [Lower, Upper] of values which has a high probability of containing the unknown parameter μ.

$$P(\text{Lower} \leq \boldsymbol{\mu} \leq \text{Upper}) = 1\text{-}\boldsymbol{\alpha}$$

Consider the case with a normal population or a large sample size (n ≥ 30); and population variance σ^2 is known. Under these conditions:
We use a $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ has a standard normal distribution

That is N(0, 1) distribution. It is approximative if n is large but the population is not normal. We are first interested in finding a point z such that $P(-z \le Z \le z) = 1 - \alpha$. This means that $P(Z > z) = \alpha/2$ and hence $P(Z \le z) = 1 - \alpha/2$. By the first definition, we call this z as $z_{\alpha/2}$.

The 95% confidence interval for the mean μ: From Table A.3, we know that

$$P(-1.96 \le Z \le 1.96) = 0.95.$$

Using this and the formula above we derive that the 95% CI for the mean is then the interval $[\overline{X}-(1.96)(\sigma/\sqrt{n}), \overline{X}+(1.96)(\sigma/\sqrt{n})]$.

Likewise, all other CI of means and proportions follow the same format. That is the mean (or proportion) plus or minus the standard error of the estimate. As noted in class the standard error of the estimate is the denominator on your Z or t statistic.

- Statistic used for a mean μ when σ is unknown:
$$T_0 = \frac{\bar{X}-\mu}{S/\sqrt{n}} \text{ has a t distribution with } n - 1 \text{ d. f.}$$
- Statistic used for a difference when observations are paired and σ is unknown:
$$T_0 = \frac{\bar{d}-d_0}{S_d/\sqrt{n}} \text{ has a t distribution with } n - 1 \text{ d. f.}$$
- Statistic used for a proportion p when n is large:
$$Z = \frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}} \text{ has a standard normal distribution,}$$
where $\hat{p}$ is the sample proportion.
- Statistic used for a proportion p_0:
$$Z_0 = \frac{X - np_0}{\sqrt{np_0(1-p_0)}} = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0/n}} \text{ has a standard normal distribution}$$
- Statistic used for two independent populations with variances are known:
$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{(\sigma_1^2/n_1 + \sigma_2^2/n_2)}} \text{ has a standard normal distribution}$$
- Statistic used for the means μ_1 and μ_2 of two small independent normal populations with equal variances:
$$T_0 = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{(1/n_1 + 1/n_2)}} \text{ has a } t \text{ distribution with } n_1 + n_2 - 2 \text{ d. f.}$$
where

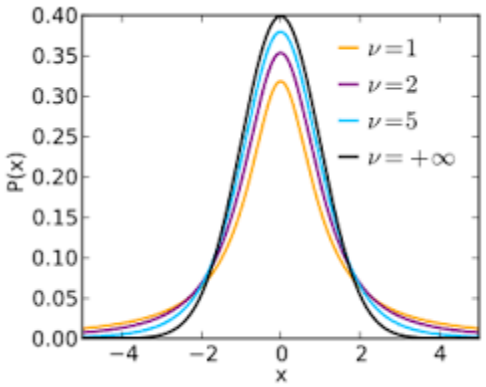
$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

- Statistic used for μ_1 and μ_2 of two independent large arbitrary populations:

$$Z_0 = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \text{ has a standard normal distribution}$$

Example of types of questions:
A sample of 9 observations from a normal population with known standard deviation σ = 5 yields a sample mean of 19.93. Provide 95% and 99% confidence intervals for the unknown population mean μ based on this sample.

Note – the degrees of freedom on the t statistic are associated with the sample size, as the sample size increases the t distribution approaches Z, in fact this occurs at n=30.



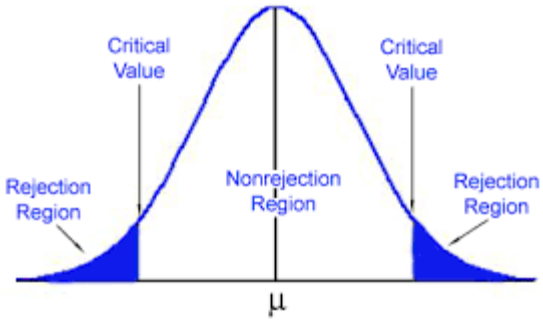
Hypothesis testing:
Hypothesis testing is a procedure that leads us to decide if experimental data supports a hypothesis concerning population(s) parameter(s). We will consider hypotheses concerning a population mean μ or a population proportion p.

Stating the Hypotheses: Often the researcher would like to verify a change in the unknown parameter under new experimental conditions. The null hypothesis, denoted by H_0, will always be a simple statement concerning the unknown parameter. The null hypothesis is often a statement of no effect. On the other hand, the alternative hypothesis, denoted by H_1, will be a composite statement concerning the unknown parameter. It is often the research hypothesis, i.e. the hypothesis that we would like to support with the data. Suppose that θ is the unknown population parameter and that θ0

is a real number. For example, θ could be the population mean. We will consider three types of alternative.
1. Suppose that a priori, our research hypothesis is that $\theta > \theta_0$. Then, we will consider a right sided alternative. That is, we will test $H_0 : \theta = \theta_0$, against $H_1 : \theta > \theta_0$.
2. Suppose that a priori, our research hypothesis is that $\theta < \theta_0$. Then, we will consider a left sided alternative. That is, we will test $H_0 : \theta = \theta_0$, against $H_1 : \theta < \theta_0$.
3. A two sided alternative would give the following test $H_0 : \theta = \theta_0$, against $H_1 : \theta = \theta_0$.

We will follow 5 steps:

1) State the Null and Alternative Hypothesis
2) Choose a fixed α
3) Choose and appropriate test statistic
4) Calculate our value and reject if it is outside of the critical region, accept if it is inside
5) State your conclusion



For example, with a large sample, suppose that the test statistic is Z_0 which follows a standard normal distribution under H_0. Let z_0 be the observed value of Z_0.

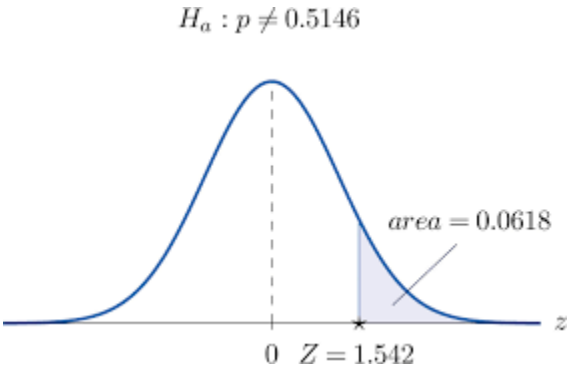
[Right-Sided Alternative] $H_1 : \mu > \mu_0$. The critical region is $z_0 > z_\alpha$.
[Left-Sided Alternative] $H_1 : \mu < \mu_0$. The critical region is $z_0 < -z_\alpha$.
[Two-Sided Alternative] $H_1 : \mu$ 6= μ_0. The critical region is $z_0 < -z_{\alpha/2}$ or $z_0 > z_{\alpha/2}$.

Example of hypothesis test type question:
Suppose buildings specifications in a certain city require that the average breaking strength of residential sewer pipe be more than 2, 400 pounds per foot of length. Each manufacturer who wants to sell pipe in this city must demonstrate that its products meets the specification. Let μ denote the mean breaking strength of residential sewer pipe. Note

that we are interested in making an inference about the mean μ of the population. We want to decide whether the mean breaking of the pipe exceeds 2, 400 pounds per linear foot. The null hypothesis is $H_0 : \mu = 2, 400$ and the alternative hypothesis is $H_1 : \mu > 2, 400$. Thus, we want to test $H_0 : \mu = 2, 400$ against $H_1 : \mu > 2, 400$.

When asked to use a P-value we will use 4 steps:

1) State the Null and Alternative Hypothesis
2) Choose an appropriate statistic
3) Calculate our value calculate the p-value for the hypothesis (remember if 2-sided you double)
4) State your conclusion



Example of p-value when the test statistic is Z which follows a N(0, 1) when H_0 is true. Then the p-value of the test is
Case I : $H_0 : \mu = \mu_0$, $H_1 : \mu > \mu_0$, $p - value = P\ (Z \geq z_0)$
Case II : $H_0 : \mu = \mu_0$, $H1 : \mu < \mu_0$, $p - value = P\ (Z \leq z_0)$
Case III : $H_0 : \mu = \mu_0$, $H1 : \mu = \mu_0$, $p—value = 2P\ (Z \geq |z_0|)$

Notes:

- A test statistic is a statistic that is used to test hypotheses.
- The critical region of the test statistic is a set of possible values of the test statistic such that if the observed of the test statistic falls in the critical region we will reject H_0 and accept H_1.
- If we reject H0 when H0 is true, we say that we have committed an error of type I and $\alpha = P(\text{type I error}) = P(\text{ reject } H_0 \text{ when } H_0 \text{ is true})$.

- If the observed value of the test statistic does not fall in the critical region, then we fail to reject H_0. If we fail to reject H_0 when H_0 is false, then we say that we have committed an error of type II and $\beta(\theta_1) = P(\text{type II error}) = P(\text{ fail to reject } H_0 \text{ when } \theta = \theta_1 \in H_1)$. The value of β will vary for a composite alternative.

We also have the case where we can test more than two proportions with a chi-squared test. You follow the same steps as listed above but use the following test statistic.

- Statistic used with more than two proportions:
$$X^2 = \sum \frac{(O_i - E_i)}{E_i} \text{ chi square with } v = \text{k-1 degrees of freedom}$$

Simple Linear Regression:

Simple linear regression is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables. First one variable, denoted x, is regarded as the **predictor**, **explanatory**, or **independent** variable. Next the other variable, denoted y, is regarded as the **response**, **outcome**, or our **dependent** variable.

As this is a linear model, we can trivially see that the model underlying a linear regression analysis is that the explanatory and outcome variables are linearly related such that the population mean of the outcome for any x value is $\beta_0 + \beta_1 x +$ error. Where the error model underlying a linear regression, analysis includes the assumptions of fixed-x, Normality, equal spread, and independent errors.

- Formulas for regression analysis:
  - Parameter estimation:

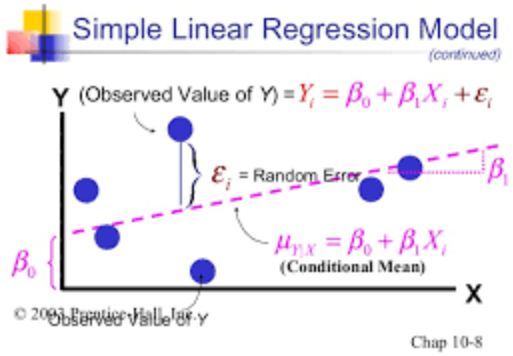
$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)$$

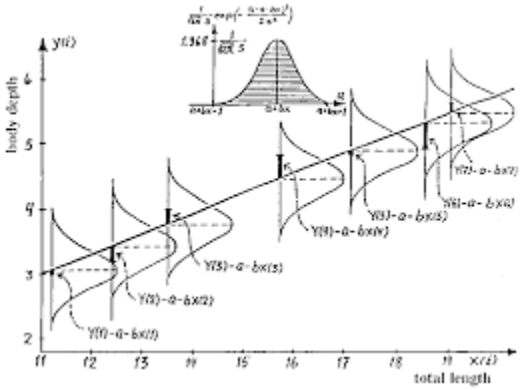
$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$$

$$S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2$$

The usual regression null hypothesis is $H_0 : \beta_1 = 0$. Sometimes it is also meaningful to test $H_0 : \beta_0 = 0$ or $H_0 : \beta_1 = 1$.



Source: https://www.slideshare.net/vermaumeshverma/linear-regression-38653351



Source: https://mobiledevmemo.com/when-why-and-how-you-should-use-linear-regression/

We note that the interpretation of the intercept is generally only interpreted if there are values of x near 0 and that it makes scientific sense. The meaning of b_0 is the estimate of the mean outcome when $x = 0$ and should always be stated in terms of the actual variables of the study.

The interpretation of b_1 is the change in the average outcome when the explanatory variable increases by one unit. This should always be stated in terms of the actual variables of the study.