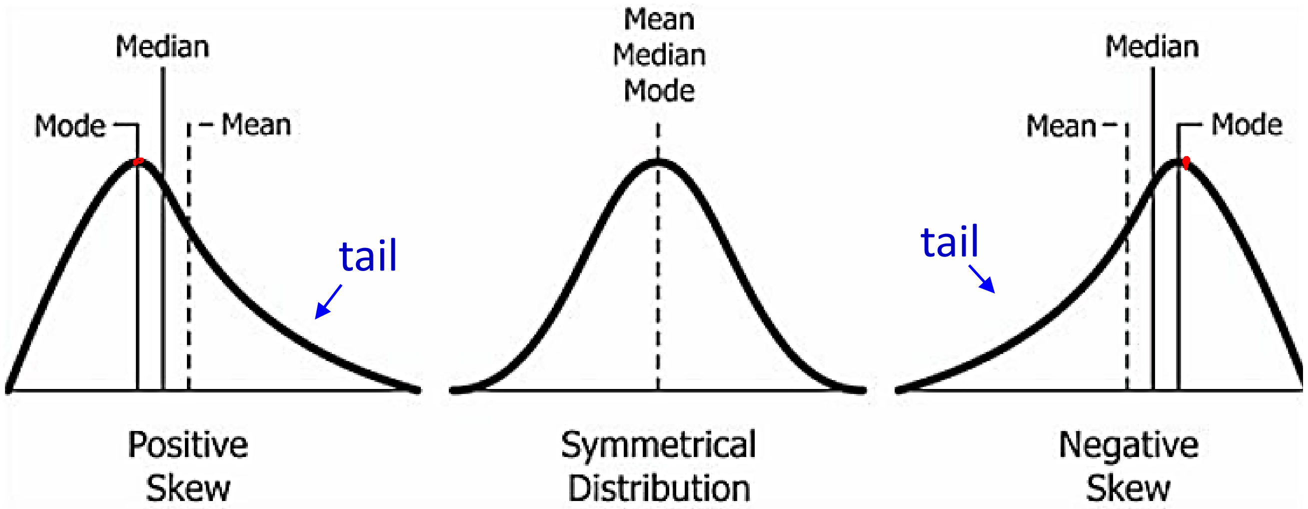
Other Statistical Parameters

Name	Definition	Symbol
mean	$E[X]$	μ
variance	$E[(X-\mu)^2]$	σ^2
standard deviation	$\sqrt{\sigma^2}$	σ
skewness	$E[(X-\mu)^3]/\sigma^3$	γ_1
kurtosis	$E[(X-\mu)^4]/\sigma^4-3$	γ_2

Skewness

skewness	$E[(X-\mu)^3]/\sigma^3$	γ_1

- Skewness can tell us about symmetry: . It measures the lack of symmetry in data distribution.
- It is the degree of distortion from the symmetrical bell curve or the normal distribution
- It differentiates extreme values in one versus the other tail
- A symmetrical distribution will have a skewness of 0

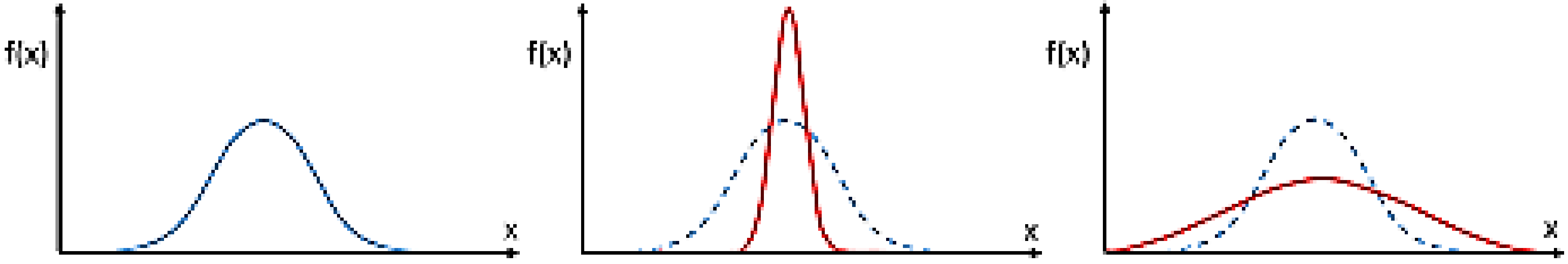


- If the skewness is between -0.5 and 0.5, the data are fairly symmetrical.
- If the skewness is between -1 and -0.5 (negatively skewed) or between 0.5 and 1 (positively skewed), the data are moderately skewed.
- If the skewness is less than -1 (negatively skewed) or greater than 1 (positively skewed), the data are highly skewed.

Kurtosis

kurtosis	$E[(X-\mu)^4]/\sigma^4 - 3$	γ_2

- Kurtosis is all about the tails of the distribution — the peakedness or flatness.
- It is used to describe the extreme values in one versus the other tail.
- It is actually the measure of outliers present in the distribution.
- High kurtosis data has heavy tails or outliers.
- Low kurtosis data has light tails or lack of outliers.



Zero kurtosis
Gaussian distribution

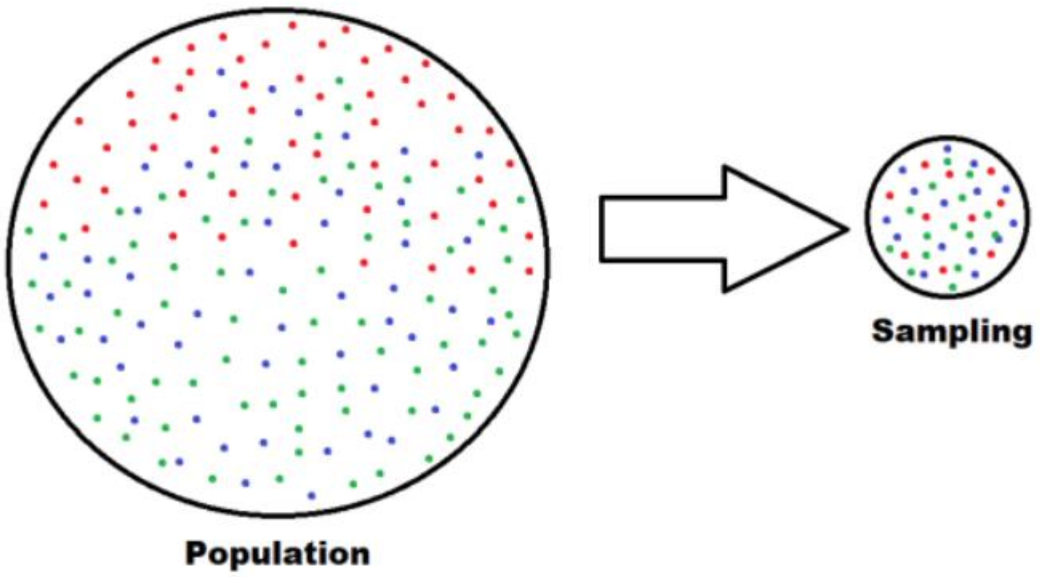
Positive kurtosis

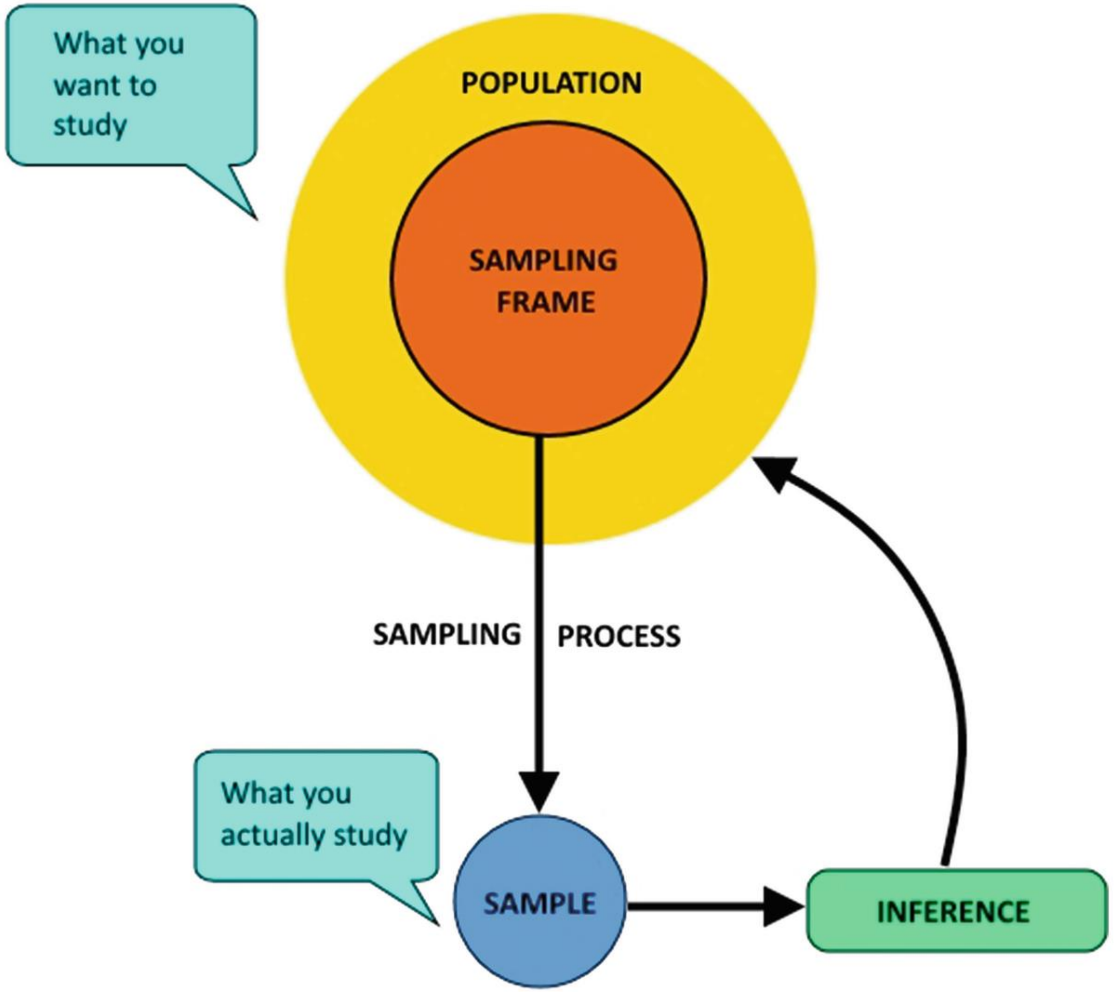
Negative kurtosis

Data Sampling

What is Sampling?

https://images.app.goo.gl/cvN1oMakejPsdhva7

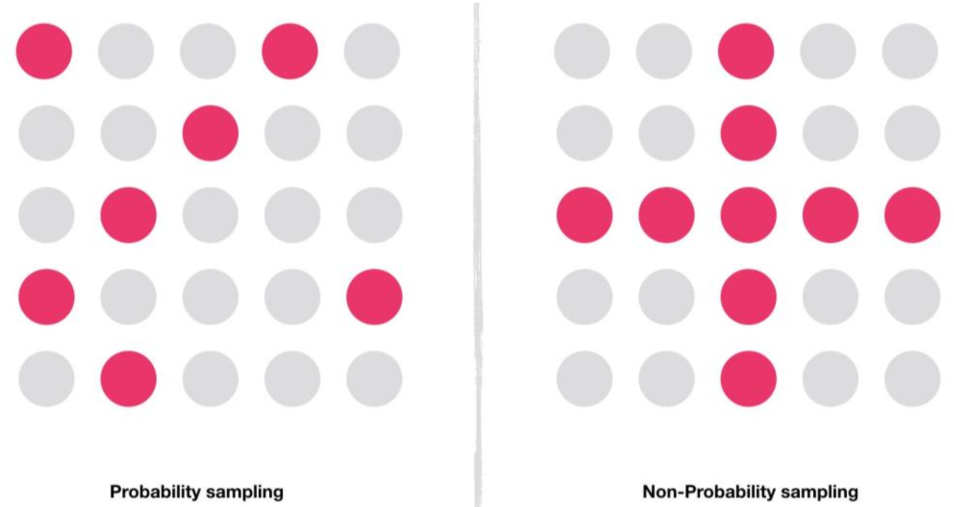




Sampling Methods

- **Probability Sampling**
  - When *each entity* of the population has *a finite, non-zero probability* of being into the sample
  - Sampling procedure involves random sampling and without bias
- **Non-probability Sampling**
  - Some units of the population have zero chance of selection
  - OR probability of selection cannot be determined accurately

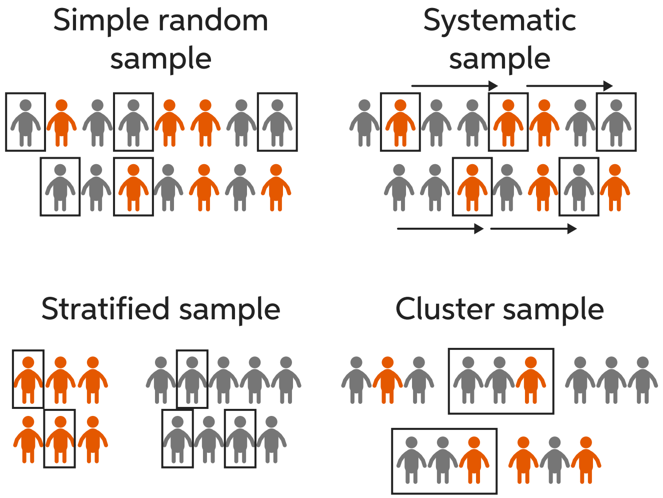
Probability sampling	Non-probability sampling
The samples are randomly selected.	Samples are selected on the basis of the researcher's subjective judgment.
Everyone in the population has an equal chance of getting selected.	Not everyone has an equal chance to participate.
Researchers use this technique when they want to keep a tab on sampling bias.	Sampling bias is not a concern for the researcher.
Useful in an environment having a diverse population.	Useful in an environment that shares similar traits.
Used when the researcher wants to create accurate samples.	This method does not help in representing the population accurately.
Finding the correct audience is not simple.	Finding an audience is very simple.



https://www.questionpro.com/blog/probability-sampling/

Probability Sampling

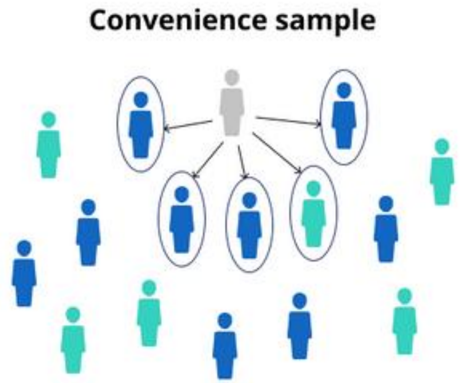
- **Simple Random Sampling**
  - Each subject/unit selected at random, independent from each other
  - Typically done when the population is large
- **Systematic Sampling**
  - Arrange the population in some order, and pick a unit at regular intervals from the list
  - When population is logically homogenous
  - E.g. You ask every 10th customer entering a shop about his purchase habits
- **Stratified Sampling**
  - Population divided into groups/stratas based on some characteristics
  - Then population is sampled randomly within each strata
  - E.g. If 38% of the population is college-educated, then 38% of the sample is randomly selected from the college-educated subset of the population
- **Cluster Sampling**
  - Random sample is drawn from a cluster of data, rather than individual samples
  - E.g. An NGO wants to create a sample of girls across five neighboring towns to provide education. Using single-stage sampling, the NGO randomly selects towns (clusters) to form a sample and extend help to the girls deprived of education in those towns.

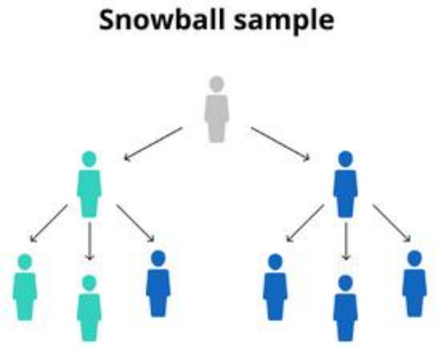


www.chegg.com

Non-Probability Sampling

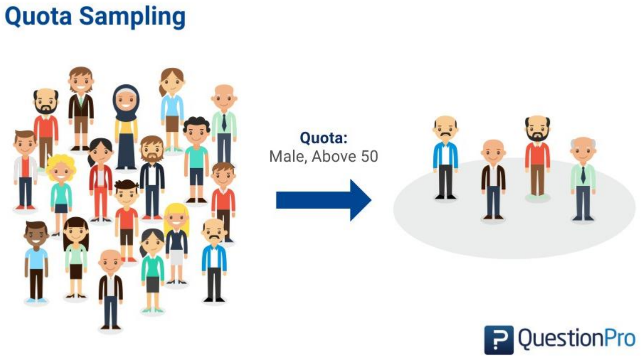
- **Convenience Sampling**
  - Each subject/unit is selected on the basis of convenience, availability, reach, etc.
  - Typically during preliminary research





- **Snowball Sampling**
  - One unit refers you to the next unit
  - Costs of sampling are lower

- **Quota Sampling**
  - Population divided into mutually exclusive subgroups and non-random set of observations chosen from each subgroup



Distribution of Sample Means

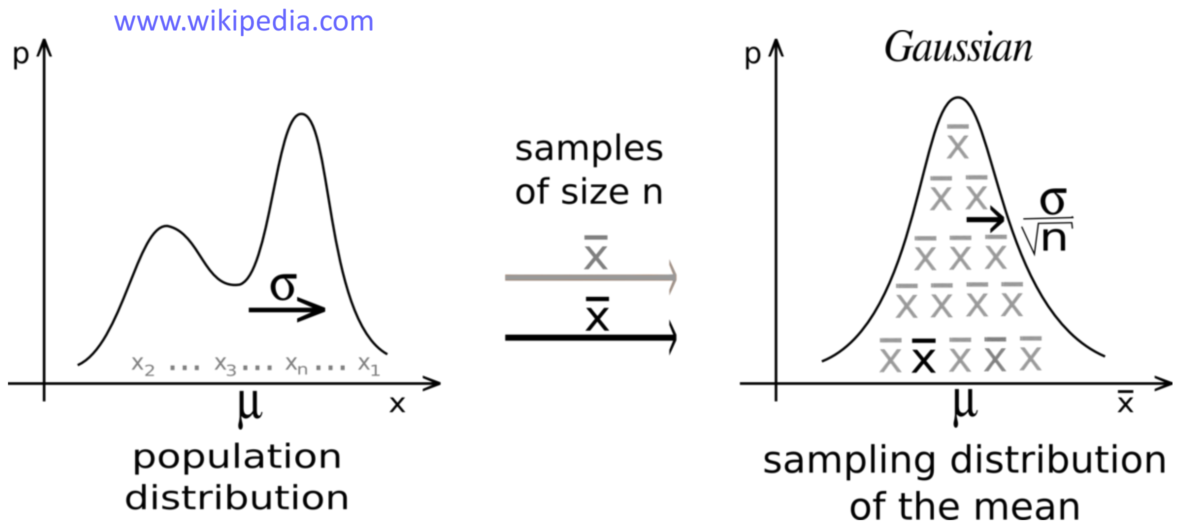
- We say that the sample mean ($\bar{y}$) gives us *an estimate* of the population mean μ
- But, since the sample is only a small subset of the entire population, *$\bar{y}$ is an uncertain estimate of μ*

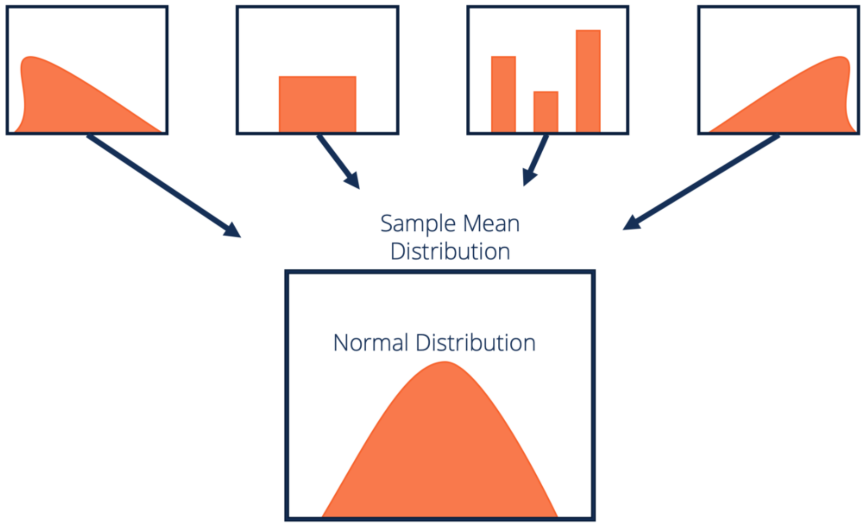
- What if, we sample the population *several times*, each time calculating the sample mean $\bar{y}$
- Let's say we do it 'k' times, we get sample means as $\bar{y}_1, \bar{y}_2, \bar{y}_3, \bar{y}_4, \ldots.., \bar{y}_k$
- How would these sample means behave?
  - How close are they from *μ?*
  - What's the *mean of sample means*?
  - What's the *standard deviation of sample means*?
  - More importantly, *what's their frequency distribution*?

Central Limit Theorem

"The *distribution of sample means* $(\bar{y}_1, \bar{y}_2, \bar{y}_3, \bar{y}_4, \ldots.., \bar{y}_k)$ follows *a normal distribution*, even when the original variable y is NOT normally distributed."

www.moriah.com





- What is the mean of distribution of sample means?

- What is the variance of distribution of sample means?

Central Limit Theorem

DIY: Read

"In non-mathematical language, the "CLT" says that whatever the PDF of a variable is, if we randomly sample a "large" number (say k) of independent values from that random variable, the sum or mean of those k values, if collected repeatedly, will have a Normal distribution.

It takes some extra thought to understand what is going on here. The process I am describing here takes a sample of (independent) outcomes, e.g., the weights of all of the rats chosen for an experiment, and calculates the mean weight (or sum of weights). Then we consider the less practical process of repeating the whole experiment many, many times (taking a new sample of rats each time). If we would do this, the CLT says that a histogram of all of these mean weights across all of these experiments would show a Gaussian shape, even if the histogram of the individual weights of any one experiment were not following a Gaussian distribution.

By the way, the distribution of the means across many experiments is usually called the sampling distribution of the mean."

- Seltman, Howard J. "Experimental design and analysis." (2012)

Distribution of Sample Means

- The Central Limit Theorem is the explanation for why many real-world random variables tend to have a Gaussian distribution. It is also the justification for assuming that if we could repeat an experiment many times, any sample mean that we calculate once per experiment would follow a Gaussian distribution over the many experiments.
- Since the *distribution of the sample means* with mean (μ) and variance (σ_y^2/n) follows a normal distribution, then the relationship between the distribution of sample means and the z-distribution is given by:

$$z = \frac{\bar{y} - \mu}{\frac{\sigma_y}{\sqrt{n}}}$$

- What does it tell about value of a random sample mean?
- But, we often don't know the population standard deviation (σ_y) or variance (!)
- **Can we estimate them?**

Estimators of Population

Ref: Montgomery, D. C., "Design and Analysis of Experiments," 8th Ed.,

We may easily show that $\bar{y}$ and S^2 are unbiased estimators of μ and σ^2, respectively.

First consider $\bar{y}$. Using the properties of expectation, we have

$$
\begin{aligned}
E(\bar{y}) &= E\left(\frac{\sum_{i=1}^{n} y_i}{n}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} E(y_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mu \\
&= \mu
\end{aligned}
$$

$$
\begin{aligned}
E(S^2) &= E\left[\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}\right] \\
&= \frac{1}{n-1} E\left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right] \\
&= \frac{1}{n-1} E(SS)
\end{aligned}
$$

$$
\begin{aligned}
E(SS) &= E\left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right] \\
&= E\left[\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right] \\
&= \sum_{i=1}^{n}(\mu^2 + \sigma^2) - n(\mu^2 + \sigma^2/n) \\
&= (n-1)\sigma^2
\end{aligned}
$$

$$
E(S^2) = \frac{1}{n-1} E(SS) = \sigma^2
$$

S^2 is an unbiased estimator of σ^2.

where $SS = \Sigma_{i=1}^{n}(y_i - \bar{y})^2$ is the **corrected sum of squares** of the observations y_i.