Fundamentals-Statistics: Population vs Sample

Intuition: A Chef cooking a pot of soup. If he is to check whether the soup is salty after adding some salts.The spoon of soup that he taste is the a SAMPLE from the entire pot of soup(POPULATION). Based on the taste from the spoon, he judges the entire pot of soup.

If the spoonful is salty, you assume the whole pot is salty. But what if you didn't stir the pot? Your spoonful might be bland while the bottom is pure salt. This problem is called Sampling Bias.

Population: This is the ground truth because it represents the entire set of all elements for a specific experiment/study. This is usually represented with N.

Samples: This is just a part or subsets of the main population. This is represented with n.

There manys of selecting your samples from the population.

1. Random sampling: As the name implies just picking randomly. The issue with the this is that introduces biasness. Think of it like this. yoy want to carry out a study that involves the Nigeria. you just enter the street and start picking any body. Now what has been sampled out does not represent the Nigerian population well, we have different tribes, different cities e.t.c.

2. Stratified sampling: This involves splitting the population into groups based on shared properties and then sampling from this strata. This groups are called strata. This ensures all subgroup are represent. Scenerio-seperating the nation into their states, and then picking from each states. With this you are sure that you have represented Nigeria in your sampling.

3. Cluster sampling: Here we divide the population into natural groups (clusters). Randomly select a few entire clusters and test every unit in them.

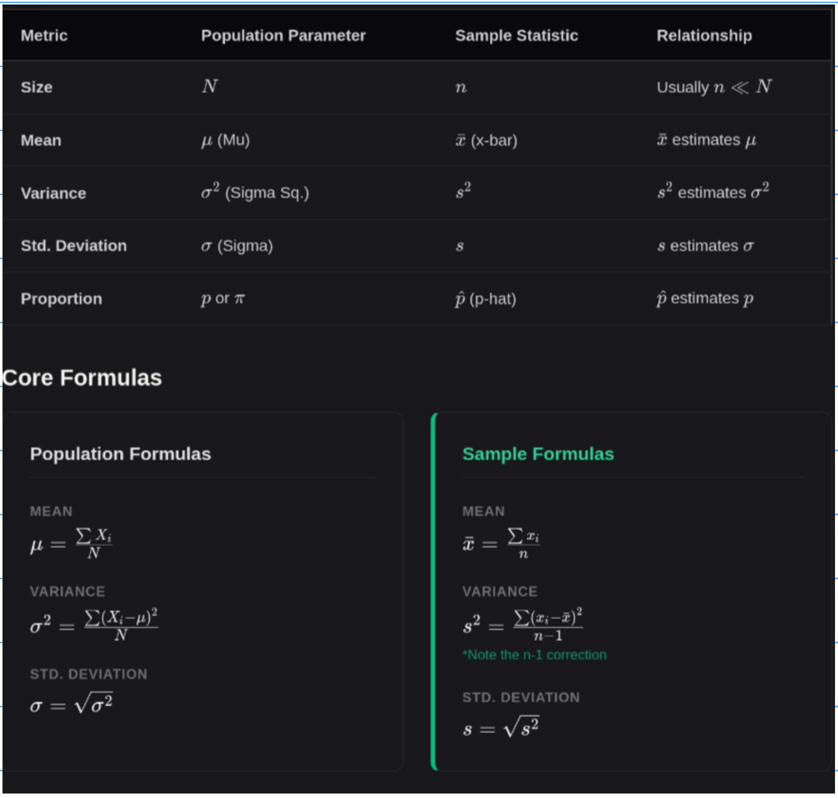
e.t.c

Bias & Representative -> A sample is Representative if its distribution of attributes matches the population. If not, it is Biased.

Note: Your data only contains records of things that made it into your dataset.

Ask yourself: What's missing? When analyzing customer behavior, you're seeing active customers, not churned ones. When studying successful startups, you're missing the 90% that failed quietly.

Important Notation and Formulas



Metric	Population Parameter	Sample Statistic	Relationship
Size	N	n	Usually $n \ll N$
Mean	μ (Mu)	$\bar{x}$ (x-bar)	$\bar{x}$ estimates μ
Variance	σ^2 (Sigma Sq.)	s^2	s^2 estimates σ^2
Std. Deviation	σ (Sigma)	s	s estimates σ
Proportion	p or π	$\hat{p}$ (p-hat)	$\hat{p}$ estimates p

Core Formulas

Population Formulas

MEAN

$$\mu = \frac{\sum X_i}{N}$$

VARIANCE

$$\sigma^2 = \frac{\sum (X_i - \mu)^2}{N}$$

STD. DEVIATION

$$\sigma = \sqrt{\sigma^2}$$

Sample Formulas

MEAN

$$\bar{x} = \frac{\sum x_i}{n}$$

VARIANCE

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

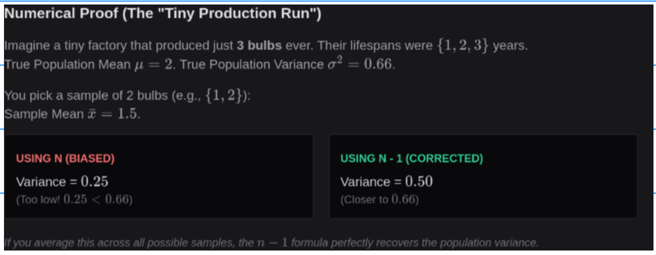
*Note the n-1 correction

STD. DEVIATION

$$s = \sqrt{s^2}$$

The n-1 in the variance is called the Bessel correction. This is to correct biasness, the way i understand it. It when calculation the deviation for the samples, it ....

The Numerical proof makes it clear



Numerical Proof (The "Tiny Production Run")

Imagine a tiny factory that produced just **3 bulbs** ever. Their lifespans were $\{1, 2, 3\}$ years.
True Population Mean $\mu = 2$. True Population Variance $\sigma^2 = 0.66$.

You pick a sample of 2 bulbs (e.g., $\{1, 2\}$):
Sample Mean $\bar{x} = 1.5$.

USING N (BIASED)

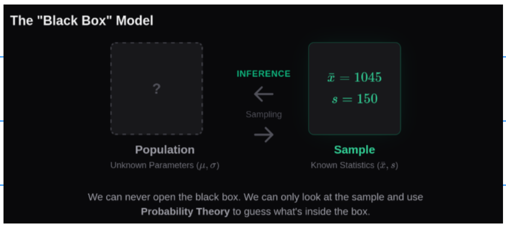
Variance = 0.25
(Too low! $0.25 < 0.66$)

USING N - 1 (CORRECTED)

Variance = 0.50
(Closer to 0.66)

If you average this across all possible samples, the $n-1$ formula perfectly recovers the population variance.

The main goal of sampling is not just to get the descriptive stats of the samples, but to also INFER. Predict the data we don't have, that of the main population.



The "Black Box" Model

?
INFERENCE
Sampling
$\bar{x} = 1045$
$s = 150$
Population
Unknown Parameters (μ, σ)
Sample
Known Statistics ($\bar{x}, s$)

We can never open the black box. We can only look at the sample and use **Probability Theory** to guess what's inside the box.

They are two main pillars of inference. Estimation and Hypothesis Testing.

Estimation: I don't know but i can give you a range; Hypothesis Testing: I have a theory does the does the data support or reject it.

Machine Learning Application:

During the process of training a model. The data is split into train and test data, the test data is assumed to be the Population(unseen data), where the model is tested upon. If the models performs well on the test data then we infer it will perform well in real production(i.e when deployed).