Chapter 2. Data and Sampling Distributions

Random Sampling

Sample: A portion of the overall population, taken when it is either too difficult to obtain or process statistics the entire dataset. (Subset from larger set)

Population: The totality of the dataset available.

N (n) : The size of the population (sample).

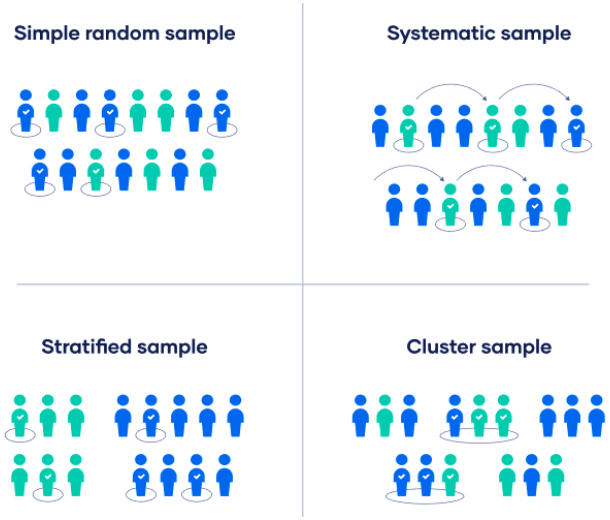
Types Of Sampling:

1.Probability sampling - Random selection techniques are used to select the sample.

- Simple random sampling: Creating a sample by randomly picking elements of the population is known as random sampling, sample that results from it are called as simple random sampling method.
- Stratified sampling: Dividing the population into strata and randomly sampling from each strata(homogeneous subgroup).

Eg. Food categories- fruits and veggie, meat and protein, dairy and grains.

- Systematic Sampling: Population is given randomly generated numbers from which the samples are chosen at regular intervals.(5, 15, 25, and so on until the sample is obtained).
- Cluster Sampling: Population is divided into subgroups instead of selecting a sample from each subgroup, you randomly select an entire subgroup.

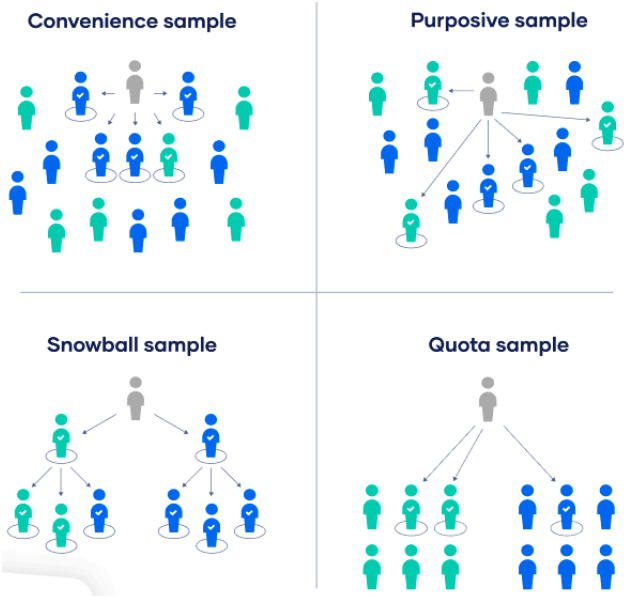


2. Non-probability sampling - Non-random selection techniques based on certain criteria are used to select the sample.This sampling method is easier and cheaper but also has high risks of sampling bias.

- Convenience Sampling: The researcher simply selects the individuals which are most easily accessible to them.The only criteria involved is that people are available and willing to participate.

- Voluntary (Quota) Response Sampling: The only criterion is people are willing to participate, instead of the researcher choosing the participants, the participants volunteer themselves.

- Purposive Sampling: The researcher uses their expertise and judgment to select a sample that they think is the best fit.



Eg.The researcher wants to know about the experiences of disabled employees at a company. So the sample is purposefully selected from this population.

- Snowball Sampling: The research participants recruit other participants for the study. It is used when participants required for the research are hard to find. It is called snowball sampling because like a snowball, it picks up more participants along the way and gets larger and larger.

Bias: Systematic error.

Sample Bias: A sample that misrepresents the population.

μ(mu) | (x̄) : The mean of the population (sample).

Types Of Sampling bias:

- **Selection bias:** Some elements of the population is excluded or underrepresented in sample. Eg. Population is teachers and doesn't include math dept.

- **Confirmation bias:** It occurs when the person performing the statistical analysis has some predefined assumption.

- **Time interval bias:** It is caused intentionally by specifying a certain time range to favour a particular outcome.

- **Non-response bias:** No data is collected from individuals who have been selected. Eg. x person is given a survey, but was too busy to fill it out.

- **Response bias:** Process distorts responses. Eg. given question asks about customer satisfaction, and the options given are Very Satisfied, Satisfied and Dissatisfied, it must be including two of each of the positive and negative options.

Selection bias

Bias that results from how observations in data are selected.

Data snooping: Searching through the data in order to find something interesting.

Tip : split apart a section of the data(train-test split method) so that we have unseen data to test.

Data splitting

Vast search effect: A form of selection bias where one repeatedly runs different models asking different questions on a large dataset which increases the chances that an outlier is deemed interesting.

Tip :**Target shuffling:** Create a model and note it's efficiency.Shuffle the target variable the model won't be performing well, it's expected that the quality of a model accuracy will decrease drastically. Repeat the process (target shuffling + model building + model evaluation) many times and record the bogus result in order to receive good estimates on how well the real model performs in comparison to randomised data.

Ref : ***https://towardsdatascience.com/shuffling-rows-in-pandas-dataframes-eda052275635***

Regression to the mean

Refers to phenomenon that if a variable is extreme the first time you measure it, it will be closer to the average the next time you measure it.
In technical terms, it describes how a random variable that is outside the norm eventually tends to return to the norm.
For example, your odds of winning on a slot machine stay the same. You might hit a "winning streak" which is, technically speaking, a set of random variables outside the norm. But play the machine long enough, and the random variables will regress to the mean (i.e. "return to normal") and you'll end up losing.

Ref : *https://www.statisticshowto.com/regression-mean/*

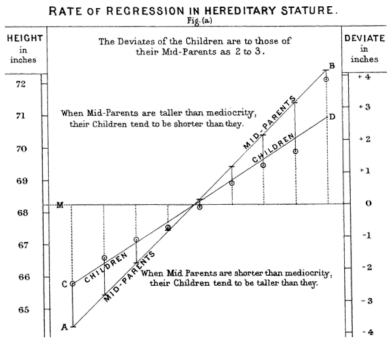


Figure 2-5. Galton's study that identified the phenomenon of regression to the mean

Sampling Distribution of a Statistic

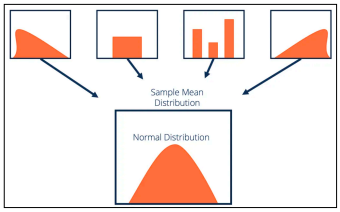
Working with large dataset sampling becomes very imp and a sample is not a perfect representation of the population it is important to understand *sampling variability*.

Sample statistics: statistics that is gathered on the sample data like sample mean, sample median, etc.

Data distribution which are frequency distribution of individual values in a dataset.

Sampling Distribution: The frequency distribution of a sample statistic over many samples or resamples (combination of data distribution values and sample distribution)

Central Limit Theorem: Although many population distribution are not normally distributed, the means drawn from multiple samples from the data when plotted will resembles a bell - shaped normal curve. Larger the sample, the narrower the distribution of the sample statistic.



CLT is relevant for A/B testing and bootstrapping is more relevant in data science problems.

Standard Error(measure the variability of a sampling distribution):

It is a single metric that sums up the variability in the sampling distribution for a statistic.From CLT , increasing the sample size decreases the error.

$$SE = \frac{\sigma}{\sqrt{n}}$$ *(Sample std deviation)* *(Number of samples)*

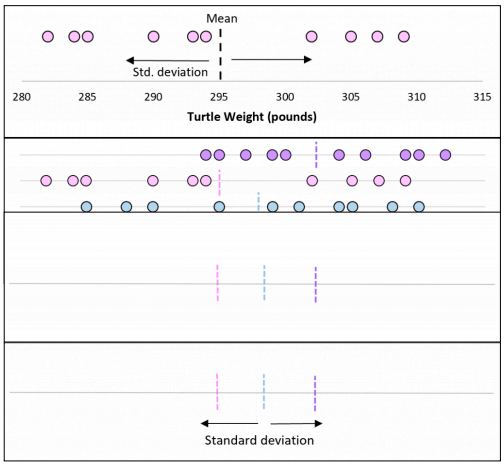
Ref: https://www.statology.org/standard-deviation-vs-standard-error/

Example and explanation:

1. Take repeated samples from the same population and record the sample mean and sample standard deviation for each sample.

2. Plot each of the sample means on the same line.

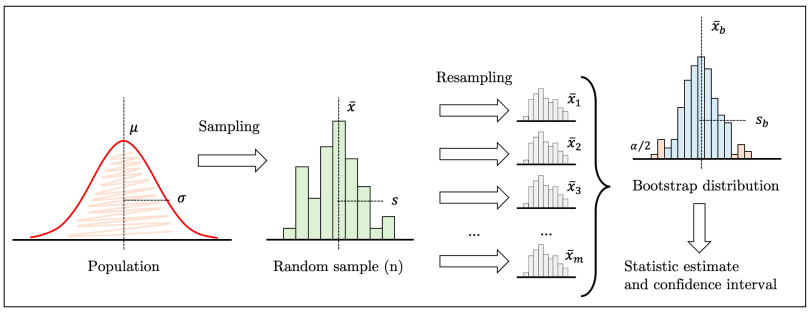
3. The standard deviation of these means is known as the standard error.



Each time drawing new sample is a vain and modern statistics favours method called bootstrap.

The bootstrap

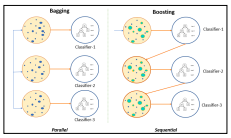
Powerful Sampling method for assessing the variability of a sample statistic, where we draw resamples(sampling with replacement) from a sample, replace the values in the resample back to the sample, then draw them again.No assumptions about the sample being normally distributed.



Resampling: the process of taking repeated samples from observed data, includes both bootstrap and permutation (shuffling) procedures.

The bootstrap does not compensate for a small sample size; it does not create new data, nor does it fill in holes in an existing data set. It merely informs us about how lots of additional samples would behave when drawn from a population like our original sample.

Bagging is the process of running multiple classification and regression trees, and then averaging their predictions. It generally performs better than a single tree.Bootstrapping is used to improve the efficiency of Decision Trees.

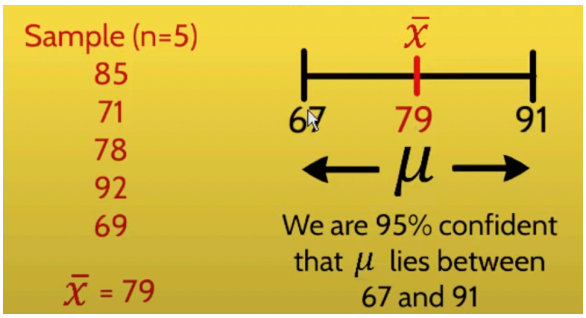


Confidence Intervals

Frequency tables, histograms, boxplots, and standard errors are all ways to understand the potential error in a sample estimate.Confidence interval is another one which presents an estimate not as a single number but as a range of values that's likely to include population value with a certain degree of confidence.
It answer's: "Given a sample result, what is the probability that (something is true about the population)?"

Confidence level: The percentage of confidence intervals, constricted in the same way from the same population, expected to contain the statistic of interest.

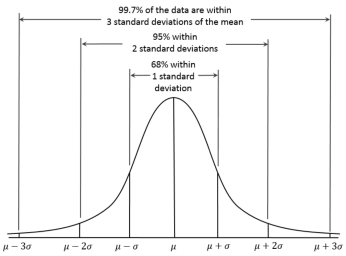
Interval endpoints: The top and bottom of the confidence interval.



The bootstrap is a general tool that can be used to generate confidence intervals generated by formulas for most statistics (especially the t-distribution), or model parameters.

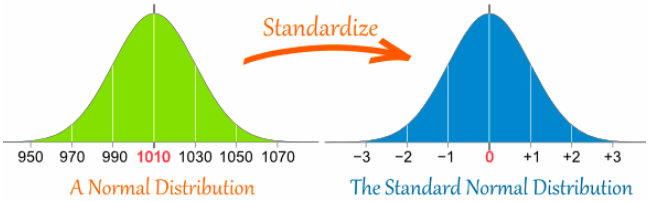
- The higher the level of confidence, the wider the interval.
- The smaller the sample, the wider the interval (i.e., the greater the uncertainty).

Normal distribution



N**ormal distribution** :Bell-shaped distribution is a normal distribution with 68% of the data lies within one standard deviation of the mean, and 95% lies within two standard deviations. The normal distribution is also referred to as a Gaussian distribution.

Standard normal distribution is one in which the units on the x-axis are expressed in terms of standard deviations away from the mean.A standard normal distribution is with mean = 0 and standard deviation = 1.



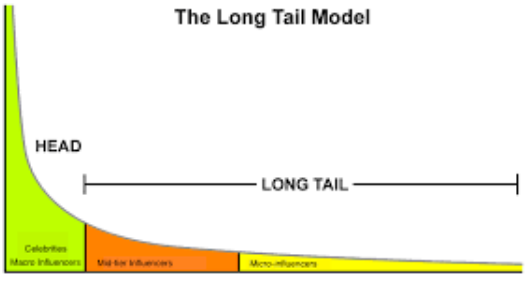
<u>To compare data to a standard normal distribution</u>, subtract the mean and then divide by the standard deviation, this is called normalization or standardization and the transformed value is termed a z-score.

QQ-Plots : Quantile-Quantile plot, used to determine if a distribution of values is close to a specified distribution.

If the QQ-plot is roughly diagonal going upwards left to right then it can be considered a normal distribution.This plot orders the z- score from low to high, plotting the z- score on the y-axis and x=axis is what quantile (percentile) that ranked value is in normal distribution. (theoretically what would the quantile be if the data were normal)

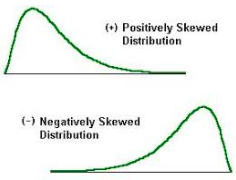
Long -Tailed distribution

Extreme values with low frequency are present on many distribution like as data being normal in the middle but having much longer tails.These tails are formed due to the points close to the line of the data within one standard deviation of the mean.



Tail : The long narrow portion of a frequency distribution, where relatively extreme values occur at low frequency.

Skew : Where one tail of a distribution is longer than the other.

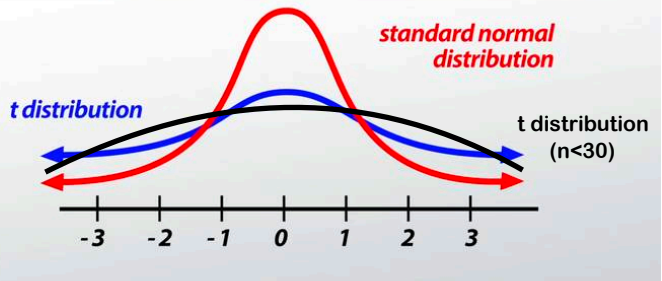


Student's t-distribution

It is a normally shaped distribution, except that it is a bit thicker and longer on the tails.Distributions of sample means are typically shaped like a t-distribution, and there is a family of t-distributions that differ depending on how large the sample is.As you take more samples, the distribution will more closely resemble a normal distribution.

t-statistics(eg. compare mean of different samples) often used in A/B test and regressions.

Degree of freedom: A parameter that allows the t-distribution to adjust to different sample sizes, statistics, and number of groups.

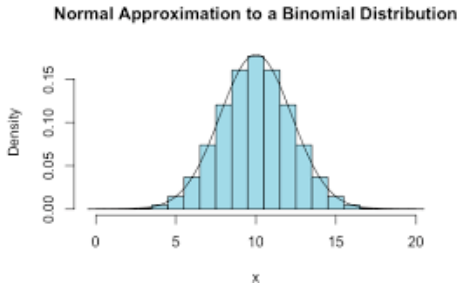


Binomial distribution (buy or don't buy, survive or die, etc)

It is the idea of a set of trials, each trial having two possible outcomes(the outcome of interest success/rare outcome assigned as 1) with definite probabilities(one with probability p and the other with probability $1 - p$).The binomial distribution is the frequency distribution of the number of successes (x) in a given number of trials (n) with specified probability (p) of success in each trial.

With large n, and provided p is not too close to 0 or 1, the binomial distribution can be approximated by the normal distribution.

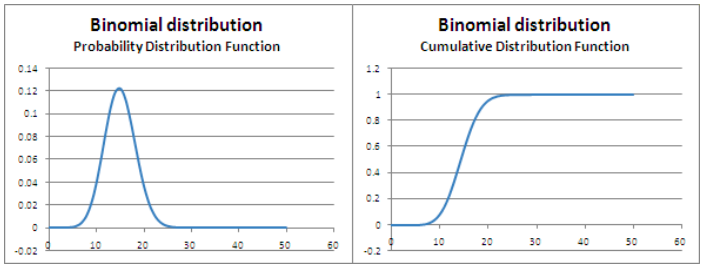
The mean of a binomial distribution is (n * p)
The variance is n * p(1-p)



PDF - Probabilities for continuous random variables.
Eg. Probabilities for range of outcome P(X= 5 TO 6)

PMF - Probabilities for discrete random variables.
Eg. Probabilities for range of outcome P(X= 5)

CDF - Cumulative Probabilities associated with a function.
Eg. Cumulative value from negative infinity up to a random variable X . (X<5)



Chi - Square distribution

Chi-square statistic is a measure of assessing the goodness of fit between observed values and those expected theoretically. It is the difference between the observed and expected values, divided by the square root of the expected value, squared, then summed across all categories. This process standardizes the statistic so it can be compared to a reference distribution.

It is useful for determining whether multiple treatments (an 'A/B/C ... test') differ from one another in their effects.

The chi-square distribution is the distribution of this statistic under repeated resampled draws from the *null model*. A *low chi-square value* for a set of counts indicates that they *closely follow the expected distribution*. A *high chi-square* indicates that they *differ markedly from what is expected.*

There are a variety of chi-square distributions associated with different degrees of freedom.

F distribution / ANOVA test

Used to compare the effect of multiple treatments (A/B/C test referred to in the chi-square distribution, except we are dealing with measured continuous values rather than counts) across different groups.

In this case we are interested in the extent to which differences among group means are greater than we might expect under normal random variation. The F-statistic measures this and is the ratio of the variability among the group means to the variability within each group (also called residual variability).This comparison is termed an analysis of variance (ANOVA)

Poisson and related distribution

Poisson distribution : The frequency distribution of events occurring per unit of time or space of many samples.Lambda is the mean number of events that occurs in a specified interval of time or space. The variance for a Poisson distribution is also λ.

Eg. cars arriving at a toll plaza (events spread over time) and typos per 100 lines of code (events spread over space)

Exponential distribution : The frequency distribution of the time or distance from one event to the next event.

Poisson and Exponential distributions are only useful as long as number of occurrence of an instance are relatively consistent over time.

Eg.Time required in between per service calls.

Weibull distribution : Used to model event occurrence when the event probability change over time.

Eg. Mechanical car failure - the risk of failure increases as time goes by.