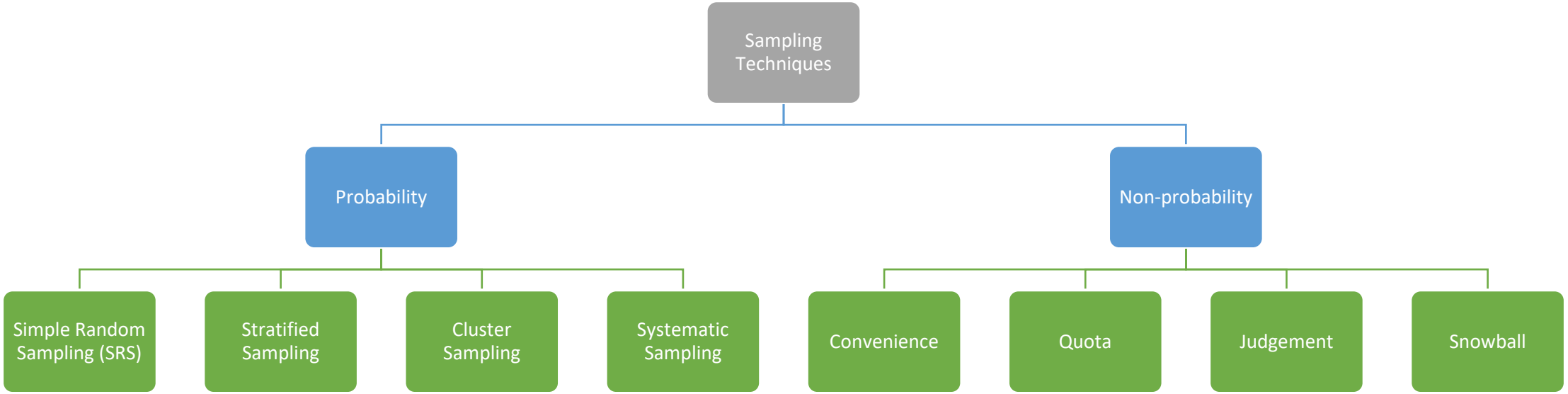
Applied Statistics for Data Scientists with R

Class 14: Sampling Techniques and Sampling Distributions Fundamentals

Sampling Techniques

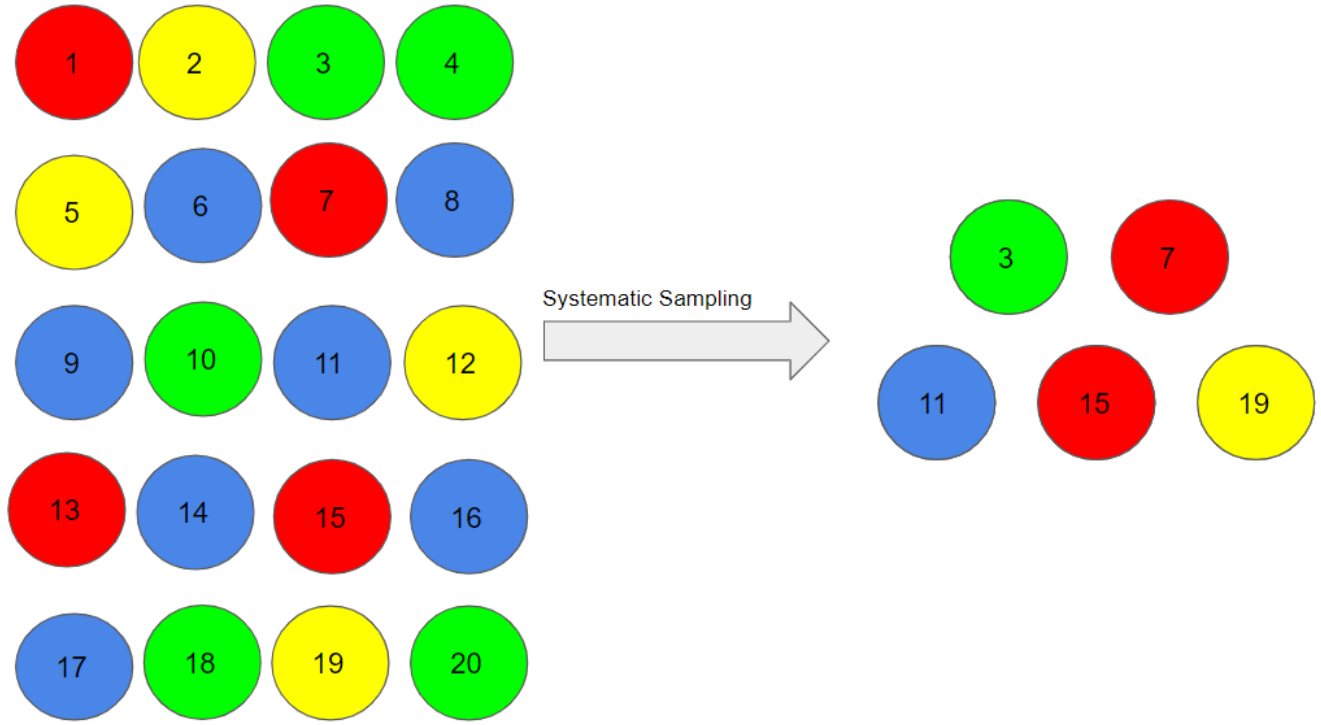


Simple Random Sampling

- Every individual has an equal and independent chance of being selected.

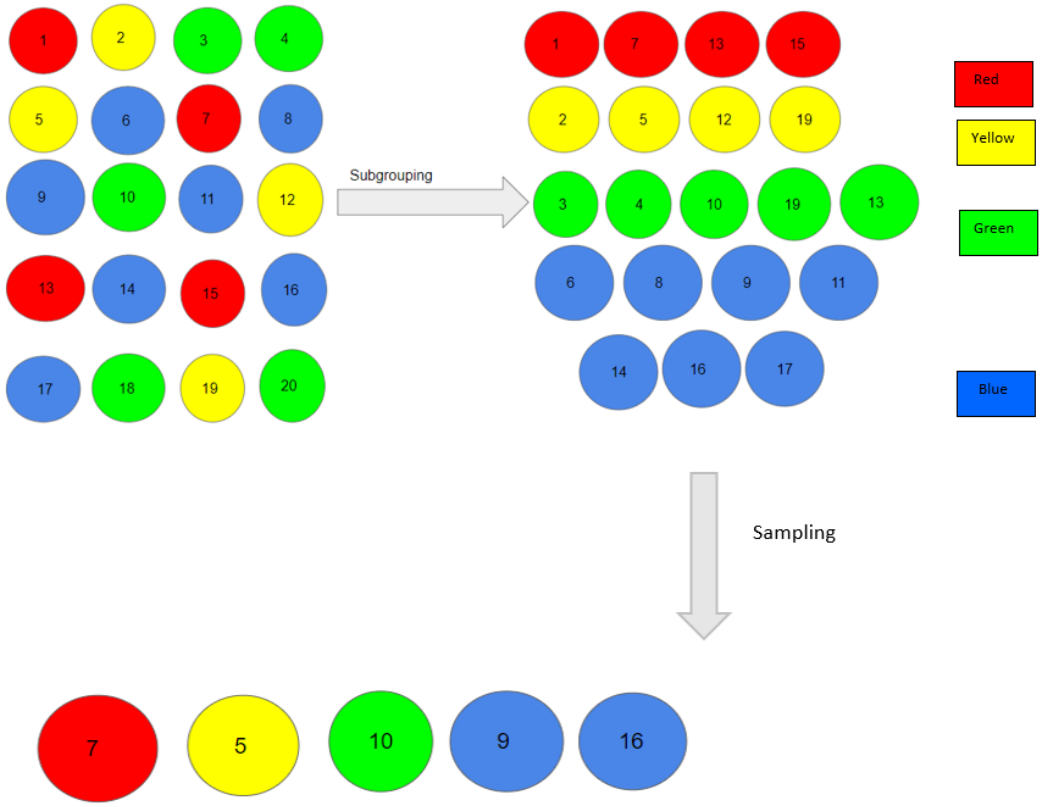
Systematic Sampling

- Select every k-th individual from a list after a random starting point.



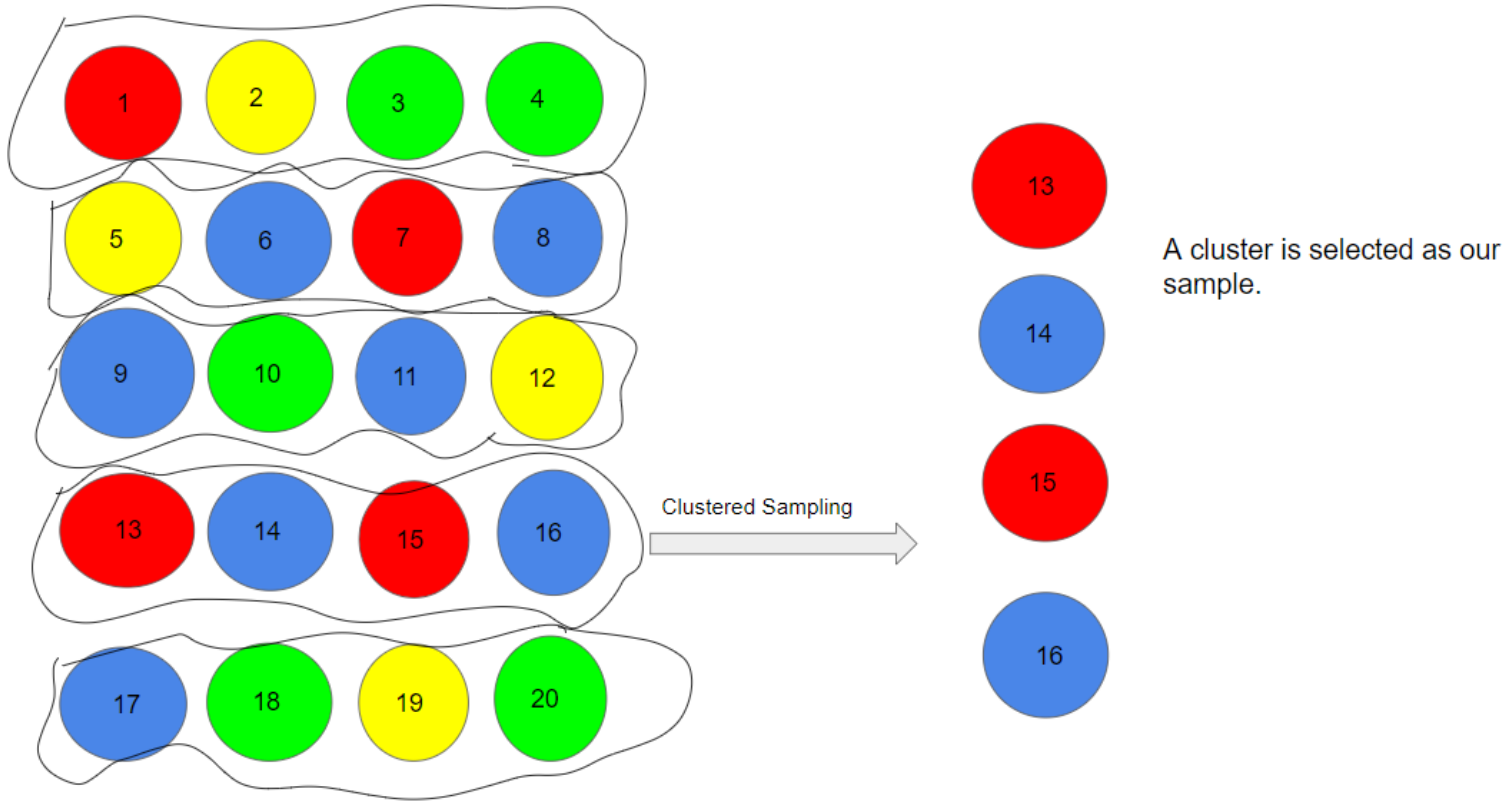
Stratified Sampling

- The population is divided into homogeneous subgroups (strata), and a random sample is taken from each stratum.



Cluster Sampling

- The population is divided into clusters (usually based on geography or natural groupings), and entire clusters are randomly selected.



Multi-stage Sampling

- Combines multiple sampling methods, often starting with cluster sampling and then applying random or systematic sampling within clusters.

Non-probability Sampling Techniques

- Convenience Sampling
  - Collect data from individuals who are *easily accessible* or *willing to participate*.
  - Surveying shoppers at a mall.
  - Online polls shared on social media.

- Quota sampling
  - Select participants to meet pre-defined *quotas* (e.g., age, gender, ethnicity) representing the population.
  - Interviewing 50% women and 50% men in a study.
  - Ensuring 20% of participants are aged 18–25.

Non-probability Sampling Techniques

- Judgement sampling
  - Researcher *handpicks* participants thinking to be most representative to the population of interest.
  - Choosing schools in specific neighborhoods for an education survey.

- Snowball sampling
  - Existing participants recruit others from their network (chain-referral).
  - Studying hidden populations (e.g., homeless individuals, rare disease patients).
  - Research on social networks (e.g., gang members).

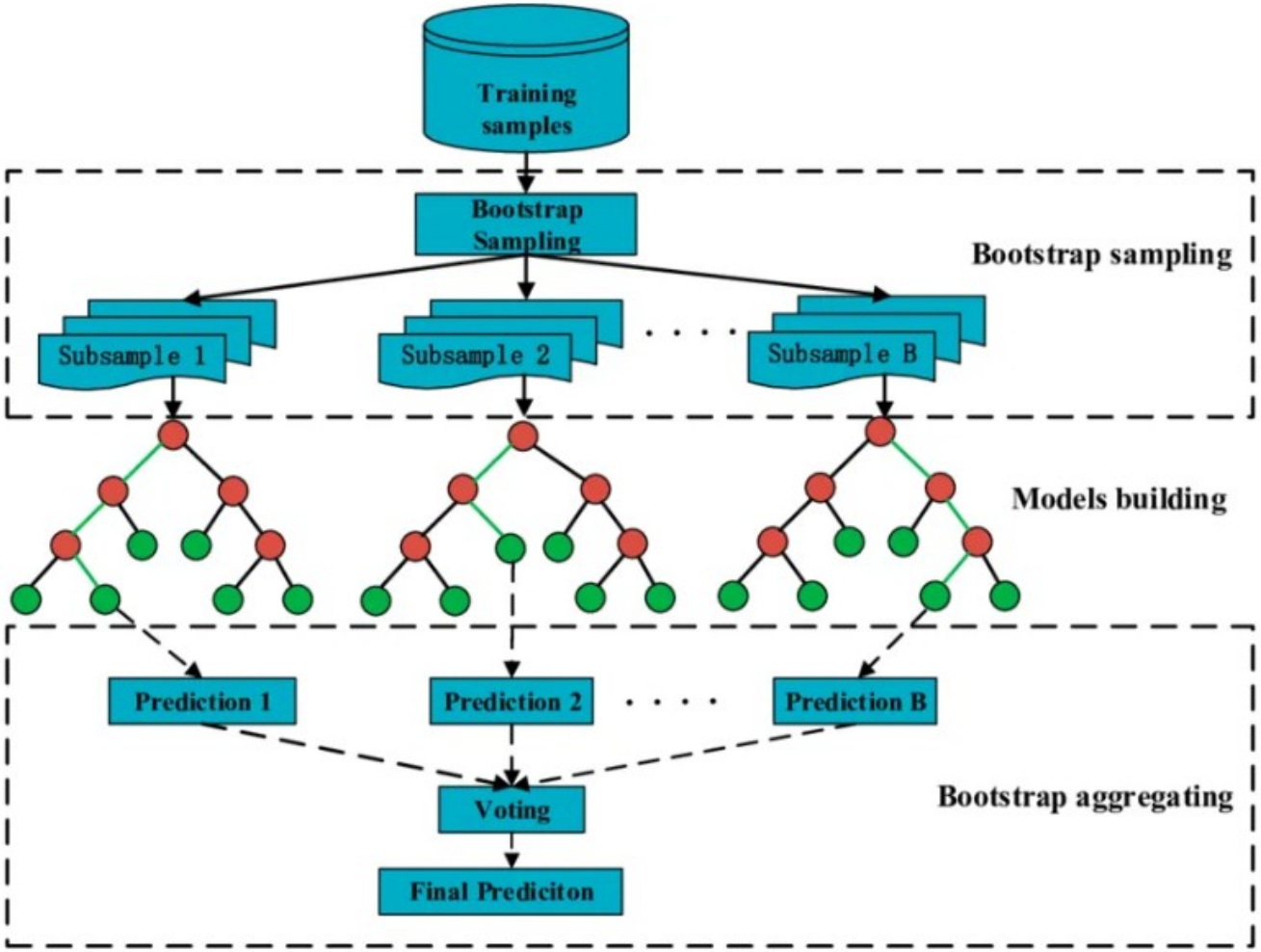
Resampling Techniques: Bootstrap and Jackknife

- Resampling is the creation of new samples based on one observed sample

- Bootstrap: Sampling with replacement from the original sample to create many new "bootstrap samples.“
  - Useful for creating sampling distribution.
  - Useful for estimating confidence interval.

- Jackknife: Systematic sampling without replacement by leaving out one observation at a time to create "jackknife samples.“
  - Useful for estimating bias and variance of an estimate.

Resampling Techniques: Bootstrap and Jackknife



Sampling Distributions

- A sampling distribution is the probability distribution of a statistic (like the mean, variance, or proportion) calculated from multiple random samples of the same size, say n, from a population of size N.

Some Terminology

- Standard Error is a measure of variability in sampling distributions.
- Smaller SE means the sample statistic is more stable across different samples.

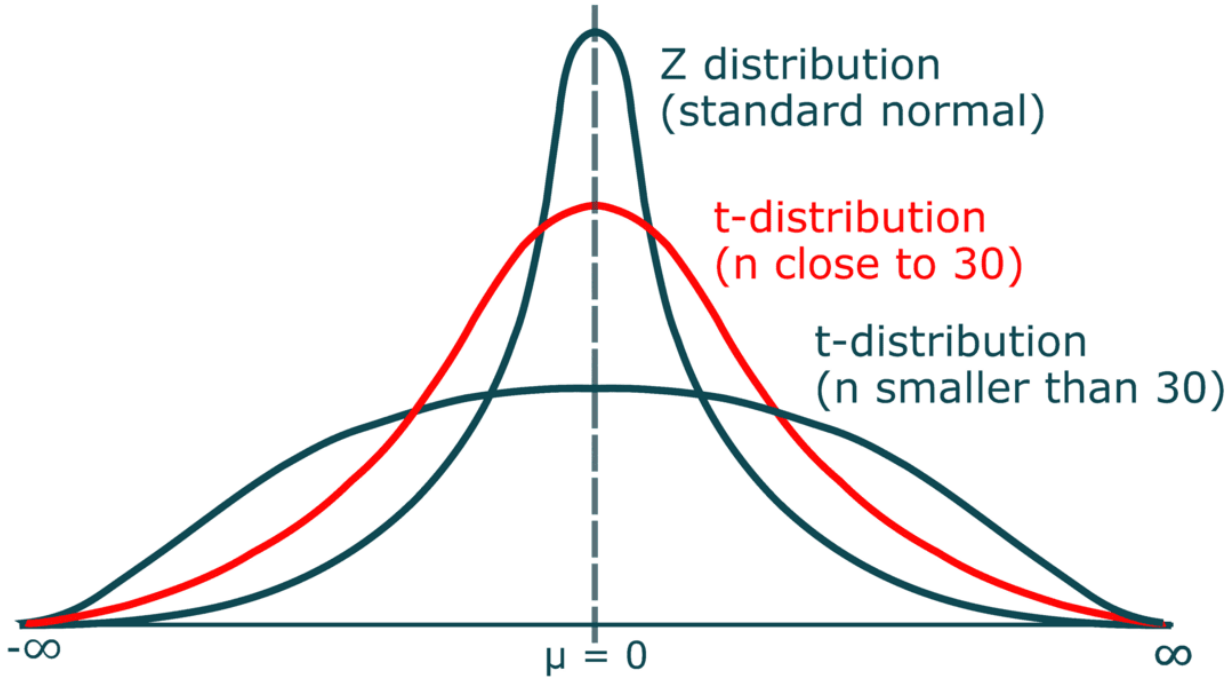
Central Limit Theorem

- Regardless of the population’s distribution, the sampling distribution of the sample mean tends to be normal as the sample size increases ($n \geq 30$ is often sufficient).

Student's t distribution

- Describes the distribution of mean of samples when the standard deviation is unknown.
- Used to compare means.
- Has degrees of freedom, $df = n - 1$

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

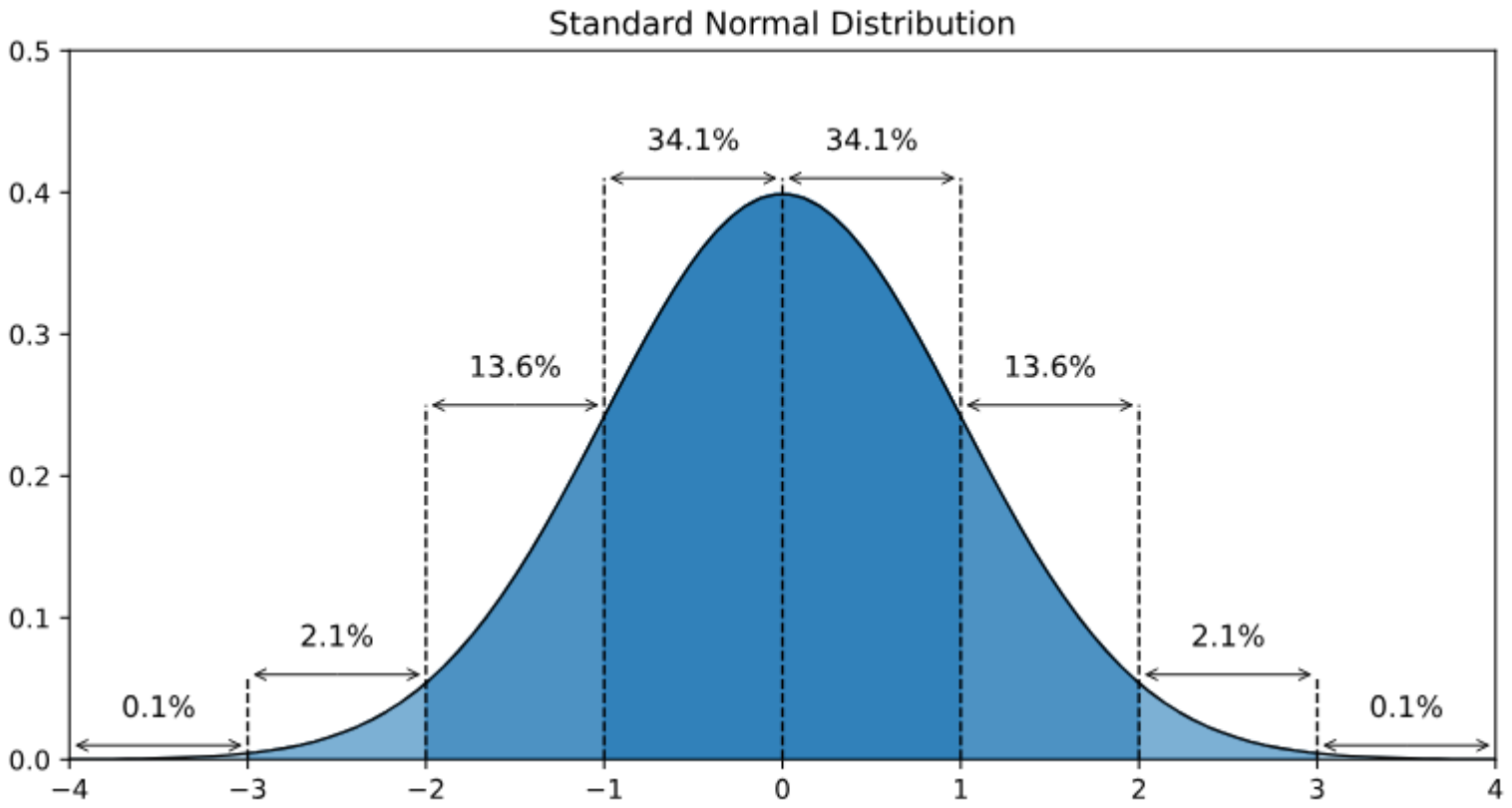


Fun fact: William Sealy Gosset's pseudonym was Student

z-distribution

- Describes the distribution of mean of samples when the standard deviation is known.

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

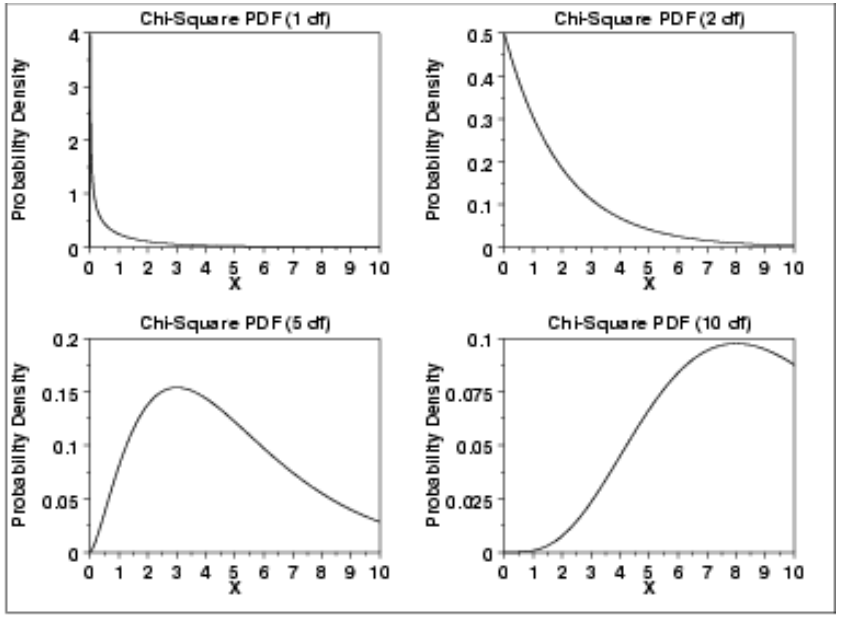


Chi-square Distribution

- Distribution of Sample Variance (s^2) relative to population variance.
- Also, sum of squared deviations (goodness-of-fit, test of independence).
- This is right skewed since variances are always positive, and their squared nature leads to right-skewness.
- $df = k - 1$, where *k* is the number of categories or groups being tested.
- For variance tests, $df = n - 1$, where *n* is the sample size

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$



F Distribution

$$F = \frac{s_1^2}{s_2^2} = \frac{\text{Between-group variance (MSB)}}{\text{Within-group variance (MSW)}}$$

- Distribution of ratio of two variances.
- Used in comparing variances of two groups, and in ANOVA.
- Has degrees of freedom:
  - $df_1 = k_1 - 1$, where k_1 is the number of groups or treatments in the numerator
    (e.g., between-group variation in ANOVA).
  - $df_2 = k_2 - 1$, where k_2 is the number of observations or groups in the denominator
    (e.g., within-group variation in ANOVA).