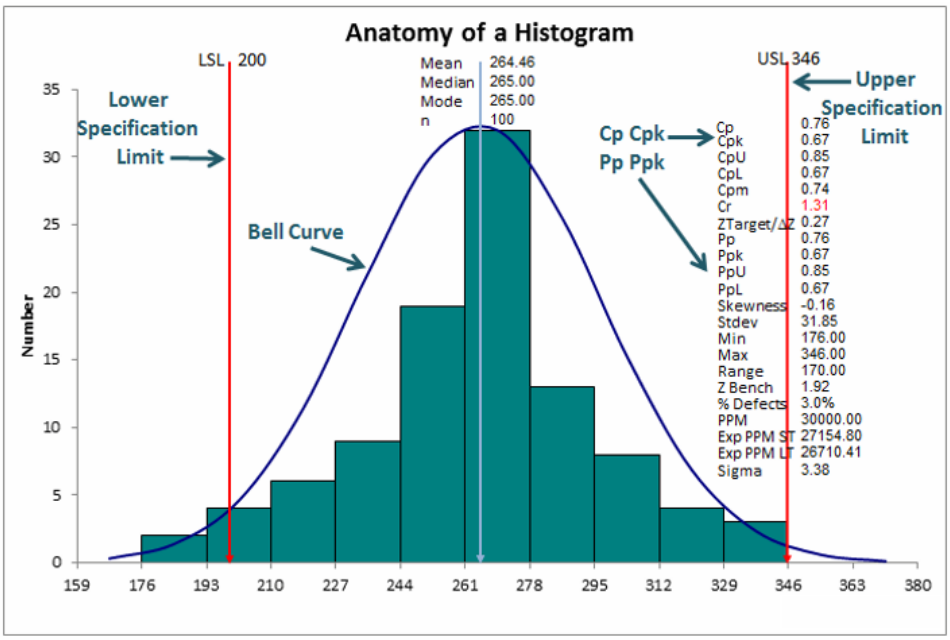
How to interpret Histogram for Univariate analysis – Numerical variable



A. Interpretation of the Histogram Components:

- **Horizontal Axis (X-axis):** Represents the range of values for the numerical variable being analyzed. In this case, it spans from approximately 159 to 380. The axis is divided into intervals or **bins**.
- **Vertical Axis (Y-axis):** Represents the **frequency** or **number** of data points that fall within each bin on the x-axis. Here, the frequency ranges from 0 to 35.
- **Bars:** Each **vertical bar** represents one bin.
    - The **width** of the bar corresponds to the size of the interval for that bin on the x-axis.
    - The **height** of the bar indicates the number of data points that fall within that specific interval. Taller bars indicate a higher frequency of values within that range.
- **Shape of the Distribution:** The overall shape formed by the bars provides a visual representation of the distribution of the data:

- **Central Tendency:** We can visually estimate where the center of the data lies. In this case, the tallest bars are around 261-278, suggesting the central tendency. The provided statistics confirm this: Mean = 264.46, Median = 265.00, Mode = 265.00.
- **Spread (Variability):** The width of the overall distribution (from the leftmost to the rightmost bars with significant height) indicates the spread of the data. Here, the data is spread out over a considerable range. The standard deviation (Stdev = 31.85) quantifies this spread.
- **Symmetry:** We can assess if the distribution is symmetric around its center. The blue **bell curve** overlaid on the histogram represents a normal distribution. The histogram's shape roughly follows this bell curve, suggesting a relatively symmetric distribution, although there might be a slight lean to the right. The skewness value (-0.16, close to zero) also indicates near symmetry.
- **Modality:** The number of peaks (highest bars) indicates the mode(s) of the distribution. This histogram appears to be **unimodal** (one main peak around 265), consistent with the provided mode.
- **Outliers:** While not explicitly highlighted here, bars that are isolated and far from the main body of the distribution could indicate potential outliers.

- **Overlaid Normal Distribution (Bell Curve):** The blue curve represents a theoretical normal distribution with the same mean and standard deviation as the sample data. This helps to visually assess how well the actual data distribution approximates a normal distribution.
- **Specification Limits (LSL and USL):** The vertical red lines indicate the **Lower Specification Limit (LSL = 200)** and the **Upper Specification Limit (USL = 346)**. These are external boundaries defining acceptable values for the process or product being measured. The histogram shows how the data distribution relates to these limits. We can see that some data points fall outside these limits, indicating potential non-conforming items.
- **Process Capability Indices (Cp, Cpk, Pp, Ppk, etc.):** The text box on the right provides various process capability indices. These are

calculated metrics that assess how well the process output (the data) meets the specification limits. For example:

- **Cpk (0.76):** Indicates how centered the process is within the specification limits, considering the process variation. A value less than 1 suggests the process is not consistently within specifications.
- **Ppk (0.85):** Similar to Cpk but uses the overall process standard deviation instead of the within-subgroup standard deviation. These indices are derived from the distribution's mean, standard deviation, and the specification limits.

Histograms are the best choice for univariate analysis in the following scenarios:

- **Understanding the Distribution of a Single Numerical Variable:** Their primary purpose is to visualize the frequency distribution, making it easy to see where the data is concentrated, its spread, and its shape.
- **Assessing Normality:** By comparing the histogram's shape to a normal distribution curve (often overlaid), you can visually assess whether the data is approximately normally distributed. This is important for many statistical analyses that assume normality.
- **Identifying Skewness and Kurtosis:** The shape of the histogram clearly shows if the data is skewed (asymmetric) or has heavy or light tails (kurtosis).
- **Detecting Multiple Modes (Multimodality):** Histograms can reveal if the data has multiple peaks, suggesting the presence of different underlying groups or processes.
- **Identifying Potential Outliers:** Isolated bars far from the main distribution can indicate potential outliers.
- **Analyzing Process Capability:** When combined with specification limits, histograms are crucial for visually assessing whether a process is producing output within acceptable ranges and for calculating process capability indices.
- **Communicating the Distribution to a Non-Technical Audience:** The bar chart format is generally easy to understand, making histograms a

good way to communicate the distribution of a numerical variable to a broad audience.

In contrast to box plots or stem-and-leaf plots:

- **Histograms are better for larger datasets** as they provide a clear overall picture of the distribution without becoming cluttered.
- **Histograms emphasize the frequency of values within intervals**, while box plots emphasize summary statistics (quartiles, median, outliers).
- **Histograms provide more detail about the shape of the distribution** (e.g., the number of peaks), which might be less apparent in a box plot.

In summary, histograms are the go-to visualization for understanding the shape, central tendency, spread, and presence of multiple modes in a single numerical variable, especially for larger datasets. When combined with specification limits, they are essential for process capability analysis.