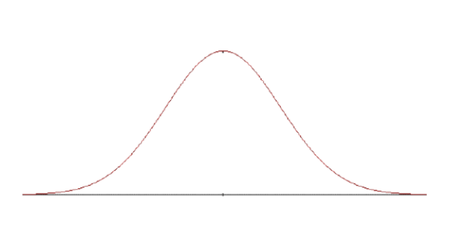
ISCB20.05–Introduction to Biostatistics

Md. Jubayer Hossain
https://jhossain.me/
December 19, 2020

Lead Organizer, Introduction to Scientific Computing for Biologists
Founder, Health Data Research Organization

Section–2.2: Interpreting Data Using Descriptive Statistics

Distribution

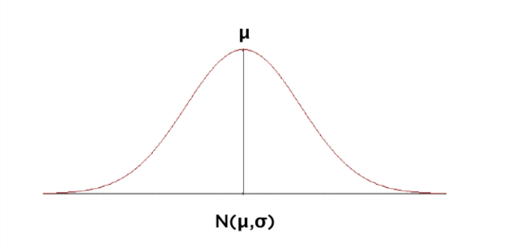


(a) Values close to the mean are more likely

(b) All values are equally likely

The Normal Curve

- The distributions of most continuous random variables will follow the shape of the normal curve.
- Mean, Median and Mode all exist at the center.

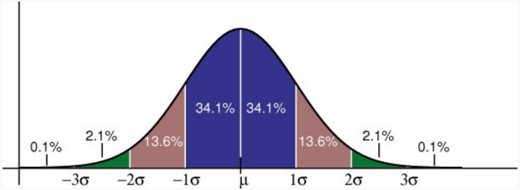


A formula which tells how likely a particular value is to occur in your data.

The Empirical Rule–1

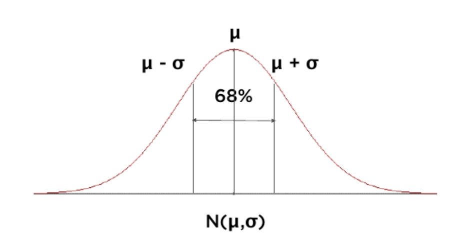
The empirical rule tells you what percentage of your data falls within a certain number of standard deviations from the mean.

- 68% of all values fall within 1 standard deviation of the mean.
- 95% of the all values fall within 2 standard deviation of the mean.
- 99.7% of the all values fall within 3 standard deviation of the mean.



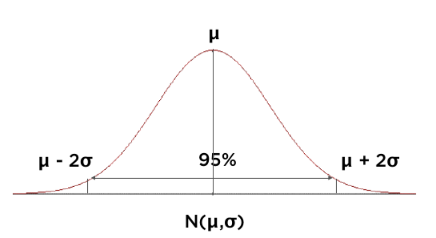
Source:
`https://www.statisticshowto.com/probability-and-statistics/normal-distributions/`

The Empirical Rule–2



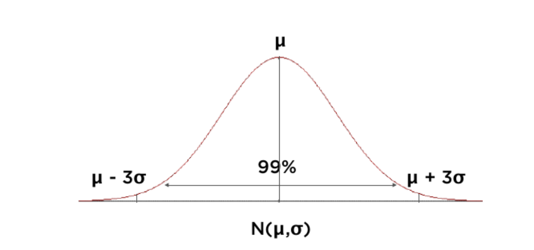
68% within 1 standard deviation of mean

The Empirical Rule–3



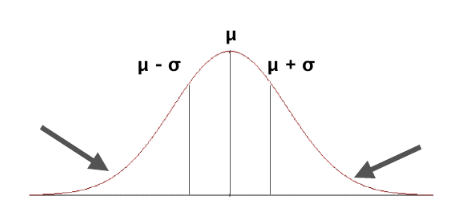
95% within 2 standard deviations of mean

The Empirical Rule–4



99% within 3 standard deviations of mean

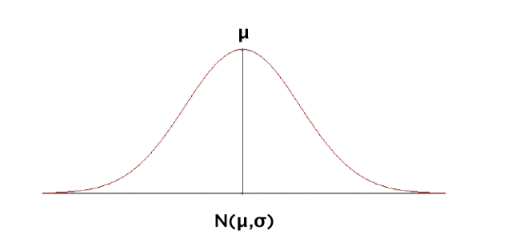
Impact of Outliers



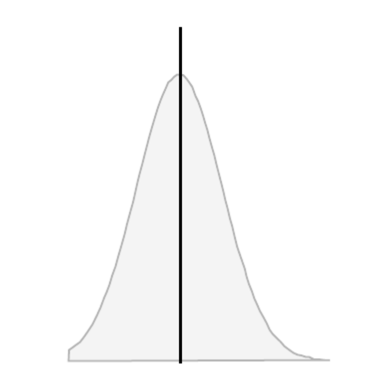
There will be few extreme values - the number of extreme values at either side of the mean will be the same.

Properties of a Normal Distribution

- The mean, mode and median are all equal.
- The curve is symmetric at the center (i.e. around the mean, μ).
- Exactly half of the values are to the left of center and exactly half the values are to the right.
- The total area under the curve is 1.



Role of Sigma(σ)



(a) Small Standard Deviation(σ)

Few points far from the mean

(b) Large Standard Deviation(σ)

Many points far from the mean

Z-Scores

- Z-Scores are standardized values that can be used to compare scores in different distributions.
- Simply put, a z-score (also called a standard score) gives you an idea of how far from the mean a data point is. But more technically it's a measure of how many standard deviations below or above the population mean a raw score is.
- A z-score can be placed on a normal distribution curve. Z-scores range from -3 standard deviations (which would fall to the far left of the normal distribution curve) up to $+3$ standard deviations (which would fall to the far right of the normal distribution curve).
- In order to use a z-score, you need to know the population mean μ and also the population standard deviation σ.

Calculating Z-Score

$$Z = \frac{\overline{x} - \mu}{\sigma}$$

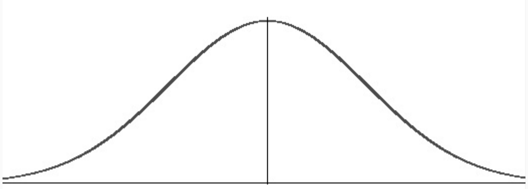
- $\overline{x} \rightarrow$ mean
- $\mu \rightarrow$ population mean
- $\sigma \rightarrow$ population standard deviation

Calculating Z-Score

For example, let's say you have a test score of 190. The test has a mean (μ) of 150 and a standard deviation (σ) of 25. Assuming a normal distribution, your z score would be

Skewness–1

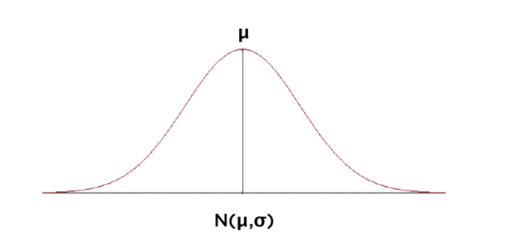
- A measure of asymmetry around the mean.
- If one tail is longer than another, the distribution is skewed.
- These distributions are sometimes called asymmetric or asymmetrical distributions as they don't show any kind of symmetry.
- Symmetry means that one half of the distribution is a mirror image of the other half.



Source: https://www.statisticshowto.com/probability-and-statistics/

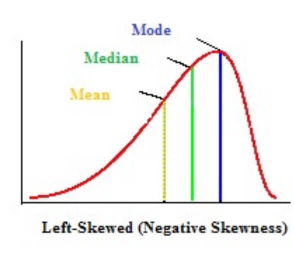
Skewness–2

- Normally distributed data: skewness $= 0$
- Extreme values are equally likely on both sides of the mean.
- Symmetry about the mean



Negative Skewness

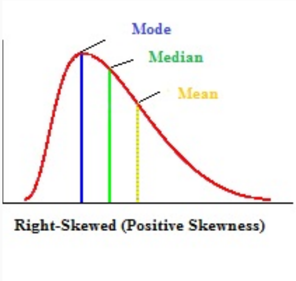
- A left-skewed distribution has a long left tail.
- Left-skewed distributions are also called negatively-skewed distributions.
- That's because there is a long tail in the negative direction on the number line. The mean is also to the left of the peak.



Source: `https://www.statisticshowto.com/probability-and-statistics/`

Positive Skewness

- A right-skewed distribution has a long right tail.
- Right-skewed distributions are also called positive-skew distributions.
- That's because there is a long tail in the positive direction on the number line. The mean is also to the right of the peak.

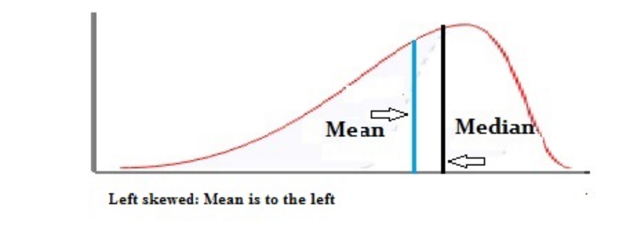


Source: `https://www.statisticshowto.com/probability-and-statistics/`

Mean and Median in Skewed Distributions

In a normal distribution, the mean and the median are the same number while the mean and median in a skewed distribution become different numbers.

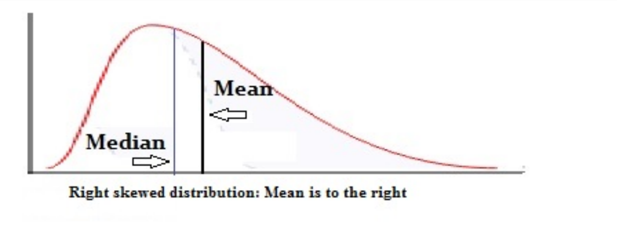
- A left-skewed, negative distribution will have the mean to the left of the median.



Source: `https://www.statisticshowto.com/probability-and-statistics/`

Mean and Median in Skewed Distributions

- A right-skewed, negative distribution will have the mean to the right of the median.



Source: `https://www.statisticshowto.com/probability-and-statistics/`

Kurtosis

Kurtosis is a statistical measure that defines how heavily the tails of a distribution differ from the tails of a normal distribution. In other words, kurtosis identifies whether the tails of a given distribution contain extreme values.

Kurtosis

- Normally distributed data: kurtosis = 3
- Excess kurtosis = kurtosis - 3
- Kurtosis   Tail risk
- High kurtosis =¿ extreme events more likely than in normal distribution.

Point Estimation

The value of any statistic of any that estimates the value of a parameter is called a point estimation.
$\overline{x} = 2.9 \rightarrow \mu = 3.00$

We rarely know if our point estimate is correct because it is merely an estimation of the actual value.

Confidence Interval

A Confidence Interval is a range of values we are fairly sure our true value lies in.

Confidence Interval	Z-Value
90%	1.65
95%	1.69
99%	2.58
99.9%	3.291

Calculating Confidence Intervals

We measure the heights of 40 randomly chosen men, and get a mean height of 175cm,We also know the standard deviation of men's heights is 20cm.

- **Step-1**
  - the number of observations(n)
  - the mean $\overline{x}$
  - the standard deviation s
- **Step-2:**
  - number of observations n = 40
  - mean X = 175
  - standard deviation s = 20
- **Step-3:** decide what Confidence Interval we want: 95% or 99% are common choices. Then find the "Z" value for that Confidence
- **Step-4:** use that Z value in this formula for the Confidence Interval.

$$X \pm Z \frac{s}{\sqrt{n}}$$

Calculating Confidence Intervals

$$X \pm Z \frac{s}{\sqrt{n}}$$

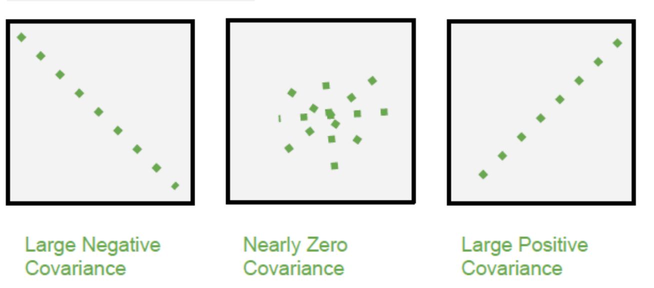
$$175 \pm 1.960 \times \frac{20}{40} = 175cm \pm 6.20$$

Bivariate Analysis

- **Covariance:** Measures relationship between two variables specially whether greater values of one variable correspond to greater values in the other.
- **Correlation:** Similar to covariance; measures whether greater values of one variable correspond to greater values in the other. Scaled to always lie between $+1$ and -1

Covariance

- Covariance is a measure of how much two random variables vary together.
- It's similar to variance, but where variance tells you how a single variable varies, covariance tells you how two variables vary together.



Source: https://www.statisticshowto.com/covariance/

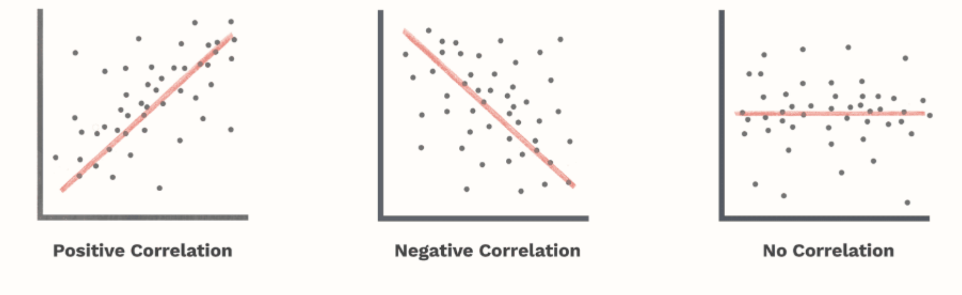
Covariance

$$cov(x, y) = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{n - 1}$$

- $cov(x, y) \rightarrow$ covariance between x and y
- $x_i \rightarrow$ data value of x
- $y_i \rightarrow$ data value of y
- $\overline{x} \rightarrow$ mean of x
- $\overline{y} \rightarrow$ mean of y
- $n \rightarrow$ number of data values.

Correlation

- When two sets of data are strongly linked together we say they have a High Correlation.
- Correlation is **Positive** when the values increase together.
- Correlation is **Negative** when one value decreases as the other increases
- A correlation is assumed to be linear.



Interpretation

- 1 is a perfect positive correlation
- 0 is no correlation (the values don't seem linked at all)
- -1 is a perfect negative correlation

Pearson's r Correlation

- Pearson's r measures the strength of the linear relationship between two variables.
- Pearson's r is always between -1 and 1

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- $r \rightarrow$ correlation between x and y
- $x_i \rightarrow$ data value of x
- $y_i \rightarrow$ data value of y
- $\overline{x} \rightarrow$ mean of x
- $\overline{y} \rightarrow$ mean of y

Correlation Is Not Causation

- A common saying is "Correlation Is Not Causation".
- What it really means is that a correlation does not prove one thing causes the other.
- Causation means that one variable causes something to happen in another variable.
- To say that two things are correlated is to say that they are not some kind of relationship.
- In order to imply causation, a true experiment must be performed where subjects are randomly assigned to different conditions.

References

- https://www.formpl.us/blog/research/home
- https://www.scribbr.com/methodology/
- https://www.questionpro.com/blog/
- https://www.scribbr.com/category/research-process/
- https://ocw.jhsph.edu/index.cfm/go/viewCourse/course/FundEpi/coursePage/index/

Thank You

☺