Together for Tomorrow! Enabling People
Education for Future Generations

Chapter 4.

Probability and Statistics

UNIT 3.

Understanding of Statistics II

Samsung Innovation Campus

Unit 3. Understanding of Statistics II

- What this unit is about:
  - You will learn about descriptive statistics.
  - You will learn about the central limit theorem.
  - You will learn about point estimation and interval estimation.
- Expected outcome:
  - Ability to summarize the statistical properties of a sample.
  - Ability to differentiate the population from the sample.
  - Ability to apply interval estimation.
- How to check your progress:
  - Coding Exercises.
  - Quiz.

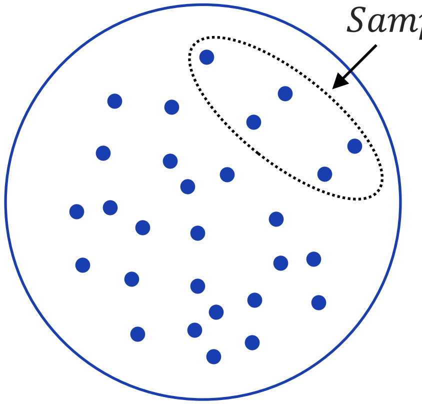
Chapter 4.

Probability and Statistics

Together for Tomorrow! Enabling People
Education for Future Generations

Together for Tomorrow! Enabling People
Education for Future Generations

Population vs Sample



Population

- Whole data set subject to the analysis. It can be real or conceptual.

Sample:

- A subset of the population.
- You analyze samples because the population is hard or impossible to reach.
- You would like to draw conclusions about the population using its samples.

Descriptive vs Inferential

- Descriptive statistics:
  - Summarizes the data without generalization as a primary goal.
  - Extracts the properties of data as is: sample statistics.
- Inferential statistics:
  - Analysis of the samples with the purpose of making generalized statements about the population.
- **The main difference between descriptive and inferential statistics lies in the goal.**

Descriptive Statistics (1/10)

Sample statistics:

- Mean value: $\bar{x}$
- Median: m
- Variance: s^2
- Standard deviation: $s = \sqrt{s^2}$
- Covariance: s_{XY}
- Correlation: r
- Skewness: ς
- Kurtosis: κ

Averages

Statistic of a collection of numerical observations is a single number that describes the entire collection.
Average is a statistic.

There are THREE kinds of Averages:

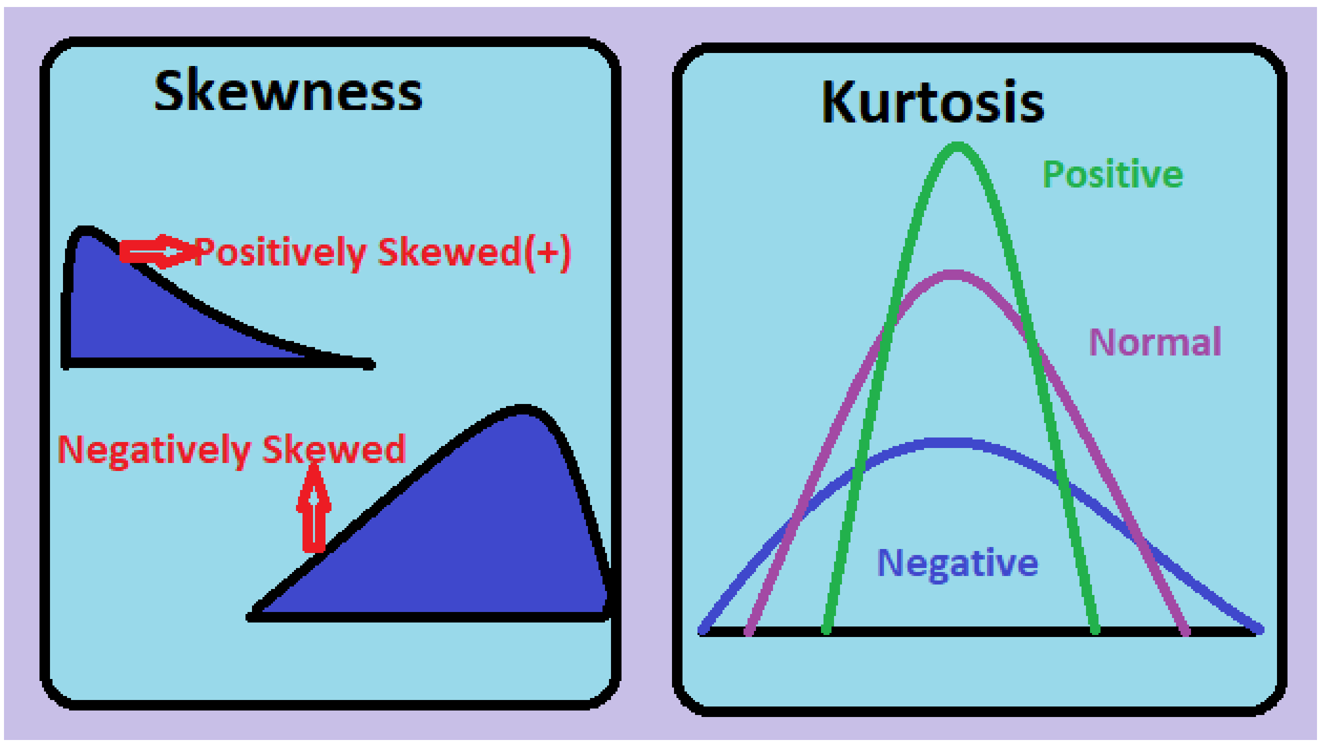
- Mean or Expected Value
- Median
- Mode

Example: For one quiz, 10 students have the following grades (on a scale of 0 to 10): 9, 5, 10, 8, 4, 7, 5, 5, 8, 7
Find Mean, Median, Mode

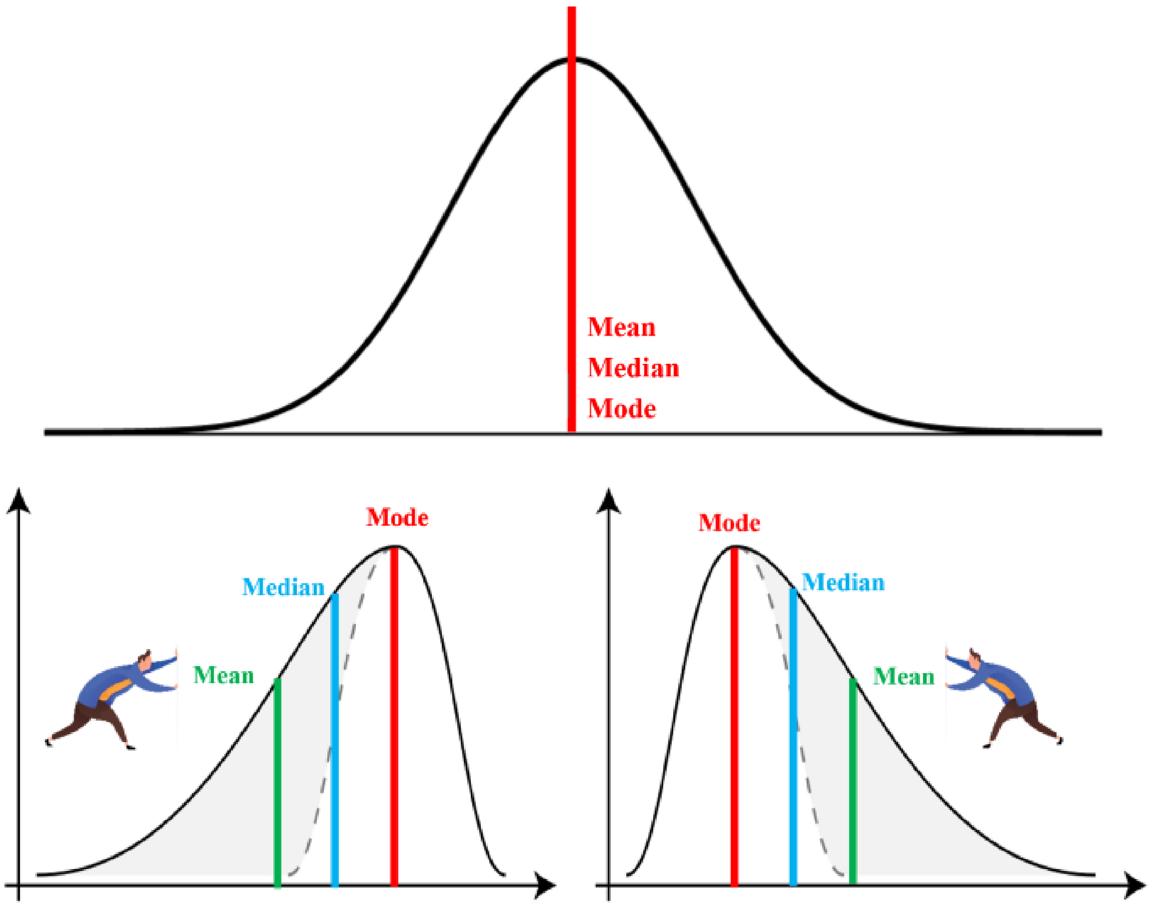
Together for Tomorrow! Enabling People
Education for Future Generations

Descriptive Statistics (1/10)

Skewness refers the lack of symmetry and kurtosis refers the peakedness of a distribution.

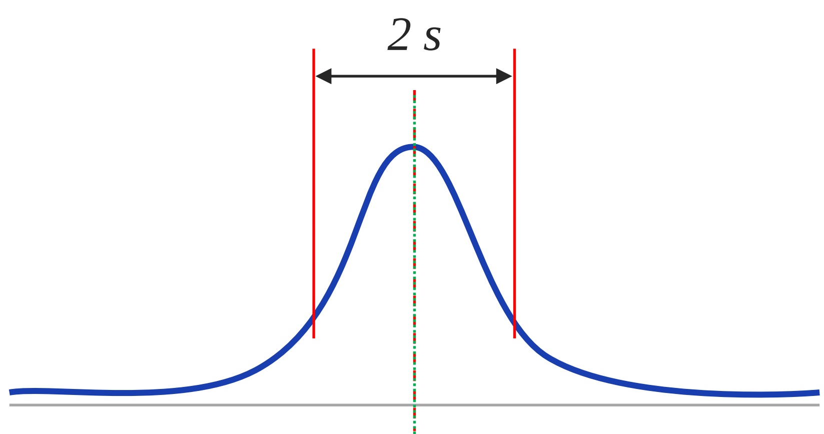


Descriptive Statistics (1/10)



Descriptive Statistics (2/10)

- Sample statistics:



$$\bar{x} = m$$

$$\varsigma = 0$$

Symmetric

Together for Tomorrow! Enabling People
Education for Future Generations

Descriptive Statistics (2/10)

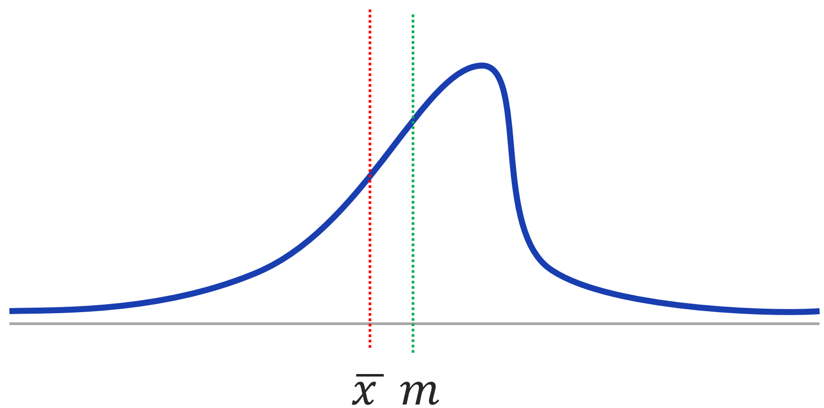
$$\text{Skewness} = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}}$$

$$\text{Mode} = 3(\text{Median}) - 2(\text{Mean})$$

$$\text{Skewness} = \frac{3(\text{Mean} - \text{Median})}{\text{Standard Deviation}}$$

Descriptive Statistics (3/10)

- Sample statistics:



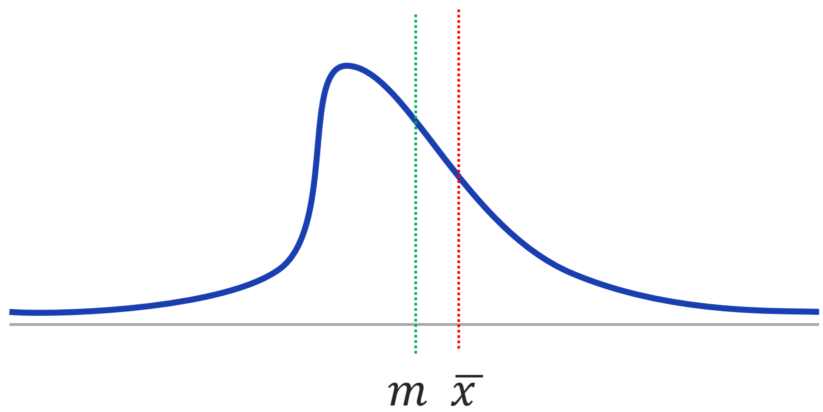
$$\varsigma < 0$$

Asymmetric

Together for Tomorrow! Enabling People
Education for Future Generations

Descriptive Statistics (4/10)

- Sample statistics:



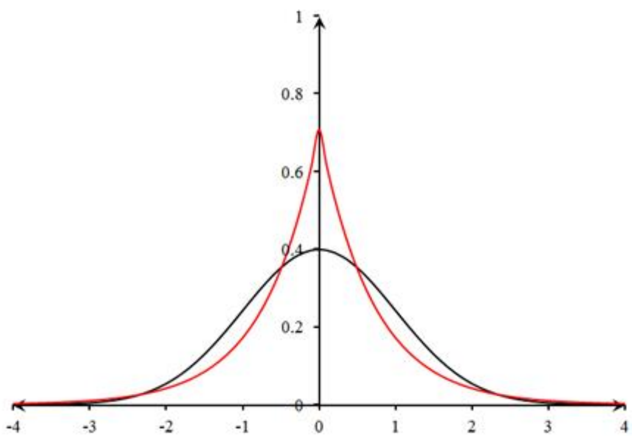
$$\varsigma > 0$$

Asymmetric

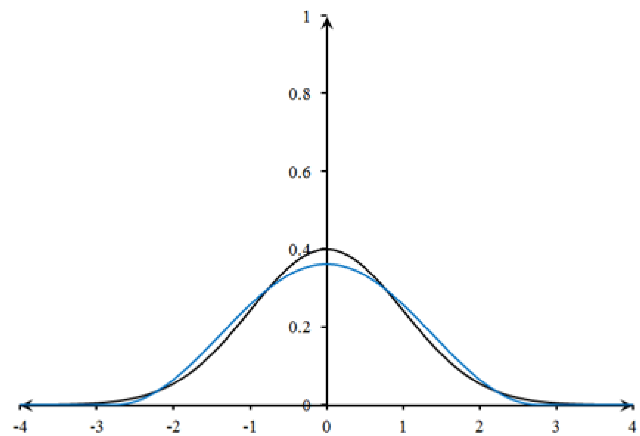
Together for Tomorrow! Enabling People
Education for Future Generations

Descriptive Statistics (5/10)

- Sample statistics:



$\kappa > 0$



$\kappa < 0$

Descriptive Statistics (6/10)

Sample statistics:

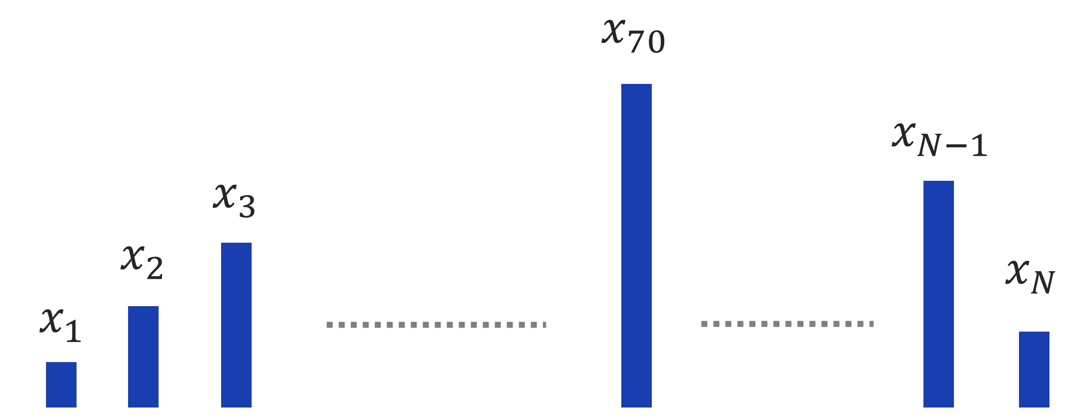
- Sample variance: $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
- Sample covariance: $s_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$
- Sample correlation: $r = \frac{s_{XY}}{s_X s_Y}$

$$\sigma^2 = \frac{\Sigma (x - \mu)^2}{N}$$ Population Variance

$$s^2 = \frac{\Sigma (x - \bar{x})^2}{n - 1}$$ Sample Variance

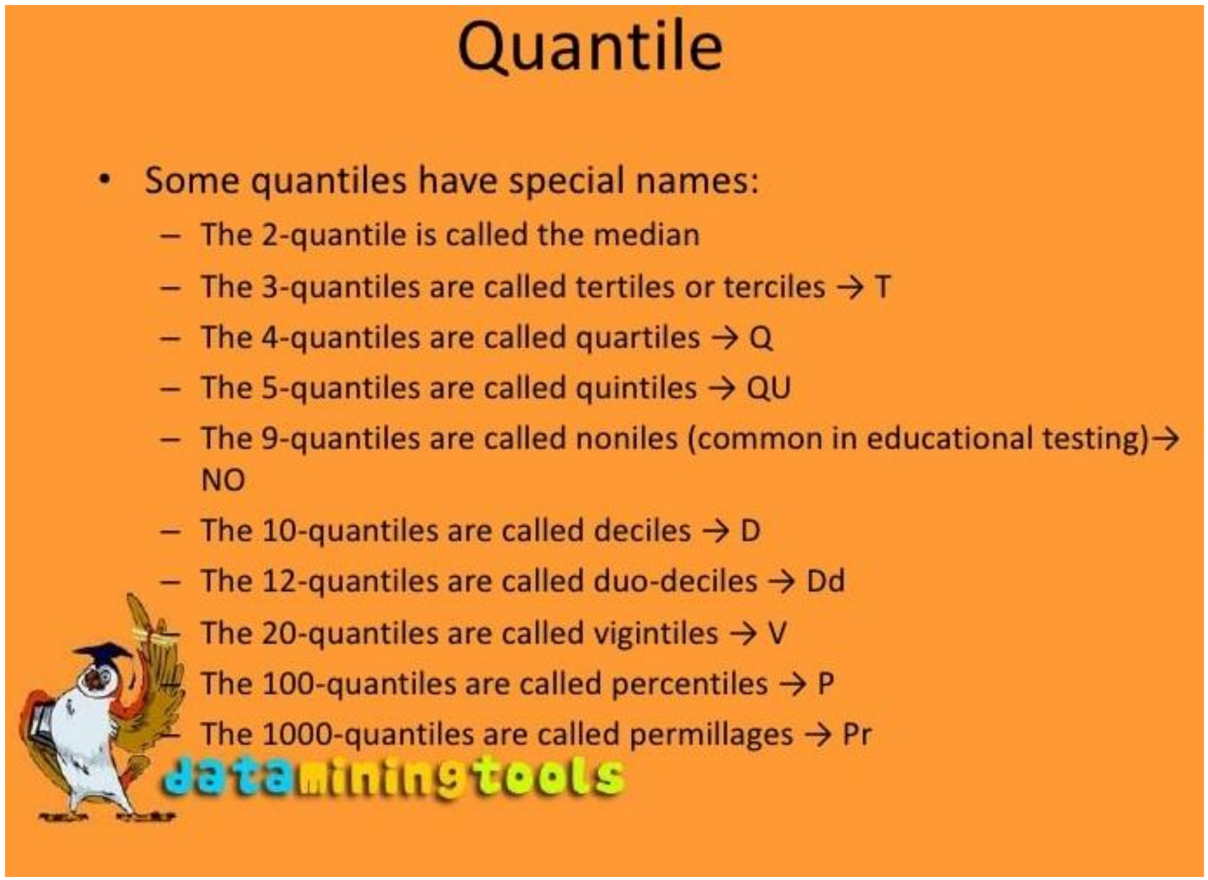
Descriptive Statistics (7/10)

- Sample statistics: quantile
  - Let's suppose that you have a sample of values $x_1, x_2, x_3, \cdots, x_N$. You can represent them as bars.



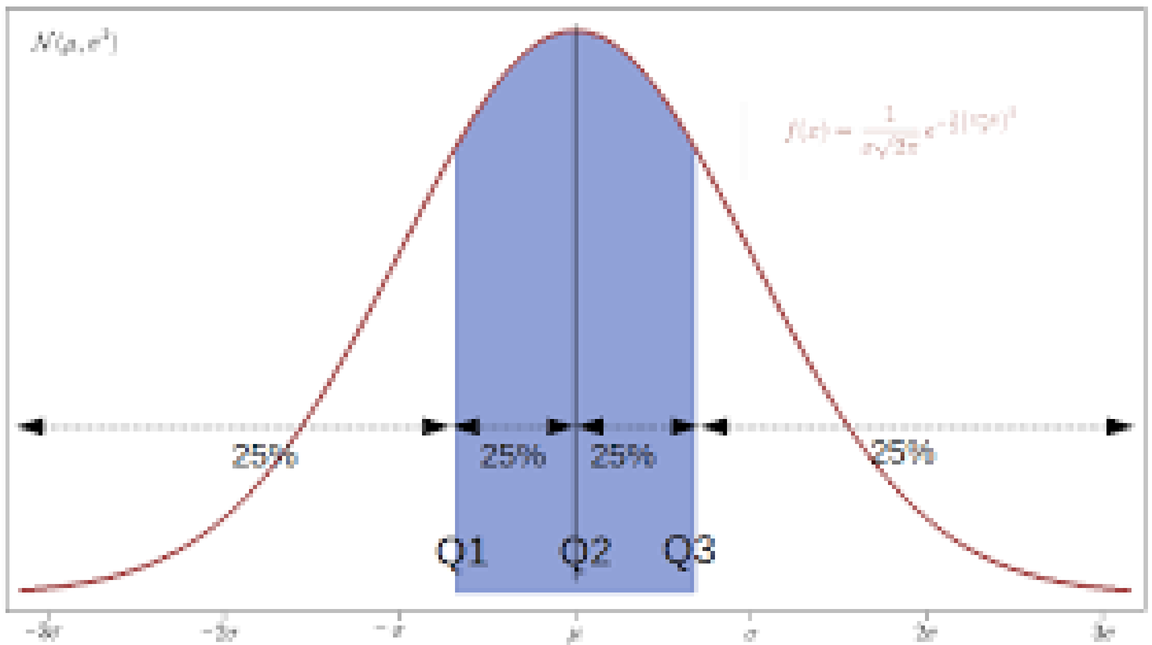
Descriptive Statistics (7/10)

Quantile



Descriptive Statistics (7/10)

Quantile Vs Quartile



Descriptive Statistics (7/10)

Quantile Vs Quartile

Example 1

0 1 5 6 7 8 9 10 12 12 13 14 16 19 19

- Find the Q_1, Q_2, and Q_3 of the following set of data.

19	**12**	**16**	**0**	**14**
9	**6**	**1**	**12**	**13**
10	**19**	**7**	**5**	**8**

$$Q_i = \left(\frac{i \cdot (n+1)}{4}\right)^{th}$$

$$Q_3 = \left(\frac{3 \cdot (15+1)}{4}\right)^{th}$$

$$Q_3 = \left(\frac{48}{4}\right)^{th}$$

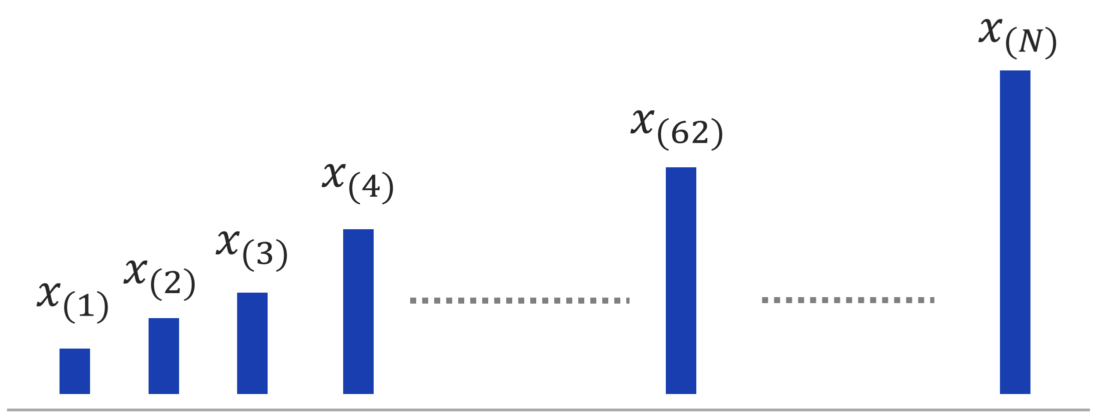
$$Q_3 = 12^{th}$$

$$\boxed{Q_3 = 14}$$

Descriptive Statistics (8/10)

Sample statistics: quantile

- Now, let us sort the values from the smallest to the largest and get $x_{(1)}, x_{(2)}, x_{(3)}, \cdots, x_{(N)}$.



- α *quantile*= value at the position $int(N \times \alpha)$

Descriptive Statistics (9/10)

Sample statistics: quantile

- Quantile: α is probability $0 \sim 1$.
- Percentile: α is given as percentage $0\% \sim 100\%$.

  → Minimum = 0% percentile.

  → Maximum = 100% percentile.

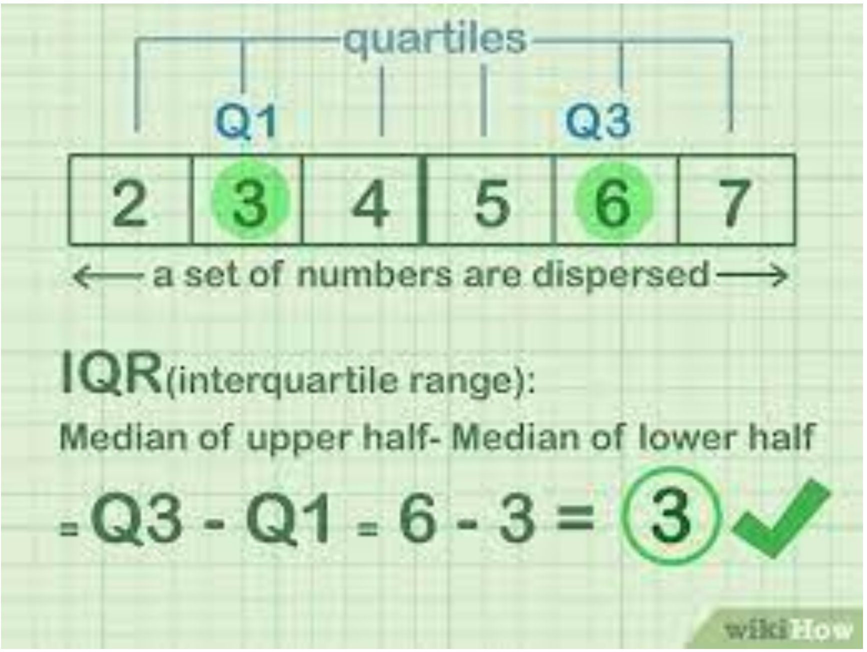
  $\text{Range} = \text{Maximum} - \text{Minimum}$
- Quartile: α subdivided into four equal intervals.

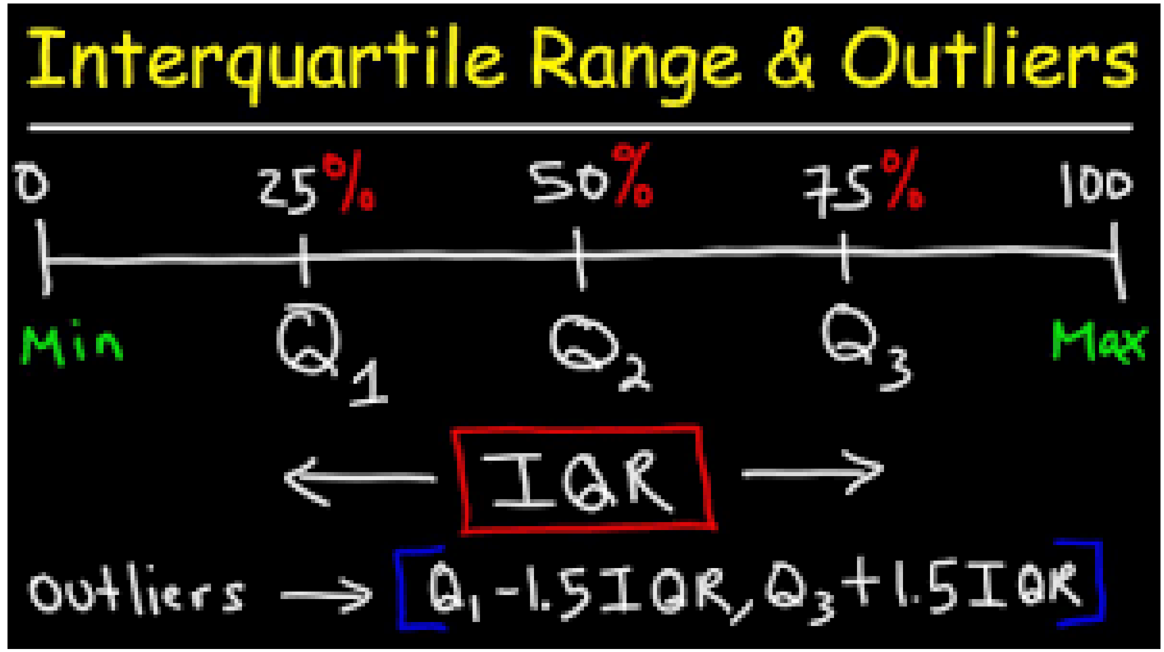
  → 1st quartile (Q1): 25% percentile.

  → 2nd quartile (Q2): 50% percentile = Median.

  → 3rd quartile (Q3): 75% percentile.

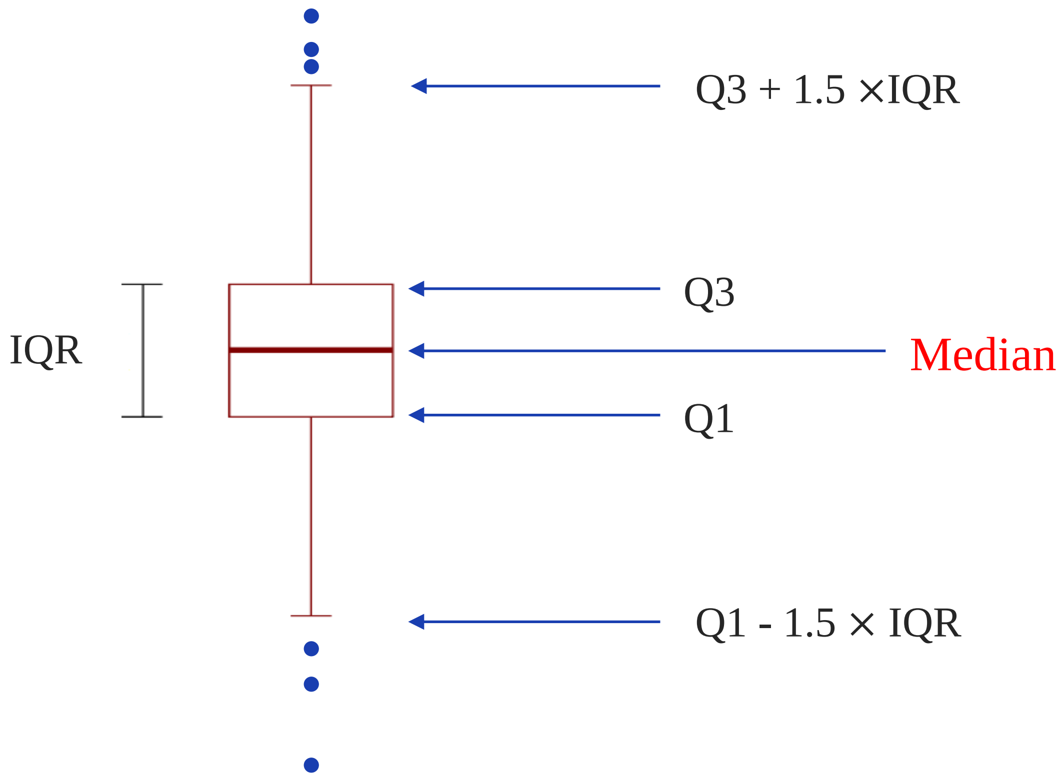
Descriptive Statistics (10/10)





Descriptive Statistics (10/10)

- Boxplot:



Coding Exercise #0303

Follow practice steps on 'ex_0303.ipynb' file.

Chapter 4.

Probability and Statistics

Together for Tomorrow! Enabling People

Education for Future Generations

Inferential Statistics (1/3)

Population vs Sample:

- You analyze samples because the population is hard or impossible to reach.
- You would like to draw conclusions about the population using its samples.
- However, you should carefully analyze the uncertainties associated with the samples.

Population parameters vs Sample statistics:

- The population parameters are properties of a given population.
- There is one population: and the population parameters are calculated once for each.
- Sample statistics are properties of a given sample.
- In principle, you can draw several samples and calculate the same sample statistics many times.

Inferential Statistics (2/3)

Sampling method:

- Simple random sampling: values drawn with equal probability with or without replacement.
- Weighted random sampling: values drawn with varying probabilities.
- Stratified sampling: reflects the proportions of the strata.
- Systematic sampling: values are drawn with an implicit periodicity.
- Cluster sampling: takes a representative cluster.

Inferential Statistics (3/3)

Population parameters vs Sample statistics:

	Population	**Sample**
Size	N	n
Mean	$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
Variance	$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$	$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
Standard Deviation	$\sigma = \sqrt{\sigma^2}$	$s = \sqrt{s^2}$

- Connect the sample statistics with the population parameters through statistical inference.

Together for Tomorrow! Enabling People
Education for Future Generations

Central Limit Theorem (1/10)

Coin flipping experiment:

- Let us flip the coin twice and group the results as a sample (n=2). We assign T = 0 and H =1.
- As we draw several samples, we get the corresponding sample means: $\overline{x_1}$, $\overline{x_2}$, $\overline{x_3}$,...

Ex)

i	Sample	$\overline{x}_i$
1	1,1	1
2	0,1	0.5
3	1,0	0.5
4	0,0	0
⋮	⋮	⋮

Central Limit Theorem (2/10)

Coin flipping experiment:

- Let us flip the coin three times and group the results as a sample (n=3).
- As we draw several samples, we get the corresponding sample means: $\overline{x_1}, \overline{x_2}, \overline{x_3}$,...

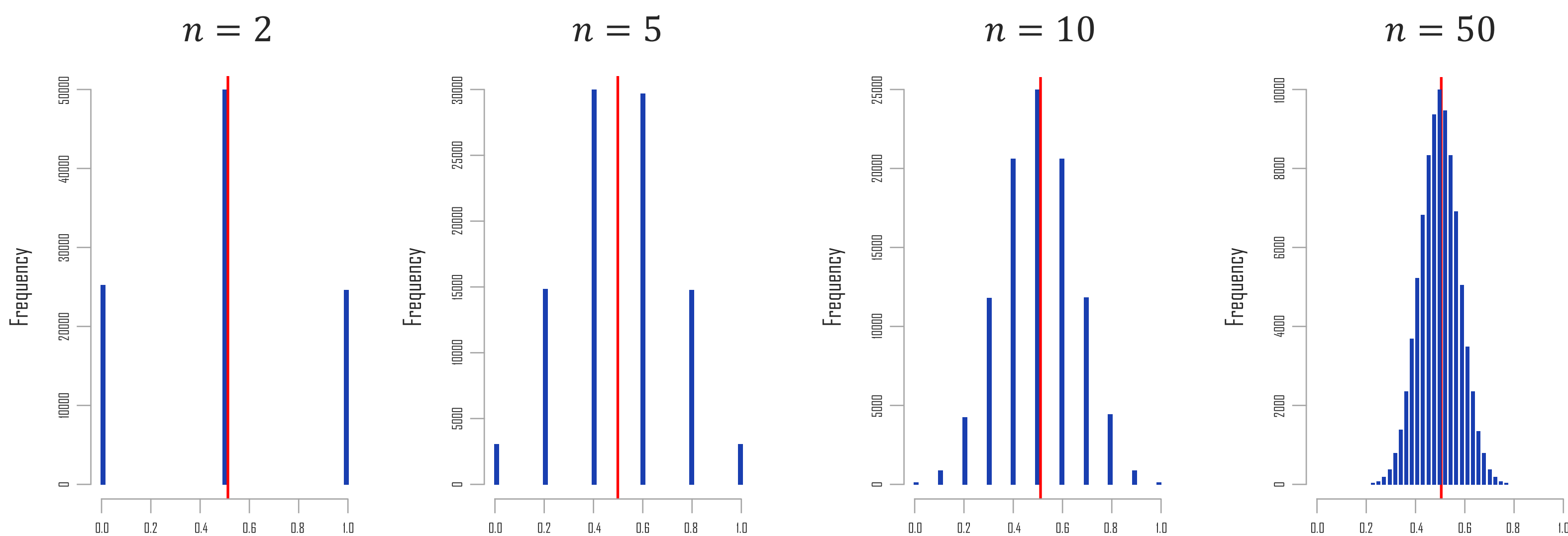
Ex)

i	Sample	$\overline{x}_i$
1	1,0,1	2/3
2	0,1,0	1/3
3	1,0,0	1/3
4	0,0,0	0
⋮	⋮	⋮

Together for Tomorrow! Enabling People
Education for Future Generations

Central Limit Theorem (3/10)

Coin flipping experiment:



- Histograms of the sample means of varying sample size.
- As the sample size increases, the histograms converge to the normal distribution.

Together for Tomorrow! Enabling People
Education for Future Generations

Central Limit Theorem (4/10)

Coin flipping experiment:

- The sample means $\overline{x}$ are also randomly distributed.
- You can denote the sample mean by $\overline{\boldsymbol{X}}$ (upper case) and treat it as a random variable.
- Then, we obtain the following results.

a) Expected value of the $\overline{X}$ is the population mean : $E[\overline{X}]=\mu$

b) Variance and standard deviation of the $\overline{X}$:

$$Var(\overline{X}) = \sigma_{\overline{X}}^2 = \frac{\sigma^2}{n} = \frac{p(1-p)}{n} = \frac{0.25}{n}$$

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{p(1-p)}}{\sqrt{n}} = \frac{0.5}{\sqrt{n}}$$

"**Standard error**"

- σ^2= population variance, s^2= sample variance, $\sigma_{\overline{X}}^2$= variance of the sample mean.

Together for Tomorrow! Enabling People
Education for Future Generations

Central Limit Theorem (5/10)

Central limit theorem:

- The distribution of the sample means converges to the Normal distribution as n increases.

  1) This is true not just for the coin flipping (Binomial).

  2) You can verify the same phenomena for other underlying distributions (discrete and continuous).

- For large enough n, analogous tendency can be verified for the sample statistics beside the mean.

Together for Tomorrow! Enabling People
Education for Future Generations

Central Limit Theorem (6/10)

Realistic considerations:

- In reality, you will not have several samples but **one**.
- You apply the central limit theorem and assume that the sample distribution is Normal.
- For the statistical inference, you will take advantage of the properties of the Normal distribution.

Together for Tomorrow! Enabling People
Education for Future Generations

Central Limit Theorem (7/10)

Standardization:

- When the sample size n is big enough $\overline{X}$ follows the Normal distribution $N(\mu,\sigma$^2).
- You can standardize $\overline{X}$ and get

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

which follows the Standard Normal distribution N(0,1).

- When the sample size n is small or the σ is not known, you define

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

which follows the Student-t probability density with degree of freedom n−1.

Together for Tomorrow! Enabling People
Education for Future Generations

Central Limit Theorem (8/10)

Sampling distribution of proportions:

- Let us consider the case where the population follows Bernoulli distribution.
- Let us suppose that the success probability is p.
- We denote the sample proportion (of success) as $\widehat{P}$ and have

  a) Expected value: $E\left[\widehat{P}\right] = p$

  b) Standard error: $\sigma_{\widehat{P}} = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{p(1-p)}{n}}$

- When $n\,p$ > 10 and $n(1-p)$ > 10 then, we can apply the central limit theorem: $\frac{\hat{P}-p}{\sigma_{\hat{P}}} \sim N(0,1)$

Central Limit Theorem (9/10)

Sampling distribution of the difference:

- Let us consider two populations. Samples of size n_1 and n_2 are drawn from each population.
- Then for the difference of the sample means, you have

  a) Expected value: $E[\overline{X_1} - \overline{X_2}] = E[\overline{X_1}] - E[\overline{X_2}] = \mu_1 - \mu_2$

  b) Standard error: $\sigma_{\overline{X_1}-\overline{X_2}} = \sqrt{\sigma_{\overline{X_1}}^2 + \sigma_{\overline{X_2}}^2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

- When n_1 and n_2 are large enough, you can apply the central limit theorem: $\frac{\overline{X_1}-\overline{X_2}-(\mu_1-\mu_2)}{\sigma_{\overline{X_1}-\overline{X_2}}} \sim N(0,1)$

Central Limit Theorem (10/10)

Standard errors:

Statistic	Standard Error	Explanation
Mean	$\frac{\sigma}{\sqrt{n}}$	When $n \geq 30$, the sample mean approximately follows the Normal distribution.
Proportion	$\sqrt{\frac{p(1-p)}{n}}$	When $n\,p > 10$ and $n(1-p) > 10$, the sample proportion approximately follows the Normal distribution.
Median	$\sigma\sqrt{\frac{\pi}{2n}}$	When $n \geq 30$, the sample mean approximate follows the Normal distribution.
Standard Deviation	a) $\frac{\sigma}{\sqrt{2n}}$ b) $\sqrt{\frac{\mu_4-\sigma^4}{4n\sigma^2}}$	a) When the population follows the Normal distribution. b) When the population does not follow the Normal distribution.
Variance	a) $\sigma^2\sqrt{\frac{2}{n}}$ b) $\sqrt{\frac{\mu_4-\sigma^2}{n}}$	a) When the population follows the Normal distribution. b) When the population does not follow the Normal distribution. The sample variance follows the Chi-square distribution.
Correlation	$\sqrt{\frac{1-r^2}{n-2}}$	r is the sample correlation. Fisher's z-transformation required.

Chapter 4.

Probability and Statistics

Together for Tomorrow! Enabling People

Education for Future Generations

Together for Tomorrow! Enabling People
Education for Future Generations

Estimation Theory

Point estimation vs Interval estimation:

Ex) Say you want to survey the mean weight of adult males. The possible answers are:

a) 72 kg

b) 71 kg ~ 74 kg

c) 70 kg ~ 75 kg

- There many ways of answering to the same question.
- Providing a single value like in a) is called "point estimation". ⇐ using an estimator.
- Providing an interval like in b) and c) is called "interval estimation".

Estimator (1/4)

What makes a "good" estimator?

1) Unbiasedness.

2) Efficiency.

3) Consistency.

Estimator (2/4)

Unbiasedness:

- Let us suppose that θ is the population parameter and θ ˆis the corresponding estimator.
- A unbiased estimator satisfies the following condition: $E\left[\hat{\theta}\right] = \theta$

Ex) $\frac{\sum_{i=1}^{n} x_i}{n}$ is an unbiased estimator of the population mean μ.

Ex) $\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\boldsymbol{n-1}}$ is an unbiased estimator of the population variance σ^2.

Estimator (3/4)

Efficiency:

- Let us suppose that there are several unbiased estimators $\hat{\theta}_1, \hat{\theta}_2, \ldots$

  $E[\hat{\theta}_1] = \theta \;\; , \; E[\hat{\theta}_2] = \theta \;\; , \; E[\hat{\theta}_3] = \theta \;\; , \ldots$

- The estimator with the least uncertainty (variance or standard deviation) is the efficient estimator:

  Ex) If $Var(\hat{\theta}_1)$ is at the minimum, then $\hat{\theta}_1$ is the efficient estimator.

Together for Tomorrow! Enabling People
Education for Future Generations

Estimator (4/4)

Consistency:

- A biased estimator for small n, may become unbiased for large enough n because of the consistency.

Ex) $\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}$ is biased estimator of σ^2 for small n but becomes unbiased for large enough n.

$$\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n} \cong \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$$

- When the sample size n can be increased, the consistency becomes an important criteria.

Interval Estimation of the Mean (1/11)

Confidence interval:

- It is the interval that may contain the true value of a population parameter with a given confidence level.

  a) Significance level: α

  b) Confidence level: $(1-\alpha)$ $\Leftarrow$ A probability.

Together for Tomorrow! Enabling People
Education for Future Generations

Interval Estimation of the Mean (2/11)

Confidence interval of the mean:

- Let us suppose that the sample size is big enough that the sample mean is distributed normally.
- The 95% confidence interval can be obtained in the following way.

$$P(-1.96 < Z < 1.96) = 0.95$$

$$P\left(-1.96 < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < 1.96\right) = 0.95$$

$$P\left(-1.96\,{}^{\sigma}\!/_{\sqrt{n}} < \bar{X}-\mu < 1.96\,{}^{\sigma}\!/_{\sqrt{n}}\right) = 0.95$$

Interval Estimation of the Mean (3/11)

Confidence interval of the mean:

- Let us suppose that the sample size is big enough that the sample mean is distributed normally.
- The 95% confidence interval can be obtained in the following way.

$$P\left(-1.96\,{}^{\sigma}\!/_{\sqrt{n}} < \bar{X} - \mu < 1.96\,{}^{\sigma}\!/_{\sqrt{n}}\right) = 0.95$$

$$P\left(-\bar{X} - 1.96\,{}^{\sigma}\!/_{\sqrt{n}} < -\mu < -\bar{X} + 1.96\,{}^{\sigma}\!/_{\sqrt{n}}\right) = 0.95$$

$$P\left(\bar{X} - 1.96\,{}^{\sigma}\!/_{\sqrt{n}} \le \mu \le \bar{X} + 1.96\,{}^{\sigma}\!/_{\sqrt{n}}\right) = 0.95$$

Interval Estimation of the Mean (4/11)

Confidence interval of the mean:

- The 95% confidence interval can be obtained in the following way.

Lower bound: $\bar{X} - 1.96\,{}^{\sigma}/_{\sqrt{n}}$ Upper bound: $\bar{X} + 1.96\,{}^{\sigma}/_{\sqrt{n}}$

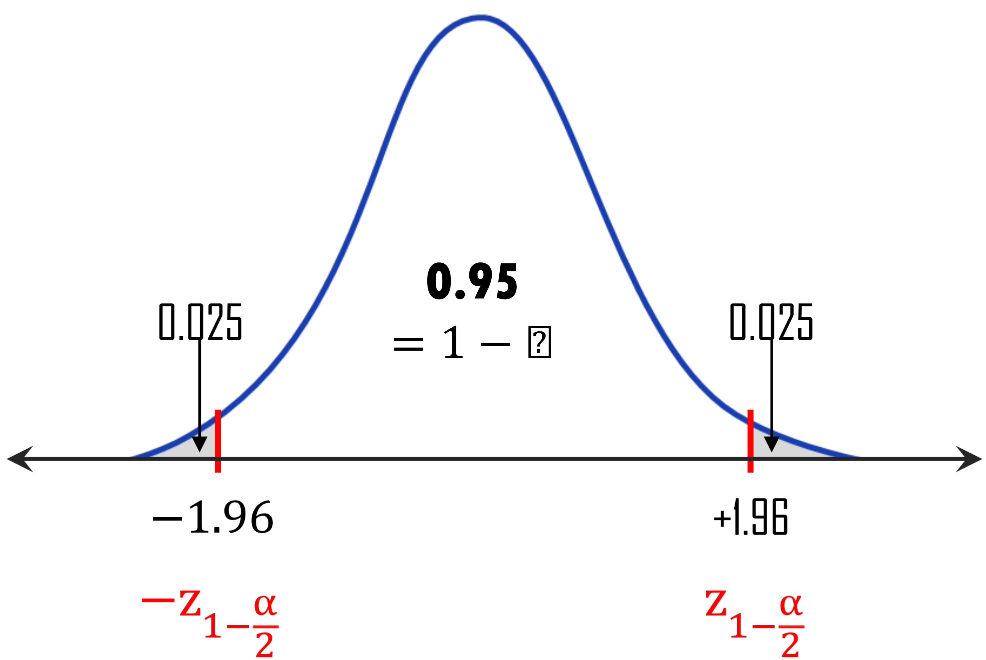
[← Confidence Interval →]

- But, where did the number 1.96 come from?
- 1.96 is the 0.975 quantile of the normal distribution ⇔ the position where the CDF = 0.975.

Together for Tomorrow! Enabling People
Education for Future Generations

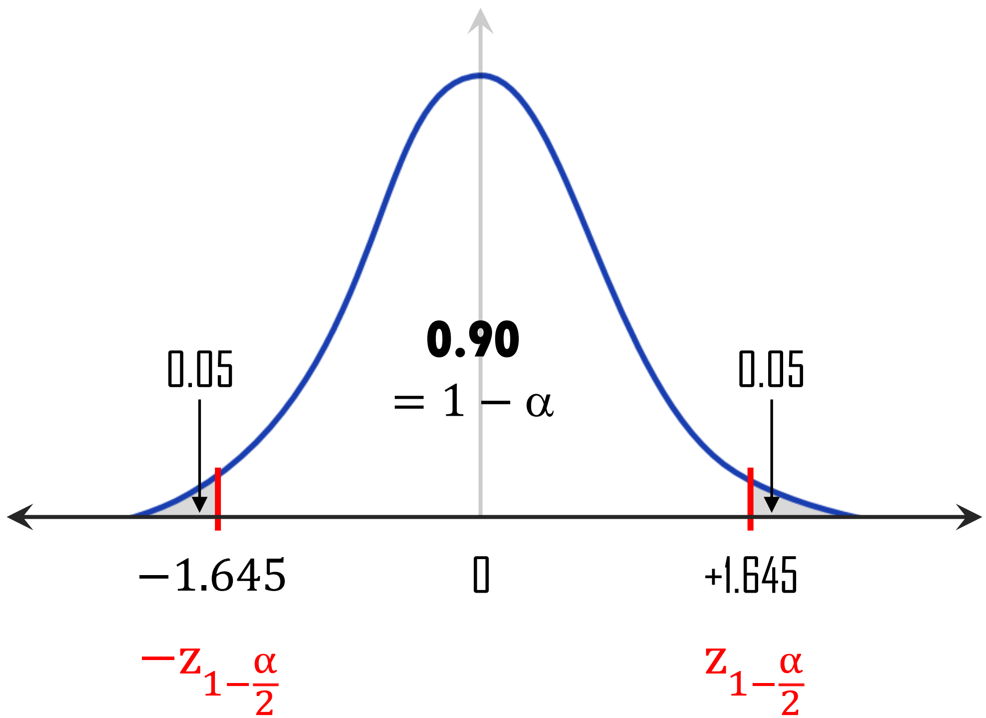
Interval Estimation of the Mean (5/11)

- Confidence interval of the mean: 95%



Interval Estimation of the Mean (6/11)

- Confidence interval of the mean: 90%



Interval Estimation of the Mean (7/11)

Confidence interval of the mean:

- Of arbitrary confidence level = (1−α).

Lower bound: $\bar{X} - z_{1-\frac{\alpha}{2}} \sigma/\sqrt{n}$ Upper bound: $\bar{X} + z_{1-\frac{\alpha}{2}} \sigma/\sqrt{n}$

[← Confidence Interval →]

- When the sample size is small and the population variance is not known, use Student-t.

Lower bound: $\bar{X} - t_{1-\frac{\alpha}{2}} s/\sqrt{n}$ Upper bound: $\bar{X} + t_{1-\frac{\alpha}{2}} s/\sqrt{n}$

[← Confidence Interval →]

Interval Estimation of the Mean (8/11)

Confidence interval of the mean:

Question:

Isn't it always good to have as large a confidence level as possible?

Interval Estimation of the Mean (9/11)

Confidence interval of the mean:

Answer:

All the other conditions being equal, larger confidence level means larger uncertainty.

The confidence interval broadens.

Interval Estimation of the Mean (10/11)

- Confidence interval of the mean:

[99% confidence interval]

[95% confidence interval]

[90% confidence interval]

Interval Estimation of the Mean (11/11)

How to narrow the confidence interval of the mean while keeping the confidence level high?

- Let us remember that the bounds of the confidence interval was given by:

$$\bar{X} \pm z_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} \quad \text{or} \quad \bar{X} \pm t_{1-\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}}$$

- Smaller σ or s can narrow the confidence interval $\Leftarrow$ **you cannot** control it!
- Larger sample size n can narrow the confidence interval $\Leftarrow$ in principle **you can** control it!
- If W is the targeted half width of the confidence interval, apply the formula:

$$z_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} = W \quad \Rightarrow \quad \boldsymbol{n} = \left[\frac{z_{1-\alpha/2} \times \sigma}{W}\right]^2$$

Coding Exercise #0304

Follow practice steps on 'ex_0304.ipynb' file.

Coding Exercise #0305

Follow practice steps on 'ex_0305.ipynb' file.

Together for Tomorrow! Enabling People
Education for Future Generations

Interval Estimation of the Correlation (1/3)

Confidence interval of the correlation:

- There are different kinds of correlation: Pearson, Spearman, Kendall, etc.
- Spearman and Kendall are the so-called "rank correlations".
- Let us focus on the "usual type", that is, Pearson correlation.

Interval Estimation of the Correlation (2/3)

Confidence interval of the correlation:

- The distribution of the sample correlation r deviates significantly from the Normal distribution.
- This is true even for large enough sample size, because the correlation is bound to be $[-1, +1]$.
- To overcome these "distortions", it is necessary to do Fisher's z-transformation.

a) Transformed correlation: z=$arctanh(r)$

b) Standard error after the transformation: $\sigma_z = \frac{1}{\sqrt{n-3}}$

c) Inverse transformation: r= $tanh(z)$

Interval Estimation of the Correlation (3/3)

Confidence interval of the correlation:

- Of arbitrary confidence level = (1−α).

$$tanh\left(z - z_{1-\frac{\alpha}{2}} \times \sigma_z\right) \qquad\qquad tanh\left(z + z_{1-\frac{\alpha}{2}} \times \sigma_z\right)$$

[← Confidence Interval →]

Ex) 95% confidence interval: $[tanh(z - 1.96\ \sigma_z), tanh(z + 1.96\ \sigma_z)]$

Coding Exercise #0306

Follow practice steps on 'ex_0306.ipynb' file.