Statistics Review Session

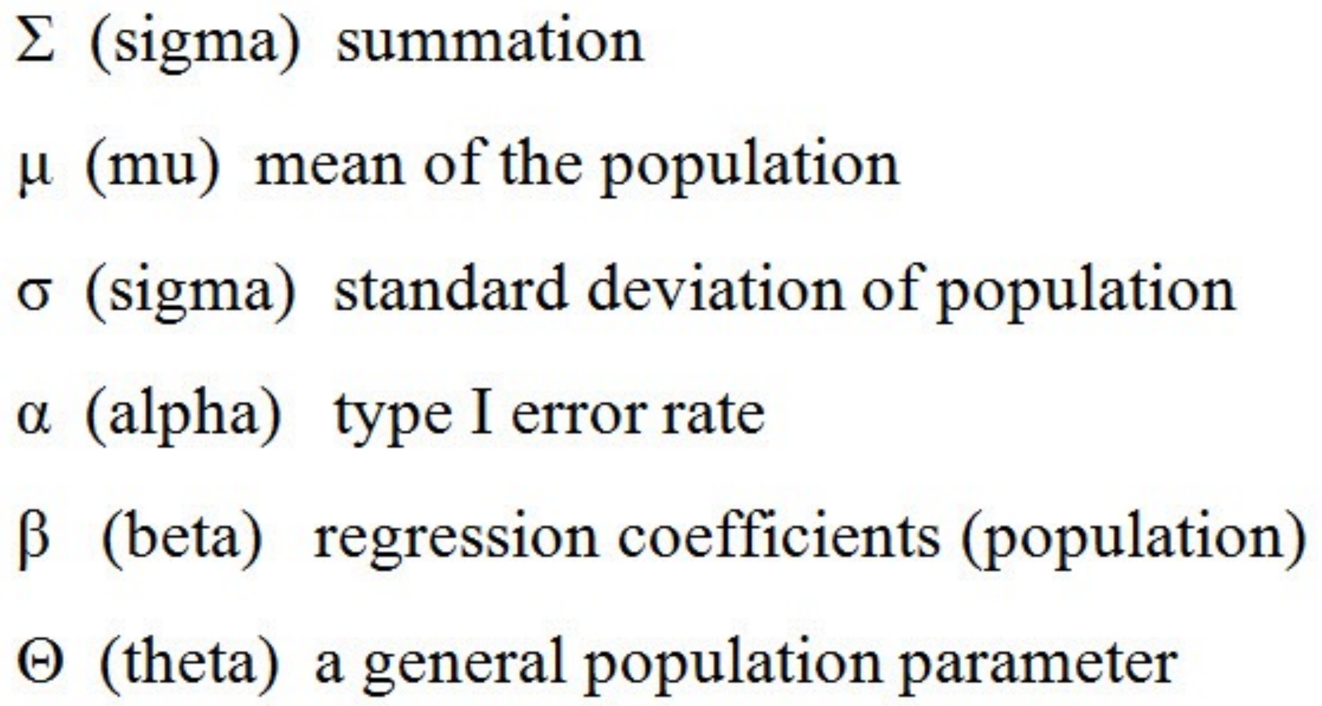
Miao Hu
miao.hu@duke.edu

- This is designed only as a review of basic statistics. I assume that you have already been exposed to this material and simply need a quick refresher.

Content

- 1. Rounding and Significant Digits
- 2. Calculating Summary Statistics
- 3. Exponent Rules
- 4. Logarithm Rules
- 5. Factorials
- 6. Descriptive Statistics
- 7. Probability
- 8. The normal distribution
- 9. T-test
- 10. Standard Error vs. Standard Deviation
- 11. Central Limit Theorem
- 12. Calculating Z-Scores
- 13. Reading and using a Z-table
- 14. Confidence Intervals
- 15. Inference (Comparison of Means)

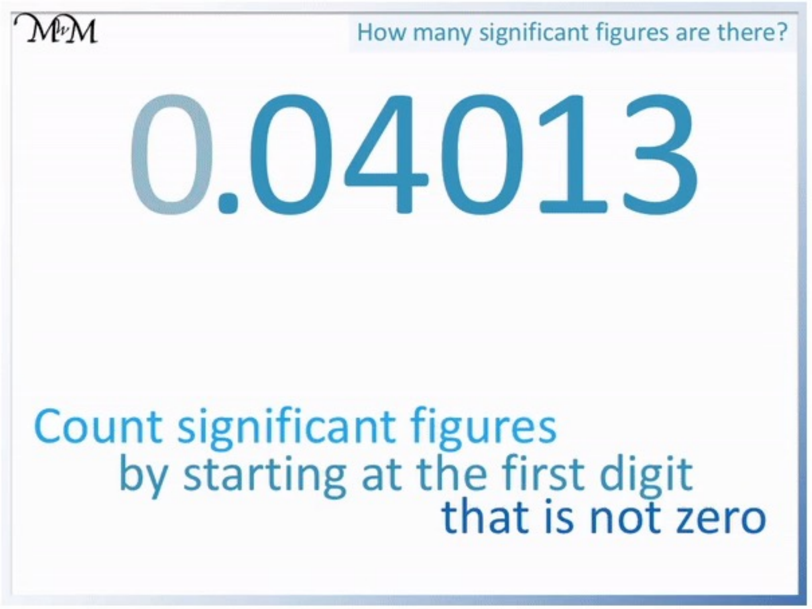
Basic Math Symbols



Σ	(sigma)	summation
μ	(mu)	mean of the population
σ	(sigma)	standard deviation of population
α	(alpha)	type I error rate
β	(beta)	regression coefficients (population)
Θ	(theta)	a general population parameter

Rounding and Significant Digits

- Non-zero digits are always significant


MM

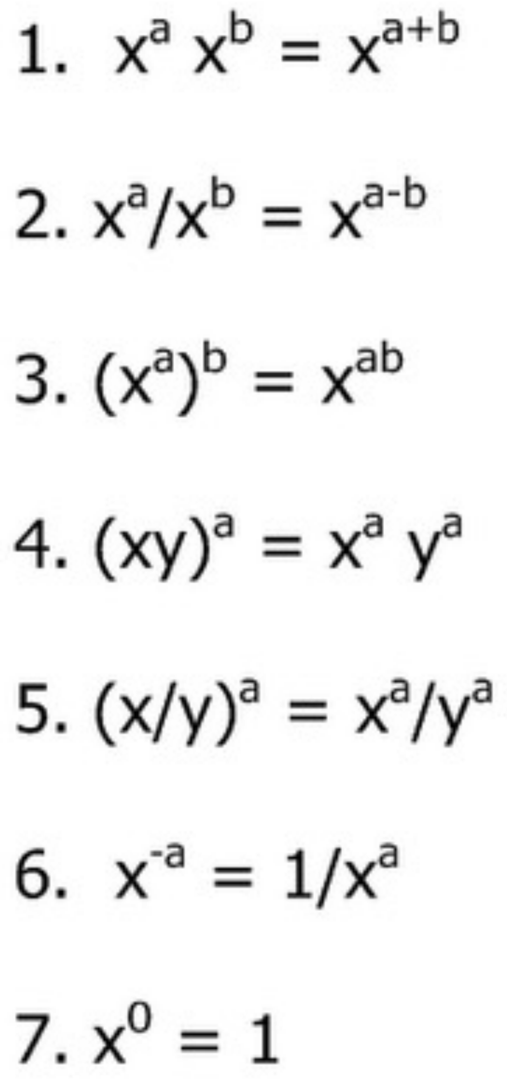
How many significant figures are there?

0.04013

Count significant figures
by starting at the first digit
that is not zero

- One frequently used rule of thumb is to round a mean (or a standard deviation) to one additional decimal than the data.
- Example:
- Dataset: 4.2, 3.8, 4.5, 4.1, 3.9.
- If the calculated mean is 4.1 and the standard deviation is 0.26, you should report:
- Mean= 4.10
- Standard deviation= 0.26

Exponents



1. $x^a x^b = x^{a+b}$

2. $x^a/x^b = x^{a-b}$

3. $(x^a)^b = x^{ab}$

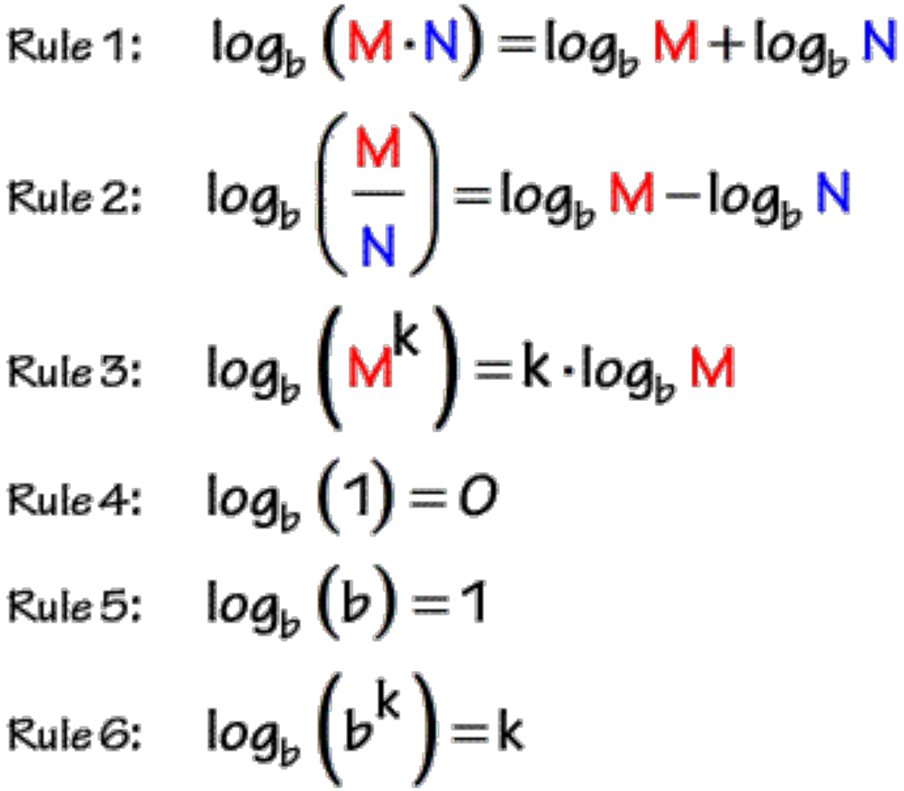
4. $(xy)^a = x^a y^a$

5. $(x/y)^a = x^a/y^a$

6. $x^{-a} = 1/x^a$

7. $x^0 = 1$

Logarithms



Rule 1: $\log_b(M \cdot N) = \log_b M + \log_b N$

Rule 2: $\log_b\left(\frac{M}{N}\right) = \log_b M - \log_b N$

Rule 3: $\log_b\left(M^k\right) = k \cdot \log_b M$

Rule 4: $\log_b(1) = 0$

Rule 5: $\log_b(b) = 1$

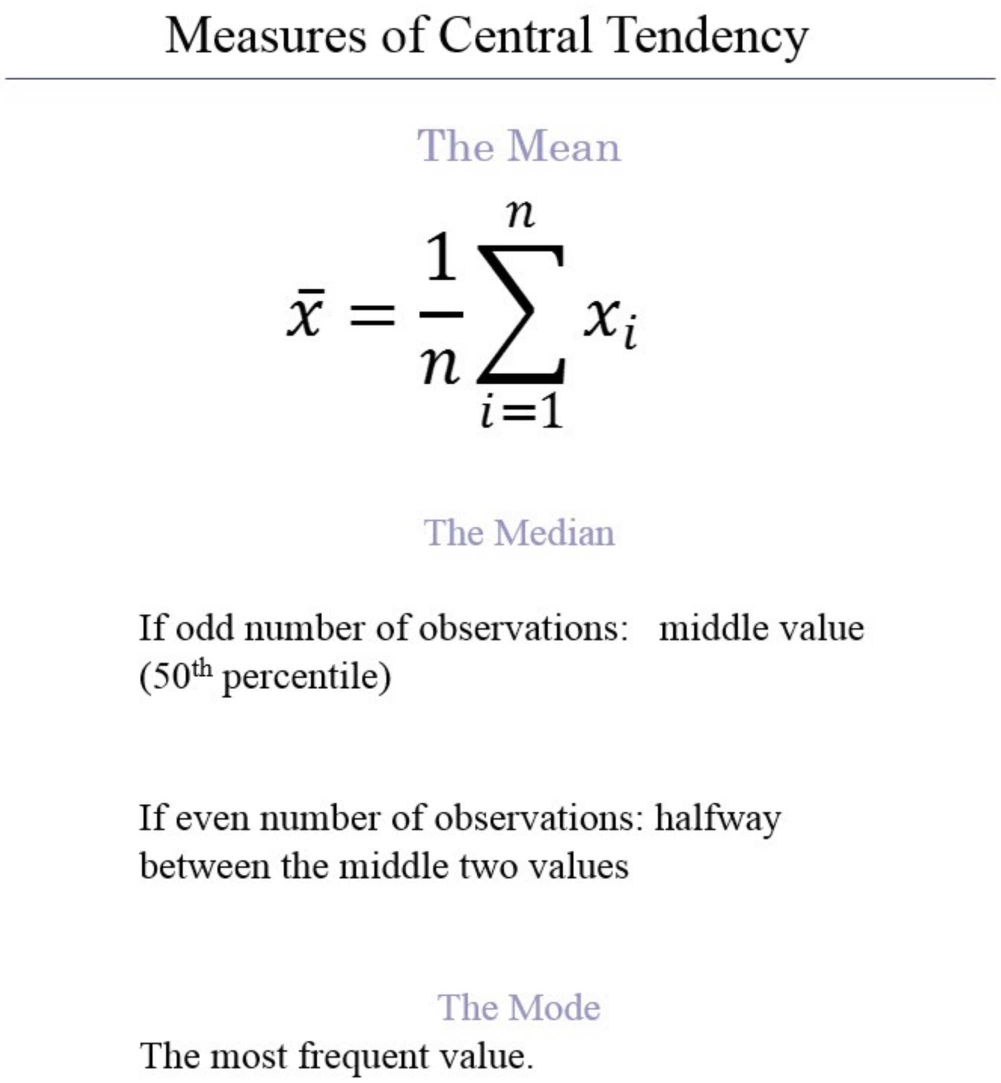
Rule 6: $\log_b\left(b^k\right) = k$

Factorials

- A factorial, represented by an explanation mark (!), denotes a multiplication of the sequence of descending (natural) numbers.

- Example: $\frac{\mathbf{6\,!}}{\mathbf{4!3!}}$

Descriptive Statistics



Measures of Central Tendency

The Mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The Median

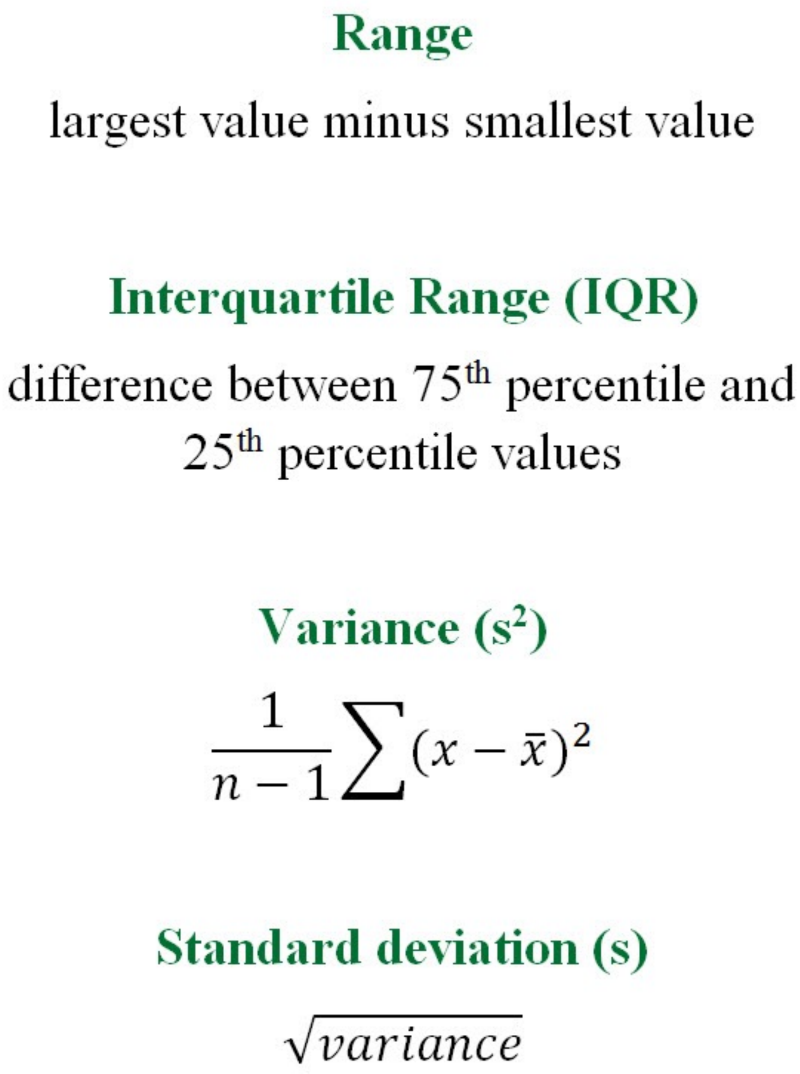
If odd number of observations: middle value (50th percentile)

If even number of observations: halfway between the middle two values

The Mode

The most frequent value.

Descriptive Statistics



Range

largest value minus smallest value

Interquartile Range (IQR)

difference between 75th percentile and 25th percentile values

Variance (s^2)

$$\frac{1}{n-1}\sum(x-\bar{x})^2$$

Standard deviation (s)

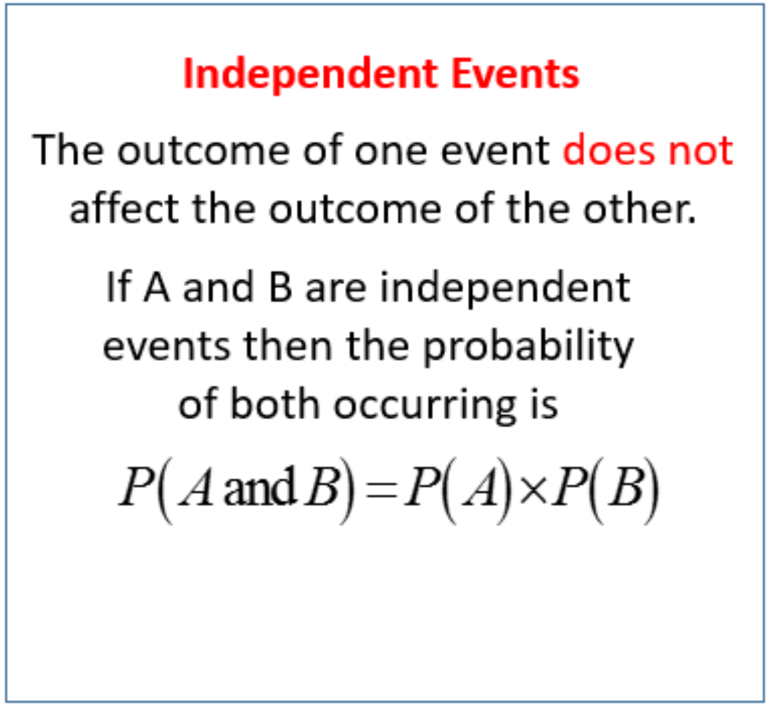
$$\sqrt{variance}$$

Probability

- A probability ranges from 0 to 1 (or 0% to 100%)
- Therefore, the probabilities of all outcomes must sum to 1, e.g.

$$\sum_{i=0}^{n} p(A_i) = 1$$

- The probability of an event plus the probability of its complement must equal 1.

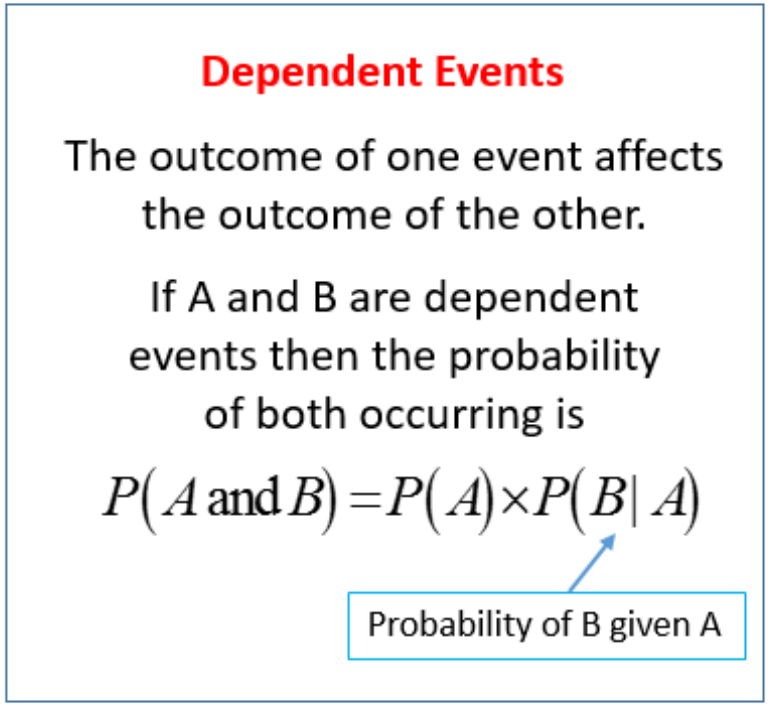


Independent Events

The outcome of one event does not affect the outcome of the other.

If A and B are independent events then the probability of both occurring is

$$P(A \text{ and } B) = P(A) \times P(B)$$



Dependent Events

The outcome of one event affects the outcome of the other.

If A and B are dependent events then the probability of both occurring is

$$P(A \text{ and } B) = P(A) \times P(B| A)$$

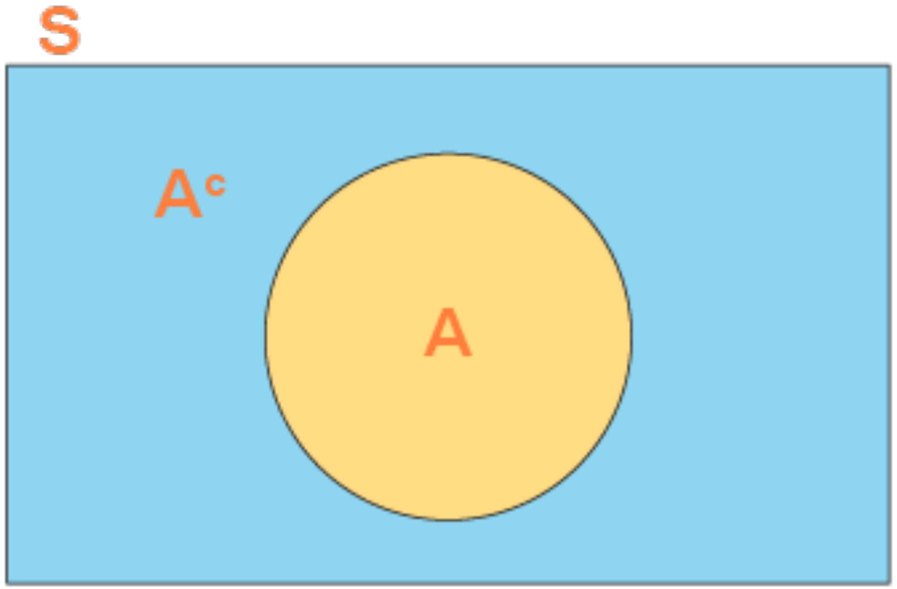
Probability of B given A

Probability

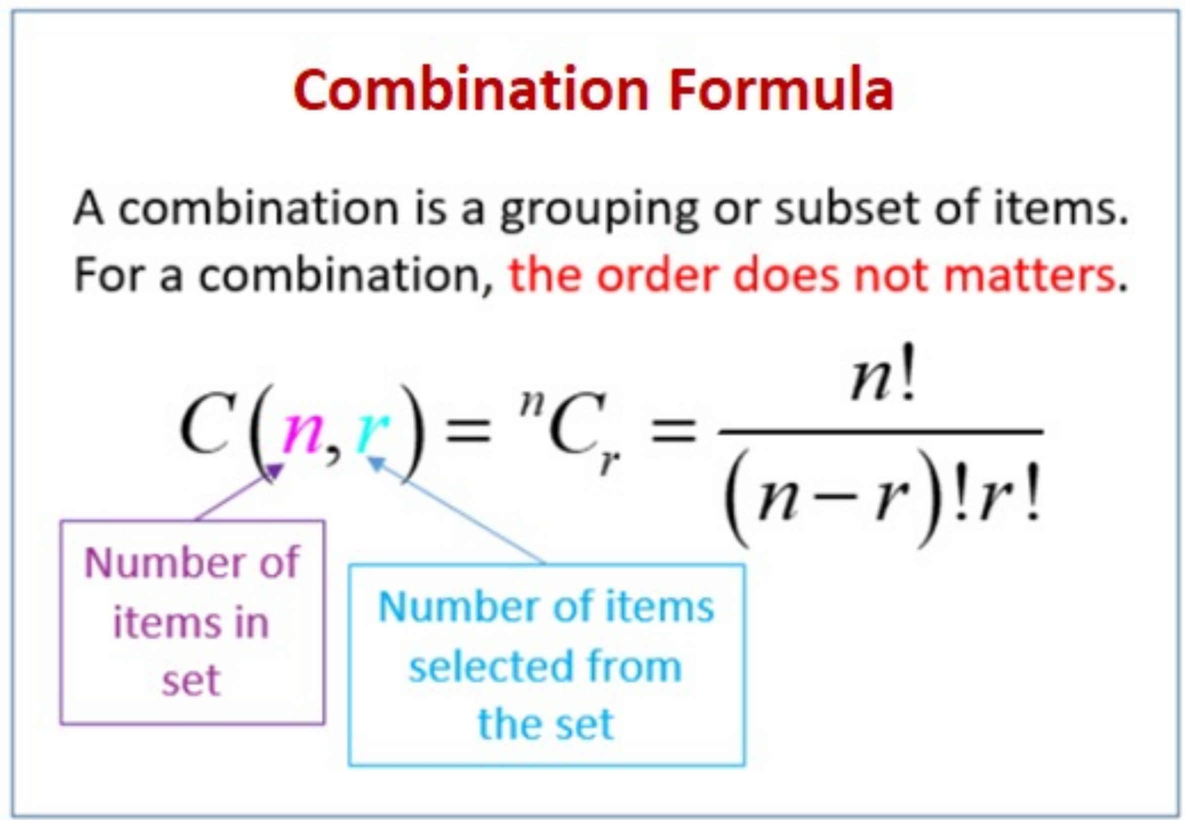
Independent Events & Dependent Events

Probability

- **Complementary Events**: Two events A and A' are complementary if they are mutually exclusive and their union covers all possible outcomes.
- P(A) + P(A') = 1



Probability Example



Combination Formula

A combination is a grouping or subset of items.
For a combination, the order does not matters.

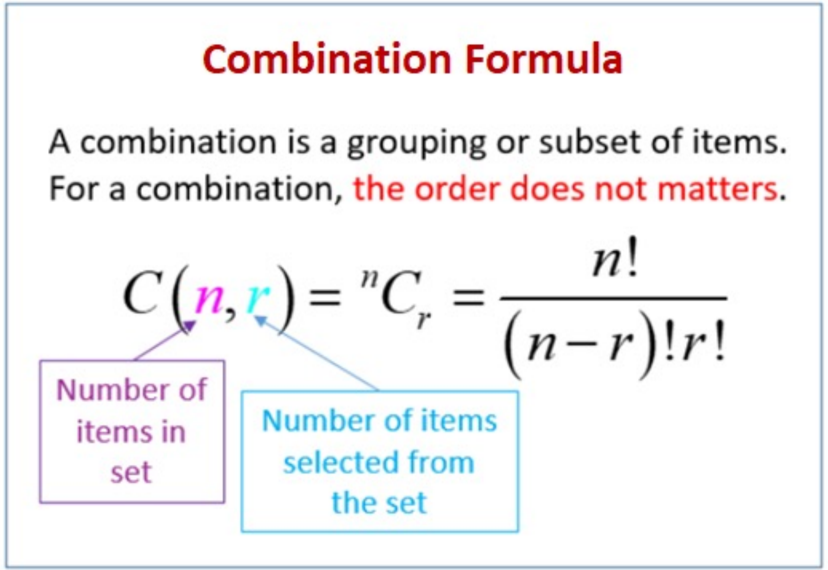
$$C(n,r) = {}^{n}C_{r} = \frac{n!}{(n-r)!r!}$$

Number of items in set

Number of items selected from the set

- If you have a standard deck of 52 cards and you randomly select 5 cards (without replacement), what is the probability that NONE of the selected cards are hearts?

- 1: Calculate the total number of ways to select 5 cards from a 52-card deck (52C5).
- 2: Calculate the number of ways to select 5 cards from the 39 non-heart cards (39C5).
- 3: Divide the result from Step 2 by the result from Step 1 to find the probability of selecting no hearts.
- 4: Round the answer to two significant digits.



Combination Formula

A combination is a grouping or subset of items. For a combination, the order does not matters.

$$C(n,r) = {}^{n}C_{r} = \frac{n!}{(n-r)!\,r!}$$

Number of items in set

Number of items selected from the set

Binomial Distribution is A discrete probability distribution that describes the number of successes (k) in a fixed number of independent trials (n), each with a constant probability of success (p).

Our random variable

Num trials

Probability of success on each trial

$$X \sim \mathrm{Bin}(n, p)$$

Is distributed as a

Binomial

With these parameters

Probability Mass Function for a Binomial

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Probability that our variable takes on the value k

Probability Mass Function (PMF):

- n: number of trials
- k: number of successes
- p: probability of success in a single trial

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!} \; for\ 0 \le k \le n$$

Example

- If the probability of a student passing a test is 4/5, what is the probability that out of 5 students, exactly 3 will pass the test? Assume that the students' test results are independent events.

- n = 5 (total number of students)
- k = 3 (number of students passing the test)
- p = 4/5 (probability of a single student passing the test)
- Simplify the fraction: P(X = 3) = 128 / 625

$$\underset{\substack{\uparrow \\ \text{Probability that our} \\ \text{variable takes on the} \\ \text{value k}}}{P(X=k)} = \overbrace{\binom{n}{k} p^k (1-p)^{n-k}}^{\substack{\text{Probability Mass Function} \\ \text{for a Binomial}}}$$

Hypergeometric Distribution

- **The hypergeometric distribution** is a discrete probability distribution that calculates the likelihood an event happens k times in n trials when you are sampling from a small population without replacement.
- This distribution is like the binomial distribution except for the sampling without replacement aspect.

$$P(X = x) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}, \qquad P(X = x) = \frac{n!k!(N-n)!(N-k)!}{N!x!(k-x)!(n-x)!(N-k-n+x)!}.$$

Example

- You have six marbles numbered 1 to 6 in a bag. You pull out four marbles without replacement. What is the probability that exactly three of the marbles pulled out are even numbers? Report your answer to two significant digits.
- **Dependence of Trials**: Each trial is dependent, meaning the outcome of one trial affects the outcome of subsequent trials.
- **No Replacement**: Trials are conducted without replacement. This means once an item is drawn, it is not placed back into the population, altering the probabilities for subsequent trials.
- Total number of balls (N) = 6
- Number of even balls (M) = 3 (balls numbered 2, 4, 6)
- Number of balls drawn (n) = 4
- Number of successful draws (even balls, x) = 3

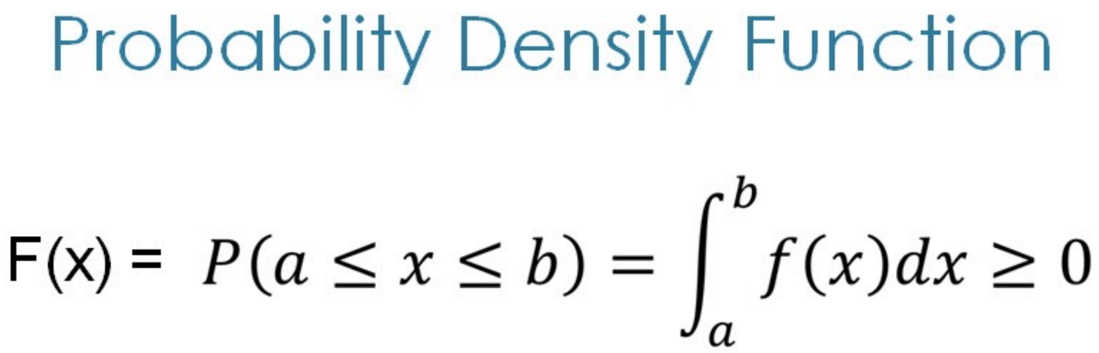
$$P(X = x) = h(x; n, M, N) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

Probability Distributions

Continuous

- *A probability distribution in which the random variable X can take on any value (is continuous). Because there are infinite values that X could assume, the probability of X taking on any one specific value is zero.*

- The probability that X falls between two values (a and b) equals the integral (area under the curve) from a to b



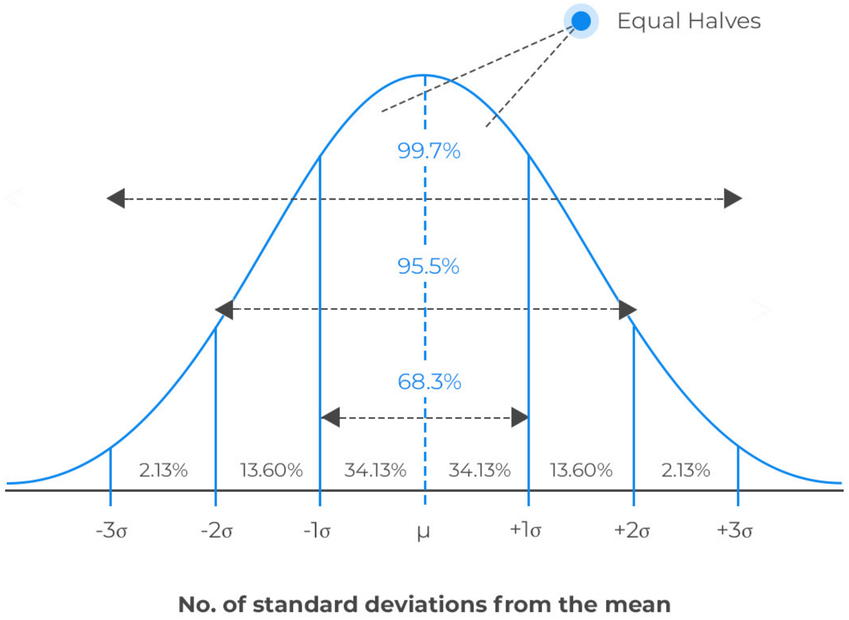
Probability Density Function

$$F(x) = P(a \leq x \leq b) = \int_a^b f(x)dx \geq 0$$

The Normal Distribution

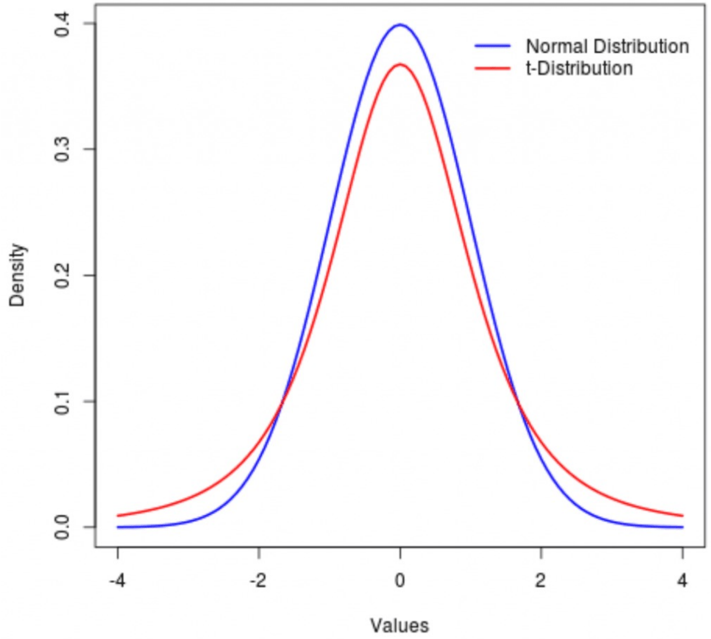
The normal distribution is the most common type of distribution. The standard normal distribution has two parameters: the mean and the standard deviation.

- In a normal distribution, mean, median, and mode are equal.
- The normal distribution is symmetric and centered on the mean.
- While the x-axis ranges from negative infinity to positive infinity.
- Nearly all of the X values fall within +/- three standard deviations of the mean (99.7% of values), while ~68% are within +/-1 standard deviation and ~95% are within +/- two standard deviations.



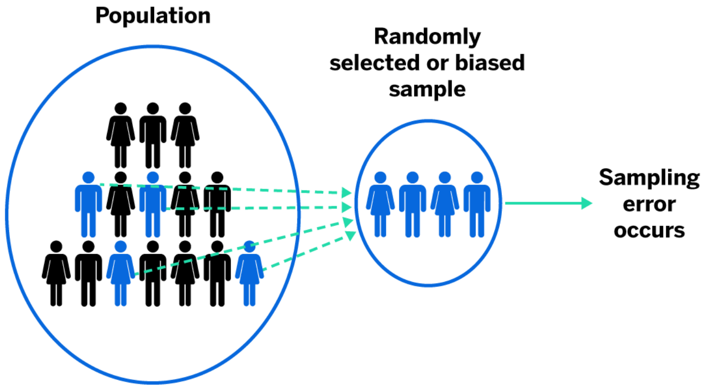
Student’s t-distribution

- The T distribution (also called Student’s T Distribution) is a family of distributions that look almost identical to the normal distribution curve, only a bit shorter and fatter.
- It is used for estimating population parameters for small sample sizes or unknown variances.
- The larger the sample size, the more the t distribution looks like the normal distribution. In fact, for sample sizes larger than 20, the distribution is almost exactly like the normal distribution.



Sampling

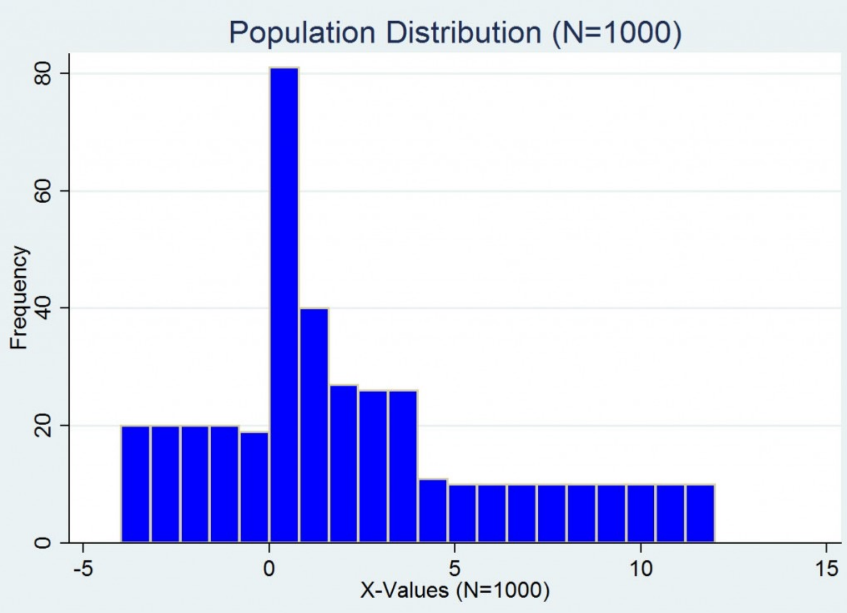
- **Simple random sampling** is one of the core concepts to much of data collection and analysis.
- *In simple random sampling, each individual or object in a population has an equal probability of being selected into the sample.*
- **Sampling error** is the difference between a sample statistic and the true population parameter
- With a large sample size, the sample means are normally distributed with a mean of µ (mu) and a standard deviation of σ/sqrt(n). **The standard deviation of the sample means** is called the standard error of the mean (σ/sqrt(n)).

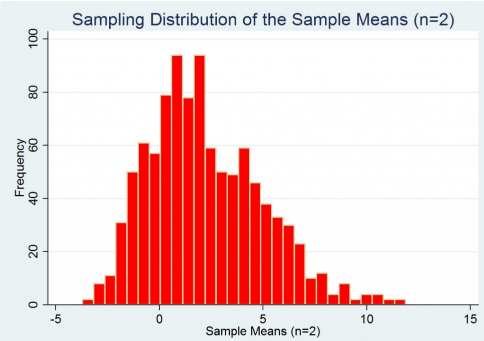


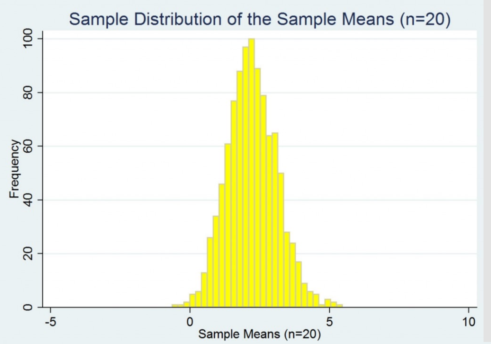
$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

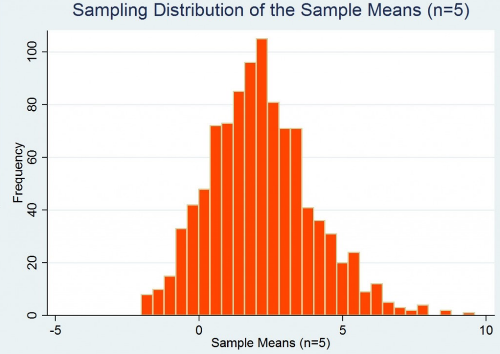
The Central Limit Theorem

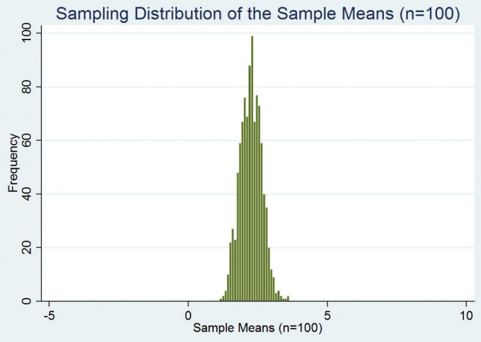
- The central limit theorem says that the sampling distribution of the mean will always be **normally distributed**, as long as the sample size is large enough.
- Regardless of whether the population has a normal, Poisson, binomial, or any other distribution, the sampling distribution of the mean will be normal.


Population Distribution (N=1000)

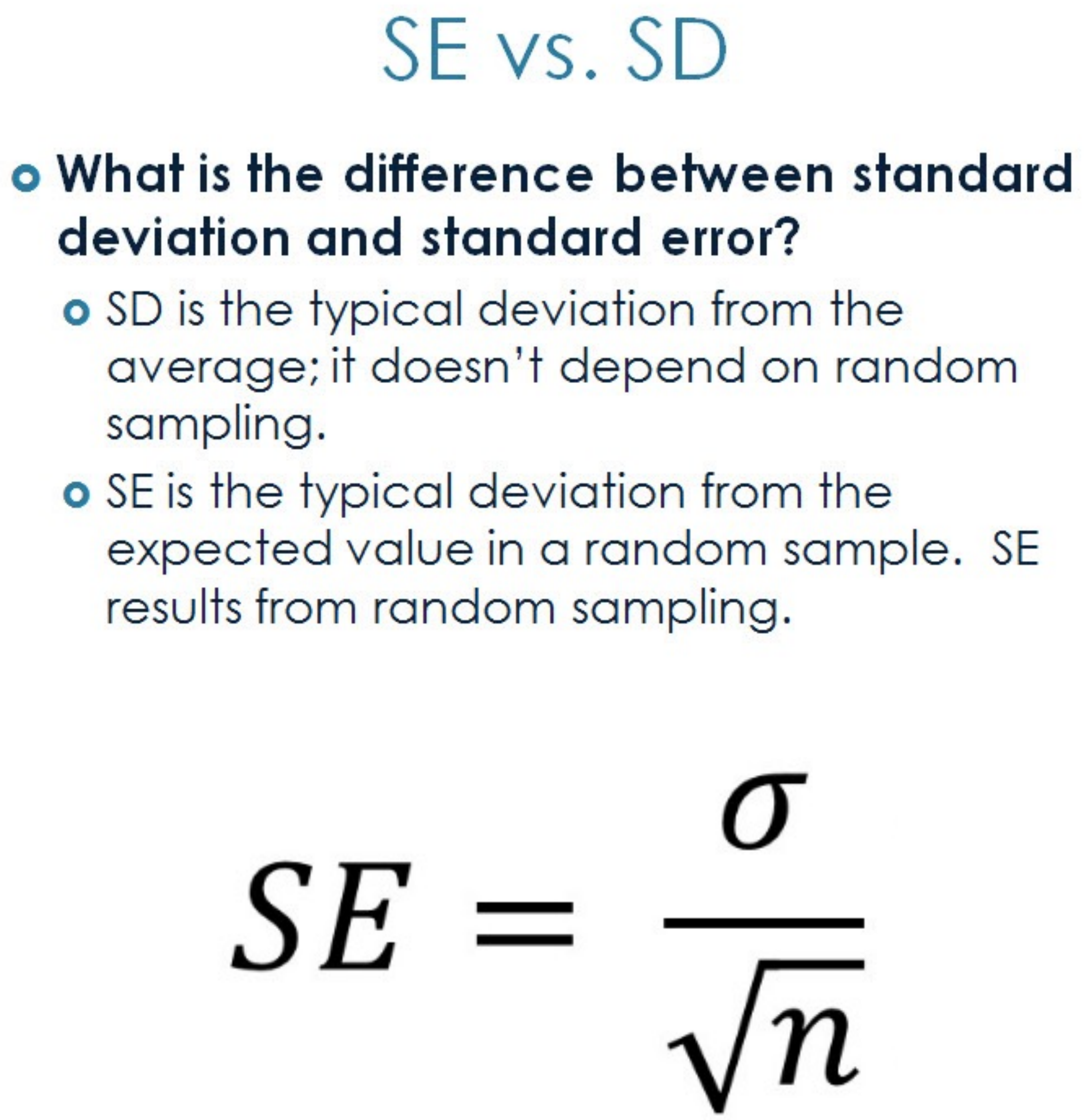

Sampling Distribution of the Sample Means (n=2)


Sample Distribution of the Sample Means (n=20)


Sampling Distribution of the Sample Means (n=5)


Sampling Distribution of the Sample Means (n=100)

Standard Deviation vs. Standard Error



SE vs. SD

- **What is the difference between standard deviation and standard error?**
  - SD is the typical deviation from the average; it doesn't depend on random sampling.
  - SE is the typical deviation from the expected value in a random sample. SE results from random sampling.

$$SE = \frac{\sigma}{\sqrt{n}}$$

Example

$$SE = \frac{\sigma}{\sqrt{n}}$$

- A biologist is studying the number of fish in a large lake. She sets up 50 traps across the lake and counts the number of fish caught in each trap after 24 hours. Assume the population distribution of fish counts across all traps is discrete and positively skewed with a mean of 12 and a standard deviation of 6. Assuming the Central Limit Theorem holds, find the probability that the mean number of fish in 50 randomly selected traps is less than 10 fish.

Solution:

Population mean (μ) = 12

Population standard deviation (σ) = 6

Sample size (n) = 50

Event of interest: Mean number of fish < 10

$SE = \sigma/\sqrt{n} = 6/\sqrt{50} \approx 0.85$

$z = (\bar{x} - \mu) / SE = (10 - 12) / 0.85 \approx -2.35$

$P(z < -2.35) \approx 0.0094$ or 0.94%

Example

- The average lifespan of the bulbs produced is 1,500 hours with a population standard deviation of 300 hours. So…….
- (1) We select 100 light bulbs at random. What is the standard deviation of the sample means?
- (2) What is the probability that one bulb, selected at random, will last longer than 1,800 hours?
- (3) What is the probability that the average of 100 randomly selected bulbs is greater than 1,800 hours?

Solution:

(1) The ***sd*** of the sample means equals the known population standard deviation divided by the square root of the sample size (n). $sd(\bar{x}) = \frac{300}{\sqrt{100}} = 30$

Example

- (2) What is the probability that one bulb, selected at random, will last longer than 1,800 hours?

Solution:
Because we are interested in just one bulb (not an average), we use this z-score formula:

$$Z = \frac{X - \mu}{\sigma}$$

Therefore, Z=(1,800 – 1,500)/300 = 300/300 = 1.

P(X>1800) ➡️ P(Z > 1)

We look this up in our z table and find that P(Z > 1)= P(Z < -1) =
= 0.1587.

There is approximately 15.9% chance that one light bulb chosen at random will last longer than 1,800 hours.

Example

- (3) What is the probability that the average of 100 randomly selected bulbs is greater than 1,800 hours?

Solution:
Because we are interested in the average, we use the following z-score formula:

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

Therefore, $\mathrm{Z} = \frac{1800-1500}{300/\sqrt{100}} = 10$

And the probability that z>10 is practically zero.

Calculating Z-Scores

- Find probability of a z-score from z-table
- https://www.ztable.net
- This table gives a probability that a statistic is less than Z (i.e. between negative infinity and Z).
- If you need to find the area to the right of a z-score (Z greater than some value), you need to subtract the value in the table from one.

Example

- Suppose the weight of apples from a certain orchard follows a normal distribution with a mean of 150 grams and a known standard deviation of 20 grams. Given this information, what is the probability that a randomly selected apple will weigh more than 175 grams?

$$Z = \frac{X - \mu}{\sigma}$$

Example

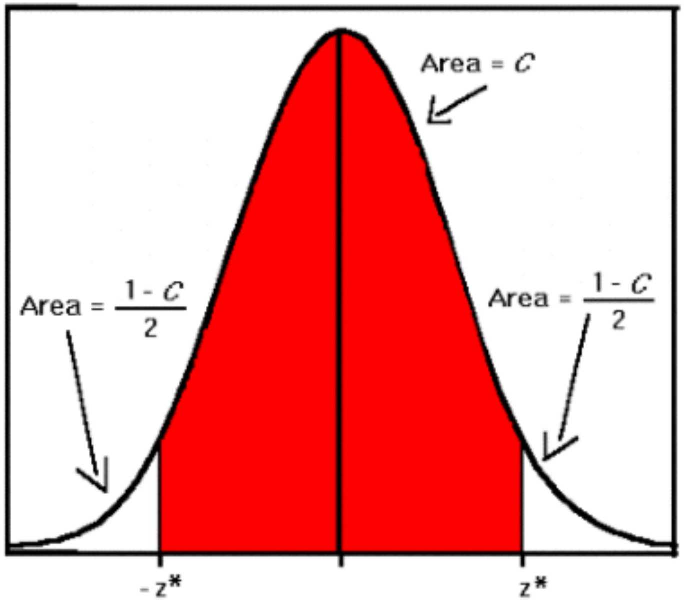
- A company manufactures a new type of phone battery. Assume that the battery life is normally distributed with a standard deviation of 2 hours. Based on this assumption, how large should the mean (μ) battery life be in order for 95% of the batteries to last for the advertised duration of 10 hours?

Solution:

1. Advertised duration: 10 hours
2. Standard deviation (σ): 2 hours
3. Desired percentage: 95% (5th percentile)
4. Find the z-score associated with the 5th percentile (z = -1.645). https://www.ztable.net
5. Use the inverse z-score formula to find the required mean battery life.
   $\mu = x - z \cdot \sigma$
6. μ = 10 - (-1.645)*2 = 13.29

Confidence Intervals

- In statistical inference, one wishes to estimate population parameters using observed sample data.
- ***A confidence interval*** gives an estimated range of values which is likely to include an unknown population parameter, the estimated range being calculated from a given set of sample data. (Definition taken from Valerie J. Easton and John H. McColl's Statistics Glossary v1.1).
- Common choices for the confidence level C are 0.90, 0.95, and 0.99. These levels correspond to percentages of the area of the normal density curve. Because the normal curve is symmetric, half of the area is in the left tail of the curve, and the other half of the area is in the right tail of the curve.



- A 95% confidence interval covers 95% of the normal curve -- the probability of observing a value outside of this area is less than 0.05.
- The value z* representing the point on the standard normal density curve such that the probability of observing a value greater than z* is equal to p is known as the upper p critical value of the standard normal distribution.
- For example, if p = 0.025, the value z* such that P(Z > z*) = 0.025, or P(Z < z*) = 0.975, is equal to 1.96. A 95% confidence interval for the standard normal distribution, then, is the interval (-1.96, 1.96), since 95% of the area under the curve falls within this interval.

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

$$(\bar{x} - Z^* \frac{\sigma}{\sqrt{n}}, \bar{x} + Z^* \frac{\sigma}{\sqrt{n}})$$

But if we rearrange this equation to solve for μ, we get…

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

Example

- A quality control manager at a factory wants to estimate the mean weight of the products produced. A random sample of 200 products is selected, and their weights are recorded. The sample mean weight is 5.2 pounds with a known population standard deviation of 0.8 pounds. Construct a 99% confidence interval for the mean weight of the products. Please enter the confidence interval in the following format (lower bound, upper bound) on the answer page.

Solution:

1. Identify the sample statistics and confidence level:

$\bar{x}$ = 5.2 pounds, σ = 0.8 pounds, n = 200, confidence level = 99%

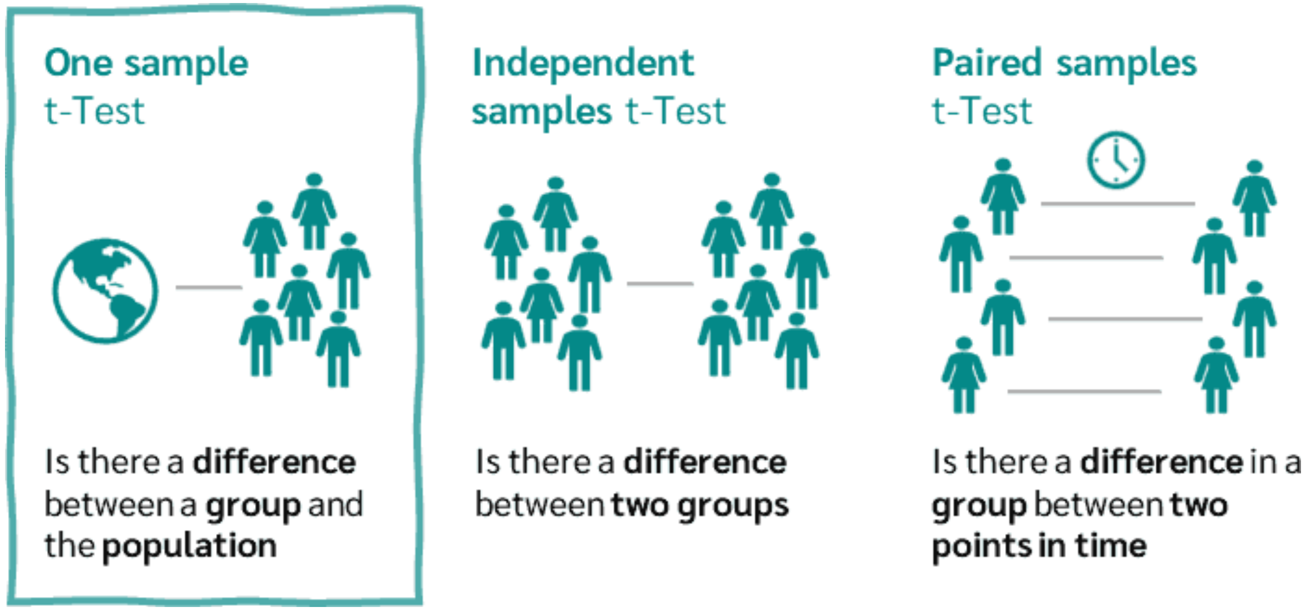
2. Find the critical z-value for a 99% confidence level (z = 2.58).

3. CI = 5.2 ± 2.58*(0.8/√200)

$$\left(\bar{x} - Z^* \frac{\sigma}{\sqrt{n}}, \bar{x} + Z^* \frac{\sigma}{\sqrt{n}}\right)$$

Inference: Comparison of Means

- Three major types of comparison of means tests:
  1. **One sample test**: We make an inference to a population in comparison to some set value
  2. **Two independent sample test:** In this test, we collect two independent samples to test whether there is a difference in means between two populations (or if one population mean is greater or less than the other) **.**
  3. **Paired or Repeated measure test**: This test compares paired data, such as data collected before and after a treatment.



One sample t-Test

Is there a **difference** between a **group** and the **population**

Independent samples t-Test

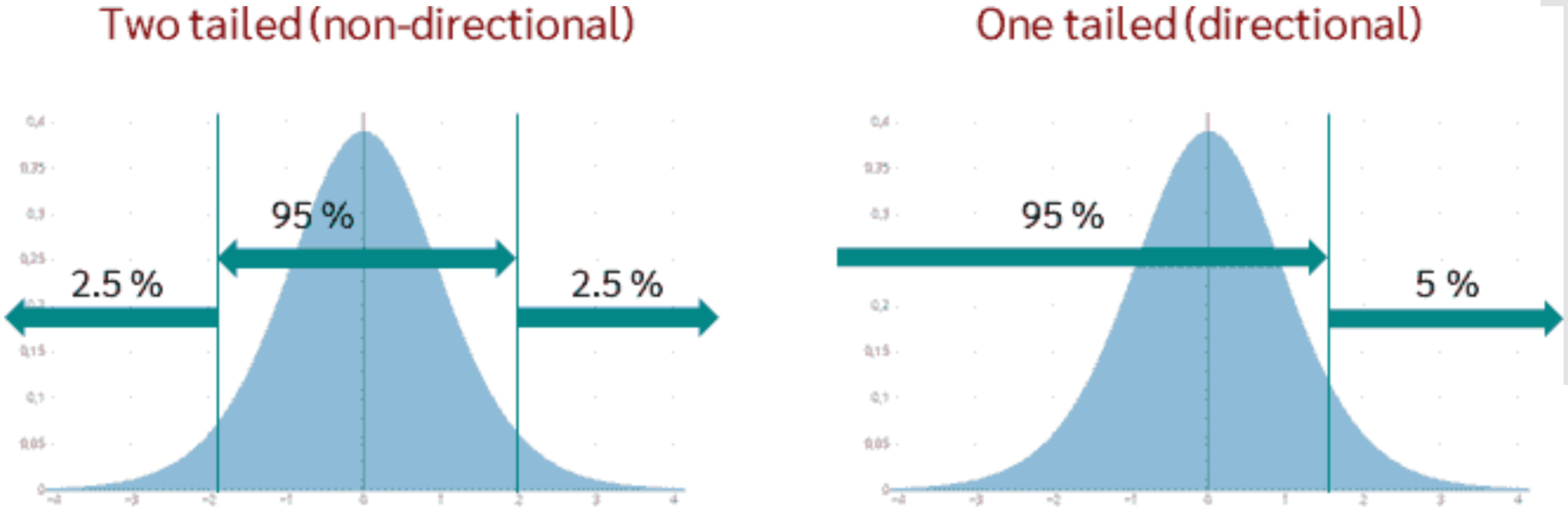
Is there a **difference** between **two groups**

Paired samples t-Test

Is there a **difference** in a **group** between **two points in time**

Inference: Comparison of Means

- For a comparison of means test, you may use either a **one-sided** or **two-sided** test.
    - A one-sided test (leading to a one-sided p-value) examines whether one mean is greater (or less than) the other mean.
    - If you want to test whether there is a difference between two means (without any directionality), then you use a two-sided test (and subsequently a two-sided p-value (see below).
    - The null and alternative hypotheses should reflect whether or not you are using a one- or two-sided comparison of means test.



Two tailed (non-directional)

One tailed (directional)

Example 1

- A manufacturer claims that their new type of car tire has a mean lifetime of 50,000 miles with a standard deviation of 5,000 miles. A random sample of 50 tires is tested. Is there evidence to suggest that the mean lifetime of the tires is less than the claimed 50,000 miles?
- ***What type of test is used for this?***
- ***One-sample t-test***: Since the sample size is given (50 tires) and the population standard deviation is known, a one-sample t-test should be used to make the inference.
- Null and alternative hypotheses: The null hypothesis would be that the mean lifetime is equal to 50,000 miles, while the alternative hypothesis would be that the mean lifetime is less than 50,000 miles (one-tailed test).

Example 2

- A gym wants to evaluate the effectiveness of a new 6-week fitness program. They randomly select 20 members and record their body fat percentage before and after completing the program. The gym wants to determine if there is a significant reduction in body fat percentage after participating in the fitness program.
- ***What type of test is used for this?***
- ***Matched pairs test:*** The study involves comparing two sets of measurements from the same group of participants. Each participant serves as their own control, as their body fat percentage is measured before and after the intervention
- **Null and alternative hypotheses**:
- Null hypothesis would be the mean difference in body fat percentage before and after the program is greater than or equal to zero. (μ_diff >= 0).
- Alternative hypothesis would be the mean difference in body fat percentage before and after the program is less than zero. (μ_diff < 0) .
- One-tailed test.

Example 3

- A university wants to compare the effectiveness of two different teaching methods on student performance in a calculus course. The university randomly assigns 100 students to two groups: one group attends traditional lectures, while the other group participates in interactive workshops. At the end of the semester, the university compares the mean scores of the final exams between the two groups.

- ***What type of test is used for this?***
- **Two-sample t-test (Independent Samples t-test):**The study involves comparing the means of two independent groups. The students were randomly assigned to the two groups, ensuring independence between the samples.
- **Null and alternative hypotheses:** Null hypothesis would be the mean final exam scores are equal for both teaching methods ($\mu_1 = \mu_2$).Alternative hypothesis would be the mean final exam scores are different for the two teaching methods ($\mu_1 \neq \mu_2$).
- Two-tailed test.