# Lecture 3

Probability and Statistics

# Statistics for Data Science

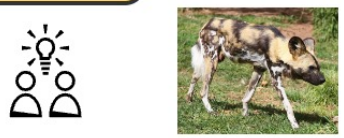
- Statistics involves techniques for analyzing data to gain insights.

- Data science uses statistics for data-driven decisions. Ex, predicting sales using a statistical model:
  - Sampling methods gather data, reducing bias.
  - Descriptive stats explore data via visuals and summaries.
  - Inferential stats draw population conclusions from samples, using probability.
  - Knowing stats in data science ensures accurate interpretation and communication.

# Statistics for Data Science

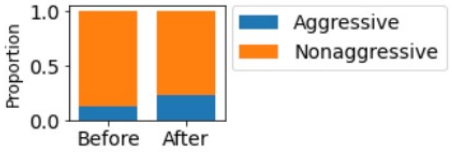
1. A data scientist teams with a zookeeper to examine behavior changes in African-painted dogs due to a new pack member.
2. An observational study gathers behavior data from pack dogs before/after the newcomer.
3. Data summaries reveal a 0.096 increase in aggressive behaviors (0.128 to 0.224) post-introduction.
4. Statistically, this increase exceeds 0, showing aggression rise with new dog.
5. The scientist informs the zookeeper: new dogs relate to increased aggression.

**Data collection**



| Dog ID | Date | Behavior | New dog | Aggressive |
|---|---|---|---|---|
| 54672 | 3/2 | grooming | no | no |
| 43284 | 3/2 | aggression | no | yes |
| 12114 | 3/6 | sleeping | no | no |
| ... | ... | ... | ... | ... |
| 43284 | 5/18 | aggression | yes | yes |
| 87401 | 5/19 | aggression | yes | yes |
| 12114 | 5/19 | eating | yes | no |

**Descriptive statistics**




Observed difference: 0.096

**Inferential statistics**

```
observed difference = 0.096
test statistic = -1.993
p-value = 0.046
```

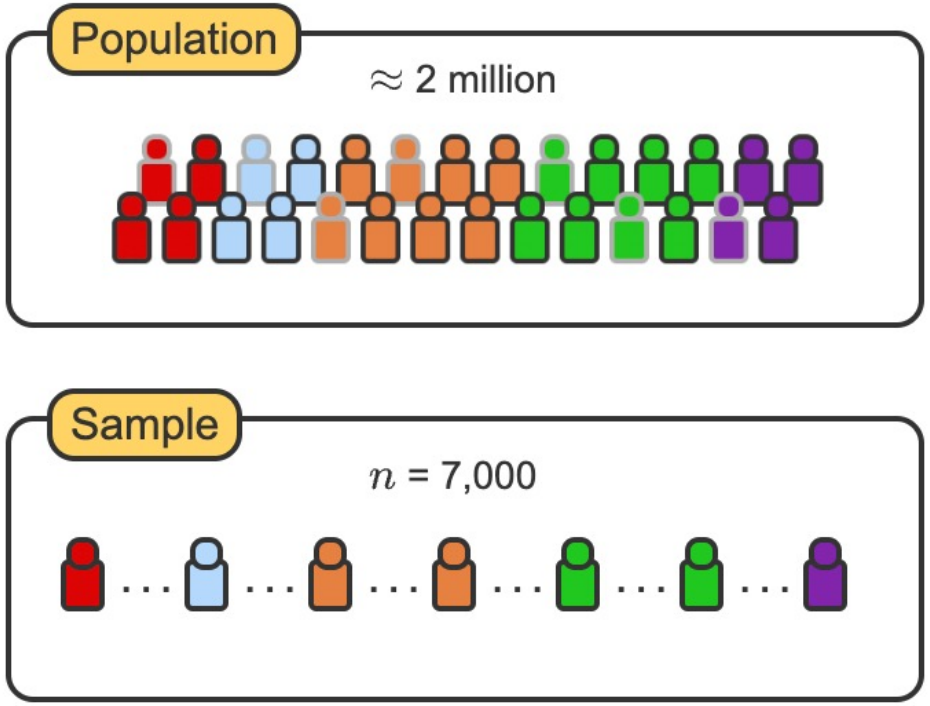
**Conclusions**

# Statistic vs. Parameter

- A **statistic** is a number that describes some characteristic of a sample.
- A **parameter** is a number that describes some characteristic of a population.

# Sampling

- Data science analyzes data to grasp populations.
- Population: all relevant individuals, items, or events.
- Samples: subsets due to resource limits, comprising observational units (n).



# Sampling Methods

- Sampling picks units from a population for a sample. Units should ideally mirror the population. Common methods:
  - **Random Sampling**: Units are chosen randomly, giving all an equal chance. Example: Picking names from a hat for a survey.
  - **Stratified Sampling**: Population divides into subgroups (strata), then random sampling from each. Example: Surveying students by grade level (freshmen, sophomores, etc.).
  - **Cluster Sampling**: Population divides into clusters, randomly choosing some clusters, and then including all units within those selected clusters. Example: Surveying households by selecting random neighborhoods.
  - **Systematic Sampling**: Choosing every nth unit from the population. Example: Surveying every 10th customer entering a store.
  - **Convenience Sampling**: Easily accessible units are chosen. Example: Surveying people passing by in a park.

# Observational Studies vs. Experiments

- Data can be collected either through an **observational** study or an **experiment**.
- *Observational study*: data is collected by recording the responses as they occur without any direct influence on the observed data. Ex: Home features data collection.
- *Experiment*: treatments are first assigned to observational units and then responses are recorded. Ex: A/B test with webpage layouts for click data.
  - Enables **causal conclusions** with random treatment assignment

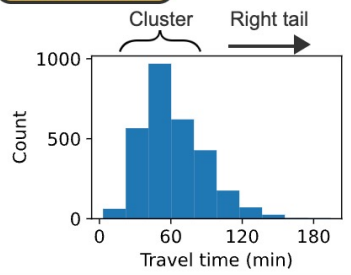
# Descriptive Statistics

- A feature is a measurable trait on an observational unit. Descriptive statistics capture its key aspects. For instance, the median home price summarizes home prices effectively.
- Understanding a feature starts with exploring its distribution. This reveals potential values and their frequencies. Graphs visually convey distribution shapes, especially for numerical features:
  - A cluster is a distinct, frequent group within a distribution.
  - Tails are distribution ends: left for low, right for high values.

Dataframe

| ID | License | Cost | TotalTime | Satisfaction |
|---|---|---|---|---|
| 126 | 0 | 5.0 | 95 | 1 |
| 131 | 1 | 3.6 | 60 | 3 |
| 142 | 1 | 2.5 | 80 | 4 |
| 145 | 1 | 5.5 | 43 | 3 |
| 168 | 1 | 2.3 | 33 | 3 |
| ... | ... | ... | ... | ... |
| 22598 | 1 | 0.0 | 68 | 4 |
| 22620 | 1 | 0.0 | 50 | 4 |
| 22642 | 1 | 3.4 | 26 | 2 |
| 23003 | 1 | 1.2 | 35 | 4 |
| 23014 | 1 | 0.0 | 21 | 4 |

Visual summary

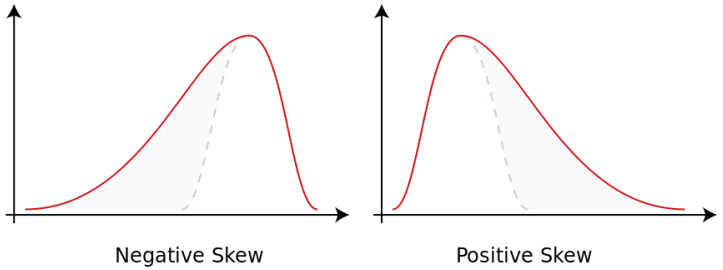



Numerical summary

| | |
|---|---|
| count | 2927.00 |
| mean | 62.60 |
| std | 25.83 |
| min | 3.00 |
| 25% | 44.50 |
| 50% | 59.00 |
| 75% | 78.00 |
| max | 194.00 |

# 4 Moments in Statistics

- *First* moment, the **mean ($\mu$)**, or average, of a numerical feature is the sum of all values divided by the total number of values.
  - The **median ($\nu$)** of a numerical feature is the middle value of the ordered data (it is not considered as the first moment)
- *Second* moment, the **variance ($\sigma^2$)** is the average squared distance a numerical feature's values lie from the distribution's mean.
- *Third* moment, **Skewness ($\gamma$)** is a measure of the amount and direction of skew, or departure from symmetry. Positive values indicate the distribution has an extended right tail and negative values indicate the distribution has an extended left tail.




- *Fourth* moment, **Kurtosis ($\kappa$)** is a measure of tail heaviness. Larger values of kurtosis indicate a greater presence of extreme values in the distribution.

# 4 Moments in Statistics

- **mean ($\mu$)**:
  - Arithmetic mean: $\bar{x} = \frac{1}{n}(\sum_{i=1}^{n} x_i)$ (good for sharing things equally among a group)
  - Geometric mean: $\bar{x} = (\prod_{i=1}^{n} x_i)^{\frac{1}{n}}$ (good for finding an average that considers different parts of a whole)
  - Harmonic mean: $\bar{x} = n\left(\sum_{i=1}^{n} \frac{1}{x_i}\right)^{-1}$ (good for understanding how things grow or change when they're being multiplied)
- **variance ($\sigma^2$)**:
  - (population) $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - (sample) $s^2 = \frac{1}{n-1}\sum_{i=1}^{n} x_i$
- **Skewness ($\gamma$)**:
  - (population) $\gamma = \sqrt{n} \times \frac{\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^{3/2}}$
  - (sample) $\gamma = \frac{n\sqrt{n-1}}{n-2} \times \frac{\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^{3/2}}$

**Kurtosis ($\kappa$)**:

(population) $\kappa = n \times \frac{\sum_{i=1}^{n}(x_i-\bar{x})^4}{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^2}$

(sample) $\kappa = \frac{n(n+1)(n-1)}{(n-2)(n-3)} \times \frac{\sum_{i=1}^{n}(x_i-\bar{x})^4}{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^2}$

# $\sigma$ vs. $\sigma^2$

| | Variance | Standard Deviation |
|---|---|---|
| **Meaning** | Variance is a numerical value that describes the variability of observations from its arithmetic mean. | Standard deviation is a measure of dispersion of observations within a data set. |
| **What is it?** | It is the average of squared deviations. | It is the root mean square deviation. |
| **Labelled as** | $\sigma^2$ | $\sigma$ |
| **Expressed in** | Squared units | Same units as the values in the set of data. |
| **Indicates** | How far individuals in a group are spread out. | How much observations of a data set differs from its mean. |

# Random Variable

- A **random variable** is a rule that assigns a number to every outcome in the sample space of an experiment.
  - E.g. coin toss (H,T) or (1,0)

- A random variable is typically defined using a **capital letter**
  - E.g. X = 1, if the coin toss yielded heads
- Random variables can be used to model quantities that can change if an experiment is repeated.

# Types of Random Variable

- **Discrete random variable**: it can take on a countable number of distinct values like the integers between 0 and 100.
  - It is typically counted
  - E.g. the possible outcomes of a dies roll is {1,2,3,4,5,6}

- **Continuous random variable**: it can take on any value within a range of values.
  - It is typically measured
  - E.g. measuring temperature

# Expectation

- Expectation is the **first moment** of every distribution
- Some distributions do not have any moments:
  - E.g. Cauchy distribution
- The expected or average value of a random variable is a useful way to summarize the information in the distribution.
- Discrete random variable:
  - $\mu = E[X] = \sum_i x_i p(x_i)$ , where $p(x_i)$ is the probability of outcome $x_i$.
- Continuous random variable:
  - $\mu = E[X] = \int_{-\infty}^{\infty} x f(x) dx$, where $f(x)$ is the probability density function of $x$.

# Linearity of Expectation

- If X and Y are two random variables defined on the same sample space S, and c is a real number:

$$E[X + Y] = E[X] + E[Y]$$
$$E[cX] = cE[X]$$

- Example: What is the expected number of aces in the 5-card hand?

$$E[A_1] = [x = ace]p(x = ace) + [x\ != ace]p(x != ace) = 1/13$$
$$\mathrm{A}_1 = \mathrm{A}_2 = \mathrm{A}_3 = \mathrm{A}_4 = \mathrm{A}_5 \rightarrow E[\mathrm{A}] = 5.E[\mathrm{A}_1] = 5/13$$

# Intro to Probability

- An **experiment** is a procedure that results in one out of a number of possible outcomes.
  - E.g. rolling a six-sided die
- An **outcome** is the result of an experiment.
  - E.g the number of dots displayed in a six-sided die
- The **set** of all possible outcomes is called the sample space of the experiment and is denoted S.
  - E.g. the sample space of a six-sided die roll is S = {1,2,3,4,5,6}
- An **event** is a subset of the sample space.
  - E.g. for a die roll, the event A is rolling an even number {2,4,6}

# Intro to Probability

- **Probability** is a measure of how likely an event is to occur.
- The probability of an event A is denoted P(A) and is the sum of the probabilities of each outcome in the event.
- Another definition of probability: the number of desired outcomes divided by the total number of outcomes in the sample space, assuming that all outcomes are equally likely.

# Unions of Events

- Sometimes it is easier to determine the probability of an event by defining the event in terms of other events.
  - E.g. to calculate the probability that a 5-card hand has at least four face cards (jack, queen, or king), we determine the probability that the 5-card hand has exactly four face cards or exactly five face cards:

$$p(F_4 \cup F_5) = p(F_4) + p(F_5) - p(F_4 \cap F_5) = \frac{\binom{12}{4}40}{\binom{52}{5}} + \frac{\binom{12}{5}}{\binom{52}{5}} - 0$$

# Complement of Events

- Sometimes it is easier to determine the probability that an event doesn't happen than to determine that the event does happen.
- The complement of an event $E$ is $S - E$ and is denoted by $\bar{E}$
  - $p(E) + p(\bar{E}) = 1$
  - $p(E) \cap p(\bar{E}) = \emptyset$
- Example: In the experiment in which the red and blue dice are thrown, define E to be the event that at least one of the dice comes up 6. The $\bar{E}$ is the event that neither die comes up 6. So,

$$p(E) = 1 - p(\bar{E}) = 1 - \frac{25}{36} = \frac{11}{36}$$

# Properties of Discrete Probability Distributions

- **PMF**: a probability mass function (pmf) assigns the probability that a discrete random variable is exactly equal to some value.
  - Typically depicted as a table, plot, or equation
  - $p(X = x)$ or $p(x)$ is typically used for the PMF of $X$.

- **CDF**: the cumulative distribution function (cdf) of a discrete random variable is the probability that for any number $x$, the observed value of the random variable will be at most $x$ or $p(X \leq x)$.
  - E.g. in rolling a fair die, the probability of getting less than or equal to 3 is:

$$F(3) = p(X \leq 3) = p(1) + p(2) + p(3)$$

# Expected Value, Mean, Average or the First Moment

- For a **discrete random variable** $X$ is:

$$\mu = E[X] = \sum_i x_i p(x_i)$$

- For a **continuous random variable** $X$ is:

$$\mu = E[X] = \int_{-\infty}^{\infty} x f(x) dx$$

# Variance or the Second Moment

- For a discrete random variable $X$ is:

$$\sigma^2 = V[X] = \sum_i \left((x_i - \mu)^2 . p(x_i)\right)$$

- For a continuous random variable $X$ is:

$$\sigma^2 = V[X] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$$

- The standard deviation is:

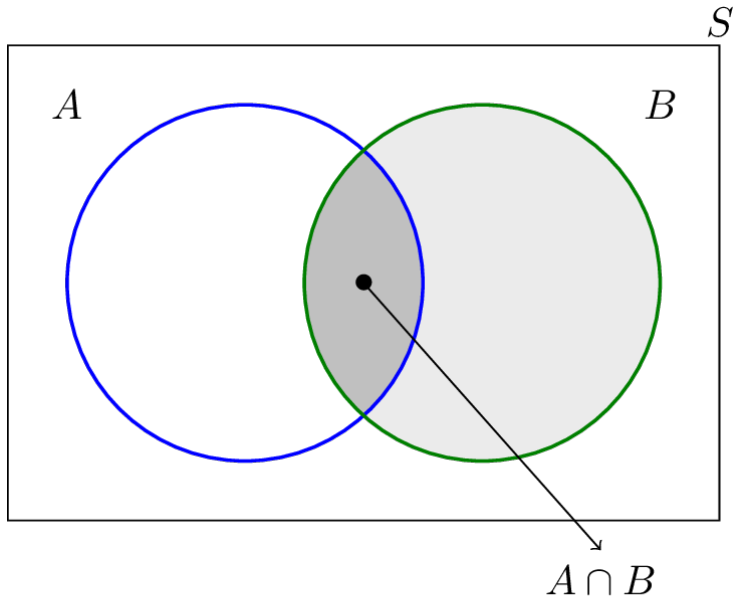
$$\sigma = \sqrt{\sigma^2}$$

- The precision is the inverse of variance:

$$p = \frac{1}{\sigma^2}$$

# Conditional Probability

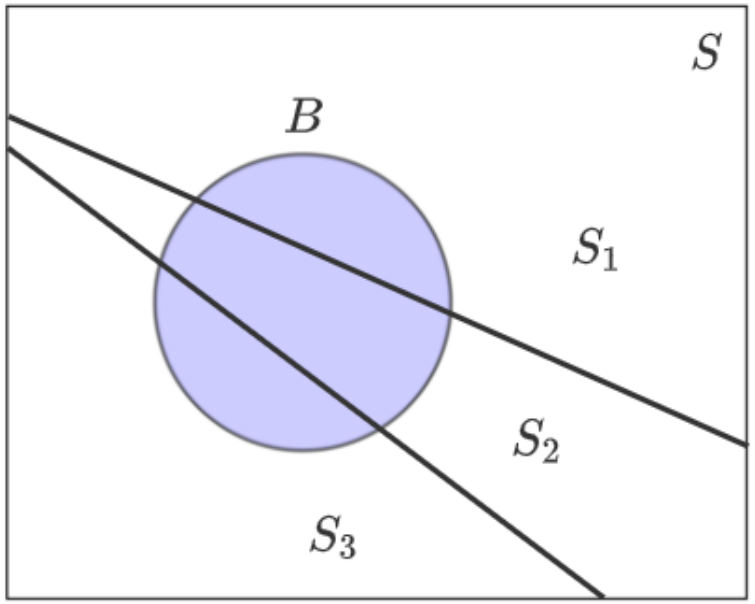
- **Conditional probability** is a measure of the probability of one event given that another event has occurred.




$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

# The Law of Total Probability

- The **law of total probability**: if the sample space is partitioned into two or more mutually exclusive subevents, the probability of an event B can be expressed in terms of conditional probabilities given each of the subevents.




3 subregions $S_1$, $S_2$, and $S_3$

$$B = (B \cap S_1) \cup (B \cap S_2) \cup (B \cap S_3)$$

$$P(B) = P(B \cap S_1) + P(B \cap S_2) + P(B \cap S_3)$$

$$P(B) = P(B|S_1)P(S_1) + P(B|S_2)P(S_2) + P(B|S_3)P(S_3)$$

# Independent Events

- Two events are **independent** if the probability of one event does not affect the probability of the other.
  - E.g. flipping the nickel and flipping the dime are independent events, because whether the nickel comes up heads or tails is not affected by whether the dime comes up heads or tails.
- If $A$ and $B$ are independent events, the probability that both $A$ and $B$ occur is (multiplication rule):

$$P(A \cap B) = P(A)P(B)$$

The generalized multiplication rule is:

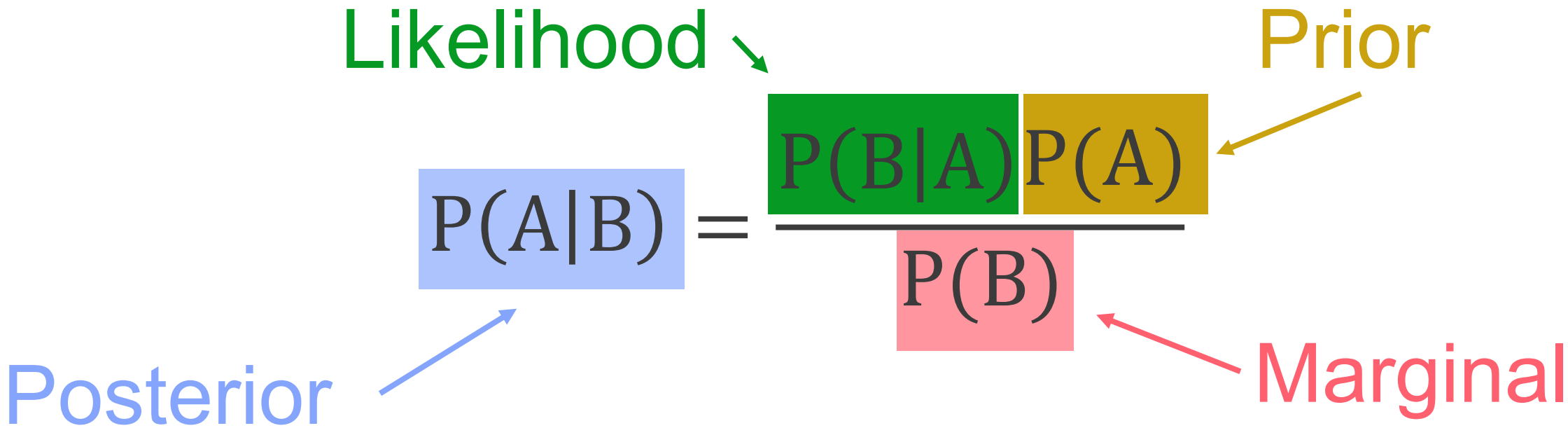
$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)P(A_2) \ldots P(A_n)$$

# Note

Independent events are different from mutually exclusive events. Mutually exclusive events are dependent events by definition, because if one event occurs, the other event cannot occur.

# Bayes' Theorem

Bayes' Theorem relates the probability of an event $A$ occurred. That is, Bayes' Theorem allows $P(B|A)$ to be calculated from $P(A|B)$.



# Bernoulli Trials & Binomial Distribution

- Random Experiment: 1/0, T/F, H/T, -1/1 etc.
  - e.g. tossing a coin: P(H) = p; P(T) = q
- Sequence of Bernoulli trials: n independent repetitions
  - n consecutive execution of an *if-then-else* statement
- $S_n$: sample space of n Bernoulli trials:

$S_1 = \{0,1\}$
$S_2 = \{(0,0),(0,1),(1,0),(1,1)\}$
$S_n = \{2^n$ n-tuple of 0s and 1s $\}$

- For $S_1$: $P(0) = q, P(1) = p, p + q = 1$, with $p, q \geq 0$
- E[X]=p
- Var[X]=pq

# Bernoulli Trials & Binomial Distribution

- The distribution over the random variable defined by the number of successes in a sequence of independent Bernoulli trials is called the **binomial distribution**.

$$b(k; n, p) = \binom{n}{k} p^k q^{n-k}, \text{ where } q = 1 - p$$

- E[X] = np
- Var[X] = npq

# Hypergeometric Distribution

- The hypergeometric distribution is a discrete probability distribution that describes the probability of success in draws, **without replacement**.

$$P(k) = \frac{\binom{x}{k}\binom{N-x}{n-k}}{\binom{N}{n}}$$

- $N$ = population size
- $n$ = number of draws
- $x$ = elements with a specific property in the population
- $k$ = hits for element with specific property

# Hypergeometric Distribution $\mu, \sigma^2$

- $E[X] = \mu = n\left(\frac{M}{n}\right) = np$
- $Var[X] = \sigma^2 = \left(\frac{N-n}{N-1}\right)(n)\left(\frac{M}{N}\right)\left(1 - \frac{M}{N}\right) = \left(\frac{N-n}{N-1}\right)npq$

# Poisson Distribution

- The Poisson distribution gives the probability of $k$ independent, randomly occurring events happening over a period or area where $\lambda$ events happen on average.

$$P(k) = \frac{e^{-\lambda}\lambda^{k}}{k!}$$

$k$ = 0, 1, 2, …

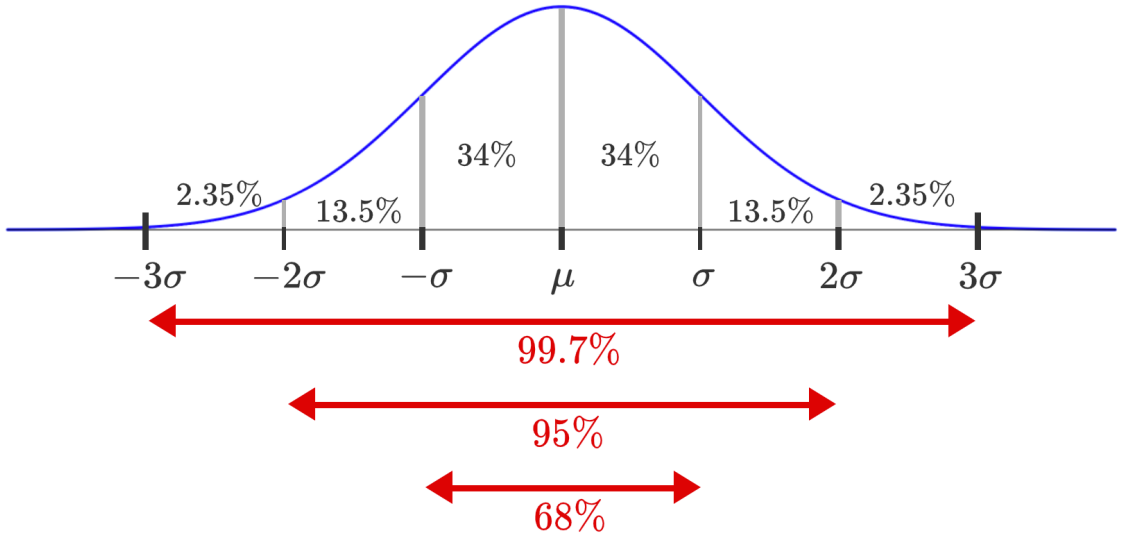
$\lambda$ = mean number of occurrences in the interval

$e$ = Euler's constant

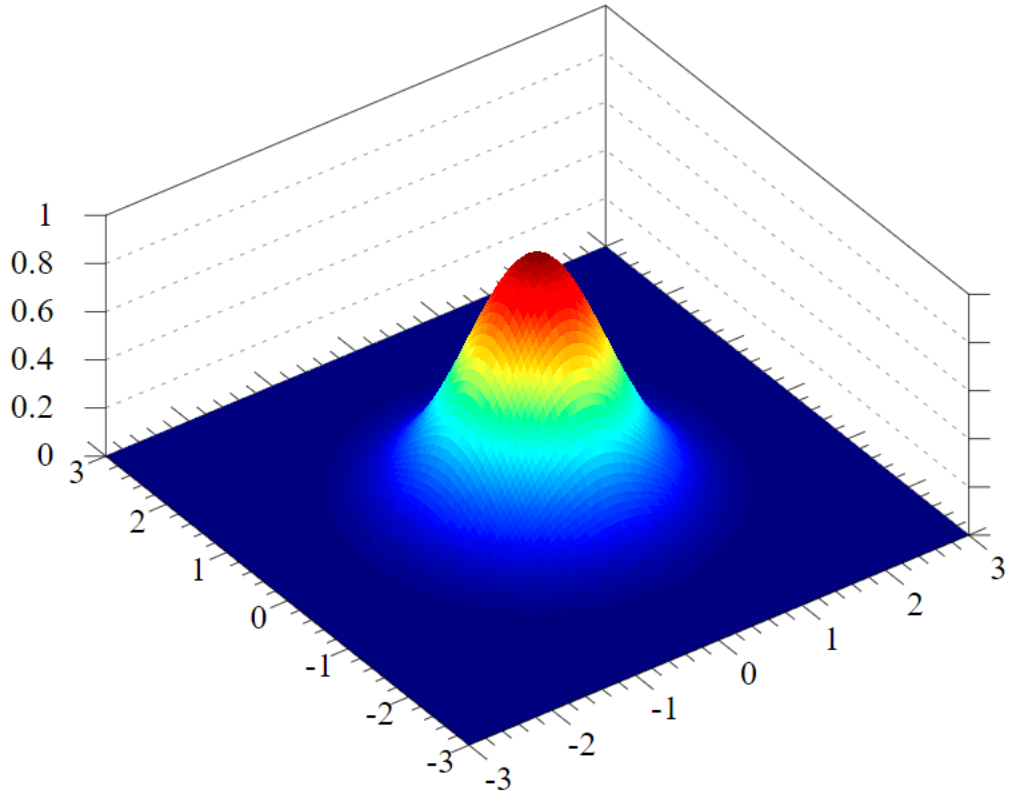
- $E[X] = \mu = \lambda$
- $Var[X] = \sigma^2 = \lambda$

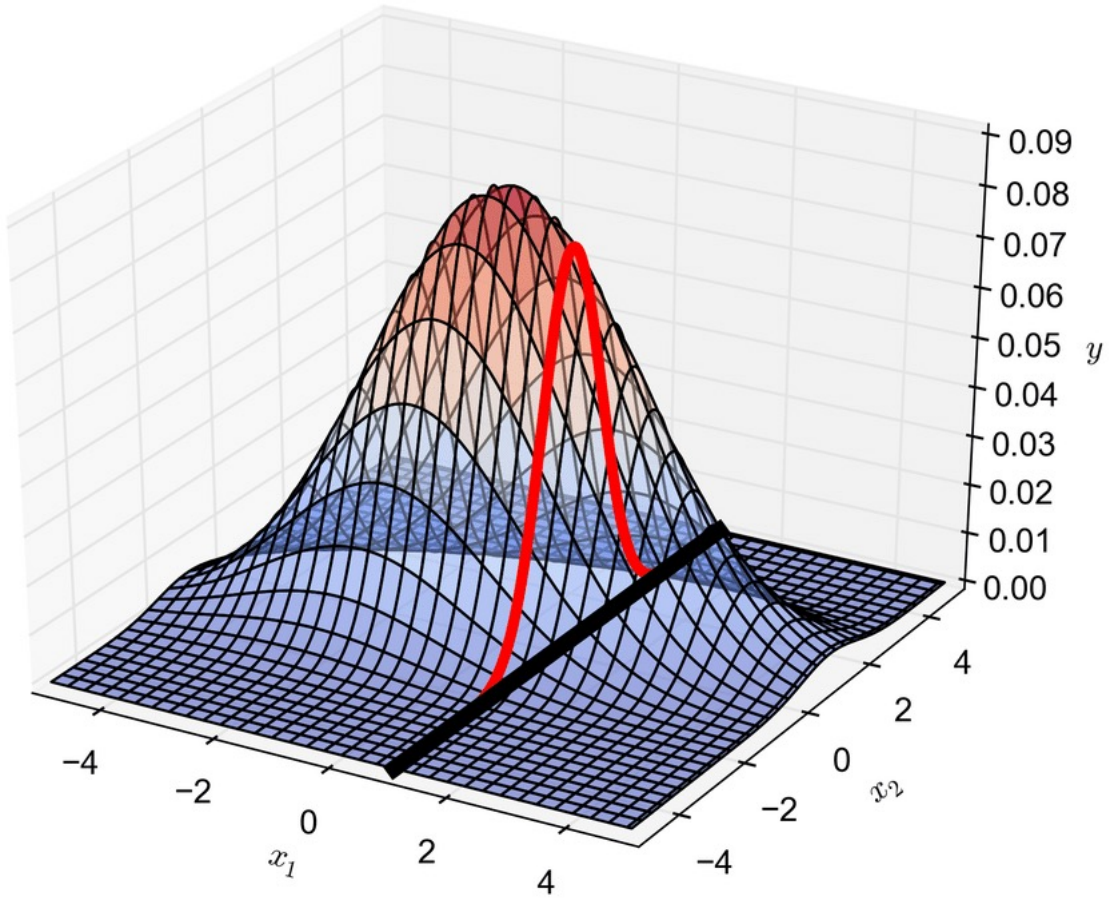
# Gaussian Distribution

$$\mathcal{N}(x;\mu,\sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$



# Marginal Distribution



# $z$-score Distribution

- A $z$-score is a signed value that indicates the number of standard deviations a quantity is from the mean $\mu$.
- A positive $z$-score indicates that the quantity is above the mean and a negative $z$-score indicates that the quantity is below the mean.
- A $z$-score with high absolute value implies that the quantity is farther from the mean, and thus more unusual.

$$z = \frac{x - \mu}{\sigma}$$

$\mu$ = mean

$x$ = raw score

$\sigma$ = standard deviation

# *t*-Distribution

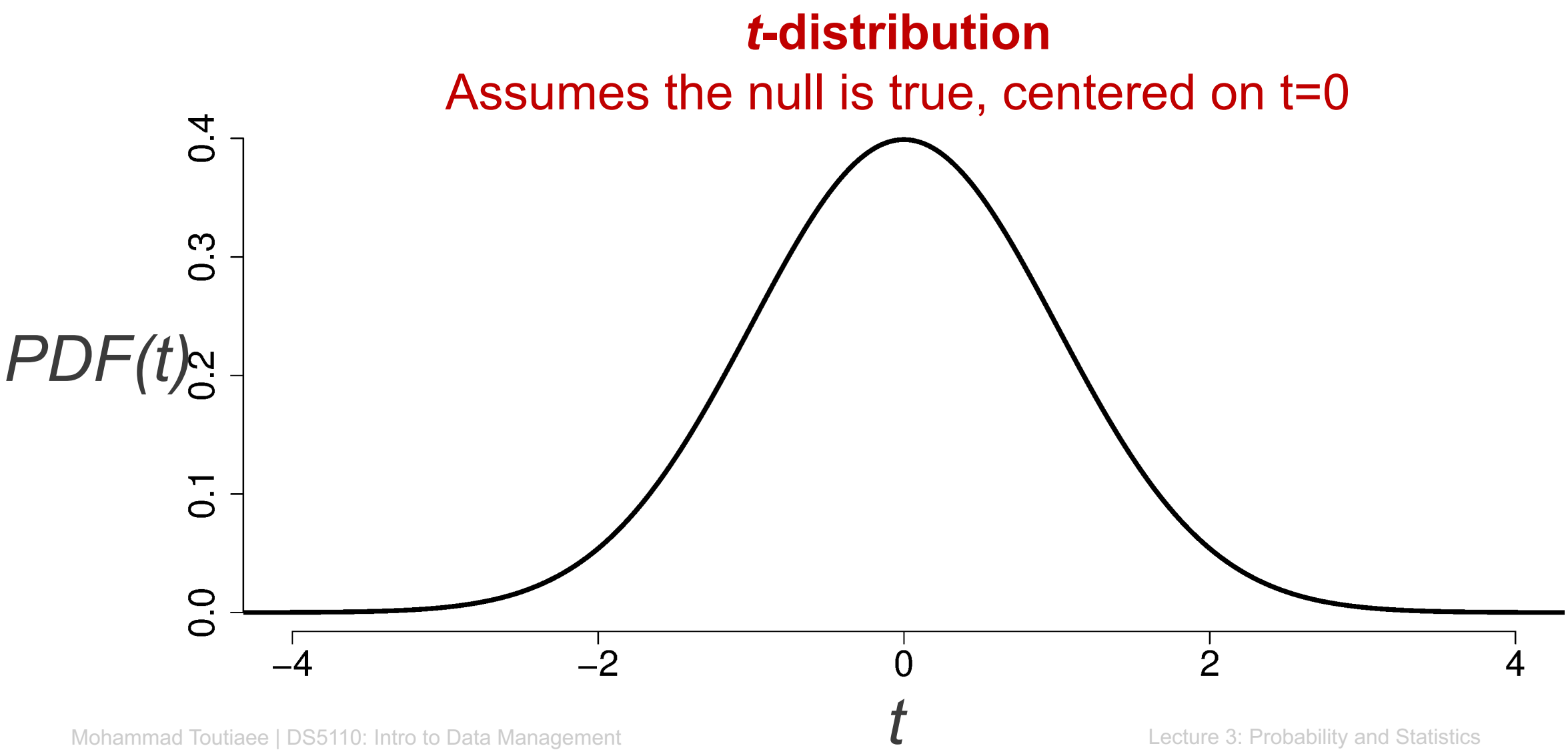
Was developed by William Gosset (under the pseudonym "*Student*") in 1908.

We use *t*-distribution in place of Gaussian when:

- the sample size $n$ is **too small** (usually less than 30), or
- the population standard deviation $\sigma$ is unknown.

The *t*-value is calculated via: $t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$

# *t*-Distribution; examples



# Sampling Distribution

- The sampling distribution of the mean, denoted by $\overline{X}$, is the distribution of sample means when taking random samples of the same size.
- The mean of the sample means, denoted by $\mu_{\overline{X}}$, is the population mean (i.e. $\mu_{\overline{X}} = \mu$).
- The standard error (SE) is the standard deviation of the sampling distribution, denoted by $\sigma_{\overline{X}}$, when sampling with replacement (i.e. $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$).
- The standard deviation requires a correction factor when sampling without replacement. This correction factor is $\sqrt{\frac{N-n}{N-1}}$, where N is the population size.

# Central Limit Theorem

- The distribution of sample means can be assumed to be approximately normal because of a powerful result of the Central Limit Theorem (CLT).

- The CLT is the basis for assuming averages and totals follow the normal distribution and underlies many of the tests and results used in data analysis.

- The CLT states that as the sample size drawn from the population with distribution $X$ becomes larger, the sampling distribution of the means $\overline{X}$ approaches that of a normal distribution $\mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$

# Central Limit Theorem Assumptions

- **Randomness assumption:** samples must be randomly selected.
- **Independence condition:** sample values must be independent from each other.
- **Sample size assumption:** sample size must be large enough. A rule of thumb is that sample sizes should be at least 30.

# Central Limit Theorem for $z$-score

- The $z$-score calculated via CLT is:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

# Central Limit Theorem for Sample Proportion

- The Central Limit Theorem for proportions states that if $\mathrm{X} \sim \mathcal{B}(\mathrm{n}, \mathrm{p})$ where $n$ is the number of trials and $\mathrm{p}$ is the probability of success, then the sampling distribution for proportions $\hat{\mathrm{p}}$ follows a normal distribution $\mathcal{N}(\mathrm{p}, \sqrt{\frac{\mathrm{p}(1-\mathrm{p})}{\mathrm{n}}})$.
- The $z$-score in this case is:

$$z = \frac{\hat{\mathrm{p}} - \mathrm{p}}{\sqrt{\frac{\mathrm{p}(1-\mathrm{p})}{\mathrm{n}}}}, \text{ where } \mathrm{np}, \mathrm{n}(1-\mathrm{p}) \geq 5$$

# P-value

Technical definition:

- When you perform a hypothesis test in statistics, the p-value is the probability of the observed outcome or something more extreme than the observed outcome, computed under the assumption that the null hypothesis is true.
- The lower the p-value, the more confident we are that the alternate hypothesis is true.

Non-technical definition:

- It quantifies surprise in a universe when there is no effect or no change.

# Type I and Type II Errors

- A hypothesis test decides between rejecting or not rejecting the null hypothesis. The **significance level**, $\alpha$, determines when to reject based on the p-value. Reject if p-value ≤ $\alpha$, otherwise fail to reject.

- In reality, either the null or alternative hypothesis is true. A test's conclusion is either right or wrong:
  - A **type I** error is rejecting the null hypothesis in favor of the alternative when in reality the null hypothesis is true.
  - A **type II** error is failing to reject the null hypothesis when in reality the alternative hypothesis is true.

# Type I and Type II Errors

- A **type I** error is rejecting the null hypothesis in favor of the alternative when in reality the null hypothesis is true.
- A **type II** error is failing to reject the null hypothesis when in reality the alternative hypothesis is true.

**Decision errors**

| | | Reality | |
|---|---|---|---|
| | | $H_0$ true | $H_a$ true |
| Hypothesis test decision | p-value $\leq \alpha$<br>Reject $H_0$ | Type I error | Correct |
| | p-value $> \alpha$<br>Fail to reject $H_0$ | Correct | Type II error |

**Example**

p-value $\approx 0 \leq \alpha = 0.05$ Reject $H_0 : \pi = 0.25$ Correct or Type I error

# Estimation

- Confidence Interval: is a likely range around a sample statistic showing the real population value. Ex., a 95% confidence interval of 160 cm to 170 cm for average student height means the true height is likely in this range.

- Margin of Error: is how much a sample statistic could differ from the true population value, considering data variation and confidence. Ex. a poll with 45% support and a 3% margin of error means the candidate's actual support might range from 42% to 48% due to poll uncertainty.

- Relation: Confidence intervals are based on sample stats like means or proportions, their width set by margin of error. Ex., a 60% tea preference with 4% margin of error means the real tea lover proportion probably lies between 56% and 64% with 95% confidence.


# Case Study

Flight Delays


# Next Lecture

SQL for Data Science

# Hypothesis Testing

Supplementary Notes

# Introduction

- What is Hypothesis Testing?
  - Hypothesis testing is a fundamental method in statistics that helps us determine if there is a significant difference or effect in our data. It's like being a detective, trying to find evidence to support a claim or hypothesis.

- What are P-Values?
  - P-values are like the clues in our detective work. They tell us how strong the evidence is against a claim we are testing. If you find strong clues, you can be more confident in your conclusions.

# Hypothesis Testing

- We use hypothesis testing to make decisions about data.
  - When we collect data, we often have a hypothesis or educated guess about what it means.
  - Hypothesis testing helps us decide if our data supports this guess or if it's just a coincidence.

- It helps us answer questions like 'Is there a difference?' or 'Is there an effect?
  - Think of hypothesis testing as a way to answer questions such as, "Does a new medicine work better than the old one?" or "Is there a significant change in sales after a marketing campaign?"
  - It gives us a structured approach to find answers backed by evidence.

# The Two Hypotheses Testing

- Null Hypothesis ($\mathbf{H_0}$): Assumes no effect or difference.
- Alternative Hypothesis ($\mathbf{H_a}$): Assumes an effect or difference.

# The Two Hypotheses Testing; Example

- Example 1: Medical Research
  - Research Question: Does a new drug reduce cholesterol levels in patients?
  - H0: The new drug has no effect on reducing cholestrol levels in patients.
  - Ha: The new drug reduces cholestrol levels in patients. one-sided

    The new drug has effect on reducing cholestrol. two-sided

    (+, -)

# The Two Hypotheses Testing; Example

- Example 2: Marketing Campaign
  - Research Question: Does a new advertising campaign increase product sales?
  - H0: The new ads campaign has no effect on product sales
  - Ha: $\neg H_0$

# The Two Hypotheses Testing; Example

- Example 3: Education
  - Research Question: Does a new teaching method improve student test scores?
  - H0:
  - Ha:

# The Two Hypotheses Testing; Example

- Example 4: Manufacturing
  - Research Question: Does a change in the manufacturing process reduce defects in a product?
  - H0:
  - Ha:

# P-value

- The p-value measures the strength of evidence against the null hypothesis.
- It tells us how likely we'd see our data if the null hypothesis were true.
- (High Surprise, Low false positive) Small P-Value (e.g., < 0.05): Strong evidence against the null hypothesis. We reject the null hypothesis.
- (Low Surprise, High false positive) Large P-Value (e.g., > 0.05): Weak evidence against the null hypothesis. We fail to reject the null hypothesis.

# P-value Computation; Exact Solution

| Dog ID | Date | Behavior | New dog | Aggressive | |
|---|---|---|---|---|---|
| 54672 | 3/2 | grooming | no | no | (no, no) |
| 43284 | 3/2 | aggression | no | yes | (no, yes) →how likely this happens? (p-value) |
| 12114 | 3/6 | sleeping | no | no | (no, no) |

H0: Old dog is not aggressive
Ha: Old dog is aggressive

(no, no)
(no, yes) →how likely this happens? (p-value)
(no, no)

$$b(1; 3,1/2) = \binom{3}{1} .5^1 .5^{3-1} = \frac{3}{8} = 0.375$$

Decision: We fail to reject H0 (0.375 > 0.05)

Number of samples: 3 dogs
Number of old dogs that are aggressive: 1 sample

# P-value Computation; Exact Solution

| Dog ID | Date | Behavior | New dog | Aggressive |
|---|---|---|---|---|
| 54672 | 3/2 | grooming | no | no |
| 43284 | 3/2 | aggression | no | yes |
| 12114 | 3/6 | sleeping | no | no |
| ... | ... | ... | ... | ... |
| 43284 | 5/18 | aggression | yes | yes |
| 87401 | 5/19 | aggression | yes | yes |
| 12114 | 5/19 | eating | yes | no |

H0: Old dog is not aggressive
Ha: Old dog is aggressive

$$b(1150; 10000, 1/2) = \binom{10000}{1150} .5^{1150} .5^{10000-1150} = ?$$

Computationally expensive!

Alternative solution: use approximation methods!

Number of samples: 10000 dogs
Number of old dogs that are aggressive: 1150 samples

# P-value Computation; Approximation Methods

**(DS5020/5110) Z-Test or T-Test:** These tests are used for comparing means in a sample to a known population or between two samples. You calculate a test statistic (Z or t) and then find the corresponding p-value from a standard normal distribution or t-distribution, respectively.

**(DS5020) Chi-Square Test:** Used for categorical data analysis, such as testing independence between variables in a contingency table. The p-value is calculated by comparing the observed and expected frequencies under the null hypothesis.

**(DS5020) ANOVA (Analysis of Variance):** ANOVA is used to test differences between three or more groups. The p-value is based on the F-statistic, which measures the ratio of variation between group means to variation within groups.

**(DS5110) Regression Analysis:** In linear regression, the p-value associated with each coefficient tests the null hypothesis that the coefficient is equal to zero. For logistic regression, it tests the null hypothesis that the odds ratio is equal to one.

**Nonparametric Tests:** Tests like the Wilcoxon-Mann-Whitney U test or the Kruskal-Wallis test are used when assumptions for param