Statistics Part II

Outline

- Basic Probability Theory
- Bayes' Theorem
- Hypotheses and Statistical Tests

Basic Probability Theory

Basic Probability Theory

Probability is the measure of the likelihood that an event will occur. ***Probability*** quantifies as a number between 0 and 1, where, loosely speaking, 0 indicates impossibility and 1 indicates certainty. The higher the probability of an event, the more likely it is that the event will occur. [wikipedia]

Notation and Probability Axioms

we write $P(E) \in [0, 1]$ to denote the ***probability*** of event E. $P(E)$ follows the following axioms:

- $P(E) \in \mathbb{R}$, and $P(E) \geq 0$ for all possible E
- $P(\Omega) = 1$: it is certain that at least one event will occur
- $P\left(\cup_i E_i\right) = \sum_i P(E_i)$

Dependence of Events

In probability theory, two events are independent, statistically independent, or stochastically independent if the occurrence of one does not affect the probability of occurrence of the other (equivalently, does not affect the odds).

Dependence of Events

In probability theory, two events are independent, statistically independent, or stochastically independent if the occurrence of one does not affect the probability of occurrence of the other (equivalently, does not affect the odds).

- we write $E_1 \perp E_1$ for ***independent*** events
- $P(E_1, E_2) = P(E_1)P(E_2)$

Dependence of Events

In probability theory, two events are independent, statistically independent, or stochastically independent if the occurrence of one does not affect the probability of occurrence of the other (equivalently, does not affect the odds).

- we write $E_1 \perp E_1$ for ***independent*** events
- $P(E_1, E_2) = P(E_1)P(E_2)$

→ **multiplication** of probabilities if events are independent!

Example 1: Coin flip

Example 1: Coin flip

Example 2: Sex of children

How to compute $P(E_1, E_2)$ if events are dependent, e.g. E_1 depends on E_2 ?

How to compute $P(E_1, E_2)$ if events are dependent, e.g. E_1 depends on E_2 ?

Conditional Probability

$$P(E_1, E_2) = P(E_1|E_2)P(E_2)$$

where $P(E_1|E_2)$ is the ***conditional*** probability of event E_1 , aka the probability that E_1 occurs **after** $P(E_2) = 1$

How to compute $P(E_1, E_2)$ if events are dependent, e.g. E_1 depends on E_2 ?

Conditional Probability

$$P(E_1, E_2) = P(E_1|E_2)P(E_2)$$

where $P(E_1|E_2)$ is the ***conditional*** probability of event E_1, aka the probability that E_1 occurs **after** $P(E_2) = 1$

This results in:

- $P(E_1|E_2) = P(E_1)$ if events are independent
- $P(E_1|E_2) = P(E_1, E_2) \div P(E_2)$

Probability can be quite counter intuitive:

Examples based on child sex (assuming 50% chance for a boy or girl an independence for several children):

- **Example 1**: What is the probability of both children to be girls $P(B) := 0.25$, given the conditional event that the first child is a Girl $P(G) := 0.5$

Probability can be quite counter intuitive:

Examples based on child sex (assuming 50% chance for a boy or girl an independence for several children):

- **Example 1**: What is the probability of both children to be girls $P(B) := 0.25$, given the conditional event that the first child is a Girl $P(G) := 0.5$

$$P(B|G) = P(B, G) \div P(G) = P(B) \div P(G) = 0.25 \div 0.5 = 0.5$$

Hochschule Offenburg
offenburg.university

Probability can be quite counter intuitive:

Examples based on child sex (assuming 50% chance for a boy or girl an independence for several children):

- **Example 2**: What is the probability of both children to be girls $P(B) := 0.25$, given the conditional event that at least one child is a girl $P(L) := 0.75$

Probability can be quite counter intuitive:

Examples based on child sex (assuming 50% chance for a boy or girl an independence for several children):

- **Example 2**: What is the probability of both children to be girls $P(B) := 0.25$, given the conditional event that at least one child is a girl $P(L) := 0.75$

$$P(B|L) = P(B, L) \div P(L) = P(B) \div P(L) = 0.25 \div 0.75 = 0.333$$

Bayes' Theorem

Bayes' Theorem

In probability theory and statistics, ***Bayes' theorem*** describes the ***probability*** of an event, based on ***prior knowledge*** of conditions that might be related to the event.

Bayes' Theorem

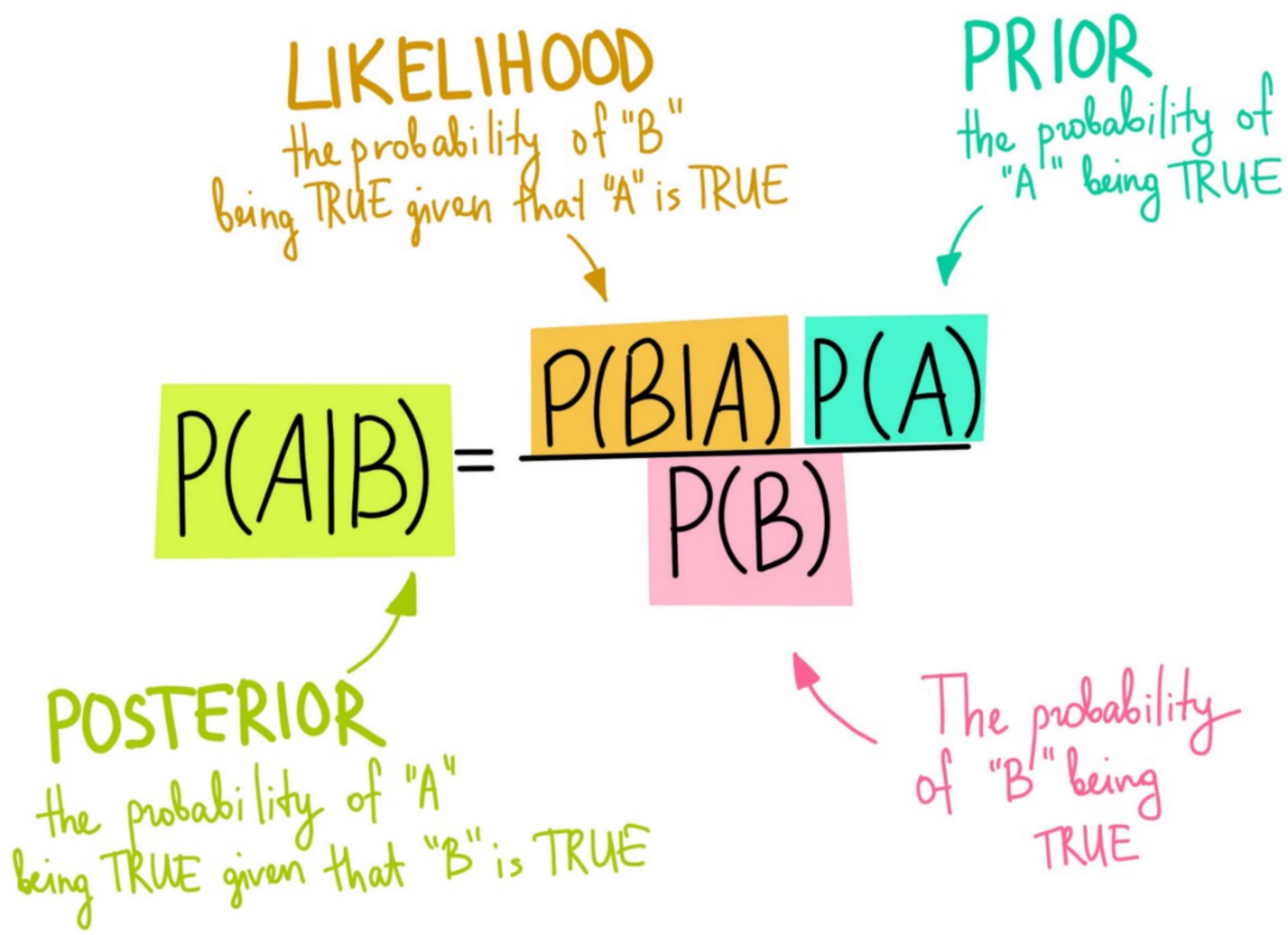
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \text{ for } P(B) > 0$$

Bayes' Theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \text{ for } P(B) > 0$$

Interpretation:

- reversing the deduction of a conditional probability
- a feature often needed in inference (machine learning) settings (later more)


LIKELIHOOD
the probability of "B" being TRUE given that "A" is TRUE
PRIOR
the probability of "A" being TRUE
P(A|B) = P(B|A) P(A) / P(B)
POSTERIOR
the probability of "A" being TRUE given that "B" is TRUE
The probability of "B" being TRUE

Example:

Suppose that a test for using a particular drug is 99% sensitive and 99% specific. That is, the test will produce 99% true positive results for drug users and 99% true negative results for non-drug users. Suppose that 0.5% of people are users of the drug. What is the probability that a randomly selected individual with a positive test is a drug user?

Example:

Suppose that a test for using a particular drug is 99% sensitive and 99% specific. That is, the test will produce 99% true positive results for drug users and 99% true negative results for non-drug users. Suppose that 0.5% of people are users of the drug. What is the probability that a randomly selected individual with a positive test is a drug user?

$$\begin{aligned}
P(\text{User} \mid +) &= \frac{P(+ \mid \text{User}) P(\text{User})}{P(+)} \\
&= \frac{P(+ \mid \text{User}) P(\text{User})}{P(+ \mid \text{User}) P(\text{User}) + P(+ \mid \text{Non-user}) P(\text{Non-user})} \\
&= \frac{0.99 \times 0.005}{0.99 \times 0.005 + 0.01 \times 0.995} \\
&\approx 33.2\%
\end{aligned}$$

Experiments: Hypotheses and Statistical Tests

Ultimately, we use **statistics** and analyze **probabilities** in order to **draw conclusions**. A common approach in statistics is to start an experiment with a ***hypotheses***, which is then validated by a statistical ***test***.

Statistical Inference Pipeline

- Formulate Hypotheses
- Design Experiment
- Collect Data
- Test / draw conclusions

Null-Hypotheses

- The statement being tested in a test of statistical significance is called the ***null hypothesis***. The test of significance is designed to assess the strength of the evidence against the null hypothesis. Usually, the null hypothesis is a statement of ***'no effect'*** or ***'no difference'***.
  It is often symbolized as H_0.
- The statement that is being tested against the null hypothesis is the alternative hypothesis H_1.
- ***Statistical significance test***: Very roughly, the procedure for deciding goes like this: Take a random sample from the population. If the sample data are consistent with the null hypothesis, then do not reject the null hypothesis; if the sample data are inconsistent with the null hypothesis, then reject the null hypothesis and conclude that the alternative hypothesis is true.

Significant Tests

- p-Value (or significance): is, for a given statistical model, the probability that, when the null hypothesis is true, the statistical **test summary** would be greater than or equal to the actual observed results.
- at experiment design, a significance threshold α is chosen
- typically, $\alpha = 0.05$ for scientific experiments

t-Test

The ***t-Test*** (also called Student's t-Test) compares two averages (means) and tells you if they are different from each other. The t-Test also tells you how significant the differences are; In other words it lets you know if those differences could have happened by chance.

t-Test

The ***t-Test*** (also called Student's t-Test) compares two averages (means) and tells you if they are different from each other. The t-Test also tells you how significant the differences are; In other words it lets you know if those differences could have happened by chance.

$t = \sqrt{n}\frac{\bar{X}-\mu_0}{\sigma}$, where

- σ standard deviation (from n samples)
- $\bar{X}$ sampled mean from n samples
- μ_0 mean hypotheses

In [11]:
```
#Sample Size
N = 1000000
#Gaussian distributed data with mean = 0 and var = 0.1
a = np.random.normal(0,0.3,N)
#Gaussian distributed data with with mean = 0 and var = 0.2
b = np.random.normal(-0.0001,0.1,N)
plt.hist(a,bins=100)
plt.hist(b,bins=100)
## Cross Checking with the internal scipy function
t2, p2 = stats.ttest_ind(a,b)
print("means: ", a.mean(), b.mean())
print("t = " + str(t2))
print("p = " + str(p2))
```

```
means:  0.00026424482988052976 -3.7931362338061 98e-07
t = 0.836169199886757
p = 0.40305982287283526
```