Statistics
Sampling
And
Hypothesis Testing

Sum of IID RV

- Suppose we flip a coin n times. Coin's bias is $p = P(heads)$
- Define $X_1, X_2, \ldots, X_n$ to be independent binary random variables (IID RV) such that $P(X_i = 1) = p,\ P(X_i = 0) = 1 - p$
- Define the average $Y_n = \frac{1}{n}\sum_{i=1}^{n} X_i$
- We can use Y_n as an estimate of p because:
- We have shown that $\mu = E(Y_n) = E(X_i) = p;\ \ \sigma = \sqrt{Var(Y_n)} = \sqrt{\frac{p(1-p)}{n}}$
- Using Chebyshev we get

$$P(|Y_n - p| > k\sigma) \leq \frac{1}{k^2}$$

Central Limit Theorem

Let $X_1, X_2, \ldots, X_n$ be IID Random variables with common mean μ and variance σ^2

Define: $$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

Then the CDF of Z_n converges to the standard normal CDF:

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2} dx,$$

In the sense that

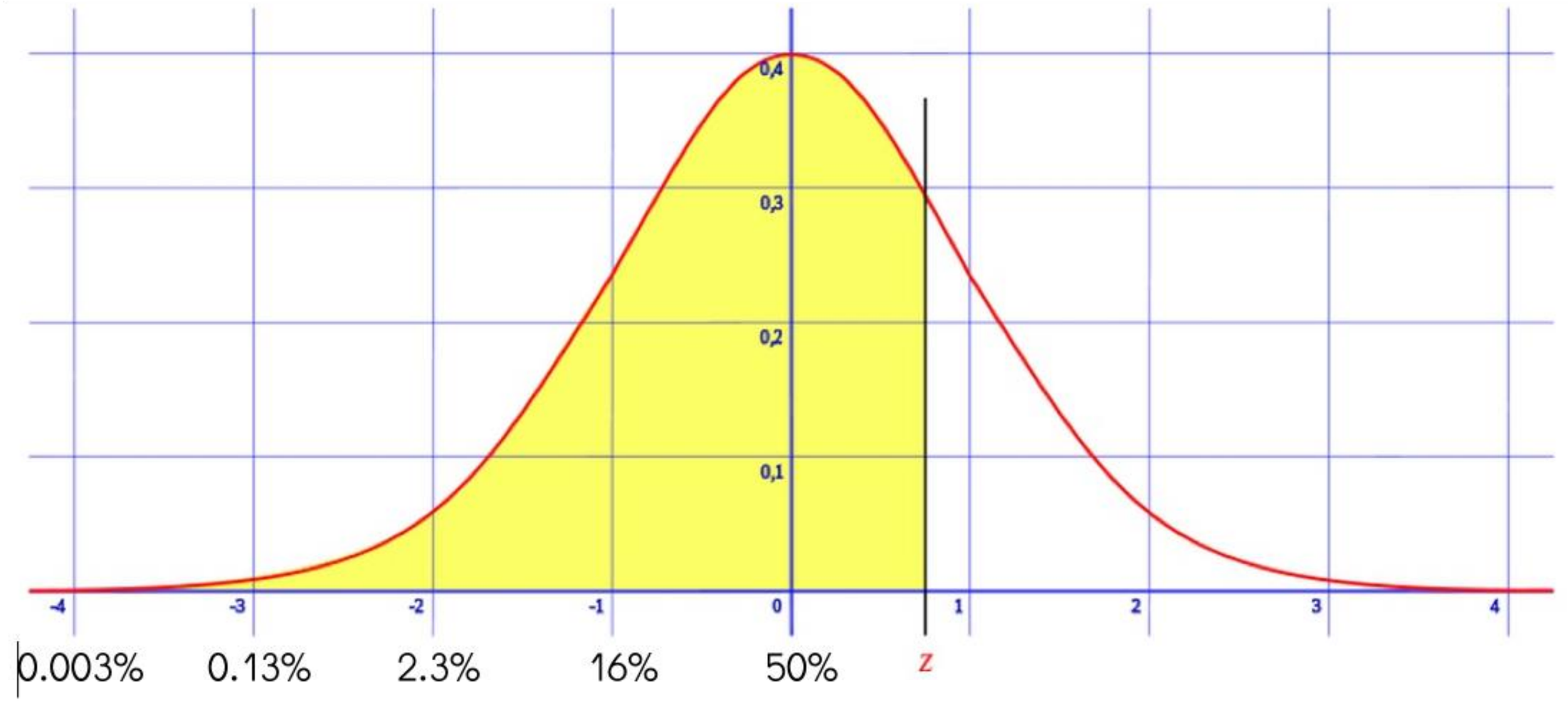
$$\forall z, \quad \lim_{n \to \infty} P\left(Z_n \leq z\right) = \Phi(z)$$

The central limit theorem is a strong justification for assuming that a distribution is normal.

Assuming normality is very common in practice.

Gives rise to the common use of Z-scores and Z-tables.

$$Z = \frac{X - \mathrm{E}[X]}{\sigma(X)}$$



- A few standard definitions: $Q(z) = P(X > z)$, $\Phi(z) = P(X < z)$, $Q^{-1}(p)$ is the inverse function to Q. In other words: $Q(Q^{-1}(p)) = p$.

- few useful values:

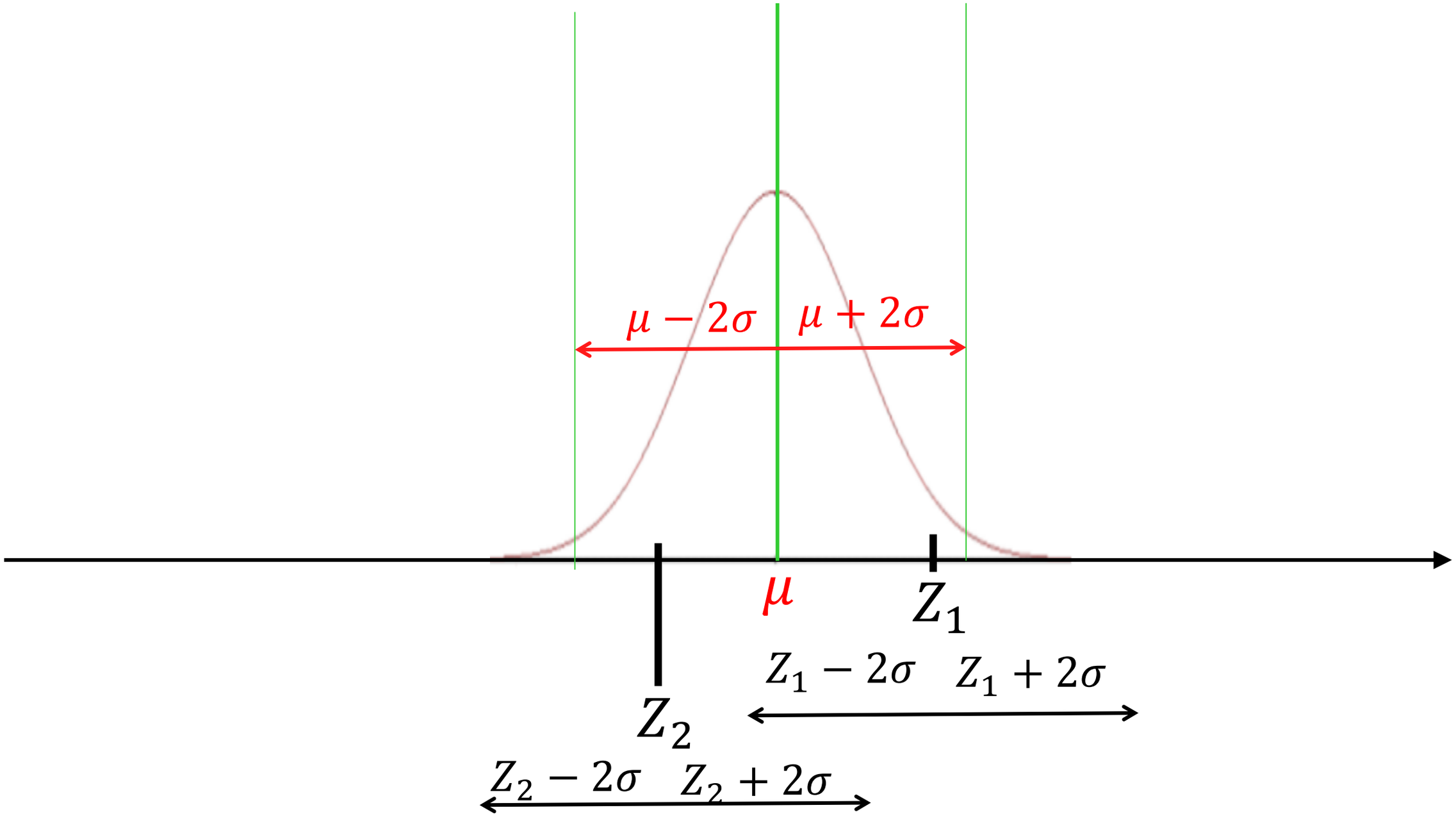
$$Q(1) \approx 15\%, \quad Q(2) \approx 2.5\%,$$

$$Q(3) \approx 0.15\%, \quad Q(4) \approx 0.003\%$$

Probability vs. Statistics

- In **probability** calculations we start from true probabilities and we calculate: Probabilities of events, conditional probabilities, Expected values, variance, correlation etc.
- In **statistical inference** we start from measurements of empirical probabilities, empirical mean, empirical correlations etc and make inferences about the true probabilities, true mean, true correlations etc.
  - Statistical inferences never have 100% certainty. There is always some probability that the inferences are wrong.
  - That probability (or one minus this probability) is called the "confidence level" of the statement.
  - Even regarding a simple property, such as the mean, there are many different inferences we can make.

Confidence interval inference



With probability ~95% the actual expected value μ is within $\pm 2\sigma$ of the observed value Z

Example application of Hypothesis testing

We want to prove that seat-belts save a significant number of lives.

Suppose first that we know that, in general, the probability of a fatality in a car accident is $q=1\%$. (one out of 100 car accidents is fatal, ignoring whether or not seat-belts were used)

We examine $n=1000$ randomy selected records of accidents where the driver wore a car seat. We find that $k=5$ of these accidents were fatal.

With what confidence can we conclude that claim is correct?

X_i = i'th accident was fatal.

The null hypothesis is that seat belts make no difference.

P(X_i=1)=q=0.01

Alternative hypothesis, P(X_i=1)<q

$$S_n = \sum_{i=1}^{n} X_i;$$

$$\mu = E(S_n) = nq = 1000 \times 0.010 = 10$$

$$\text{var}(S_n) = nq(1-q) \approx nq = 10$$

$$\sigma(S_n) = \sqrt{10} \approx 3.16$$

Therefore z-value = (10-5)/3.16= 1.58,
and the probability of getting 5 or higher is Q(1.58) = 5.94% ~ 6%

The p-value is 6%

The p value is a *random variable*

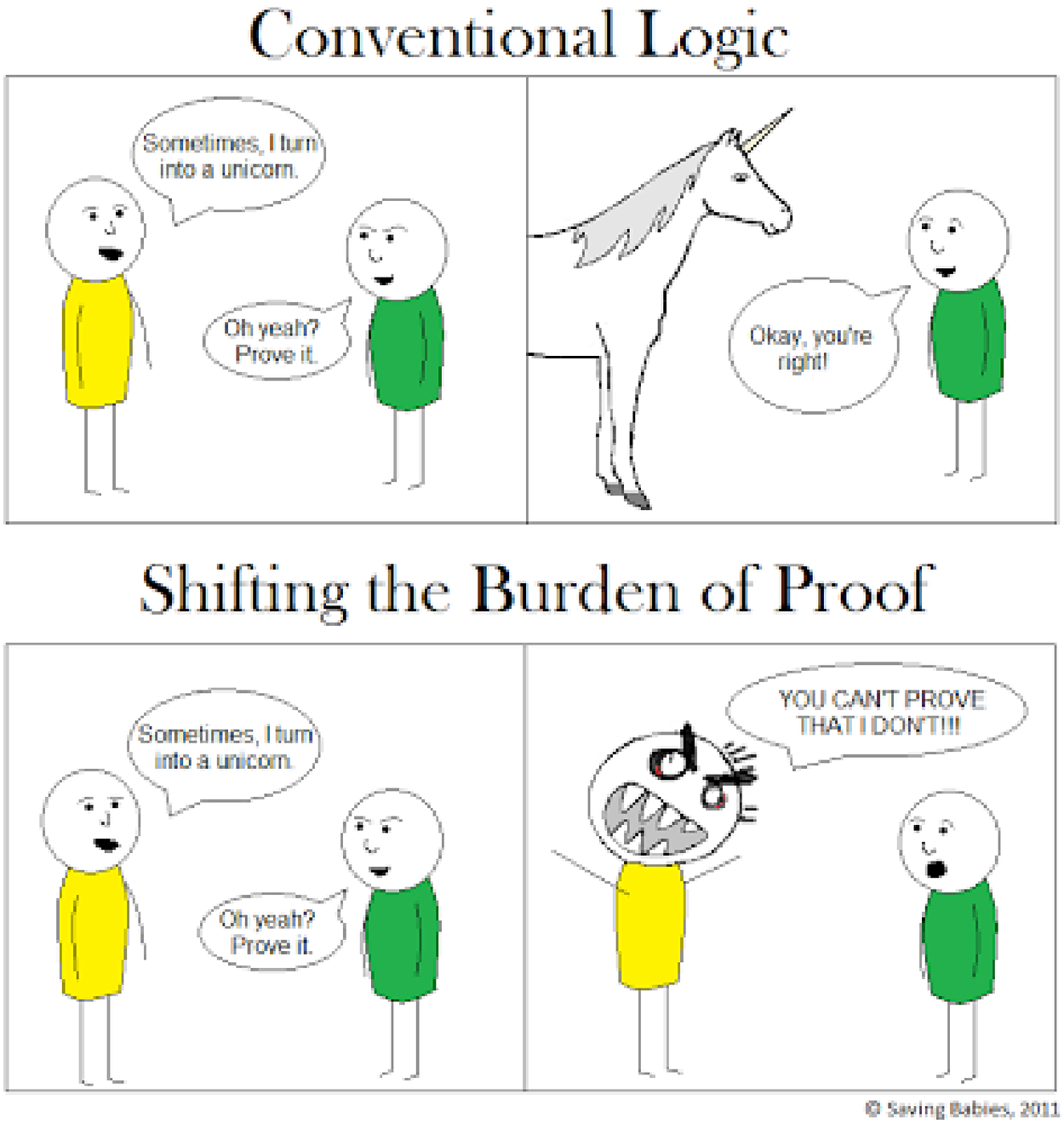
Assymetry in decisions

- In some decisions there is no a-priori bias:
  - which person runs faster
  - Which party will win the upcoming election
- In others, there is a strong bias to one side:
  - In court: innocent until proven guilty
  - In Science: Current theory holds until proven incorrect.
  - In medicine: A new treatment is assumed ineffective until clinical trials prove that it is effective.
- When there is a strong bias, one side carries "burden of proof".

Burden of Proof

- Legal: The burden of proof (Latin: onus probandi) is the imperative on a party in a trial to produce the evidence that will shift the conclusion away from the default position to one's own position.
  - Example: Innocent until proven guilty.
- Statistical: The imperative on a scientist arguing for a new theory to provide sufficient evident to reject the Null Hypothesis = the prevailing theory and establish his own theory: the Alternative Hypothesis.
  - Example: proving that the police is guilty of racial profiling.

Burden of Proof



Is 6% small enough? maybe yes maybe no, how is this decided?

Suppose that an important decision needs to be made, for example, to increase the fine for not wearing a seat-belt. We need to choose a significance level (α). And this choice has to be made before looking at the data. It should NOT be a random variable.

If $p \leq \alpha$ we say that the test rejected the null hypothesis

If $p > \alpha$ we say that the test failed to reject the null hypothesis.

In other words, if we chose α =0.05, then the test failed. We did not show that seatbelt save lives.

One might hope that in this case we gained some evidence that seat-belts do not save lives. Unfortunately, that is not the case, we did NOT gain evidence that seat-belts are useless. We are in a situation in which no significant conclusions can be drawn.

* A test cannot provide evidence towards the null hypothesis.
* We did not disprove that belts save lives.

Are these details important?

Yes!
Statistical tests are at the foundation of the scientific method, medicine, and public policy.

Scientific method = repeatability of experiments. We need to decide how many successful repetitions are needed to be convinced.

Medicine = The most expensive part of drug development are the human trials.

What does it mean that the standard value of alpha used in medical journals is 5% ?

Public policy = Seat-belts? What level of chemicals in public water deems it unsafe?

Gullability

Fact: most articles published in medical journals use an alpha value of 0.05=%5

The hypothesis testing protocol

1. Define null hypothesis and alternative hypothesis. The null hypothesis represents the default or prevailing opinion which you want to overturn. the alternative hypothesis represents your opinion/belief.
   Both hypotheses are distributions over experimental outcomes.

2. Define an experiment and a statistical test:
   test: experimental outcome --> score.

3. Compute the p-value of each score.
   p(S) = prob. that a random score > S under the null hypothesis distribution.
4. Decide on a value of alpha - smaller - more convincing, larger - higher chance of rejecting the null.
5. Run experiment
6. Reject null hypothesis if p<alpha, else experiment failed.

Hypothesis Testing Protocol for the effect of Seat-Belts

1. Define null hypothesis and alternative hypothesis. The null hypothesis represents the default or prevailing opinion which you want to overturn. the alternative hypothesis represents your opinion/belief.
   Both hypotheses are distributions over experimental outcomes.

2. Define an experiment and a statistical test:
   test: experimental outcome --> score.

3. Compute the p-value of each score.
   p(S) = prob. that a random score > S under the null hypothesis distribution.
4. Decide on a value of alpha - smaller - more convincing, larger - higher chance of rejecting the null.
5. Run experiment
6. Reject null hypothesis if p<alpha, else experiment failed.

1. Null hypothesis: probability of a fatality in an accident is q=1%, whether or not you wear a seat-belt.
   alternative hypothesis: seat belts reduce chance of fatality.
2. Experiment: Collect 1,000 records of experiments in which seat-belts were used.
3. $p(S) = Q\left(\frac{nq - S}{\sqrt{nq(1-q)}}\right)$
4. alpha=5%
5. outcome was S=7
6. p=6%, null hypothesis not rejected, hypothesis failed

What can statistics can prove?

Can

* Driving under the influence increases the chance of an accident
* Driving under the influence does not increase the chance of an accident by more than 2%.
* Members of the Kalenjin tribe run faster than the average.
* $E(X)>2$
* $E(X)<7$

Cannot

* Driving under the influence does not change the chance of an accident.
* The probability of an accident when DUI is 1.2%
* $E(X)=7$
* $P(X=3)=0.23$
* Prove the null hypothesis

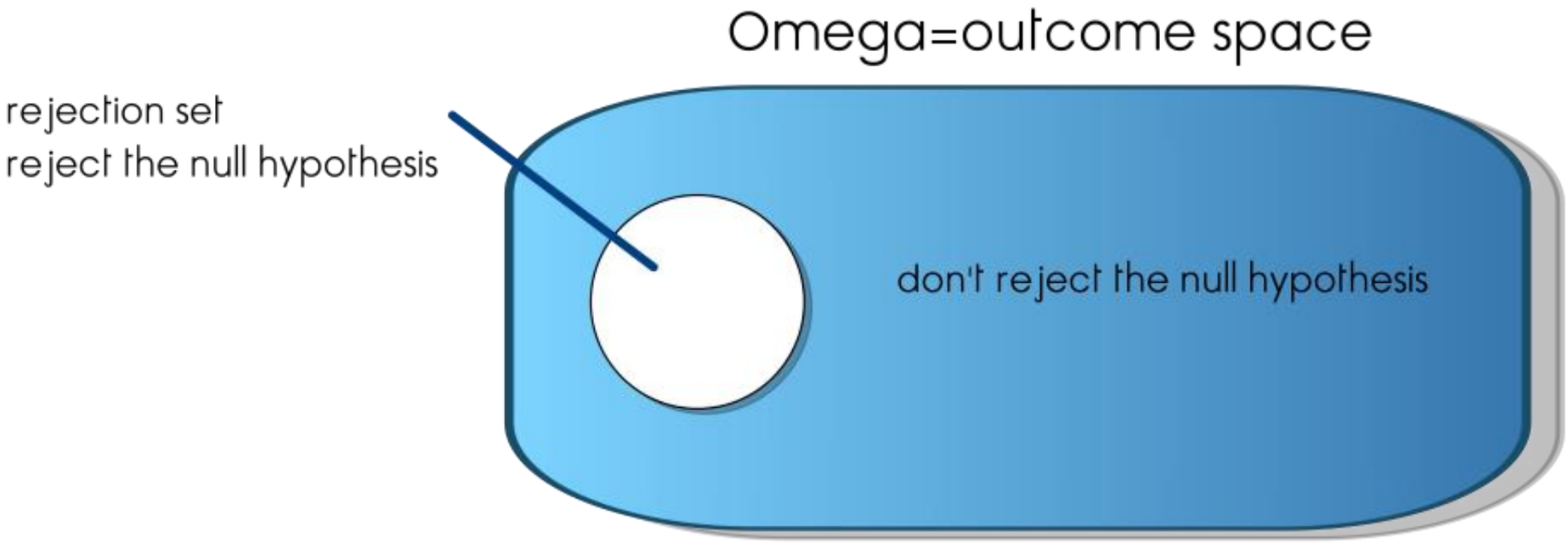
Choosing alpha is a compromise between two types of errors:
Type I error: Rejecting the null hypothesis when it is correct
Type II error: Failing to reject the null hypothesis when it is incorrect.

	H_0 true Seatbelts don't help	H_1 true Seatbelts help
Fail to reject	+	type II
Reject Null	Type I	+

Question: Increasing alpha:

A Increases Type II error, Decreases type I
B Increases Type I Error, Decreases type II
C Decreases both.
D Increases both.

The probabilty theory of statistical tests



1. The each point in the outcome space corresponds to outcomes of a complete experiment - we observe only one!
2. The white circle represents the set of outcomes that will cause us to reject the null hypothesis.
3. alpha = The probability of the rejection set under the distribution defined by the null hypothesis.

Designing a statistical test

- The statistical test is function that maps the experimental outcome to True/False:
  - True: reject the null hypothesis
  - False: don't reject the null hypothesis.
- The "reject set" is the set of outcomes on which the test is True.
- The null hypothesis is a distribution over outcomes.
- The probability assigned to the reject set when the null hypothesis is the true distribution is at most α
- The alternative hypothesis is also a distribution. In a well designed test the probability of the rejection set is large under the alternative hypothesis.
- There are many statistical tests, each for a particular pair of Null hypothesis / alternative hypothesis.

The internals of a statistical test

- Most statistical test are based on a **statistic**
- The statistic is a random variable that maps the experimental outcome to a real number..
- The distribution of the statistic under the null hypothesis is known.
- Large values of the statistic correspond to a large deviation from the null hypothesis.
- Computing the p-value: Compute the tail of the distribution that is larger or equal to the value attained by the statistic.
- Given α the test rejects the null hypothesis if $p < \alpha$ otherwise the null hypothesis is not rejected (the experiment failed).

Tests, dependence, correlation and causation

correlation vs. dependence

- If X, Y are independent then $Cov(X, Y) = 0$ and $Corr(X, Y) = 0$
- If $Corr(X, Y) \neq 0$ then X, Y are dependent.
- No implications in the opposite directions
- If X is a random variable that takes n discrete values, and Y is a random variable that takes m discrete values, Then checking for independence requires checking nm values.
- Checking for zero correlation requires calculation of just one value.
- Correlation is the quick and dirty way to detect strong dependencies, but it cannot find them all.

Correlation vs. Dependence

- Non-zero Correlation implies dependence
- Dependence does not imply correlation

Conditioning vs Causation

- Causation is directions mosquitoes cause malaria, malaria does not cause mosquitoes.
- Correlation is non-directional: $Cov(X, Y) = Cov(Y, X)$
  - Correlation does not imply causation
- Conditioning seems directional, but this is superficial:

$$P(A \wedge B) \neq P(A)P(B) \Leftrightarrow P(A|B) \neq P(A) \Leftrightarrow P(B|A) \neq P(B)$$

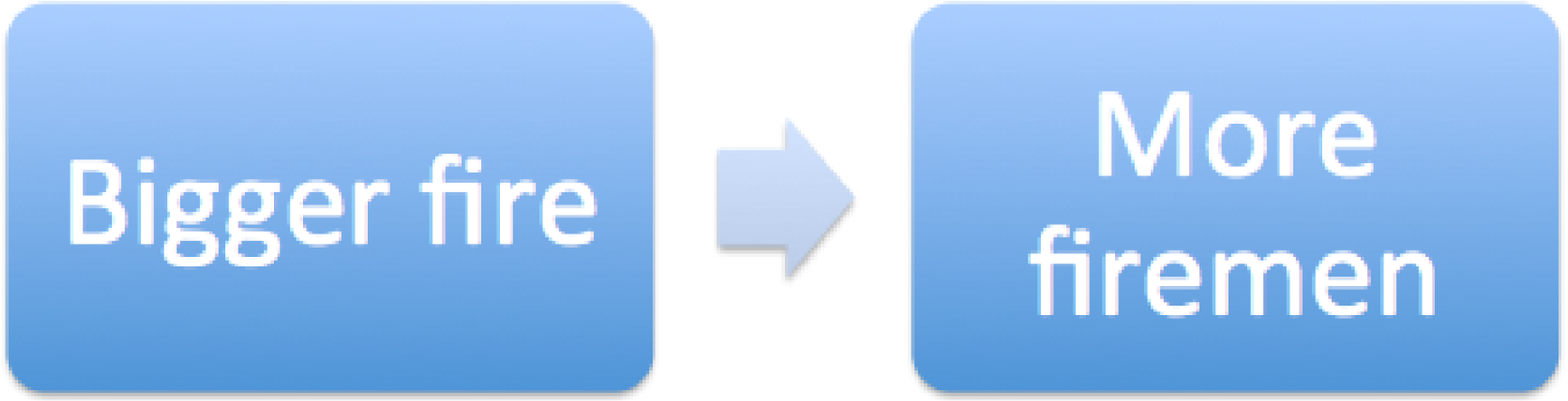
  - Dependence does not implyu causation

Correlation vs Causation

- Using correlation because common, same can be said regarding Dependence vs. causation.
- The simple case is: the number of mosquitoes is correlated with the number of malaria cases. Therefor mosquitoes cause malaria. Which is true.
- However, one can deduce that malaria causes mosquitoes, which is false.

Correlation vs. causation 1

- The more firemen fighting a fire, the bigger the fire.
- Therefore firemen cause an increase in the size of a fire.



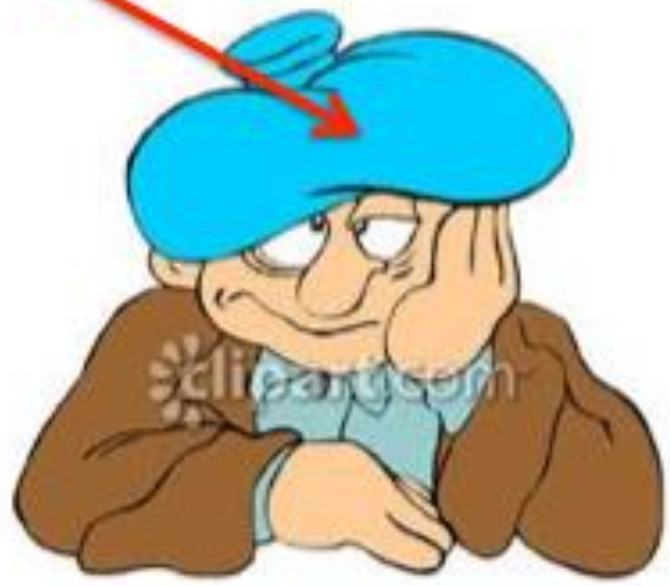
- Causation reversal. Correlation cannot distinguish between **A causes B** and **B causes A**

Correlation vs. causation 2

Excessive Drinking

Sleeping with shoes on

Common Cause



Morning Headache

Determining causation

- Can be very hard.
- Usually required intervention
- How can you do determine whether or not sleeping with shoes causes headaches?
  - A. Stop drinking.
  - B. Flip a coin to decide whether to wear shoes to bed.
  - C. Flip a coin to decide whether or not to drink.
  - D. Observe that every time you drank, you both slept with shoes and got up with a headache.

Correlation and Causation 3

- Suppose we want to know whether store brand (generic) medicine is worse than name brand medicine.
- Suppose we have access both to medical records and to drug-store records.
- We observe a correlation between consuming generic brand and the time until the patient gets well. In other words, the time to heal increases when taking generic brand relative to name brand.
- Can we deduce that name brand is better?
- Not necessarily. Alternative explanation: name brands are more expensive, therefor more wealthy people buy them, and more wealthy people tend to be more healthy and heal more quickly.
- We need to control for external variables such as wealth.

Clinical trials

- Developing a new drug takes 10-15 years and hundreds of millions of dollars.
- After tests on animals, the final stage is a clinical trial – a test on human patients.
- Very expensive: usually limited to a few hundred patients.
- Need to demonstrate **causality:** that taking the medication causes an improvement in the patient's health, and is not just correlated with it. Requires a **controlled study**:
  - **Placebo**
  - **Double blindness**
  - **Fixing the protocol before the start of the trial.**

Observational Studies

- With access to all electronic medical records, it is becoming possible to measure the effectiveness of a drug on millions of people (contrast with a few hundred in controlled studies).
- **Potential of revolutionizing medical research.**
- **Challenge: hard to control for non-causal correlations.**
- Challenge can be met by controlling for potential causes: wealth, age, race …

Counting fish

- Suppose we are studying a lake, and we want to estimate how many fish are in it.
- It is not realistic to try and catch all, or even most of them.
- Instead we follow a 3 step process:
  1. Catch m fish, mark them, and release back to the lake.
  2. Let some time pass, so that the marked fish mix with the unmarked.
  3. Catch l fish, count the number of marked fish, call that number the random variable Y.
- Let n be the number of fish in the lake. The probability that a random fish is marked is
  - m/n
- Y is the sum of l IID Binary random variables whose mean is m/n

Counting Fish continued

- Y is the sum of l IID Binary random variables whose mean is m/n E(Y)=?
  - $E(Y) = \frac{lm}{n}$
  - $Var(Y) = l\frac{m}{n}\left(1 - \frac{m}{n}\right) \leq$
    - $\leq l\frac{1}{2}\frac{1}{2} = \frac{l}{4}$; $\sigma(Y) = \frac{\sqrt{l}}{2}$
- The 95% confidence interval for $\frac{lm}{n} = \mathrm{E(Y)}$ is
  - $[Y - \sqrt{l}\,, Y + \sqrt{l}\,]$
- Therefor the 95% confidence interval on the number of fish is
  - $\left[\frac{lm}{Y+\sqrt{l}}, \frac{lm}{Y-\sqrt{l}}\right]$
- If l is too small then the estimate would be weak.
- If m is too small relative to n then we might not catch any marked fish,

Be a skeptic:

When you read that something has been proven statistically:

1. Ask what was the null hypothesis
2. Ask what was the statistical significance
3. Ask whether similar tests were performed that did not succeed.

Be a skeptic:

When you read that something has been proven statistically:

1. Ask what was the null hypothesis
2. Ask what was the statistical significance
3. Ask whether similar tests were performed that did not succeed.

Examples of common statistical tests

From the matlab Statistics module

ranksum	Wilcoxon rank sum test. Tests if two independent samples come from identical continuous distributions with equal medians, against the alternative that they do not have equal medians.
runstest	Runs test. Tests if a sequence of values comes in random order, against the alternative that the ordering is not random.
signrank	One-sample or paired-sample Wilcoxon signed rank test. Tests if a sample comes from a continuous distribution symmetric about a specified median, against the alternative that it does not have that median.
signtest	One-sample or paired-sample sign test. Tests if a sample comes from an arbitrary continuous distribution with a specified median, against the alternative that it does not have that median.
ttest	One-sample or paired-sample *t*-test. Tests if a sample comes from a normal distribution with unknown variance and a specified mean, against the alternative that it does not have that mean.
ttest2	Two-sample *t*-test. Tests if two independent samples come from normal distributions with unknown but equal (or, optionally, unequal) variances and the same mean, against the alternative that the means are unequal.

One-sample t-test

`h = ttest(x)` performs a *t*-test of the null hypothesis that data in the vector `x` are a random sample from a normal distribution with mean 0 and unknown variance, against the alternative that the mean is not 0. The result of the test is returned in `h`. `h = 1` indicates a rejection of the null hypothesis at the 5% significance level. `h = 0` indicates a failure to reject the null hypothesis at the 5% significance level.

tests the mean
Assumes normality

High power test - can identify a small deviation from the mean using few samples.

two-sample t-test

`h = ttest2(x,y)` performs a *t*-test of the null hypothesis that data in the vectors `x` and `y` are independent random samples from normal distributions with equal means and equal but unknown variances, against the alternative that the means are not equal. The result of the test is returned in `h`. `h = 1` indicates a rejection of the null hypothesis at the 5% significance level. `h = 0` indicates a failure to reject the null hypothesis at the 5% significance level. `x` and `y` need not be vectors of the same length.

tests difference between mean
Assumes normality

Lilliefors test

Description

`h = lillietest(x)` performs a Lilliefors test of the default null hypothesis that the sample in vector `x` comes from a distribution in the normal family, against the alternative that it does not come from a normal distribution. The test returns the logical value `h = 1` if it rejects the null hypothesis at the 5% significance level, and `h = 0` if it cannot. The test treats `NaN` values in `x` as missing values, and ignores them.

alternative hypothesis is the complement of the null hypothesis.

Ansari bradley test

`h = ansaribradley(x,y)` performs an Ansari-Bradley test of the hypothesis that two independent samples, in the vectors `x` and `y`, come from the same distribution, against the alternative that they come from distributions that have the same median and shape but different dispersions (e.g. variances). The result is `h = 0` if the null hypothesis of identical distributions cannot be rejected at the 5% significance level, or `h = 1` if the null hypothesis can be rejected at the 5% level. `x` and `y` can have different lengths.

alternative hypothesis is not the complement of the null hypothesis.

Multiple Hypothesis testing

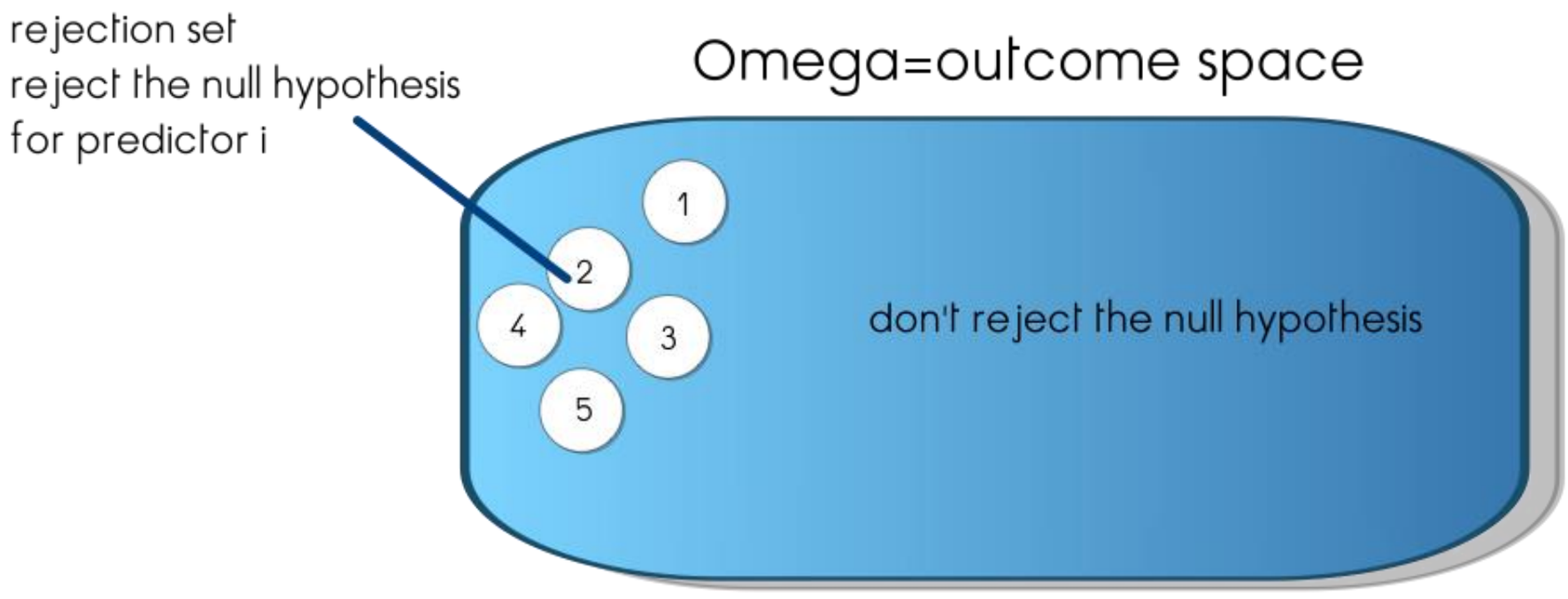
Consider the online ad problem, our goal is to maximize click-through rate. Our null hypothesis is that nothing performs better than picking one of the ads uniformly at random each time.

We have a large number of click-prediction algorithms. Each such algorithm takes as input information about the person, the web page and the ad and predicts the probability that the person will click on the ad.

We can go back in time and compute the expected number of errors each method would have made. We can use a statistical test to quantify the statistical significance of the performance of the method.

Suppose we have 100 methods and use an alpha value of 1%
Suppose for our data we found that one of the 100 methods rejects the null hypothesis at the 1% significance level. How sure can we be that the predictor that we found is better than random?

The probabilty theory of statistical tests



We don't know what would happen of different samples than the one we observe.
In the worst case the rejection sets are disjoint.

The Bonferroni correction for multiple-hypothesis testing:

If n statistical tests are performed using the same data
and the significance threshold used for all tests is α
Then the probability that at least one of the tests will
reject the null hypothesis can be as high as $n\alpha$

Be a skeptic:

When you read that something has been proven statistically:
1. Ask what was the null hypothesis
2. Ask what was the statistical significance
3. Ask whether similar tests were performed that did not succeed.