Correlation, Causation and Statistical tests

Covariance

Recall $\mu_X \doteq E(X),\ \mu_Y \doteq E(Y)$

$$\mathrm{Cov}(X,Y) = E((X-\mu_X)(Y-\mu_Y)) =$$

$$= E(XY) - \mu_X E(Y) - \mu_Y E(X) + \mu_X \mu_Y = E(XY) - E(X)E(Y)$$

Recall $\mathrm{Var}(X) = E(X^2) - E(X)^2 = \mathrm{Cov}(X,X)$

Cov(X,Y)≠0 implies that X and Y are not independent

but Cov(X,Y)=0 **does not imply** that X and Y are independent

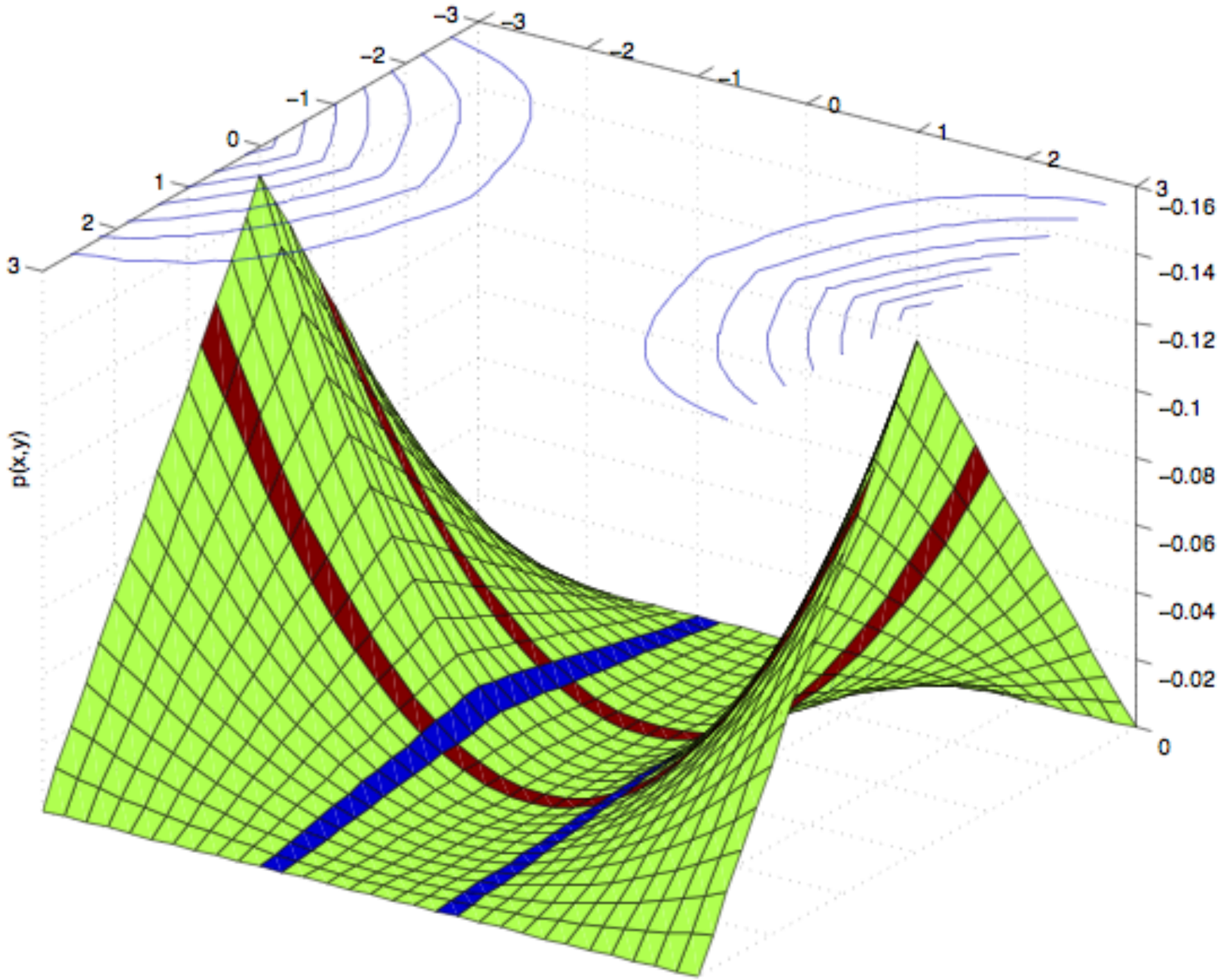
Go back to circle example

Correlation coefficient

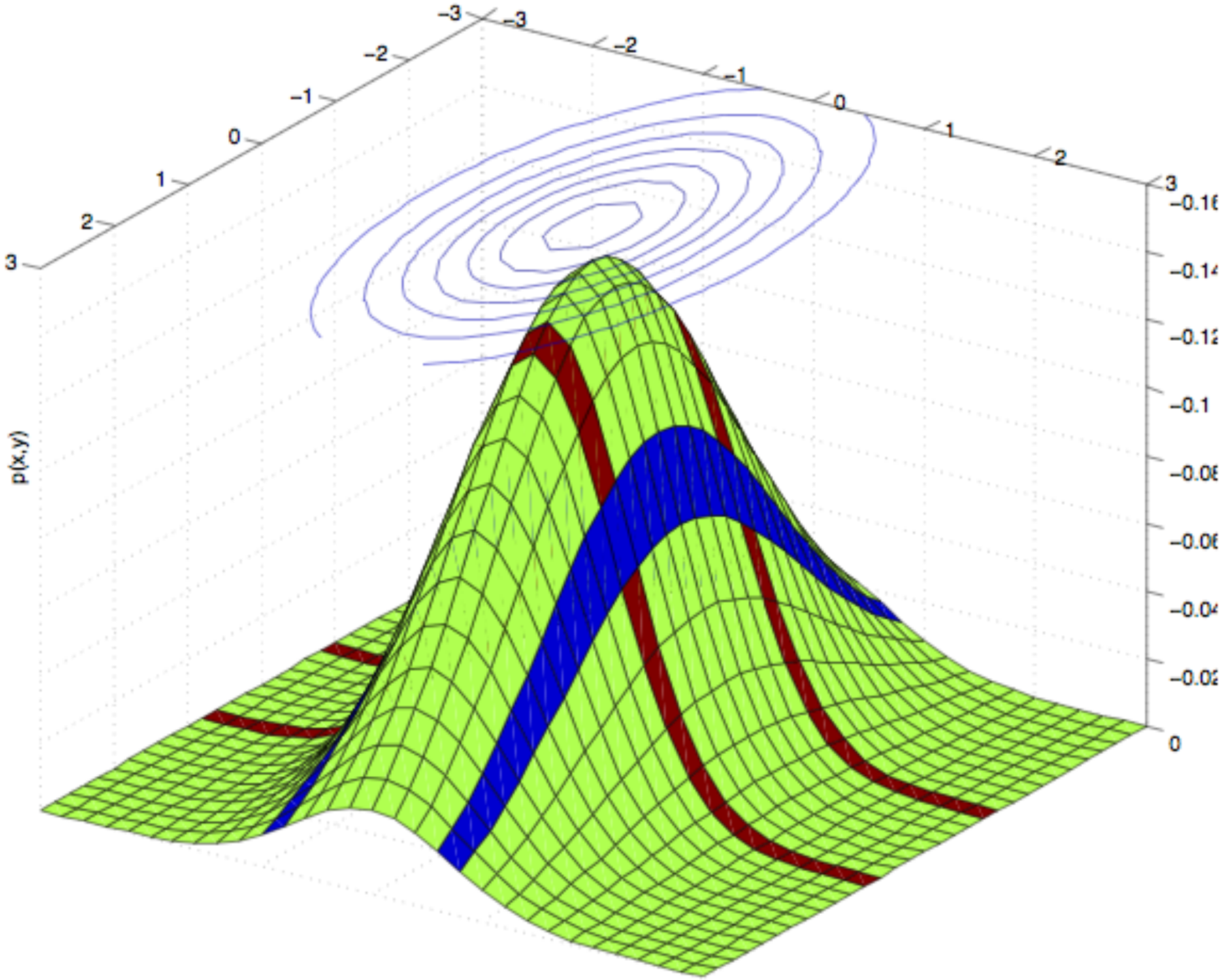
$$\mathrm{Corr}(X,Y) \doteq \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

- Corr(aX+c,bY+d)=Corr(X,Y) if a,b>0
- Corr(X,Y) varies from -1 to +1
- Corr(X,Y)>0 ⇔ X and Y are "correlated"
- Corr(X,Y)<0 ⇔ X and Y are "anti-correlated"
- Corr(X,Y)=1 ⇔ X=aY, a>0
- Corr(X,Y)=-1 ⇔ X=aY a<0

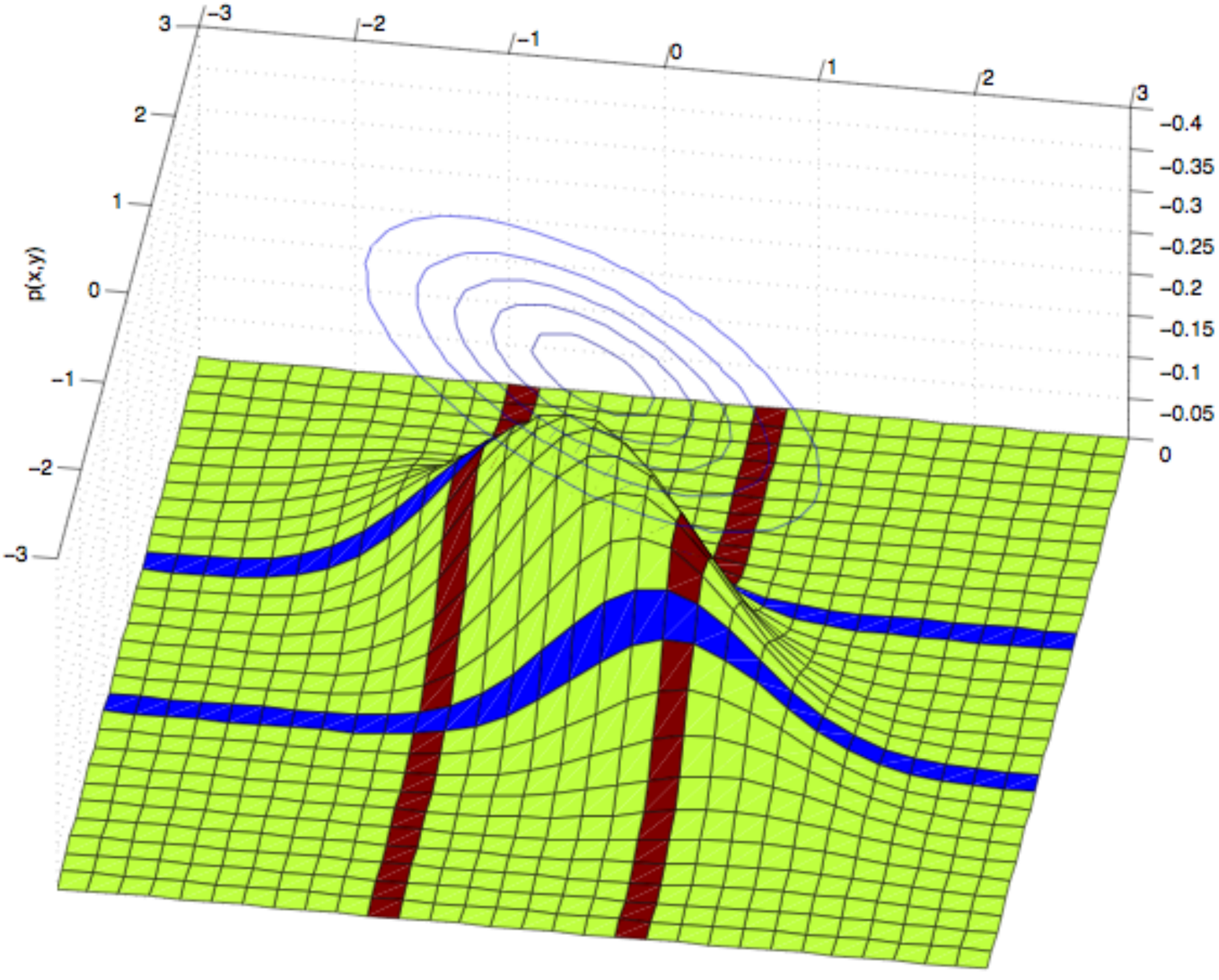
Example I, independent RVs



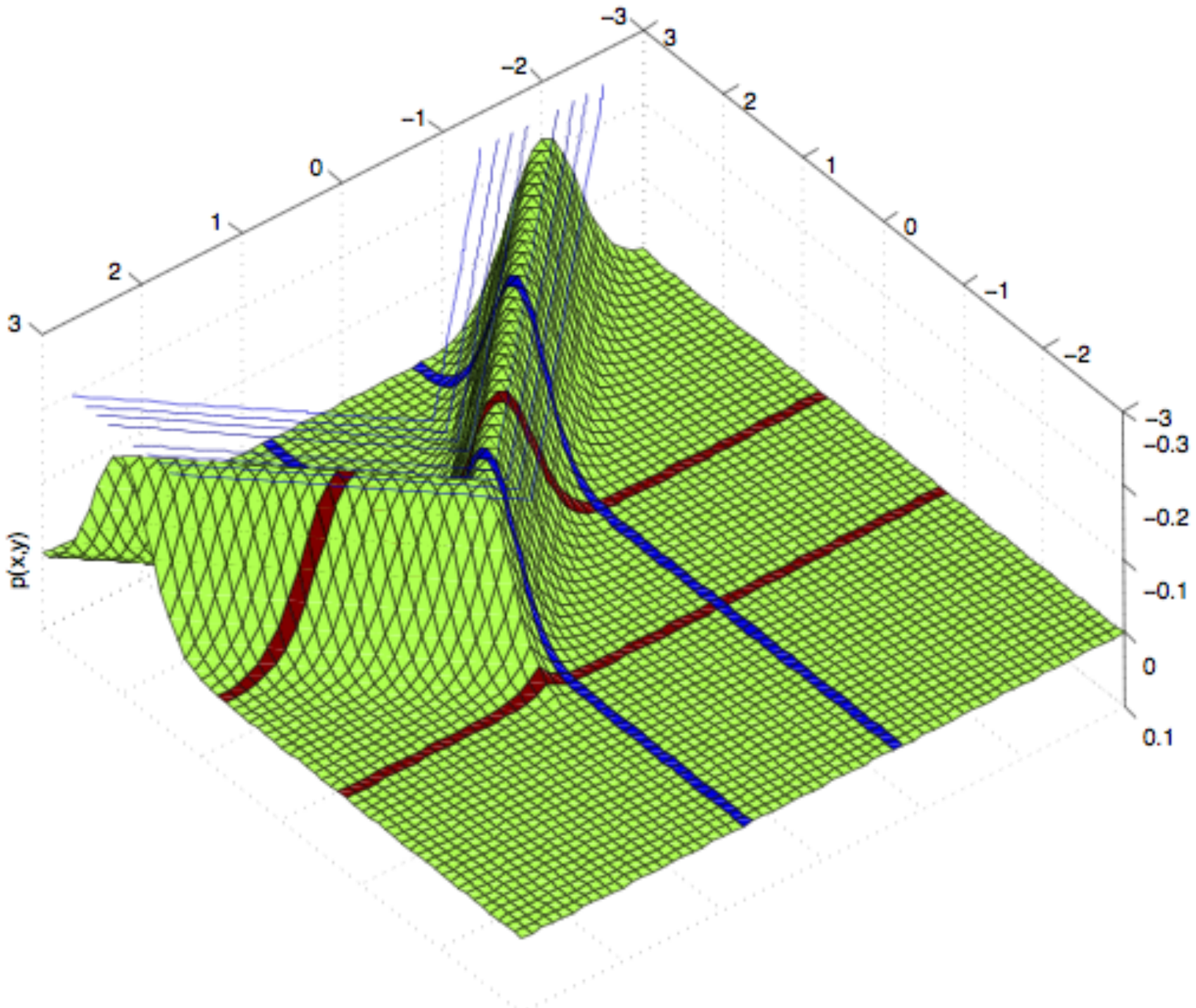
Example 2 independent RVs



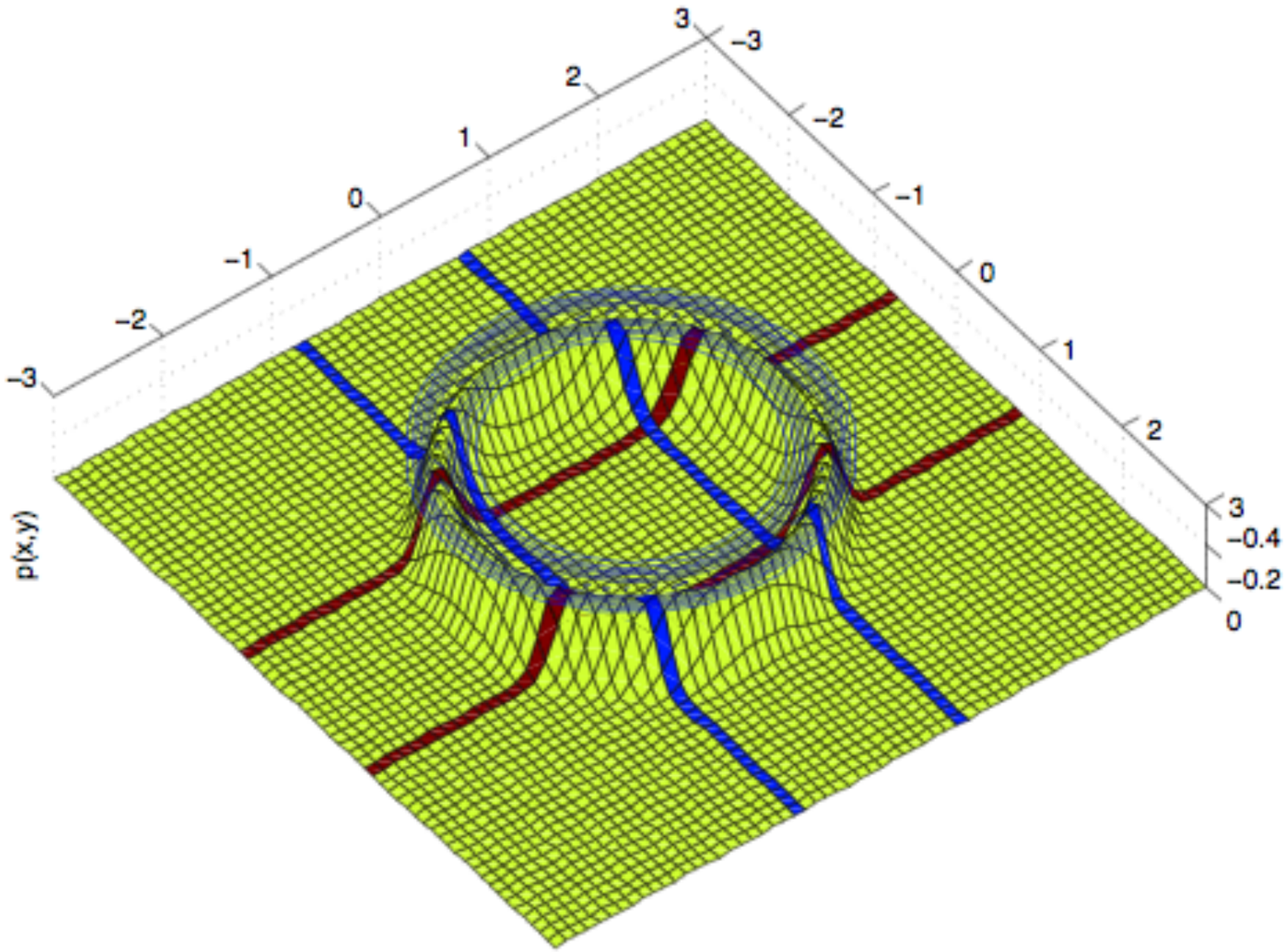
Example 3, Dependent RVs



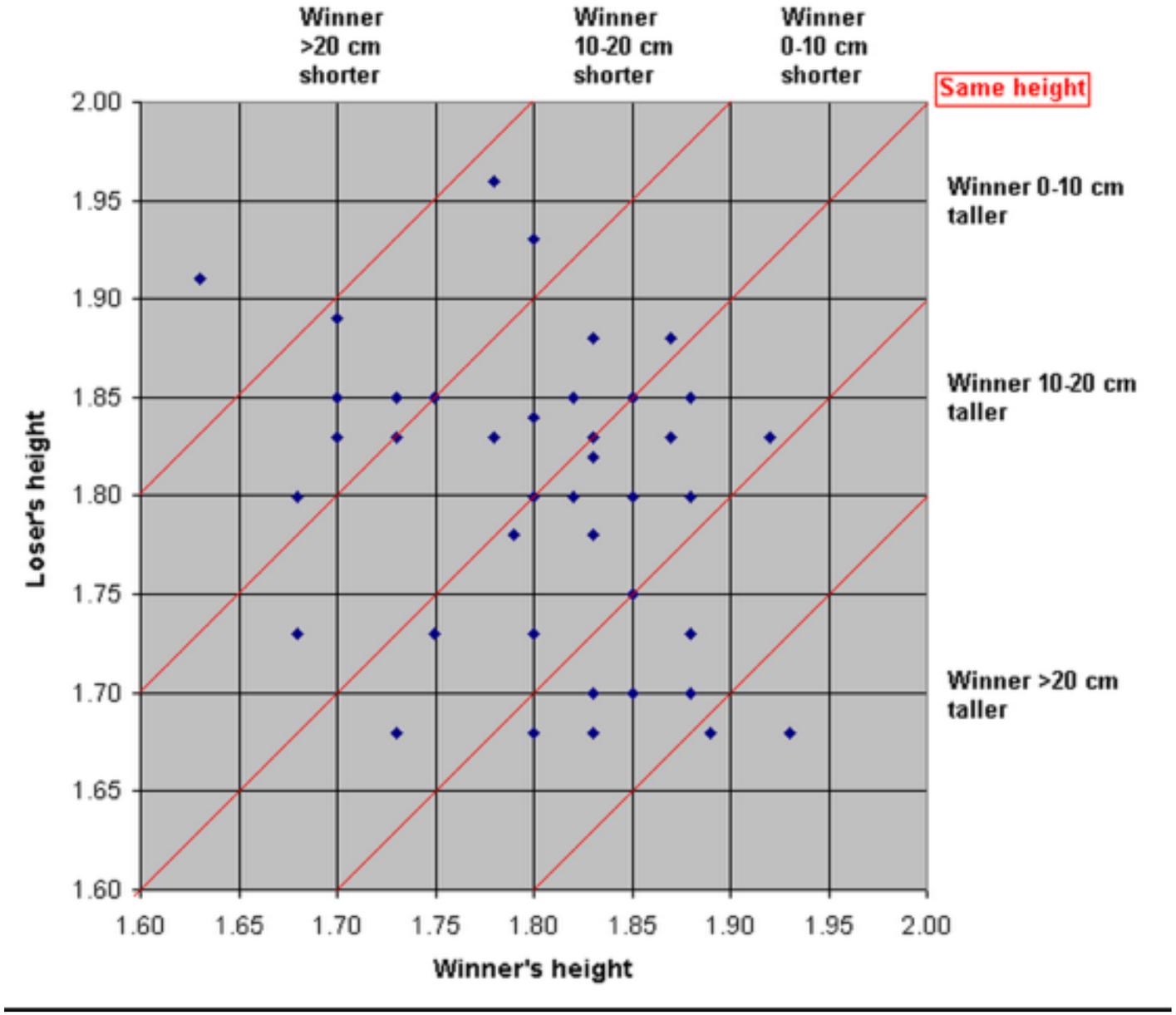
Example 4, functional dependence



Example 5, Circle



Question: are winners in presidential elections taller than their opponent?



Correlation, dependence and causation

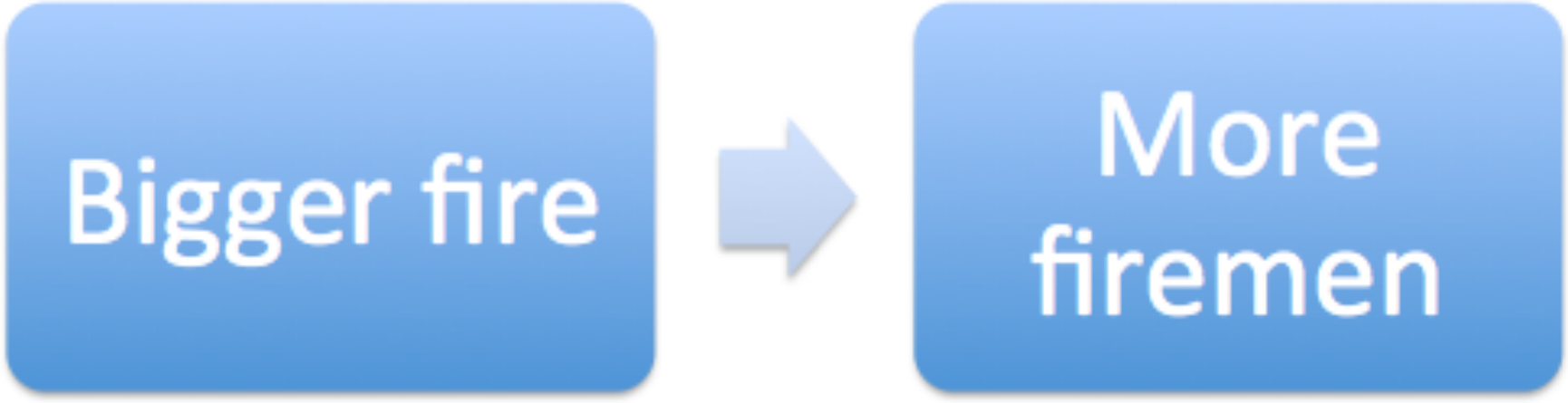
- Correlation → Dependence
- Dependence ↛ Correlation
- Causation can → dependence or Correlation
- dependence or Correlation ↛ Causation

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.

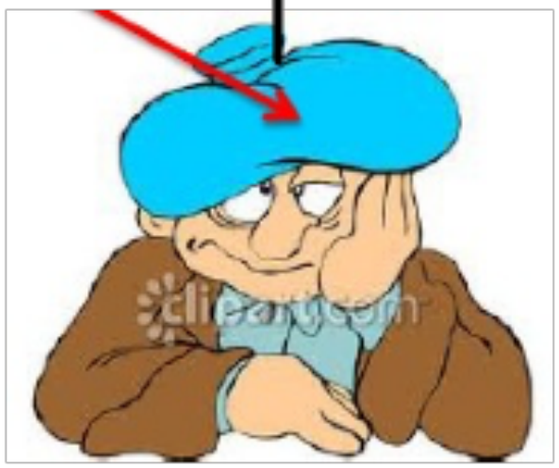
Correlation vs. causation 1

- 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**

Dependence: Sleeping with shoes on is correlated with having a headache in the morning



Correlation vs. causation 2

Excessive Drinking

Sleeping with shoes on

Common Cause

Morning Headache

Correlation vs. Causation 3

- For an ideal gas in a fixed volume, temperature is correlated with pressure.
- Gas, volume and temperature are related by the equation PV=nRT.
- Pressure and Temperature or co-dependent.
- Causation is bi-directional or not well defined.

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.

Determining Causation

In praise of randomized experiments

Randomized Trials

- The final (and most expensive) step in drug development is drug development.
- A drug development company needs to prove to the FDA that their drug **D** reduces blood pressure.
- We select a representative set of patients, randomly select half to be “cases” and half to be “control”
- “Cases” receive the drug.
- “Control” do not receive the drug.

The power of belief

- Studies show that patients that receive a sugar pill (placebo) do significantly better than patients that receive nothing.
  - Solution: give all patients an identical pill, half of the patients get a pill with the drug and half get placebo.
- Studies show that patients that receive treatment from **a doctor who knows** that their pill is real do better than patients whose doctor knows that the pill is a placebo.
  - Solution: double blind trials: neither the doctor nor the patient know whether the pill is real.

Hypothesis testing

- The a-priori opinion of the FDA is that the drug has no effect = the **null hypothesis** is that the mean blood pressure is the same for case and control.
- The goal of the drug company is to reject that hypothesis with a predefined confidence level (typically 5% or 1%).
- The drug company has to declare it's plans for conducting the trial and for evaluating the dependence **before starting the trial**.

The 2 sample t-test

- A standard test used in this kind of case is the 2-sample t-test.
- Two samples are given (case and control)
  - We assume that the size of the sample is large enough that the distribution of the mean is normal (CLT)
  - Null Hypothesis: case and control have the same distribution (same mean and same std)
  - Alternative hypothesis: mean (blood pressure) of case is smaller than mean of control. STDs are equal.
- A confidence level is set: $\alpha = 0.05$

Day of the trial!

- The trial is done according to the pre-authorized protocol.
- The results from the trial are put into a two-sample, equal std t-test.
- p-value: The probability that the results were generated by the null hypothesis. p is a random variable. α is not.
- If $p \leq \alpha$: the null hypothesis is rejected and the drug **D** is "FDA approved"
- If $p > \alpha$ the trial failed – **D** not approved - many unhappy investors.

Observational Study – comparing the effectiveness of approved blood pressure drugs

- Get from CVS/Walgreens .. The purchase orders for blood pressure medicines.
- Get from healthcare providers the records of patient lab tests and diagnostics.
- Compare the effect of different drugs.

The following is a blood pressure medication list of the most commonly prescribed:

- Diuretics
- Beta-blockers
- ACE (Angiotensin-converting enzyme) inhib
- Angiotensin II receptor blockers
- Calcium channel blockers
- Alpha blockers
- Alpha-2 Receptor Agonist
- Combined alpha and beta-blockers
- Central agonists
- Peripheral adrenergic inhibitors
- Blood vessel dilators (vasodilators)

Observational Studies

- In clinical studies sources of Bias are controlled.
  - On the other hand, it is rate to have more than ~100 patients because they are so expensive.
  - This causes large variance.
- In observational studies we cannot show causation, and there is a danger of large Bias.
  - But we can have >100,000 patients, which means small variance.
  - Bias can be reduced by correction methods:
    - Group patients by age, gender …
    - Subtract out the effect of nuisance variables.
    - ….
- Other issues:
  - Patient Privacy.
  - Big Pharma and Big Hospitals have their own interests.

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

Example question:
Suppose that the probability that a computer chip is defective is 0.1% and that we are manufacturing 1,000,000 chips. What is the probabillity that the number of defective chips is larger than 1100?

mean of single defect p=1/1000
n=1000000
mean number of defects=1000
var of single defect 999/1,000,000 approx 1/1000
var of number of defects= 1000. std is approximately 31

Z-score is 100/31 more than 3, less than 4.

Probability is smaller than 0.13% (corresponding to 3X std)

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%
* $E(X)=7$
* $P(X=3)=0.23$

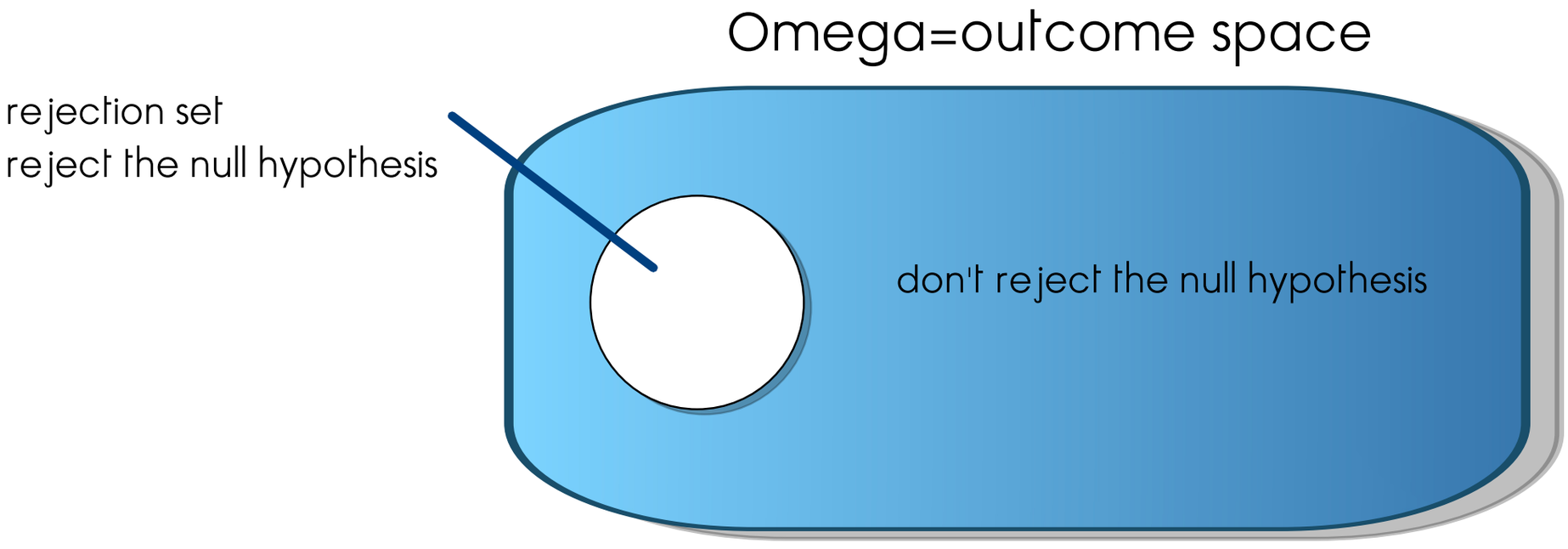
Choosing alpha is a compromise between two types of errors:
Type I error: Rejecting the null hypothesis when it is correct
Type II error: Failing to reject the null hypothesis when it is incorrect.

	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.

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

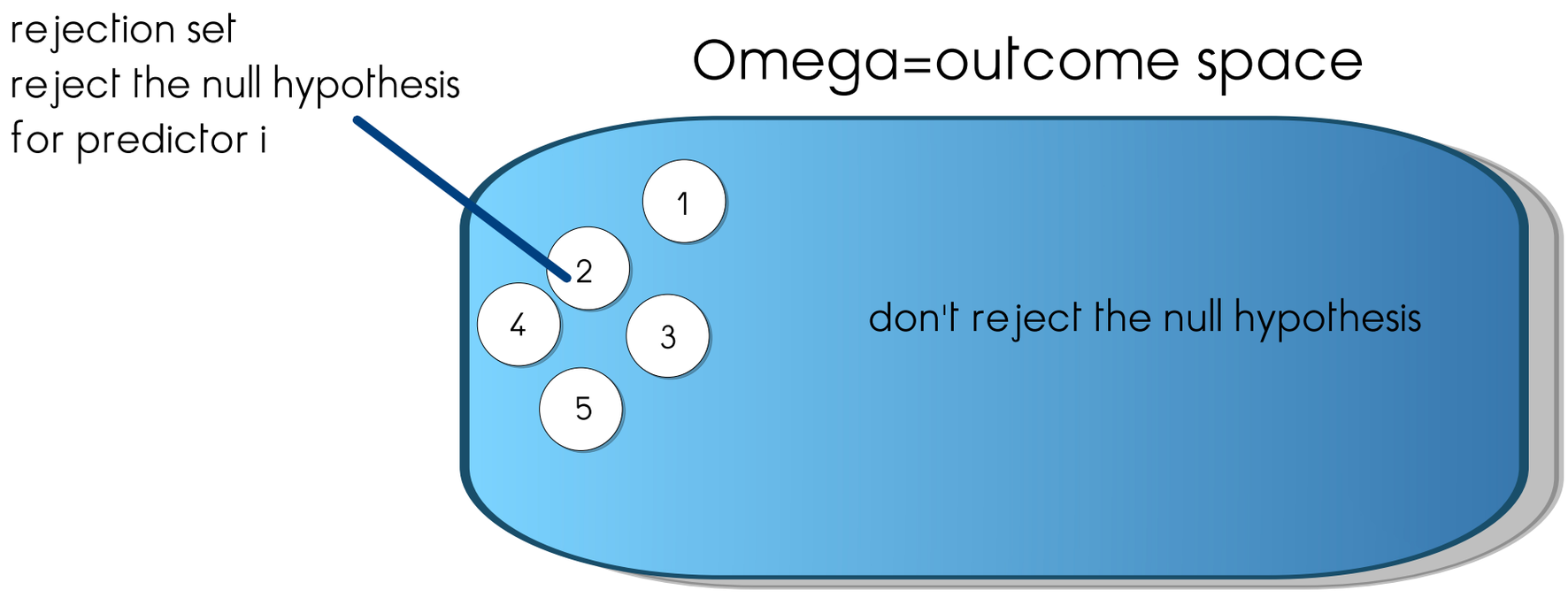
Consider the online ad problem, our goal is to maximize click-through rate. Our null hypothesis is that nothing performs better than picking one of the ads uniformly at random each time.

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.