Biostatistics

Chapter 4 Hypothesis Testing (假设检验)

Jing Li

jing.li@sjtu.edu.cn

http://cbb.sjtu.edu.cn/~jingli/courses/2017fall/bi372/

Dept of Bioinformatics & Biostatistics, SJTU

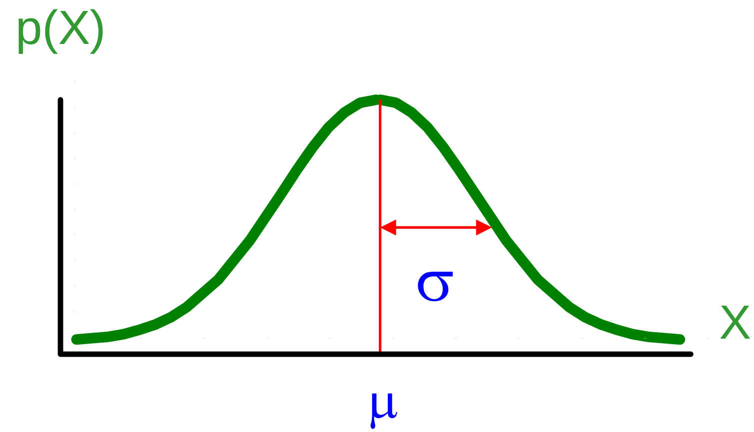
Review Questions (5 min)

- What's standard error (s.e.)? And which factor(s) affects the size of s.e.?
- What's the differences between s.d. and s.e.

Review lecture 3

Normal distribution

- μ, σ
- 68-95-99.7 Rule
- Standard normal curve



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

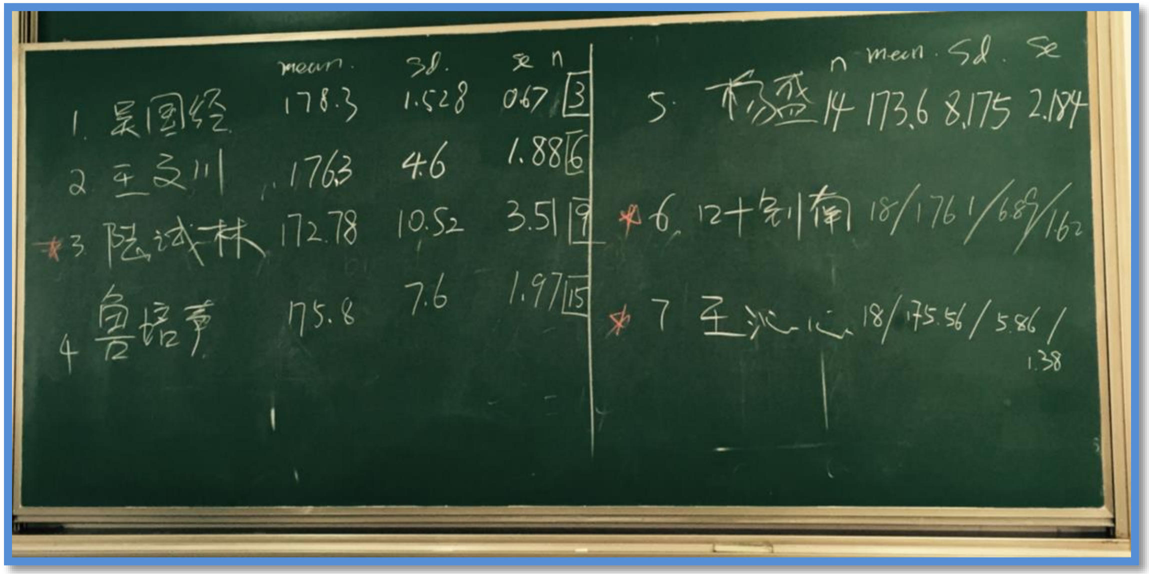
- Statistics
- Random sampling

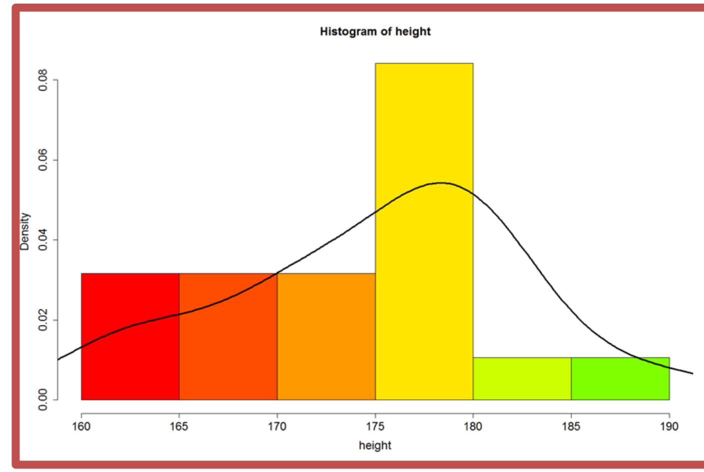
Standard error

- Standard Error is a measure of sampling variability.
- Standard error is the standard deviation of a sample statistic.
- Standard error decreases with increasing sample size and increases with increasing variability of the outcome (e.g., IQ).
- Standard errors can be predicted by computer simulation or mathematical theory (formulas).
  - **The formula for standard error is different for every type of statistic (e.g., mean, difference in means, odds ratio).**

(sample mean) s.e. = $$\frac{s}{\sqrt{n}}$$

Local data of last year-- height



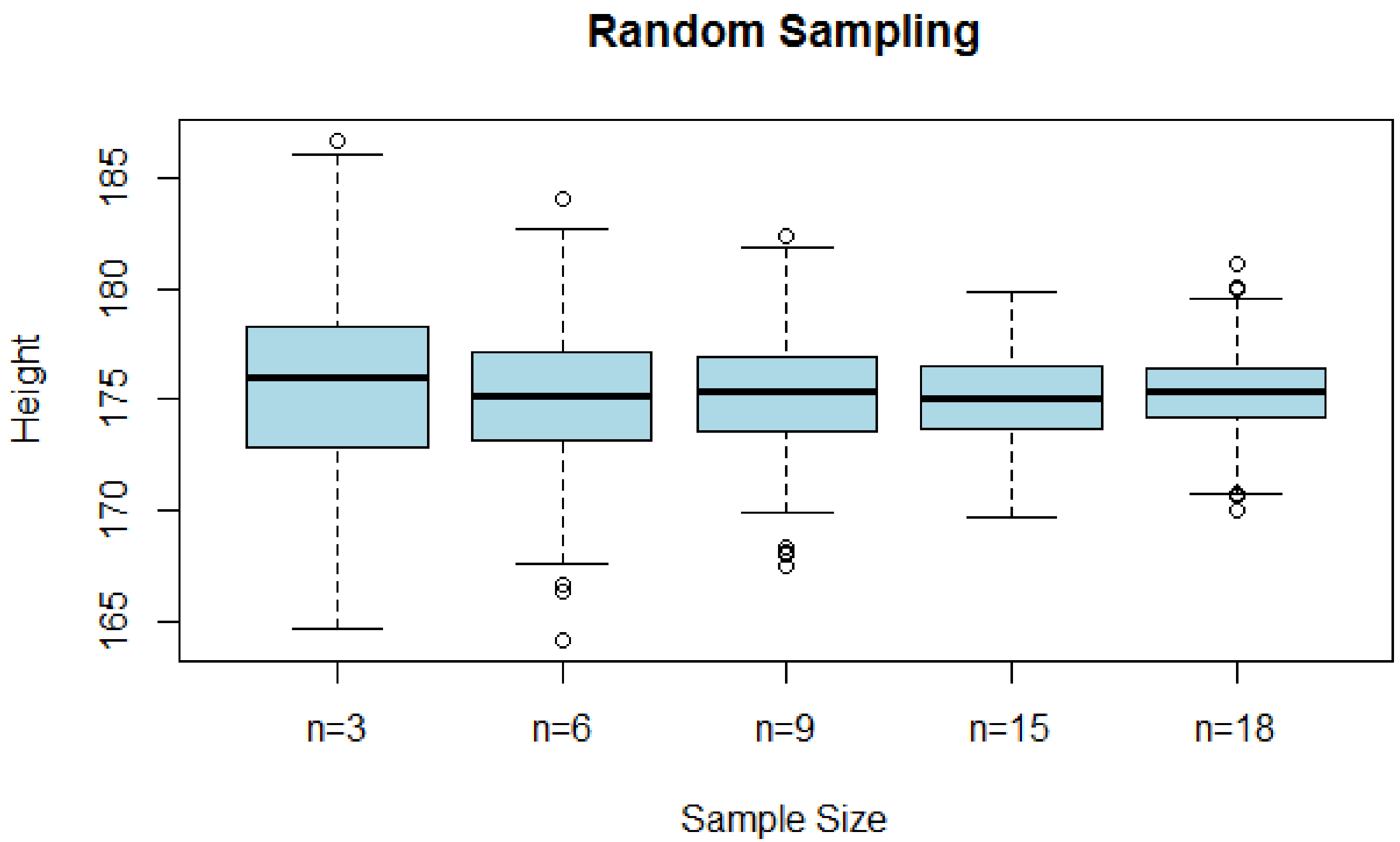


- summary(height)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
161.0	170.0	178.0	175.2	180.0	190.0

sd=7.45

Random sampling



Mean=175.2

Confidence intervals

CI, 置信区间

Example

- Cross-sectional study of 100 middle-aged and older European men.
- Estimation: What is the average serum vitamin D in middle-aged and older European men?

  - **Mean = 62 nmol/L**
  - **Standard deviation = 33 nmol/L**

Something more

- Up to this point we have drawn a sample and estimated the population value with the sample mean. This was called a point estimate.

- Now, we may want to know even more than the point estimate. We want to know an interval of plausible values for the population mean based on our sample

Something more

- Up to this point we have drawn a sample and estimated the population value with the sample mean. This was called a point estimate.

- Now, we may want to know even more than the point estimate. We want to know an interval of plausible values for the population mean based on our sample

Something more

- Up to this point we have drawn a sample and estimated the population value with the sample mean. This was called a point estimate.

- Now, we may want to know even more than the point estimate. We want to know an interval of plausible values for the population mean based on our sample

- Confidence interval (CI)

Confidence interval

- Definition is a particular kind of interval estimate of a population parameter and is used to indicate the reliability of an estimate.
- As we discussed before, when we take multiple samples, the sample mean will not be the same every time. The confidence interval is an interval around our sample mean that allows us to have a certain amount of confidence that the true mean is covered by the interval.
- We can draw conclusions about the true population mean based on our confidence interval

95% confidence interval

- Goal: capture the true effect (e.g., the true mean) most of the time.

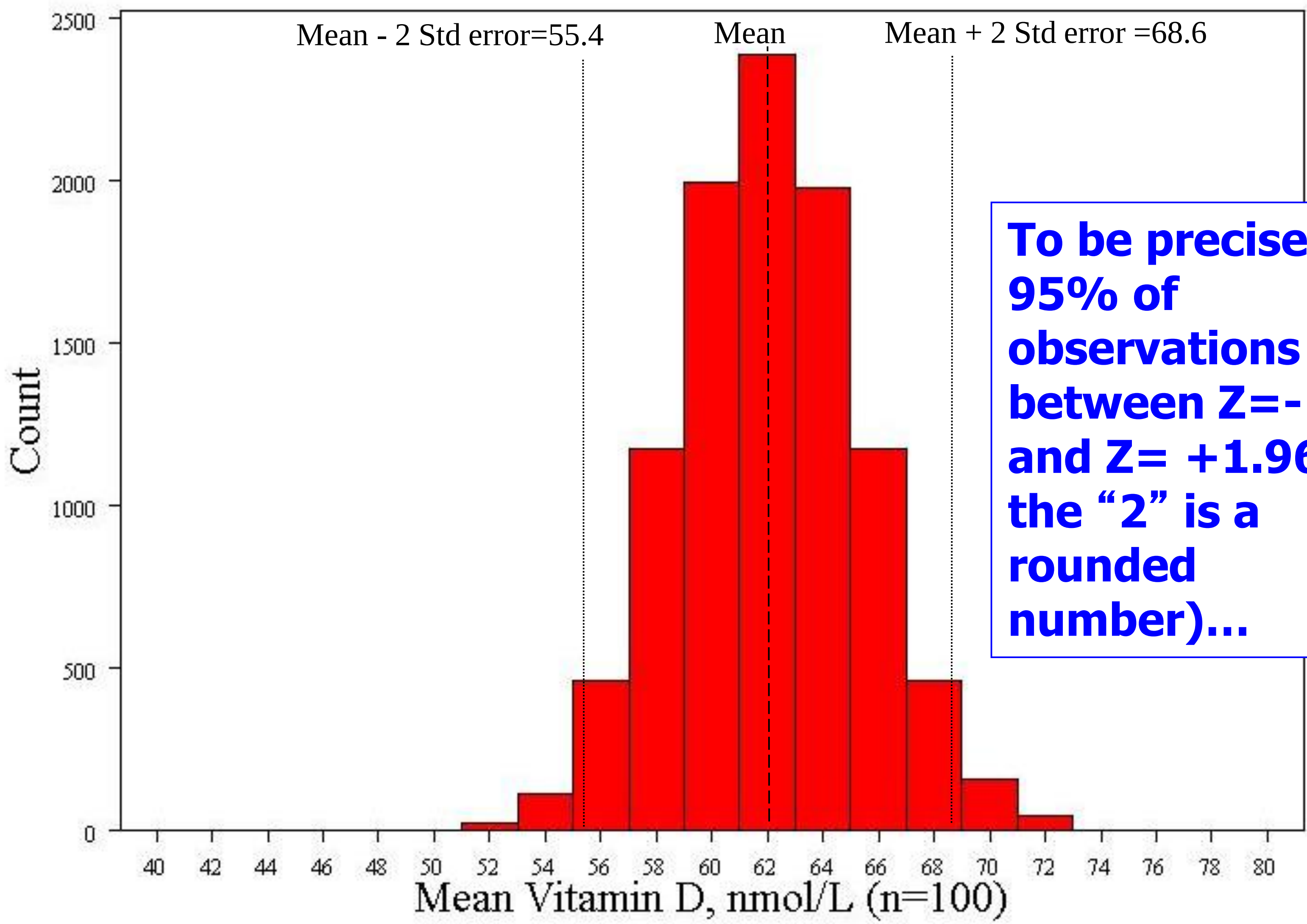
If repeated samples were taken

- A 95% confidence interval should include the true effect about 95% of the time. Naturally, 5% of the intervals would not contain the population mean.

- A 99% confidence interval should include the true effect about 99% of the time.

Recall: 68-95-99.7 rule for normal distributions!

These is a 95% chance that the sample mean will fall within two standard errors of the true mean= 62 +/- 2*3.3 = 55.4 nmol/L to 68.6 nmol/L



To be precise, 95% of observations fall between Z=-1.96 and Z= +1.96 (so the "2" is a rounded number)…

Confidence Intervals

The value of the statistic in my sample (eg., mean, odds ratio, etc.)

point estimate ± (measure of how confident we want to be) × (standard error)

From a Z table or a T table, depending on the sampling distribution of the statistic

Standard error of the statistic.

Confidence Intervals give:

*A plausible range of values for a population parameter.

*The precision of an estimate.(When sampling variability is high, the confidence interval will be wide to reflect the uncertainty of the observation.)

Standard error

- s.e= $\dfrac{s}{\sqrt{n}}$

The standard error of a mean

- s.e= $\sqrt{\dfrac{p(1-p)}{n}}$

The standard error of a proportion or percentage

Difference between means,

$x_1 - x_{2:}$ σ_{x1-x2} = sqrt [$\sigma^2{}_1$ / n_1 + $\sigma^2{}_2$ / n_2]

Difference between proportions,

$p_1 - p_{2:}$ σ_{p1-p2} = sqrt [$P_1(1-P_1)$ / n_1 + $P_2(1-P_2)$ / n_2]

Common "Z" levels of confidence

- Commonly used confidence levels are 90%, 95%, and 99%

Confidence Level	*Z value*
80%	1.28
90%	1.645
95%	1.96
98%	2.33
99%	2.58
99.8%	3.08
99.9%	3.27

99% confidence intervals…

- **99% CI for mean vitamin D (mean=63nmol/L, s.e=3.3):**

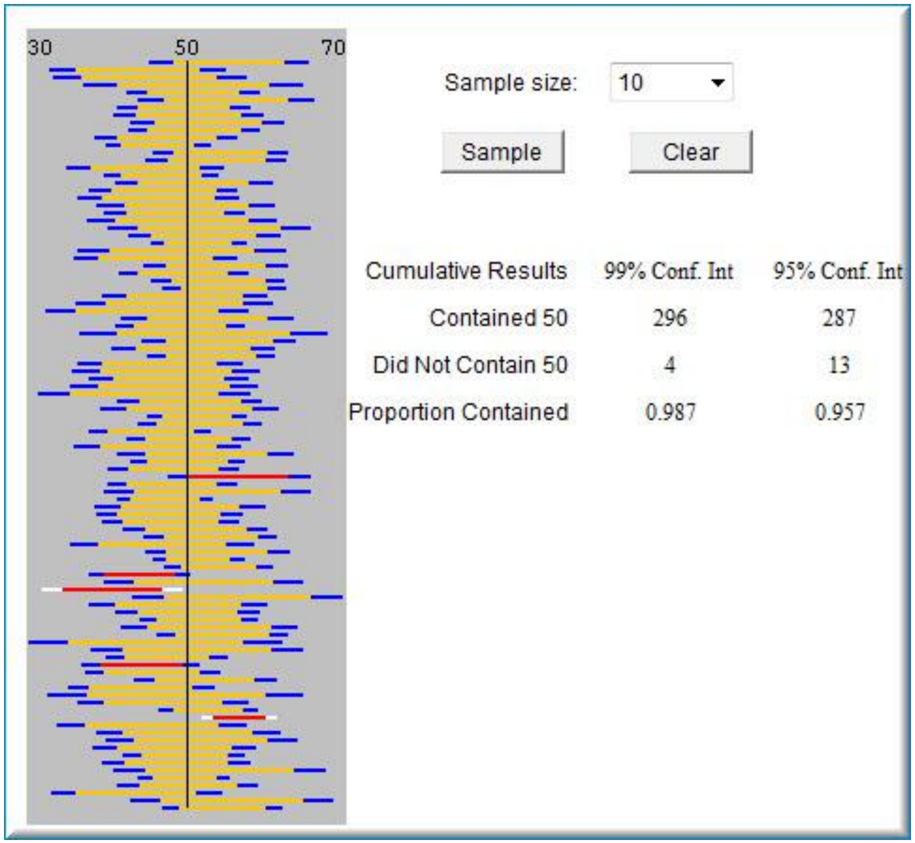
63 nmol/L $\pm$ 2.6 *x* (3.3) = 54.4 – 71.6 nmol/L

Changing the width of the confidence interval

- The width of the confidence interval is based on 3 factors
  - confidence level (z)- how confident do we want to be that the interval covers m; the higher the confidence, the wider the interval
  - variance (s)- how different might the samples be; the more variability, the wider the interval
  - sample size (n)- how many samples did we use to estimate the population mean; the larger the sample, the better the point estimate, the narrower the interval

Simulation for CI

The demonstrtation generates confidence intervals for sample experiments taken from a population with a mean of 50 and a standard deviation of 10.



The figure displays the results of 300 experiments with a sample size of 10. The 95% confidence intervals that contain the mean of 50 are shown in orange and the those that do not are shown in red. The 99% confidence intervals are shown in blue if they contain 50 and white if they do not.

http://onlinestatbook.com/2/estimation/ci_sim.html

Practice

- A student collected a large amount of demographic data from school children in a depressed area. Since this population was possibly malnourished ［营养不良的］, she was concerned that the children would have a hemoglobin ［血红素］ level below the healthy average. The healthy average is **13 g/dL**.
- She asked me to run a hypothesis test comparing the hemoglobin levels in her sample population to the healthy average value. She had collected a sample of size 127 children.

Practice

Sample hemoglobin levels:

Mean = 11.7 g/dL, Standard deviation = 1.2 g/dL, n=127

- We would like to provide a 95% confidence interval for the hemoglobin level for the children in the school.

(sqrt(127)=11.27)

Practice

Sample hemoglobin levels:

Mean = 11.7 g/dL, Standard deviation = 1.2 g/dL, n=127

- We would like to provide a 95% confidence interval for the hemoglobin level for the children in the school.

$$\left(11.7 - 1.96\frac{1.2}{\sqrt{127}}, 11.7 + 1.96\frac{1.2}{\sqrt{127}}\right) = (11.49, 11.91)$$

- For a 99% interval

$$\left(11.7 - 2.58\frac{1.2}{\sqrt{127}}, 11.7 + 2.58\frac{1.2}{\sqrt{127}}\right) = (11.43, 11.97)$$

(sqrt(127)=11.27)

Conclusions

We are 95% confident that the true mean level of hemoglobin in school children is between 11.49 and 11.91. Beyond that, we are 99% confident that the true mean level is between 11.43 and 11.97.

Statistics Primer（统计入门）

- Statistical Inference
- Hypothesis testing
- P-values
- Type I error
- Type II error
- Statistical power

What is statistical inference?

- The field of statistics provides guidance on how to make conclusions in the face of chance variation (sampling variability).

Example 1: Difference in proportions

- Research Question: Are antidepressants a risk factor for suicide attempts in children and teenagers ?

▪Example modified from: "Antidepressant Drug Therapy and Suicide in Severely Depressed Children and Adults "; Olfson et al. *Arch Gen Psychiatry.* 2006;63:865-872.

Example 1:

- Design: Case-control study
- Methods: Researchers used Medicaid records to compare prescription histories between 263 children and teenagers (6-18 years) who had attempted suicide and 1241 controls who had never attempted suicide (all subjects suffered from depression).
- Statistical question: Is a history of use of antidepressants more common among cases than controls?

Example I

- Statistical question: Is a history of use of particular antidepressants **more common** among depress cases than controls?

What will we actually compare?

Proportion of cases who used antidepressants in the past vs. proportion of controls who did

Results

	Cases (n=263)	Controls (n=1241)
Any antidepressant drug ever	120 (46%)	448 (36%)

46%

36%

Difference=10%

What does a 10% difference mean?

- Before we perform any formal statistical analysis on these data, we already have a lot of information.

- Look at the basic numbers first; THEN consider statistical significance as a secondary guide.

Is the association statistically significant?

- This 10% difference could reflect a true association or it could be a fluke (偶然事件) in this particular sample.

- The question: is 10% bigger or smaller than the expected sampling variability?

What is hypothesis testing?

- Statisticians try to answer this question with a formal hypothesis test

Hypothesis testing

Step 1: Assume the null hypothesis（无效假设）.

Null hypothesis: there is no association between antidepressant use and suicide attempts in the target population (= the difference is 0%)

Hypothesis Testing

Step 2: Predict the sampling variability assuming the null hypothesis is true—math theory (formula):

The standard error of the difference in two proportions is:

$$= \sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}}$$

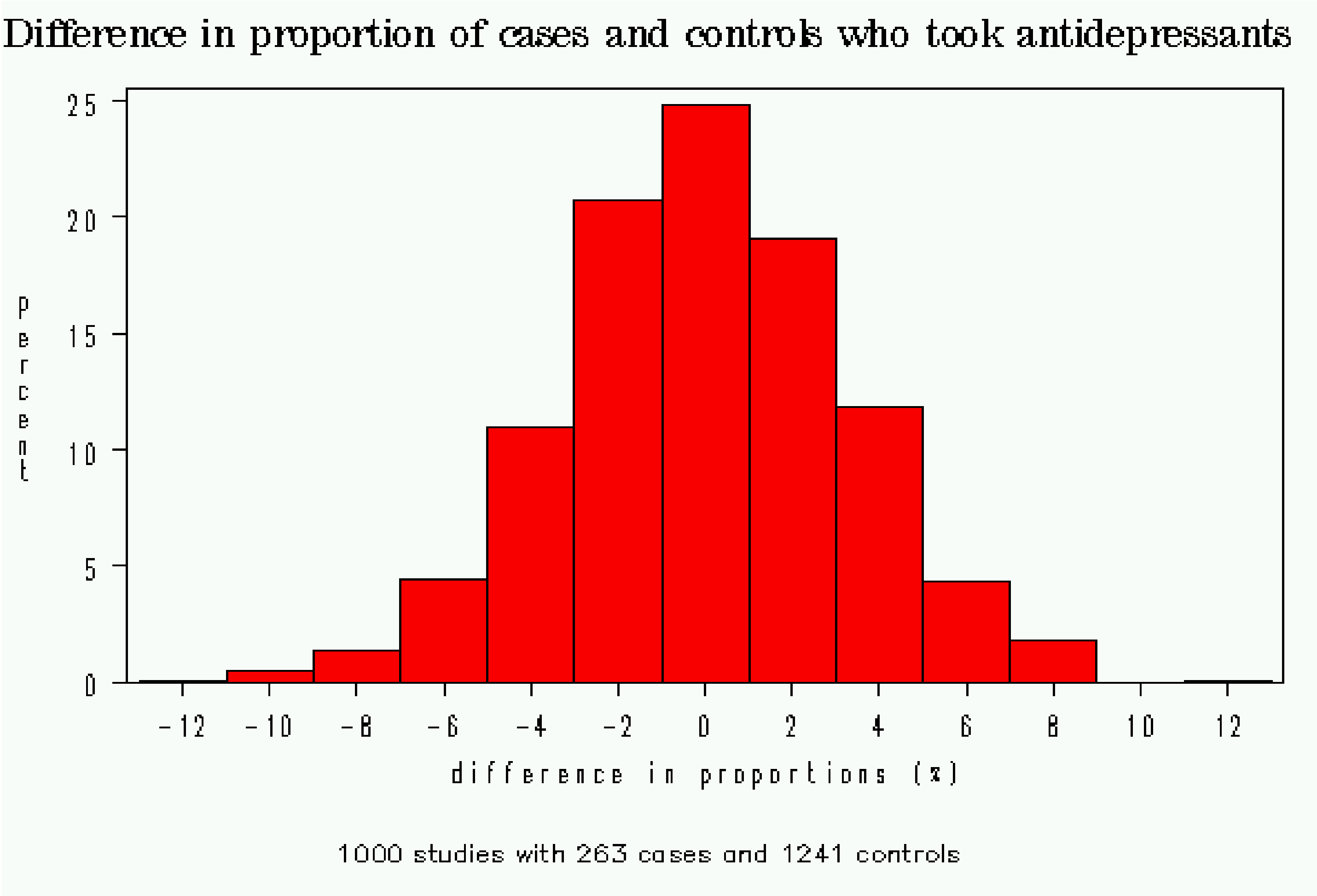
$$= \sqrt{\frac{\frac{568}{1504}(1-\frac{568}{1504})}{263} + \frac{\frac{568}{1504}(1-\frac{568}{1504})}{1241}} = .033$$

Hypothesis Testing

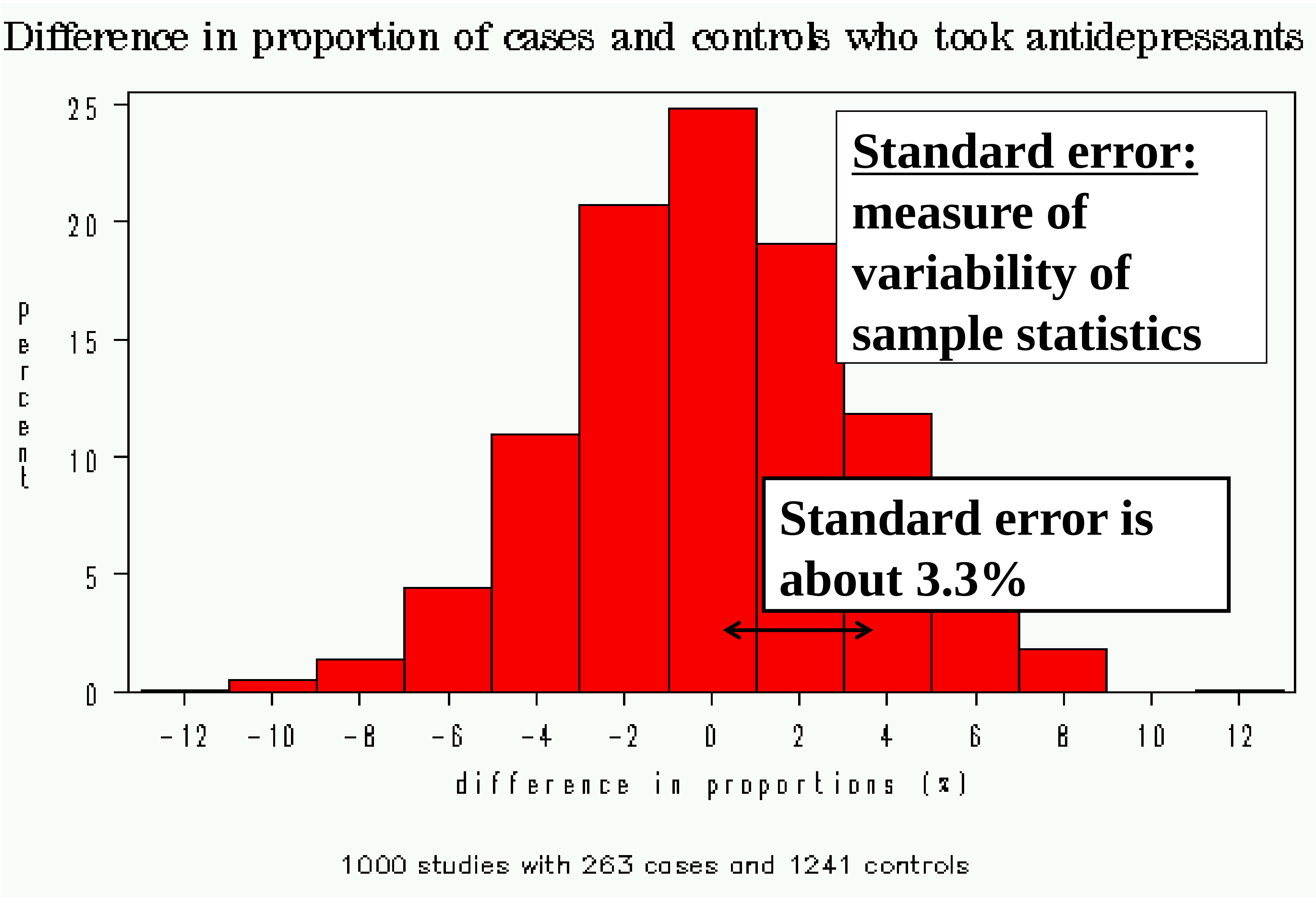
Step 2: Predict the sampling variability assuming the null hypothesis is true—computer simulation:

- In computer simulation, you simulate taking repeated samples of the same size from the same population and observe the sampling variability.
- I used computer simulation to take 1000 samples of 263 cases and 1241 controls assuming the null hypothesis is true (e.g., no difference in antidepressant use between the groups).

Computer Simulation Results



What is standard error?



Hypothesis Testing

Step 3: Do an experiment

We observed a difference of 10% between cases and controls.

Hypothesis Testing

Step 4: Calculate a p-value

P-value=the probability of your data or something more extreme under the null hypothesis.

Hypothesis Testing

Step 4: Calculate a p-value—mathematical theory:

Difference in proportions follows a normal distribution.

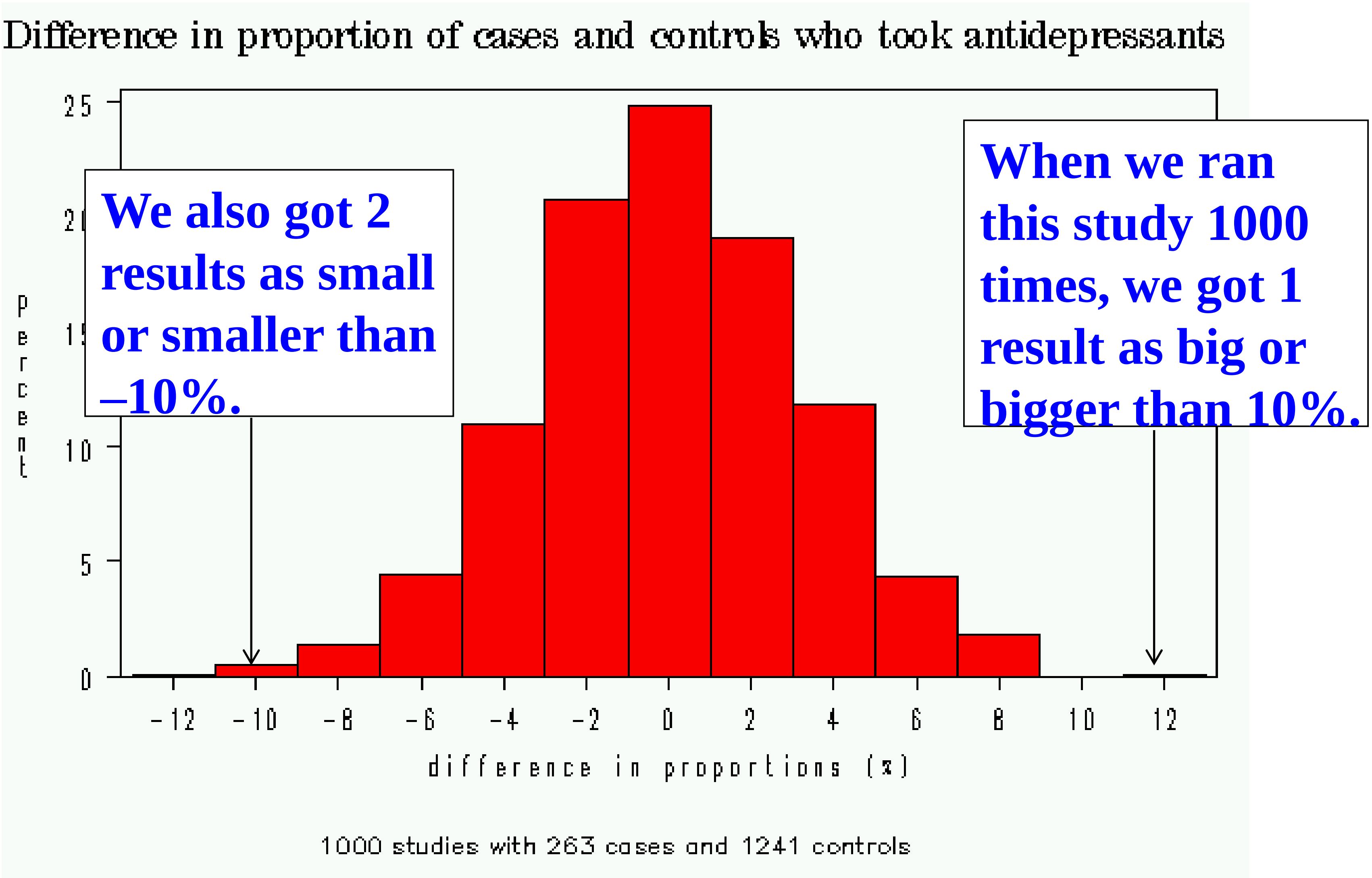
Observed difference between the groups.

A Z-value of 3.0 corresponds to a p-value of .003.

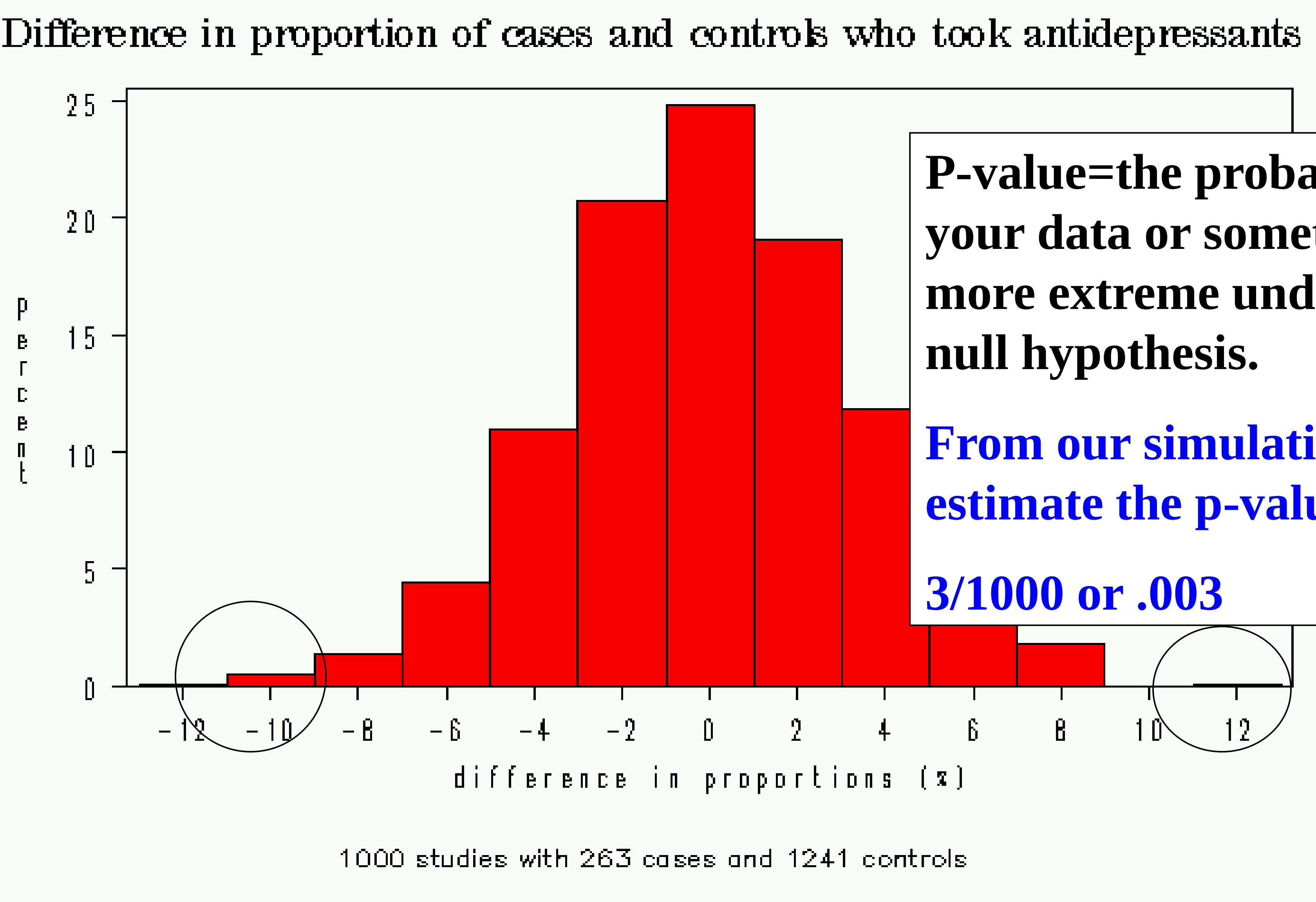
$$Z = \frac{.10}{.033} = 3.0; p = .003$$

Standard error.

The p-value from computer simulation…



P-value



P-value=the probability of your data or something more extreme under the null hypothesis.

From our simulation, we estimate the p-value to be:

3/1000 or .003

Hypothesis Testing

Step 5: Reject or do not reject the null hypothesis.

Here we reject the null.

Alternative hypothesis（备择假设）:There is an association between antidepressant use and suicide in the target population.

What does a 10% difference mean?

- Is it “statistically significant”?
- Is it clinically significant?
- Is this a causal association?

What does a 10% difference mean?

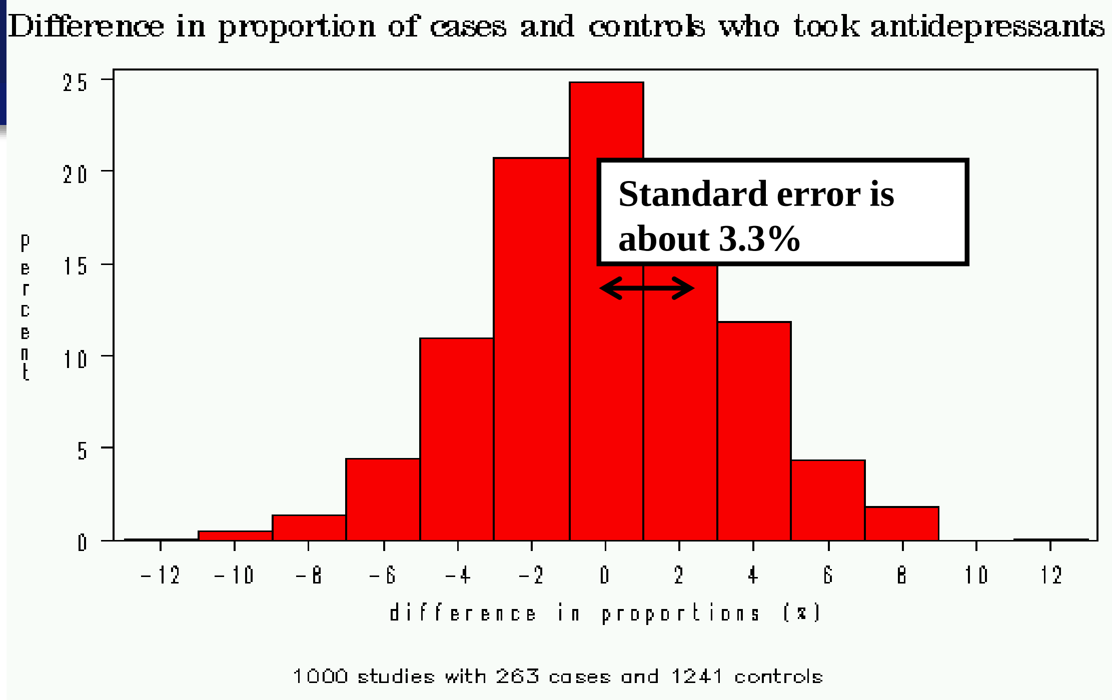
- Is it "statistically significant"? YES
- Is it clinically significant? MAYBE
- Is this a causal association? MAYBE

Statistical significance does not necessarily imply clinical significance.

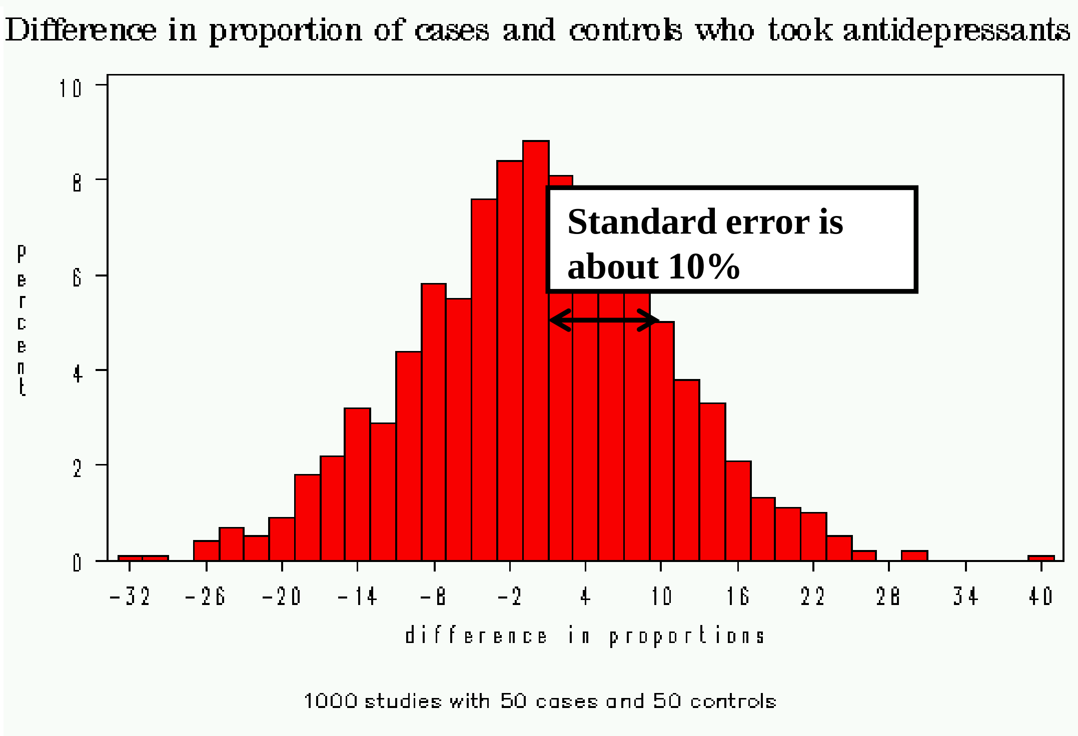
Statistical significance does not necessarily imply a cause-and-effect relationship.

What would a lack of statistical significance mean?

- If this study had sampled only 50 cases and 50 controls, the sampling variability would have been much higher—as shown in this computer simulation…

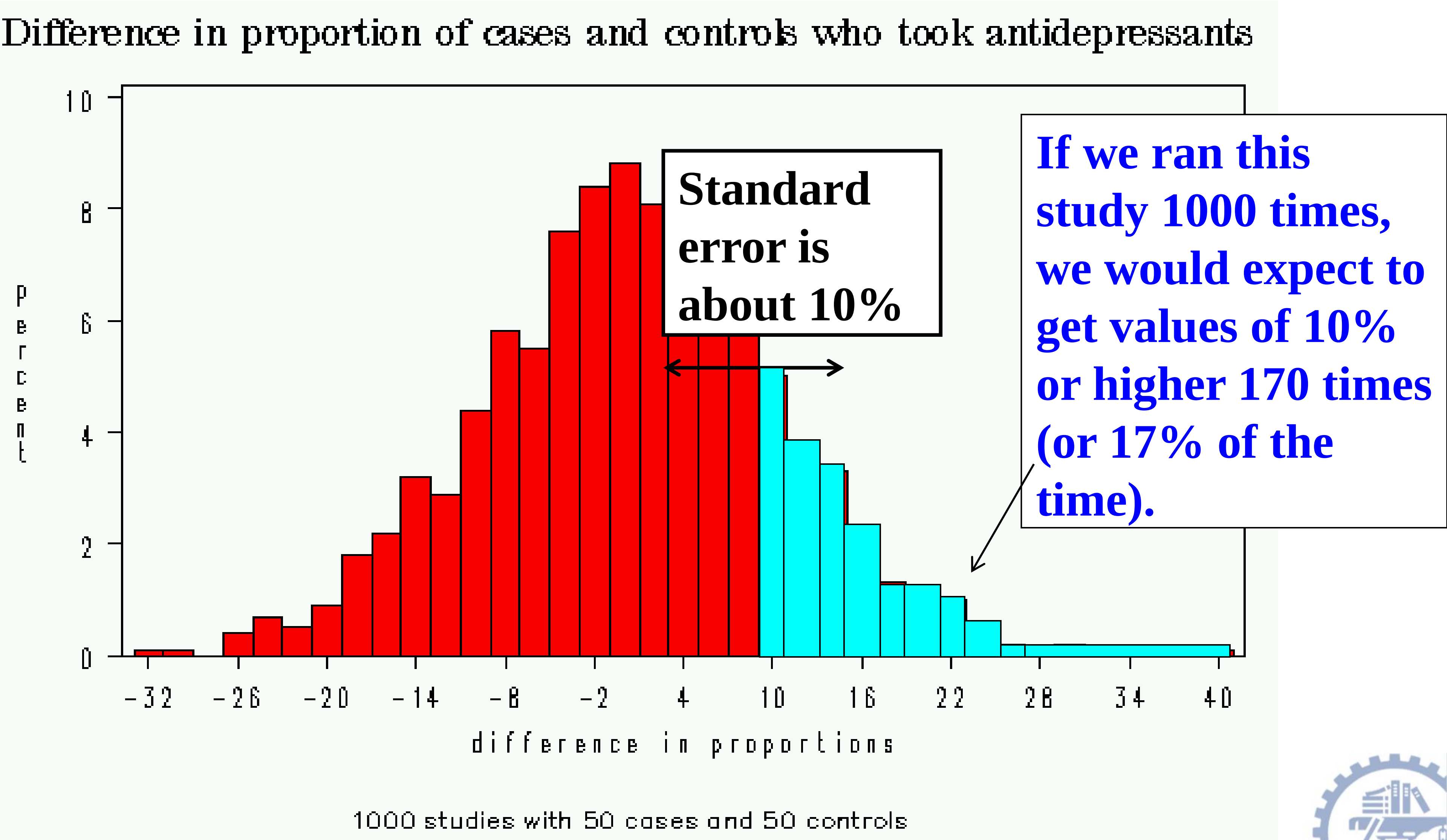


263 cases and 1241 controls.

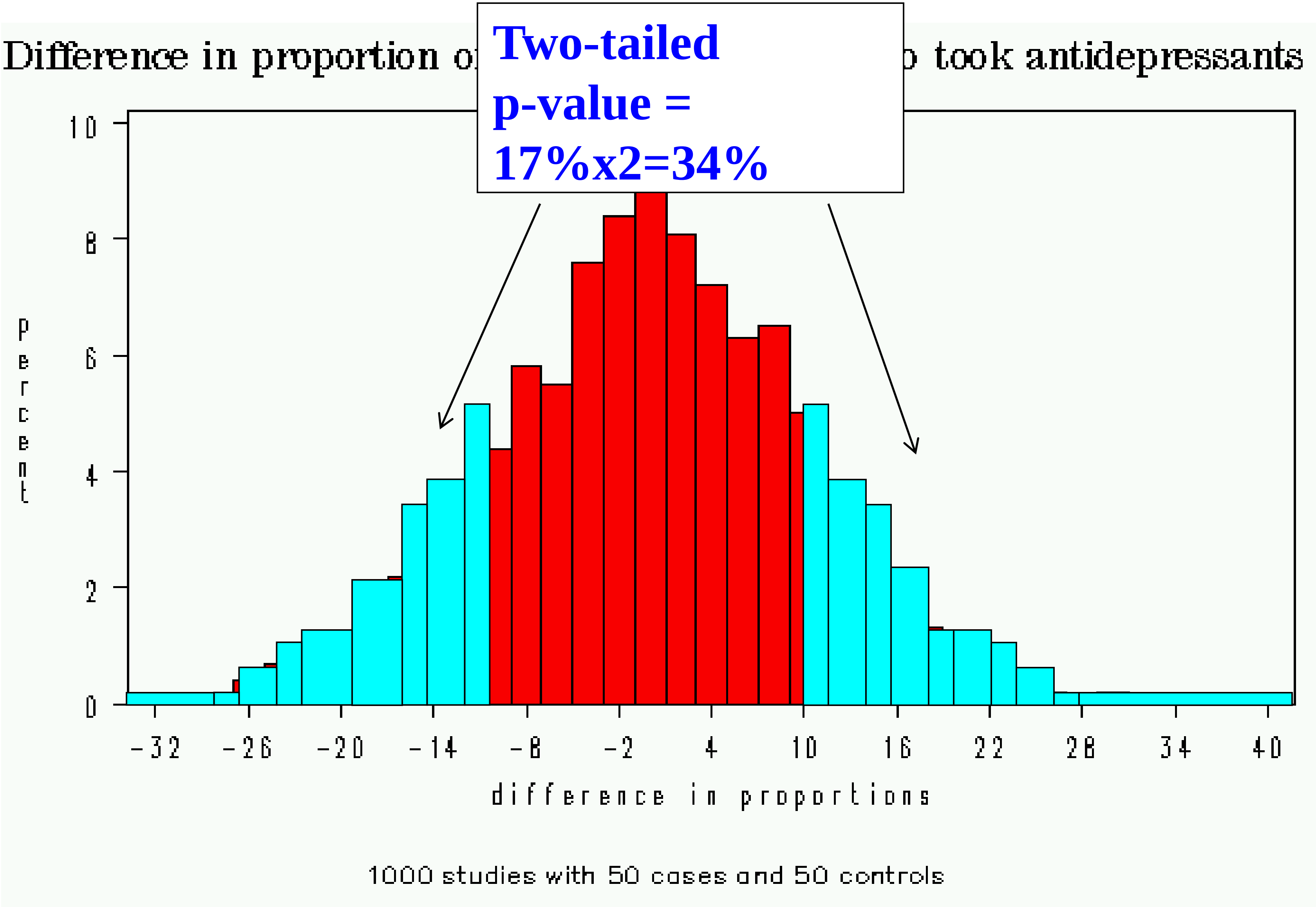


50 cases and 50 controls.

With only 50 cases and 50 controls…



Two-tailed p-value



With only 50 cases and 50 controls…

What does a 10% difference mean (50 cases/50 controls)?

- Is it "statistically significant"? NO

No evidence of an effect ≠ Evidence of no effect.

Example 2: Difference in means

- Rosental, R. and Jacobson, L. (1966) Teachers' expectancies: Determinates of pupils' I.Q. gains. *Psychological Re ports*, 19, 115-118. (皮格马利翁效应/罗森塔尔效应/期待效应)

The Experiment

- Grade 3 at a school were given an IQ test at the beginning of the academic year (n=90).
- Classroom teachers were given a list of names of students in their classes who had supposedly scored in the top 20 percent; these students were identified as "academic bloomers" (n=18).
- *BUT: the children on the teachers lists had actually been randomly assigned to the list.*
- At the end of the year, the same I.Q. test was re-administered.

Example 2

- Statistical question: Do students in the treatment group have **more improvement** in IQ than students in the control group?

What will we actually compare?

- One-year change in IQ score in the treatment group vs. one-year change in IQ score in the control group.

The standard deviation of change scores was 2.0 in both groups. This affects statistical significance…

Results:

	"Academic bloomers" (n=18)	Controls (n=72)
Change in IQ score:	12.2 (2.0)	8.2 (2.0)

12.2 points

8.2 points

Difference=4 points

What does a 4-point difference mean?

- Before we perform any formal statistical analysis on these data, we already have a lot of information.

- Look at the basic numbers first; THEN consider statistical significance as a secondary guide.

Is the association statistically significant?

- This 4-point difference could reflect a true effect or it could be a fluke.
- The question: is a 4-point difference bigger or smaller than the expected sampling variability?

Hypothesis testing

Step 1: Assume the null hypothesis.

Null hypothesis: There is no difference between "academic bloomers" and normal students (= the difference is 0%)

Hypothesis Testing

<u>Step 2: Predict the sampling variability assuming the null hypothesis is true—math theory:</u>

The standard error of the difference in two means is:

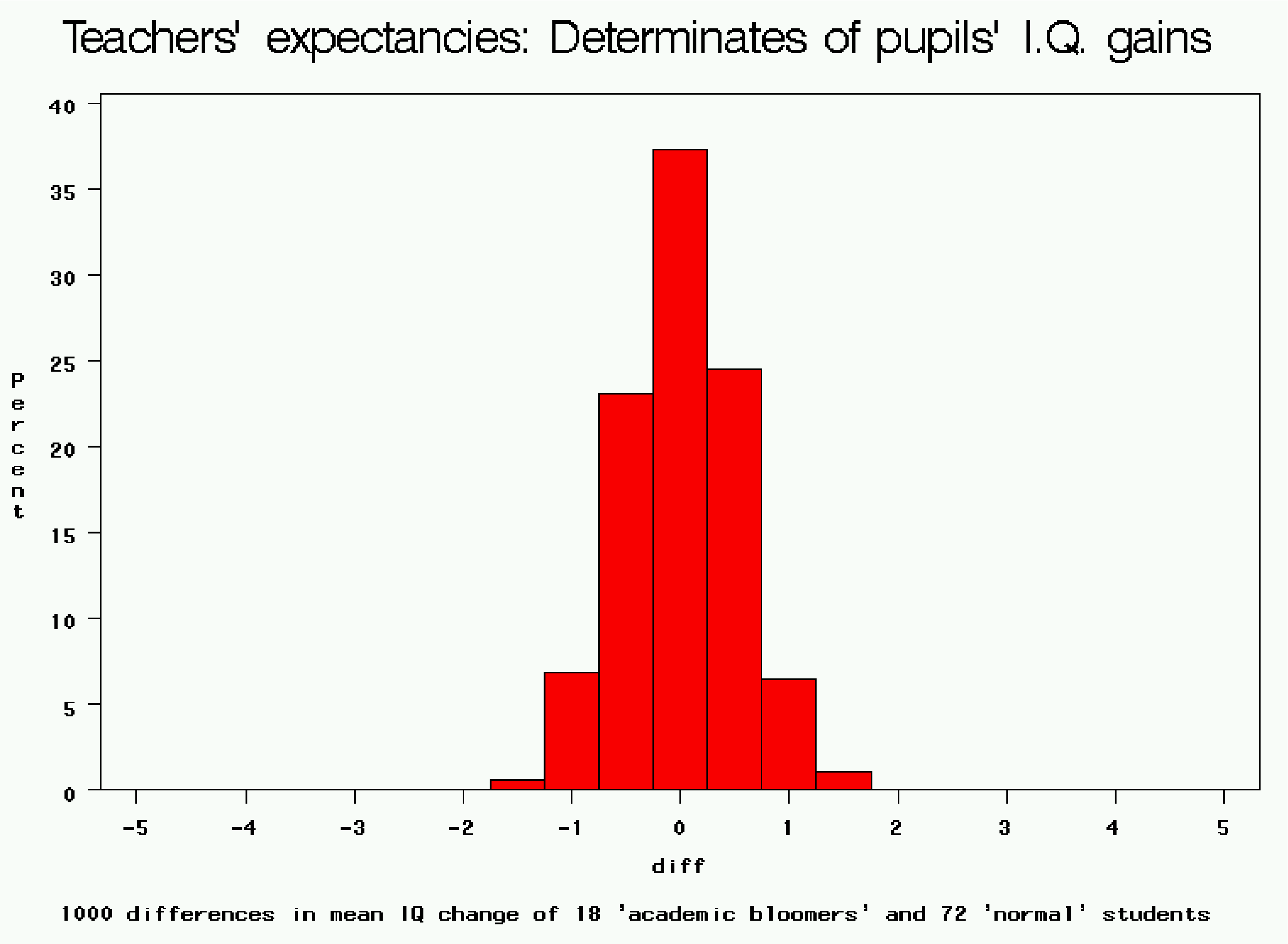
$$= \sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}} = \sqrt{\frac{4}{18} + \frac{4}{72}} = 0.52$$

Hypothesis Testing

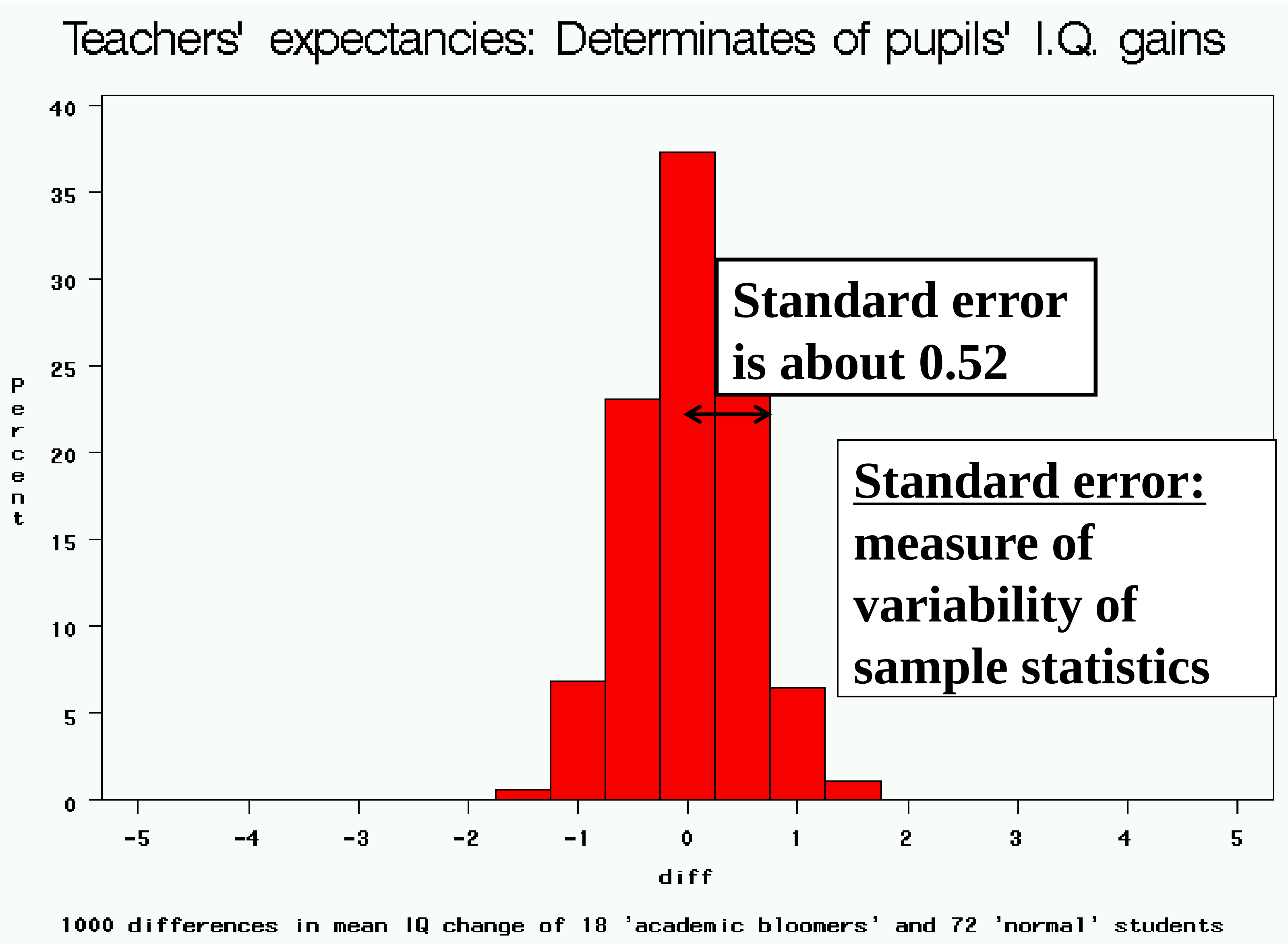
Step 2: Predict the sampling variability assuming the null hypothesis is true—computer simulation:

- In computer simulation, you simulate taking repeated samples of the same size from the same population and observe the sampling variability.
- I used computer simulation to take 1000 samples of 18 treated and 72 controls, assuming the null hypothesis (that the treatment doesn't affect IQ).

Computer Simulation Results



What is the standard error?



Hypothesis Testing

Step 3: Do an experiment

We observed a difference of 4 between treated and controls.

Hypothesis Testing

Step 4: Calculate a p-value

P-value=the probability of your data or something more extreme under the null hypothesis.

Hypothesis Testing

<u>Step 4: Calculate a p-value—mathematical theory:</u>

Difference in means follows a T distribution (which is very similar to a normal except with very small samples).

Observed difference between the groups.

A T_{88}-value of 8.0 corresponds to a p-value of <.0001

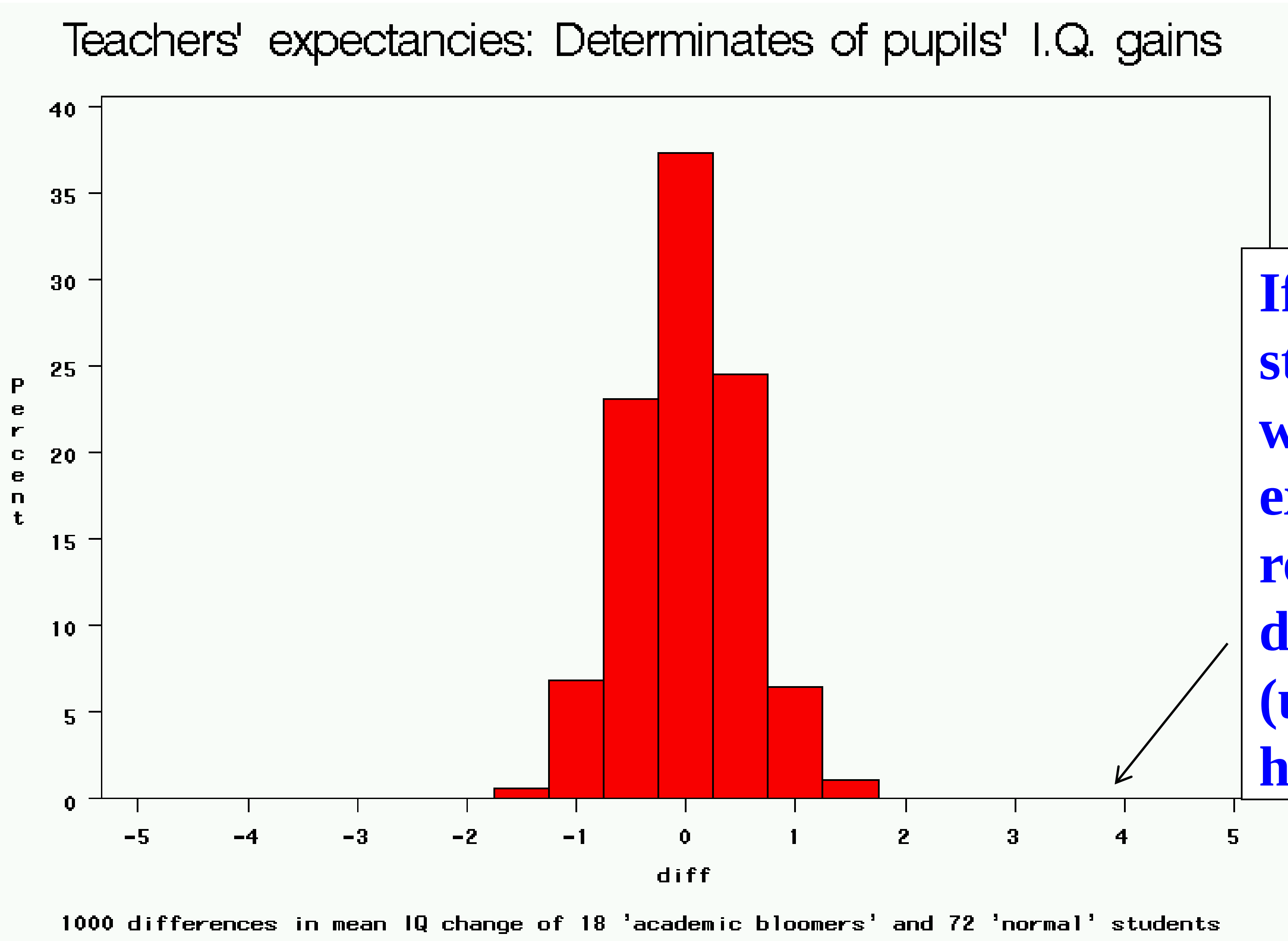
$$t_{88} = \frac{4}{.52} = 8$$

p-value <.0001

Standard error.

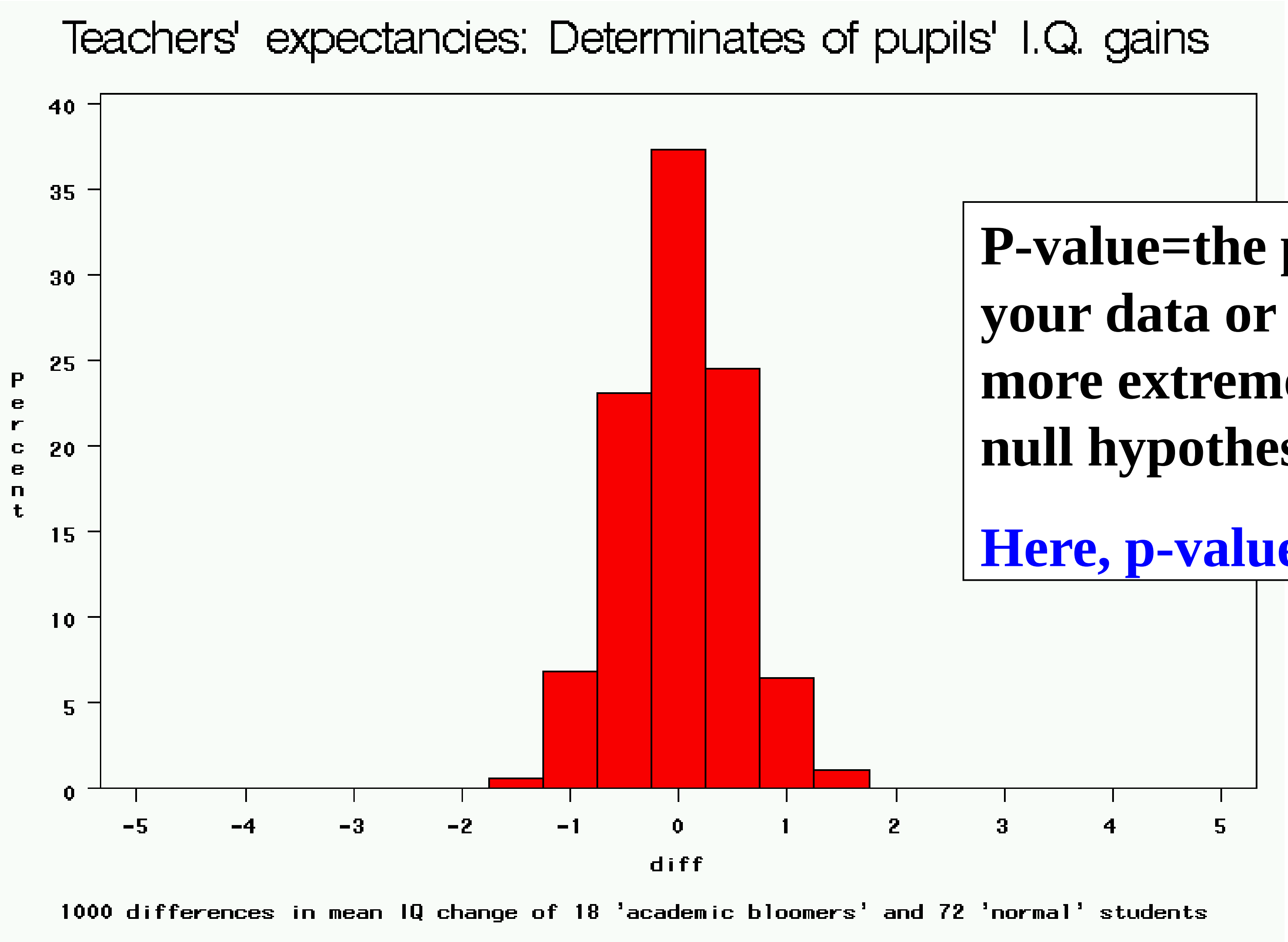
A t-curve with 88 df's has slightly wider cut-off's for 95% area (t=1.99) than a normal curve (Z=1.96)

Getting the P-value from computer simulation…



If we ran this study 1000 times we wouldn't expect to get 1 result as big as a difference of 4 (under the null hypothesis).

P-value



P-value=the probability of your data or something more extreme under the null hypothesis.

Here, p-value<.0001

Hypothesis Testing

Step 5: Reject or do not reject the null hypothesis.

Here we reject the null.

Alternative hypothesis: There is an association between being labeled as gifted and subsequent academic achievement.

What does a 4-point difference mean?

- Is it "statistically significant"? YES
- Is it clinically significant?
- Is this a causal association?

What does a 4-point difference mean?

- Is it "statistically significant"? YES
- Is it clinically significant? MAYBE
- Is this a causal association? MAYBE

Statistical significance does not necessarily imply clinical significance.

Statistical significance does not necessarily imply a cause-and-effect relationship.