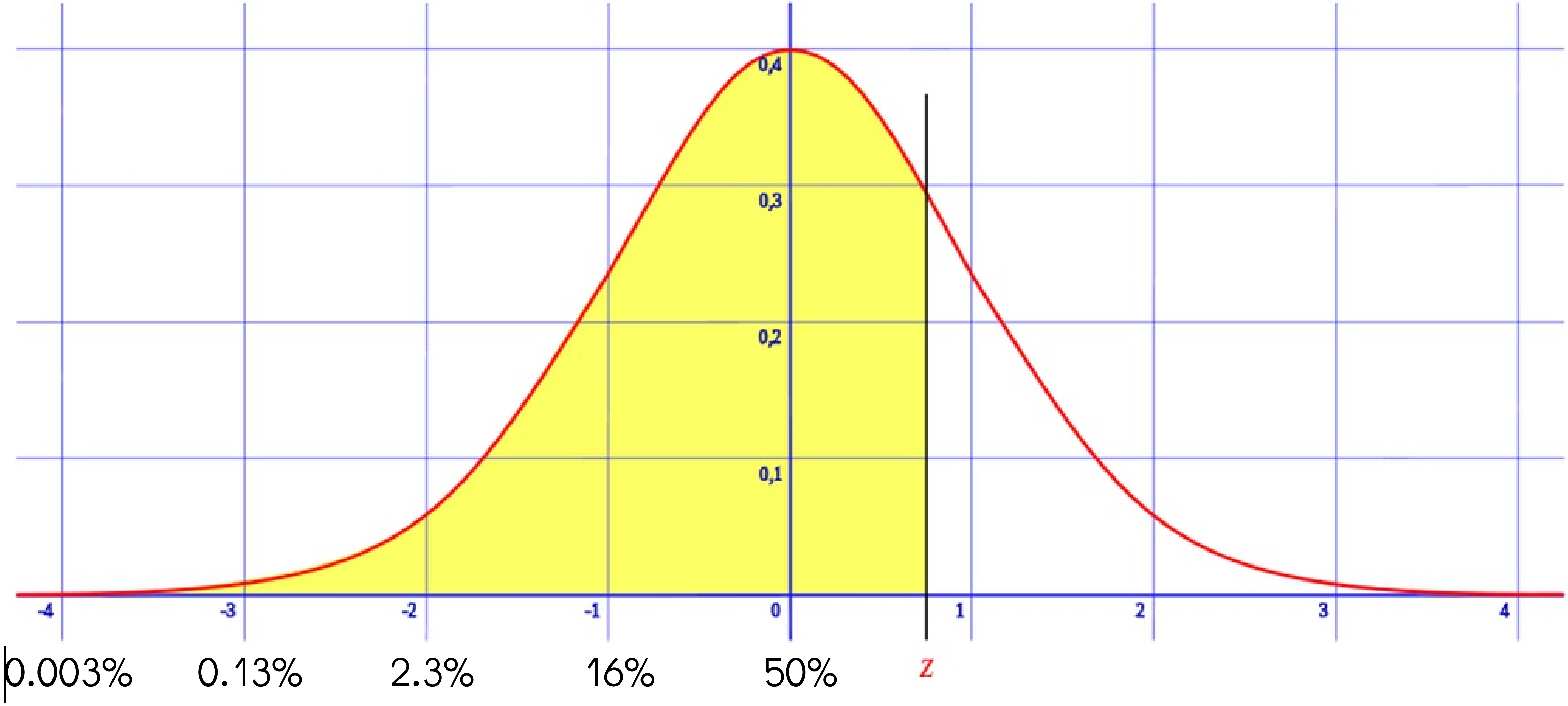
Hypothesis Testing 2

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:

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

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

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

Question: Increasing alpha:

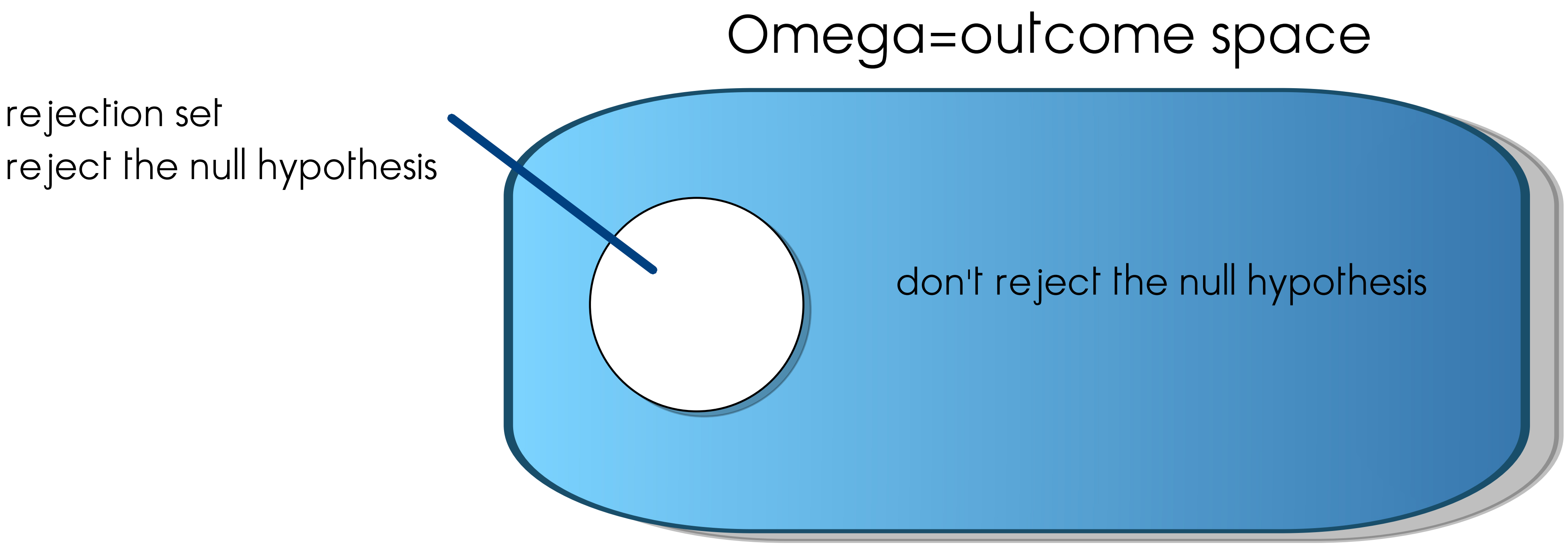
A Increases Type II error, Decreases type I

B Increases Type I Error, Decreases type II

C Decreases both.

D Increases both.

The probabilty theory of statistical tests



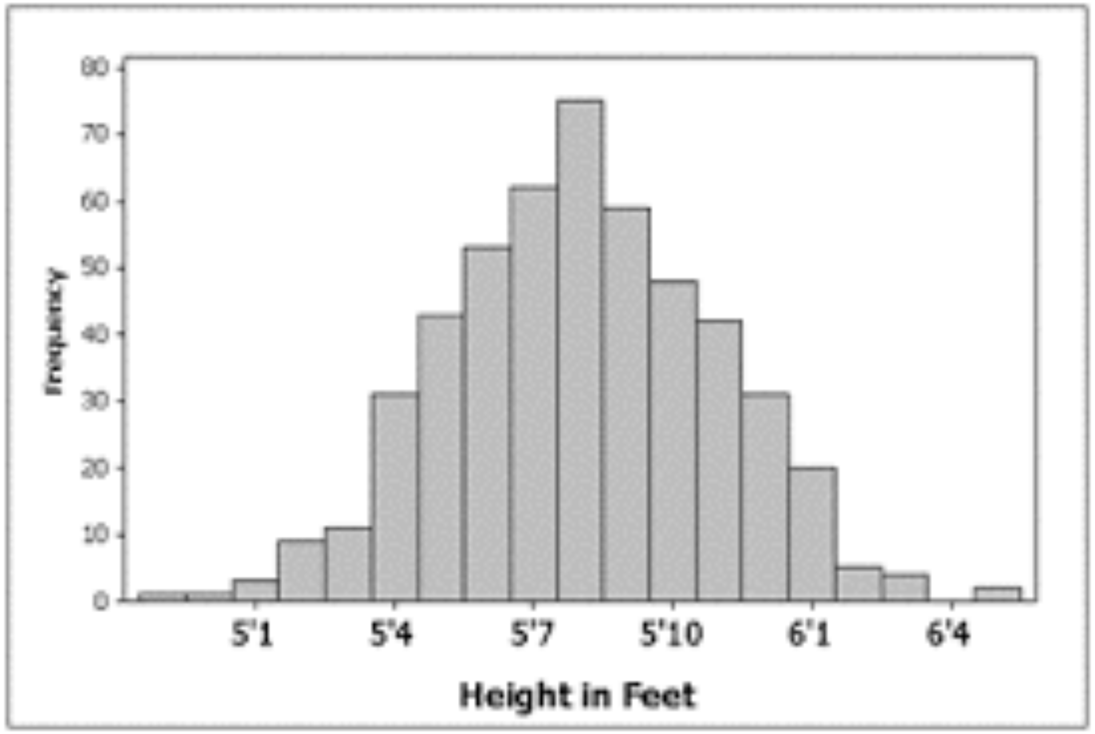
Omega=outcome space

rejection set
reject the null hypothesis

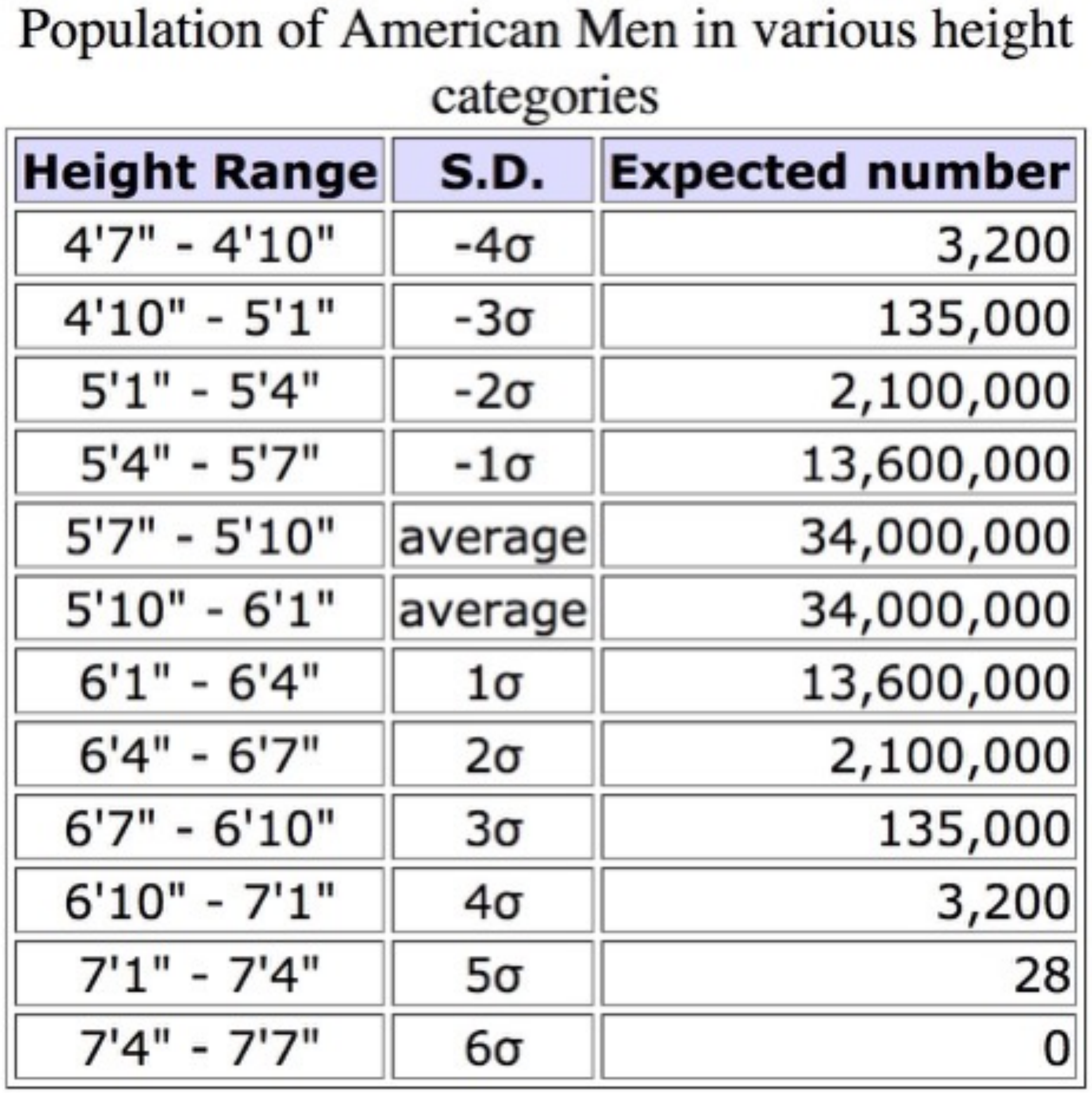
don't reject the null hypothesis

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.

Distribution of heights

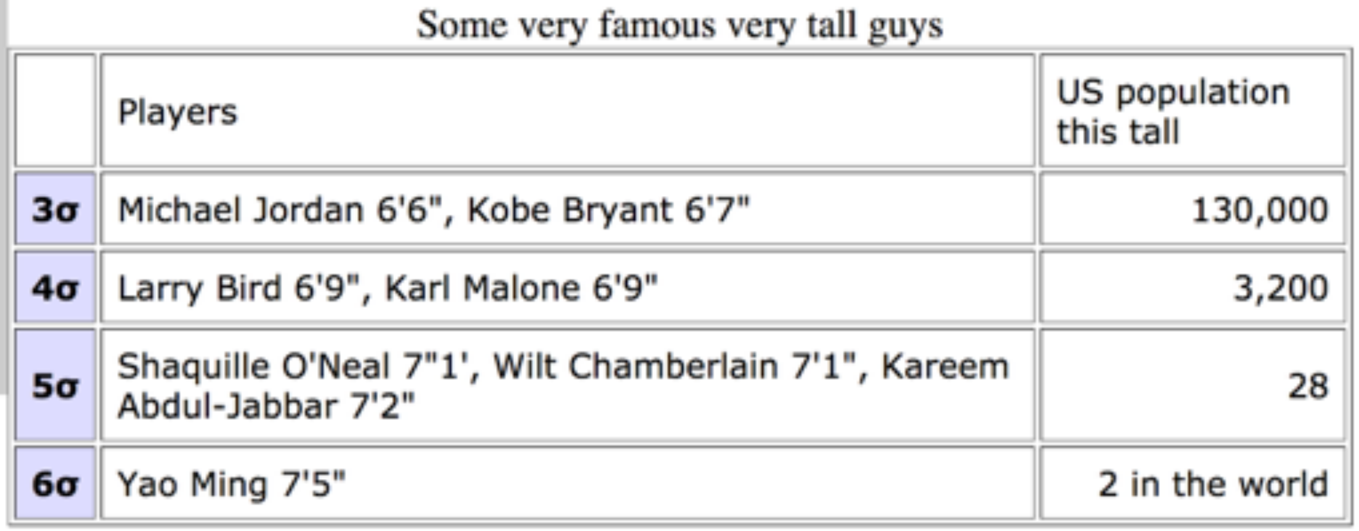


$\mu = 5'10''$ $\sigma = 3''$



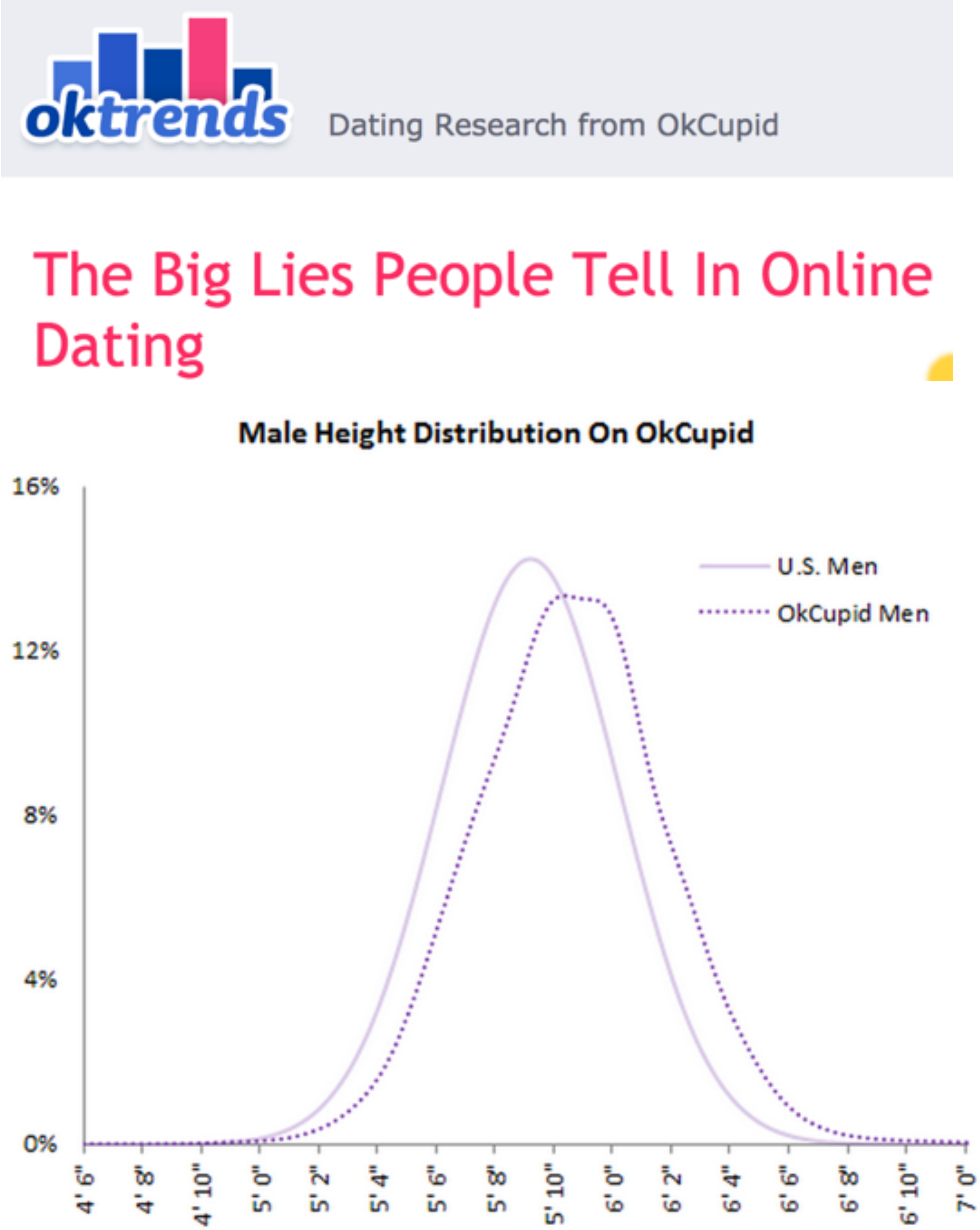
Population of American Men in various height categories

Height Range	S.D.	Expected number
4'7" - 4'10"	-4σ	3,200
4'10" - 5'1"	-3σ	135,000
5'1" - 5'4"	-2σ	2,100,000
5'4" - 5'7"	-1σ	13,600,000
5'7" - 5'10"	average	34,000,000
5'10" - 6'1"	average	34,000,000
6'1" - 6'4"	1σ	13,600,000
6'4" - 6'7"	2σ	2,100,000
6'7" - 6'10"	3σ	135,000
6'10" - 7'1"	4σ	3,200
7'1" - 7'4"	5σ	28
7'4" - 7'7"	6σ	0



Some very famous very tall guys

	Players	US population this tall
3σ	Michael Jordan 6'6", Kobe Bryant 6'7"	130,000
4σ	Larry Bird 6'9", Karl Malone 6'9"	3,200
5σ	Shaquille O'Neal 7"1', Wilt Chamberlain 7'1", Kareem Abdul-Jabbar 7'2"	28
6σ	Yao Ming 7'5"	2 in the world

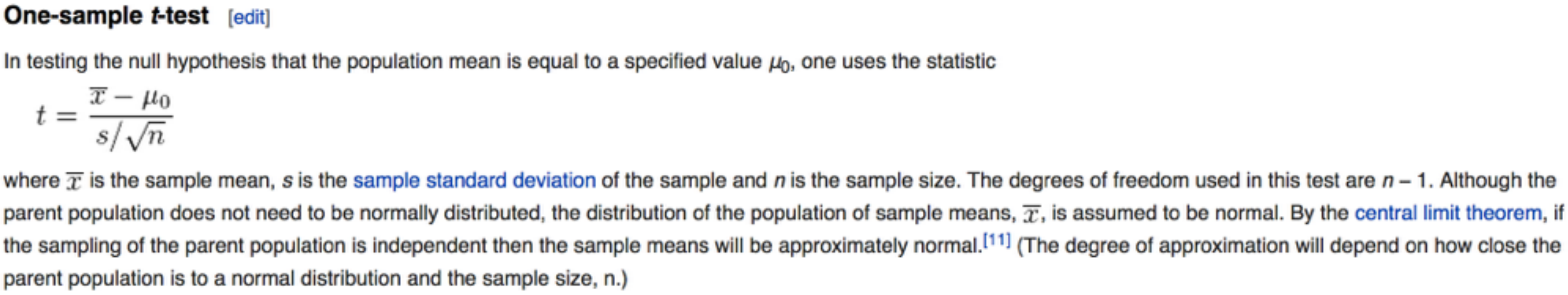


oktrends Dating Research from OkCupid

The Big Lies People Tell In Online Dating

Male Height Distribution On OkCupid

Question: Are men telling their true height on dating sites?



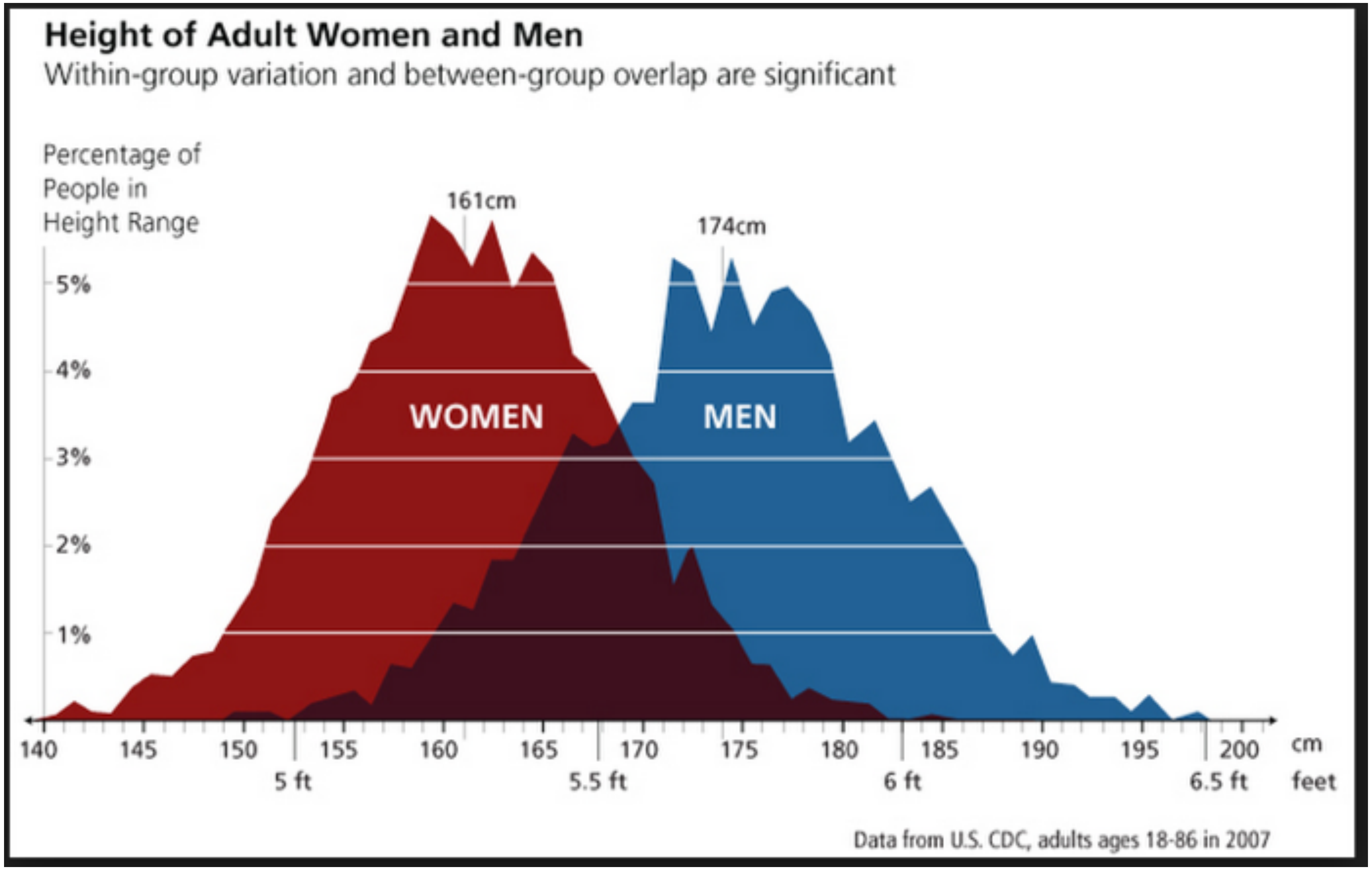
One-sample *t*-test [edit]

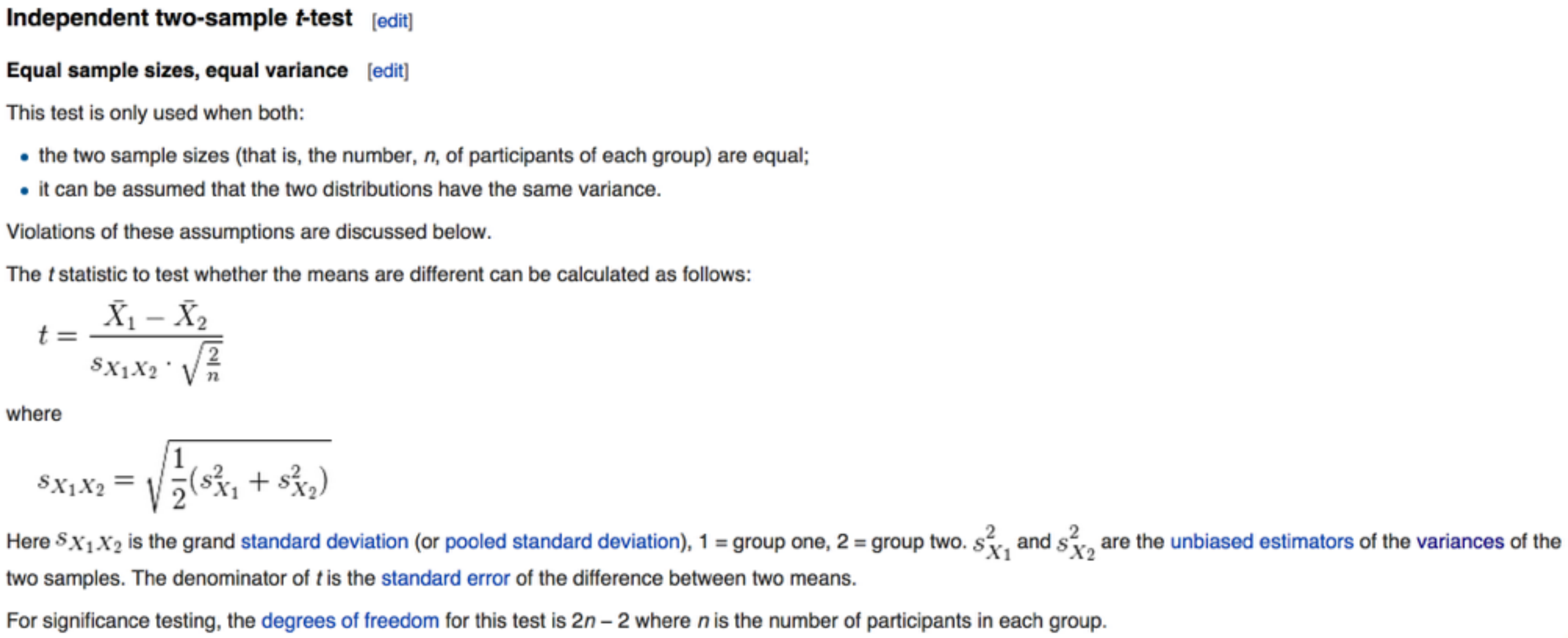
In testing the null hypothesis that the population mean is equal to a specified value μ_0, one uses the statistic

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

where $\overline{x}$ is the sample mean, *s* is the sample standard deviation of the sample and *n* is the sample size. The degrees of freedom used in this test are *n* – 1. Although the parent population does not need to be normally distributed, the distribution of the population of sample means, $\overline{x}$, is assumed to be normal. By the central limit theorem, if the sampling of the parent population is independent then the sample means will be approximately normal.[11] (The degree of approximation will depend on how close the parent population is to a normal distribution and the sample size, n.)

Question: are men, on average, taller than women?





Independent two-sample *t*-test [edit]

Equal sample sizes, equal variance [edit]

This test is only used when both:

- the two sample sizes (that is, the number, *n*, of participants of each group) are equal;
- it can be assumed that the two distributions have the same variance.

Violations of these assumptions are discussed below.

The *t* statistic to test whether the means are different can be calculated as follows:

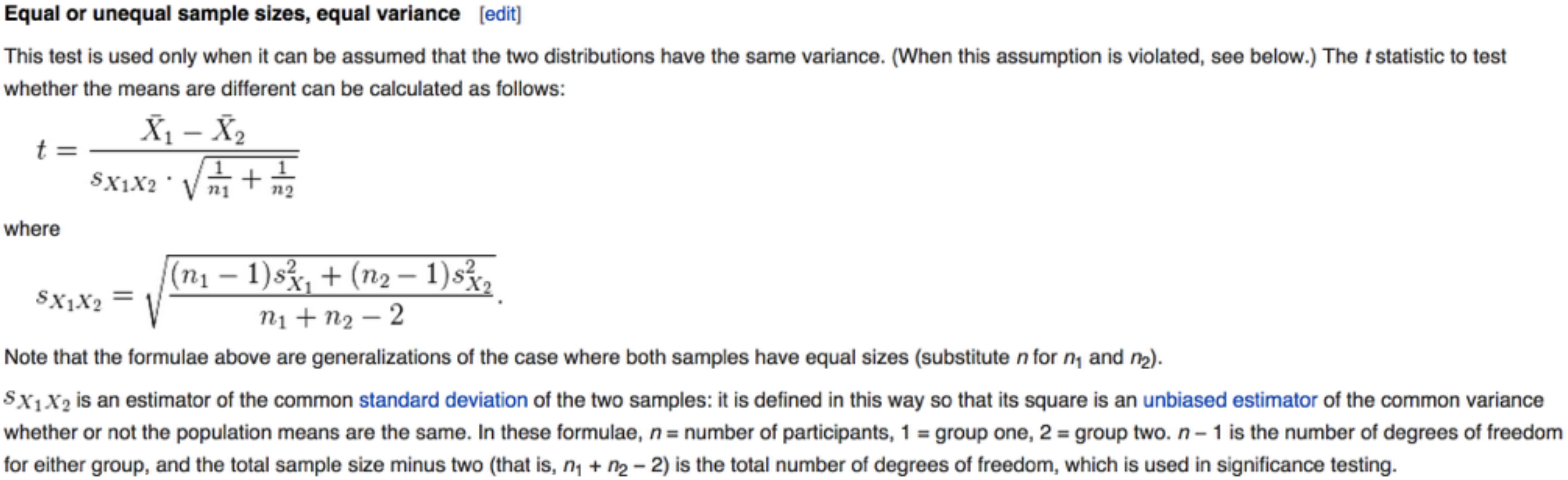
$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{X_1X_2} \cdot \sqrt{\frac{2}{n}}}$$

where

$$s_{X_1X_2} = \sqrt{\frac{1}{2}(s_{X_1}^2 + s_{X_2}^2)}$$

Here $s_{X_1X_2}$ is the grand standard deviation (or pooled standard deviation), 1 = group one, 2 = group two. $s_{X_1}^2$ and $s_{X_2}^2$ are the unbiased estimators of the variances of the two samples. The denominator of *t* is the standard error of the difference between two means.

For significance testing, the degrees of freedom for this test is 2*n* – 2 where *n* is the number of participants in each group.



Equal or unequal sample sizes, equal variance [edit]

This test is used only when it can be assumed that the two distributions have the same variance. (When this assumption is violated, see below.) The *t* statistic to test whether the means are different can be calculated as follows:

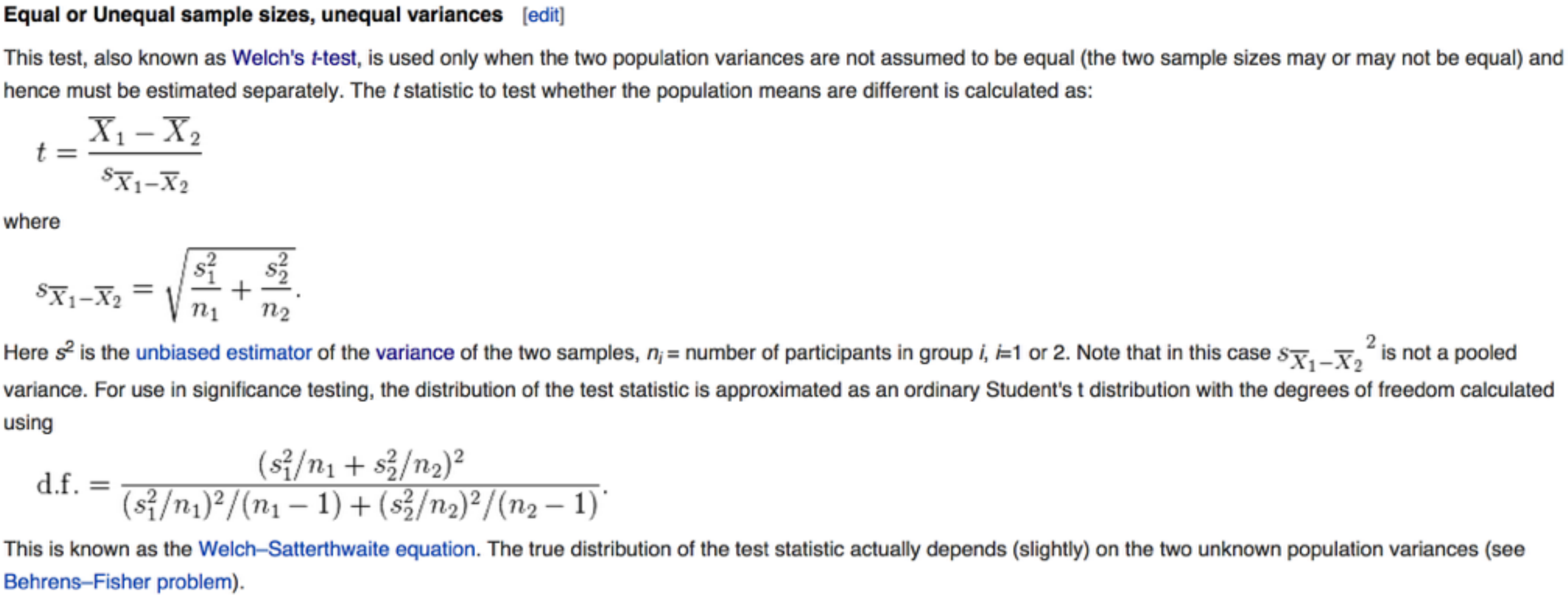
$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{X_1X_2} \cdot \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

where

$$s_{X_1X_2} = \sqrt{\frac{(n_1-1)s_{X_1}^2+(n_2-1)s_{X_2}^2}{n_1+n_2-2}}.$$

Note that the formulae above are generalizations of the case where both samples have equal sizes (substitute *n* for n_1 and n_2).

$s_{X_1X_2}$ is an estimator of the common standard deviation of the two samples: it is defined in this way so that its square is an unbiased estimator of the common variance whether or not the population means are the same. In these formulae, *n* = number of participants, 1 = group one, 2 = group two. *n* – 1 is the number of degrees of freedom for either group, and the total sample size minus two (that is, $n_1 + n_2 - 2$) is the total number of degrees of freedom, which is used in significance testing.



Equal or Unequal sample sizes, unequal variances [edit]

This test, also known as Welch's *t*-test, is used only when the two population variances are not assumed to be equal (the two sample sizes may or may not be equal) and hence must be estimated separately. The *t* statistic to test whether the population means are different is calculated as:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}}$$

where

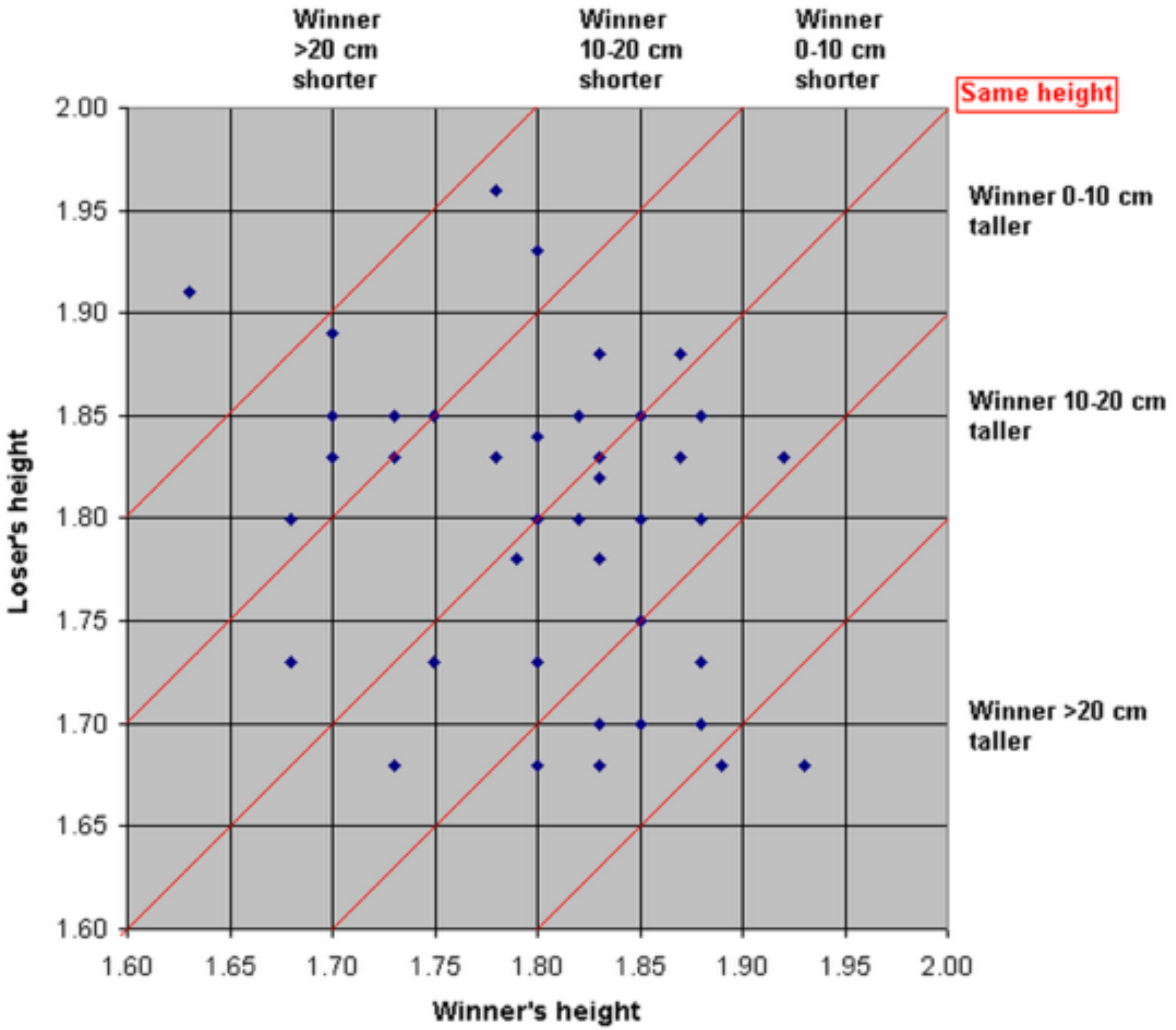
$$s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

Here s^2 is the unbiased estimator of the variance of the two samples, n_i = number of participants in group i, i=1 or 2. Note that in this case ${s_{\overline{X}_1 - \overline{X}_2}}^2$ is not a pooled variance. For use in significance testing, the distribution of the test statistic is approximated as an ordinary Student's t distribution with the degrees of freedom calculated using

$$\mathrm{d.f.} = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1-1) + (s_2^2/n_2)^2/(n_2-1)}.$$

This is known as the Welch–Satterthwaite equation. The true distribution of the test statistic actually depends (slightly) on the two unknown population variances (see Behrens–Fisher problem).

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



Dependent *t*-test for paired samples [edit]

This test is used when the samples are dependent; that is, when there is only one sample that has been tested twice (repeated measures) or when there are two samples that have been matched or "paired". This is an example of a paired difference test.

$$t = \frac{\overline{X}_D - \mu_0}{s_D/\sqrt{n}}.$$

For this equation, the differences between all pairs must be calculated. The pairs are either one person's pre-test and post-test scores or between pairs of persons matched into meaningful groups (for instance drawn from the same family or age group: see table). The average (X_D) and standard deviation (s_D) of those differences are used in the equation. The constant μ_0 is non-zero if you want to test whether the average of the difference is significantly different from μ_0. The degree of freedom used is $n - 1$.

Example of matched pairs

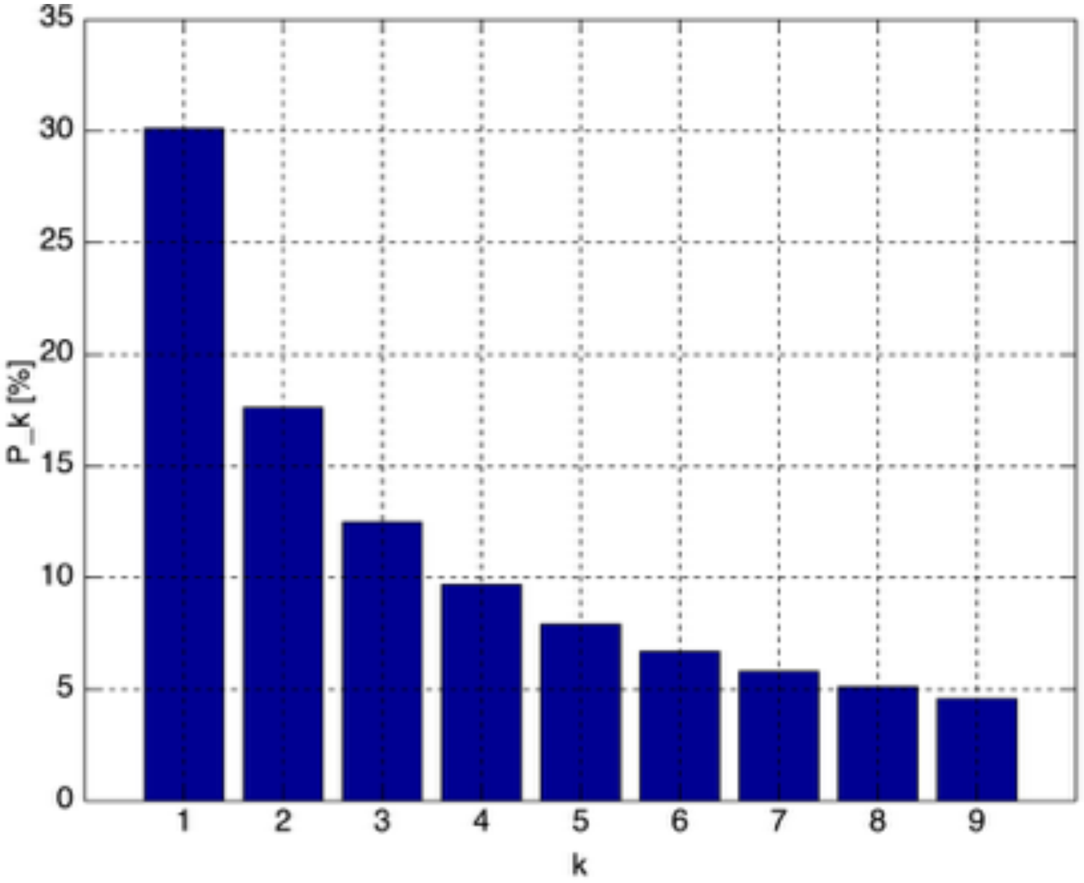
Pair	Name	Age	Test
1	John	35	250
1	Jane	36	340
2	Jimmy	22	460
2	Jessy	21	200

Example of repeated measures

Number	Name	Test 1	Test 2
1	Mike	35%	67%
2	Melanie	50%	46%
3	Melissa	90%	86%
4	Mitchell	78%	91%

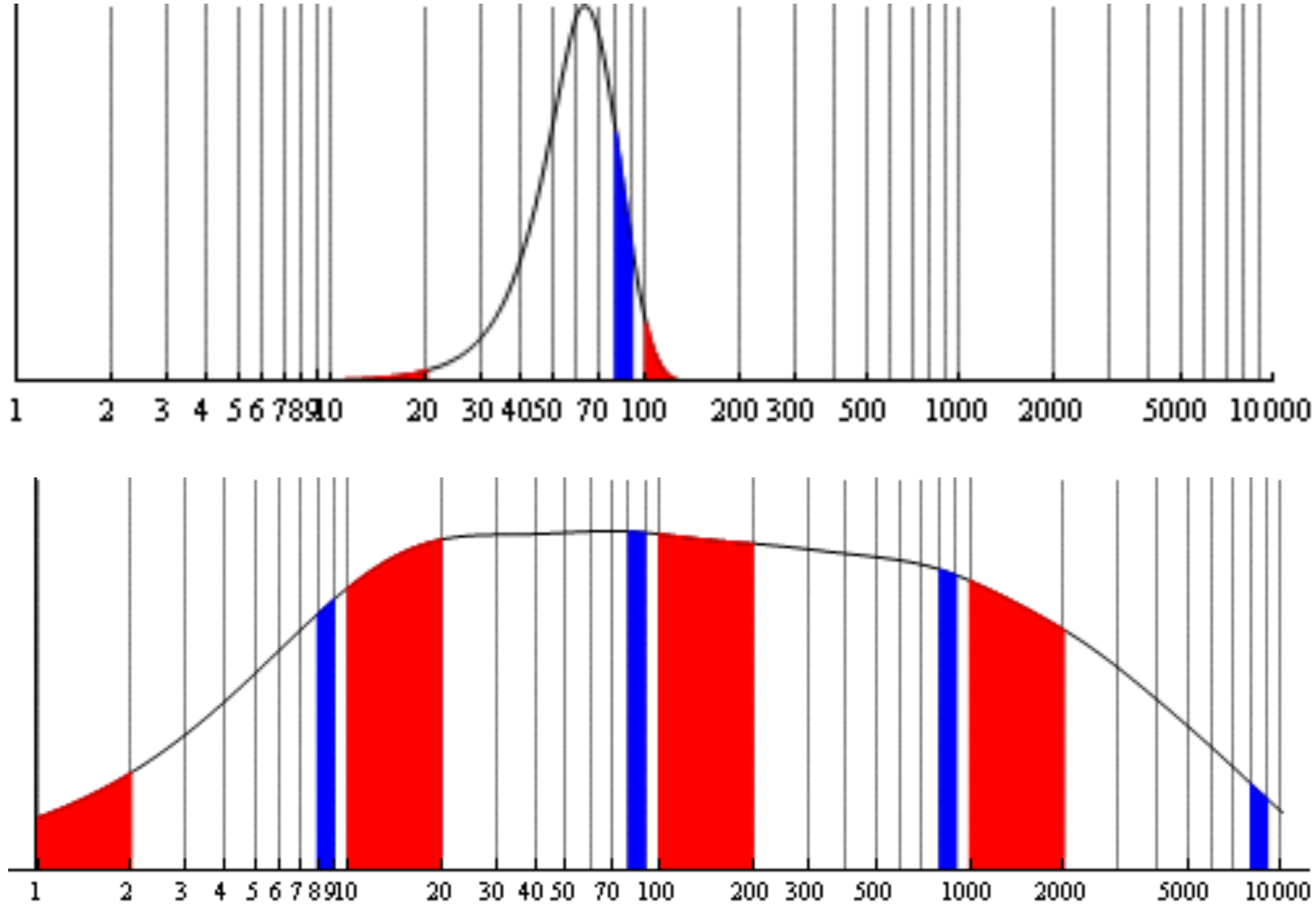
The Benford distribution

Describes the distribution of the most significant digit in large collection of financial numbers



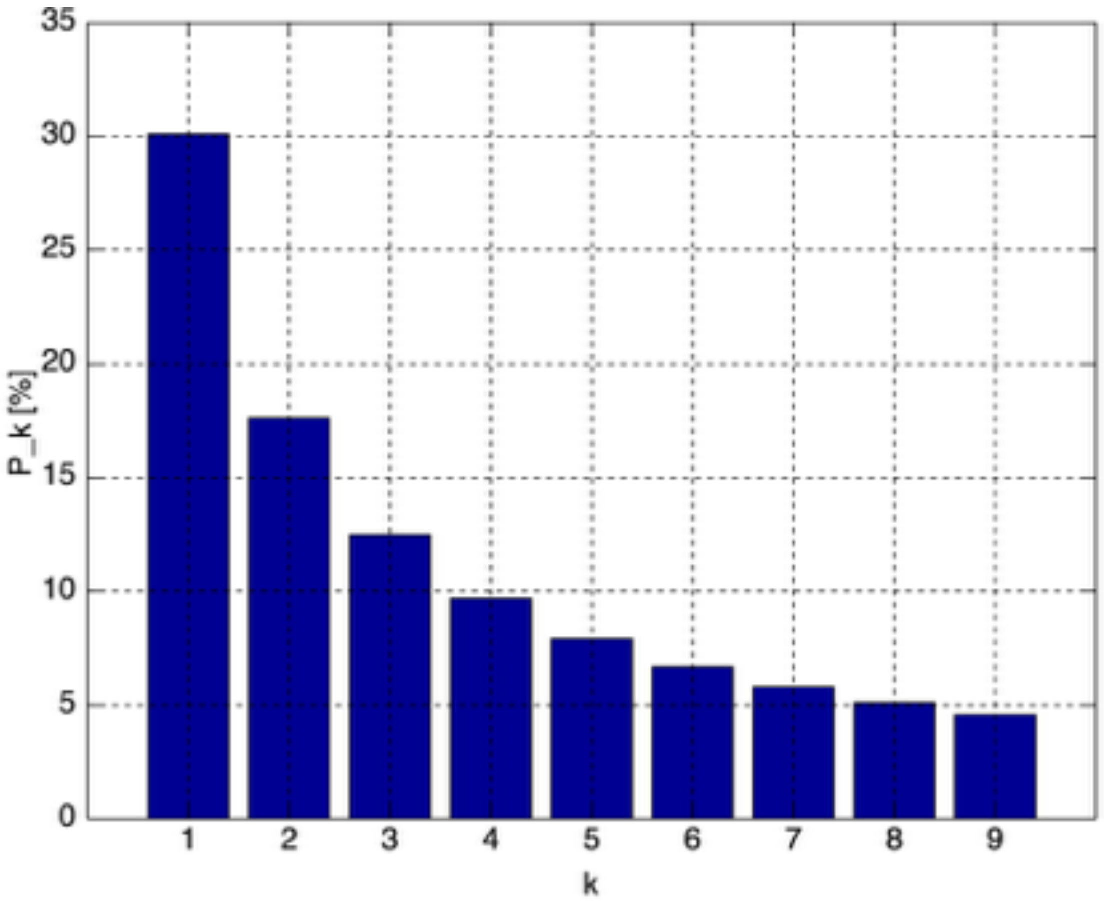
How can we explain Benford law?

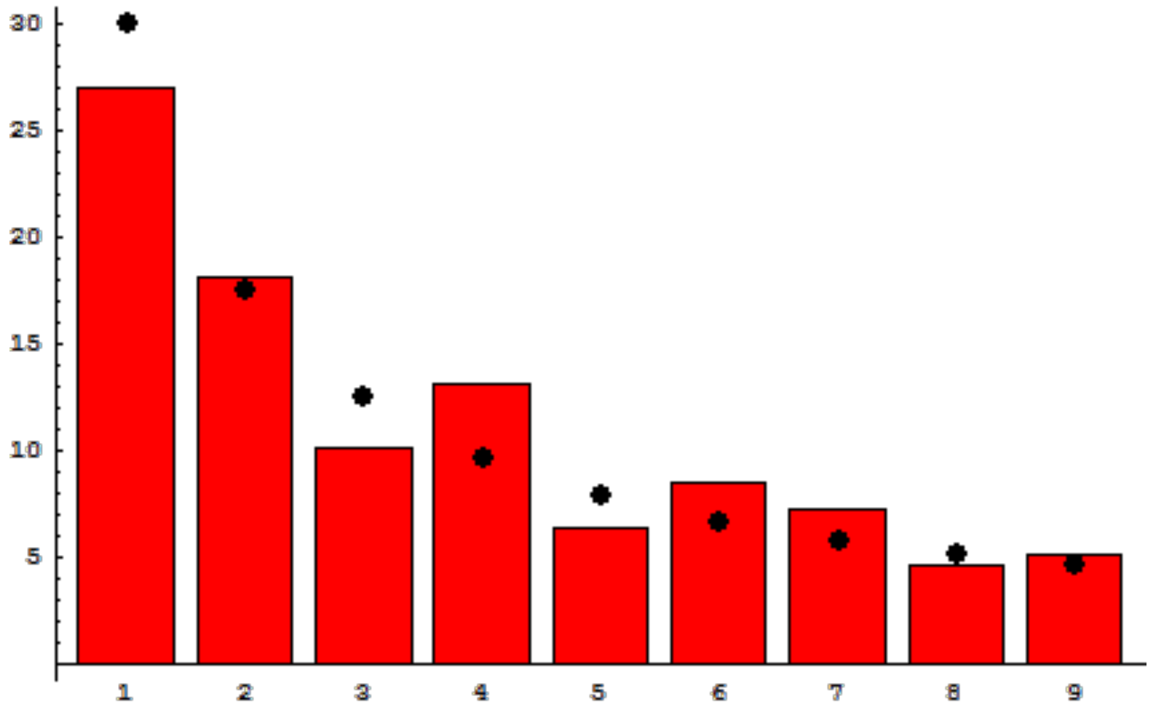
- As currency units are arbitrary, changing the definition of a the currency unit does not fundamentally change the distribution.
- The distribution is approximately constant on a logarithmic scale.
- If the distribution spans several orders of magnitude (from single dollars thousands of dollars) we get the Benford distribution



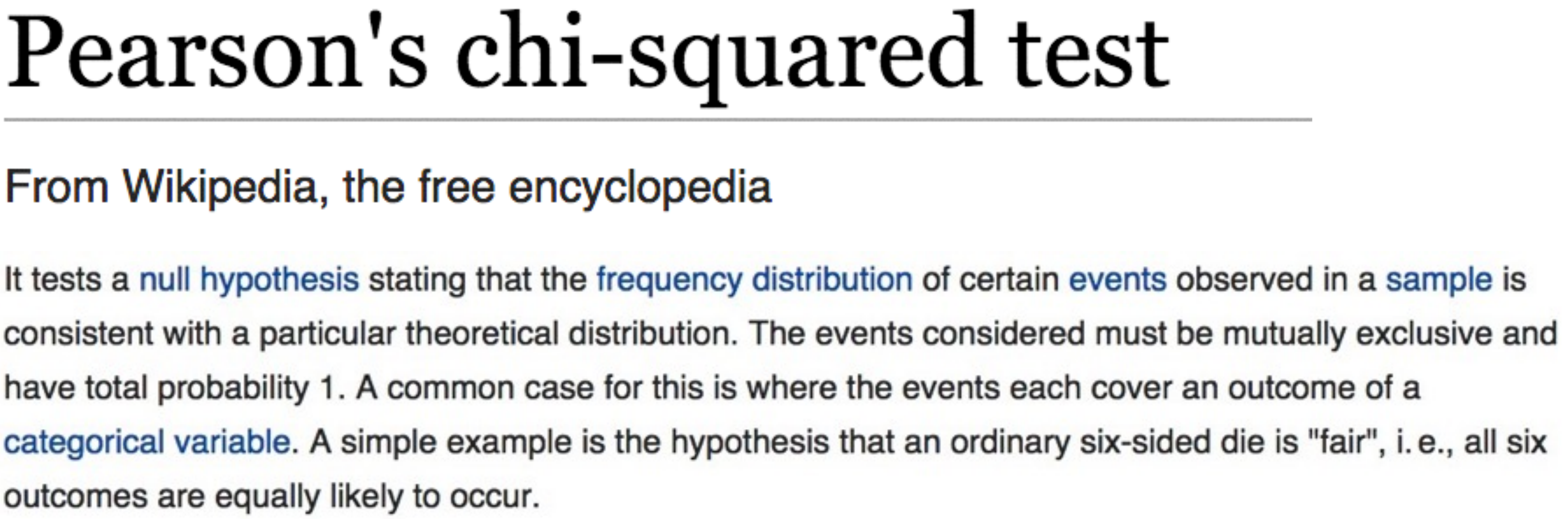
Can we detect accounting fraud using the Benford distribution?

Null Hyp: dist is Benford





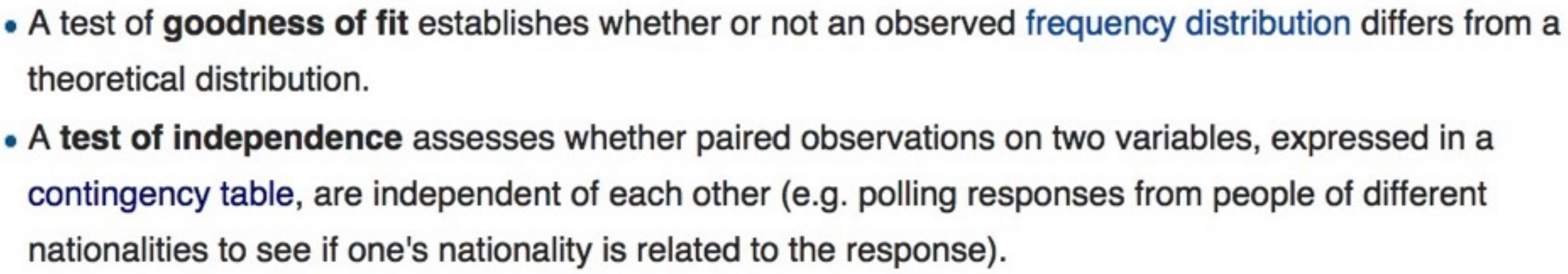
Distribution of top digits in a tax return



Pearson's chi-squared test

From Wikipedia, the free encyclopedia

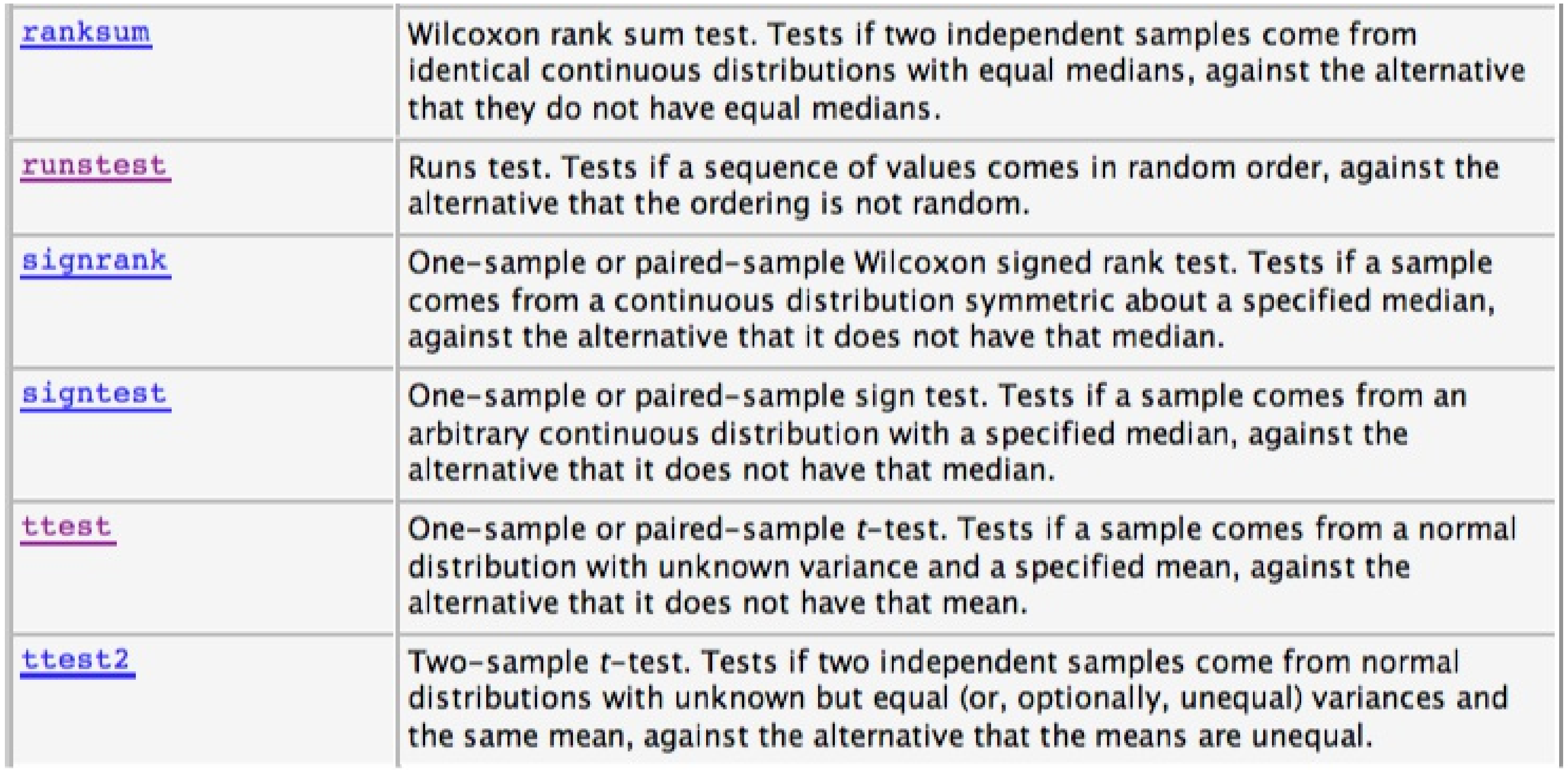
It tests a null hypothesis stating that the frequency distribution of certain events observed in a sample is consistent with a particular theoretical distribution. The events considered must be mutually exclusive and have total probability 1. A common case for this is where the events each cover an outcome of a categorical variable. A simple example is the hypothesis that an ordinary six-sided die is "fair", i. e., all six outcomes are equally likely to occur.



- A test of **goodness of fit** establishes whether or not an observed frequency distribution differs from a theoretical distribution.
- A **test of independence** assesses whether paired observations on two variables, expressed in a contingency table, are independent of each other (e.g. polling responses from people of different nationalities to see if one's nationality is related to the response).

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.

Multiple Hypothesis testing

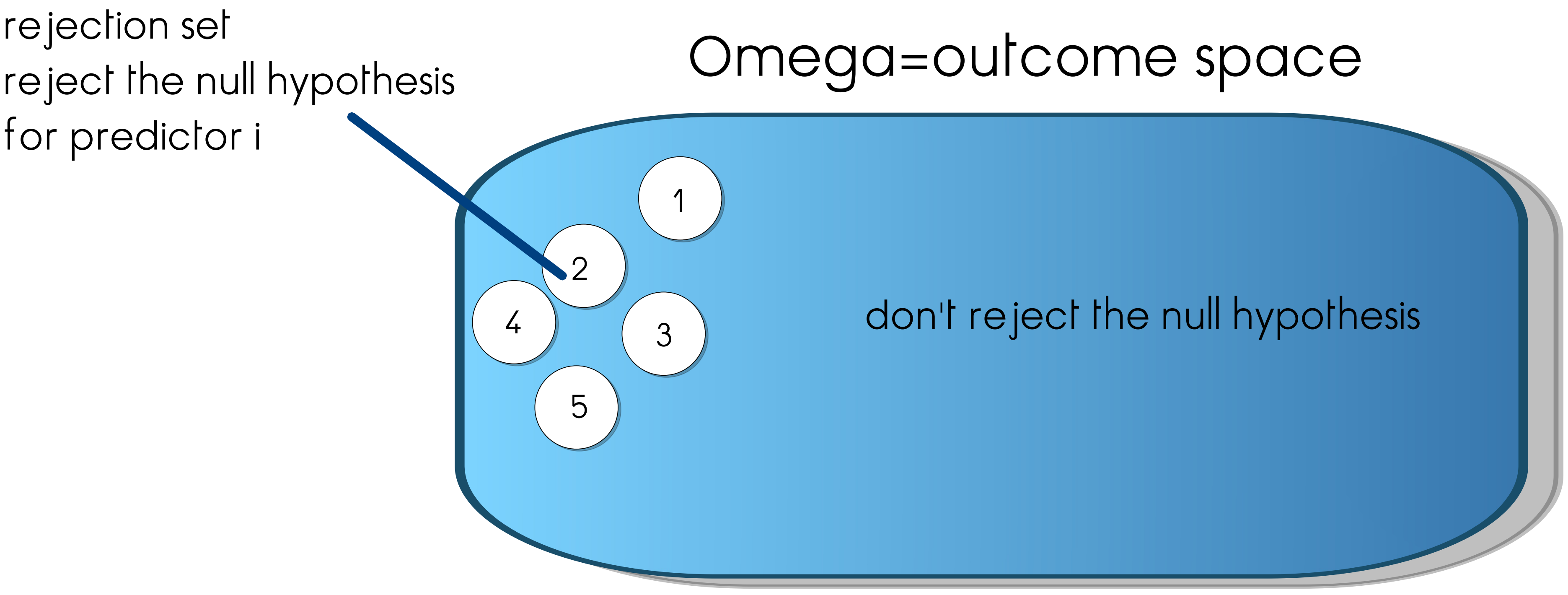
Consider the online ad problem, our goal is to maximize click-through rate. Our null hypothesis is that nothing performs better than picking one of the ads uniformly at random each time.

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.