FEEG6017 lecture: Hypothesis testing, t-tests, p-values, type-I and type-II errors

mb8@ecs.soton.ac.uk

The t-test

- This lecture introduces the t-test -- our first real statistical test -- and the related t-distribution.

- The t-test is used for such things as:
  - determining the likelihood that a sample comes from a population with a specified mean
  - deciding whether two samples come from the same population or not, i.e., do their means appear to be significantly different?

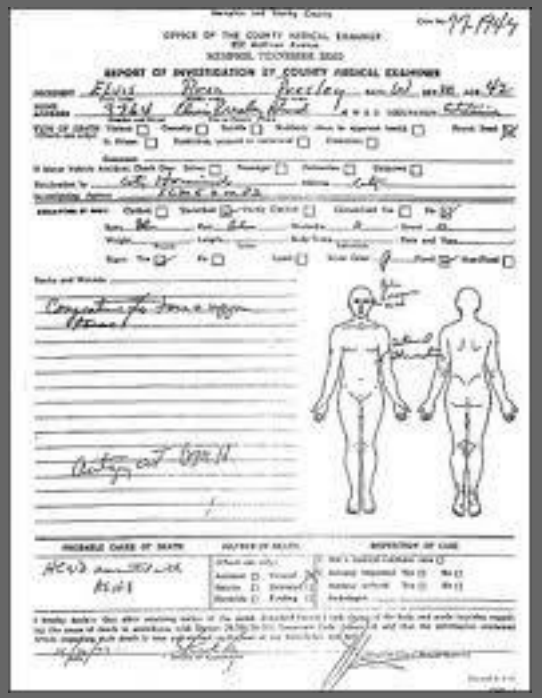
- The t-test is not mysterious; its logic follows on from the sampling experiments we discussed last time.

Hypothesis testing

- Before getting into the details of the t-test, we need to place it in the wider context of statistical hypothesis testing.
- You may already know the terms "null hypothesis" and "alternative hypothesis".
- These terms fit into the pattern of statistical inference we discussed right at the start of the module: suppose that the world works in a certain way, calculate the chances of seeing our data in a world that really did work that way, use the result of the calculation to reflect on how much we like our original supposition.

Hypothesis testing

- Consider a forensic example: a person has died and we're not sure whether it was natural causes (H_{null}) or poisoning ($H_{alternative}$).



- If we suppose it was natural causes, we can ask what the probability is of a person of that age, fitness, medical history, etc., having their heart stop without warning.

- If that probability is very small, we will naturally look more favourably on the poisoning hypothesis.

Hypothesis testing: investment example

- Another example: you inherit some money, and you ask a friend who knows about the stock market to invest it for you.
- One year later, your friend tells you all the money is gone.

Hypothesis testing: investment example

- How angry are you? Do you trust your friend?
- The null hypothesis: your friend simply had an unlucky run of investments. It could have happened to anyone.
- The alternative hypothesis: your friend has cheated you.

Hypothesis testing: investment example

- What we really want to know is whether your friend has cheated you, of course.
- We might express that as "what's the probability that he cheated me, given that he's claiming to have lost all the money?" (The Bayesian version of the question.)
- We could also say "he either cheated me or he didn't: which one should I believe given that he's claiming to have lost all the money?" (The frequentist version of the question.)
- But there may be no direct way to know the answer to those questions.

Hypothesis testing: investment example

- We could turn the question around and say "He is either unlucky (H_{null}) or a cheat (H_{alt}). Under each of those hypotheses, what are the chances of seeing a loss of all the money?"

- Consider H_{alt}. Assume he's cheating you; what are the chances that he'd report a total loss of the money after one year? This is quite hard to answer: it depends on just how sophisticated a cheat he is.

- Consider H_{null}. What are the chances that an honest investor would lose the money in the market as it's been over the last year? A more tractable question...

Hypothesis testing: investment example

- You could ask some independent experts just how tough the market has been that year.
- You could simulate a range of investment strategies using historical market data.
- You could look empirically at how many people out of the wider population of investors lost all their money over the last year.

Hypothesis testing: investment example

- Using one or all of these methods, let's say you find that it's been a very tough year, and in fact there's a 50% chance of an honest investor having lost all their money.
- It's therefore hard to rule out the null hypothesis. You're forced to conclude something like "He may well be honest."
- But let's say you find that it's been a great year, and that only 1 honest investor in 1000 lost money.
- If you want to hang onto the null hypothesis (honesty) under these circumstances, you have to accept that a very unlikely thing has happened.

Hypothesis testing: investment example

- So because of the small probability of the observed data (total loss) given the hypothesis (honesty) you are nudged towards the conclusion that the alternative hypothesis (cheating) is likely to be true.

- Let's say you're in this situation all the time: you run a hedge fund and you have many traders working for you, any of whom might decide to cheat you one year.

- You could choose to adopt some threshold for p(total loss given assumption of honesty) that means you'll reject the honesty assumption.

Hypothesis testing: investment example

- Let's say you decide (somewhat arbitrarily) to go with 1% as your threshold.

- If someone loses money, and the probability of an honest trader losing their funds is 1% or lower, you conclude that the person is not honest.

- This is not a bad decision rule, but note that it can't be perfect. Sometimes an honest person will get extremely unlucky and then be unfairly accused by you, and sometimes a cheat will steal from you but p(loss | honesty) will be above the 1% threshold.

Statistical tests

- In the terms of our investment example, a statistical test is just a procedure for calculating p(loss | honesty) or its equivalent.
- The t-test is one such test.
- In general we want to know p (observed data | H_{null}). This is all a "p-value" is.
- Because of a throwaway remark by Ronald Fisher, the threshold for rejecting H_{null} has been set at $p \leq 5\%$ in many fields. There is nothing magical about this value, however.

Logic of the one-sample t-test

- We found in previous lectures that if we take repeated samples from a population, even of quite small size, the distribution of the *means of those samples* quickly approximates the bell curve of the normal distribution.

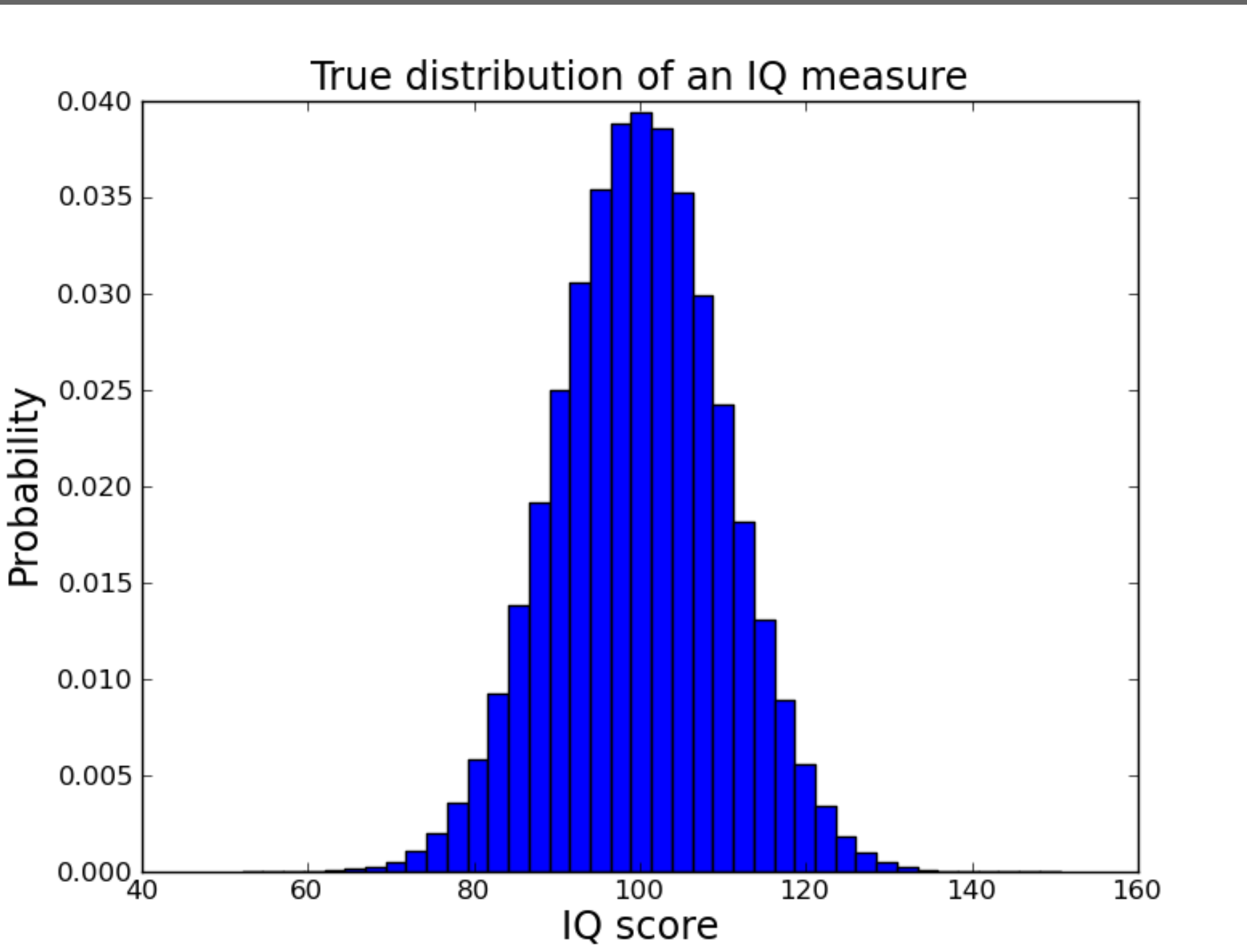
- If we're dealing with big sample sizes, the distribution of the sample means is as close as makes no difference to being the normal distribution.

- But the match is not perfect for small samples though. This is where the t-distribution comes in.

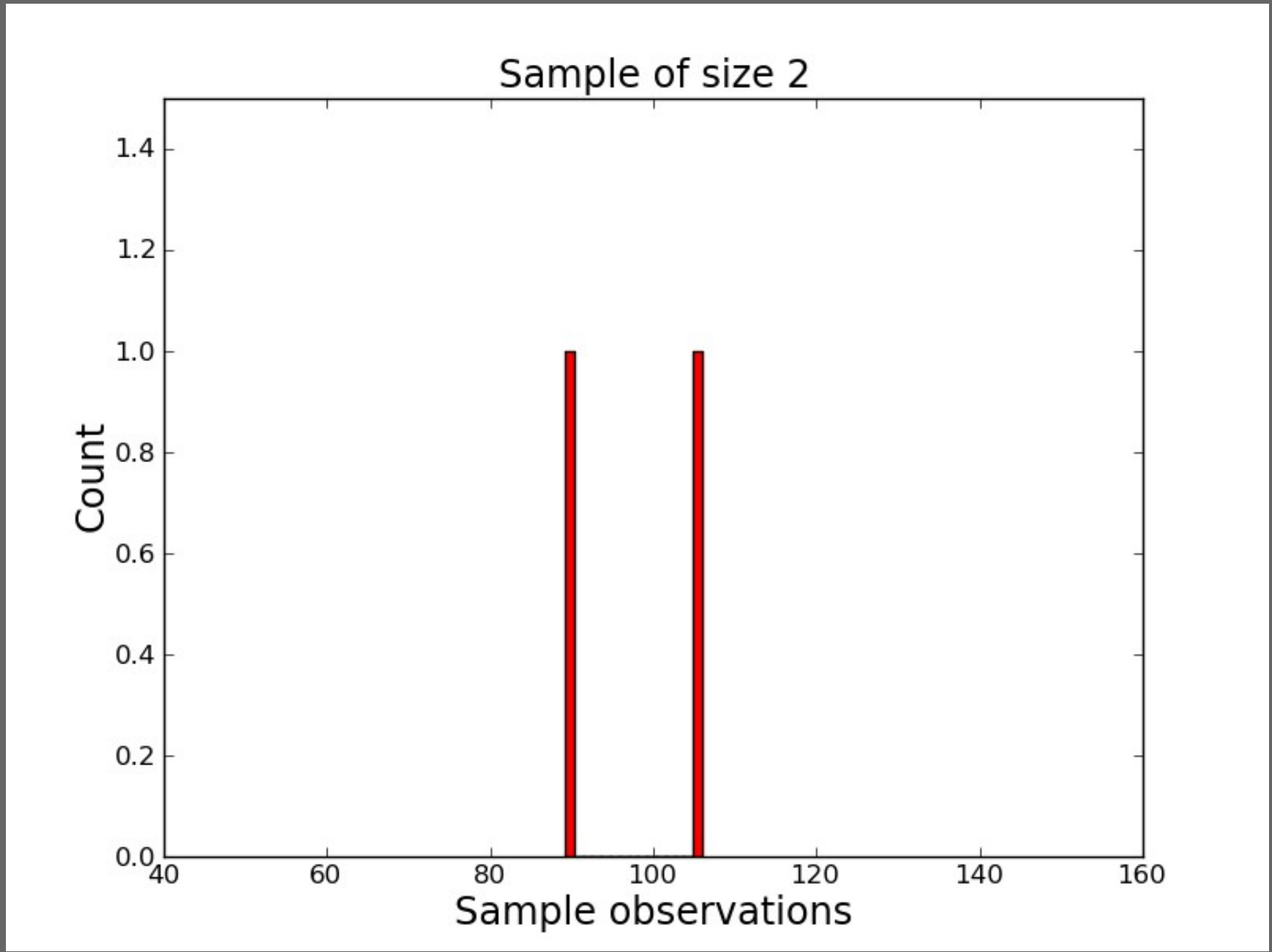
Fictional data exercise

- We can demonstrate the logic of the t-test by working through an exercise in sampling from a fictional distribution.

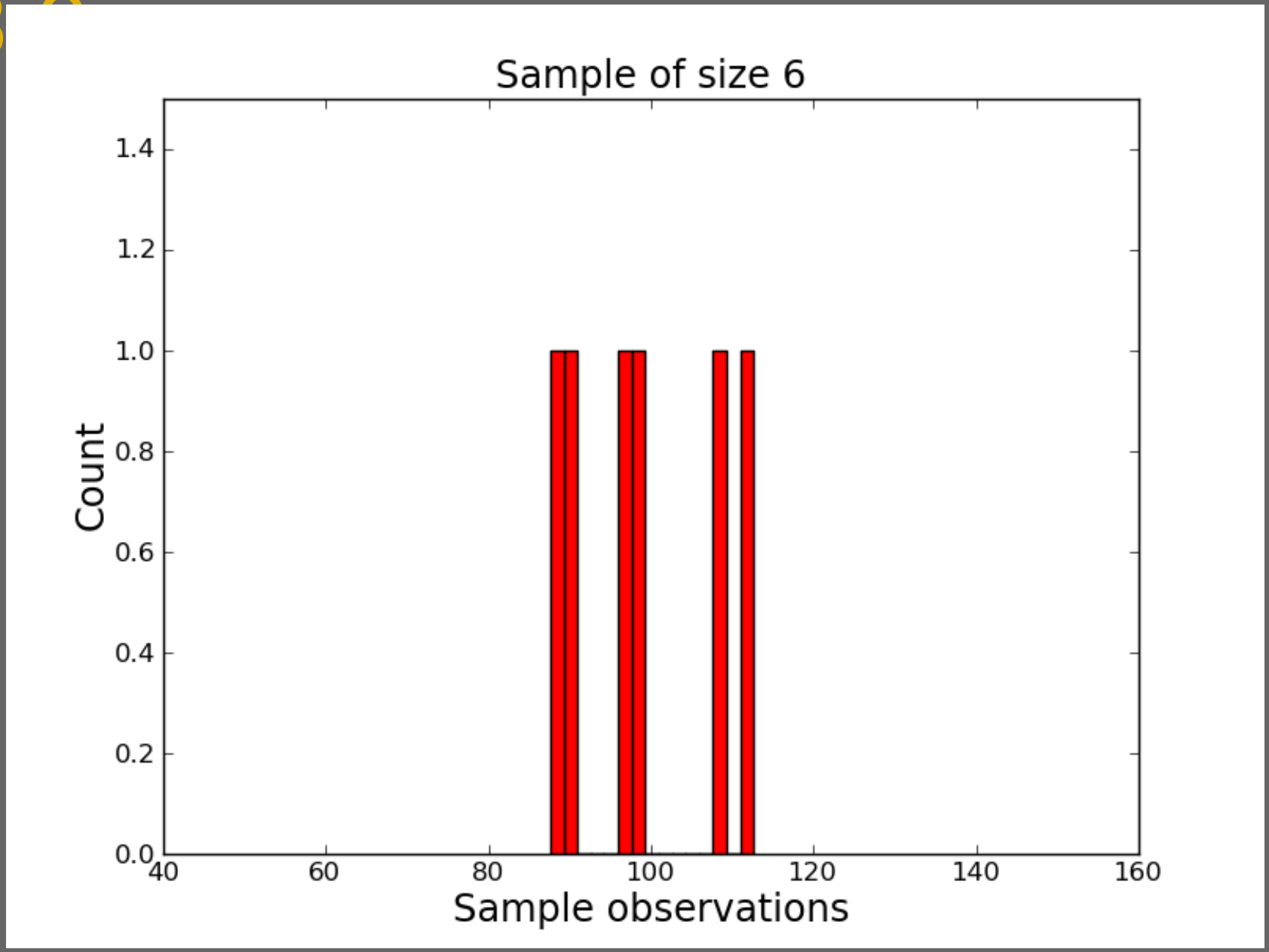
- Let's say we have some kind of IQ measure where the true distribution is normal, with a mean of 100 and a standard deviation of 10.


True distribution of an IQ measure
Probability
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
40
60
80
100
120
140
160
IQ score

Particular sample: size 2, mean = 97.6

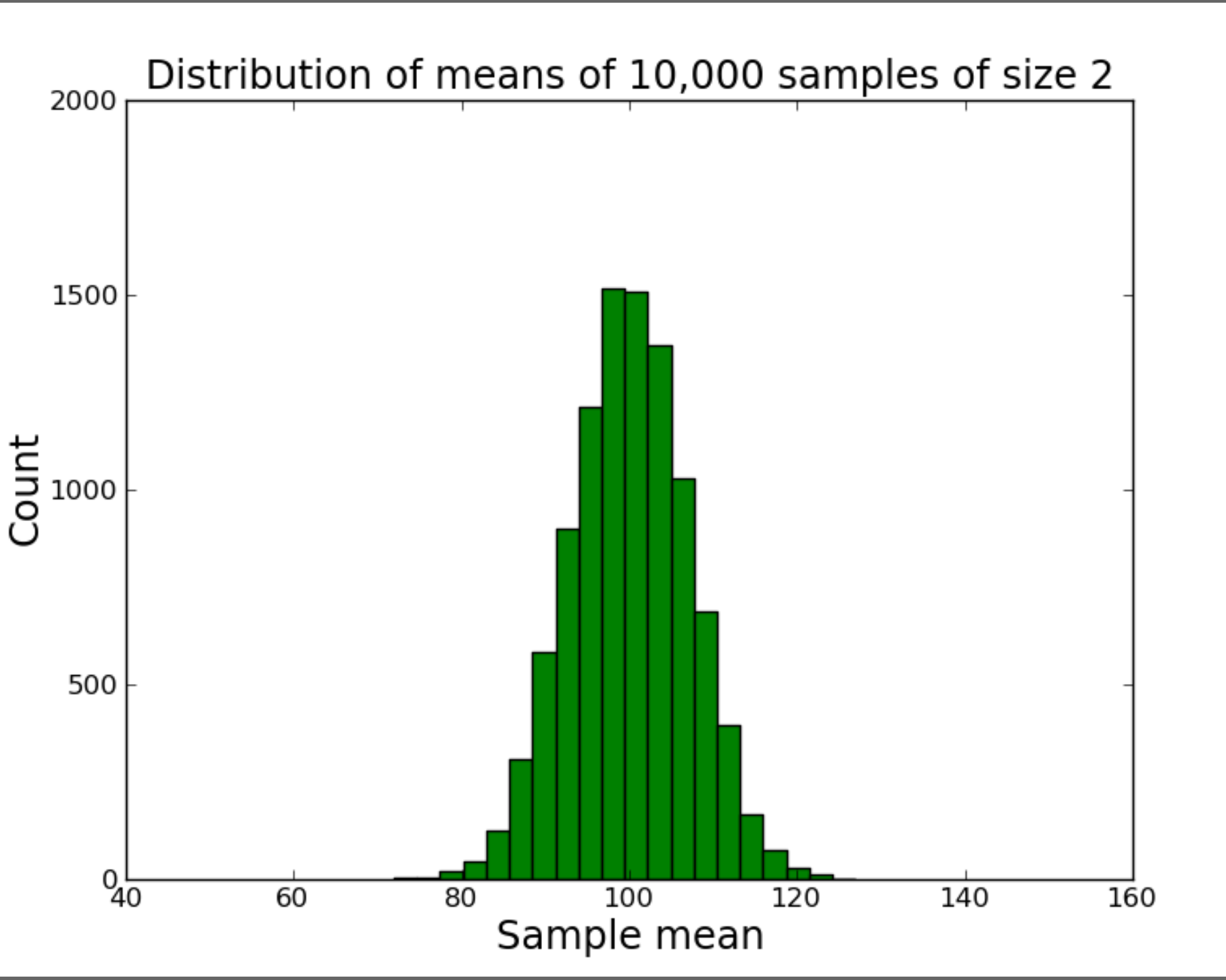


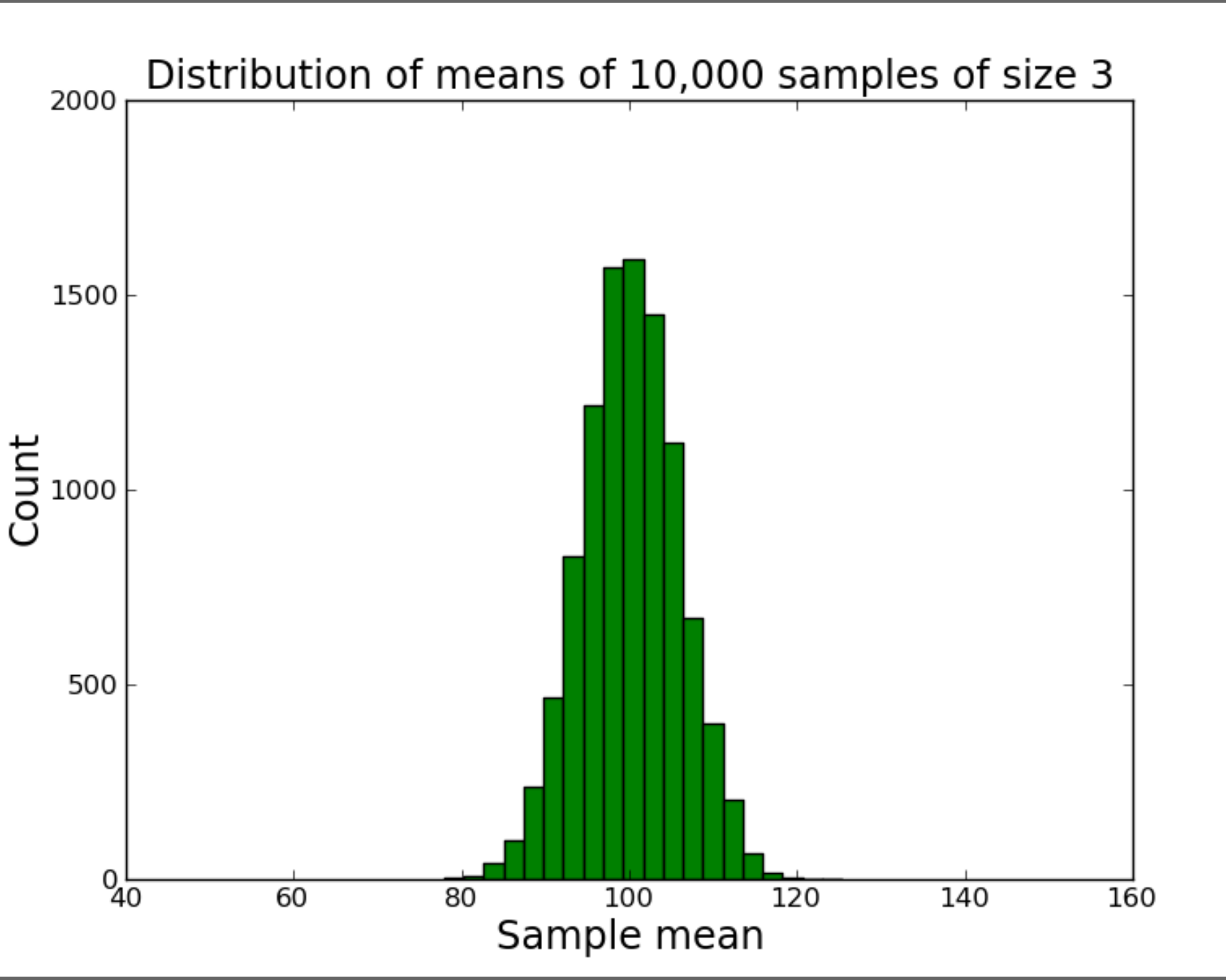
Particular sample: size 6, mean = 98.0

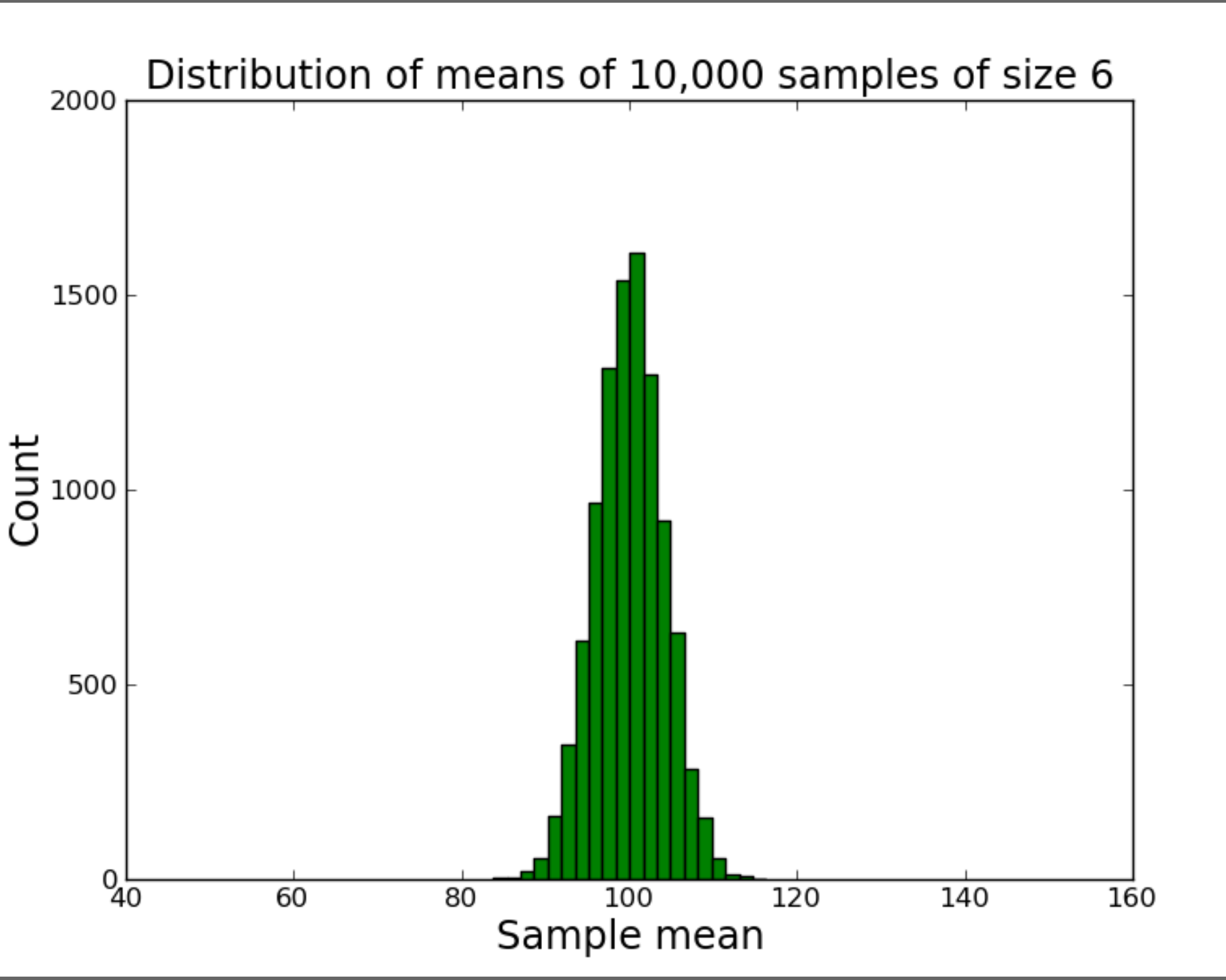


What is a sample?

- What are we doing when we take a sample of size N and calculate the mean?
- We know that there's a "meta-distribution" that describes the mean and standard deviations of the sample means. (Think of the green histograms.)
- The mean of this meta-distribution is the original population mean, and the standard deviation is the population standard deviation divided by the square root of the sample size (i.e., the standard error).
- So when we calculate the mean of one sample, we are *drawing a random variate from this meta-distribution*.


Distribution of means of 10,000 samples of size 2
Count
Sample mean
0
500
1000
1500
2000
40
60
80
100
120
140
160


Distribution of means of 10,000 samples of size 3
Count
Sample mean
0
500
1000
1500
2000
40
60
80
100
120
140
160


Distribution of means of 10,000 samples of size 6
Count
Sample mean
0
500
1000
1500
2000
40
60
80
100
120
140
160

Distribution of the sample means

- This distribution of the sample means "tightens up" as the sample size gets bigger. We've seen this before.
- We can characterize this meta-distribution in terms of its own mean and standard deviation, although of course we need to estimate those from our sample in real situations.
- Calculating the mean of a sample equates to drawing one variate from the meta-distribution.
- Therefore we can ask how often we're likely to see extreme values of the sample mean, i.e., values that lie in the tails of the meta-distribution.

How to calculate the probability of extreme sample means?

- If our sample size was big enough, we could use the normal distribution to make this calculation.
- We would ask how many standard deviations away from the overall mean our particular sample mean was: this is called a Z-score if the distribution is normal.
- In our case we're going to calculate basically the same thing and call it a t-score because we can't assume normality.

Calculating a t-score

- We want to know how many standard deviations away from the overall mean of the sampling distribution our one particular sample is.
- Let's flesh out the example: we take our IQ testing scheme to a new country, and we want to know whether the people here are any smarter or dumber than they were at home.
- This gives us our null and alternative hypotheses.
- The null hypothesis is that there's no difference in the IQ scores between the two countries: the mean is 100 in both cases.

Calculating a t-score

- Our alternative hypothesis is simply the converse, that there is *some* difference in IQ scores between the countries.
- Note that we usually don't have a commitment about whether the difference, if there is one, will be positive or negative.
- H_{null}: that $\mu = 100$. You also see "H_0:$\mu_0 = 100$".
- We collect a sample of 6 people, and give them IQ tests.
- They score: 101, 112, 100, 107, 94, 104.

Calculating a t-score

- The mean of the six scores is 103. This is higher than the null hypothesis suggests, but should we get excited?
- The standard deviation of the six scores is 5.66.
- But we don't want the plain SD, we want the *sample* standard deviation (division by N-1) because we're trying to estimate the population standard deviation.
- Remember we have to work with what we have. In this case, that's the tiny sample of 6 scores.

Calculating a t-score

- So the *sample* standard deviation is 6.20.
- The standard error is going to be 6.20 / sqrt(6), which is 2.53.
- We now have our best guess at the meta-distribution of the sample-of-size-six means: in the absence of any other information, we'd say that its mean is 103 and its standard deviation is 2.53.
- However, our null hypothesis is that our six numbers come from a distribution with a mean of 100, i.e., the same as back home.

Calculating a t-score

- We might have helped ourselves to the assumption that the standard deviation of IQ scores in this new country is 10, the same as at home. But we're not going to do that: who is to say that IQ doesn't have a different spread here?

- So our null hypothesis says: let's imagine that our sample mean comes from a distribution of sample means with mean of 100 and standard deviation of 2.53.

- This gives us our t-statistic:

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

Calculating a t-score

- So a t-score is a lot like a z-score: it's essentially measuring the number of standard deviations from an expected mean that some measurement is.
- In our case, the t-score is (103 - 100) / 2.53 = 1.18.
- That's not a great distance from the mean: recall the 1.96 threshold for z-scores that equates to the most extreme 5% of the distribution.
- Similarly, it turns out that a t-score of ±1.18, or a more extreme value, happens 28.9% of the time (this is our p-value). So we're not motivated to reject H_{null}.

Linking a t-score to a p-value

- In the old days you would look up a table of critical p-values for the t-distribution with an appropriate number of *degrees of freedom*.

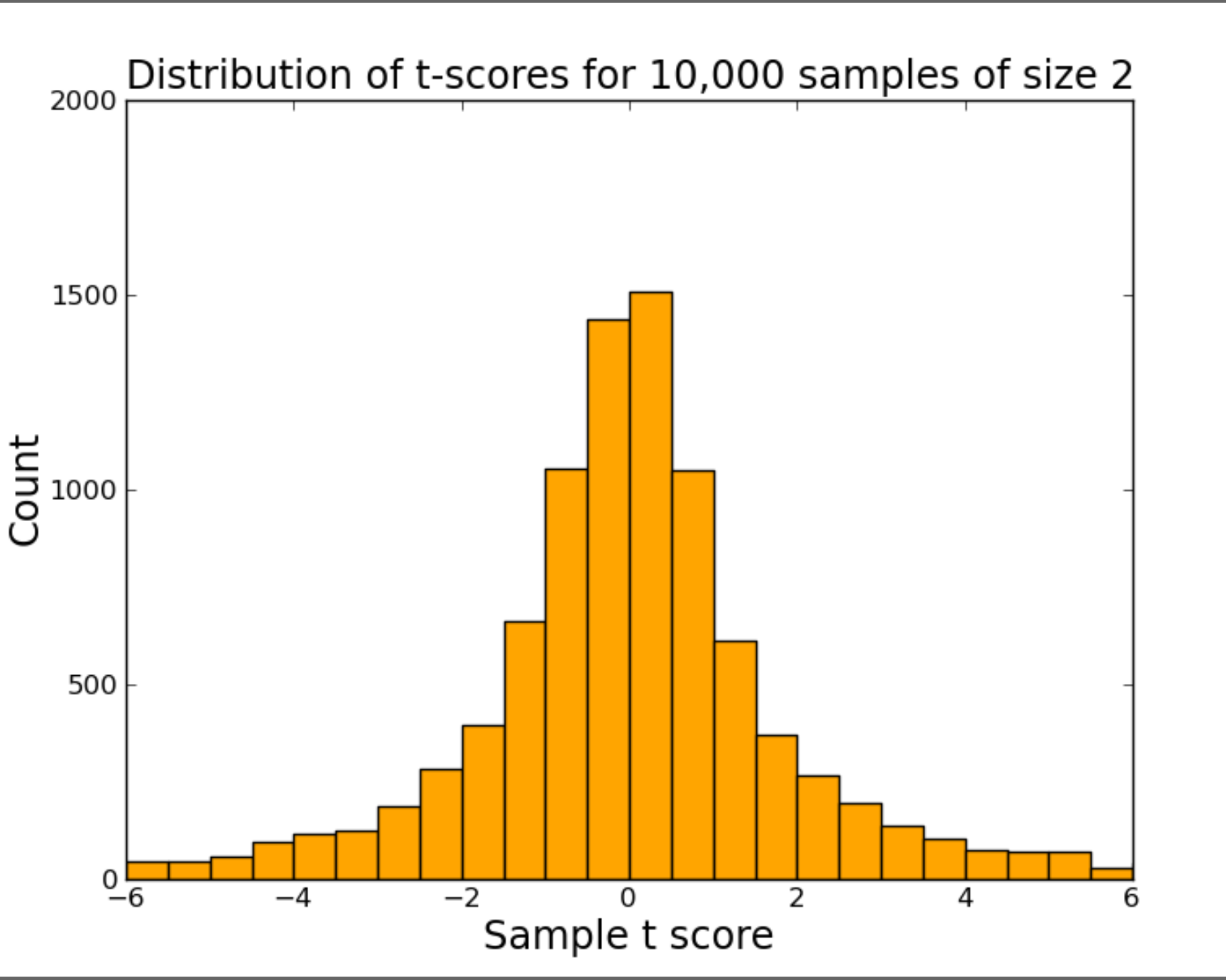
- Degrees of freedom come up a lot in statistics. It's just a measure of how many free parameters something has. For one-sample t-tests, the degrees of freedom are N-1, where N is the sample size. This is because to get a particular value of t, the last score in the sample is not free to vary.

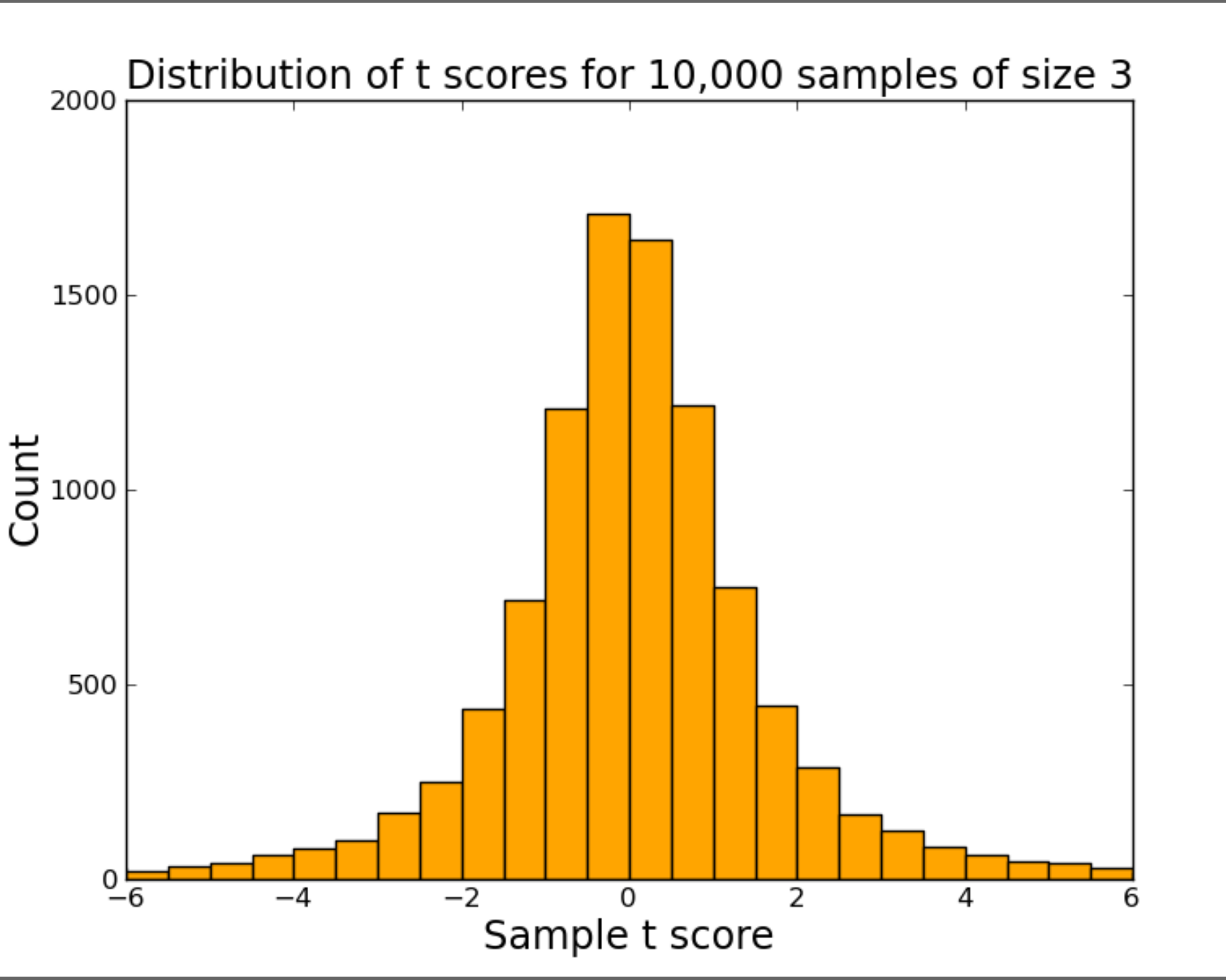
Linking a t-score to a p-value

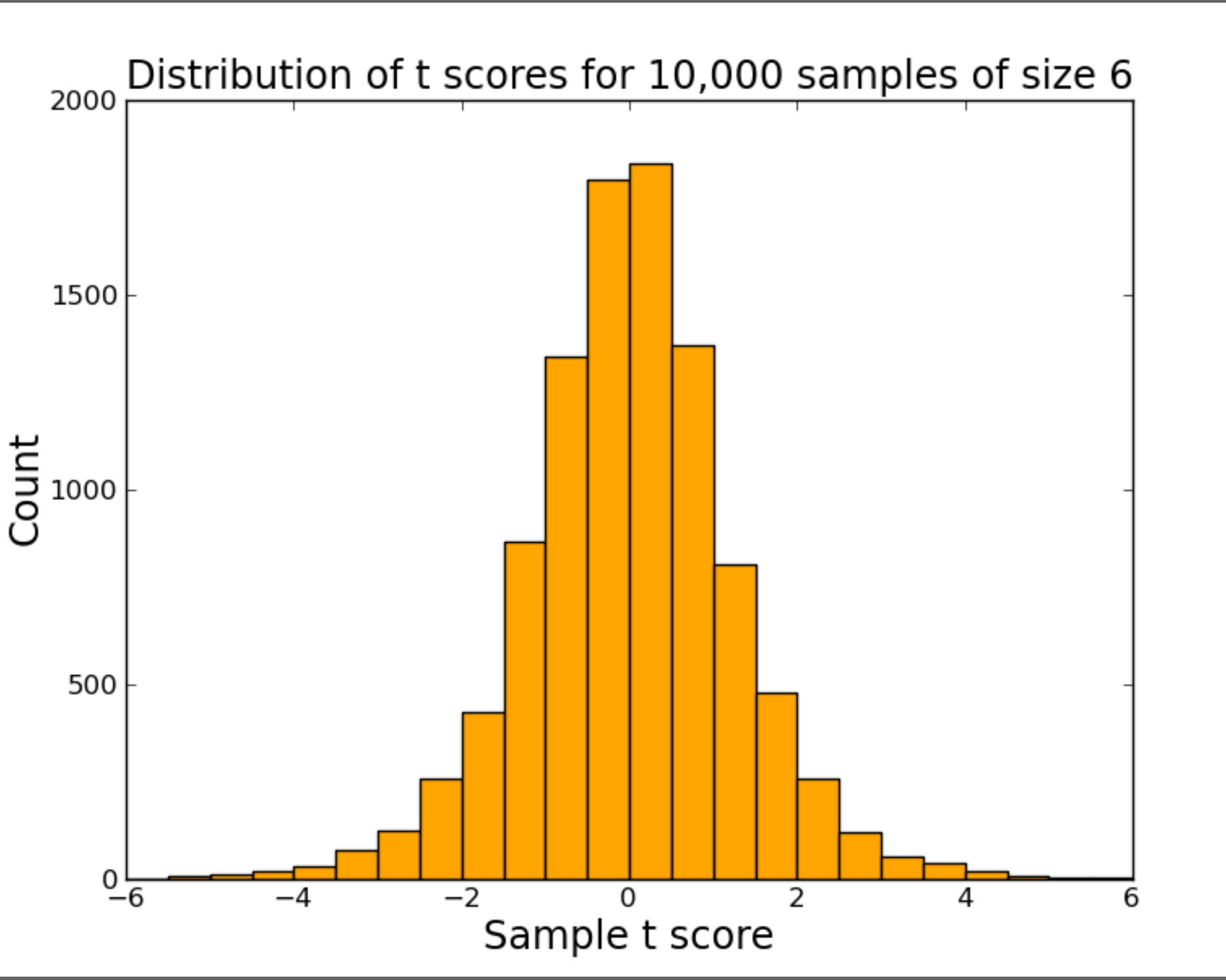
- We need to specify whether we're interested in a one-tailed or a two-tailed test.
- A two-tailed test is the default option. This means that we have no strong commitment on whether the sample mean is likely to be higher or lower than the mean specified in the null hypothesis. Thus we include both extreme tails of the distribution when figuring out our p-value.
- If for some reason we only cared about evidence for an alternative hypothesis that the mean score was (e.g.) higher, we could use a one-tailed test and look at only one side of the t-distribution in figuring out p.

Linking a t-score to a p-value

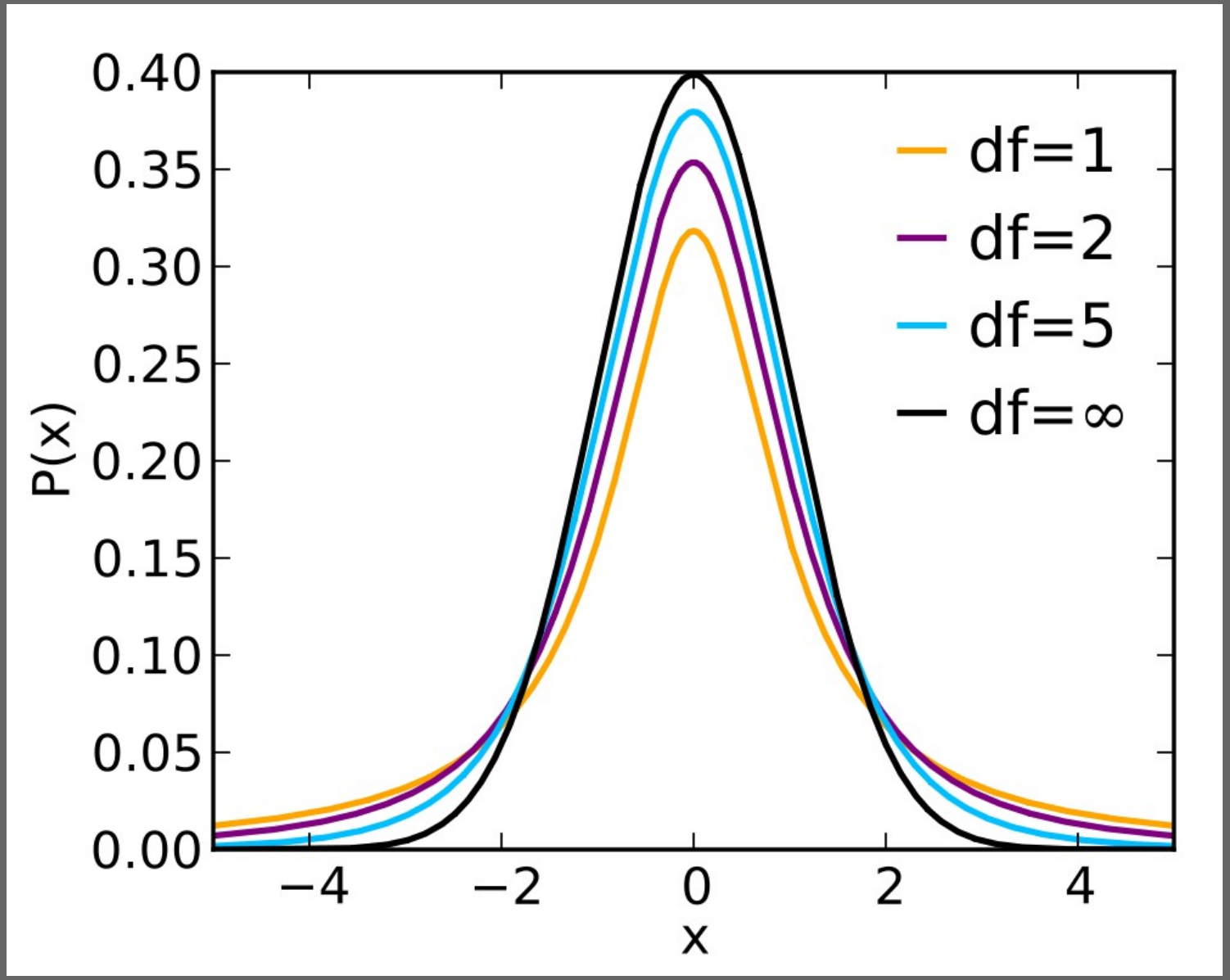
- In Python, use `from scipy import stats` and then `stats.ttest_1samp(IQscores, 100)`.
- In R, `t.test(IQscores, mu=100)`.
- Or we could do it "empirically" through simulating the sampling process...


Distribution of t-scores for 10,000 samples of size 2
Count
2000
1500
1000
500
0
−6
−4
−2
0
2
4
6
Sample t score


Distribution of t scores for 10,000 samples of size 3
Count
2000
1500
1000
500
0
−6
−4
−2
0
2
4
6
Sample t score


Distribution of t scores for 10,000 samples of size 6
Count
Sample t score
0
500
1000
1500
2000
−6
−4
−2
0
2
4
6

The t-distribution



The t-distribution summarized

- With modest sample sizes, we need to make a correction for the fact that our distribution of sample means is not actually normal.
- The t-distribution achieves this. It's really a family of distributions, one for each sample size.
- It has a lower peak and fatter tails than the normal distribution (especially so for really small sample sizes, such as 2) to capture the fact that small samples will produce extremely inaccurate estimates more often.

Some history and a gratuitous link to beer

- The t-test is also known as "Student's t-test". Why?

Some history and a gratuitous link to beer

- It was devised by William Sealy Gosset, a chemist working for the Guinness brewery in Dublin.
- Gosset devised the t-test as a way of cheaply monitoring the quality of batches of beer by taking small samples from those batches.
- Gosset published the test in 1908, but was forced to use a pseudonym ("Student") by Guiness, who regarded their use of statistics as a trade secret.

Other kinds of t-tests

- The one-sample t-test is what we've covered so far.
- The one-sample test can be used to deal with simple experimental designs in which we measure something before and after an intervention. For example, does drug X lower blood pressure, or does diet Y lead to weight loss?
- For each case in the study, we subtract the "before" score from the "after" score to get a difference.
- We can then examine the null hypothesis that the mean of the differences is zero, i.e., that the intervention makes no difference.

The two-sample t-test

- The two-sample t-test is an extension of the same idea.
- It is used to test the null hypothesis that two different samples in an experiment are drawn from the same population, i.e., that they have the same mean.
- For example, does drug A work any better or worse than drug B in reducing blood pressure? Do men and women systematically differ on their IQ scores?

The two-sample t-test

- There are some mathematical complications based on whether or not the sample sizes are the same and whether or not we can assume equal variances across the two samples.

- However: in practice you're unlikely to do many two-sample t-tests. You are more likely to use an analysis of variance (ANOVA) or a regression.

Type-I and type-II errors

	Statistical test finds something, H_{null} rejected	Statistical test is negative, can't reject H_{null}
There's an effect in the real world	A hit; you've found something	Missed a real effect, type II error
There is in fact no effect	False alarm, type I error	Correct to remain sceptical of H_{alt}

Type-I and type-II errors

- The key idea in statistics is to calculate p (data | H_{null}) and then reject H_{null} if this p-value is very low.
- But what counts as "very low"? What's the right threshold is for rejecting H_{null}?
- There's no right answer. Thresholds of 0.05 and 0.01 have been adopted in some quarters.
- It really depends on what the consequences of different kinds of errors might be.

Type-I and type-II errors

- By setting your p-value threshold for rejecting H_{null}, also known as an "alpha level", you can directly control your type-I error rate.
- How much does it bother you to believe that something is true when it isn't? (This is a type-I error.)
- If you're in the business of building aircraft navigation systems, you probably want to make sure you don't fall into this kind of error, and so you'll set you alpha level very low, perhaps 0.0001.

Type-I and type-II errors

- The difficulty is that by adopting a very conservative type-I error rate, you necessarily increase your type-II error rate.
- So you may now miss some things are actually true, e.g., you dismiss the idea of adopting a new part that could have slightly improved the performance of your system.
- If you are in the venture capital business, perhaps you're OK with making lots of type-I errors (backing companies that won't do well) but want to make sure you don't miss out on the chance to buy into the next Google. So you'd use a generous alpha level ($p < 0.1$) to minimize your type-II error rate.

Additional material

- A great video lecture on thinking critically about p-values (Geoff Cumming, LaTrobe University).
- An argument that simulation allows us to determine p-values empirically and that we shouldn't be obsessed with choosing the right statistical test (Allen Downey).
- The Python code for the graphs and simulations in this lecture.