Confidence intervals, t tests, P values

Joe Felsenstein

Department of Genome Sciences and Department of Biology

Normality

> Everybody believes in the normal approximation, the experimenters because they think it is a mathematical theorem, the mathematicians because they think it is an experimental fact! *G. Lippman*

- We can use the Gaussian (normal) distribution, assumed correct, and estimate the mean (which is the expectation). It turns out that, not surprisingly, the best estimate of the mean is the mean of the sample.

Normality

> Everybody believes in the normal approximation, the experimenters because they think it is a mathematical theorem, the mathematicians because they think it is an experimental fact! *G. Lippman*

- We can use the Gaussian (normal) distribution, assumed correct, and estimate the mean (which is the expectation). It turns out that, not surprisingly, the best estimate of the mean is the mean of the sample.
- (The median of the sample is a legitimate estimate too, but it is noisier).

Normality

> Everybody believes in the normal approximation, the experimenters because they think it is a mathematical theorem, the mathematicians because they think it is an experimental fact! *G. Lippman*

- We can use the Gaussian (normal) distribution, assumed correct, and estimate the mean (which is the expectation). It turns out that, not surprisingly, the best estimate of the mean is the mean of the sample.
- (The median of the sample is a legitimate estimate too, but it is noisier).
- The sample mean is the optimal estimate as it is the Maximum Likelihood Estimate – for which see later in the course.

Normality

> Everybody believes in the normal approximation, the experimenters because they think it is a mathematical theorem, the mathematicians because they think it is an experimental fact! *G. Lippman*

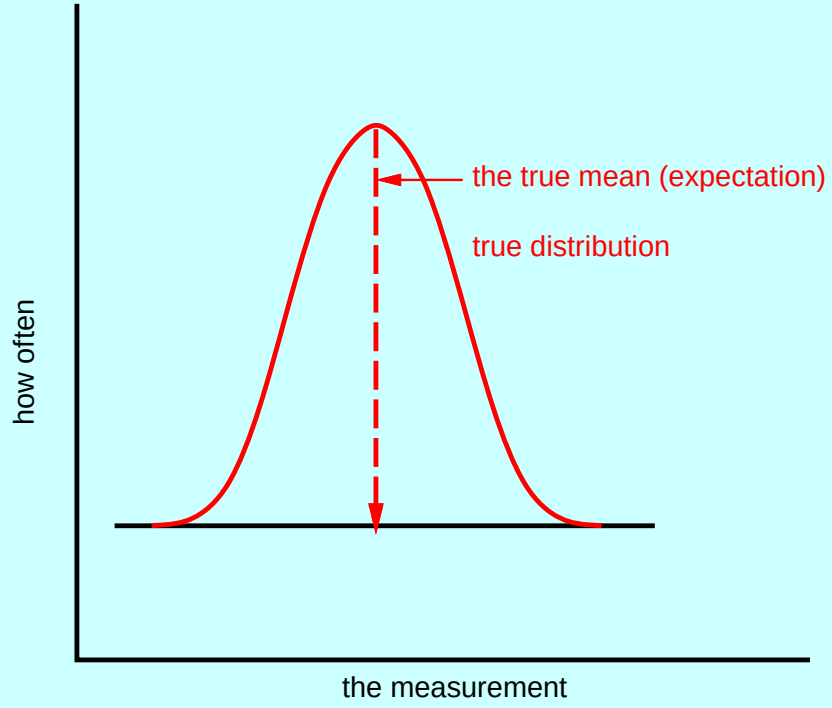
- We can use the Gaussian (normal) distribution, assumed correct, and estimate the mean (which is the expectation). It turns out that, not surprisingly, the best estimate of the mean is the mean of the sample.
- (The median of the sample is a legitimate estimate too, but it is noisier).
- The sample mean is the optimal estimate as it is the Maximum Likelihood Estimate – for which see later in the course.
- But how do we figure out how noisy the estimate is?

Normality

> Everybody believes in the normal approximation, the experimenters because they think it is a mathematical theorem, the mathematicians because they think it is an experimental fact! *G. Lippman*

- We can use the Gaussian (normal) distribution, assumed correct, and estimate the mean (which is the expectation). It turns out that, not surprisingly, the best estimate of the mean is the mean of the sample.
- (The median of the sample is a legitimate estimate too, but it is noisier).
- The sample mean is the optimal estimate as it is the Maximum Likelihood Estimate – for which see later in the course.
- But how do we figure out how noisy the estimate is?
- Can we make an *interval estimate*?

A normal distribution (artist's conception)



Uncertainty of the mean

- Let's go forward (distribution to data).

Uncertainty of the mean

- Let's go forward (distribution to data).
- if the (unknown) standard deviation of the true distribution is σ, the variance is σ^2.

Uncertainty of the mean

- Let's go forward (distribution to data).
- if the (unknown) standard deviation of the true distribution is σ, the variance is σ^2.
- The variance of the mean of a sample of n points is σ^2/n,

Uncertainty of the mean

- Let's go forward (distribution to data).
- if the (unknown) standard deviation of the true distribution is σ, the variance is σ^2.
- The variance of the mean of a sample of n points is σ^2/n,
- so its standard deviation is $\sigma/\sqrt{n}$.

Uncertainty of the mean

- Let's go forward (distribution to data).
- if the (unknown) standard deviation of the true distribution is σ, the variance is σ^2.
- The variance of the mean of a sample of n points is σ^2/n,
- so its standard deviation is $\sigma/\sqrt{n}$.
- The 2.5% point of a normal distribution is 1.95996 standard deviations below the mean.

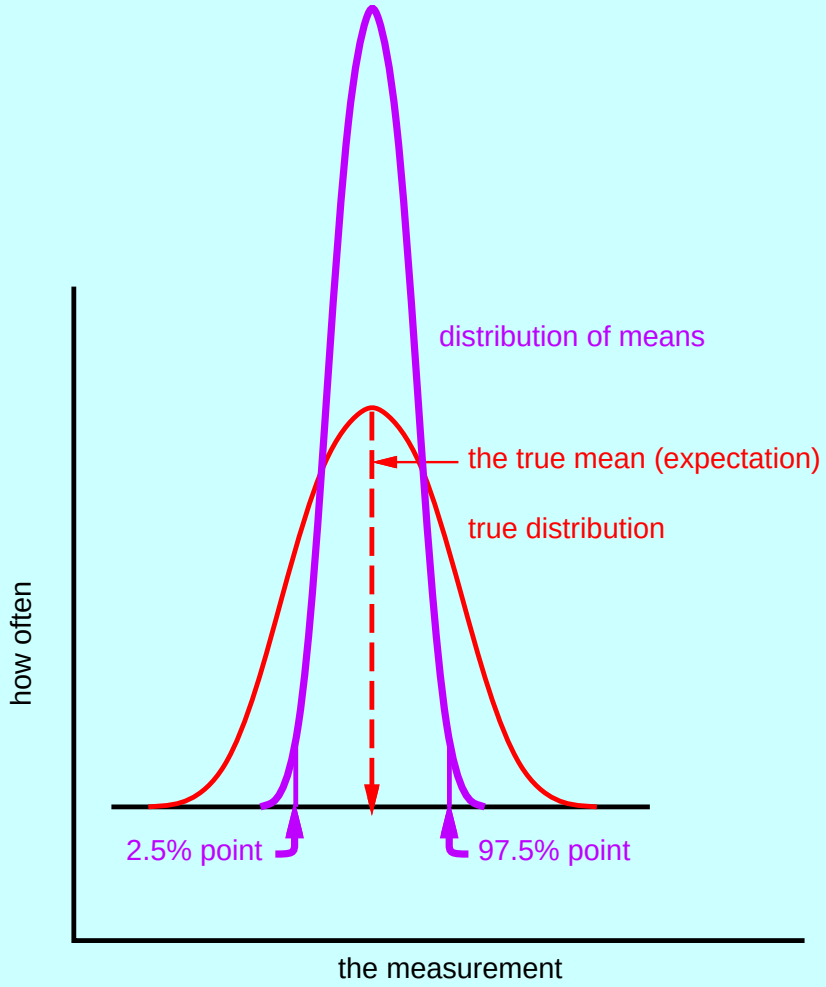
Uncertainty of the mean

- Let's go forward (distribution to data).
- if the (unknown) standard deviation of the true distribution is σ, the variance is σ^2.
- The variance of the mean of a sample of n points is σ^2/n,
- so its standard deviation is $\sigma/\sqrt{n}$.
- The 2.5% point of a normal distribution is 1.95996 standard deviations below the mean.
- The 97.5% point of a normal distribution is 1.95996 standard deviations above the mean.

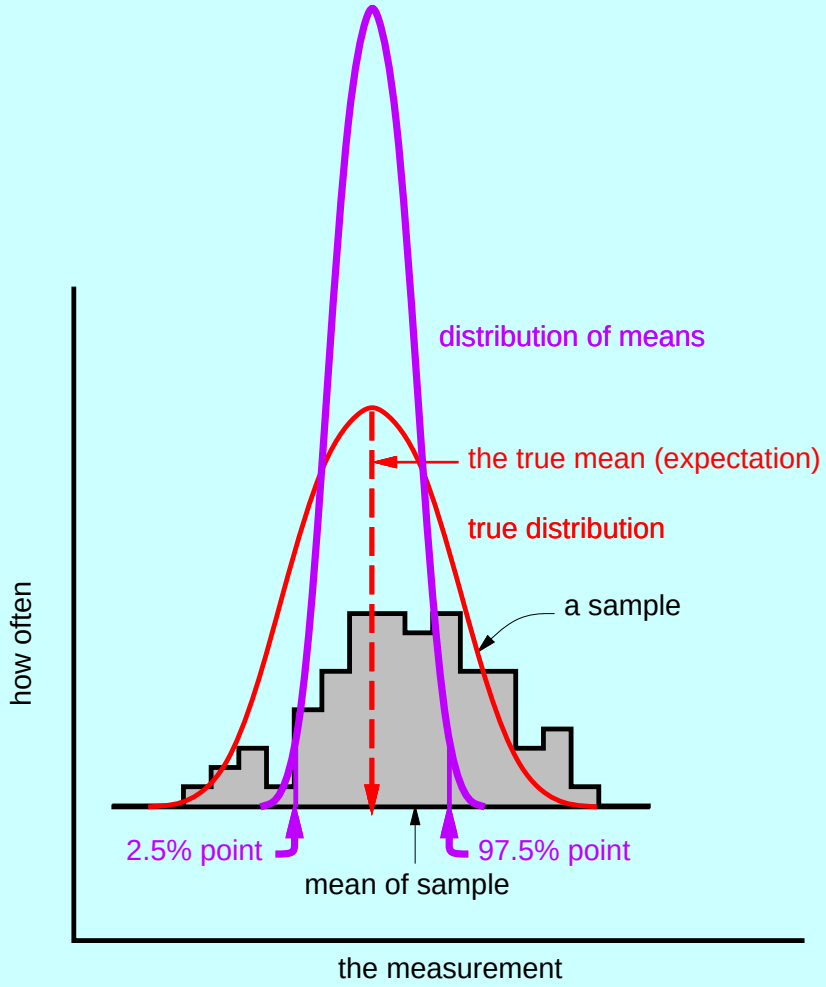
Uncertainty of the mean

- Let's go forward (distribution to data).
- if the (unknown) standard deviation of the true distribution is σ, the variance is σ^2.
- The variance of the mean of a sample of n points is σ^2/n,
- so its standard deviation is $\sigma/\sqrt{n}$.
- The 2.5% point of a normal distribution is 1.95996 standard deviations below the mean.
- The 97.5% point of a normal distribution is 1.95996 standard deviations above the mean.
- So if the (unknown) true mean is called μ, 95% of the time the mean you calculate from a sample, will lie between $\mu - 1.95996\,\sigma/\sqrt{n}$ and $\mu + 1.95996\,\sigma/\sqrt{n}$.

The distribution of means of n points



A particular sample



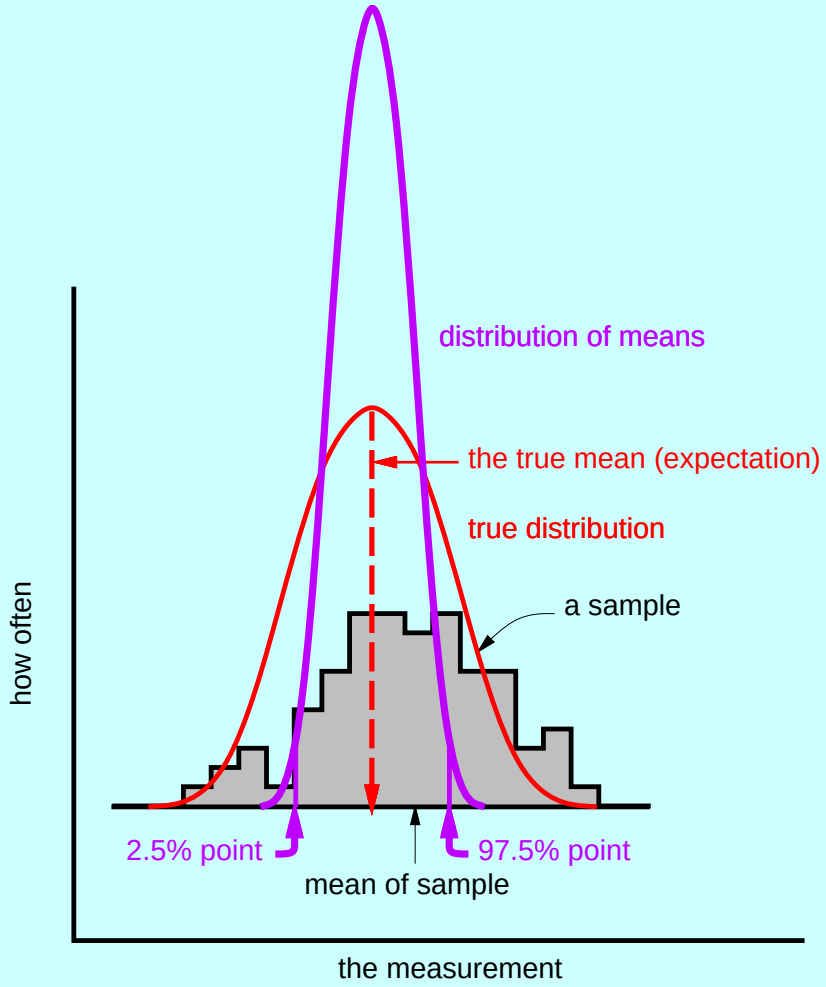
Confidence interval

So, that solves it, right? No! We don't know μ (which is what we want to know). We have calculated how the limits on the sample mean, not the limits on the true mean.

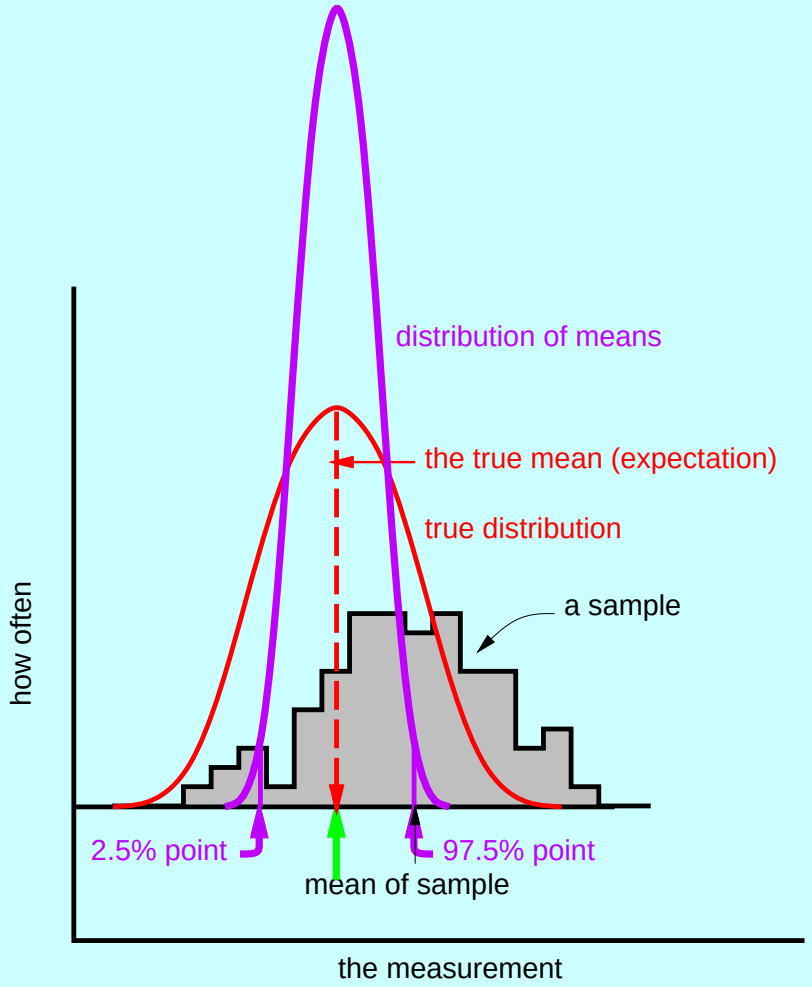
But we can say that, 95% of the time, the empirical mean $\bar{x}$ that we calculate is below that upper limit, and above that lower limit.

In that sense (*in what sense?*) the true mean is ...

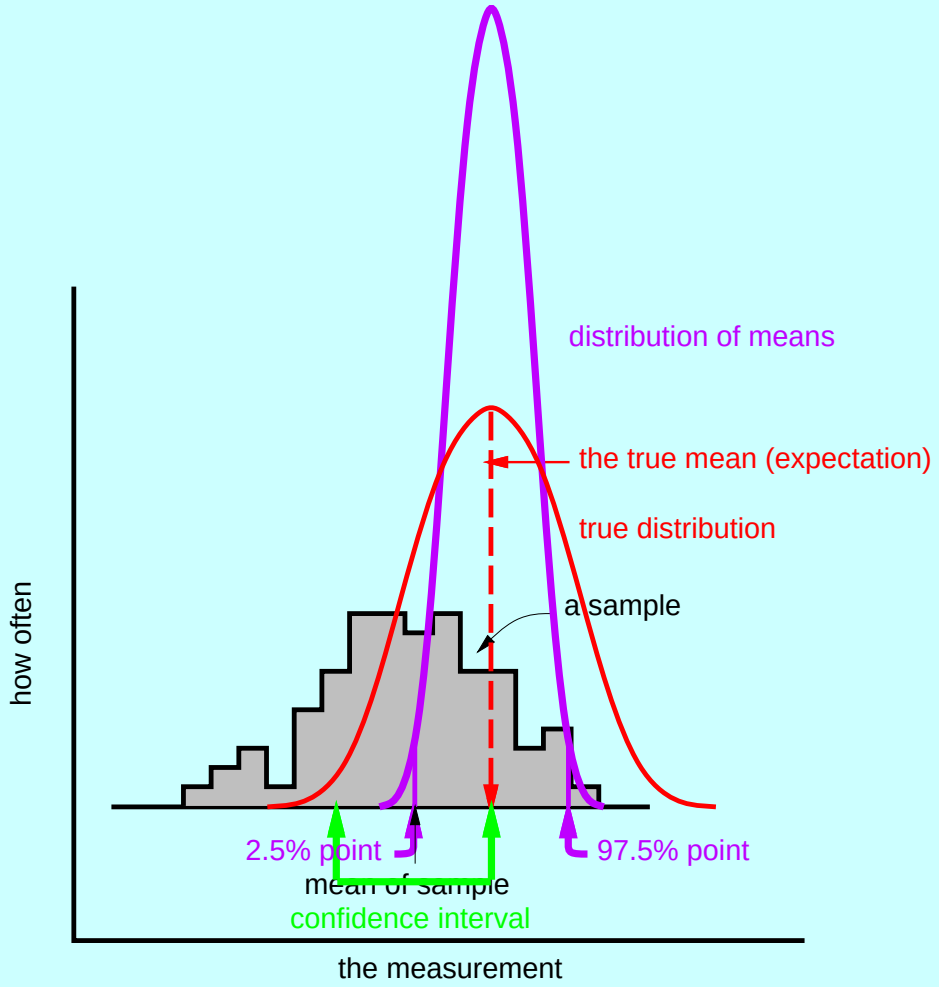
Let’s get ready to slide the true stuff left



Not any lower than this ...

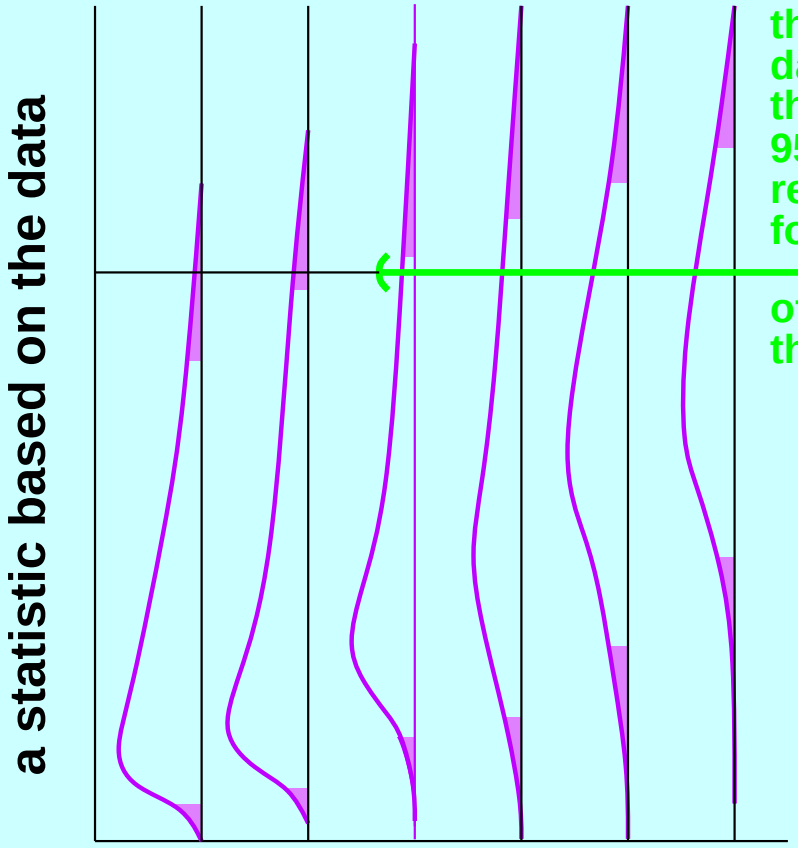


Not any higher than this ...



Here we see true values, and data statistics

for each parameter value, find data values (unshaded) that account for 95% of the probability



then, given some data, the parameters that are in the 95% confidence region are those for which that value of the statistic is in the unshaded region

so 95% of the time the statistic is in the region where the confidence interval based on it contains the truth.

In what sense??

You could say that if you do this throughout your career, 95% of these intervals will contain the true value.

- This is fairly unsatisfactory, as it acts as if the outcome of some other, unrelated, experiment is relevant.

In what sense??

You could say that if you do this throughout your career, 95% of these intervals will contain the true value.

- This is fairly unsatisfactory, as it acts as if the outcome of some other, unrelated, experiment is relevant.
- Maybe best this way: if you did this experiment again and again and again, each time making a confidence interval for the mean, 95% of those intervals would contain the true value.

The Bayesian alternative

We will cover this later (lecture 6). It involves assuming that we know a prior probability of the parameters (in this case of the mean μ of the true distribution and the standard deviation σ).

Then, using Bayes' Formula (which we see in that lecture) we can compute the posterior probability of the parameter, given the data.

This gives us what we really wanted: the probabilities (or probability densities) of different values of the parameters. But it is achieved at the cost of having to have a prior distribution for them that we are assuming is true.

The present approach instead makes a different assumption, that we can regard the experiment as a repeatable trial. This Frequentist approach and the Bayesian approach both have supporters among statisticians – the issue is one's philosophy of science, not a technical disagreement about statistics.

Coping with not knowing the standard deviation

So far we have casually assumed we know the standard deviation σ of the true distribution. But generally we don't. The upper (lower) 2.5% point of the sample means is 1.95996 σ. But if we look at the estimate of σ^2, its unbiased estimate is (for reasons connected with also estimating the mean)

$$\hat{s}^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

(the $n-1$ instead of n is to correct for x_i being a bit closer to $\bar{x}$ than it should be, because x_i is part of $\bar{x}$ as well).

The standard deviation we use is the (estimated) standard deviation of the mean, which is $\hat{s}/\sqrt{n}$.

The quantity 1.95996 needs to be replaced by something that corrects for us sometimes having a $\hat{s}$ that is smaller than σ, and sometimes larger.

Here's why, in this case "Guiness is good for you":

Student's t distribution

W. S. Gosset (1876-1937) was a modest, well-liked Englishman who was a brewer and agricultural statistician for the famous Guinness brewing company in Dublin. It insisted that its employees keep their work secret, so he published under the pseudonym 'Student' the distribution in 1908. This was one of the first results in modern small-sample statistics.

Why we have to make the confidence interval wider

We can't just say that the estimated $\hat{\sigma}$ is sometimes 20% smaller than the true value, and sometimes 20% larger, so it all cancels out.

It doesn't, because σ doesn't have a symmetrical distribution, but something derived from a chi-square distribution (which we have not introduced yet).

The t statistic

This is the number of (estimated) standard deviations of the mean that the mean deviates from its expected value.

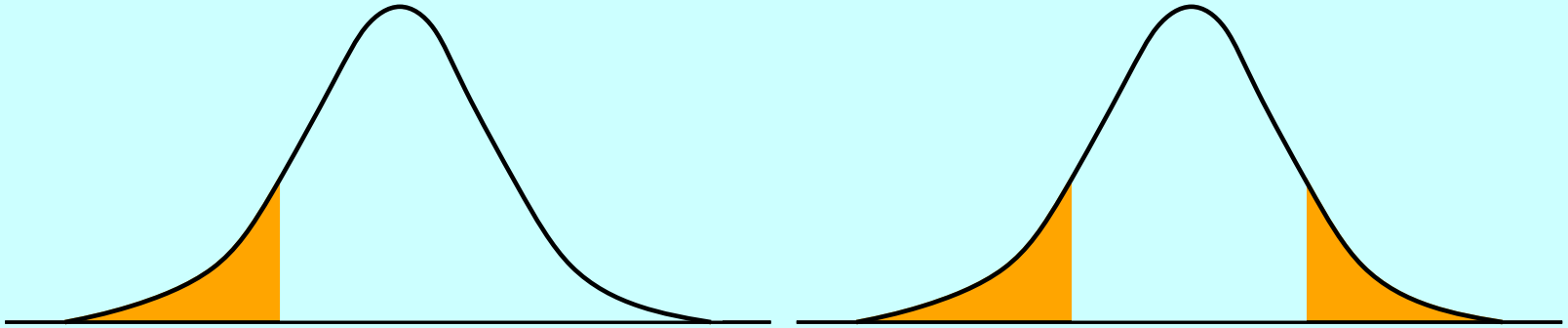
If $\hat{s}$ is the estimated standard deviation, from a sample of n values, and the mean of the sample is $\bar{x}$, while the expectation of that mean is μ, then

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

This does not have a normal distribution but it is closer to normal the bigger n is. The quantity $n - 1$ is called the *degrees of freedom* of the t value.

The P value is the (one- or two-tailed) tail probability

The P value for a t-test is the probability of the value of a t variable falling farther from the mean than the value of t that we observed.



This can be one-tailed or two-tailed (these differ by a factor of 2 since the t distribution is symmetrical) depending on whether we are interested in departures in both directions.

Values of t for 0.05 tail probability

Degrees of freedom $(n-1)$	t	Degrees of freedom $(n-1)$	t
1	12.71	9	2.262
2	4.303	10	2.228
3	3.182	15	2.131
4	2.776	20	2.086
5	2.571	30	2.042
6	2.447	40	2.021
7	2.365	50	2.009
8	2.306		

Note that, as the number of degrees of freedom $(n-1)$ rises, the value of t falls toward 1.95996 which would be the multiple to use for the true σ. This is because our estimated standard deviation $\hat{s}$ is getting close to the true value.

Elaborate tables exist to compute values of t for other values of P. Or of course you can just let R calculate it for you.

Why 0.05?

No real reason.

It is mostly an historical accident of the values R. A. Fisher chose to use in his incredibly influential 1922 book *Statistical Methods for Research Workers*.

The more cautious you need to be the smaller tail probability P you should choose. (Note: sometimes people describe the non-tail probability, such as 95%, instead).

Statistical tests and confidence intervals

If someone suggests that the true μ is (say) 3.14159, we can test this by seeing whether 3.14159 is within the confidence interval. If it isn't, we reject that hypothesis (the *null hypothesis*).

Alternatively, we can calculate for what value of P the suggested value (3.14159 in this case) is just within the confidence interval. We might get, for example, 0.09. That means it is within the interval calculated for a more extreme tail probability 0.05, Departure of the mean from 3.14159 is *not significant*. We can report the 0.09 to indicate to the reader how close to significance it was.

Getting all worried about whether 0.09 is "significantly larger than 0.05" is not worth it. If we think about the "significance of the significance", that way lies madness!

The 0.05 is the "type I error" of the test. It is the fraction of the time (when the null hypothesis is true) that we get a "false positive" and reject it.

One-tailed test

If we just want to know whether our mean is significantly bigger than 3.14159, but do not want to get excited if it is smaller, we do a one-tailed t test. We get the tail probability of 0.05 on one side, and 0.05 on the other (before we used 0.025 on each). Then we ignore the uninteresting tail (in this case the upper tail – think about it ...)

This will make sense, for example, if we want to get excited about a drug and develop it further if it causes us to lose weight, but not if it causes us to gain weight.

Paired two-sample t-test

Suppose we have two samples of the same size (say 25 numbers) and the corresponding numbers pair naturally. That is, there are some sources of error ("noise") that are shared by both members of a pair. For example, we might do before-and-after pairs of measurements after giving a drug. Or measure gene expression levels of a gene on two samples (one European and one African) on the same day, and do pairs of indviduals, one from each continent, on 25 different days.

The difference between two independent normally-distributed quantities is itself normally distributed. So to measure whether the "after" member of the pair is different from the "before" member, we just subtract them, and test whether the mean of this sample of differences is 0. It is then just a one-sample t test of the sort we used above. The standard deviation you use is the standard deviation of the differences, not of the individual members of the pair.

If the shared noise is large, this can be incredibly more appropriate than the next test ...

(Unpaired) two-sample t test

Suppose the two samples, which might even have different numbers of points, are drawn independently – there is no connection between point 18 from one sample, and point 18 from another. We want to compare the means of the two samples.

Assuming both are drawn from normal distributions *with the same (unknown) true variance* we can do a t test this way:

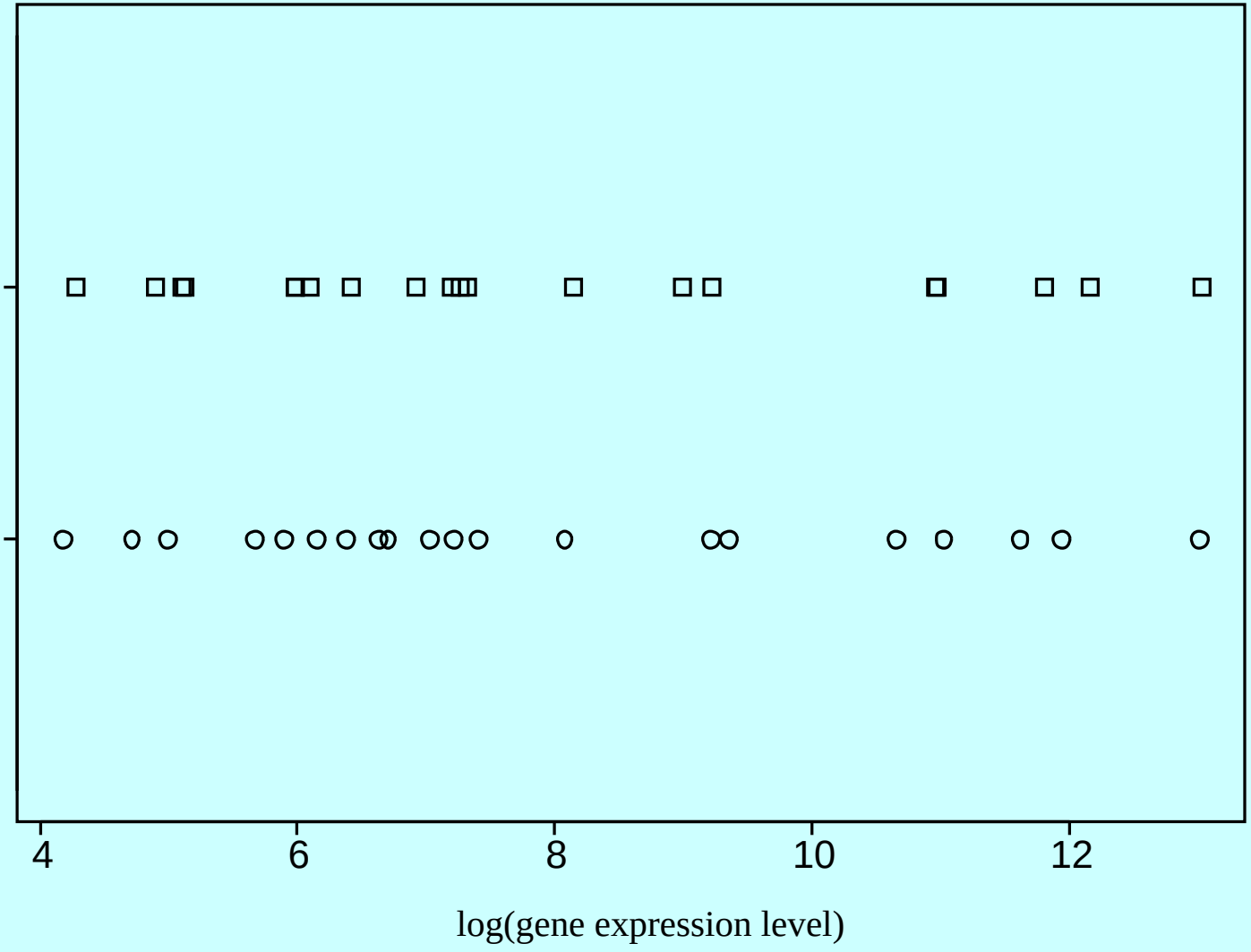
- Suppose the samples are $x_1, x_2, \ldots x_m$ and $y_1, y_2, \ldots, y_n$, compute the means of the two samples ($\bar{x}$ and $\bar{y}$).
- Compute a pooled estimate of the variance:

$$\hat{s}^2 = \frac{\sum_{i=1}^{m} (x_i - \bar{x})^2 + \sum_{j=1}^{n} (y_i - \bar{y})^2}{n + m - 2}$$

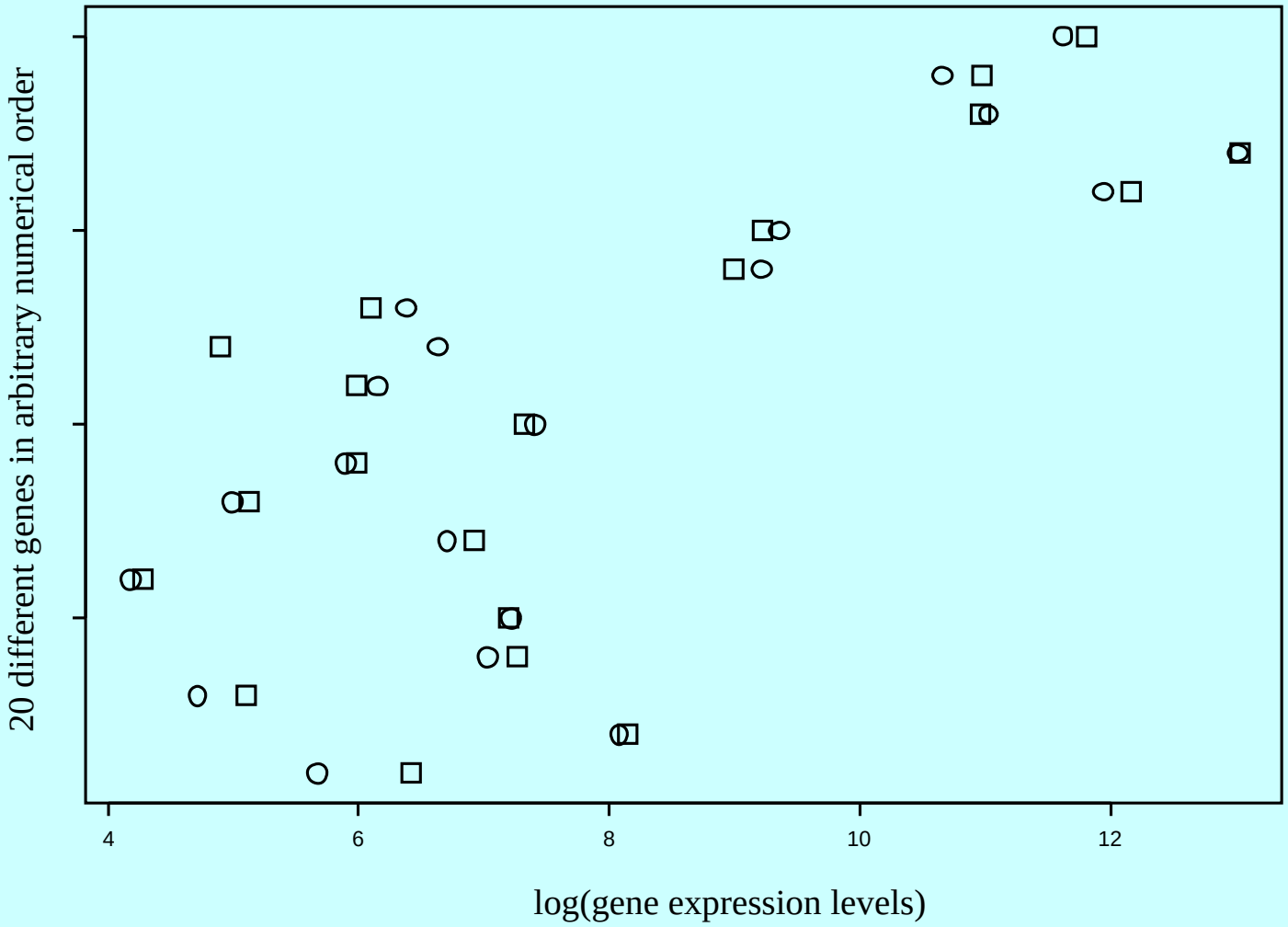
- The degrees of freedom is that denominator: $(n-1) + (m-1)$ (or $n + m - 2$).
- Compute the confidence interval on $\bar{x} - \bar{y}$ (using that estimate of the variance, and its square root times $\sqrt{1/m + 1/n}$ as the standard deviation). That is appropriate under the null hypothesis that the difference is 0. See whether 0 is in it. You can do either a two-tailed or a one-tailed test.

Gene expression levels in two individuals

Again from the Storer and Akey results, here are the gene expression levels for 20 genes, shown as unpaired. One individual is squares, the other is circles. (Storey et al., 2007, *Amer. J. Human Genet.*)



The same gene expression levels – paired



Using R to do a one-sample t test

```
> t.test(x[,1])

        One Sample t-test

data:  x[, 1]
t = 13.5484, df = 19, p-value = 3.251e-11
alternative hypothesis:  true mean is not equal to 0
95 percent confidence interval:
 6.674685 9.113771
sample estimates:
mean of x
 7.894228
```

A two-sample t-test, unpaired

```
> t.test(x[,1],x[,2])

        Welch Two Sample t-test

data:  x[, 1] and x[, 2]
t = -8e-04, df = 37.984, p-value = 0.9994
alternative hypothesis:  true difference in means is
not equal to 0
95 percent confidence interval:
 -1.686258 1.684984
sample estimates:
mean of x mean of y
 7.894228 7.894865
```

A two-sample t-test, which pairs corresponding values

```
> t.test(x[,1],x[,2],paired=TRUE)

        Paired t-test

data:  x[, 1] and x[, 2]
t = -0.006, df = 19, p-value = 0.9953
alternative hypothesis:  true difference in means is
not equal to 0
95 percent confidence interval:
 -0.2223615 0.2210875
sample estimates:
mean of the differences
                -0.000637
```

From ScienceNews, last year

Statistical insignificance

Nowhere are the problems with statistics more blatant than in studies of genetic influences on disease. In 2007, for instance, researchers combing the medical literature found numerous studies linking a total of 85 genetic variants in 70 different genes to acute coronary syndrome, a cluster of heart problems. When the researchers compared genetic tests of 811 patients that had the syndrome with a group of 650 (matched for sex and age) that didn't, only one of the suspect gene variants turned up substantially more often in those with the syndrome — a number to be expected by chance.

"Our null results provide no support for the hypothesis that any of the 85 genetic variants tested is a susceptibility factor" for the syndrome, the researchers reported in the *Journal of the American Medical Association*.

How could so many studies be wrong? Because their conclusions relied on "statistical significance," a concept at the heart of the mathematical analysis of modern scientific experiments.

Tom Siegfried. 2010. Odds are, it's wrong. *ScienceNews*, March 27.

Miscellaneous

If the two samples seem to have different variances, maybe you are on the wrong scale. Maybe their logarithms are normally distributed and have equal variances (more nearly anyway). If we are dealing with measurements that can't be negative, such as weights, that is a more natural assumption, and sometimes it homogenizes the variances nicely.

Get familiar with the lognormal distribution, which just means the logs are normally distributed. (It doesn't matter whether natural logs or common logs as those are just multiples of each other).