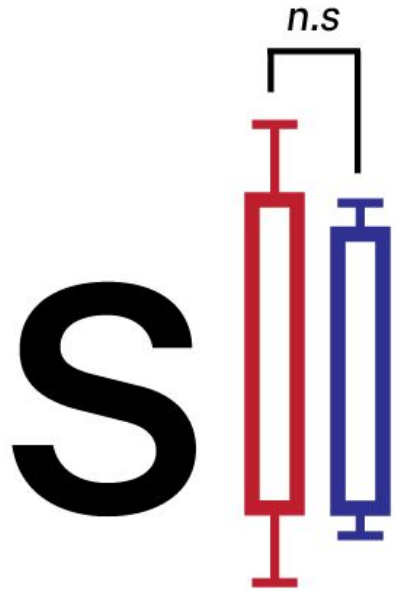


sllobs

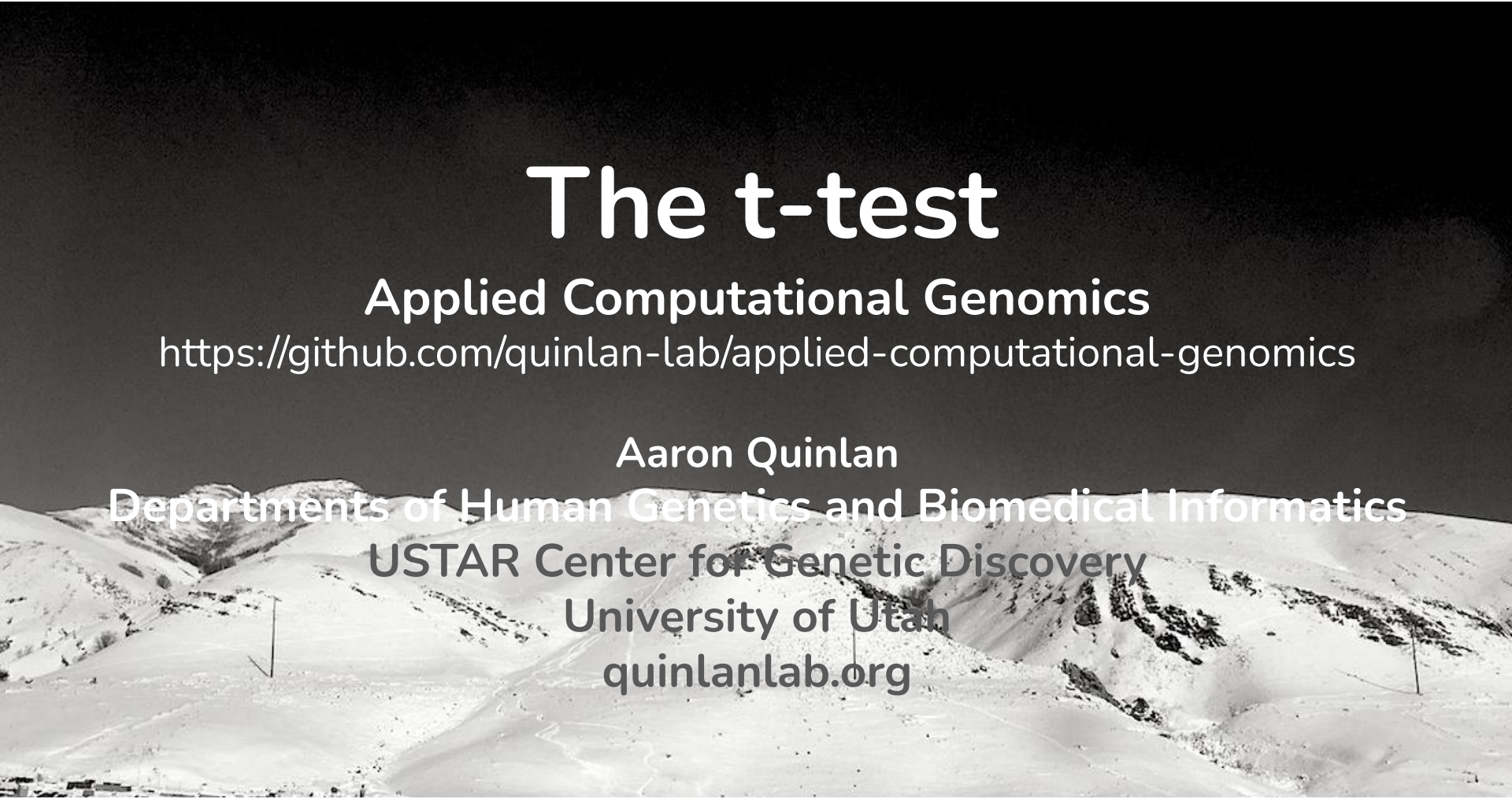
> salt lake learners of biostatistics

Meeting #14. The t-statistic, t-distribution, t-tests, and p-values

Aaron Quinlan

November 4, 2019

bit.ly//sllobs



The t-test

Applied Computational Genomics

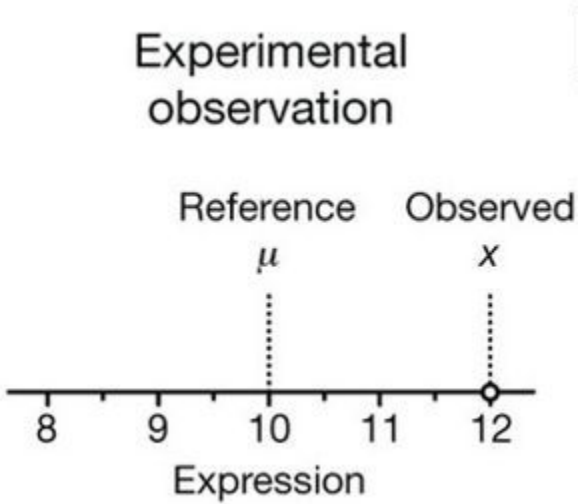
https://github.com/quinlan-lab/applied-computational-genomics

Aaron Quinlan
Departments of Human Genetics and Biomedical Informatics
USTAR Center for Genetic Discovery
University of Utah
quinlanlab.org

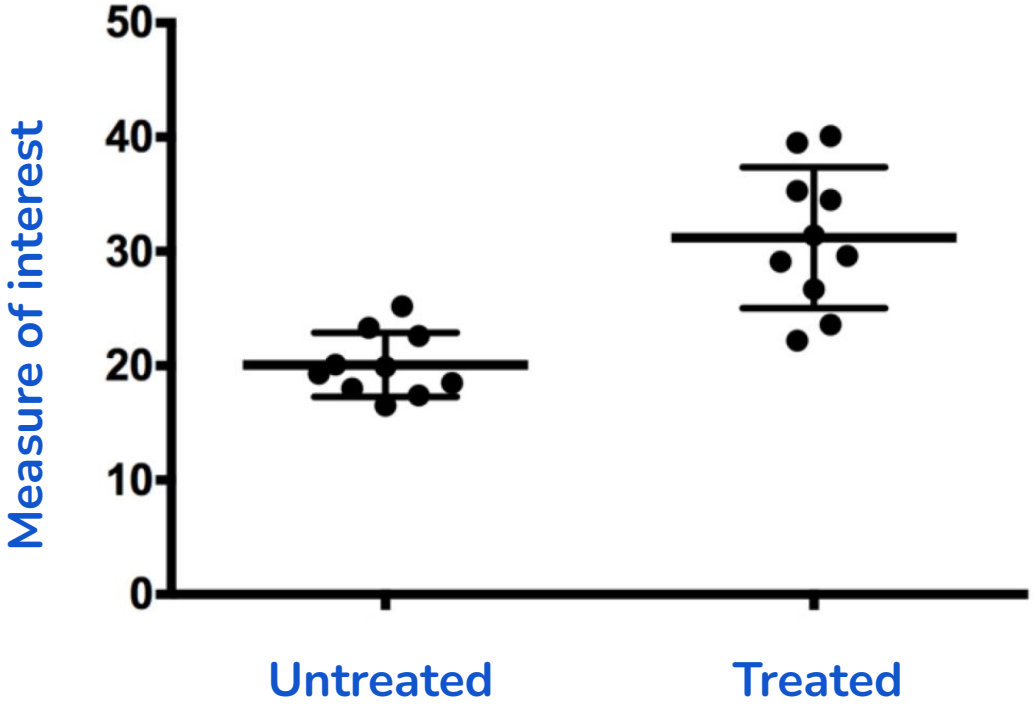
In biology, we often use samples to estimate the behavior of the full population (of molecules, cells, individuals). Sampling introduces uncertainty.

Comparing observations

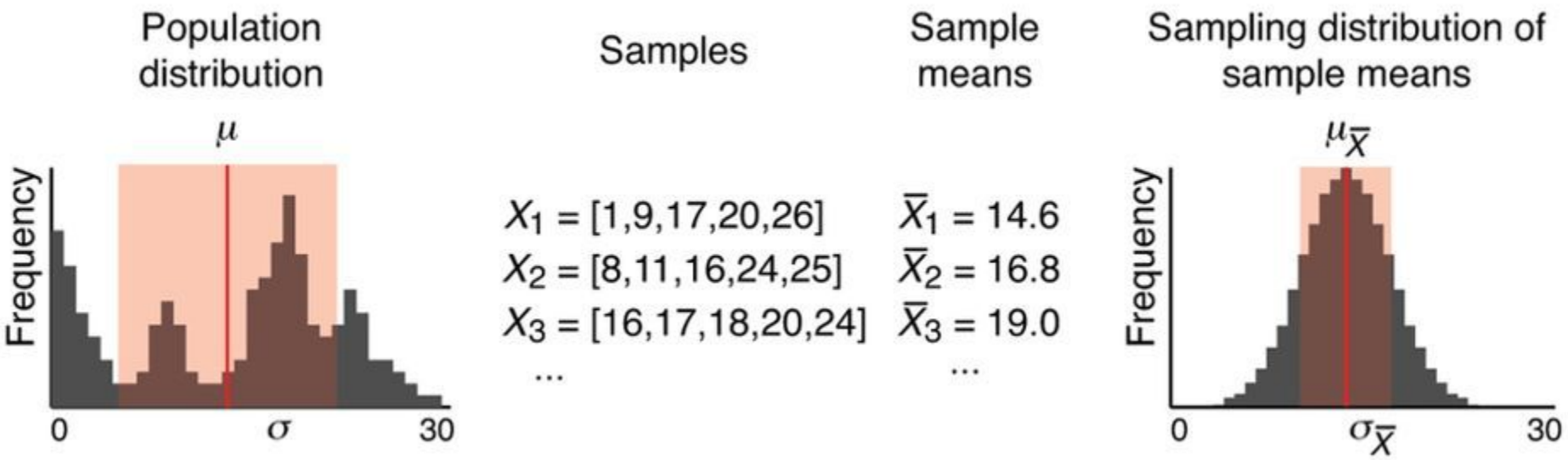
One sample t-test



Two sample t-test



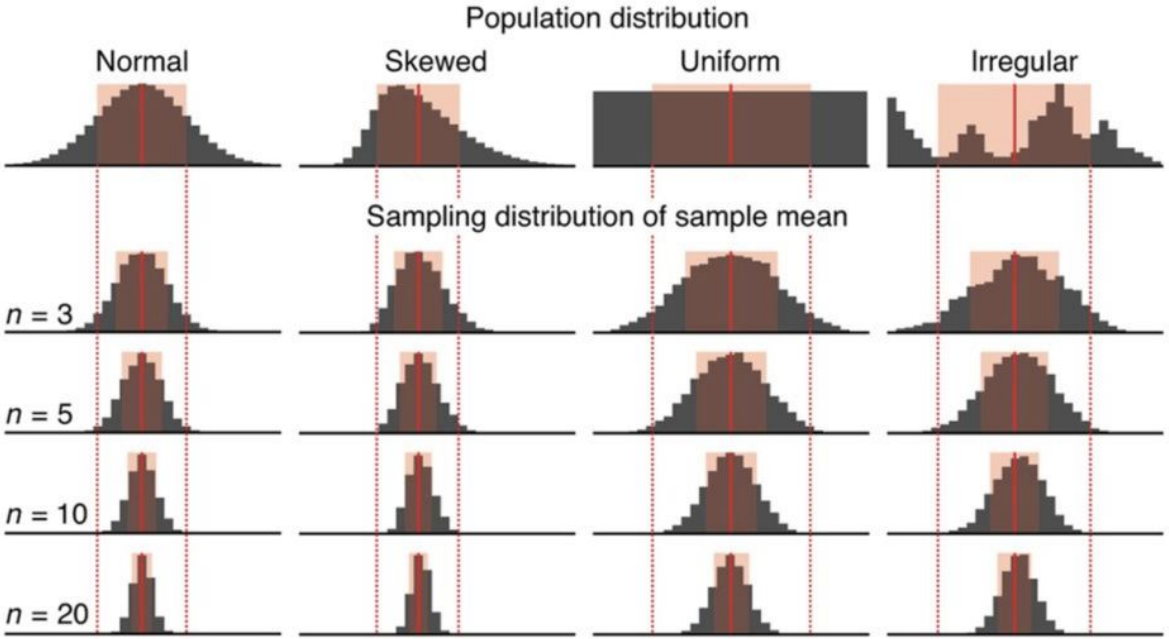
Recall: Central Limit Theorem holds true for any distribution



As the sample size increases, the means of samples will become increasingly close to a normal distribution with a mean (***μ***) equal to the population sample!

https://www.nature.com/articles/nmeth.2613/figures/2

With larger sample sizes (*n*), the standard deviation of the sample means decreases



Shown are sampling distributions of sample means for 10,000 samples for indicated sample sizes drawn from four different distributions. Mean and s.d. are indicated as in Figure 1.

https://www.nature.com/articles/nmeth.2613/figures/3

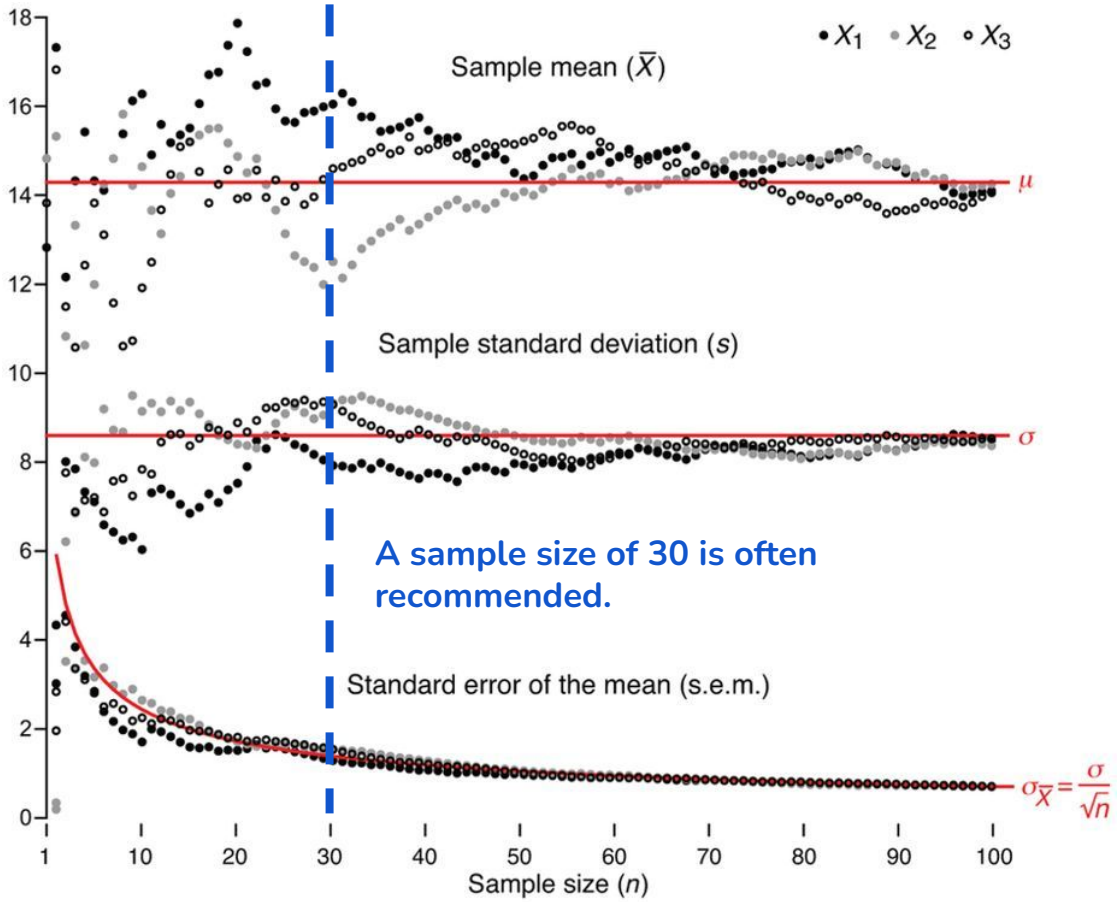
Standard deviation of the sample means.

$$\sigma_{\bar{X}} = \sigma / \sqrt{n}$$

Standard deviation of the population

As *n* increases, $\sigma_{\bar{x}}$ decreases. That is, the samples will have more similar means. Most importantly, when using a larger n, **your sample mean is much more likely to be close to the true population mean. This is important in biology because we typically do one sample of size n (i.e., n replicates).**

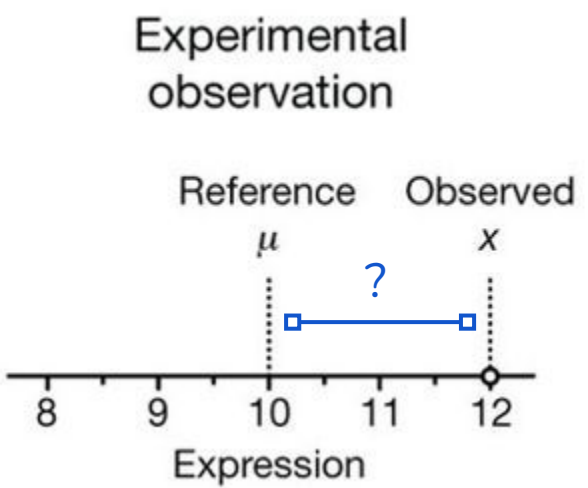
Samples better approximate population as *n* increases.



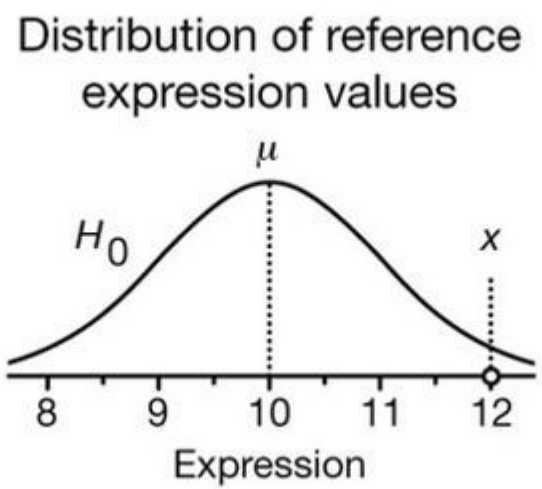
The measured spread of sample means is also known as the standard error of the mean (s.e.m., $SE_{\bar{x}, s/}$) and is used to estimate $\sigma_{\bar{x}}$, which we cannot know because we cannot collect all possible samples.

The s.e.m. ($s/\sqrt{n}$) measures how well the sample mean approximates the population mean.

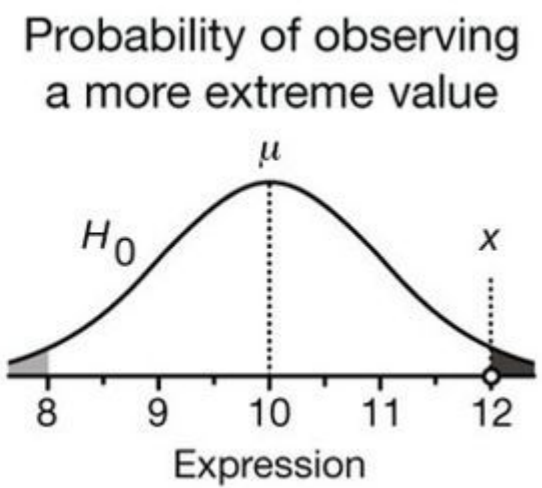
Is our observed value significantly different than a reference value?



Is this observed difference compared to the reference value due to random chance?



Assume that technical factors can lead to random fluctuations that disperse measurements forming the null distribution, which encapsulates the **null hypothesis (H_0)**. We assume that the null distribution is normally distributed.

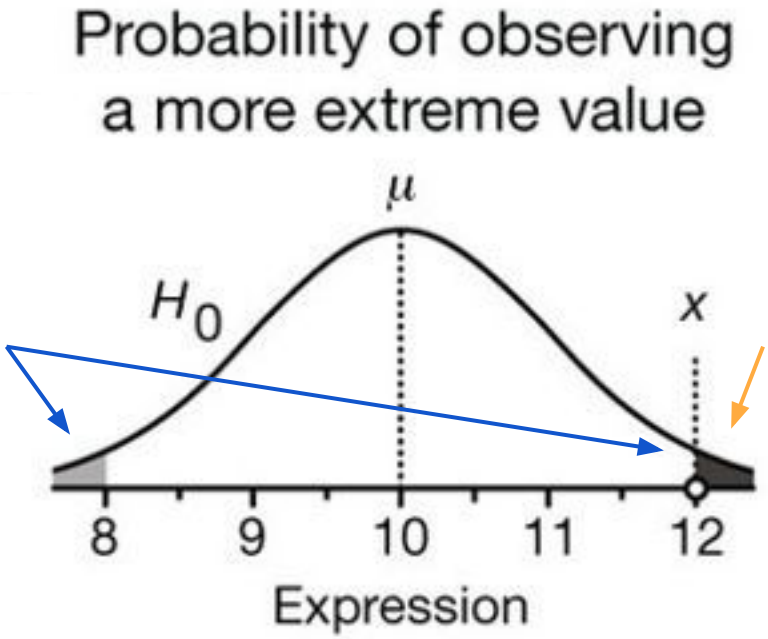


Statistical tests locate an observation on a null distribution to measure the degree to which it is an outlier. In other words, what is the probability of sampling another observation from the null distribution that is as far away from μ? **This probability is the P-value**

https://www.nature.com/articles/nmeth.2698/figures/1

P-values from two-tailed versus one-tailed tests

Two tailed test: what is the probability of measuring a value at least as far as x from μ? That is, the gray and the black areas under the null distribution

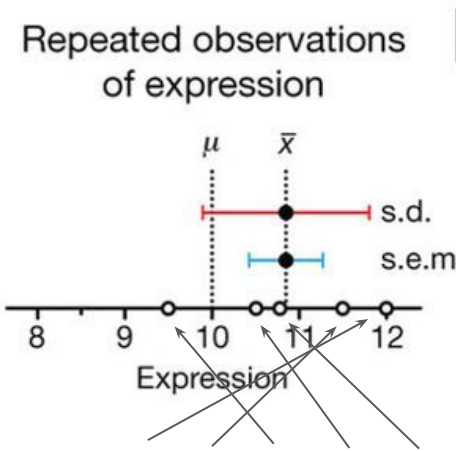


One tailed test: what is the probability of measuring a value greater than x? That is, only the black area under the null distribution

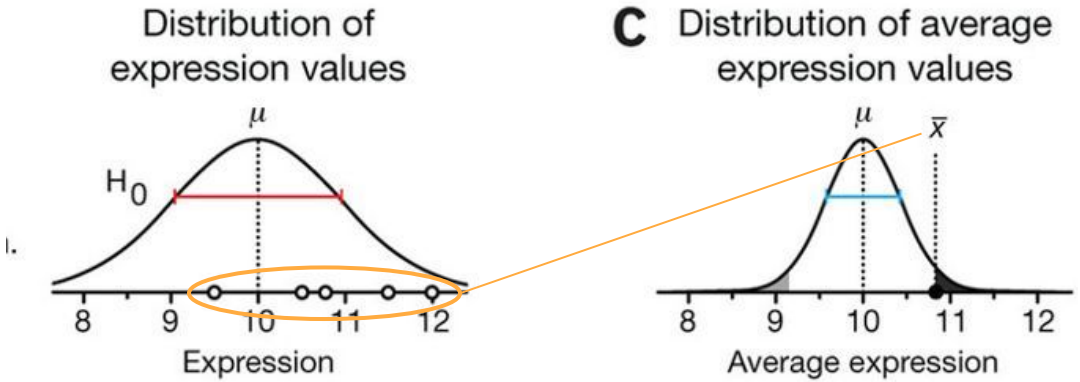
https://www.nature.com/articles/nmeth.2698/figures/1

But we need to control for the noise in our measurement!

We assume the reference value (that is, the population mean, μ=10) comes from many precise estimates and is correct. We need to know the "noise" in our measurement of protein expression to assess whether our observed expression value (x=12) is expected under the null. We therefore estimate the spread of our measurement with repeated measurements.



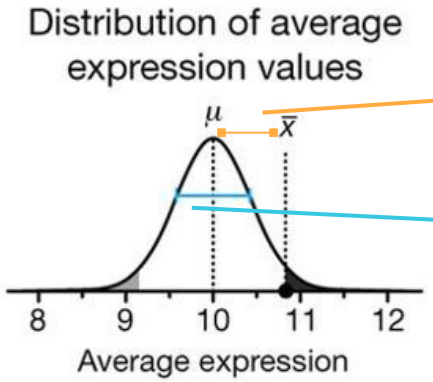

```
obs_exp = c(12.0,11.5,9.5,10.5,10.8)
mean(obs_exp)
[1] 10.86
sd(obs_exp) # s.d.
[1] 0.9607289
sd(obs_exp)/sqrt(length(obs_exp)) # s.e.m
[1] 0.429651
```



We assume that our sample standard deviation is representative of the standard deviation of the null distribution, even if our sample mean is not. This assumption of "equal variances" is commonly used.

As we learned from the Central Limit Theorem, since the null distribution is normal, we know the the sampling distribution of means will also be normal. Therefore, we can use the s.e.m. to estimate the s.d. on the null distribution of the sample means. We then locate the average expression on this revised null to compute a p-value.

The test statistic (*t*) measures "signal" versus "noise"



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

"signal"

"Noise" (s.e.m.)

As we learned from the Central Limit Theorem, since the null distribution is normal, we know the the sampling distribution of means will also be normal. Therefore, we can use the s.e.m. to estimate the s.d. on the null distribution of the sample means. We then locate the average expression on this revised null to compute a p-value.

```
obs_exp = c(12.0,11.5,9.5,10.5,10.8)
xhat = mean(obs_exp) # xhat
[1] 10.86
s = sd(obs_exp) # s.d.
[1] 0.9607289
sem = sd(obs_exp)/sqrt(length(obs_exp)) # s.e.m
[1] 0.429651
tstat = (xhat - 10.0)/sem # u, the reference value is 10
[1] 2.001624
```

https://www.nature.com/articles/nmeth.2613/figures/3

William Gosset, a.k.a. "Student". Worked at Guinness

"In his job as Head Experimental Brewer at Guinness, the self-trained Gosset developed new statistical methods. In his job as Head Experimental Brewer at Guinness, the self-trained Gosset developed new statistical methods. He published under the pseudonym 'Student' (to avoid difficulties with his employer, Guinness) in his work on optimizing barley yields. Trained with Karl Pearson. Developed most famously Student's t-distribution (originally called Student's "z") and "Student's test of statistical significance". https://en.wikipedia.org/wiki/William_Sealy_Gosset

THE PROBABLE ERROR OF A MEAN

BY STUDENT

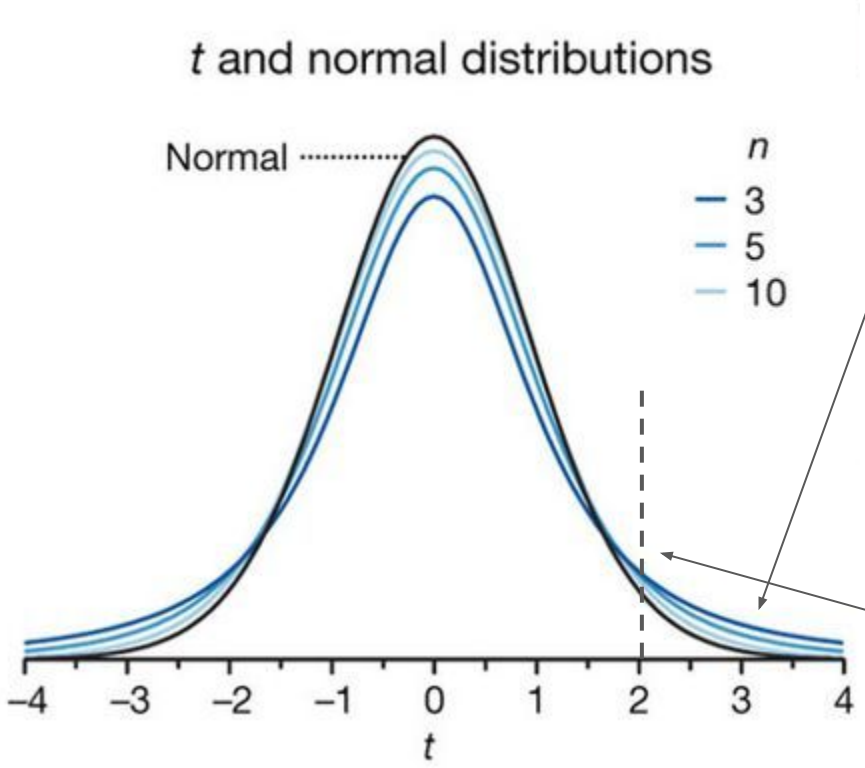
Introduction

Any experiment may he regarded as forming an individual of a "population" of experiments which might he performed under the same conditions. A series of experiments is a sample drawn from this population.

Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a greater number of cases the question finally turns on the value of a mean, either directly, or as the mean difference between the two quantities.

If the number of experiments be very large, we may have precise information as to the value of the mean, but if our sample be small, we have two sources of uncertainty: (1) owing to the "error of random sampling" the mean of our series of experiments deviates more or less widely from the mean of the population, and (2) the sample is not sufficiently large to determine what is the law of distribution of individuals. It is usual, however, to assume a normal distribution, because, in a very large number of cases, this gives an approximation so close that a small sample will give no real information as to the manner in which the population deviates from normality: since some law of distribution must he assumed it is better to work with a curve whose area and ordinates are tabled, and whose properties are well known. This assumption is accordingly made in the present paper, so that its conclusions are not strictly applicable to populations known not to be normally distributed; yet it appears probable that the deviation from normality must be very extreme to load to serious error. We are concerned here solely with the first of these two sources of uncertainty.

The shape of the sampling distribution is close to, but not quite normal. Instead, we use the t-distribution.



The t distribution has higher tails than the normal that take into account that most samples will underestimate the variability in a population. Better estimate of "noise"

The test statistic (t) is compared to this distribution and is thus called the t-statistic.

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

```
t = (xhat - 10.0)/sem
[1] 2.001624
```

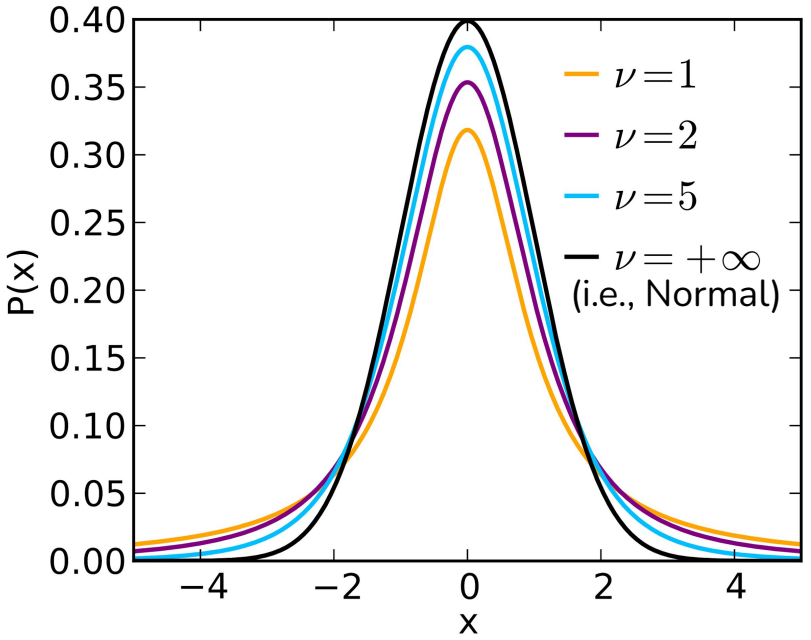
https://www.nature.com/articles/nmeth.2613/figures/3

T-distribution parameters: mean, degrees of freedom (*v*)

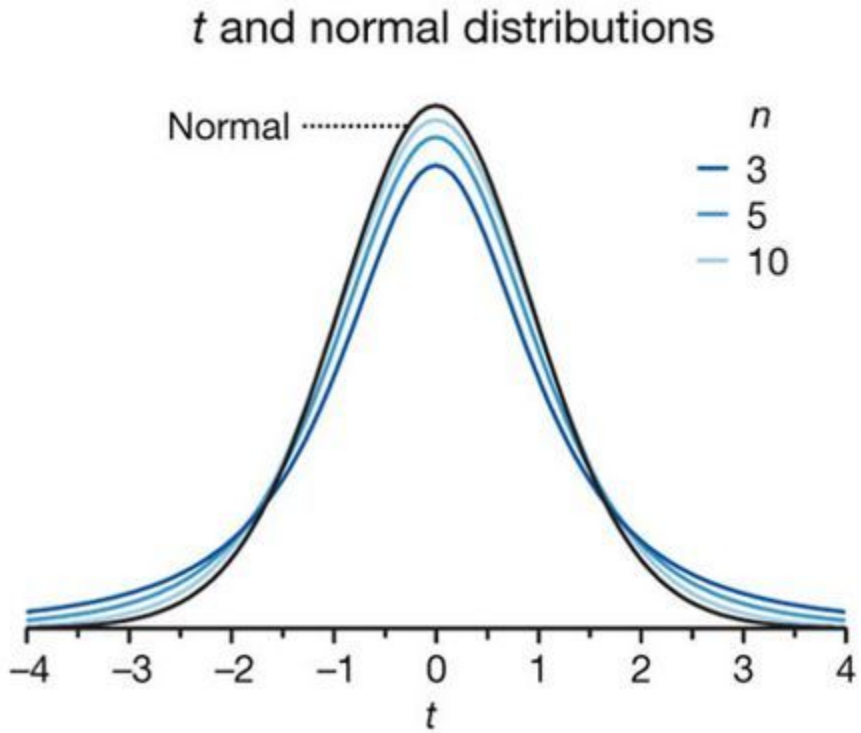
The mean is assumed to be 0, unless otherwise stated. Why?
Well, the test statistic, t, is itself a random variable and it comes from the difference between the sample mean and the population mean. The null distribution of this random variable should be centered at 0 since the sample mean should be greater than the population mean just as frequently as it is less than the population mean.

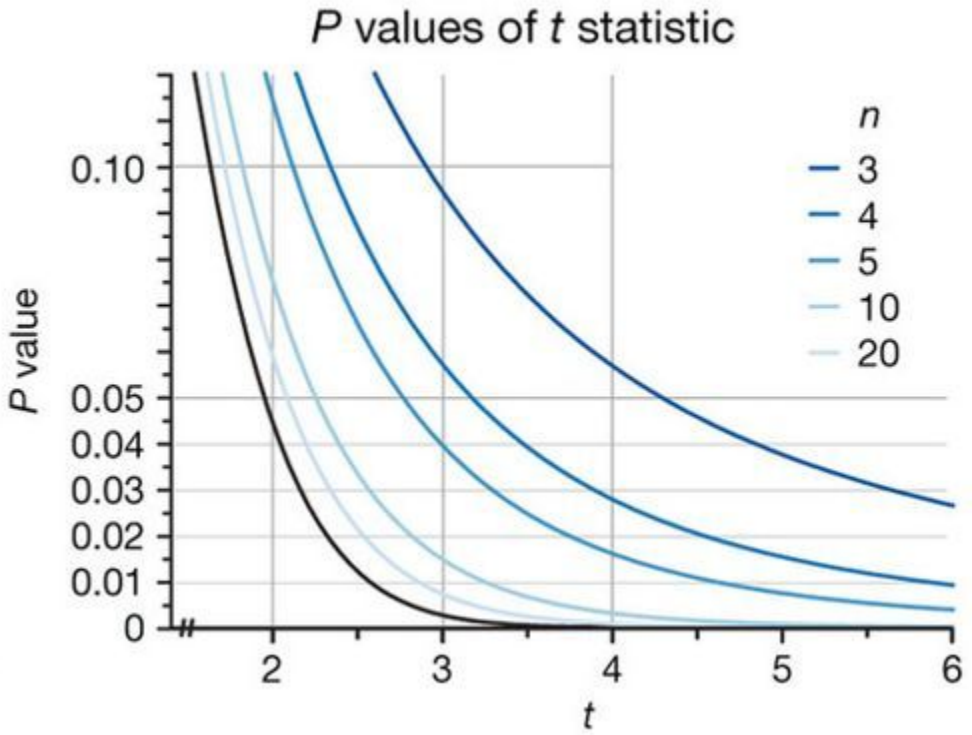
$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

The second parameter is the degrees of freedom (*v*), which is the number of variables that are free to independently vary. For the t-distribution, the degrees of freedom (*v*) is n-1. **Why n-1?** If you know sample mean, you only need n-1 of the samples to infer the nth sample (algebra). It has no "freedom" to vary.



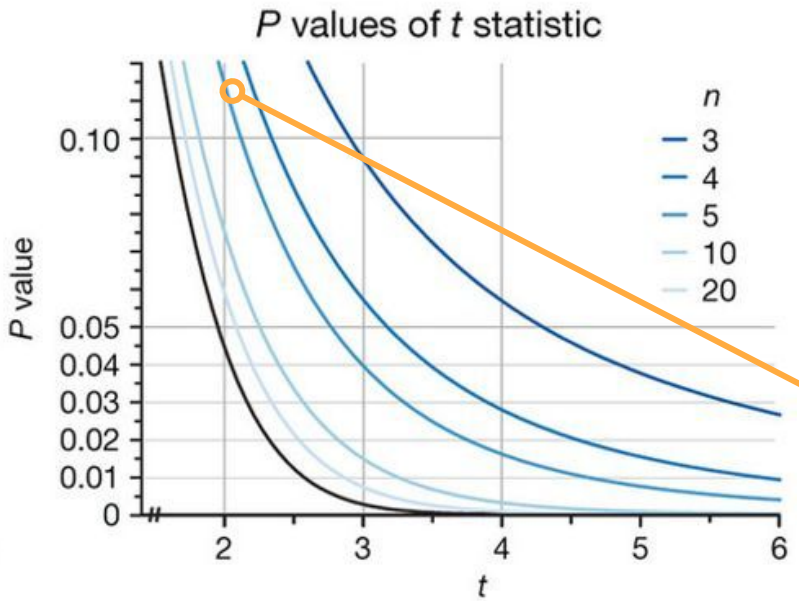
"When n is small, P values derived from the t-distribution vary greatly as n changes."





https://www.nature.com/articles/nmeth.2613/figures/3

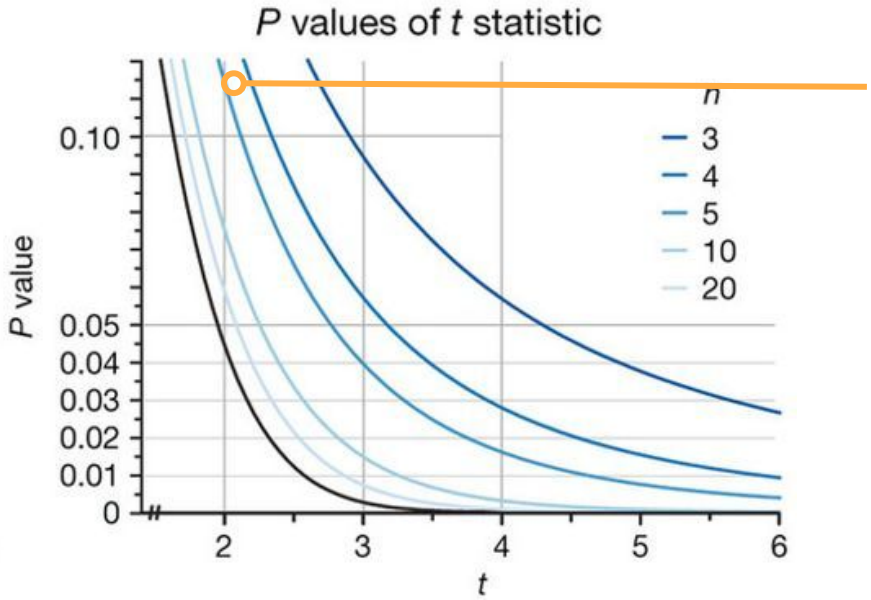
The test statistic is compared to the t-distribution with the correct n (and thus degrees of freedom ($v = n-1$)




```
obs_exp = c(12.0,11.5,9.5,10.5,10.8)
xhat = mean(obs_exp) # xhat
[1] 10.86
s = sd(obs_exp) # s.d.
[1] 0.9607289
sem = sd(obs_exp)/sqrt(length(obs_exp))
[1] 0.429651
t = (xhat - 10.0)/sem
[1] 2.001624
```

https://www.nature.com/articles/nmeth.2613/figures/3

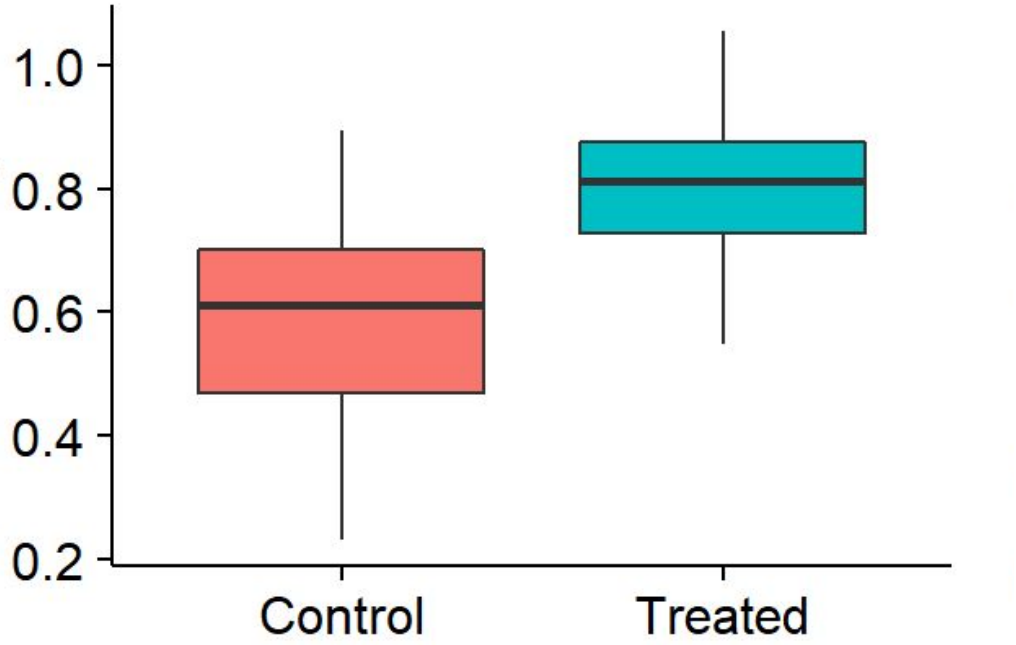
Use the R function t.test() to compute a one-sample t-test!



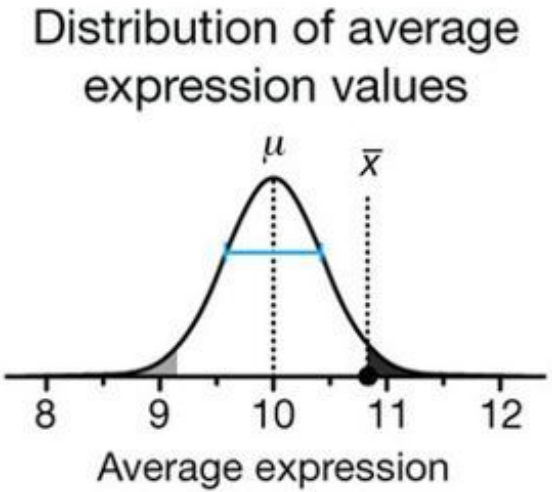

```
t = (xhat - 10.0)/sem
[1] 2.001624

t.test(obs_exp, mu = 10, alternative = "two.sided")

     One Sample t-test

data:  obs_exp
t = 2.0016, df = 4, p-value = 0.1159
alternative hypothesis: true mean is not equal to 10
95 percent confidence interval:
  9.667098 12.052902
sample estimates:
mean of x
    10.86
```

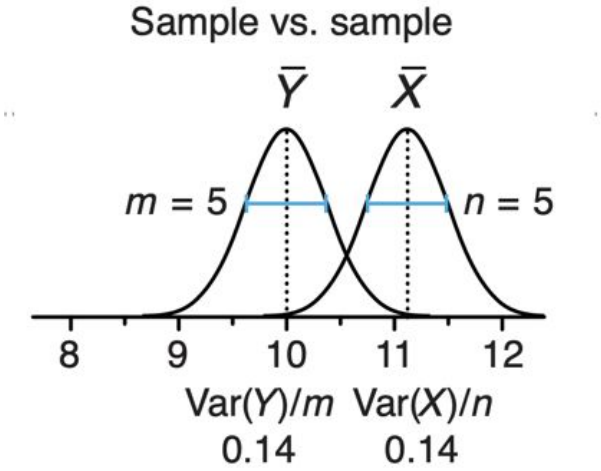
https://www.nature.com/articles/nmeth.2613/figures/3

In biology, we often want to compare the means of two different samples.



One-sample versus two-sample t-test



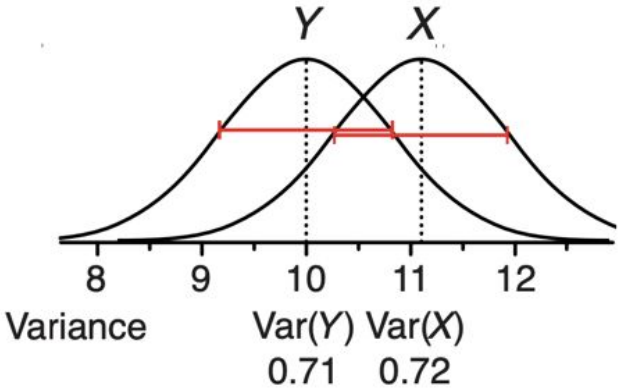


One sample t-test: Compare observed sample mean to reference value while accounting for uncertainty in the sample mean.

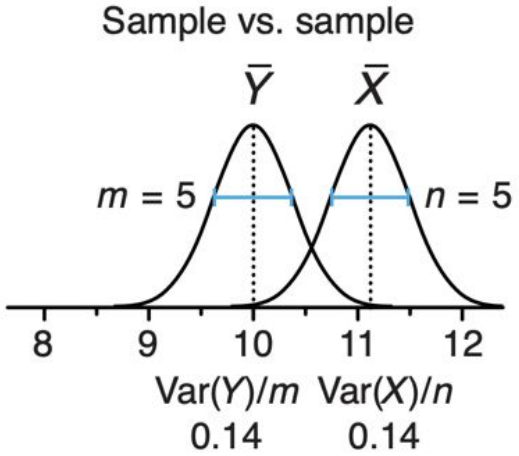
Two sample t-test: Compare two observed samples means while accounting for the **combined uncertainty in the sample means**.

https://www.nature.com/articles/nmeth.2613/figures/3

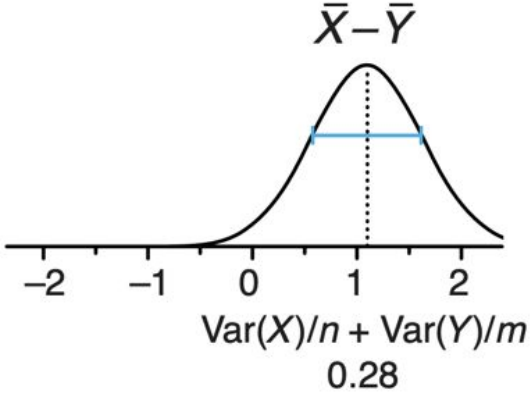
Two-sample t-test



Start with the variance of the two conditions, X (untreated) and Y (treated)



Compute the mean of each condition and adjust the variance by the sample size to assess how uncertain the variance for each condition is.



The test statistic (t) now reflects the difference of the means (approximately 1) with respect to the "pooled variance" of the two conditions

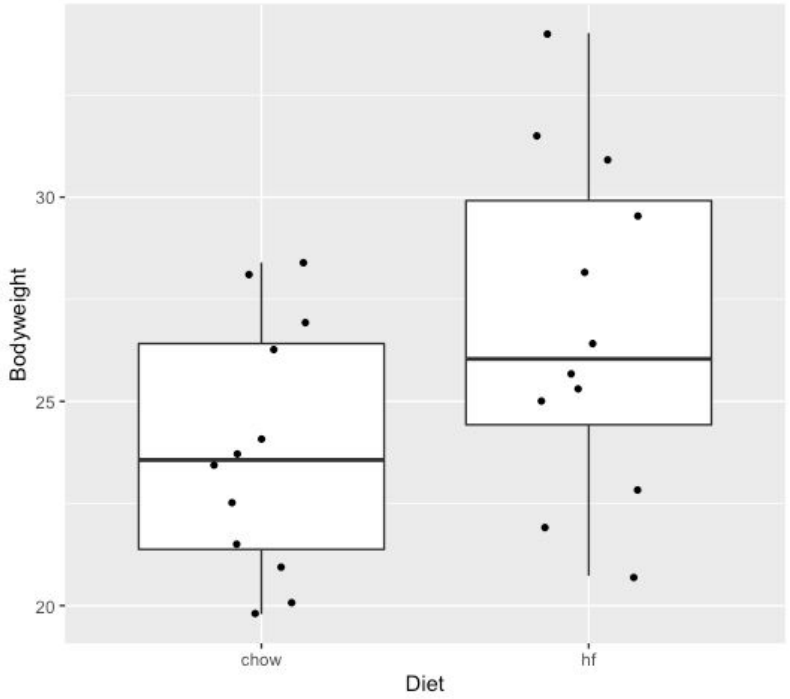
Let's work through an example. Compare mouse fed different diets. Modified from Love and Irizarry.

```
# read in weight data from mice fed two different diets.
      dat =
read.csv("https://raw.githubusercontent.com/genomicsclass/da
gdata/master/inst/extdata/femaleMiceWeights.csv")

# peek
head(dat, n=3)
  Diet Bodyweight
1 chow      21.51
2 chow      28.14
3 chow      24.04
```

http://genomicsclass.github.io/book/pages/clt_and_t-distribution.html, http://genomicsclass.github.io/book/pages/t-tests_in_practice.html

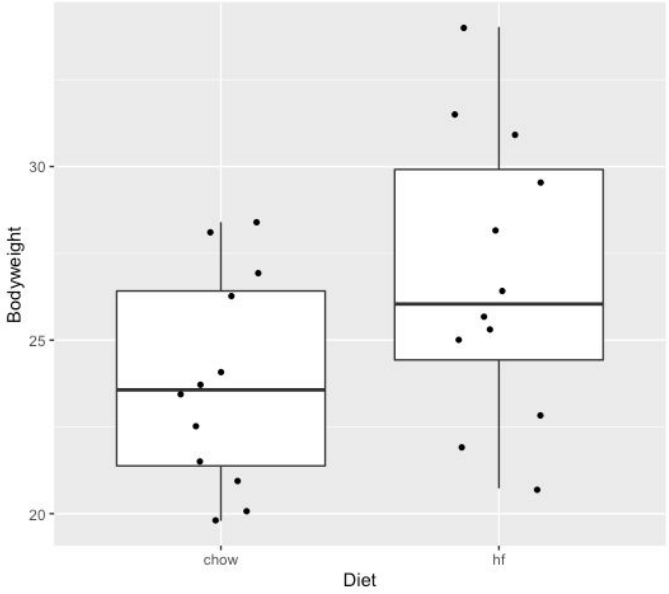
Plot the female mice fed different diets.

```
# read in weight data from mice fed two different diets.
dat =
read.csv("https://raw.githubusercontent.com/genomicsclass/da
gdata/master/inst/extdata/femaleMiceWeights.csv")

# peek
head(dat, n=3)
  Diet Bodyweight
1 chow      21.51
2 chow      28.14
3 chow      24.04

# bring in useful libraries
library(dplyr)
library(ggplot2)

# box plot the weights of the control and
# high fat female mice
dat %>% ggplot(aes(x=Diet, y=Bodyweight)) +
    geom_boxplot() +
    geom_jitter(shape=16, position=position_jitter(0.2))
```



http://genomicsclass.github.io/book/pages/clt_and_t-distribution.html, http://genomicsclass.github.io/book/pages/t-tests_in_practice.html

Welch's two-sample t-test (unequal variances)



Reminder: standard deviation is the square root of the variance

The test statistic (t) for one sample is a measure of signal versus noise

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

"signal"

"Noise" (s.e.m.)

Same goes for the two-sample test, but we must account for "pooled" noise.

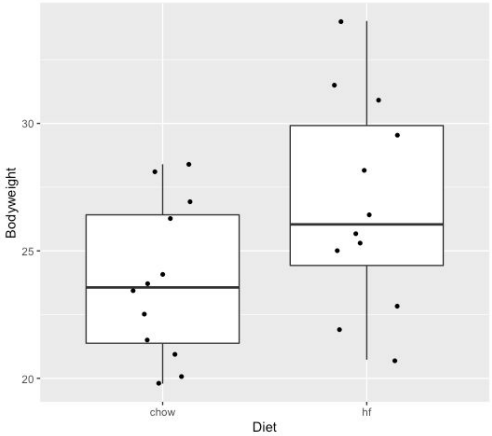
$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

"signal"

"Noise" pooled

http://genomicsclass.github.io/book/pages/clt_and_t-distribution.html, http://genomicsclass.github.io/book/pages/t-tests_in_practice.html

Welch's two-sample t-test (unequal variances)



$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

"signal"

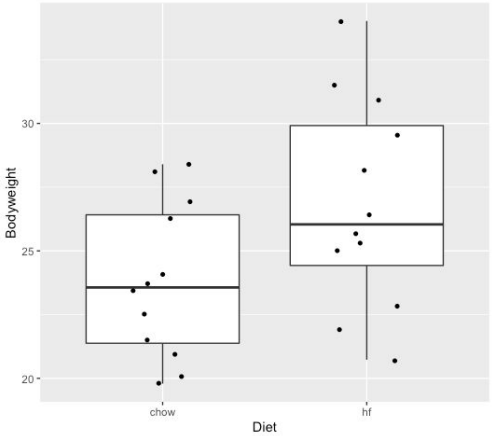
"Noise"

pooled

```
# subset the two groups
control = dat %>% filter(Diet=="chow")
highfat = dat %>% filter(Diet=="hf")
```

http://genomicsclass.github.io/book/pages/clt_and_t-distribution.html, http://genomicsclass.github.io/book/pages/t-tests_in_practice.html

Welch's two-sample t-test (unequal variances)



$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

"signal"

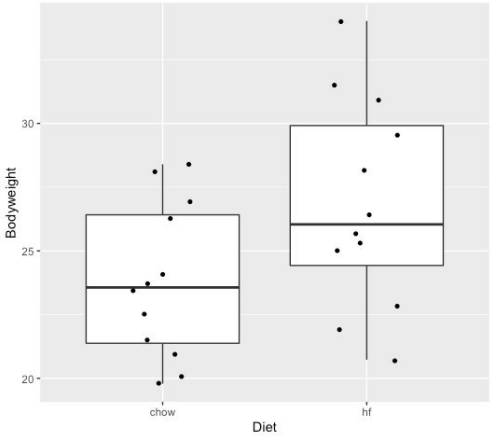
"Noise"

pooled

```
# subset the two groups
control = dat %>% filter(Diet=="chow")
highfat = dat %>% filter(Diet=="hf")
# compute the difference of means. The numerator of t
diff_of_means = mean(highfat$Bodyweight) - mean(control$Bodyweight)
# [1] 2.375517
# theory tells us that the variance of the difference of two random variables is the
sum of its variances, so we compute the variance and take the square root
pooled_se_noise <- sqrt(var(highfat$Bodyweight)/nrow(highfat) +
                  var(control$Bodyweight)/nrow(control))
```

http://genomicsclass.github.io/book/pages/clt_and_t-distribution.html, http://genomicsclass.github.io/book/pages/t-tests_in_practice.html

Welch's two-sample t-test (unequal variances)



$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

"signal"

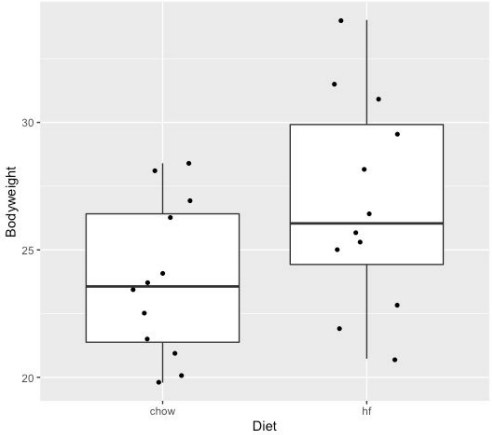
"Noise"

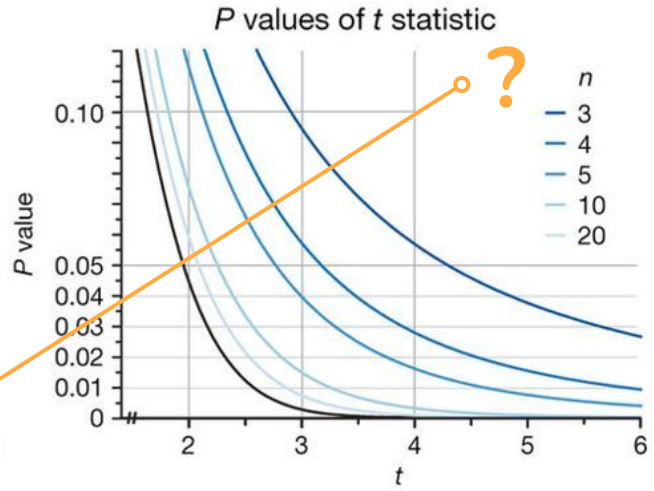
pooled

```
# subset the two groups
control = dat %>% filter(Diet=="chow")
highfat = dat %>% filter(Diet=="hf")
# compute the difference of means. The numerator of t
diff_of_means = mean(highfat$Bodyweight) - mean(control$Bodyweight)
# [1] 2.375517
# theory tells us that the variance of the difference of two random variables is the
sum of its variances, so we compute the variance and take the square root
pooled_se_noise <- sqrt(var(highfat$Bodyweight)/nrow(highfat) +
                  var(control$Bodyweight)/nrow(control))
# compute tstat
tstat = diff_of_means / pooled_se_noise
```

http://genomicsclass.github.io/book/pages/clt_and_t-distribution.html, http://genomicsclass.github.io/book/pages/t-tests_in_practice.html

Welch's two-sample t-test (unequal variances)

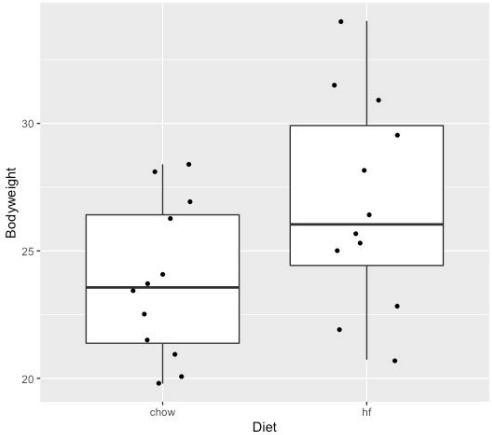





```
# subset the two groups
control = dat %>% filter(Diet=="chow")
highfat = dat %>% filter(Diet=="hf")
# compute the difference of means. The numerator of t
diff_of_means = mean(highfat$Bodyweight) - mean(control$Bodyweight)
# [1] 2.375517
# theory tells us that the variance of the difference of two random variables is the
sum of its variances, so we compute the variance and take the square root
pooled_se_noise <- sqrt(var(highfat$Bodyweight)/nrow(highfat) +
                        var(control$Bodyweight)/nrow(control))
# compute tstat
tstat = diff_of_means / pooled_se_noise
#[1] 2.055174
```

http://genomicsclass.github.io/book/pages/clt_and_t-distribution.html, http://genomicsclass.github.io/book/pages/t-tests_in_practice.html

Welch's two-sample t-test (unequal variances)



$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

"signal"

"Noise"

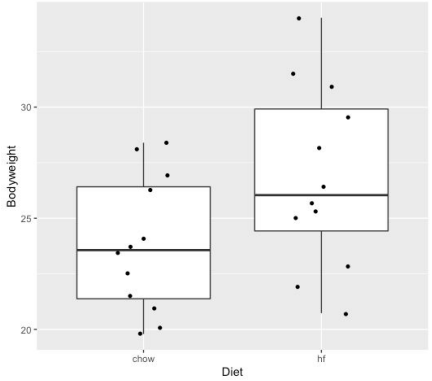
pooled

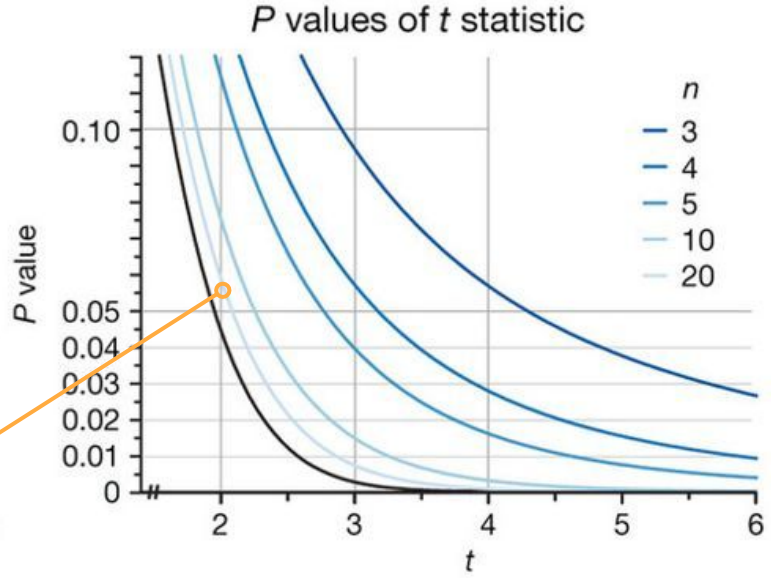
```
# compute the two sample t-test
t.test(highfat$Bodyweight, control$Bodyweight)

      Welch Two Sample t-test
data:  highfat$Bodyweight and control$Bodyweight
t = 2.0552, df = 20.236, p-value = 0.053
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.04296563  6.08463229
sample estimates:
mean of x mean of y
 26.83417  23.81333
```

Here, the **degrees of freedom is close to but not quite N-1 + M-1**. Instead it is a more complicated "pooled" degrees of freedom. Read this for more info.

http://genomicsclass.github.io/book/pages/clt_and_t-distribution.html, http://genomicsclass.github.io/book/pages/t-tests_in_practice.html

What could be done to increase confidence in rejecting the null hypothesis?



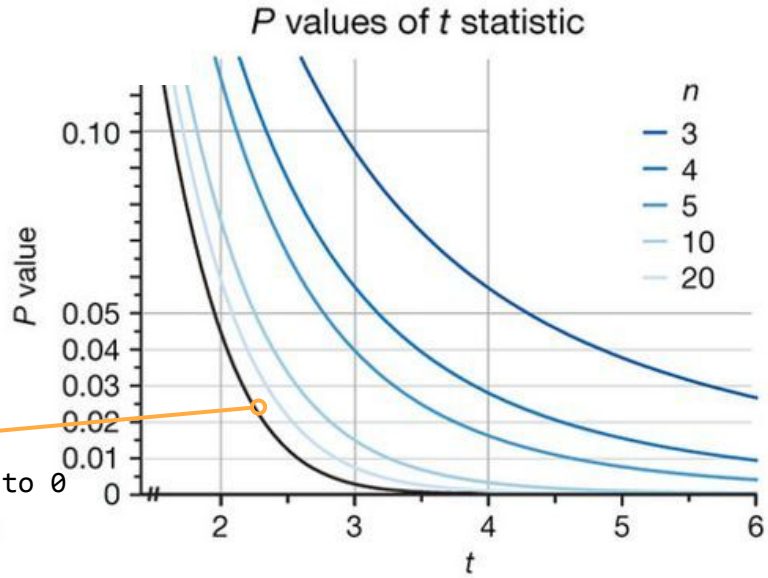



```
# compute the two sample t-test
t.test(highfat$Bodyweight, control$Bodyweight)

      Welch Two Sample t-test
data:  highfat$Bodyweight and control$Bodyweight
t = 2.0552, df = 20.236, p-value = 0.053
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.04296563  6.08463229
sample estimates:
mean of x mean of y
 26.83417  23.81333
```

http://genomicsclass.github.io/book/pages/clt_and_t-distribution.html, http://genomicsclass.github.io/book/pages/t-tests_in_practice.html

If we add more data supportive of alt. hypothesis, the t-statistic and degrees of freedom lead to p-value < 0.05

```
# add two new data points that support alt. Hypothesis
highfat = rbind(highfat, list("hf", as.numeric(30.22)))
control = rbind(control, list("chow", as.numeric(24.23)))

# compute the two sample t-test
t.test(highfat$Bodyweight, control$Bodyweight)

      Welch Two Sample t-test

data:  highfat$Bodyweight and control$Bodyweight
t = 2.3591, df = 21.774, p-value = 0.02771
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.3911423 6.1073193
sample estimates:
mean of x mean of y
 27.09462  23.84538
```



http://genomicsclass.github.io/book/pages/clt_and_t-distribution.html, http://genomicsclass.github.io/book/pages/t-tests_in_practice.html

Analysis of Variance (ANOVA) is essentially the t-test for more than two groups.