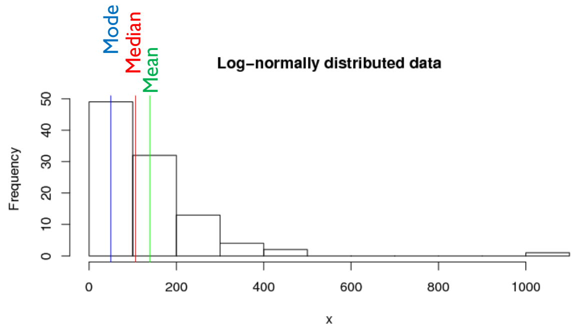
Session 5
Single sample
Why variance matters
Two samples
Hypothesis testing

Single Sample

- Central tendency (location)
  - Mode, median, mean
- Measures of variation (dispersion, spread)
  - Range, quantiles, box-and-whisker plot
  - Residuals and sum of squares
  - Degrees of freedom
  - Variance
  - Standard deviation
  - Standard error (of the mean)
  - Coefficient of Variation

Central tendency

- Mode: data values that occur most often (highest frequency)
- Median: the 'middle' of a sorted data set
- Mean (arithmetic mean): the centre of mass $m = \frac{\sum_{i=1}^{i=n} x_i}{n}$



Central tendency

- Mean (m, μ, $\bar{x}$)
  - 'Outliers' have a large influence on the mean
  - Ideal if distribution symmetric and no outliers
- Median
  - How to calculate median?
    - Example data: 17 14 20 12 10 10 13 9 6 11
    - Sort data and take the value in the 'middle' (interpolate if even)
    - 6 9 10 10 $\boxed{11\ 12}$ 13 14 17 20
    - Mode = 10 Median = 11.5 Mean = 12.2
  - Less influenced by outliers than mean: robust statistic
  - In statistics, robust always means less influenced by 'outliers'

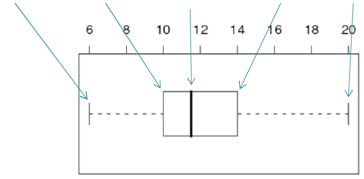
Measures of variation

- Quantiles
  - When you divide an ordered data set into (approximately) equally spaced subsets, the quantiles mark the boundaries between subsets
  - Quantiles are points taken at regular intervals from the cumulative distribution function (I'll explain that soon)
  - Quantile is the general term, special quantiles typically used are:
    - 4-quantiles are called quartiles (divide data into quarters)
    - 100-quantiles are called percentiles
    - Median = 2/4 quartile = 50% percentile

Measures of variation

- Example data set used for central tendency before
  - 17 14 20 12 10 10 13 9 6 11
  - Sorted data set
  - 6 9 10 10 11 12 13 14 17 20
  - R functions quantile(), boxplot() (no need to sort data beforehand)

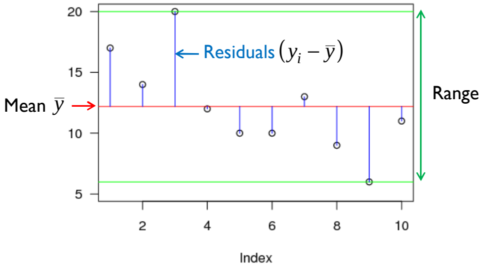
```
> quantile(x, c(0,25,50,75,100)/100)
   0%    25%    50%    75%   100%
 6.00 10.00 11.50 13.75 20.00
```



Box-and-whisker plot:
Whiskers: extremes
Box: ¼ and ¾ quartiles
Bar in box: median
NB: this is a rotated boxplot()

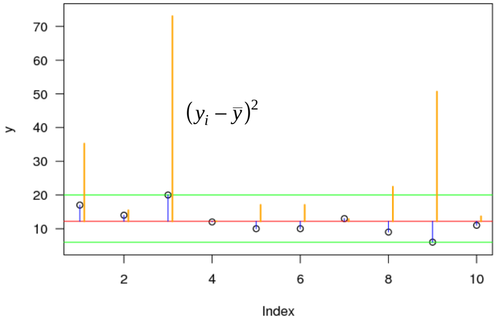
Measures of variation

- Range: max(y) – min(y)
- Residuals (errors, deviates): distances of data points to mean
  - The sum of the residuals is zero! So we need something else!



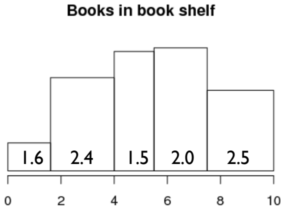
Measures of variation

- Sum of squares (SS): sum of squared residuals $SS = \sum_i (y_i - \bar{y})^2$
  - squaring makes all values positive so the sum is no longer zero
  - but large residuals now have more weight



Measures of variation

- Degrees of freedom
  - Example: n=5 books in bookshelf, total width = 10 cm
  - Measure first book = 1.6 cm
  - Measure second book = 2.4 cm
  - Measure third book = 1.5 cm
  - Measure fourth book = 2.0 cm
  - No need to measure last book!
    - There is only a gap of 2.5 cm left
  - If you calculate the sum (or anything that involves the sum, e.g. mean, variance) from your data, one value can be calculated from the sum and all the other values, it's value is no longer free
  - Degrees of freedom (d.f.) = n-1



Measures of variation

- Degrees of freedom
  - More generally, the degrees of freedom are defined as the
  - number of data points n – number of parameters p estimated from the data
  - Degrees of freedom (d.f.) = n-p

Measures of variation

- Variance

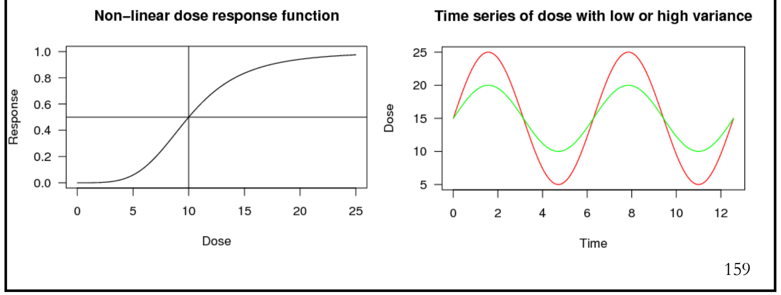
$$\text{variance} = \frac{\text{sum of squares}}{\text{degrees of freedom}}$$

$$s^2 = \frac{\sum_i (y_i - \bar{y})^2}{n-1}$$

  - In calculating the sum of squares, we have used the mean, and therefore the sum of the data, so d.f. is n-1
  - The units of the variance differ from the units of the data, e.g. if you have measured flowering times in days, then the variance is in terms of days^2, whatever that means
    - This is made obvious by using s^2 for the variance
    - This is why we use standard deviation s, which is in the same units as the data

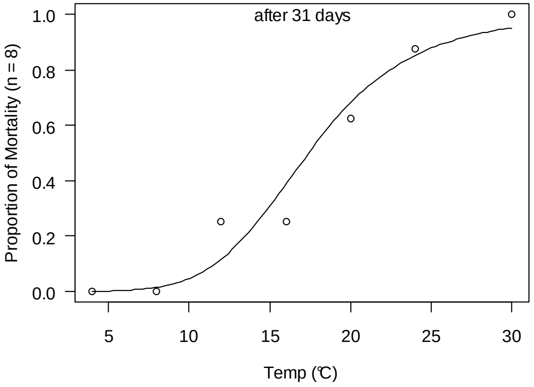
Does it matter if variance is high or low?

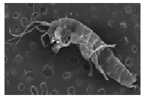
- Differences in variance can lead to differences in biological response even if the means are the same
  - If response is a non-linear function of the dose, the average of the response (0.65 for red time series with high variation, 0.78 for the green time series with low variation) is not the same as the response to the average dose of 15, which is 0.84
  - The same mean can have different outcomes if variances differ!



Does it matter if variance is high or low?

- An example of a non-linear dose response curve is survival (response) as a function of temperature (dose)





Lethality of temperature on the groundwater amphipod *Niphargus cf. bajuvaricus* after 31 days.

Data from Susanne Schmidt

Measures of variation

- Measures based on variance s^2 but generally more useful
  - Standard deviation (SD, s.d., s, σ): same units as data, so you can write e.g. 175 ± 8 cm (mean ± s.d.)

$$s = \sqrt{s^2}$$

  - Standard error (of the mean) (SE, s.e., s.e.m.): while SD is variation for individual data points, s.e.m. is variation for the sample mean

$$s.e.m. = \sqrt{\frac{s^2}{n}} = \frac{s}{\sqrt{n}}$$

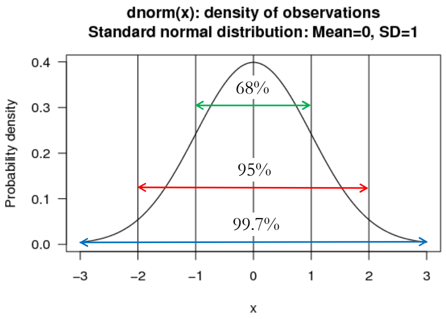
  - Coefficient of variation (CV): no units (dimensionless)!

$$CV = \frac{s}{m}$$

    - CV is SD relative to the mean, more useful measure than absolute SD when comparing variation across different populations (SD of masses of mice and elephants are different for sure, but does that hold for the CV?)

Standard deviation and Normal distribution

- How many data are between mean ± standard deviation?
  - The standard normal distribution with 0 ± 1 (mean ± s.d.)
  - Plot the probability density function with dnorm(x)



A note on symbols

- Different symbols are used for the same thing in different texts, e.g. the mean can be $\bar{x}, m, \mu$
- Often it is important to distinguish the sample mean from the population or true mean, etc.
  - The convention is to use Greek letters for the true or population statistics and Latin letters for the sample statistics, which are estimates of the true values

Mean	μ	m
Variance	σ^2	s^2
Standard deviation	σ	s

- Fortunately, the meaning of a symbol or term is clear from the mathematical description, e.g.

$$\rho = \frac{\sum_{i=1}^{i=n} x_i}{n}$$

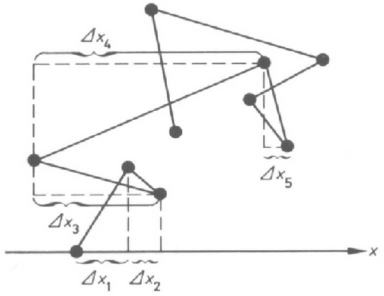
Measures of variation

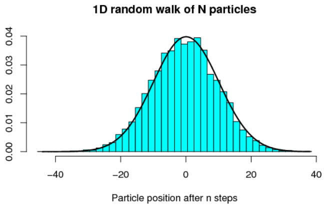
- Why squaring the residuals?
  - Why not just drop the sign (i.e. take the absolute values?)
  - This is mad: median absolute deviation
    - The median of the absolute deviations from the median
    - Robust against outliers: squaring boosts weight of outliers
    - R function: mad()

Measures of variation

- Why squaring the residuals? It's Normal!
  - Squaring many little additive errors leads to the Normal distribution
  - Example: random walk
  - Variance = mean-square displacement

$$s^2 = \frac{SS}{d.f.} = \frac{\sum(x_i - \bar{x})^2}{n-1} \cong \overline{\Delta x^2} = 2Dt$$





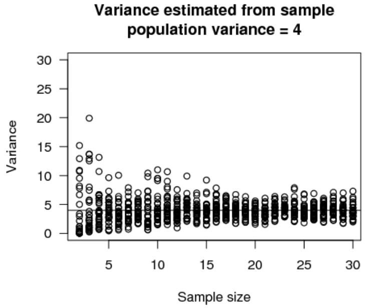
Exercise

- Data: ozone concentrations in garden A (from Crawley's book)
  - 3, 4, 4, 3, 2, 3, 1, 3, 5, 2
- Compute the following statistics
  - Mean
  - Median
  - Percentiles (0%, 25%, 50%, 75%, 100%)
  - Residuals (all residuals, i.e. a vector with the same # of elements as the data vector)
  - Median absolute deviation
  - Sum of squares
  - Variance
  - Standard deviation
  - Standard error
  - Coefficient of variation

Two samples:
Small samples and t distribution
t-test as an example hypothesis test

The problem of small samples

- Estimating variance from small samples
  - Samples drawn from normally distributed data (mean=10, SD=2)

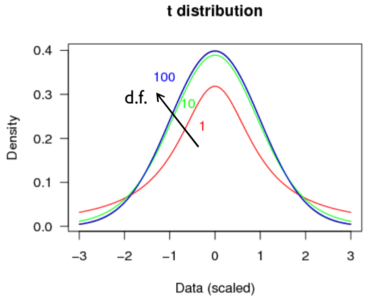


The solution:

Student's t distribution

Student's t distribution

- Student's t distribution versus the Normal distribution
  - The t distribution is like the Normal distribution with fat tails
  - The larger the sample size (degrees of freedom) the more Normal
  - Below 10 d.f. the tails of the t distribution are much fatter



NB: the Normal distribution in black is underneath the t distribution with d.f. = 100

Student's t distribution

- t distribution considers the effect of small sample sizes
  - If samples are small, the mean and variance that we estimate from the sample can be quite far from the true mean and variance of the population
  - So there is not only variance in the population but also lack of precision of the estimated mean and variance

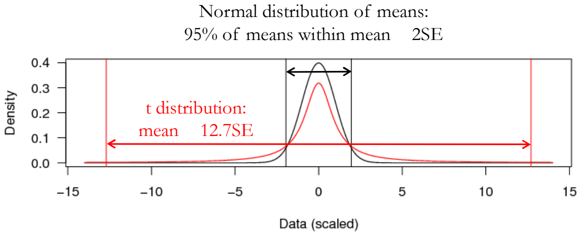
Student's t distribution

- The conception of the t distribution was a revolution for statistics as it made it possible to apply statistical techniques to everyday problems with their often very small sample sizes
- Published by William Gosset under the pseudonym Student as his employer, Guinness, regarded it as a trade secret that they were using statistics to do quality control of beer and improvements in agriculture (barley yield)

Student's t distribution

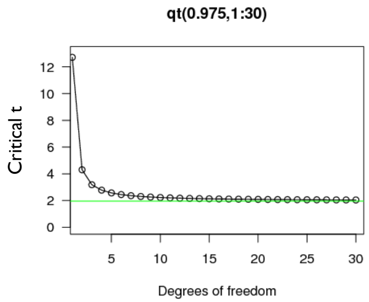
- Probability density of Normal and t distribution (df=1) again, now for the distribution of means
  - Usually we do hypothesis testing with an α error of 0.05. This corresponds to a left tail of 2.5% and a right tail of 2.5% of the values (if distribution is symmetric)

```
> qt(c(0.025,0.975),df=1)
[1] -12.70620  12.70620
```



Student's t distribution

- The critical t, `qt(0.975,df=1:30)`, asymptotically approaches 2 with increasing degrees of freedom (sample size – 1)
  - Remember experimental design: power.t.test(): t depends on d.f. and α and β



$$n = 2\frac{t^2 s^2}{\delta^2}$$

t-test

Hypothesis testing by way of example: t-test

- How does a hypothesis test work?
  - Null hypothesis of no effect: H_0 here is that the means of both populations are the same (we hypothesize about the population, no the sample – the sample is what we know!)
  - test statistic: here t-statistic
  - critical value, take from probability distribution: here t distribution
  - if test statistic > critical value, then reject H_0 and say that the results are significant. In the case of the t-test, this means that the means are significantly different

t-test

- What does the t-test test?
  - Null hypothesis that the means of the two populations are the same
- When can the t-test be used?
  - Normally distributed errors
  - Quantitative data (better than ranks)
  - Same variance (if not (known) use Welch's t test)
  - Samples independent (if not used paired t test)
- Test statistic:

$$\text{t}-\text{statistic} = \frac{\text{difference between means}}{\text{standard error of the difference of means}} = \frac{\bar{y}_A - \bar{y}_B}{s.e.d.m.}$$

- If this t-statistic > critical t, `qt(0.975,df=?)`, we reject the null hypothesis

s.e.d.m.

- The variance of differences between samples
  - Generally

$$\text{variance of difference} = \text{var(A)} + \text{var(B)} - 2\ \text{covariance(A,B)}$$

  - If the samples are independent, the covariance is zero, in other words, there is no correlation between samples, and we can simplify

$$\text{variance of difference} = \text{var(A)} + \text{var(B)} = s_A^2 + s_B^2$$

  - Going from variance of difference to standard error of difference (of the mean)

$$s.e.d.m = \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}$$

  - Even simpler, if the size and the variance of the two samples is the same, we can write (this is what we used for calculating the power of a one sample test in the experimental design session)

$$s.e.d.m = \sqrt{\frac{2s^2}{n}}$$

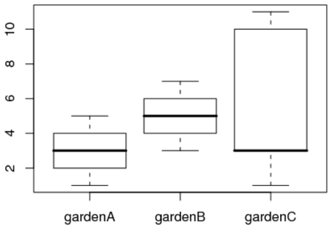
Example data

- Taken from Crawley's book website:
  - Ozone concentrations measured in gardensA, gardensB, gardensC at various times
  - We are going to use these gardens for a variety of tests
  - Download from WebCT and read text file into R

```
> g <- read.table("gardens.txt",header=TRUE)
> attach(g)
> ga <- gardenA
> gb <- gardenB
```

Example data

```
> g
   gardenA gardenB gardenC
1        3       5       3
2        4       5       3
3        4       6       2
4        3       7       1
5        2       4      10
6        3       4       4
7        1       3       3
8        3       5      11
9        5       6       3
10       2       5      10
> boxplot(g)
```



t-test by hand

- t-test for difference of means between gardenA and gardenB by hand, we just need t-statistic and critical t
  - t-statistic
    - difference between means

```
> d <- mean(ga) - mean(gb)
[1] -2
```

    - s.e.d.m.

```
> sedm <- sqrt( var(ga)/length(ga) +
  var(gb)/length(gb) )
[1] 0.5163978
```

    - t-statistic

```
> t <- d/sedm
[1] -3.872983
```

$$\text{t} - \text{statistic} = \frac{\bar{y}_A - \bar{y}_B}{s.e.d.m.}$$

$$s.e.d.m. = \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}$$

t-test by hand

- critical t: use qt(), the quantile function for the t distribution
  - degrees of freedom

```
> df <- length(ga)-1 + length(gb)-1
[1] 18
```

  - Type I error rate α=0.05, hence for two-tailed test we split 0.05 into 0.025 on both sides and use qt(0.025,df=18)

```
> ct <- qt(0.025,df)
[1] -2.100922
```

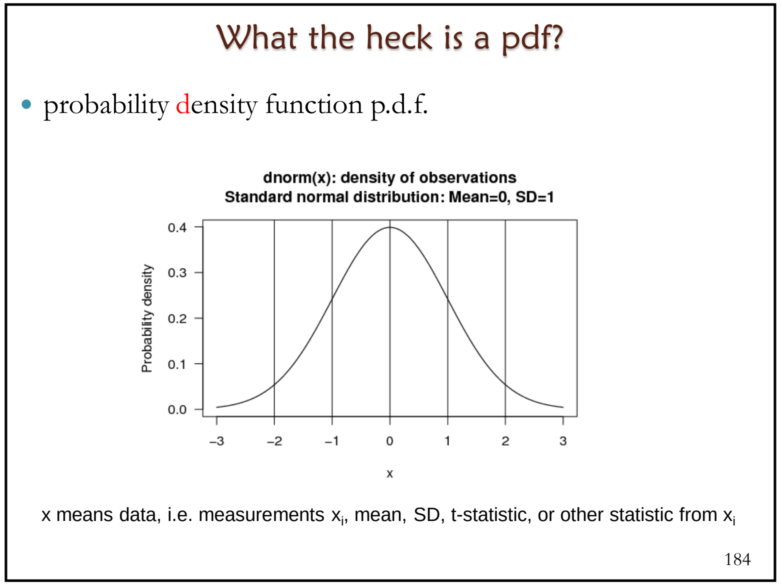
- compare t-statistic with critical t (only absolute values matter)
  - critical t is 2.100922
  - t-statistic is 3.872983
- Result: test statistic is larger than critical value so H_0 is rejected, the means are significantly different

t-test by hand

- p-value from the pt() function
  - probability that these data (or more extreme data) are observed if the null hypothesis were true (if means were the same)
    - data here are differences between means
  - use cumulative probability function pt(), which gives the summed probabilities of observations up to our value of the t-statistic which we have stored in the variable t

```
> p <- 2*pt(t,df) # note t < 0
> 2*(1-pt(abs(t),df)) # the same but more general
> p
[1] 0.001114539
```

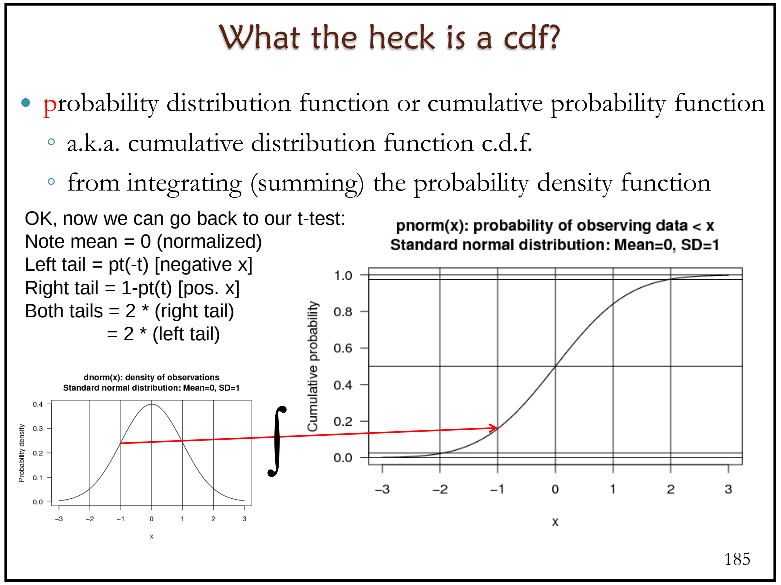
  - So it is very unlikely that our data can be explained by the null hypothesis of equal means (in other words, that our data have occurred by chance if H_0 were correct)



What the heck is a pdf?

- probability density function p.d.f.

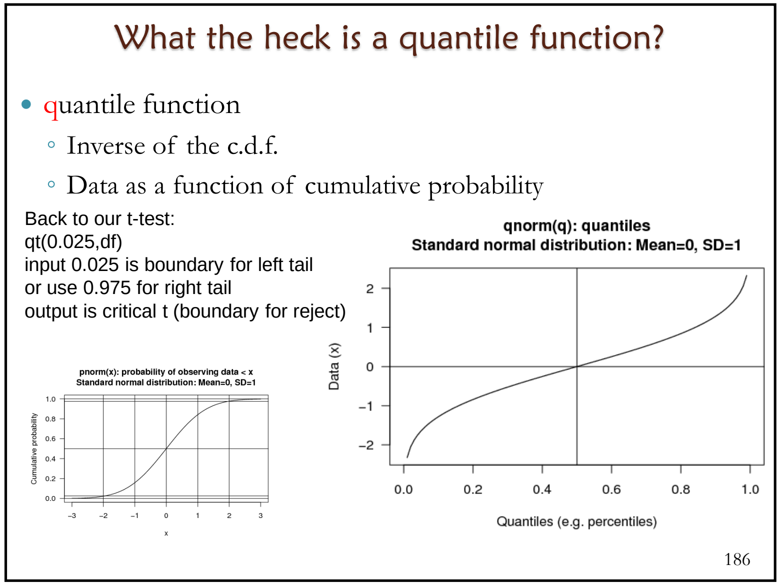
x means data, i.e. measurements x_i, mean, SD, t-statistic, or other statistic from x_i



What the heck is a cdf?

- probability distribution function or cumulative probability function
  - a.k.a. cumulative distribution function c.d.f.
  - from integrating (summing) the probability density function

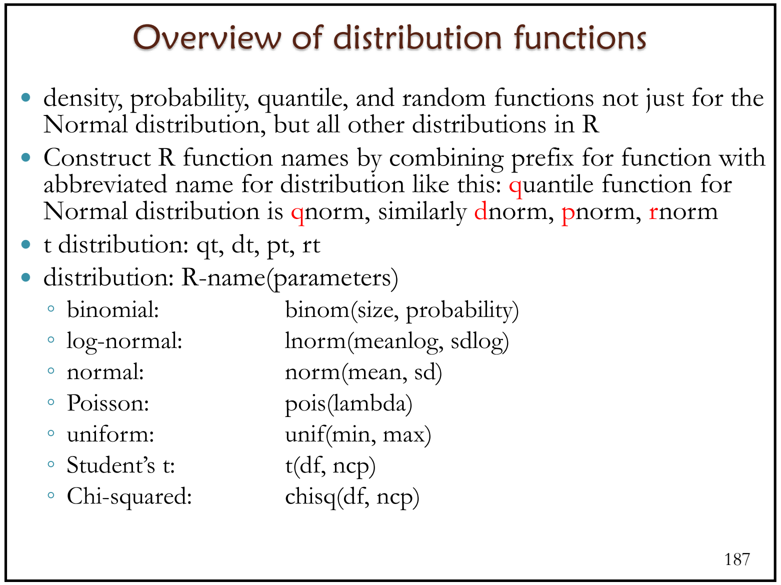
OK, now we can go back to our t-test:
Note mean = 0 (normalized)
Left tail = pt(-t) [negative x]
Right tail = 1-pt(t) [pos. x]
Both tails = 2 * (right tail)
= 2 * (left tail)



What the heck is a quantile function?

- quantile function
  - Inverse of the c.d.f.
  - Data as a function of cumulative probability

Back to our t-test:
qt(0.025,df)
input 0.025 is boundary for left tail
or use 0.975 for right tail
output is critical t (boundary for reject)



Overview of distribution functions

- density, probability, quantile, and random functions not just for the Normal distribution, but all other distributions in R
- Construct R function names by combining prefix for function with abbreviated name for distribution like this: quantile function for Normal distribution is qnorm, similarly dnorm, pnorm, rnorm
- t distribution: qt, dt, pt, rt
- distribution: R-name(parameters)

binomial:	binom(size, probability)
log-normal:	lnorm(meanlog, sdlog)
normal:	norm(mean, sd)
Poisson:	pois(lambda)
uniform:	unif(min, max)
Student’s t:	t(df, ncp)
Chi-squared:	chisq(df, ncp)

t-test by auto pilot

- Now the same t-test for difference of means between gardenA and gardenB using the ready-made function t.test()

```
> attach(g)
> t.test(gardenA,gardenB)
        Welch Two Sample t-test
data:  gardenA and gardenB
t = -3.873, df = 18, p-value = 0.001115
alternative hypothesis: true difference in means is not
  equal to 0
95 percent confidence interval:
 -3.0849115 -0.9150885
sample estimates:
mean of x mean of y
        3         5
```

t-test zoo

- One sample t-test is used to compare the mean from your single sample x with some well-known mean *w*
  - This well-known mean could be from another much larger study, or a true population mean
    - Comparing mean beer consumption of 20-30 year olds with UK average beer consumption (from brewery trade information)
  - Or this well-known mean could be a theoretical value
    - Comparing the speed of light in some substance with the speed of light in vacuum (*w* = 299,792,458 metres per second)

```
> t.test(x, mu=w)
```

t-test zoo

- One-tailed versus two-tailed
  - Use the two-tailed t-test if unsure

```
> t.test(..., alternative=“two.sided”)
```

    - This is the default option
  - Use the one-tailed t-test if you have good reasons for assuming that the mean of sample B must be higher (or lower)

```
> t.test(..., alternative=“less”) # meanA – meanB < 0
> t.test(..., alternative=“greater”)#meanA – meanB > 0
```

    - For example you know that the speed of light in any material is lower than the speed of light in vacuum
  - In our gardening example, there is no *a priori* reason why the ozone concentration in one garden should be higher than in the other garden

t-test zoo

- Paired t-test
  - If there is some connection (covariance or correlation) between your samples, you can use the more powerful paired t-test (as always, more power means better at rejecting false H_0)
  - Because positive covariance reduces the variance of the difference

$$\text{variance of difference} = \text{var}(A) + \text{var}(B) - 2\,\text{covariance}(A,B)$$

    - Measuring the same subject before and after treatment
    - Measuring biodiversity upstream and downstream a wastewater treatment plant inflow
    - Measuring oxygen in the same lake at night and during the day

```
> t.test(..., paired=TRUE) # default is paired=FALSE
```

  - There is no harm in using a paired t-test if there is no correlation between samples, so if unsure use the paired t-test
  - Paired t-test is actually identical to one-sample t-test using the differences between paired observations!

t-test zoo

- Welch's t-test
  - A generalization of the t-test
  - Use if variances are not the same
  - If unsure, use the Welch version of the t-test
  - It's actually the default method when using R's t.test() function
  - Works by correcting the degrees of freedom to account for differences in the variance (so d.f. can be non-integer)

t-test zoo

- All together now
  - Here are snippets from the help file which you get by typing `?t.test`

    Usage

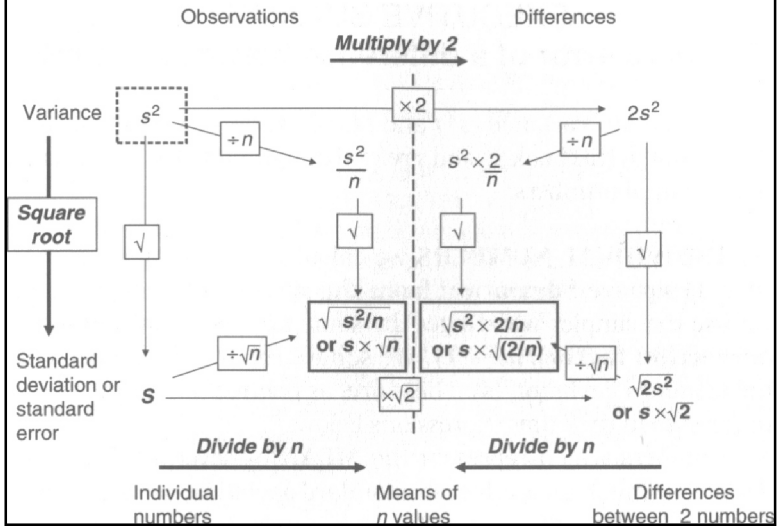
```
## Default S3 method:
t.test(x, y = NULL, alternative = c("two.sided",
  "less", "greater"), mu = 0, paired = FALSE,
  var.equal = FALSE, conf.level = 0.95, ...)
```

`var.equal` a logical variable indicating whether to treat the two variances as being equal. If TRUE then the pooled variance is used to estimate the variance otherwise the Welch (or Satterthwaite) approximation to the degrees of freedom is used.

Exercise

- Test whether the means of two samples are significantly different using an appropriate test
  - Download the text file stream.txt from WebCT and read into R
  - This contains two samples from the same river (upstream, downstream)

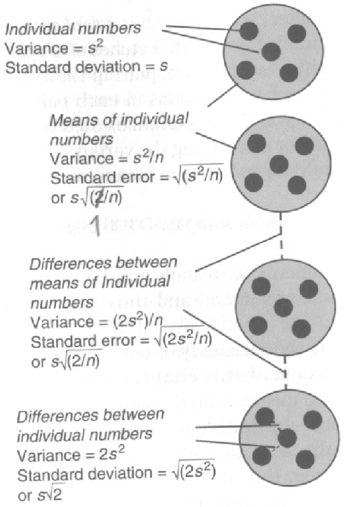
Overview of the relationships of all things variance (simplified)


Observations
Differences
Multiply by 2
Variance
s^2
×2
$2s^2$
÷n
÷n
$\frac{s^2}{n}$
$s^2 \times \frac{2}{n}$
Square root
√
√
√
√
$\sqrt{s^2/n}$ or $s \times \sqrt{n}$
$\sqrt{s^2 \times 2/n}$ or $s \times \sqrt{(2/n)}$
÷√n
÷√n
Standard deviation or standard error
s
×√2
$\sqrt{2s^2}$ or $s \times \sqrt{2}$
Divide by n
Divide by n
Individual numbers
Means of n values
Differences between 2 numbers

Visual explanation of these relationships (simplified)

These two summaries from:

Helmut Fritz van Emden (2008)
Statistics for terrified biologists
Blackwell Publishing: Oxford



Individual numbers
Variance = s^2
Standard deviation = s

Means of individual numbers
Variance = s^2/n
Standard error = $\sqrt{(s^2/n)}$
or $s\sqrt{(\cancel{2}/n)}$
1

Differences between means of Individual numbers
Variance = $(2s^2)/n$
Standard error = $\sqrt{(2s^2/n)}$
or $s\sqrt{(2/n)}$

Differences between individual numbers
Variance = $2s^2$
Standard deviation = $\sqrt{(2s^2)}$
or $s\sqrt{2}$