BIOC461

Research Design and Analysis in Biochemistry

Lecture 1

A/Prof Mik Black
Department of Biochemistry

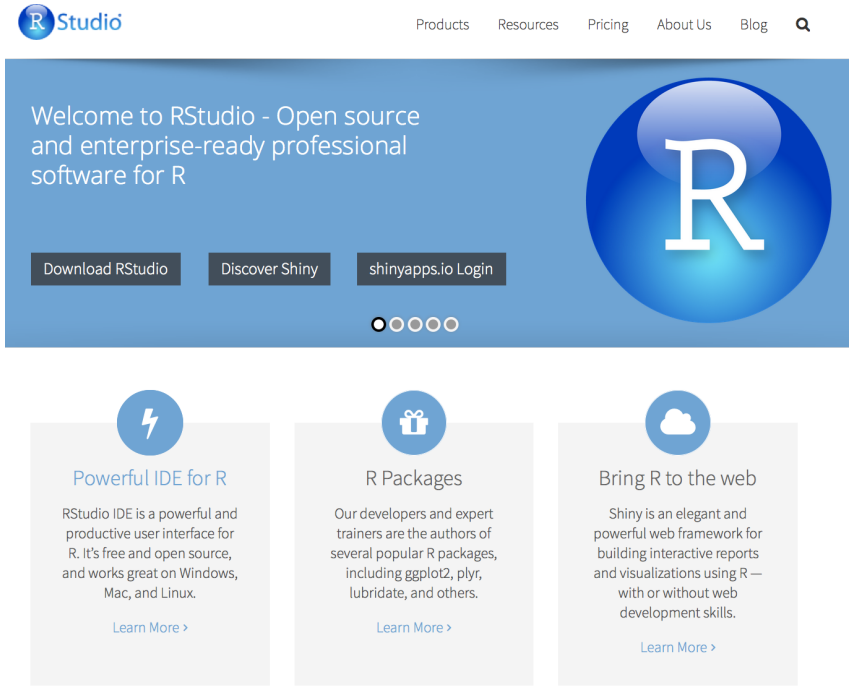
23 March 2017

What's so great about statistics?

- In science we spend much of our time generating experimental data.
- We use this data both to answer our scientific questions, and to formulate new hypotheses.
- Rigorous analysis and interpretation of data is a vital component of science.
- Statistics provides a collection of tools to help us answer questions

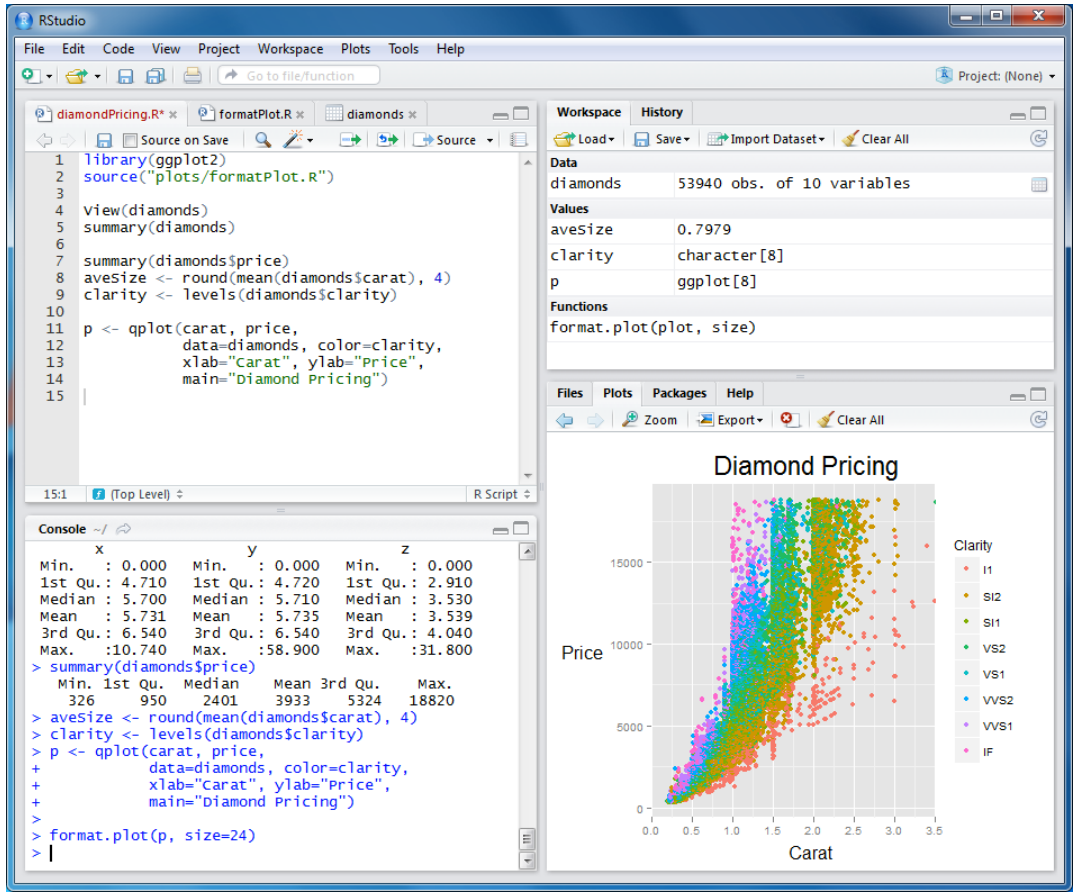
Introducing RStudio

The free RStudio application provides an "integrated development environment" (IDE) for working in R.



https://www.rstudio.com/

What does it look like?



Getting started with R

- Open RStudio and spend some time getting familiar with the layout:
  - Console: type commands here
  - Source: save your code here
  - Environment/History: session info
  - Files/Plots/Packages/Help/Viewer: everything else :)
- You can write commands/code in the source pane, and then "submit" it (i.e., run it) in the console by hitting `Ctrl+Enter` (Windows) or `Cmd+Enter` (Mac)
- Comments start with # and are ignored by R

Getting started with R

```
# This is a comment

# This adds 3 + 5
3 + 5
```

```
## [1] 8
```

```
# This adds 3 + 2 and assigns it to an object called "a"
a = 3 + 2

# This shows the value of "a"
a
```

```
## [1] 5
```

Getting started with R

```
# This also adds 3 + 2 and assigns it to an object called "a"
a <- 3 + 5

# This takes the object "a", and adds 10 to it
a + 10
```

```
## [1] 18
```

```
# This calculates the square root of "a"
# sqrt is called a "function" - we'll use functions a lot
sqrt(a)
```

```
## [1] 2.828427
```

Lets do something useful

- While it is fun to use R as a giant calculator, that's not going to help (much) with your research.
- We want to do some data analysis in R
  - First, we need data
  - Good, clean, well-organised data

Data organization

- A spreadsheet application (e.g., Microsoft Excel) is often used to “organize” data.
  - Excel has the ability to make raw data look very pretty.
  - Statistics applications (e.g., SPSS, R, SAS, Prism) often have trouble reading in the pretty data.
- Try to keep it simple:
  - Samples in rows, variables in columns
  - No blank cells or rows (unless data are missing)

Example data:

Animal	Gender	Diet	WeightGain_grams
Mouse1	Male	HighFat	100
Mouse2	Female	HighFat	145
Mouse3	Female	HighFat	135
Mouse4	Male	LowFat	104
Mouse5	Female	LowFat	120
Mouse6	Male	LowFat	116
Mouse7	Male	LowFat	109

Note the lack of units in the Weight Gain column – stats programs expect numeric values for continuous data.

Loading the data

```
## Load the data into R from a CSV file
mouseData <- read.csv('DataSets/mouseExample.csv')
```

```
## what sort of data object is it?
class(mouseData)
```

```
## [1] "data.frame"
```

```
## What are its dimensions?
dim(mouseData)
```

```
## [1] 7 4
```

Variable names

```
## List the variables names in the dataframe
names(mouseData)
```

```
## [1] "Animal"           "Gender"           "Diet"
## [4] "WeightGain_grams"
```

Inspecting the data

```
## Typing the name of the object displays its contents
mouseData
```

```
##   Animal Gender    Diet WeightGain_grams
## 1 Mouse1   Male HighFat              100
## 2 Mouse2 Female HighFat              145
## 3 Mouse3 Female HighFat              135
## 4 Mouse4   Male  LowFat              104
## 5 Mouse5 Female  LowFat              120
## 6 Mouse6   Male  LowFat              116
## 7 Mouse7   Male  LowFat              109
```

```
## Open the built-in RStudio viewer:
View(mouseData)
```

Accessing individual variables

```
## The $ symbol allows you to reference the columns of a dataframe.
## The following produces a summary of the data in the
## "WeightGain_grams" column, and the calculates the standard
## deviation.
summary(mouseData$WeightGain_grams)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   100.0   106.5   116.0   118.4   127.5   145.0
```

```
sd(mouseData$WeightGain_grams)
```

```
## [1] 16.46063
```

Accessing individual variables

```
## You can also just use the column number
summary(mouseData[,4])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   100.0   106.5   116.0   118.4   127.5   145.0
```

```
## Or the name of the column
summary(mouseData[,"WeightGain_grams"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   100.0   106.5   116.0   118.4   127.5   145.0
```

Summarizing data

- Frequency Distribution: display the number of times each value occurs in the data set
- Example (for discrete quantitative data)
  - Size of litters for n=36 sows
  - Litter size (number of piglets) is an integer (discrete)

```
pigs<-read.csv('DataSets/pigs.csv')
head(pigs)
```

```
##   Weight
## 1     10
## 2     12
## 3     10
## 4      7
## 5     14
## 6     11
```

Frequency tables and histograms

- Summarize with a frequency distribution (as a table):

```
## Since there is only one variable, can just specificy the
## "pigs" data.frame
table(pigs)
```

```
## pigs
##  5  7  8  9 10 11 12 13 14
##  1  2  3  3  9  8  5  3  2
```

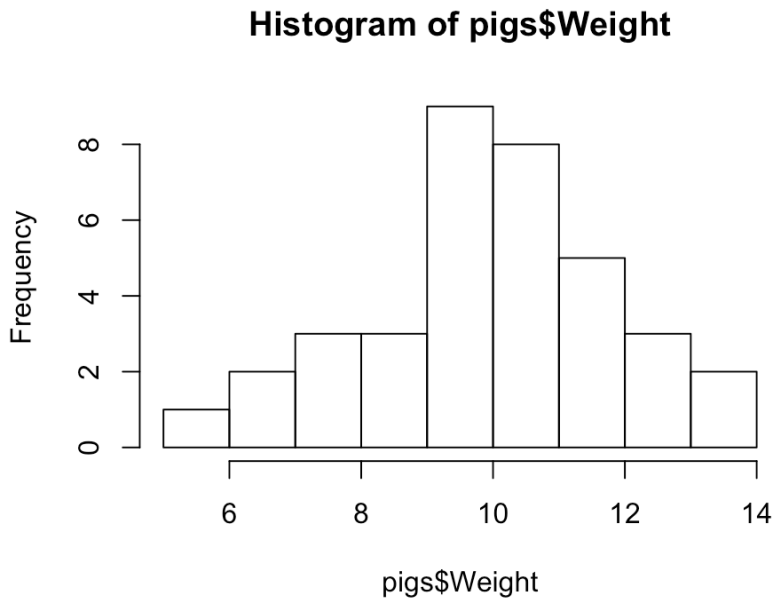
```
## Better practice to specify the variable you are interested in
table(pigs$Weight)
```

```
##
##  5  7  8  9 10 11 12 13 14
##  1  2  3  3  9  8  5  3  2
```

Frequency tables and histograms

- Generate a frequency histogram:

```
hist(pigs$Weight)
```



Example: heroin data

- A pathologist measures the amount of free morphine concentrations (ng/mL) in blood taken from 25 “acute” heroin overdose victims. The data are listed below:

```
 67  110   45   96   80  103  223   86   58
112   52  165   91  189   85  150  120  293
153  390   47  158   77  216  526
```

- Source: B Levine, D Green–Johnson, KA Moore, D Fowler, A Jenkins, (2001). Assessment of the acuteness of heroin deaths from the analysis of multiple blood specimens. Science and Justice 42(1). 17-20.

Heroin data (ordered)

```
heroin<-read.csv('DataSets/morphine.csv')
sort(heroin$Morphine)
```

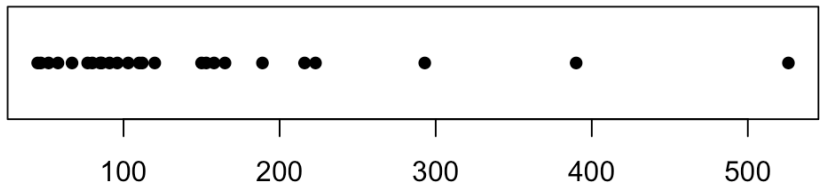
```
##  [1]  45  47  52  58  67  77  80  85  86  91  96
## [12] 103 110 112 120 150 153 158 165 189 216 223
## [23] 293 390 526
```

- Easy to see minimum and maximum
- What is a typical value?
- How variable are the data?
- Any clusters in the data?
- Any outliers?

Data visualization - dotplots

- A dotplot provides a simple 1-D representation of a quantitative variable.
  - Each dot represents one observation.
  - Provides instant visual summary of the data.

```
stripchart(heroin$Morphine, pch=16, vertical=FALSE)
```



Describing distributions

- The shape of a histogram can be described by a smooth curve that roughly follows the tops of the bars.
- This can be done by eye or by numerical algorithms on the computer.
- The shape often follows certain patterns:
  - symmetric (same right to left)
  - bell-shaped
  - skewed to the right or left
  - unimodal (one major peak)
  - bimodal (two major peaks)
  - spread (wide or narrow)

Summarizing quantitative data

- Describe the distribution of a collection of data
- Measures of centre or location
- Means or medians
- Measures of spread or scale
- Standard deviations, variances, ranges, and interquartile ranges

Measuring the centre

- Average or sample mean
- The mean is *"the sum of all the observations divided by the number of observations"*
- In R:

```
mean(heroin$Morphine)
```

```
## [1] 147.68
```

Measuring the centre

- The sample median is the middle value when the observations are ordered.
- Rule
  - If you have an odd number of observations, then the median is the middle value.
  - If you have an even number, the median is the average of the two middle observations.
- In R:

```
median(heroin$Morphine)
```

```
## [1] 110
```

Which one should we use?

- Mean
  - The “balance point” of the histogram
  - Affected by outliers (e.g., heroin data)
  - Most commonly used in testing and estimation
- Median
  - Divides the area of the histogram in half
  - Unaffected by a few outliers
- If the histogram is fairly symmetric, the mean and median will be similar.

Five number summary

- When the data are ordered
  - LQ is ¼ of the way in from the bottom
  - UQ is ¼ of the way from the top

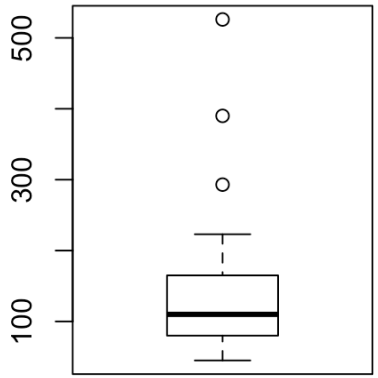
```
summary(heroin$Morphine)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    45.0    80.0   110.0   147.7   165.0   526.0
```

Visualization: boxplots

```
boxplot(heroin$Morphine)
```

calculation of “whisker length”.



- Boxplots provide another method for summarizing a quantitative variable.

Measures of Dispersion

- There are several ways to describe the variability or spread of a data set:
  - Range
  - Interquartile range
  - Standard deviation
  - Variance
- Standard deviation is the most common.

Ranges

- The Sample Range: difference between the largest and smallest observation.
  - Advantage: easy to calculate.
  - Disadvantages: highly sensitive; not useful for testing
- The Interquartile Range (IQR): difference between the first and third quartiles: IQR = Q3 – Q1.

```
range(heroin$Morphine)
```

```
## [1]  45 526
```

```
IQR(heroin$Morphine)
```

```
## [1] 85
```

Sample Standard Deviation

- Measures how far away, on average, values are from the center
- Same units as the data
- s=0 only when there is no spread, otherwise s>0.

```
sd(heroin$Morphine)
```

```
## [1] 113.3735
```

The Sample Variance

- The sample variance is just the square of the sample standard deviation. Its symbol is s^2.
- "Average squared distance from the mean."
- Units are the square of the data units.

```
var(heroin$Morphine)
```

```
## [1] 12853.56
```

```
sd(heroin$Morphine)^2
```

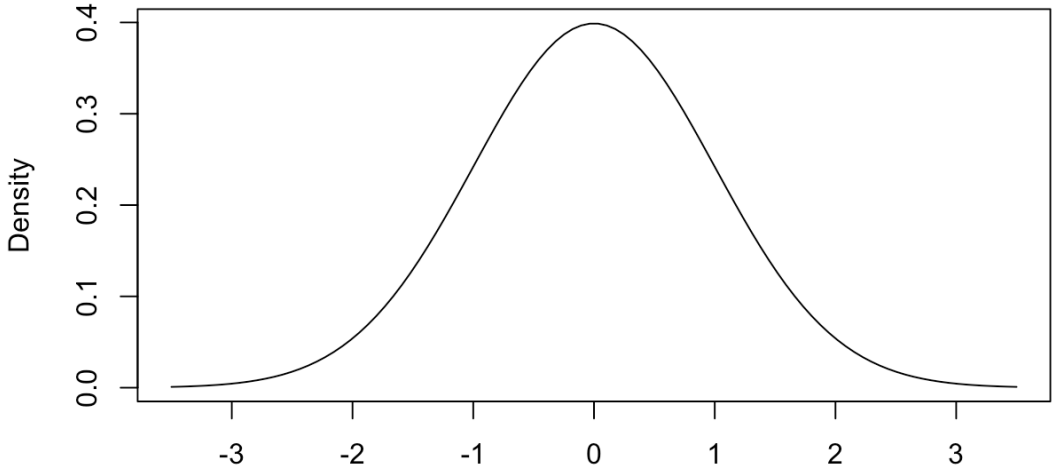
```
## [1] 12853.56
```

Summary - describing “spread”

- Range = maximum – minimum
- Not sensible – affected severely by outliers
- Interquartile range = UQ – LQ
- Describes the central half (50%) of the data
- Sample standard deviation (s.d.) = “square root of the averaged squared distance to the mean”

The Normal Distribution

- Describes a wide variety of data very well
- Good model for some continuous data
- Bell-shaped, symmetric distribution



The Normal Distribution

- It can be proved mathematically that any random variable that is the sum or average of many small independent effects will have a distribution of values that is close to normal.
- Examples:
  - measurement error
  - test scores
  - height, yield
  - body temperatures
  - levels in blood samples (e.g. cholesterol)

Normal curves

- “Normal” distribution refers to a family of curves.
- Characterized by the (population) mean, μ and the (population) SD σ.
- Roughly 68% of distribution lies between -σ and +σ, 95% between -2σ and +2σ (actually 1.96) between -3σ and +3σ.

Sampling Distributions

- Start with population with a certain distribution (e.g., normal, but doesn't have to be).
- True mean μ and SD σ (parameters describing the distribution)
- Take sample of size n from normal population
  - Get sample $y_1, \ldots, y_n$
  - Calculate mean, $\bar{y}$ and sample SD s (estimates for parameters).

Sampling Distributions (continued)

- May not be exactly μ and σ, but should be close, especially if n is large.
  - how far away can we expect them to be by chance?
  - what if we drew a different sample?
  - wouldn't get same $\bar{y}$ and s, since sample would be different
- Called sampling variability
  - How does taking different samples affect our estimates of μ, and σ?

Meta-experiment

- Think about repeating experiment many times (taking one sample of n is one repetition).
- Think of the probability distribution of all the possible outcomes for the estimates $\bar{y}$, s.
- This is called a sampling distribution for the parameter estimates (here $\bar{y}$, s).
- Usually can't actually generate the sampling distribution: can only do sampling once
- Sampling distribution represents a conceptual *meta-experiment*.

Sampling Distribution: Normal Mean

- Sample of size n from a normally distributed population
- Observe sample mean $\bar{y}$.
- How close is this likely to be to μ?
- Meta experiment: Look at the *sampling distribution* for $\bar{y}$.
  - New "sample size" is number of times we repeat experiment
  - New observations are the $\bar{y}$'s.
  - Draw a histogram of those observed $\bar{y}$'s.

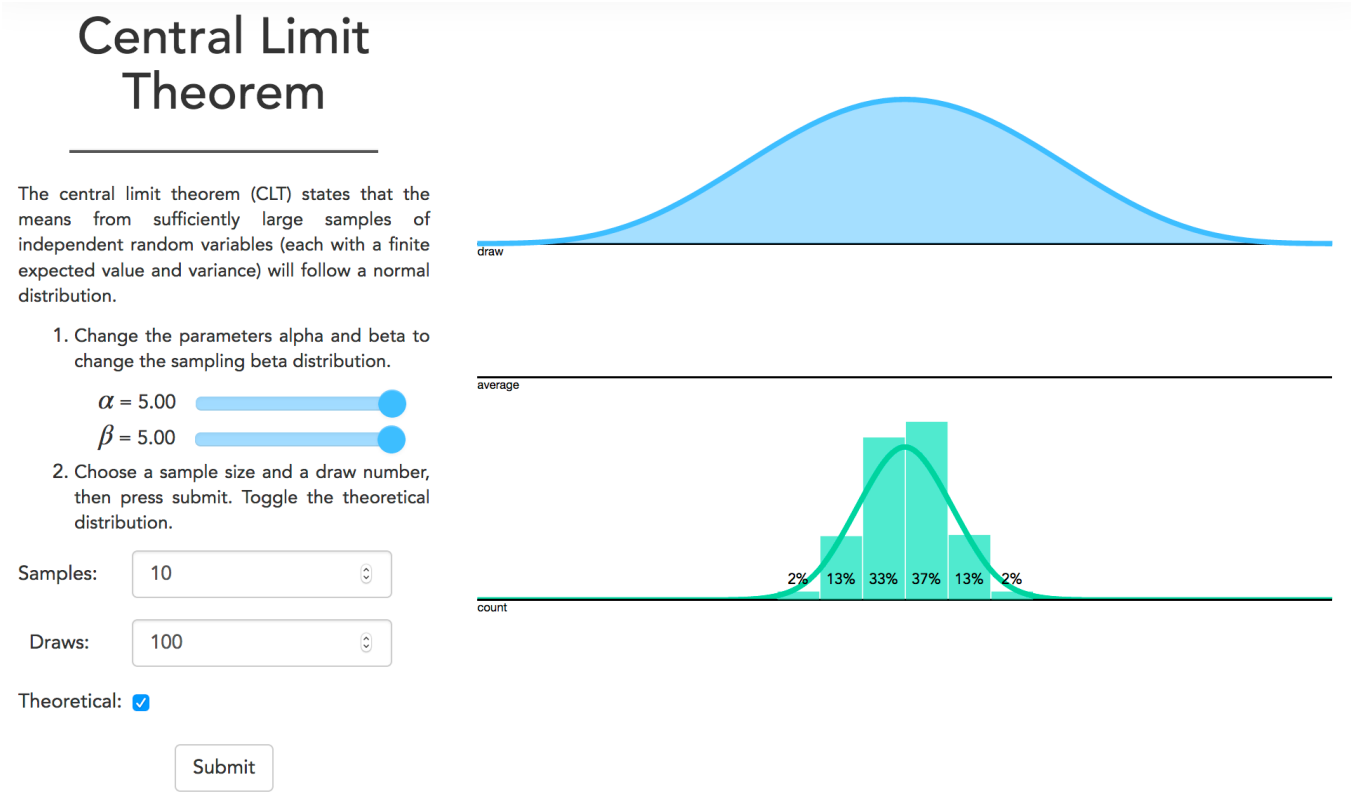
Visualising the "meta-experiment"

- Go to the website below, click on "Units" and select "Central Limit Theorem" under "Distributions":

  http://students.brown.edu/seeing-theory/index.html

- Set the parameters to $\alpha = 5$ and $\beta = 5$, with a sample size of 10 (technically noyt a normal distribution, but close enough for our purposes).
- Generate draws from this distribution, and observe the shape of the resulting distribution of the means (each "draw" generates 10 obervations, and takes the mean: the distribution of means is plotted in the lower graph).
- The shape should be suspiciously familiar…

Visualising the "meta-experiment"



http://students.brown.edu/seeing-theory/index.html

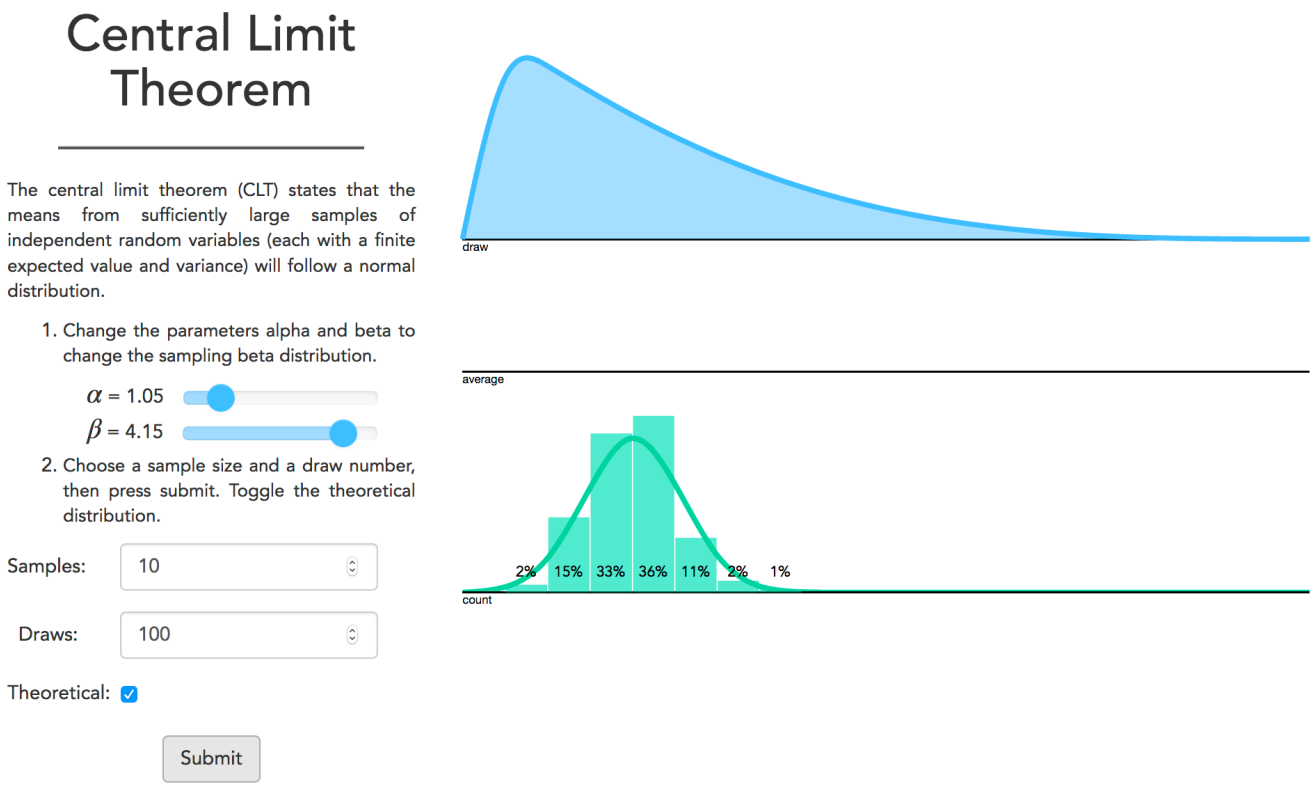
Sampling Distribution: Normal Mean

- *What does it look like?*
  - it looks like a normal curve
  - it has same mean as the original distribution: μ.
  - it has less variability: the SD is smaller.
  - This is because some of the "noise" in the individual observations is smoothed out in the $\bar{y}$'s.
  - new distribution has SD = $\sigma/\sqrt{n}$.

Central Limit Theorem

- Even if Y doesn't have a normal distribution, the sampling distribution of $\bar{y}$ will be close to normal, when n is large.
  - It also has mean μ and SD equal to $\sigma/\sqrt{n}$, i.e., it is approximately normal N(μ, $\sigma/\sqrt{n}$)
- What does "large" mean in this theorem?
- It depends how close to normal the distribution of Y is.
  - If Y is normally distributed, then n=1 is big enough.
  - If Y is not normally distributed, but is fairly symmetric and fairly bell-shaped, then n doesn't have to be too large (e.g., n=10).
  - If the distribution of Y is really weird-looking, then n may have to be really big.

Visualising the Central Limit Theorem



http://students.brown.edu/seeing-theory/index.html

Rationale for SD=$\sigma/\sqrt{n}$

- If n is large, then $\bar{y}$ will be a pretty good estimate of μ, so there should be less variability in its SD, but if n is small then $\bar{y}$ will be more variable.
- The actual square root order of the correction is a probabilistic effect related to the arithmetic of adding variances.
- The CLT applies to the above discussion on estimating means. So, even if data are not strictly normal, this method gives a good approximation to the distribution for the mean, $\bar{y}$.

Estimation

- Usually do not know the true μ and σ.
- Want to estimate μ with $\bar{y}$ and also have some idea how good our estimate is.
- Want to make an interval around $\bar{y}$ that we are, say, 95% certain contains the true value of μ.
  - called a 95% confidence interval
- In general we want to find a "1-α" confidence interval
  - For a 95% CI, have = 0.05; for a 90% CI have = 0.10, for a 99% CI have = 0.01, etc

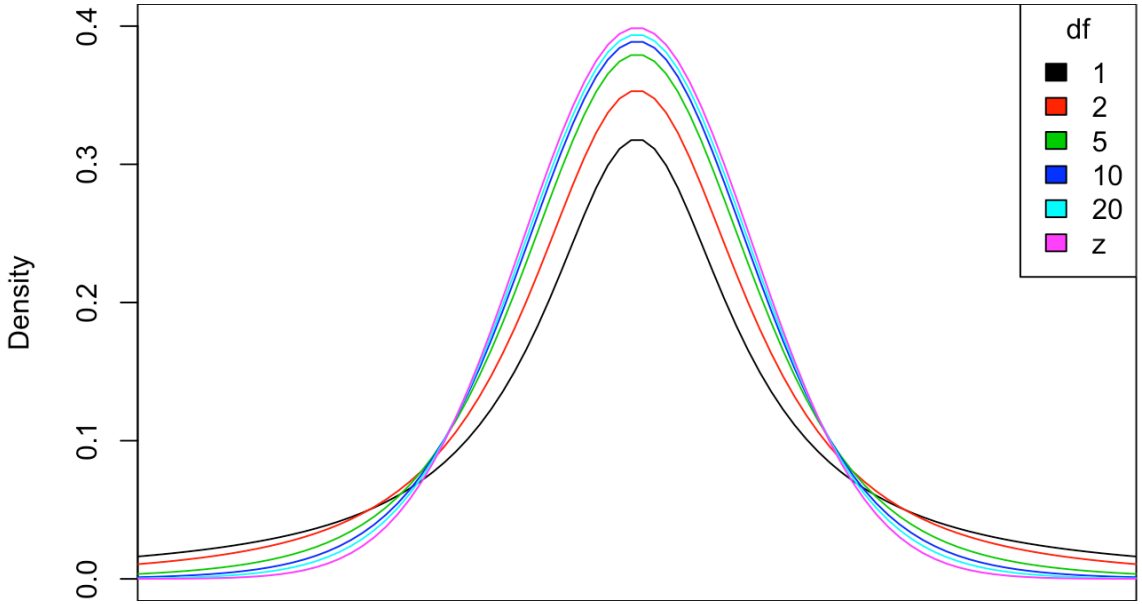
Estimation (continued)

- We do the best we can: must estimate σ with s.
- Define the *standard error of the mean* as SE = $s/\sqrt{n}$.
- SE is an estimate of the true standard deviation of the mean, which is $\sigma/\sqrt{n}$.
  - Will use the estimate SE in place of the true value .
- There is a price to pay for not using the true value: can't use normal distribution percentiles.
  - Using estimate for introduces extra uncertainty.
  - Makes the confidence interval wider.

The t distribution

- Use a distribution with a little more variability than the normal distribution: the *t-distribution*
  - a family of probability distributions for continuous variables
  - shape similar to the standard normal N(0,1) but kind of squashed so that it has a wider spread
  - the amount of squashing is determined by its "degrees of freedom" (df).
  - always centered at mean 0
- Call the random variable T and the observed values t.

The t distribution



- At df=1 the t distribution has "heavy tails".
- Tails "lighten" as df increases (mass moves to centre).
- As df$\rightarrow \infty$, the t-distribution converges to the normal (z).

t distribution for confidence intervals

- Because we are using SE instead of $\sigma/\sqrt{n}$, we have more uncertainty in our estimate for μ
- Use t-distribution percentiles instead of 1.96 (bigger)
- The better our estimates are (the bigger n is), the less variability we want to introduce.
- The t-distribution's df controls the amount of variability
- In this situation (estimating for one sample), df = n-1
- t-distribution will tell us what number to use instead of 1.96 to construct a confidence interval for μ.

Example: confidence interval

- Suppose we have the following data on daily calcium intake (mg) from a study of women's bone health:

```
calcium<-read.csv('DataSets/calcium.csv')
head(calcium)
```

```
##   ID Calcium
## 1  1     808
## 2  2     882
## 3  3    1062
## 4  4     970
## 5  5     909
## 6  6     802
```

- What is our estimate of the average daily calcium intake?

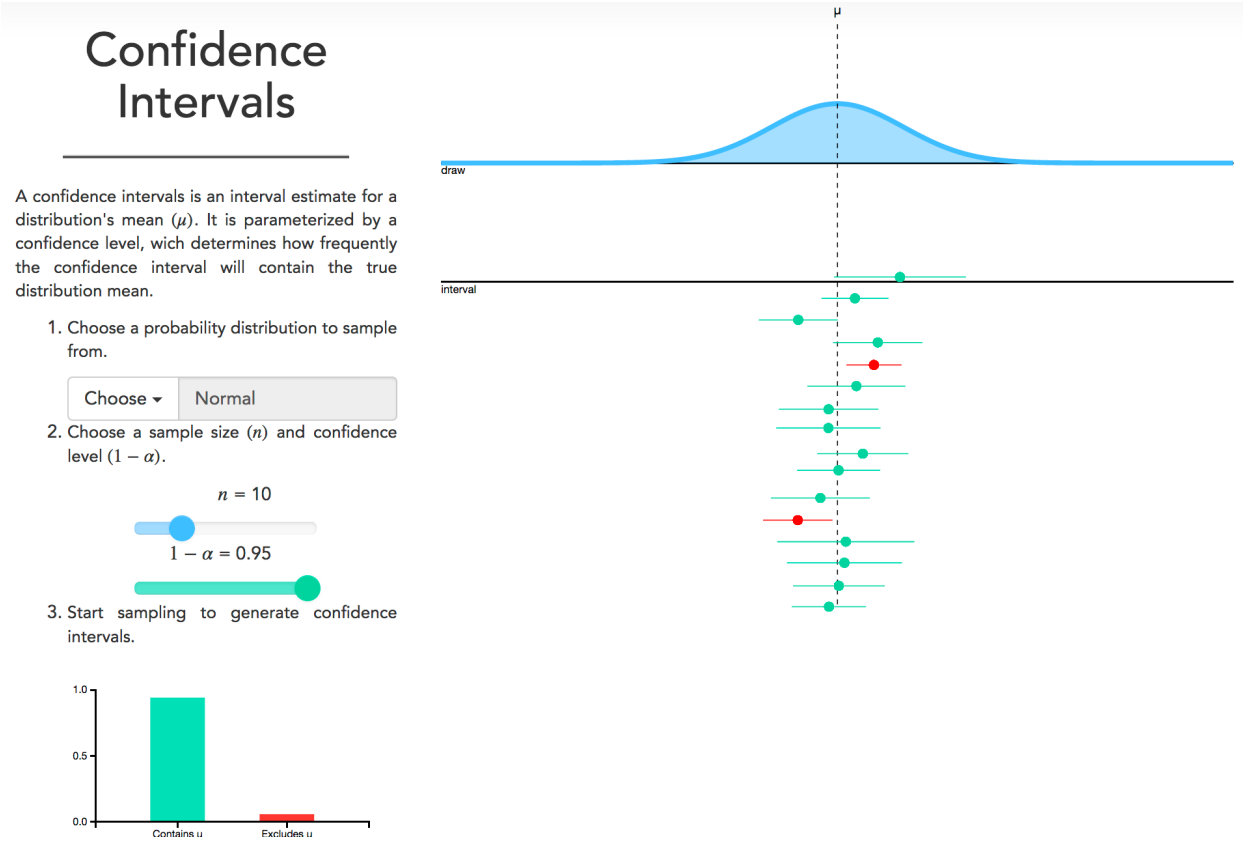
Example: confidence interval

```
t.test(calcium$Calcium)
```

```
##
##  One Sample t-test
##
## data:  calcium$Calcium
## t = 13.361, df = 37, p-value = 9.895e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##   785.5995 1066.4532
## sample estimates:
## mean of x
##  926.0263
```

Estimate: 926mg, 95% CI: (786mg, 1066mg)

Visualising Confidence Intervals



http://students.brown.edu/seeing-theory/index.html

Standard Error vs Standard Deviation

- Sample SD s:
  - measures amount of variability in data set
  - as n gets larger, SD approaches true SD ; does not tend to get smaller or larger;
  - used for making predictions about single data point.
- Sample SE = $s/\sqrt{n}$:
  - measures amount of uncertainty in estimate of mean
  - gets smaller as n gets larger (random sample)
  - used for making predictions about mean.

Summary: estimation of μ

- Estimate the population mean by the sample mean, $\bar{x}$.
  - There is still uncertainty about the true value of μ after we calculate $\bar{x}$.
  - construct a confidence interval for μ based on the observed data.
- A confidence interval is a statement about a population parameter based on the observed data, **NOT** a statement about the sample.
- The width of the confidence interval increases as:
  - the level of confidence increases (i.e., 99% *versus* 95%)
  - the sample size (i.e., n) decreases
  - the variability (i.e., s) increases

Two independent samples

- Sometimes we have two populations (groups) we want to compare, based on a continuous measurement.
- Examples:
  - cases (treatment) and controls
  - drug and placebo
  - patients taking two comparable drugs
  - males and females
  - two different genetic strains, or mutant/wildtype
- Population 1: $N(\mu_1, \sigma_1)$, sample $n_1, \bar{y}_1, s_1, SE_1 = s_1/\sqrt{n_1}$
- Population 2: $N(\mu_2, \sigma_2)$, sample $n_2, \bar{y}_2, s_2, SE_2 = s_2/\sqrt{n_2}$

Comparing μ_1 and μ_2

- Typical question: is μ_1 larger than μ_2?
  - What is the difference in the means of the two populations, $\mu_1 - \mu_2$?
  - We want to estimate $\mu_1 - \mu_2$ and get a Confidence Interval.
- What to do:
  - take random samples from each population, sizes n_1 and n_2.
  - calculate the mean from each sample
  - $\bar{y}_1 - \bar{y}_2$ is an estimate for $\mu_1 - \mu_2$
  - need a SE to find the width of a Confidence Interval

SE for difference between two means

- How to calculate SE for the estimate $\bar{y}_1 - \bar{y}_2$?
  - two ways: unpooled and pooled.
  - when $n_1 = n_2$ these are equivalent.
- Use Unpooled method most of the time.
- Use Pooled method when you think $\sigma_1 = \sigma_2$.
- Both are based on some combination of s_1 and s_2

Example: comparing two means

- Data from trial involving feeding day-old male chicks “high lysine corn”. Weight gain recorded (Cromwell, et al. Poultry Sci. 1968).

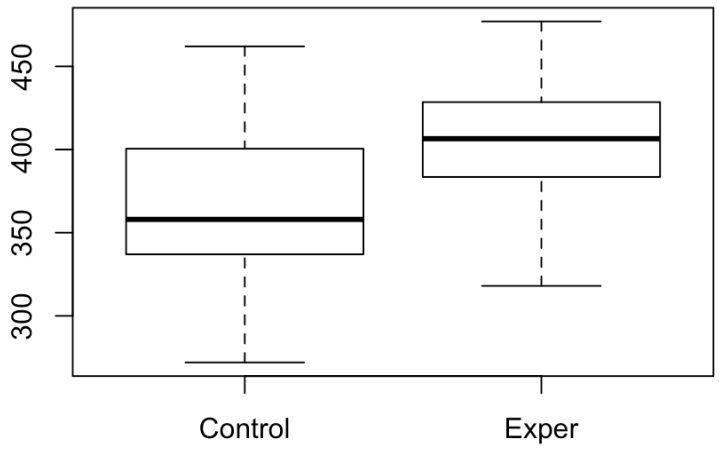
```
corn<-read.csv('DataSets/corn.csv')
head(corn)
```

```
##     Group Weight
## 1 Control    380
## 2 Control    321
## 3 Control    366
## 4 Control    356
## 5 Control    283
## 6 Control    349
```

- Is there a significance difference in weight gain between the groups?

Example: comparing two means

```
## Generate boxplots for each group
boxplot(Weight ~ Group, data=corn)
```



Comparing two means: 95% CI

```
## Confidence interval for difference between means
t.test(Weight ~ Group, data=corn)
```

```
##
##  Welch Two Sample t-test
##
## data:  Weight by Group
## t = -2.469, df = 36.915, p-value = 0.0183
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -66.729186  -6.570814
## sample estimates:
## mean in group Control   mean in group Exper
##                366.30                402.95
```

95% confidence interval for difference: (-66.73, -6.57)

Comparing two means: 99% CI

```
## Confidence interval for difference between means
t.test(Weight ~ Group, data=corn, conf.level = 0.99)
```

```
##
##  Welch Two Sample t-test
##
## data:  Weight by Group
## t = -2.469, df = 36.915, p-value = 0.0183
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  -76.962594   3.662594
## sample estimates:
## mean in group Control   mean in group Exper
##                366.30                402.95
```

99% confidence interval for difference: (-76.96, 3.66)

Interpretation of Confidence Intervals

- What does it mean to be 95% confident that the population mean is in a certain interval?
- If we were able to repeat the experiment many times, how often would we be right, on average?
- How often would the interval we calculated actually contain μ?
- How often would we miss μ entirely?

Interpretation of Confidence Intervals

- If we construct all our confidence intervals with this procedure, in the long run
  - we will be right 95% = $1 - \alpha$ of the time
  - we will be wrong 5% = α of the time
- α represents our "willingness" to be wrong
- In reality we cannot repeat the experiment, and we do not know μ.
- Therefore we NEVER KNOW for any one experiment whether we are right or wrong

Hypothesis testing: The t-test

- Can compare population means (e.g., μ_1 and μ_2) using confidence intervals.
- Often interested in whether means are the same ($\mu_1 = \mu_2$) or different ($\mu_1 \neq \mu_2$).
- Can also accomplish this comparison using Hypothesis Testing.

Null and alternative hypotheses

- Two populations, means μ_1 and μ_2.
  - Null hypothesis, $H_0 : \mu_1 = \mu_2$
  - Alternative hypothesis, $H_A : \mu_1 \neq \mu_2$
- Idea is to determine whether the data contains enough evidence to make H_0 seem unlikely.
- Example: does studying improve exam performance?
  - Two populations: the exam scores for those who didn't study (population mean 1) and the exam scores for those who did (population mean 2).
  - H_0: Studying has no effect on exam score ($\mu_1 = \mu_2$).
  - H_A: Studying has some effect on exam score ($\mu_1 \neq \mu_2$).

The t statistic

- Testing the null hypothesis:

$$H_0 : \mu_1 = \mu_2 \iff \mu_1 - \mu_2 = 0$$

- Against the alternative hypothesis

$$H_A : \mu_1 \neq \mu_2 \iff \mu_1 - \mu_2 \neq 0$$

- Use t statistic (assume H_0 is true):

$$t_s = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{SE(\bar{y}_1 - \bar{y}_2)} = \frac{(\bar{y}_1 - \bar{y}_2)}{SE(\bar{y}_1 - \bar{y}_2)}$$

Example – high lysine corn data

```
t.test(Weight ~ Group, data=corn)
```

```
##
##  Welch Two Sample t-test
##
## data:  Weight by Group
## t = -2.469, df = 36.915, p-value = 0.0183
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -66.729186  -6.570814
## sample estimates:
## mean in group Control   mean in group Exper
##                366.30                402.95
```

Example – high lysine corn data

```
aggregate(corn$Weight, list(corn$Group), mean)
```

```
##   Group.1      x
## 1 Control 366.30
## 2   Exper 402.95
```

```
aggregate(corn$Weight, list(corn$Group), sd)
```

```
##   Group.1        x
## 1 Control 50.80520
## 2   Exper 42.72861
```

$$t_s = \frac{(\bar{y}_1 - \bar{y}_2)}{SE(\bar{y}_1 - \bar{y}_2)} = \frac{366.30 - 402.95}{\sqrt{\frac{50.81^2}{20} + \frac{42.73^2}{20}}} = -2.469$$

The P-value

- The P-value of the test is the area of the t distribution in the “double tails”, beyond -t_s and +t_s.
  - Measure of compatibility between data and H_0.
- The P-value for a hypothesis test is the probability, computed under the condition that the null hypothesis is true, of the test statistic being at least as extreme as the value of the test statistic that was actually obtained.
- Basically we are asking: "what is the probability of observing these data, if the null hypothesis is true?"

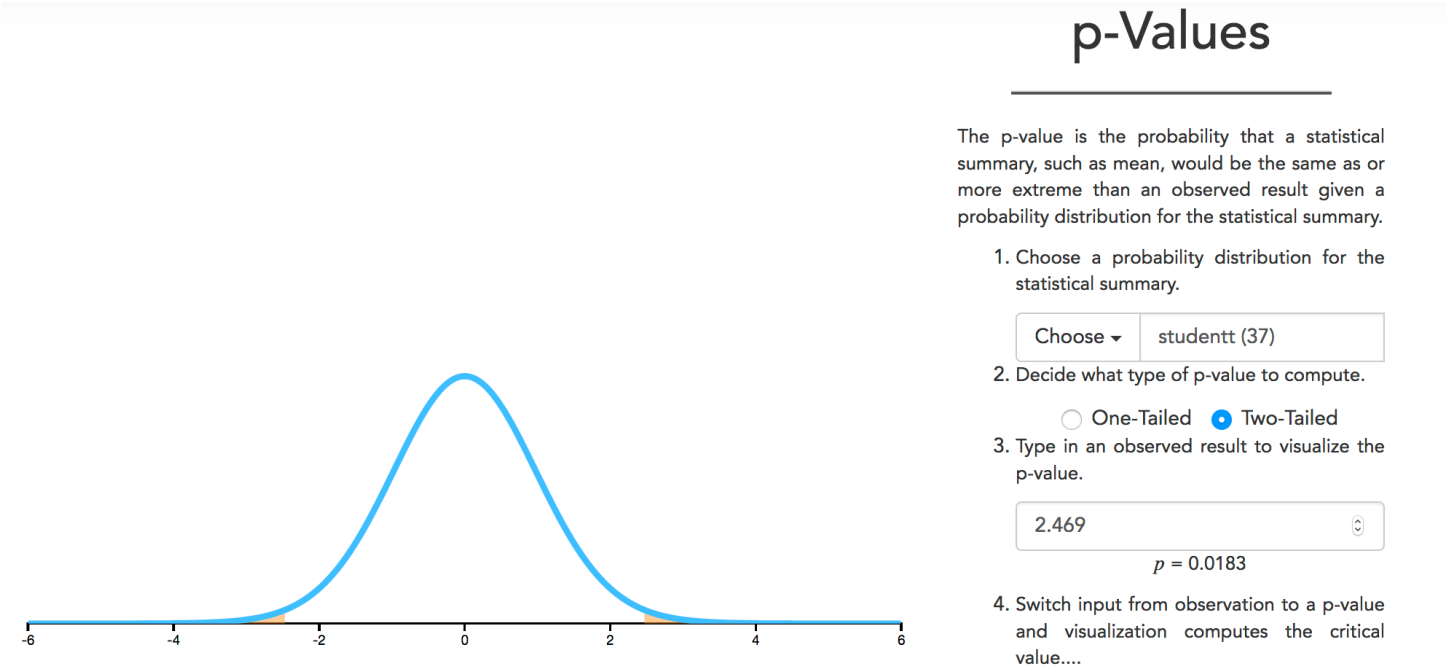
Example: high lysine corn

```
t.test(Weight ~ Group, data=corn)
```

```
##
##  Welch Two Sample t-test
##
## data:  Weight by Group
## t = -2.469, df = 36.915, p-value = 0.0183
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -66.729186  -6.570814
## sample estimates:
## mean in group Control   mean in group Exper
##                366.30                402.95
```

- So we would only expect to observe a t statistic as large as this (t_s=-2.46) in 1.83% of samples when H_0 is true.
- Based on this, we are more likely to believe H_A.

Visualising P-values



http://students.brown.edu/seeing-theory/index.html

Drawing conclusions from a t test

- How to interpret P-values?
  - Significance level: α = 0.1, 0.05, 0.01.
  - Value of is chosen by whoever is making the decision.
- If P-value is less than or equal to α, H_0 is rejected.
- For the corn example, the P-value was 0.0183.
  - Reject H_0 at $\alpha = 0.1$ or $\alpha = 0.05$, but not at $\alpha = 0.01$.
- Often say a *significant difference* exists between μ_1 and μ_2 if we reject H_0 when the P-value < α.

t tests and confidence intervals

- Close relationship between hypothesis tests and confidence intervals.
- A 100(1-α)% confidence interval for $\mu_1 - \mu_2$ which does not contain 0 tells us that a hypothesis test at level α will reject the null hypothesis.
- Both methods use the same three quantities: the difference between the sample means, the standard error of the difference between the sample means, and a t-value.

Type I and Type II errors

- Type I error: reject H_0 when H_0 is in fact true.
- Type II error: fail to reject H_0 when H_0 is false.

	True situation	
	H_0 true	H_0 false
Do not reject H_0	Correct	Type II error
Reject H_0	Type I error	Correct

Interpretting

- How do you know which value of α to use?
  - α is the probability of rejecting H_0 when H_0 is true.
  - i.e., P(Type I error) = α
- If you conducted a large number of hypothesis tests in which the null hypothesis were true, the expected proportion of tests in which H_0 would be rejected is equal to α.

Significant versus *important*

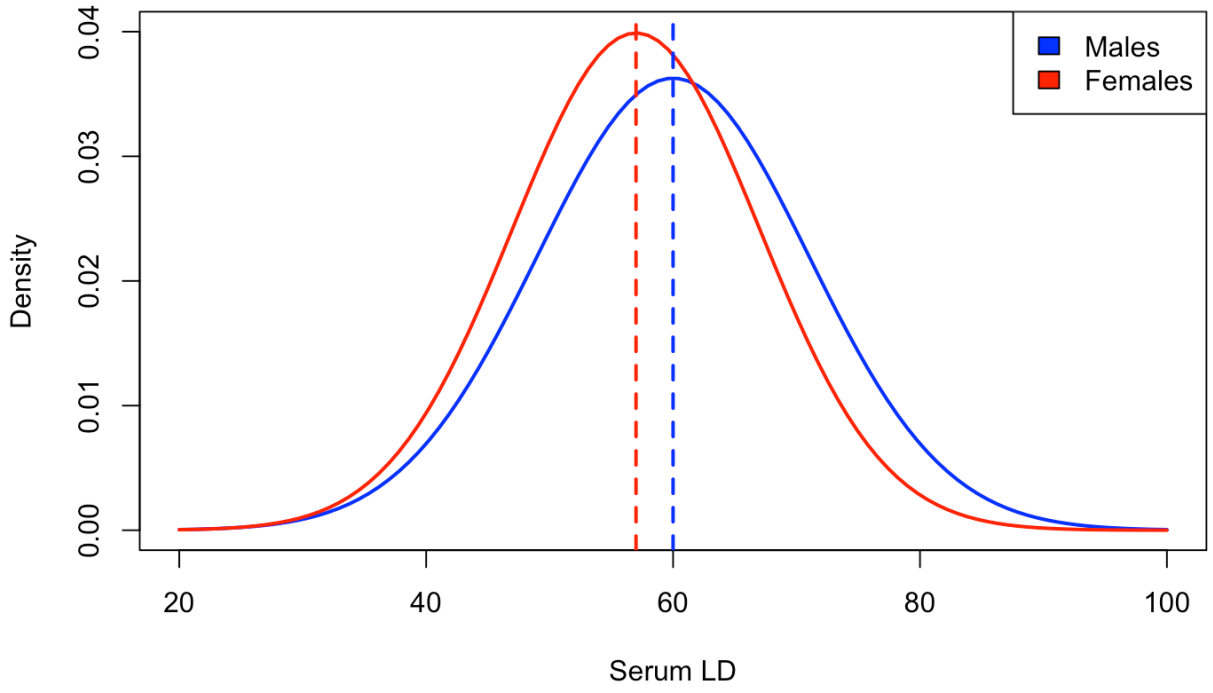
- Difference between μ_1 and μ_2 is said to be *significant* if the P-value is less than or equal to α. Otherwise the difference is said to be *non-significant*.
- **BUT**, just because a difference is significant, it doesn't make it important. Can't use the P-value to decide whether a significant difference is important. Have to think about importance in the context of the problem.

Example: Serum LD

- Serum LD: enzyme showing elevated levels after heart damage

	Males	Females
n	270	264
$\bar{y}$	60	57
s	11	10

$$t_s = \frac{(\bar{y}_1 - \bar{y}_2)}{SE(\bar{y}_1 - \bar{y}_2)} = \frac{60-57}{\sqrt{\frac{11^2}{270} + \frac{10^2}{264}}} = 3.29 \implies p \approx 0.001$$

- So the difference is statistically significant, but a difference of 3 between males and females may not be important.

Example: Serum LD