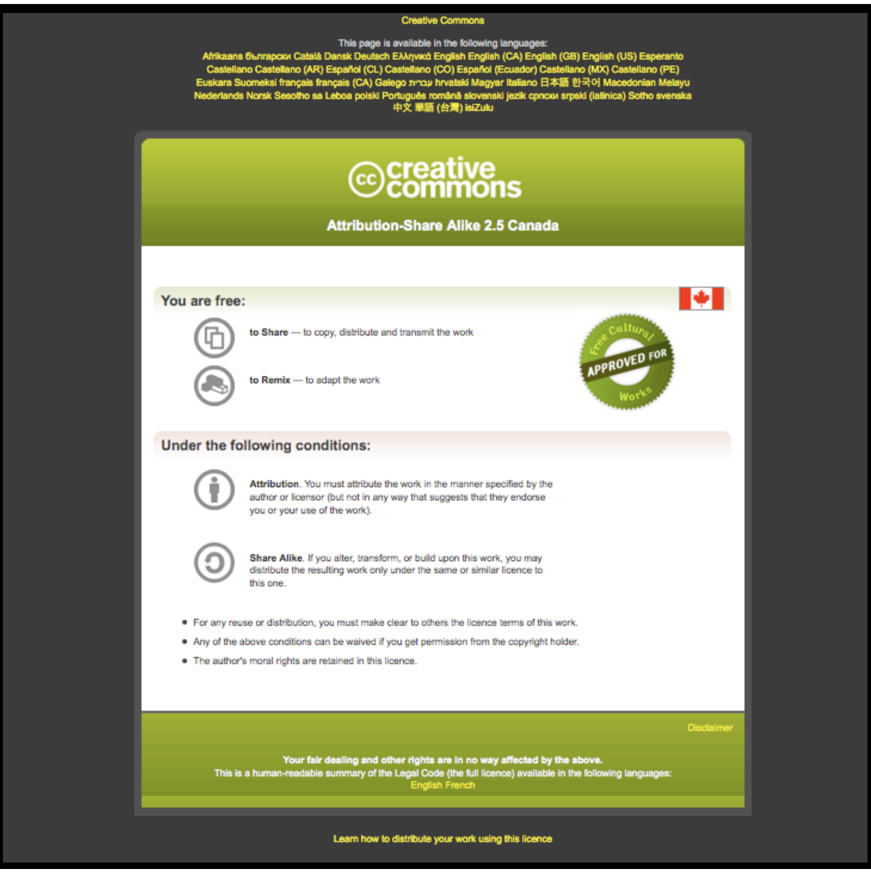
Canadian Bioinformatics Workshops

www.bioinformatics.ca



Creative Commons

This page is available in the following languages:
Afrikaans български Català Dansk Deutsch Ελληνικά English English (CA) English (GB) English (US) Esperanto Castellano Castellano (AR) Español (CL) Castellano (CO) Español (Ecuador) Castellano (MX) Castellano (PE) Euskara Suomeksi français français (CA) Galego עברית hrvatski Magyar Italiano 日本語 한국어 Macedonian Melayu Nederlands Norsk Sesotho sa Leboa polski Português română slovenski jezik српски srpski (latinica) Sotho svenska 中文 華語 (台灣) isiZulu

Attribution-Share Alike 2.5 Canada

You are free:

- **to Share** — to copy, distribute and transmit the work
- **to Remix** — to adapt the work

Under the following conditions:

Attribution. You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).

Share Alike. If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar licence to this one.

- For any reuse or distribution, you must make clear to others the licence terms of this work.
- Any of the above conditions can be waived if you get permission from the copyright holder.
- The author's moral rights are retained in this licence.

Disclaimer

Your fair dealing and other rights are in no way affected by the above.
This is a human-readable summary of the Legal Code (the full licence) available in the following languages:
English French

Learn how to distribute your work using this licence

Module 3
Hypothesis Testing

Daniele Merico

Exploratory Data Analysis and Essential Statistics using R
January 24-25, 2011

bioinformatics.ca

UNDERSTANDING TYPE I AND TYPE II ERRORS

	H_0: I AM RIGHT	H_a: I AM WRONG
ACCEPT H_0	I AM RIGHT & WAS TESTED RIGHT	I AM WRONG & WAS TESTED RIGHT (TYPE II ERROR) β
REJECT H_0	I AM RIGHT & WAS TESTED WRONG (TYPE I ERROR) α	I AM WRONG & WAS TESTED WRONG

UNIVERSITY OF TORONTO

Donnelly Centre
for Cellular + Biomolecular Research

Post-doctoral Fellow
Donnelly Centre
University of Toronto

http://baderlab.org/DanieleMerico

Outline

1. Discrete and continuous variables
2. Analytical variable distributions
3. Populations and samples, sampling distribution of the mean
4. Confidence interval of the mean
5. Inferential statistics: null hypothesis and alternative hypothesis, p-value, type-I and type–II errors
6. Power calculations
7. One-sample and two-sample t-test
8. Two-sample paired t-test
9. Permutation-based tests
10. Multiple testing correction
11. Applications to microarray data analysis

Discrete and Continuous Variables

- **Discrete**

  Values can be counted, i.e. associate an integer index

  e.g. Number of petals on the daisies in the gardens of Ottawa
  - Daisies (in the gardens of Ottawa): population units
  - Number of petals: discrete variable (numerical)

  Car brands in Sudbury
  - Cars (in Sudbury): population units
  - Car brand: discrete variable (categorical)

Discrete and Continuous Variables

- **Continuous**

  Any real value in a range (continuous)

  e.g. Blood pressure of overweight Canadians
  - Overweight Canadians: population units
  - Blood pressure: continuous variable (numerical)

  Liters of wastewater produced by each Toronto inhabitant in 2010
  - Toronto inhabitants: population units
  - Liters of wastewater (2010): continuous variable (numerical)

Analytical Variable Distributions

- **Empirical distributions**
  - We measure all the members of a *population* for some property
  - We end up having a finite number of values
  - Their distribution can be summarized using the techniques described in the previous chapter (histogram, …)
  - The probability of observing a value in a given range is just the empirically observed frequency
- **Analytical Distributions**
  - What if we can define analytically the distribution?
  - i.e. use a mathematical formula P (x) = f (x)

Discrete Analytical Distributions

- Cast a (fair) 6-face dice, observe the number on the top face
  - Population units: all the possible dice-casting events for that (fair) 6-face dice
  - Discrete variable: number on the top face of the dice
- Probability distribution

$$P(x) = 1/6, \quad x \in \{1,2,3,4,5,6\}$$

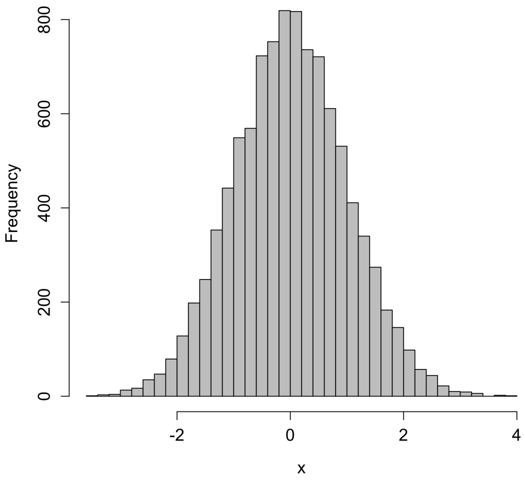
P (1) = 1/6	P (4) = 1/6
P (2) = 1/6	P (3) = 1/6
P (5) = 1/6	P (6) = 1/6

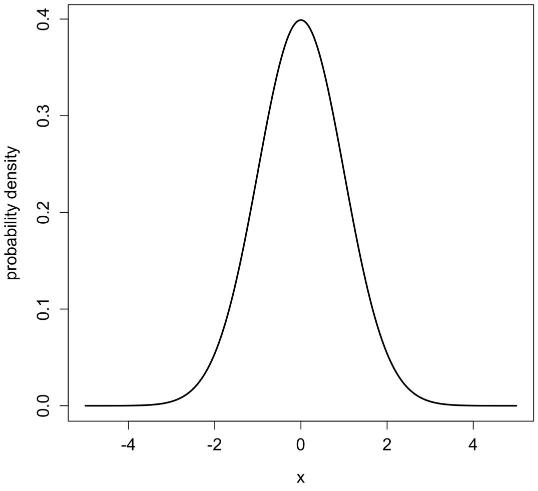
- This is a uniform discrete distribution
  - It's mathematically simple,

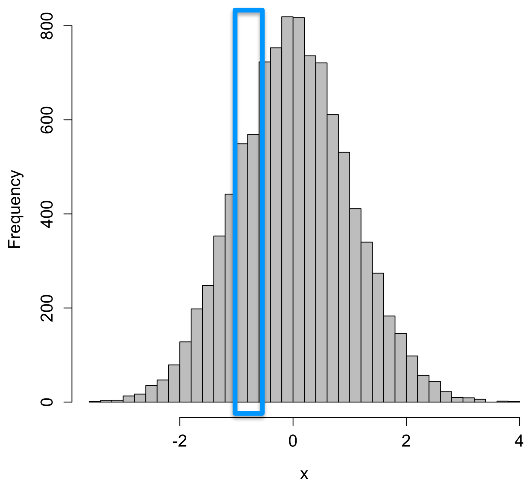
    but not all discrete analytical distributions are as simple

Continuous Analytical Distributions

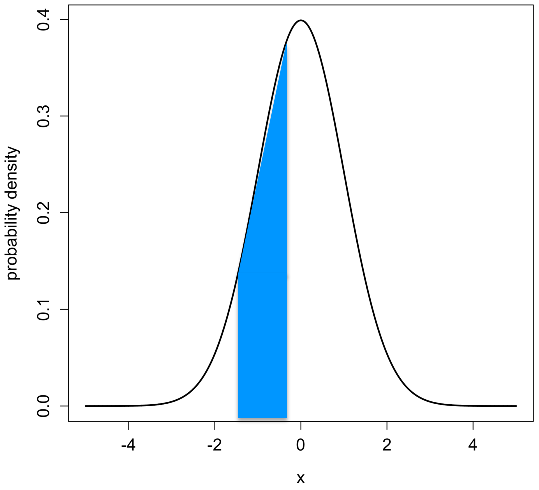
- Since the variable can have any possible value in a range, the probability of a single value in not finite
- We need calculus to correctly handle the probability distribution, which is called **density function**







Empirically observed frequency
(count the number of values observed)



Analytical probability density
(area under the curve)

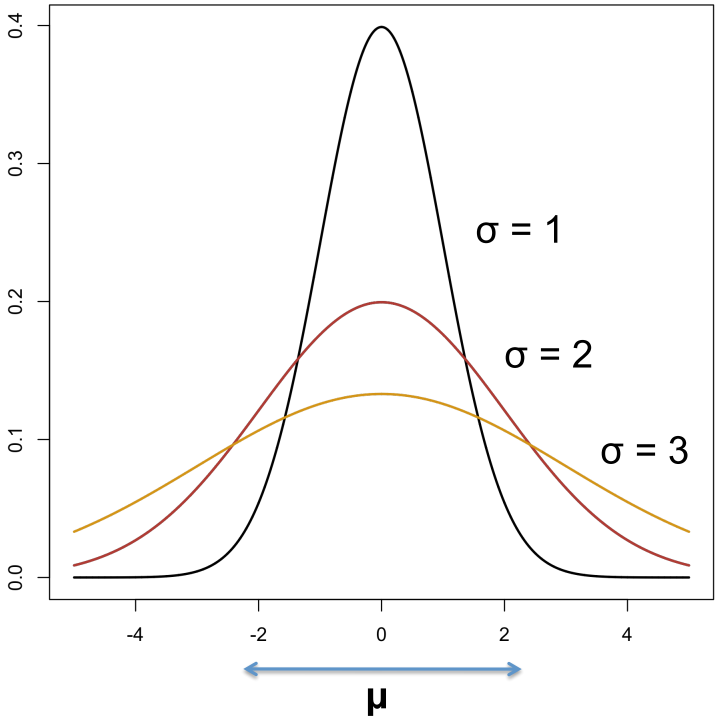
$$P\left(x_a < x < x_b\right) = \int_{x_a}^{x_b} f(x)dx$$

Normal Distribution

- The Normal is a very important distribution
  - Often found when measuring a physical property multiple times (variability due to random instrumental errors)
  - Often found for anthropometric indexes in human populations
  - The *sampling mean* follows the normal distribution

$$f\left(x\right) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Parameters:
- μ = Mean (x)
- σ = StDev (x)



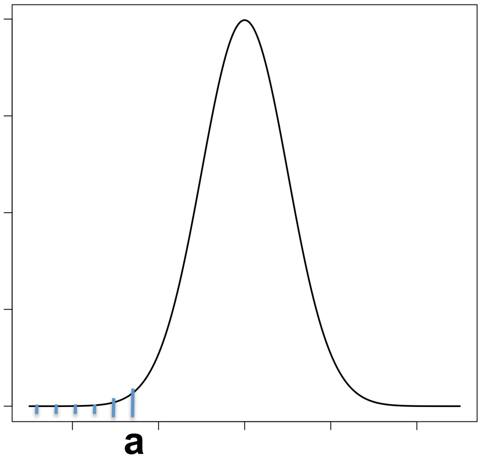
The normal is symmetric and centered on μ

σ affects the width of the curve

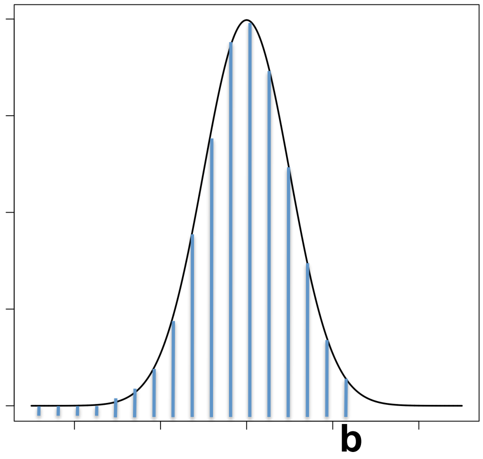
μ affects the position of the center of the curve

Normal Distribution in R: Find P given x

P (x < a) = …



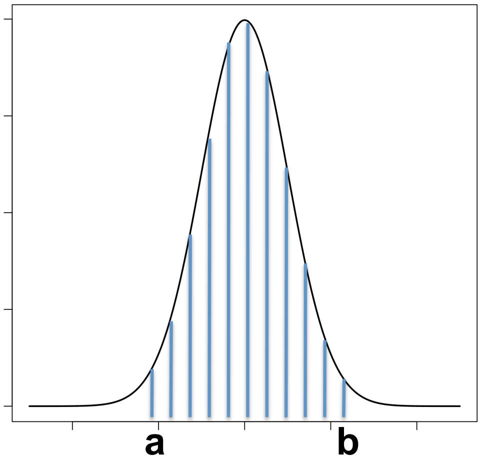
P (x < b) = …



```
pnorm (x = …, mean = …, sd = …)
```

Normal Distribution in R: Find P given x

P (a < x < b) = P (x < b) - P (x < a)

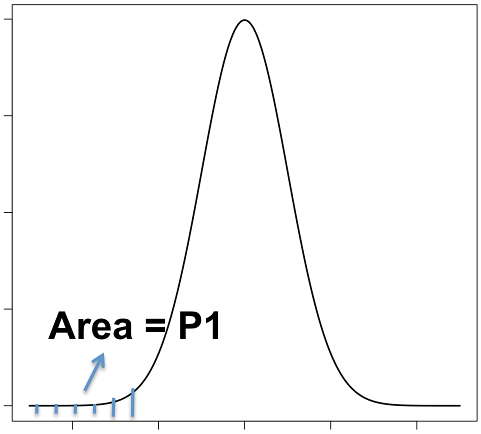


<u>Assignment:</u>
verify that for any mean and standard deviation, the probability of x falling within μ ± 2 σ is about 95%

```
pnorm (x = xa.n, …) – pnorm (x = xb.n, …)
```

Normal Distribution in R: Find x given P

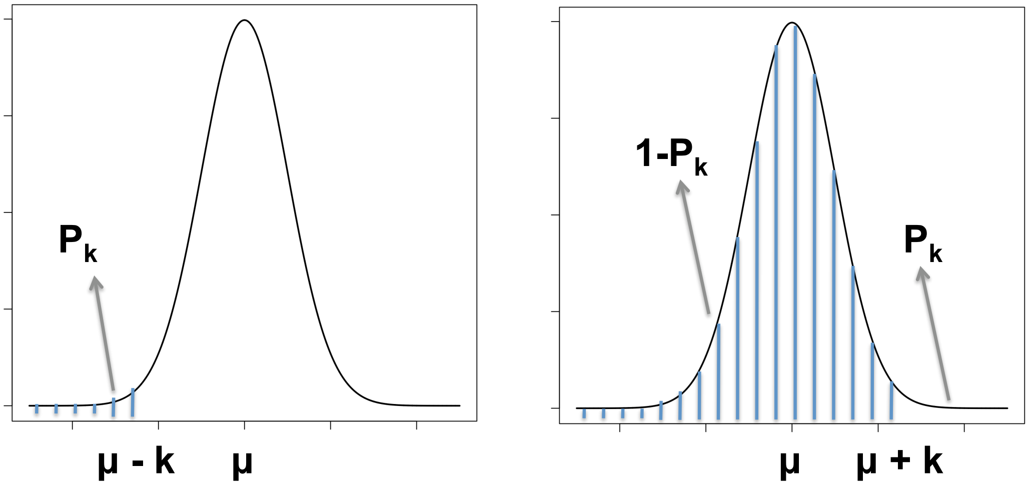
P (x < …) = P1



```
qnorm (p = …, mean = …, sd = …)
```

Normal Distribution: The Effect of Symmetry

$P\ (x < \mu - k) = P_k$ $\qquad$ $P\ (x < \mu + k) = 1 - P_k$



Assignment: *test this property using* **`qnorm ()`**

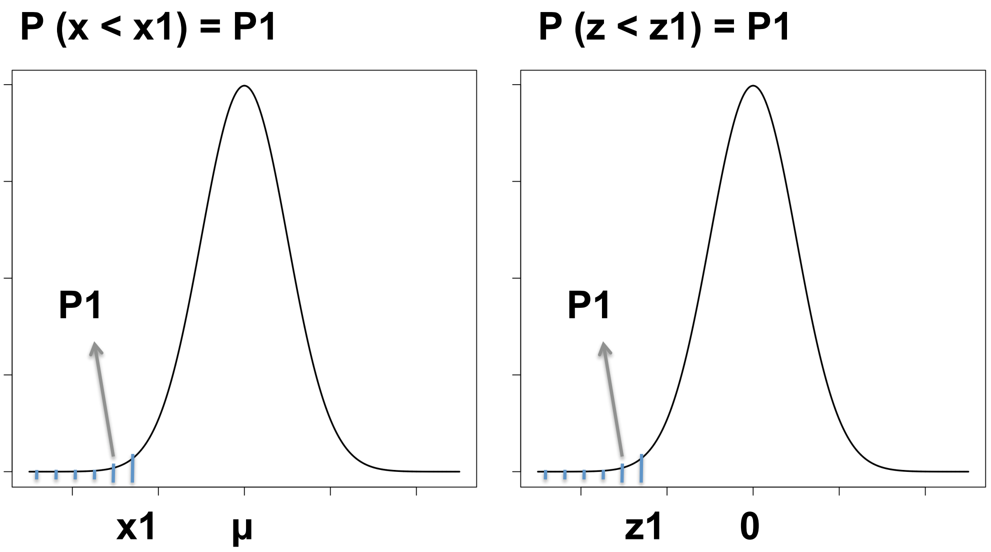
The Standard Normal and the z-score

- The Standard Normal distribution has μ = 0, σ = 1
- The z-score is used to transform normally distributed variables into a standard normal
  - Z follows the standard normal

$$z = \frac{x - \mu}{\sigma}$$

  - The z-score is often interpreted as the number of standard deviations from the mean
  - The reverse formula is also important $x = \mu + z \cdot \sigma$

Normal Distribution: Find x given P using the Standard



x1 = μ + z1*σ

- Test this relation: **x1 = μ + z1*σ**

  using the R commands you have learnt

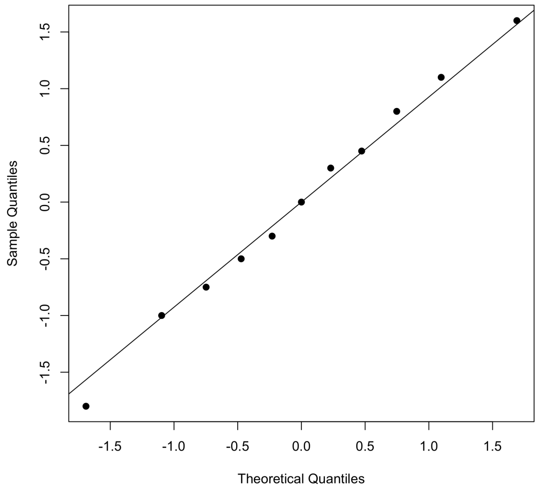
```
# Normal
x1.n <- qnorm (p = …, mean = …, sd = …)

# Standard Normal
z1.n <- qnorm (p = …)
```

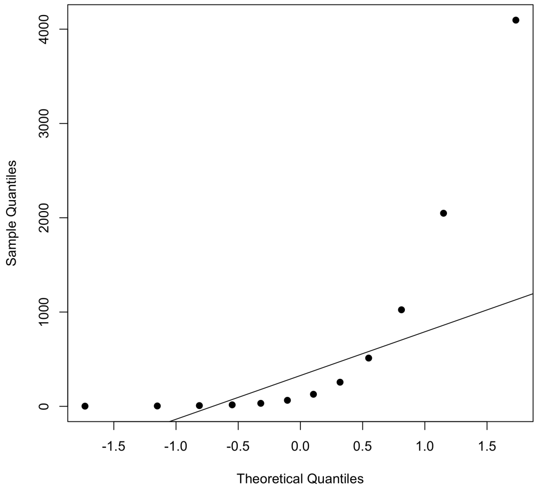
QQplot

- The qqplot of an observed distribution versus the normal can be used to evaluate how close the observed distribution is to the normal
  - The point should be lying on a line

```
# quasi-normal
x.nv <- c (-1.8, -1, -0.75,
        -0.5, -0.3, 0, 0.3,
        0.45, 0.8, 1.1, 1.6)
qqnorm (x.nv, pch = 19)
qqline (x.nv)
```

```
# not normal
x.nv <- 2 ^ (1: 12)
qqnorm (x.nv, pch = 19)
qqline (x.nv)
```





Population and Sample

- **Population**

  set of entities (individuals, objects, events)

  mean: **μ**

  stdev: **σ**

- **Sample**

  subset of a population

  mean: **m**

  stdev: **s**

Correction for Sample Stdev

- Population $$\sigma = \sqrt{\frac{1}{N}\sum_{N}^{i=1}\left(M(x) - x_i\right)^2}$$

- Sample $$s = \sqrt{\frac{1}{N-1}\sum_{N}^{i=1}\left(M(x) - x_i\right)^2}$$

The R function **`sd ()`** *uses by default the second definition*

Populations, Samples, Inferences

- Measuring a property for all the units of a population is often not practical

→ Only the units in a sub-set (i.e. sample) are measured

- If we can only measure sample,

  can we make inferences that hold at the population level?

- This is the object of **Statistical Inference**

Sampling Variable

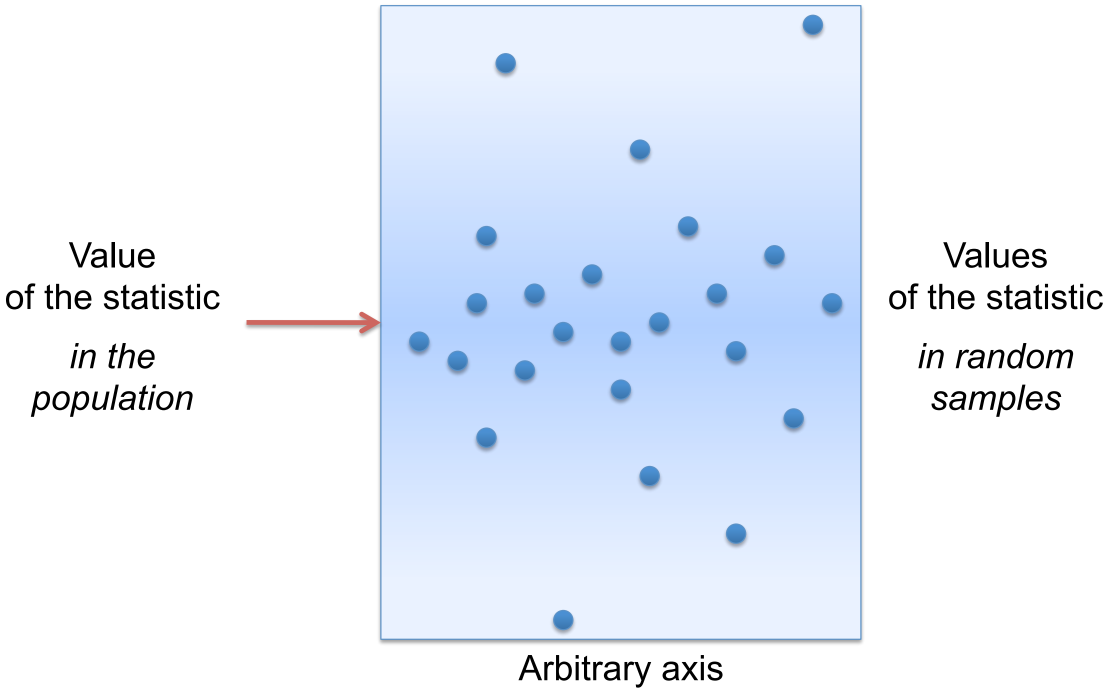
- Generate many *random* samples of a population (sample size: N)
- For each sample, measure a property → ***variable***
- For each sample, compute a statistic summarizing the variable (e.g. mean)

→ New variable (**sampling variable**)

  – New population units: samples of the original population

- *How is this useful..?!*
  – The statistic has only one value in the population (e.g. mean)
  – Different random samples will have values which cluster around the population statistic

  → Useful to study this to guide statistical inference



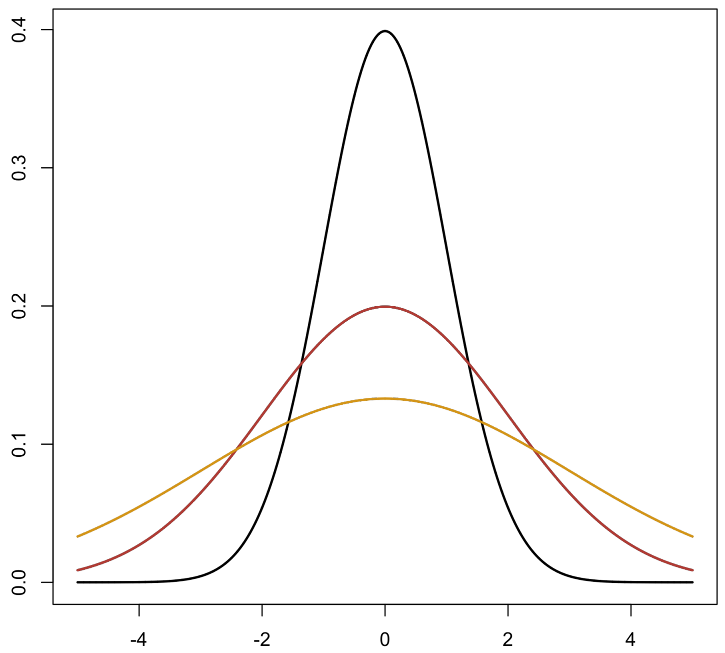
Value of the statistic *in the population*

Values of the statistic *in random samples*

Arbitrary axis

Sampling Mean

- **Sampling mean** of a variable x: $\bar{x}$
  mean of variable x for each random sample (sample size: N)
  - Mean of the sampling mean $\mu(\bar{x}) = \mu(x)$
  - Stdev of the sampling mean $\sigma(\bar{x}) = \frac{\sigma(x)}{\sqrt{N}}$
  - *What happens if sample size ≈ population size?*



As N increases, the sample means of the statistic become closer to the population value of the statistic

Sampling Mean Distribution

- If the distribution of x is normal,
  the distribution of $\bar{x}$ is normal as well

- Even if the distribution of x is not normal,
  when sample size *N is sufficiently large*
  the distribution of $\bar{x}$ is ***normal***

  (Central Limit Theorem)

- *For practical purposes, sufficiently large corresponds to N > 30*

Sampling Mean Distribution

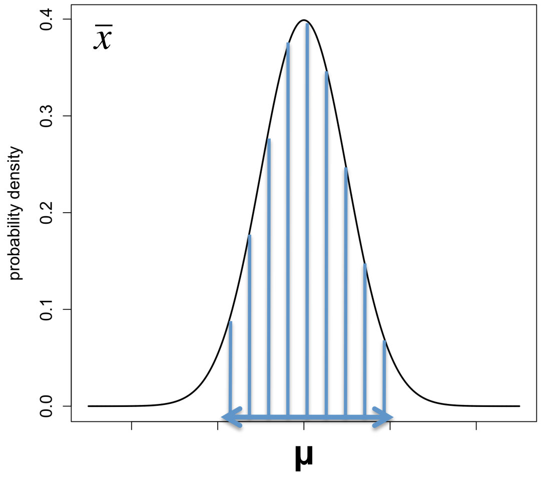
- How is this useful?
- We have a model defining a quantitative relation between the population and sample mean
- Is the sample mean probable or improbable under the population sampling mean distribution?

Confidence Interval of the Mean

- Known
  - Population parameters: **μ, σ**
  - Sample size (N): **≥ 30**
- Goal
  - Determine the range of possible sample mean values for this population
- Strategy
  - Use the sampling mean distribution (normal)

Sampling mean distribution:

- Normal
- Mean = μ, Stdev = $\frac{\sigma}{\sqrt{N}}$



Confidence Interval of the Mean

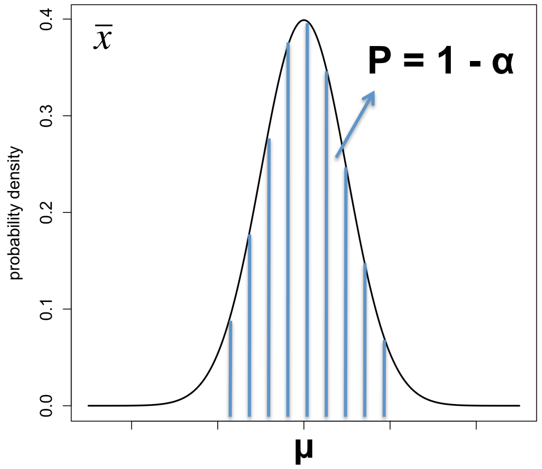
- Solution
  1. Set the probability ***α*** of $\overline{x}$ falling *outside* the interval (usually α = 0.05)

This is the confidence associated to the interval:

- *The probability of x being outside the interval is α*
- *The probability of x being within the interval is 1 - α*

Sampling mean distribution:

- Normal
- Mean = μ, Stdev = $\frac{\sigma}{\sqrt{N}}$



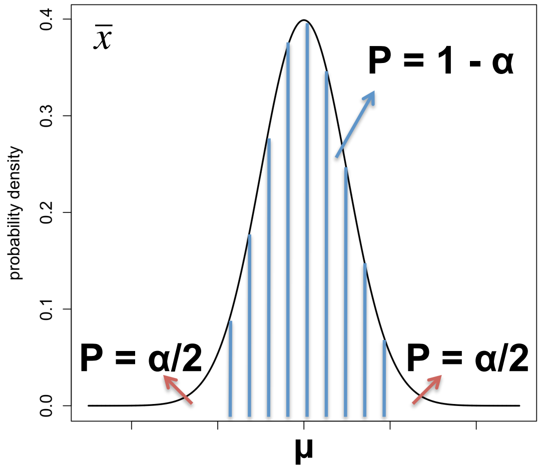
Confidence Interval of the Mean

- Solution
  1. Set the probability ***α*** of $\overline{x}$ falling *outside* the interval (usually α = 0.05)
  2. Find $z_{\alpha/2}$: P $(z < z_{\alpha/2}) = 1 - \alpha/2$ (standard normal)

We use the Standard Normal for reasons that will be clearer later. However in R we can use any normal distribution to compute x given the probability

Sampling mean distribution:

- Normal
- Mean = μ, Stdev = $\frac{\sigma}{\sqrt{N}}$



Confidence Interval of the Mean

- Solution
  1. Set the probability ***α*** of $\overline{x}$ falling *outside* the interval (usually α = 0.05)
  2. Find $z_{\alpha/2}$: P ($z < z_{\alpha/2}$) = 1 - α/2 (standard normal)
  3. The Confidence Interval is:

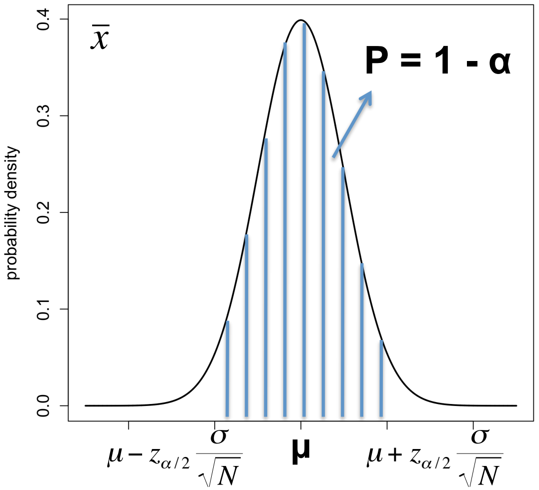
$$\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{N}} < \overline{x} < \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{N}}$$

*Since x = Mean + z * StDev*

x2 | P (x > x2) = 1 - α/2

*x2 = μ + $z_{\alpha/2}$ * σ/√N*

Sampling mean distribution:
- Normal
- Mean = μ, Stdev = $\frac{\sigma}{\sqrt{N}}$



Confidence Interval of the Mean

- This is just a way to express the confidence interval in terms of the population parameters and Standard Normal quantile

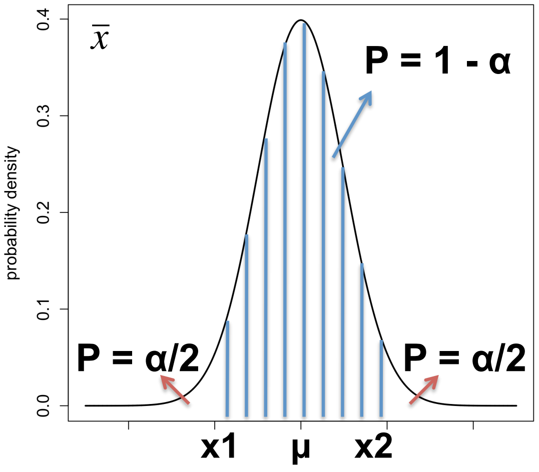
$$\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{N}} < \overline{x} < \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{N}}$$

$$\overline{x}_1 < \overline{x} < \overline{x}_2$$

$$P(\overline{x} < \overline{x}_1) = \alpha/2$$

$$P(\overline{x} < \overline{x}_2) = 1 - \alpha/2$$

Sampling mean distribution:
- Normal
- Mean = μ, Stdev = $\frac{\sigma}{\sqrt{N}}$



```
# In R,
# You can directly compute x1, x2

x1.n <- qnorm (p = 0.025, mean = …, sd = …)
x2.n <- qnorm (p = 0.975, mean = …, sd = …)

# Or use z_{α/2}

z_a2.n <- qnorm (p = 1 - 0.025)
x1.n <- mu.n - z_a2.n * sd.n / sqrt (N.n)
x1.n <- mu.n + z_a2.n * sd.n / sqrt (N.n)

# Equivalent way to compute z_{α/2} (symmetry)
z_a2.n <- - qnorm (p = 0.025)
```

Confidence Interval of the Mean

Unknown Population Parameters + Large Sample

- Known
  - Sample mean: **m**
  - Sample StDev: **s**
  - Sample size (N): **≥ 30**
- Goal
  - Determine the population mean confidence interval
- Strategy
  - Swap x and µ in the standard normal formula
  - Assume ***s*** is a good point estimate of ***σ***

$$m - z_{\alpha/2}\frac{s}{\sqrt{N}} < \mu < m + z_{\alpha/2}\frac{s}{\sqrt{N}}$$

*By extracting samples and computing their m and s for W times, M will fall in the confidence interval W * (1-α) times*

Confidence Interval of the Mean

Unknown Population Parameters + Small Sample

- For small samples (N < 30) derived from normally-distributed populations, the sample stdev is not a good estimate of the population stdev
- Instead of using the standard normal distribution, we have to use the t-student distribution
- The t-student density function depends on the degree of freedom = N – 1; for N > 30 t-student is quasi-normal

$$\mu - t(N-1)_{\alpha/2}\frac{s}{\sqrt{N}} < \bar{x} < \mu + t(N-1)_{\alpha/2}\frac{s}{\sqrt{N}}$$

```
# Confidence interval using the t-student

t_a2.n <- qt (p = 1 - 0.025, df = N.n - 1)
x1.n <- m.n - t_a2.n * s.n / sqrt (N.n)
x1.n <- m.n + t_a2.n * s.n / sqrt (N.n)
```

Hypothesis Testing

- Given a sample (with known mean and stdev), we want to test whether it may belong or not to a population (with known mean)
- We can use the framework we have derived for confidence interval, and reshape it as a ***test***
  - Application example:

    Monsanto claims that a new crop variety has a higher yield

    Compare the yield of a sample of Monsanto's new variety versus the historical yield average of the traditional variety and test Monsanto's claim

Hypothesis Testing: Null and Alternative Hypothesis

Monsanto claims that a new crop variety has a higher yield

Compare the yield of a sample of Monsanto's new variety versus the historical yield average of the traditional variety and test Monsanto's claim

Null Hypothesis ("status quo"):

the sample being tested could have been drawn form the population being tested

- Test Statistic: Mean
  - Distribution: t-student
- Null Hypothesis H_0: $\mu \leq \mu_0$
- Alternative Hypothesis H_1: $\mu > \mu_0$

μ: mean yield of the new variety

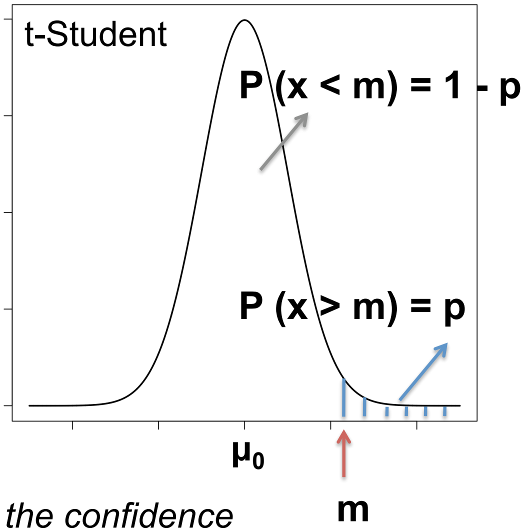
μ_0: mean yield of the traditional variety

Hypothesis Testing: p-value

- Set the confidence interval so that

  $$m = \mu_0 + t(N-1)_p \frac{s}{\sqrt{N}}$$

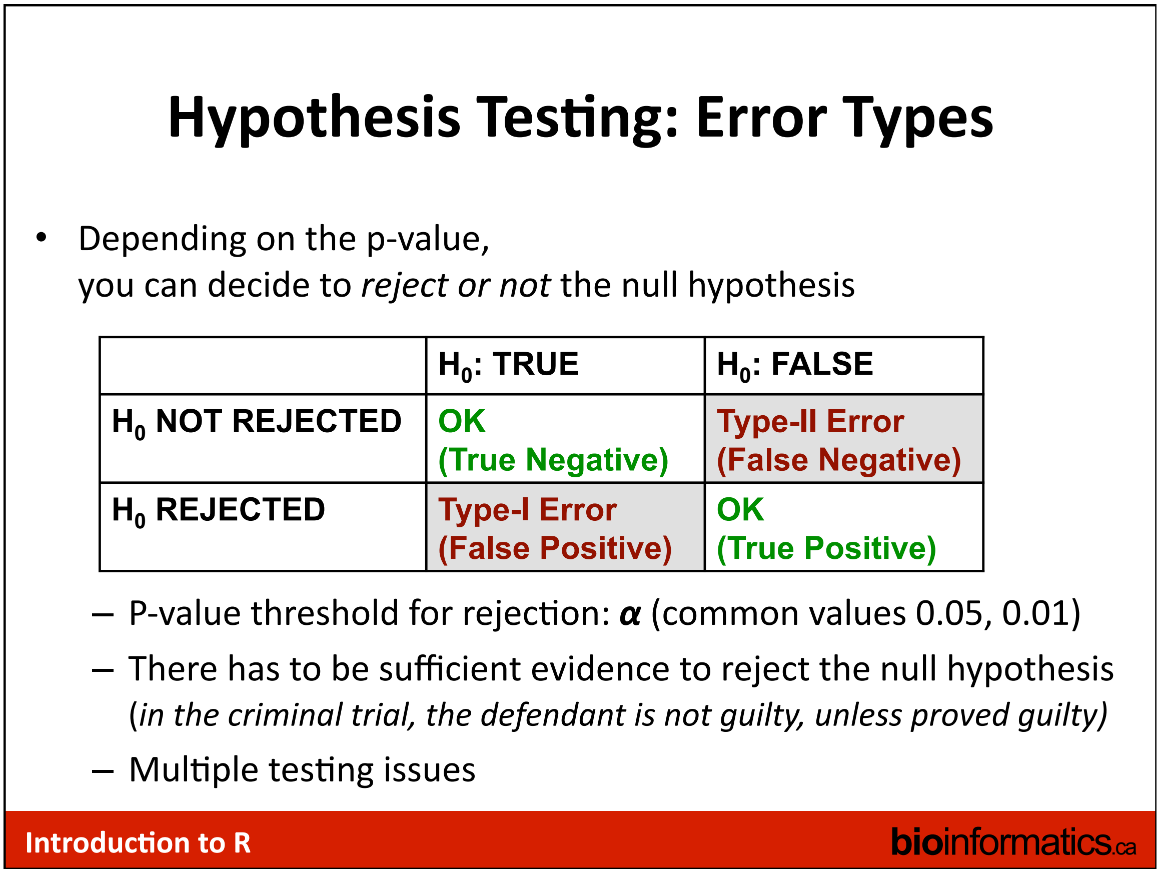
- p = probability of observing a population sample as extreme or more extreme than the one being tested when drawing from the population with mean μ_0
  - p >> 0: null hypothesis likely
  - p ~ 0: null hypothesis not likely



How much do we have to "stretch" the confidence interval to "explain" the observed sample mean?

Hypothesis Testing: p-value

- **Null Hypothesis**:
  - statistical model where differences are only due to random fluctuations (sampling)
    - If we could always work on *populations* only, we would not need inferential statistics
- **P-value**:
  - Probability that the null hypothesis model does not explain the data
  - → The differences observed are probably due to some underlying phenomenon

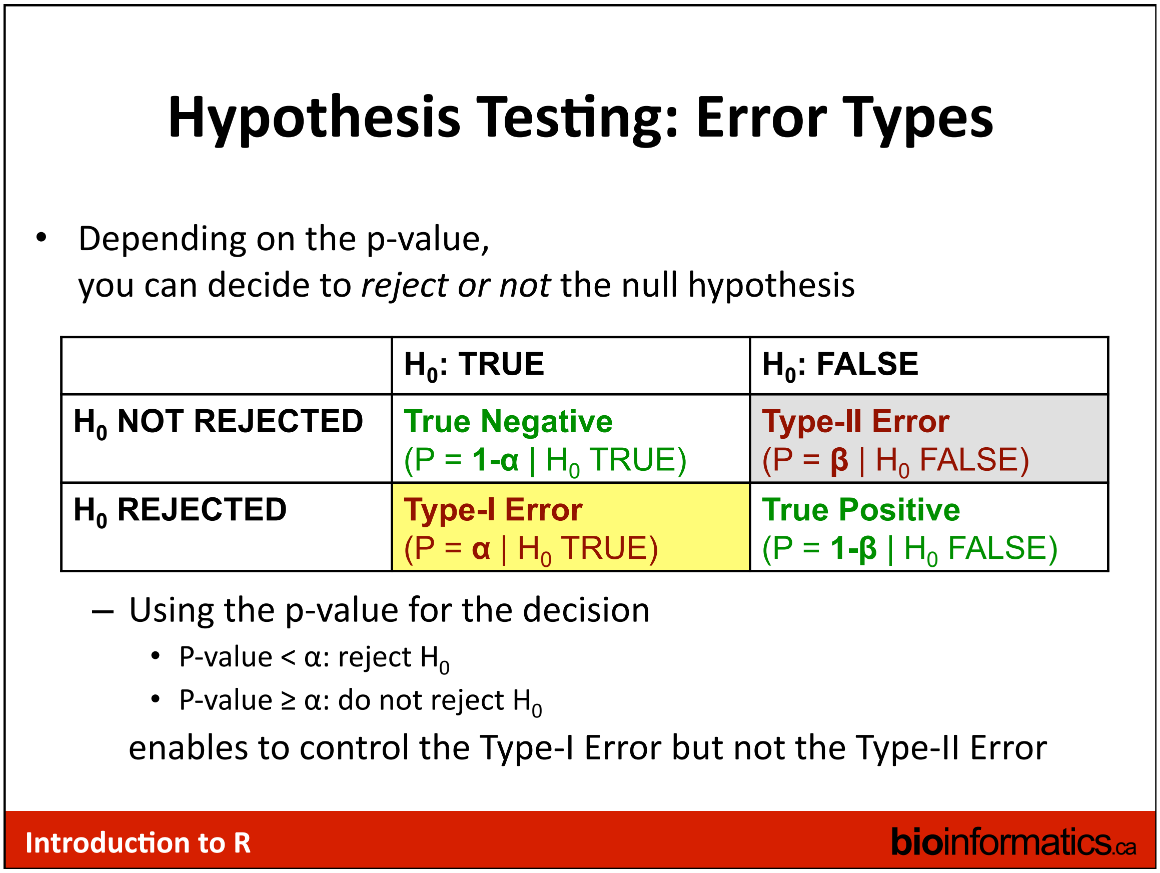


Hypothesis Testing: Error Types

- Depending on the p-value,
  you can decide to *reject or not* the null hypothesis

	H_0: TRUE	H_0: FALSE
H_0 NOT REJECTED	OK (True Negative)	Type-II Error (False Negative)
H_0 REJECTED	Type-I Error (False Positive)	OK (True Positive)

  - P-value threshold for rejection: α (common values 0.05, 0.01)
  - There has to be sufficient evidence to reject the null hypothesis (*in the criminal trial, the defendant is not guilty, unless proved guilty)*
  - Multiple testing issues



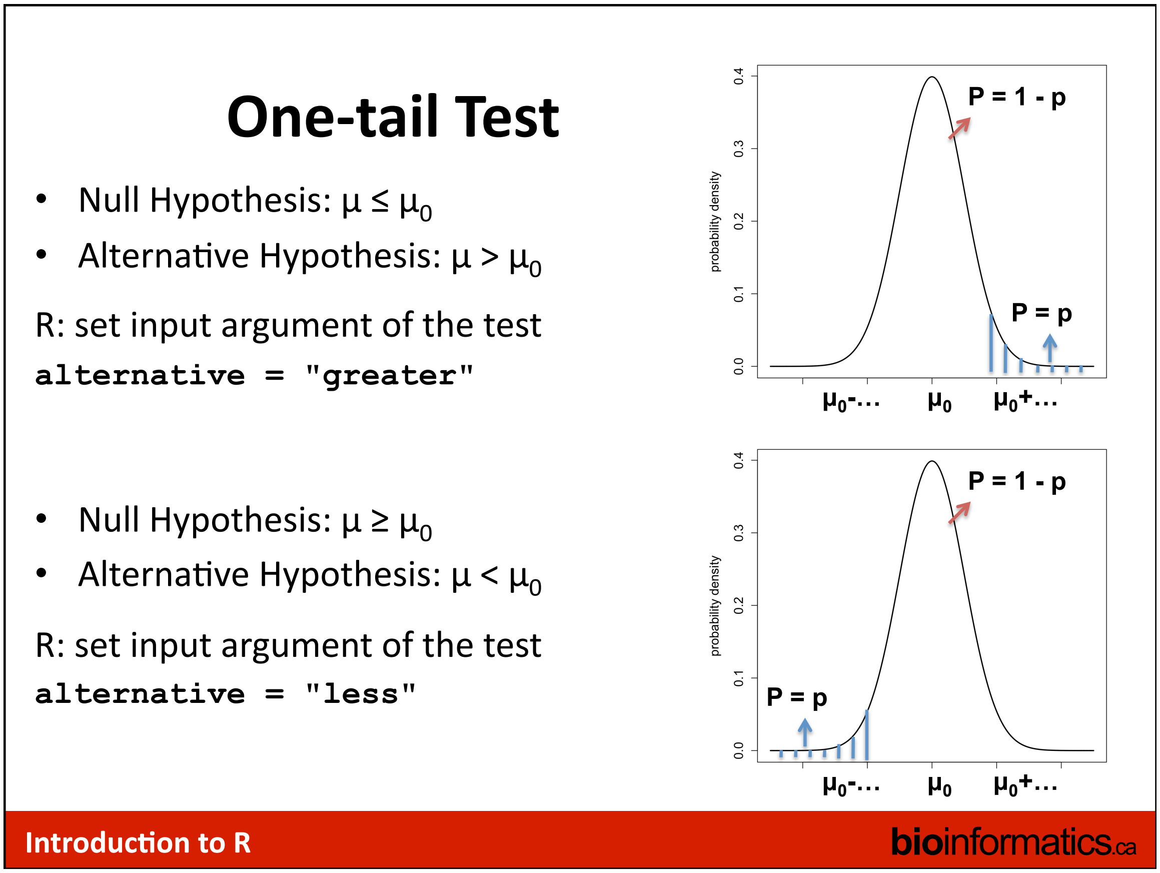
Hypothesis Testing: Error Types

- Depending on the p-value,
  you can decide to *reject or not* the null hypothesis

	H_0: TRUE	H_0: FALSE
H_0 NOT REJECTED	True Negative ($P = \mathbf{1-\alpha} \mid H_0$ TRUE)	Type-II Error ($P = \boldsymbol{\beta} \mid H_0$ FALSE)
H_0 REJECTED	Type-I Error ($P = \boldsymbol{\alpha} \mid H_0$ TRUE)	True Positive ($P = \mathbf{1-\beta} \mid H_0$ FALSE)

  - Using the p-value for the decision
    - P-value < α: reject H_0
    - P-value ≥ α: do not reject H_0

    enables to control the Type-I Error but not the Type-II Error



One-tail Test

- Null Hypothesis: $\mu \leq \mu_0$
- Alternative Hypothesis: $\mu > \mu_0$

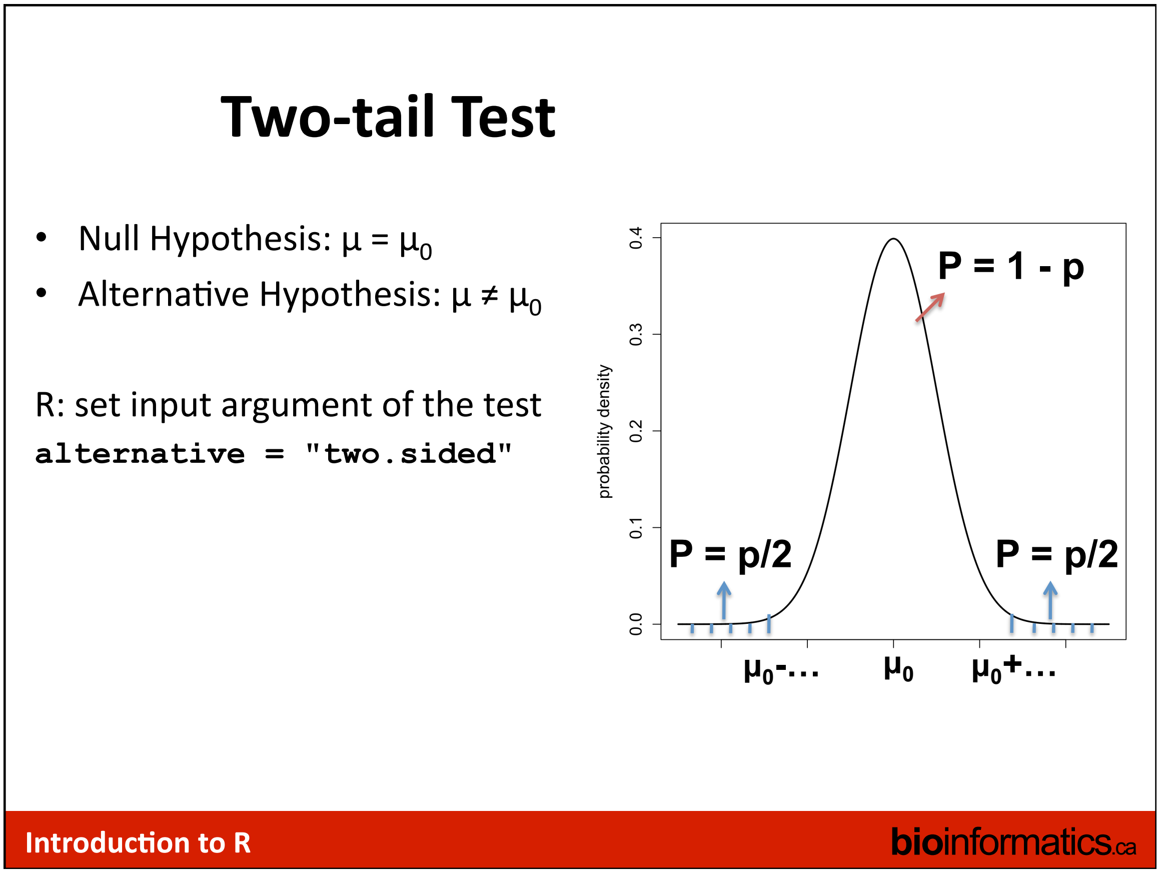
R: set input argument of the test

```
alternative = "greater"
```

- Null Hypothesis: $\mu \geq \mu_0$
- Alternative Hypothesis: $\mu < \mu_0$

R: set input argument of the test

```
alternative = "less"
```



Two-tail Test

- Null Hypothesis: $\mu = \mu_0$
- Alternative Hypothesis: $\mu \neq \mu_0$

R: set input argument of the test

```
alternative = "two.sided"
```

Power Calculations

Power = **1 - β**

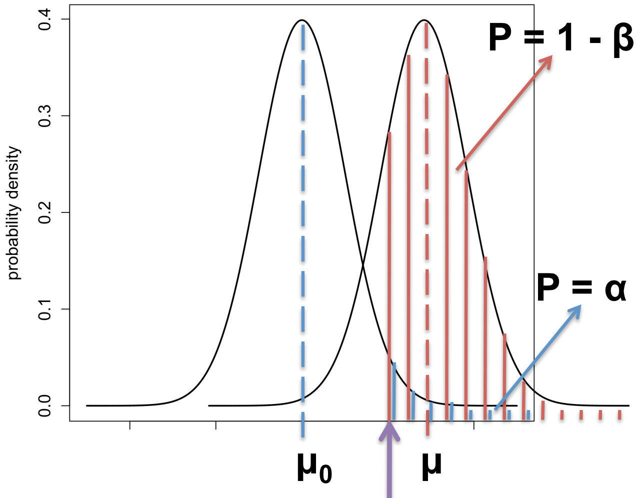
- Assume **H_0 is false**
- Set the **tail** of the test
- Set **α** (the p-value decision threshold)
- Set **μ** (mean of the sample source population)
- Set **σ** for both distributions (or use the sample estimate **s**)
- Set **N** (sample size)

→ The power is a function of all these factors

- It is common to plot the power as a function of μ-μ_0 or N

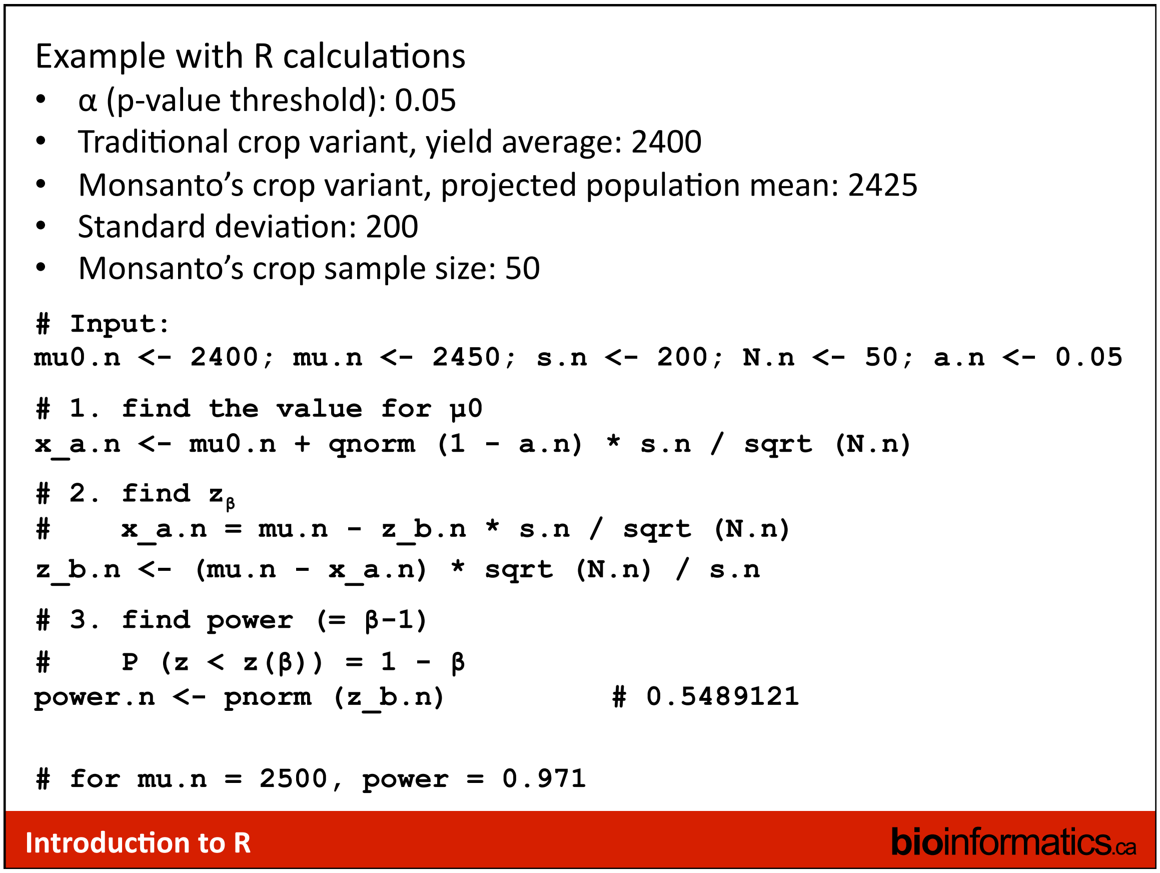
Power Calculations

1. Find the decision value on x with respect to μ_0, given α
2. Find the corresponding value for μ
3. Calculate the area under the curve (1-β)



For small samples use t-Student instead of standard normal (z)

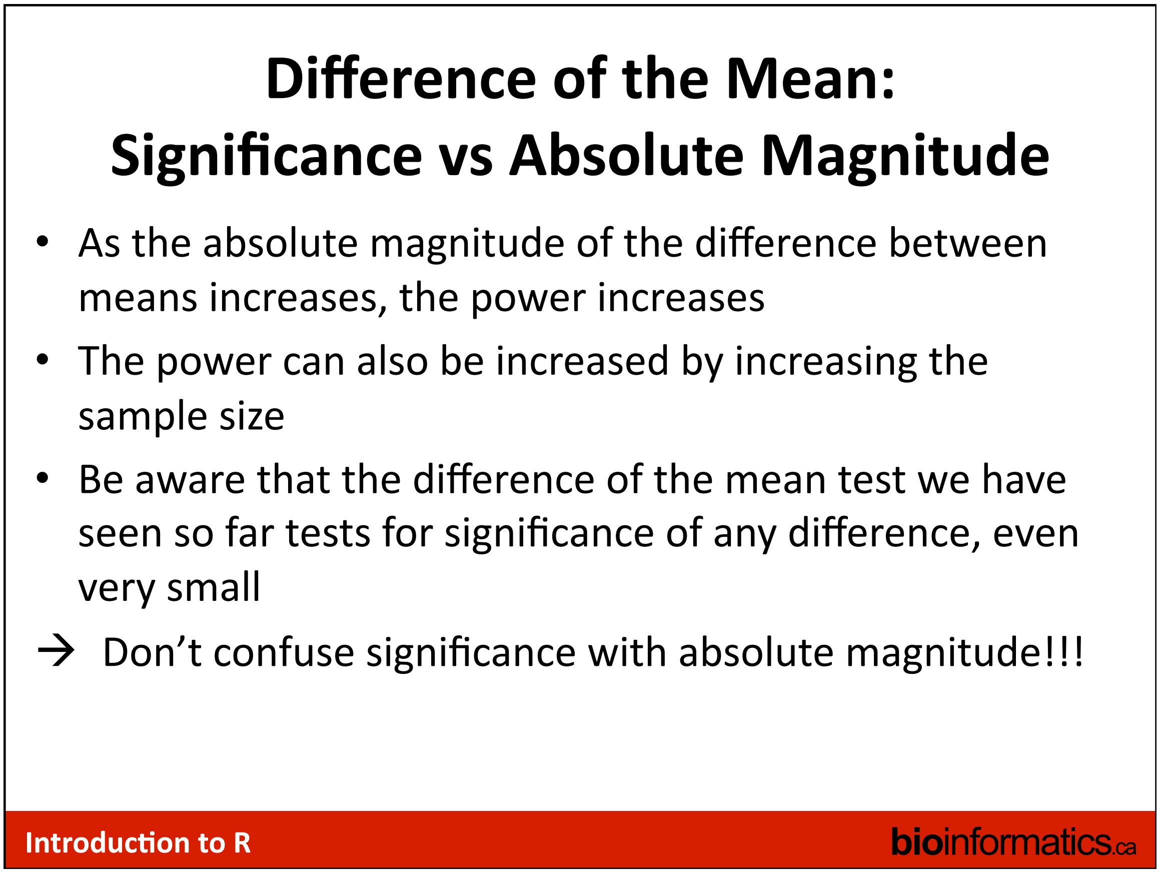
$$\mu_0 + z_\alpha \frac{\sigma}{\sqrt{N}} = \mu - z_\beta \frac{\sigma}{\sqrt{N}}$$



Example with R calculations

- α (p-value threshold): 0.05
- Traditional crop variant, yield average: 2400
- Monsanto’s crop variant, projected population mean: 2425
- Standard deviation: 200
- Monsanto’s crop sample size: 50

```
# Input:
mu0.n <- 2400; mu.n <- 2450; s.n <- 200; N.n <- 50; a.n <- 0.05

# 1. find the value for µ0
x_a.n <- mu0.n + qnorm (1 - a.n) * s.n / sqrt (N.n)

# 2. find zβ
#    x_a.n = mu.n - z_b.n * s.n / sqrt (N.n)
z_b.n <- (mu.n - x_a.n) * sqrt (N.n) / s.n

# 3. find power (= β-1)
#    P (z < z(β)) = 1 - β
power.n <- pnorm (z_b.n)          # 0.5489121


# for mu.n = 2500, power = 0.971
```



Difference of the Mean: Significance vs Absolute Magnitude

- As the absolute magnitude of the difference between means increases, the power increases
- The power can also be increased by increasing the sample size
- Be aware that the difference of the mean test we have seen so far tests for significance of any difference, even very small

→ Don’t confuse significance with absolute magnitude!!!

Difference of the Mean: Significance vs Absolute Magnitude

- Example
  - A new drug leads to a significant improvement in tumor size for a cohort of 5000 patients
  - But what's the average tumor shrinkage? Is it clinically relevant?
  - Statistically significant and clinically relevant are not the same

One-sample t-Test (Mean Difference): R

- Goal: does the sample belong to a population with mean larger/ smaller/different than a reference population with mean μ_0?
- Input
  - Reference population mean (μ_0)
  - Sample values
- Assumptions
  - Independence
    - The sample has been randomly drawn,
    - There is no dependence between sample units
  - Distribution
    - Small samples ($N < 30$): population normally distributed
    - Large samples ($N \geq 30$): none

One-sample t-Test (Mean Difference): R

- Example (Monsanto's new variety)
  - Reference yield mean: 2400
  - Sample yields: 2531, 2659, 2487, 2398, 2771
  - Alternative: Monsanto larger than reference

```
t.test (x = c (2531, 2659, 2487, 2398, 2771),
        mu = 2400,
        alternative = "greater")

# t = 2.5756, df = 4, p-value = 0.03081
# 95 percent confidence interval:
#  2429.151      Inf

# output class: list, with slots:
t.test (…)$p.value; t.test (…)$statistic
```

Two-sample t-Test (Mean Difference)

- Goal: do the samples belong to populations with mean larger/smaller/different?
- Input
  - Sample #1 values
  - Sample #2 values
- Assumptions
  - Independence
    - The samples has been randomly drawn,
    - There is no dependence between sample units
    - There is no dependence between samples
  - Distribution
    - Small samples (N < 30): population normally distributed
    - Large samples (N ≥ 30): none

Two-sample t-Test (Mean Difference): R

- Example: Monsanto compares two new varieties
  - Variety #1: 2405, 2378, 2254, 2471, 2390
  - Variety #2: 2531, 2659, 2487, 2398, 2771
  - Alternative: #1 different than #2

```
t.test (x = c (2405, 2378, 2254, 2471, 2390),
      y = c (2531, 2659, 2487, 2398, 2771),
      alternative = "two.sided")
```

```
# t = -2.5428, df = 6.129, p-value = 0.04311
# 95 percent confidence interval:
# -371.123009   -8.076991
```

  - *The confidence interval refers to the difference of the means*

Two-sample Paired t-Test

- Use instead of the standard two-sample t-test whenever sample units are highly correlated
  - E.g. patients before and after treatment

```
t.test (x = …,
      y = …,
      alternative = …,
      paired = T)
```

Non-parametric Test (Mean Difference)

- When the sample is small and the normality distribution assumption is not met,

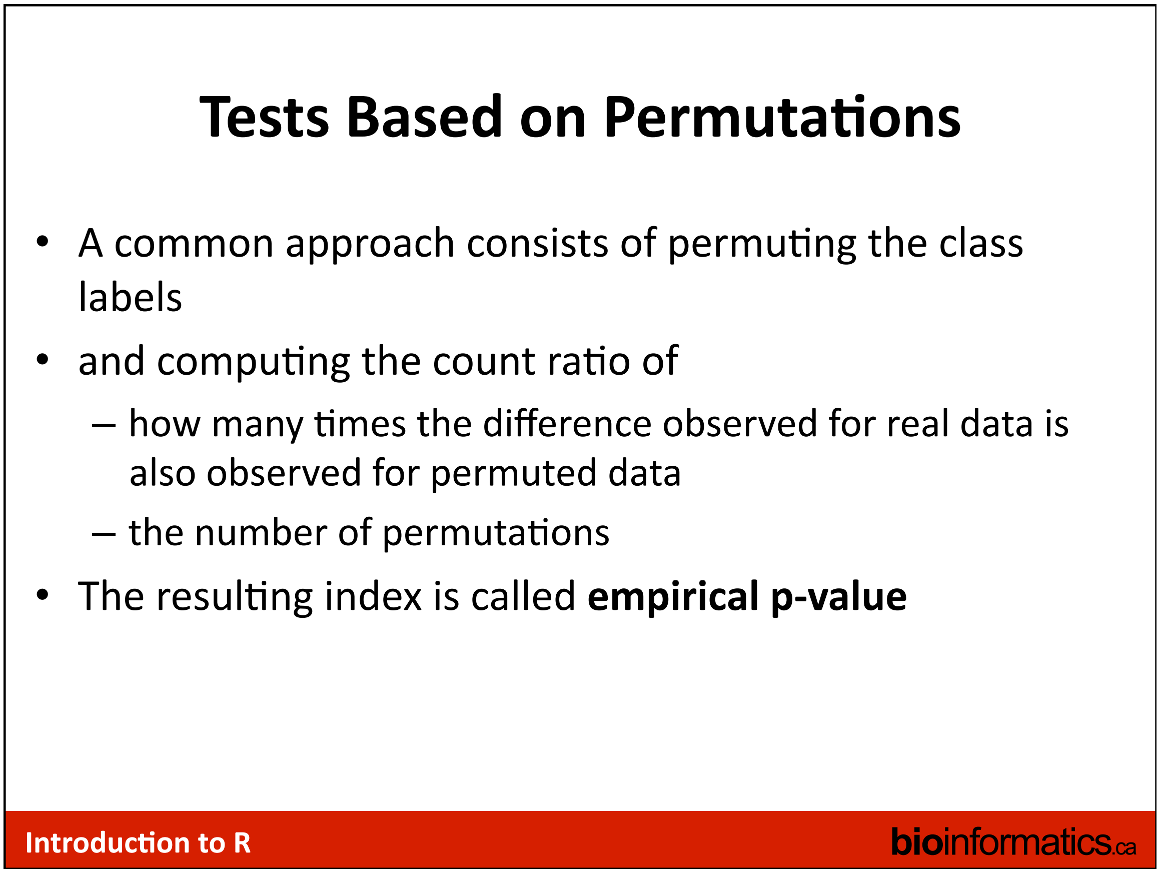
  Use the Wilcoxon test (a.k.a. Mann-Whitney test)
  - one-sample
  - Two-samples

```
wilcox.test (…)
```

  - The test works on the ***ranks*** of the values
  - The input and output is the same as the t-test

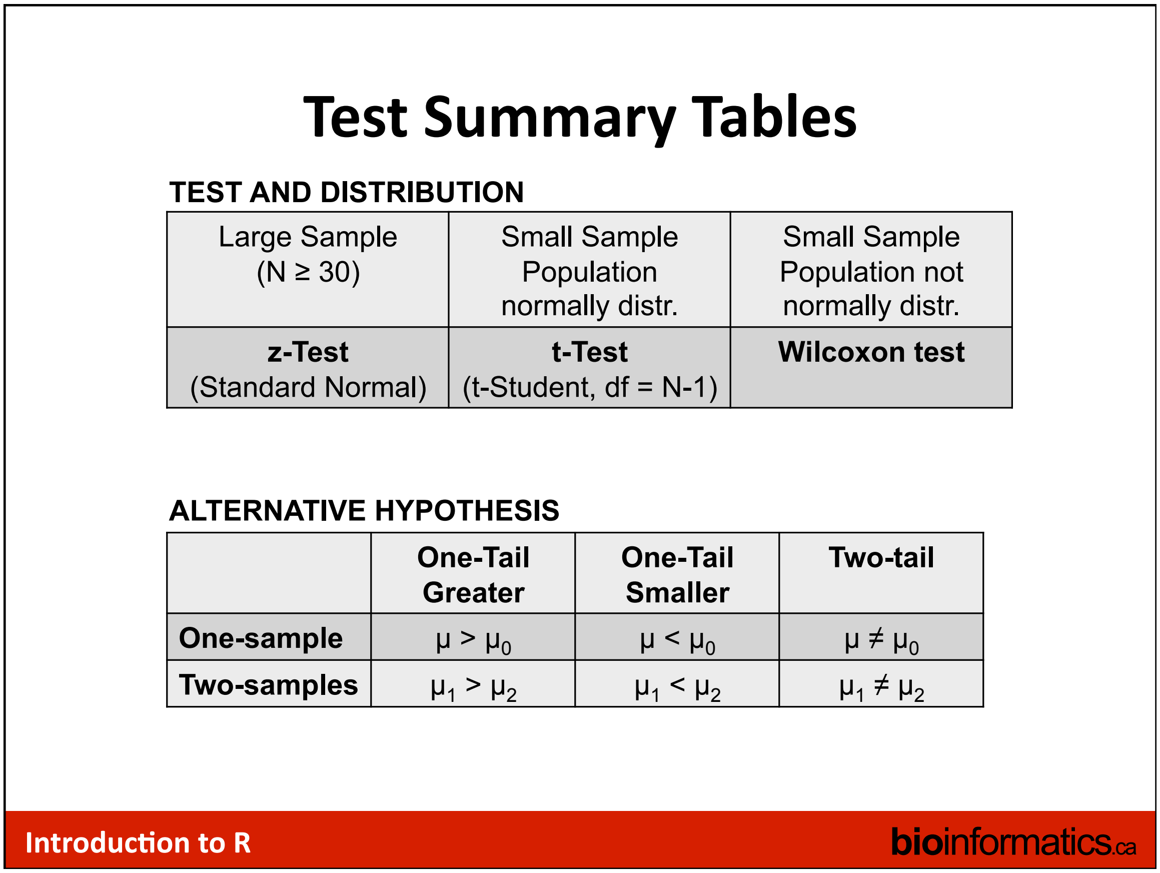
Tests Based on Permutations

- In the previous tests we have always tested the difference of means
  - between populations,
    but using limited knowledge from samples
- Thanks to the central limit theorem, we knew how the sampling mean is supposed to be distributed
  - normal or t-student, depending on sample size
- What if we are, but we don't know how the sampling distribution?



Tests Based on Permutations

- A common approach consists of permuting the class labels
- and computing the count ratio of
  - how many times the difference observed for real data is also observed for permuted data
  - the number of permutations
- The resulting index is called **empirical p-value**



Test Summary Tables

TEST AND DISTRIBUTION

Large Sample ($N \geq 30$)	Small Sample Population normally distr.	Small Sample Population not normally distr.
z-Test (Standard Normal)	**t-Test** (t-Student, df = N-1)	**Wilcoxon test**

ALTERNATIVE HYPOTHESIS

	One-Tail Greater	**One-Tail Smaller**	**Two-tail**
One-sample	$\mu > \mu_0$	$\mu < \mu_0$	$\mu \neq \mu_0$
Two-samples	$\mu_1 > \mu_2$	$\mu_1 < \mu_2$	$\mu_1 \neq \mu_2$

Test Summary Tables

TYPE OF TWO-SMAPLE TEST
(T-TEST OR WILCOXON TEST ALIKE)

Sample units: independent	Sample units: dependent
Not Paired	**Paired**

Other Tests

- **Proportion Test** (Bernoullian Probability)
- **Fisher's Exact Test** (2x2 contingency tables)
- **X^2 Test** (2x2 or larger contingency tables)
- **Kolmogorov-Smirnov** (distribution inequality)
- …

Multiple Testing

- Previously, we have always focused on single tests
- If we test many independent samples from the same population, some of them will lead to the null hypothesis rejection
- However, even if the null hypothesis is TRUE,
  we do expect a rejection rate > 0:
  $M*\alpha$, where M is the number of tests performed
- How to account for this?

Multiple Testing: Bonferroni Correction

- The Bonferroni correction is very conservative:

  after correction, the probability of finding at least one false positive at p-value $\leq \alpha$ will be exactly α
- **p' = MIN (p * M, 1)**

- This correction is usually overly conservative for most genomic applications (e.g. gene expression microarrays)
- It is sometimes recommended for biomarkers and risk factors

Multiple Testing: Benjamini-Hochberg's FDR

- The Benjamini-Hochberg FDR transforms the p-value into a q-value
- Let's consider the q-value q_i,

  that is the false positive rate when considering all tests with $q \leq q_i$

- **q_i = MIN (p_i * M / i, 1)**

  followed by monotonicity correction (i.e. values have to be monotonically increasing)

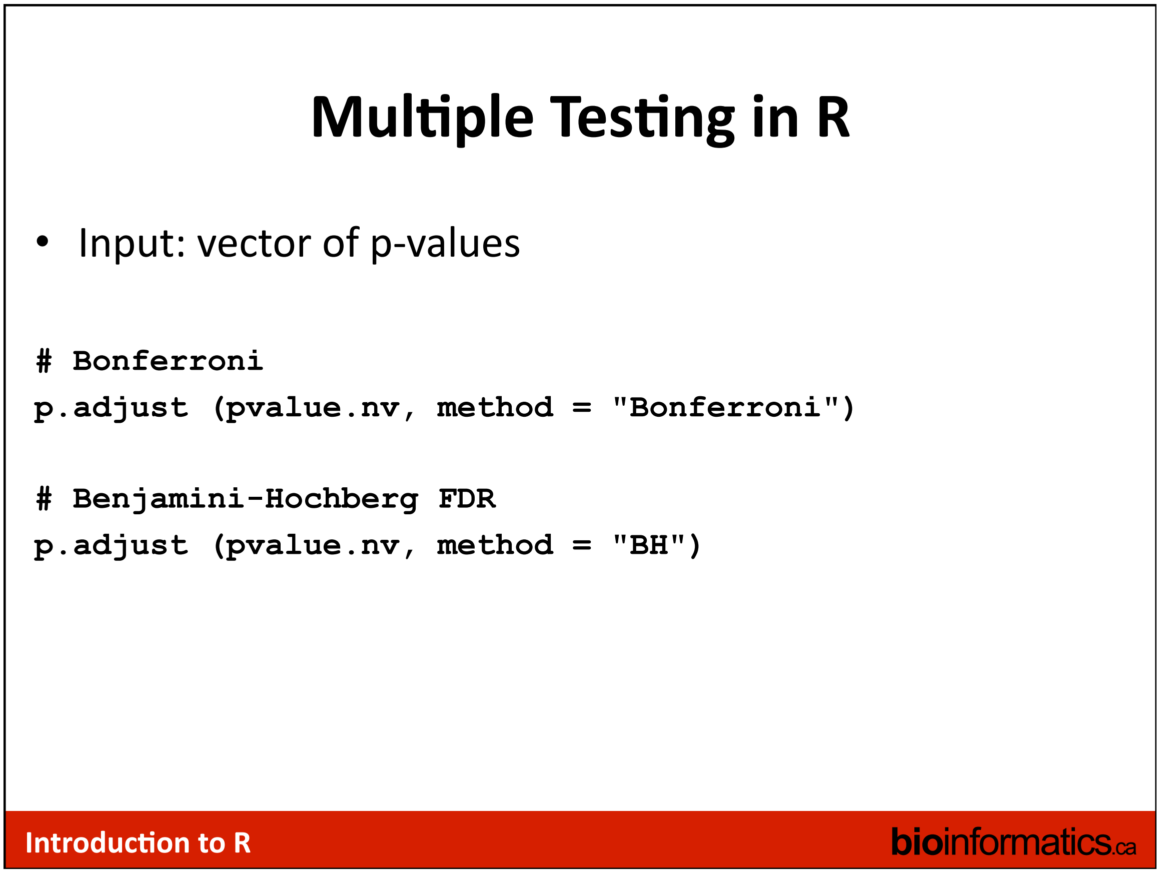
Multiple Testing: Benjamini-Hochberg's FDR

- For each p-value p_i
  - Expected number of false positives if the null hypothesis is true:

    p_i * M   ($\alpha = p_i$)
  - Observed number of positives:

    i   ($p_1, \ldots, p_i \leq \alpha$)
  - Ratio between expected false positives and observed positives:

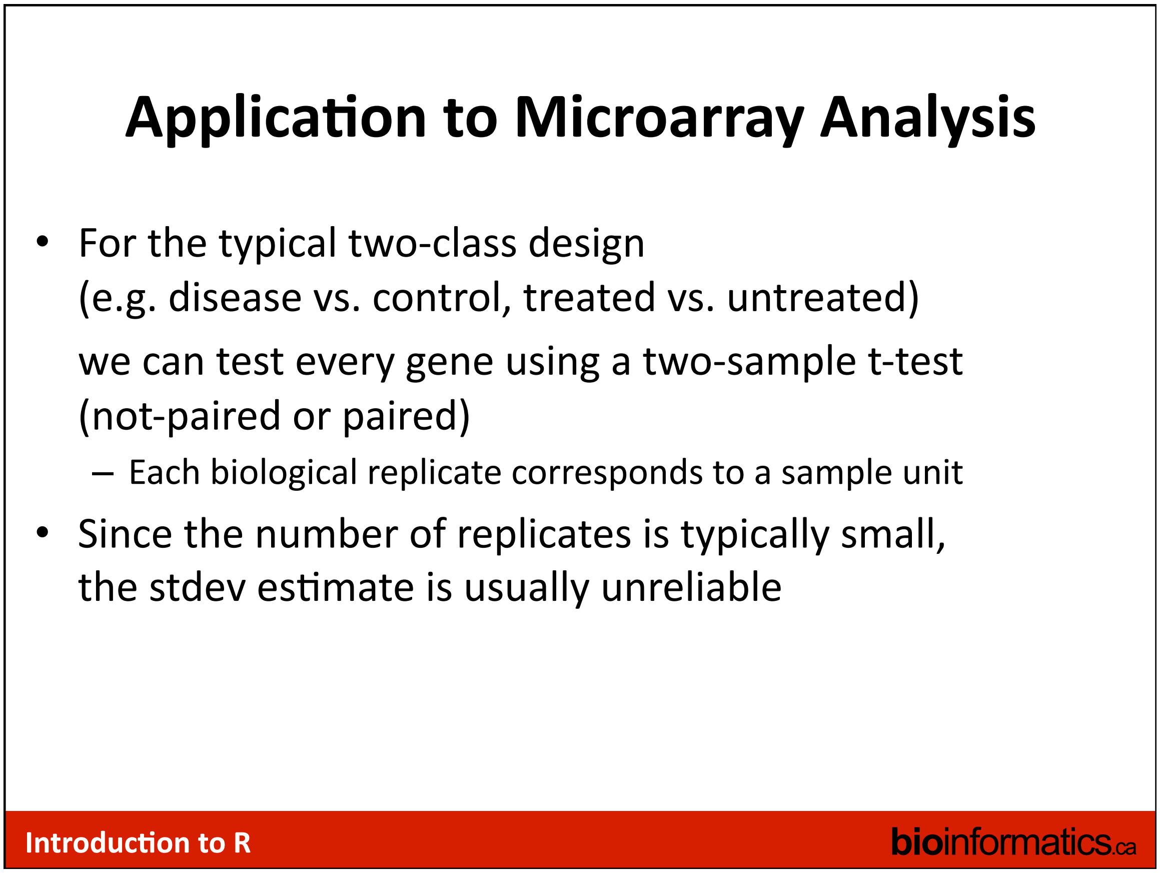
    p_i * M / i



Multiple Testing in R

- Input: vector of p-values

```
# Bonferroni
p.adjust (pvalue.nv, method = "Bonferroni")

# Benjamini-Hochberg FDR
p.adjust (pvalue.nv, method = "BH")
```



Application to Microarray Analysis

- For the typical two-class design
  (e.g. disease vs. control, treated vs. untreated)

  we can test every gene using a two-sample t-test (not-paired or paired)
  - Each biological replicate corresponds to a sample unit
- Since the number of replicates is typically small, the stdev estimate is usually unreliable

Application to Microarray Analysis

- To address the stdev estimation problem, several ***moderated t statistics*** have been introduced
  - Recommended: ***limma*** package

- P-values are usually corrected using Benjamini-Hochberg FDR

We are on a Coffee Break & Networking Session