STAT588/BIOL588: Genomic Data Science
Lecture 5: Review Statistics (part I)

Dr. Yen-Yi Ho (hoyen@stat.sc.edu)

Objectives of Lecture 5

- Introduction to data analysis
- Uncertainty
- Random Variable
- Probability
  - Conditional Probability
- Likelihood
- Maximum Likelihood Estimation
- Law of Large Numbers
- Association between Variables: one continuous variable
  - Two sample Test
  - Permutation Test
  - ANOVA

Data Analysis

Data analysis should begin by examining the types of variables collected in the dataset. We distinguish between numerical and categorical variables.

- Numerical: continous or discrete variable
- Categorical: Nominal or ordinal variable

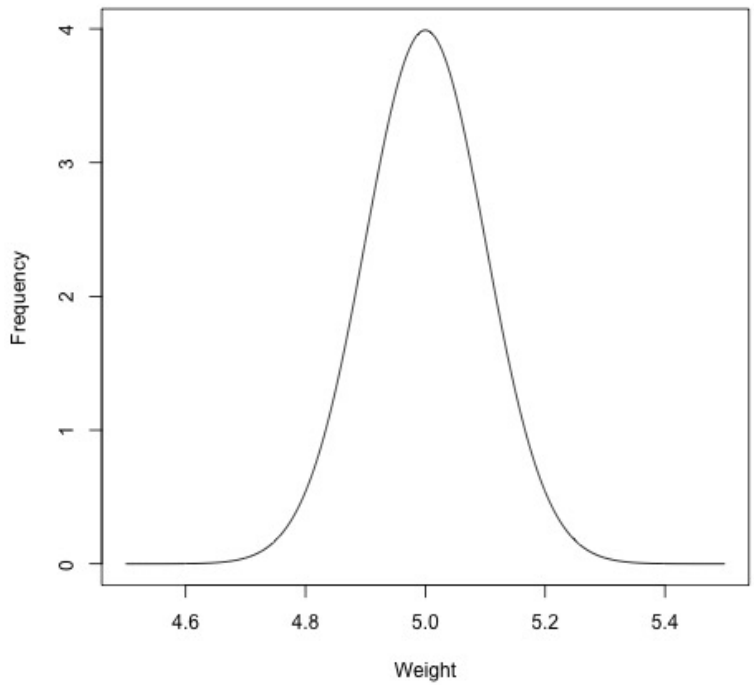
A continuous variable has infinite and uncountable number of possible values while a categorical variable has a finite and countable number of possible values.

In R

```
fmsURL<-"http://people.stat.sc.edu/hoyen/BIOL599/Data/FMS_data.txt"
fms<-read.delim(file=fmsURL, header=TRUE, sep="\t")
colnames(fms)
dim(fms)  ## check the dimension of the data
str(fms[,1:10]) ## check the structure of the data
'data.frame': 428 obs. of  10 variables:
 $ id             : Factor w/ 428 levels "FA-1801","FA-1802",..: 1 2 3 4 5 6 7
 $ acdc_rs1501299 : Factor w/ 3 levels "AA","CA","CC": 2 2 2 3 2 3 3 3 3 3 ...
 $ ace_id         : Factor w/ 3 levels "DD","ID","II": 1 2 2 1 2 2 3 2 2 2 ...
 $ actn3_r577x    : Factor w/ 3 levels "CC","CT","TT": 1 2 2 2 1 2 3 2 2 1 ...
```

Uncertainty

Data is the realization of a random process.



$$\text{weight} = 5lbs \pm 0.1lbs$$

Uncertainty is an interval around the measurement in which repeated measurements will fall.

Random Variables

Random variable: A number assigned to each outcome of a random experiment.

Example 1: I toss a brick at my neighbor's house

D= distance the brick travels

X= 1 if I break a window; 0 otherwise

Y= cost of repair

T= time until the police arrive

N= number of people injured

Example 2: Sample 20 students from the school

H_i = height of student i.

$\overline{H}$= mean of the 20 student heights

S_H = sample deviation of heights

Q: Which of the variables are continuous, which are discrete?

Simulate Random Numbers in R

```
>set.seed(1234)
>runif(10)
[1] 0.113703411 0.622299405 0.609274733 0.623379442
 [5] 0.860915384 0.640310605 0.009495756 0.232550506
 [9] 0.666083758 0.514251141
>rnorm(100)
>rbinom(100, size=1, prob=0.5)
[1] 1 1 0 1 0 1 0 0 0 0 0 0 0 0 0 1 1 1 1 0 0 0 0 1
 [25] 0 1 0 0 1 1 1 1 0 1 0 1 1 0 0 1 0 0 1 1 0 1 0 1
```

Probability

Experiment: a well-defined process with an uncertain outcome.
Example: Draw 2 balls with replacement from an urn containing 4 red and 6 blue balls.

Sample Space (S): The set of all possible outcomes.
{RR, RB, BR, BB}

X: Number of red balls observed in our experiment.
{RB}

Pr(X=1) (Probability can be assigned to outcome event)

Probability Rules

$0 \leq \Pr(A) \leq 1$	for any event A
$\Pr(S)=1$	where S is the sample space
$\Pr(A \text{ or } B) = \Pr(A) + \Pr(B)$	if A and B are mutually exclusive
$\Pr(\text{not } A) = 1 - \Pr(A)$	complement rule

Independence

Two events are independent if

$$\mathrm{P}(\mathrm{A \text{ and } B}) = \mathrm{P}(\mathrm{A}) \times \mathrm{P}(\mathrm{B}).$$

Example 1: flip a coin and draw a card from a random deck

$$\Pr\{\textit{head} \text{ and } \spadesuit A\} = \frac{1}{2} \times \frac{1}{52}$$

Example 2: Genotype at a autosomal SNP locus with two alleles, A and a, from a pair of randomly selected chromosomes.
Events: {genotype AA}, {genotype Aa}, {genotype aa}

Let p_A be the allele frequency of A allele, and assume independence

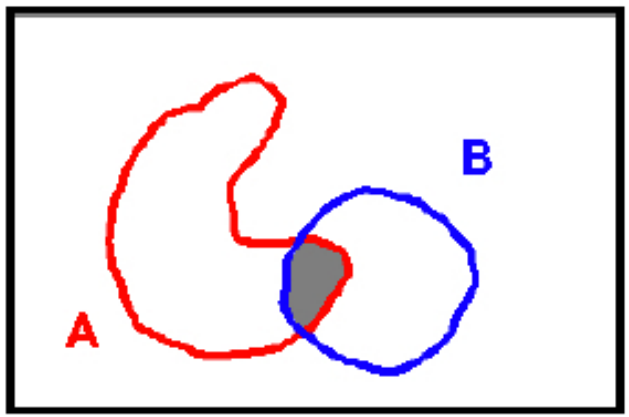
Genotype	AA	Aa	aa
Probability	p_A^2	$2p_A(1-p_A)$	$(1-p_A)^2$

Hardy-Weinberg equilibrium: independence of alleles across two homologous chromosomes.

Conditional Probability

$$\begin{aligned} \Pr(A \mid B) &= \text{Probability of A given B} \\ &= \frac{\Pr(A \text{ and } B)}{\Pr(B).} \end{aligned}$$

If A and B are independent, $\Pr(A \mid B) = \Pr(A)$.



Probability

What is the probability of obtaining a head and a tail tossing a fair coin twice? Let X be the random variable denoting the number of heads.

$$\Pr(X = 1) = \binom{2}{1} \times (\frac{1}{2}) \times (\frac{1}{2}) = 0.5$$

In R

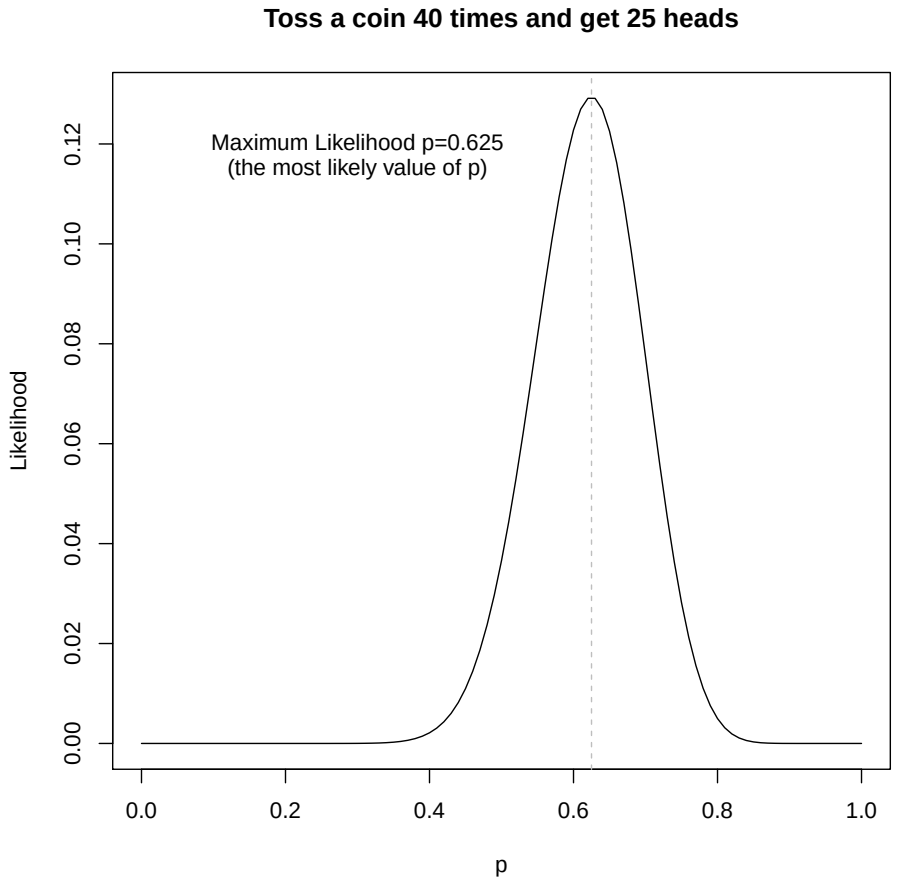
```
> dbinom(1, 2, 0.5)
[1] 0.5
> rbinom(1, 2, 0.5)  ### toss a fair coin twice
[1] 0
```

Likelihood

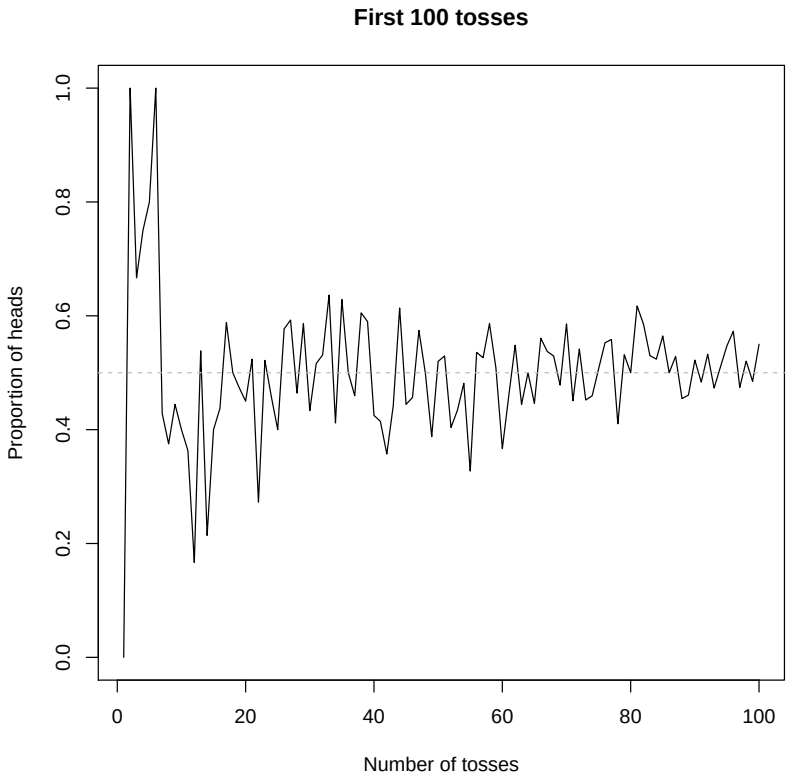
The likelihood is the probability of observing the data.

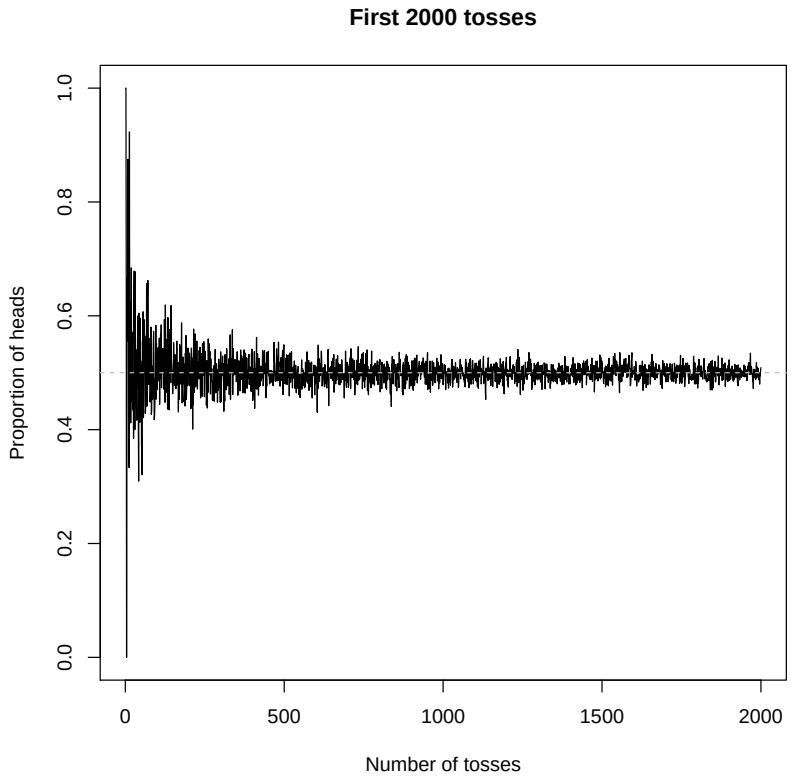
What is the likelihood of tossing a coin 40 times and get 25 heads?

$$\text{Likelihood}(\text{data}|p) = \binom{40}{25} \times p^{25} \times (1-p)^{15} = dbinom(25, 40, p)$$



The Law of Large Numbers





Toss a fair coin a lot of times . . .

Association between Variables

		Independent Categorical	Variable Continuous
Outcome Variable	Continuous	T-Test, ANOVA (A)	Regression (C)
	Categorical	χ^2, Fisher (B)	GLM (D)

- Difference in gene expression in patients with/without a mutation (yes/no):

Association between Variables

		Independent Categorical	Variable Continuous
Outcome Variable	Continuous	T-Test, ANOVA (A)	Regression (C)
	Categorical	χ^2, Fisher (B)	GLM (D)

- Difference in gene expression in patients with/without a mutation (yes/no): A
- Determine the association between disease Status (yes/no) and genotype (AA, Aa, aa):

Association between Variables

		Independent Categorical	Variable Continuous
Outcome Variable	Continuous	T-Test, ANOVA (A)	Regression (C)
	Categorical	χ^2, Fisher (B)	GLM (D)

- Difference in gene expression in patients with/without a mutation (yes/no): A
- Determine the association between disease Status (yes/no) and genotype (AA, Aa, aa): B
- Predict father's height from daughter's height:

Association between Variables

		Independent Categorical	Variable Continuous
Outcome Variable	Continuous	T-Test, ANOVA (A)	Regression (C)
	Categorical	χ^2, Fisher (B)	GLM (D)

- Difference in gene expression in patients with/without a mutation (yes/no): A
- Determine the association between disease Status (yes/no) and genotype (AA, Aa, aa): B
- Predict father's height from daughter's height: C
- Determine the relationship between smoking status (yes/no) and lung cancer (yes/no):

Association between Variables

		Independent Variable Categorical	Variable Continuous
Outcome Variable	Continuous	T-Test, ANOVA (A)	Regression (C)
	Categorical	χ^2, Fisher (B)	GLM (D)

- Difference in gene expression in patients with/without a mutation (yes/no): A
- Determine the association between disease Status (yes/no) and genotype (AA, Aa, aa): B
- Predict father's height from daughter's height: C
- Determine the relationship between smoking status (yes/no) and lung cancer (yes/no): B

The ALL Dataset

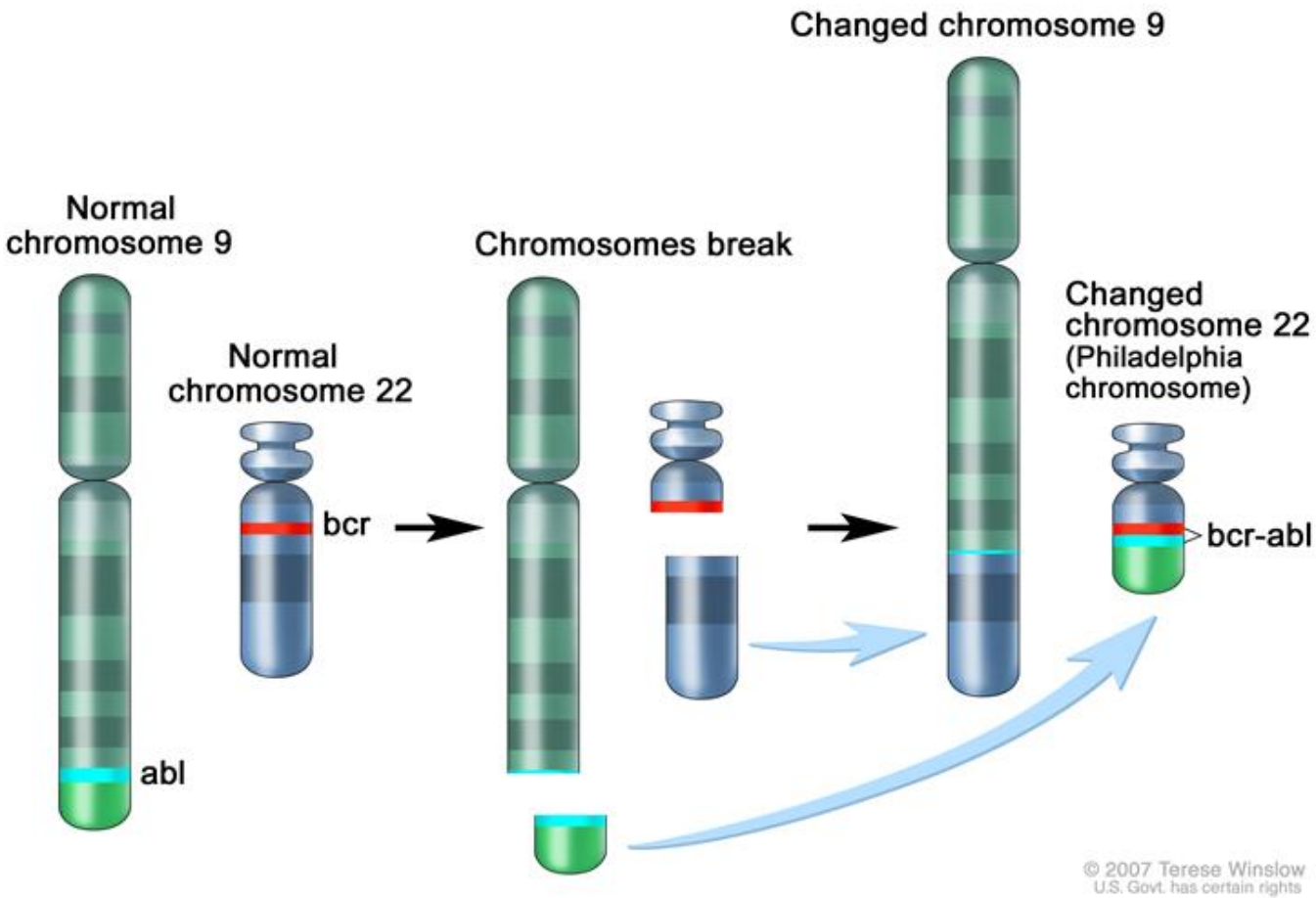
- Microarrays data with 12,625 gene expression probes (features) from 128 individuals with acute lymphoblastic leukemia (ALL).
- individual specific covariates: gender, age, tumor type and stage, translocation mutationÉ

	01005	01010	03002	04006	04007
1000_at	7.60	7.48	7.57	7.38	7.91
1001_at	5.05	4.93	4.80	4.92	4.84
1002_f_at	3.90	4.21	3.89	4.21	3.42
1003_s_at	5.90	6.17	5.86	6.12	5.69
1004_at	5.93	5.91	5.89	6.17	5.62

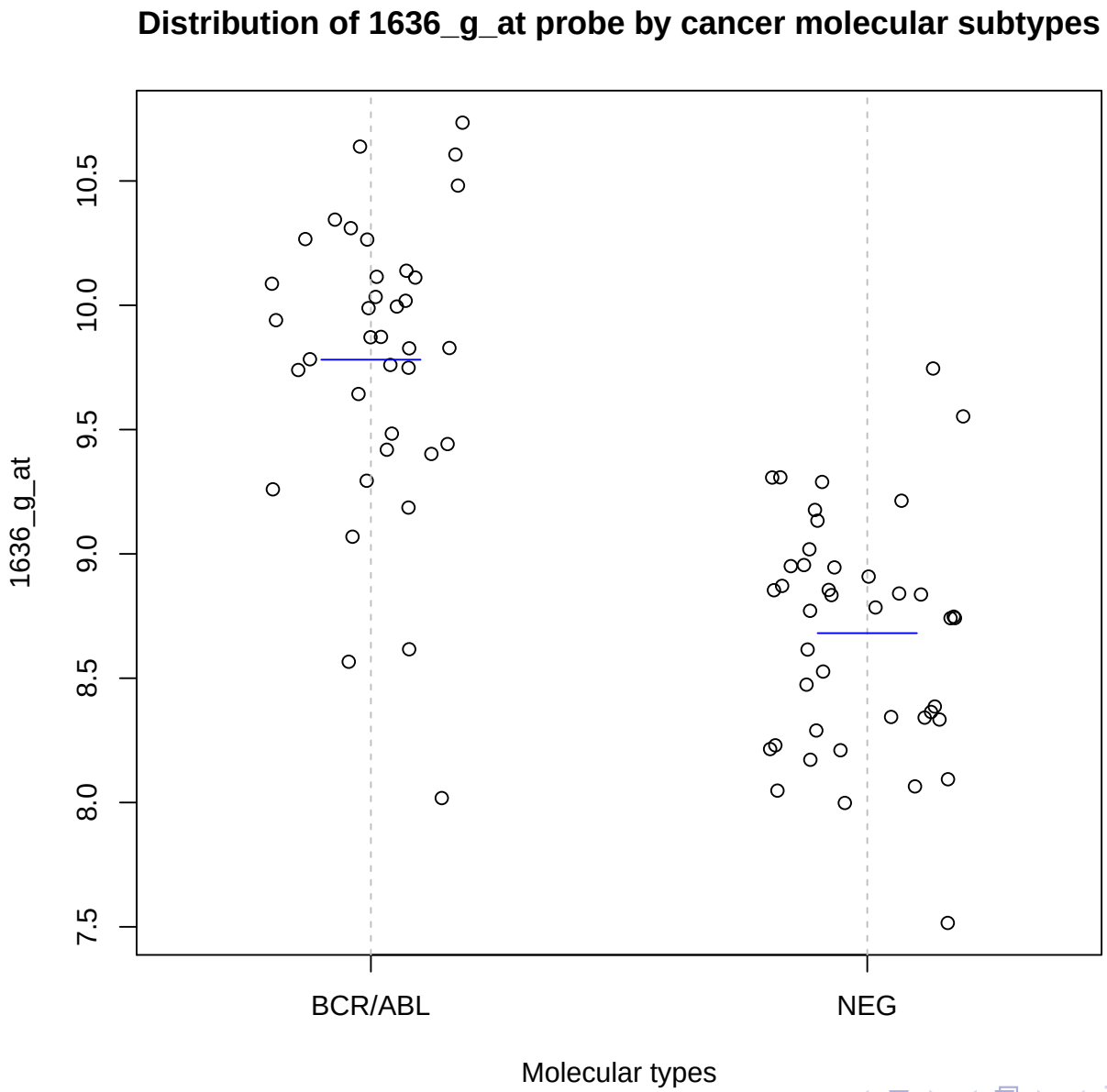
Access the ALL dataset

```
>source("http://www.bioconductor.org/biocLite.R")
>biocLite("ALL")
>library("ALL")
>data("ALL")
>table(ALL$BT)
B  B1  B2  B3  B4   T  T1  T2  T3  T4
 5  19  36  23  12   5   1  15  10   2
>table(ALL$mol.biol)
ALL1/AF4  BCR/ABL E2A/PBX1       NEG   NUP-98  p15/p16
      10       37        5        74        1        1
```

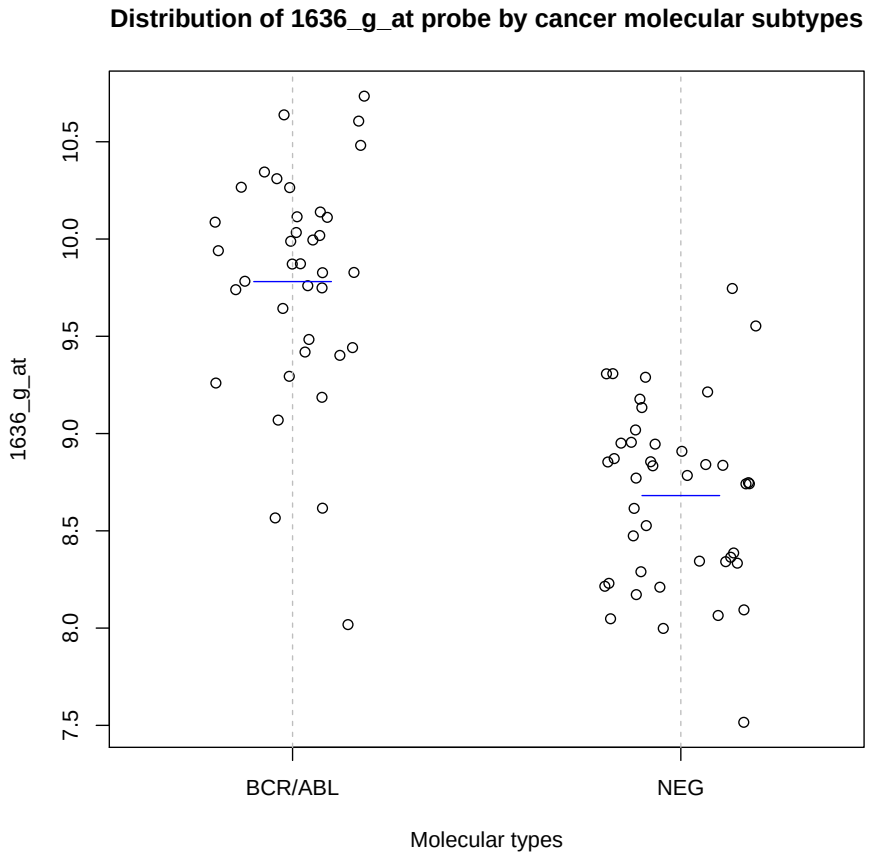
Philadelphia Chromosome



Gene Expression Example (ABL1 gene)

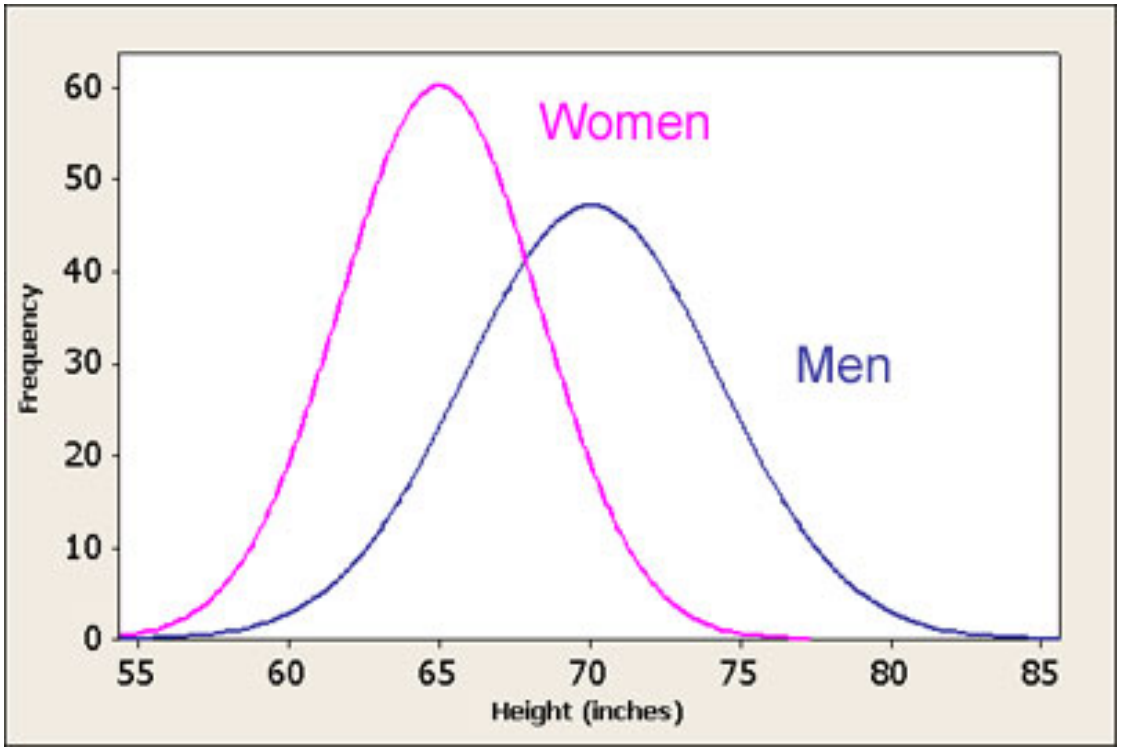


Gene Expression Example (ALL Data)



- Is this difference worth reporting?
- Some journal requires statistical significance. What does it mean?

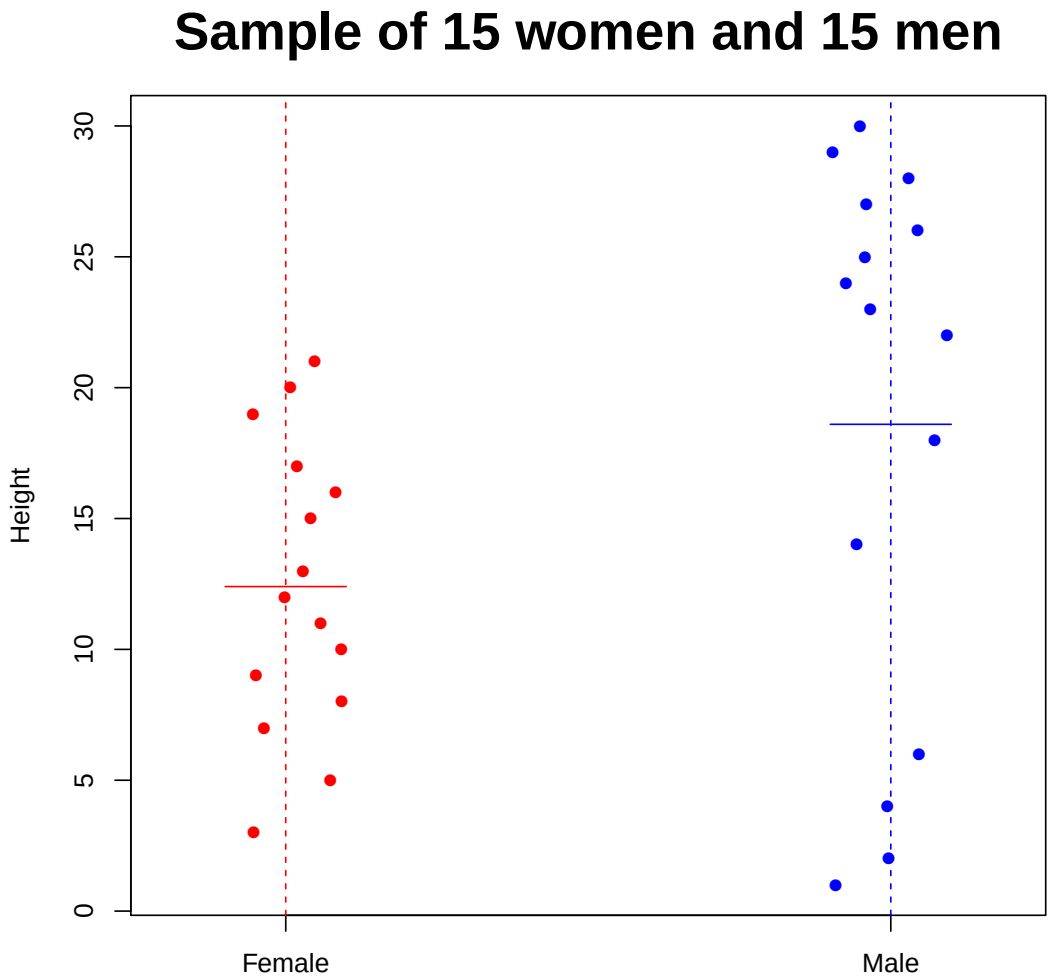
Men are taller than women



This statement refers to population averages: the population average of men's height is larger than the population average of women

One Data Point

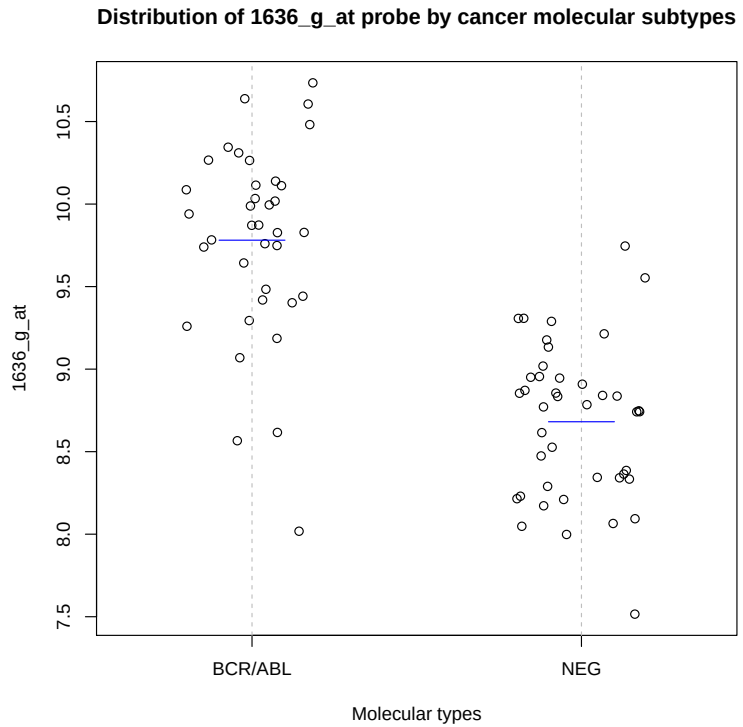
Law of Large Numbers



Hypothesis Testing

Test of hypothesis: answer a yes, or no question regarding a population parameter.

Example: Does the ABL1 (measured by 1636_g_at) gene expression from the two molecular groups (BCR/ABL vs. NEG) have the same population mean?



Two Sample T-Test

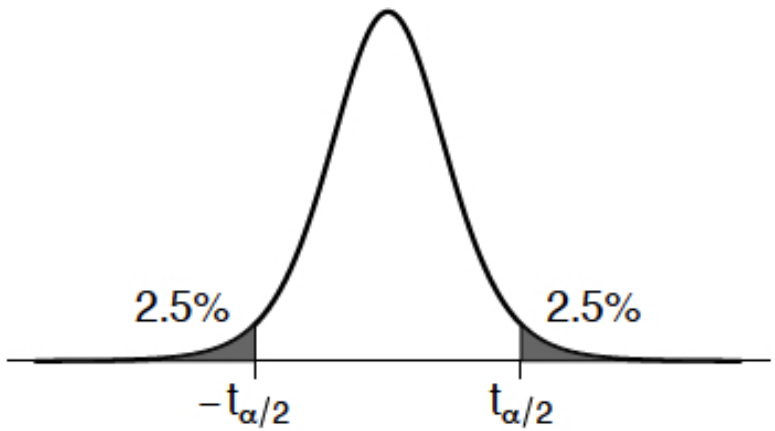
$$H_0 : \mu_1 = \mu_2$$

versus

$$H_a : \mu_1 \neq \mu_2$$

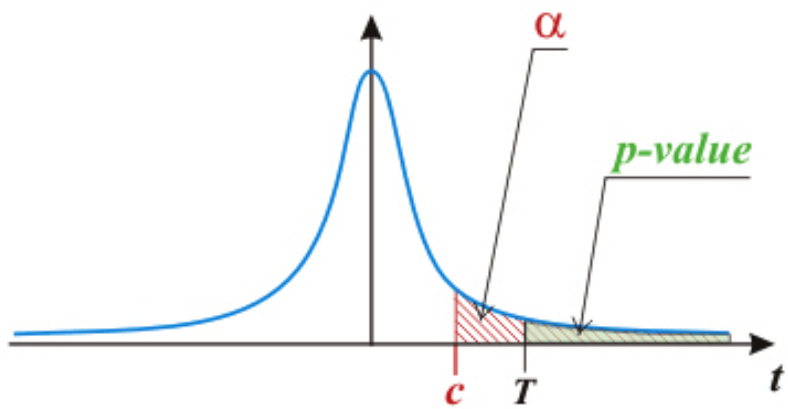
$$\text{Test Statistic: T} = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \text{ (signal to noise ratio)}$$

Reject H_0, if $|T| > t_{\alpha/2,k}$



p value

$$\text{Test Statistic: } T = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \text{ (signal to noise ratio)}$$



p value: the probability of observing a test statistic more extreme as the one that was actually observed under the null distribution.

Two Sample T-Test

- When reject H_0:
  - The difference is statistically significant.
  - The observed difference can not be explained by chance variation.
- When fail to reject H_0:
  - The difference is not statistically significant.
  - There is insufficient evidence to conclude that $\mu_1 \neq \mu_2$
  - The observed difference could reasonably be the result of chance variation.

Two Sample T-Test

```
>g1<- data[whp, ALL_bcrneg$mol.biol==‘‘BCR/ABL"]
>g2 <- data[whp,ALL_bcrneg$mol.biol==‘‘NEG"]
>t.test(g1, g2)
```

Welch Two Sample t-test

data: g1 and g2

t = 9.1304, df = 68.717, p-value = 1.792e-13

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

0.8596467 1.3403765

sample estimates:

mean of x mean of y

9.781236 8.681225

Wilcoxon Rank-Sum Test (Nonparametric Test)

Small sample setting when normality assumption is not reasonable

> wilcox.test(g1,g2)

Wilcoxon rank sum test
data: g1 and g2
W = 1432, p-value = 8.306e-13
alternative hypothesis: true location shift is not equal to 0

Permutation

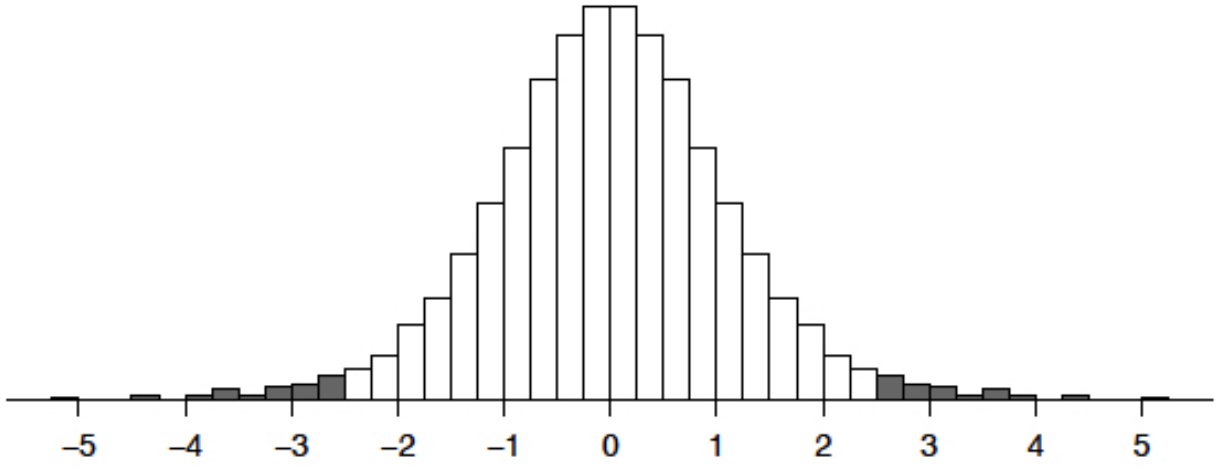
Null distribution: Distribution of the test statistic when the null hypothesis is true.

Idea: generate the null distribution by random shuffling group label.

Data	T	T	T	T	T	C	C	C	C	C
	5.4	6.2	3.8	4.4	3.3	8.1	8	7.2	7.8	7.9

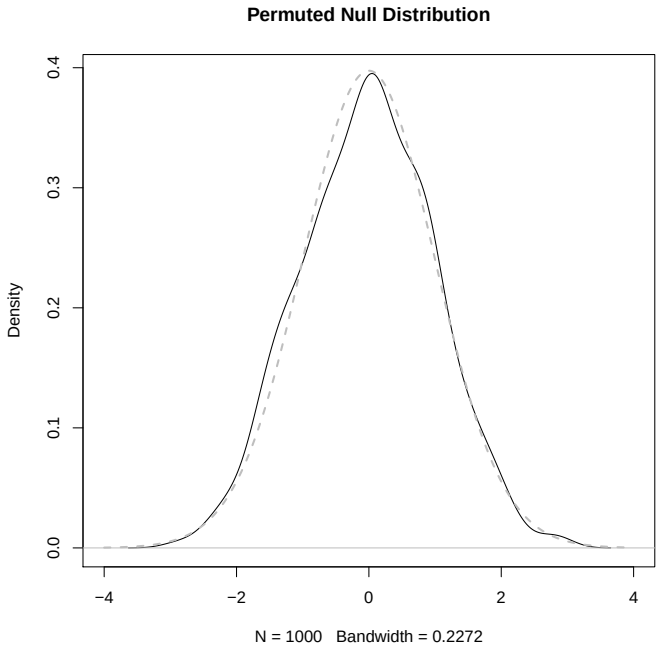
Permutate	T	C	C	C	T	T	T	C	T	C
	5.4	6.2	3.8	4.4	3.3	8.1	8	7.2	7.8	7.9

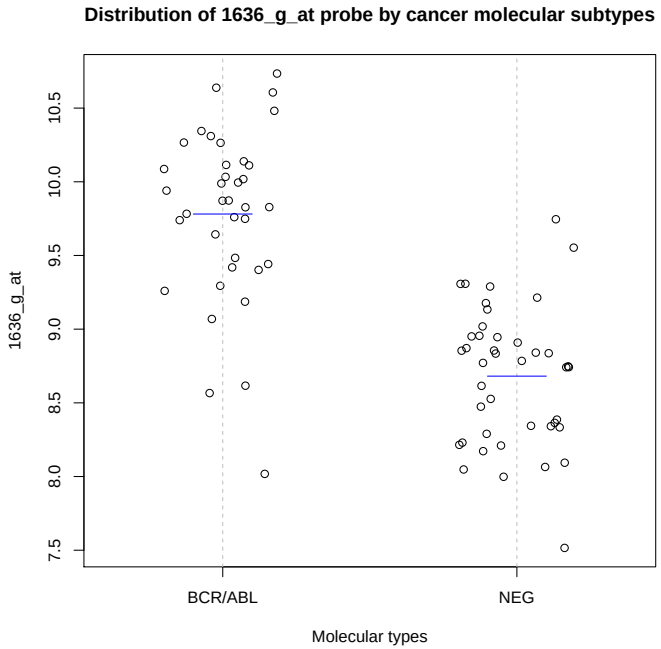
Randomly assign the group labels $\rightarrow T^*$



P-value $\Pr(|T^*| \geq |T_{obs}|)$

Permutation Test





Permutation Test is A Good Friend

Good: Do not assume distribution for the test statistic
Bad: Computational intense (longer computation time)

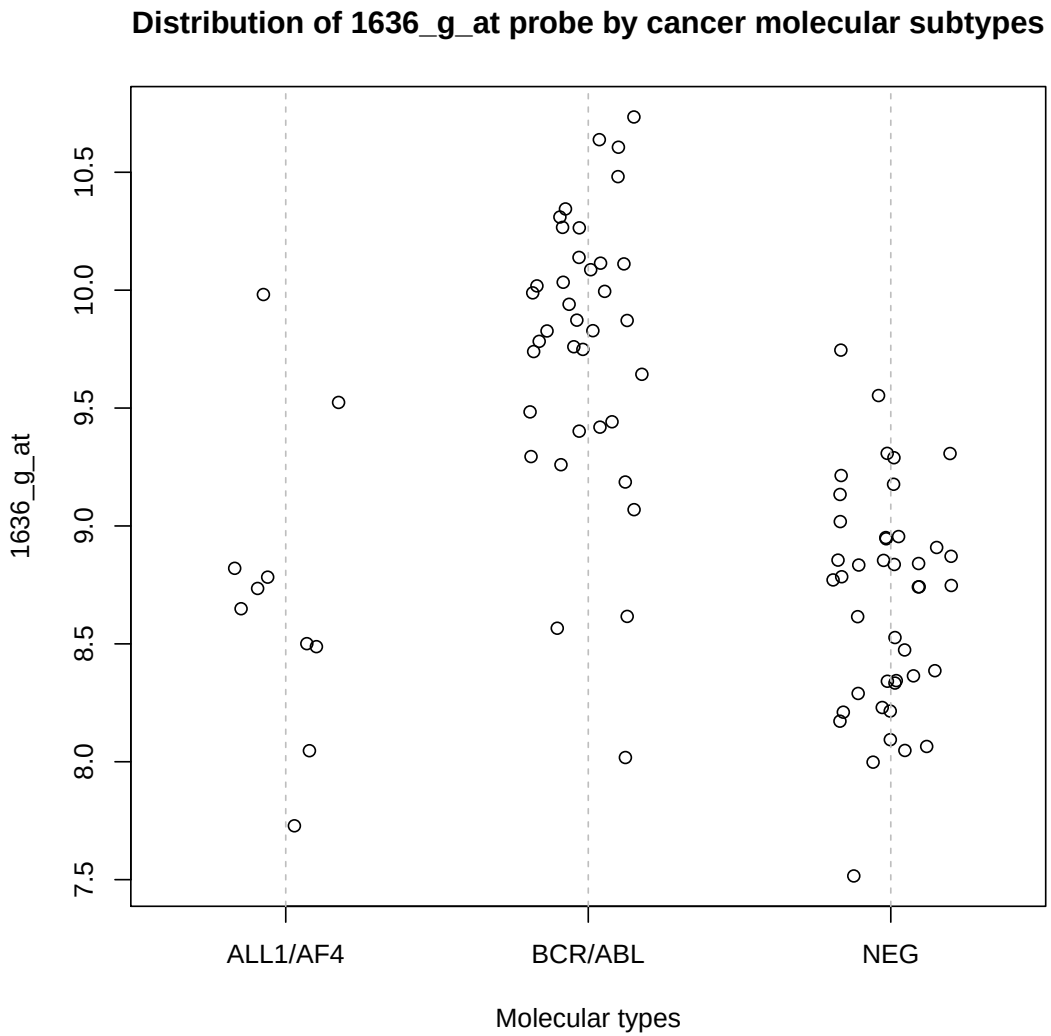
What to Use

The t-test relies on a normality assumption. When sample size is small, consider:

- Wilcoxon Rank Sum Test
- Permutation Test

$\rightarrow$ The crucial assumption is independence between observations.

Multiple Groups Comparison



Multiple groups comparison: Hypothesis

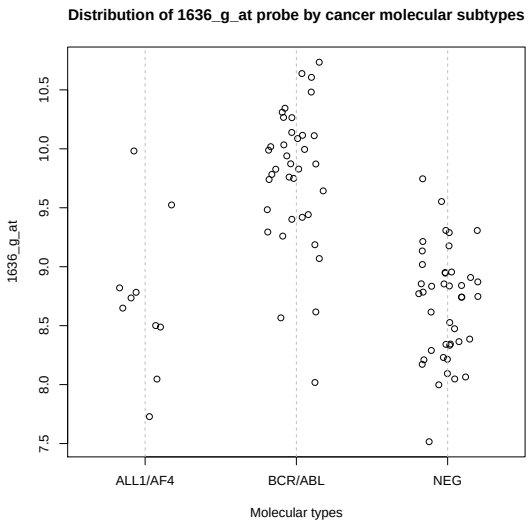
Are there differences in the means of gene expression among the three molecular groups (ALL1/AF4, BCR/ABL, NEG) ?

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 = \mu_3, \\ H_a &: H_0 \text{ is false.} \end{aligned}$$

Two Sample T Test

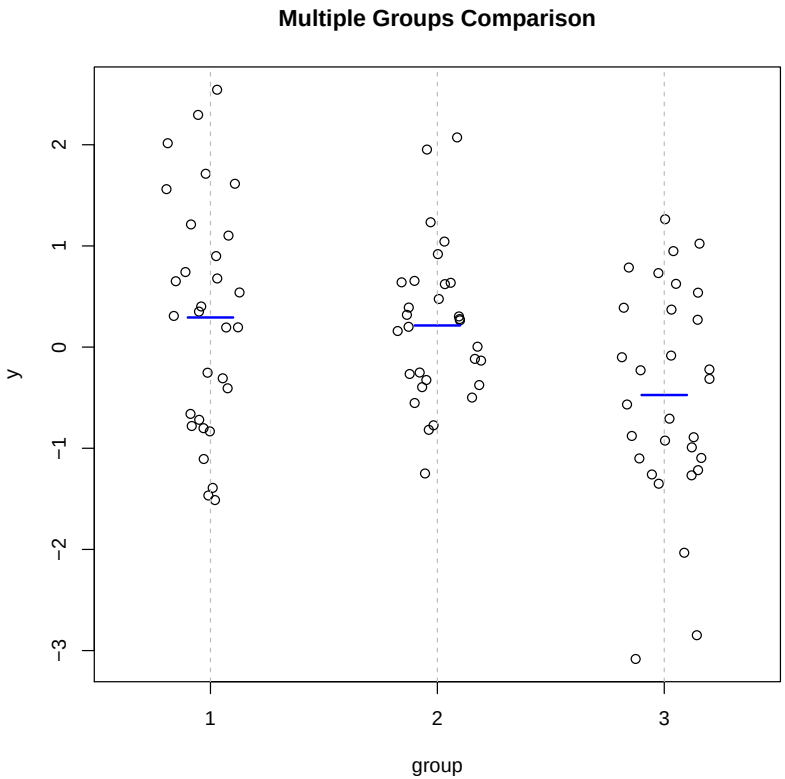
$$\text{Two Sample Test Statistic: } T = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \text{ (signal to noise ratio)}$$

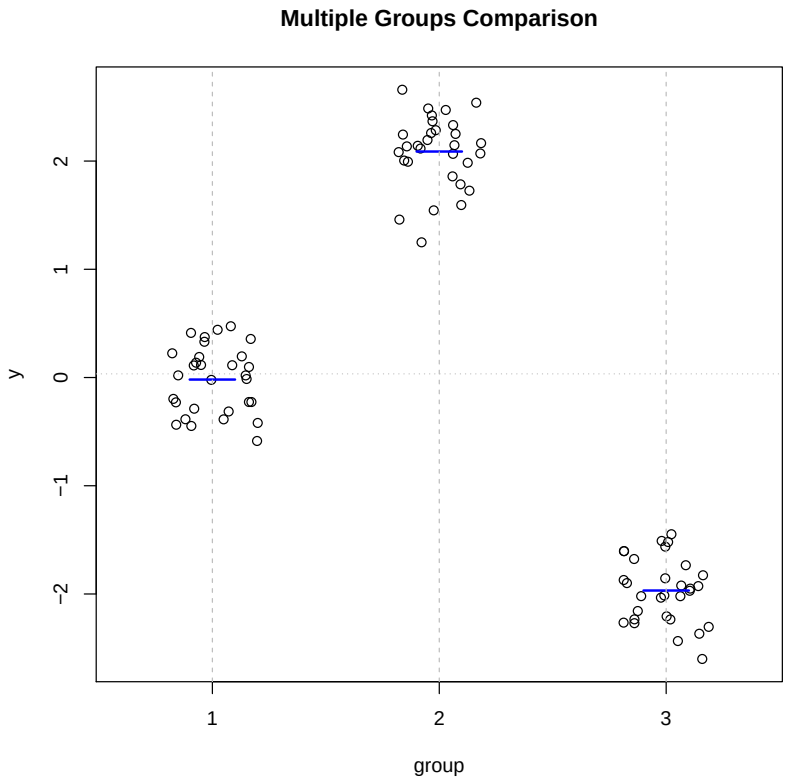
$$\text{Three Samples } F = \frac{(\overline{X_1} - \overline{X_2}) + (\overline{X_2} - \overline{X_3}) + (\overline{X_1} - \overline{X_3})}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} + \frac{s_3^2}{n_3}}} \text{ (Is this a good idea?)}$$



ANOVA (ANalysis Of VAriance)

Grouping variable is important if there is large between group variation, and small within group variation.





ANOVA: Gene Expression Example

>summary(aov(all[whs,] ~ ALL3$mol.biol))

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
ALL3$mol.biol	2	25.77	12.88	44.04	0.0000
Residuals	86	25.16	0.29