Statistics for Human Genetics and Molecular Biology
Lecture 3: Some Statistical Tools

Dr. Yen-Yi Ho (yho@umn.edu)

Sep 16, 2015

Objectives of Lecture 3

- Continuous Data
  - Summarizing and Presenting Continuous Data
  - Two sample Test
  - Permutation Test
- Categorical Data
  - Tabulating and Plotting Categorical Data
  - Test for Contingency Tables
  - Cochran-Armitage Trend Test

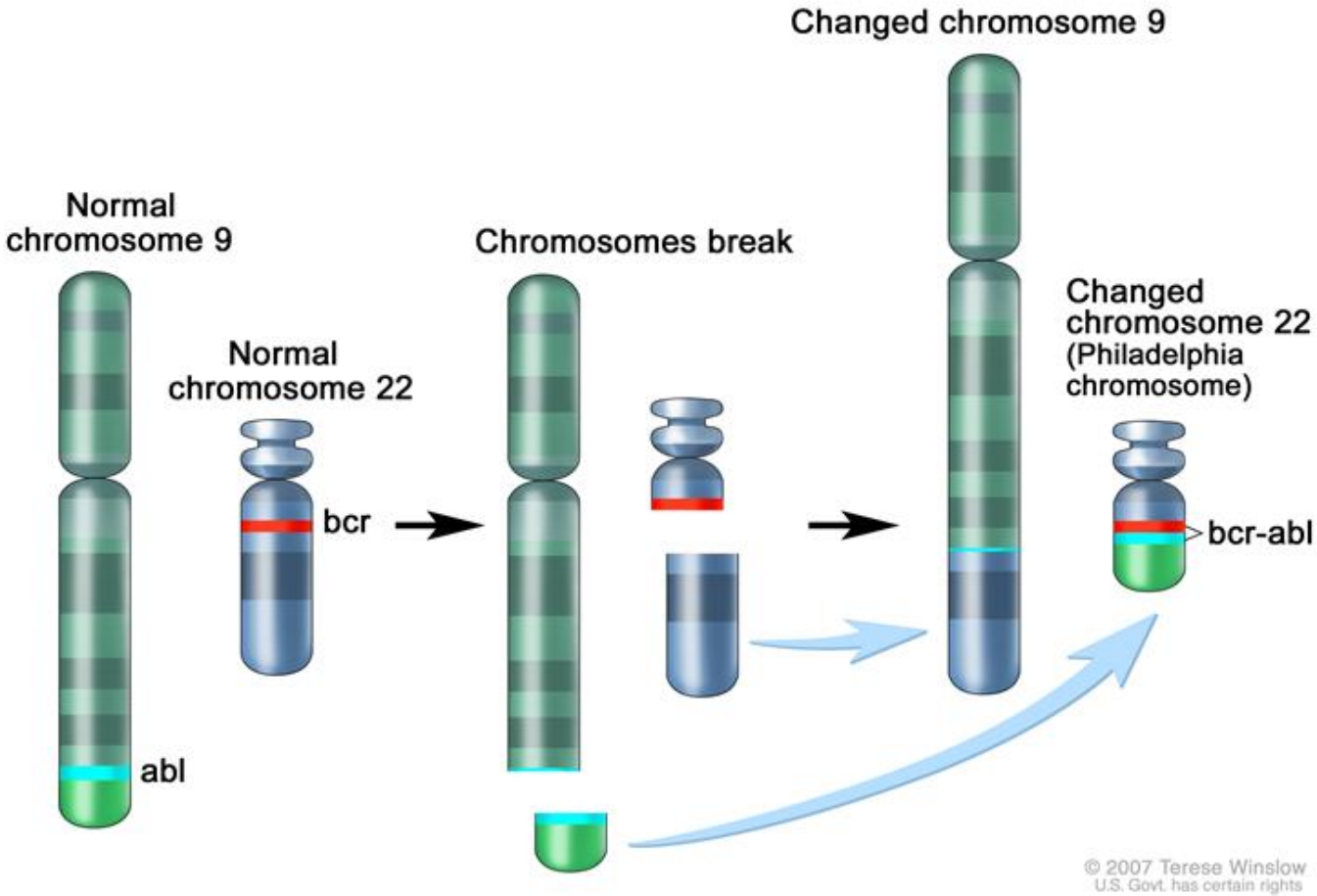
Summarizing and Presenting Continuous Data

The ALL Dataset

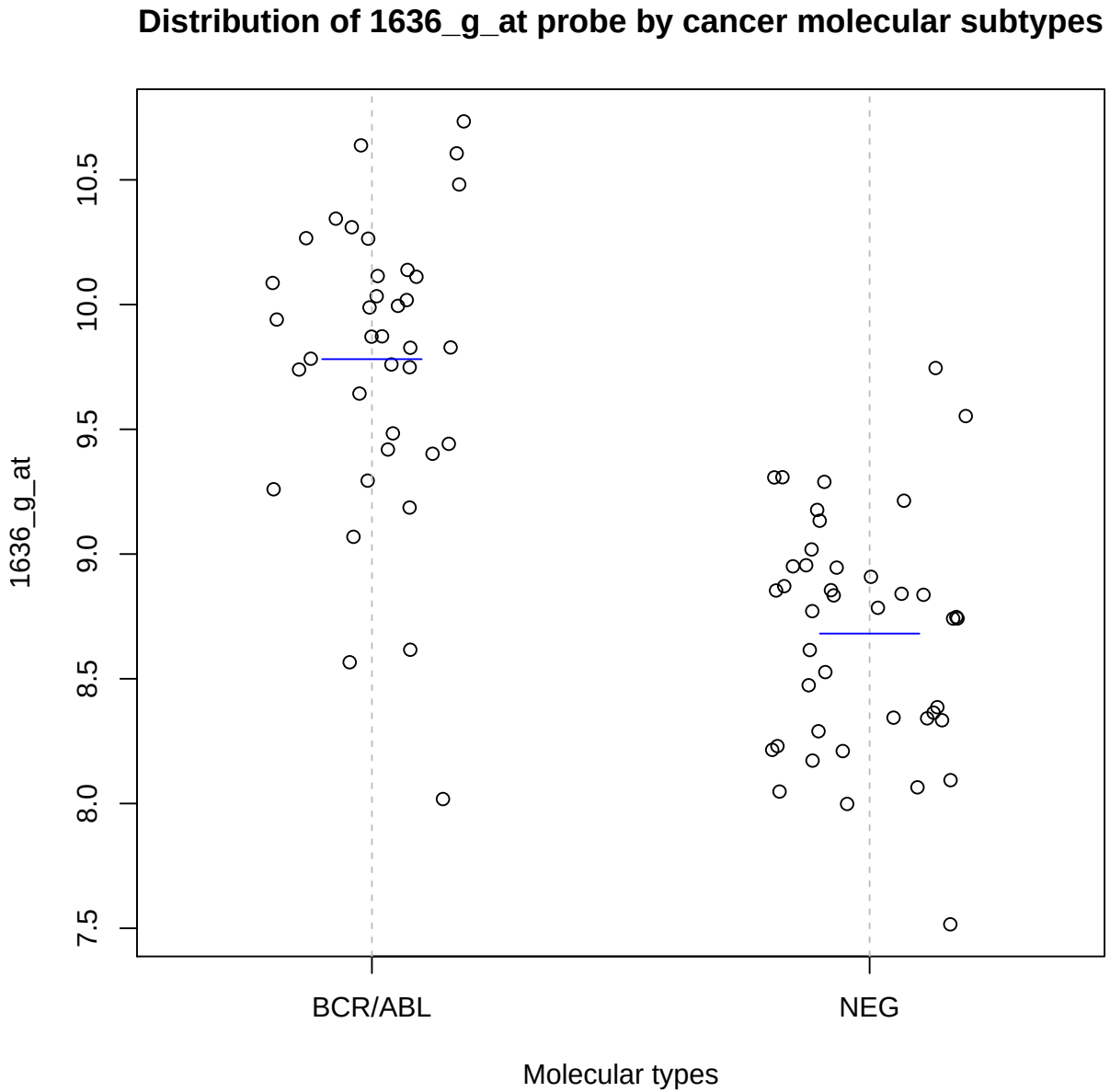
- Microarrays data with 12,625 gene expression probes (features) from 128 individuals with acute lymphoblastic leukemia (ALL).
- individual specific covariates: gender, age, tumor type and stage, translocation mutations (Philadelphia chromosome), molecular types, ...

	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

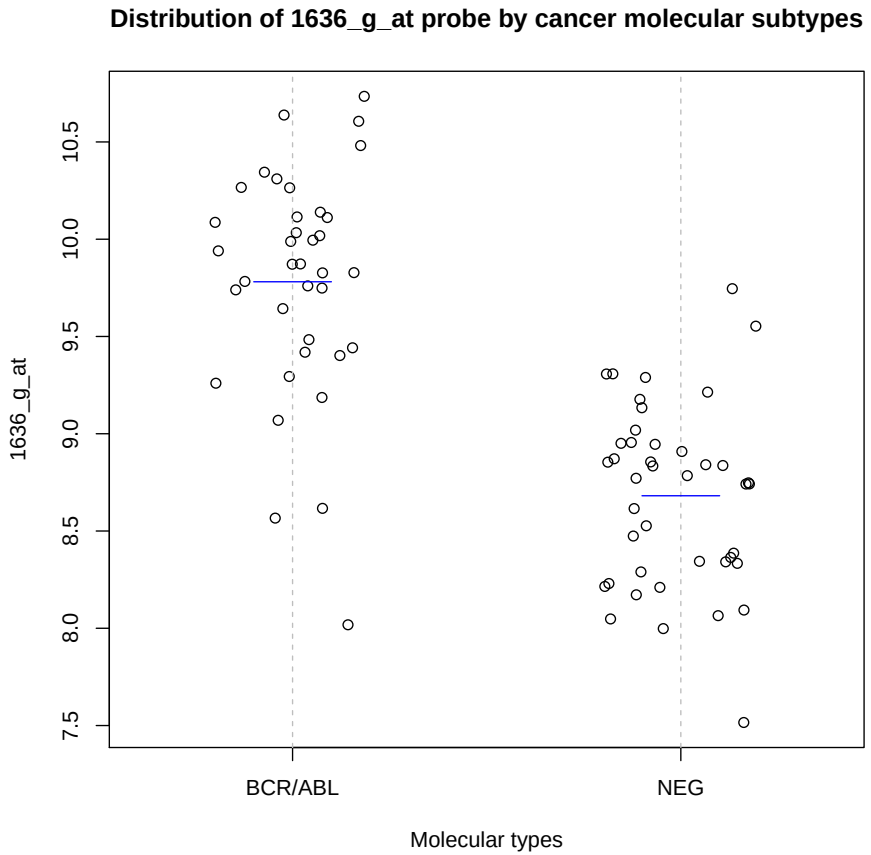
Philadelphia Chromosome



Gene Expression Example

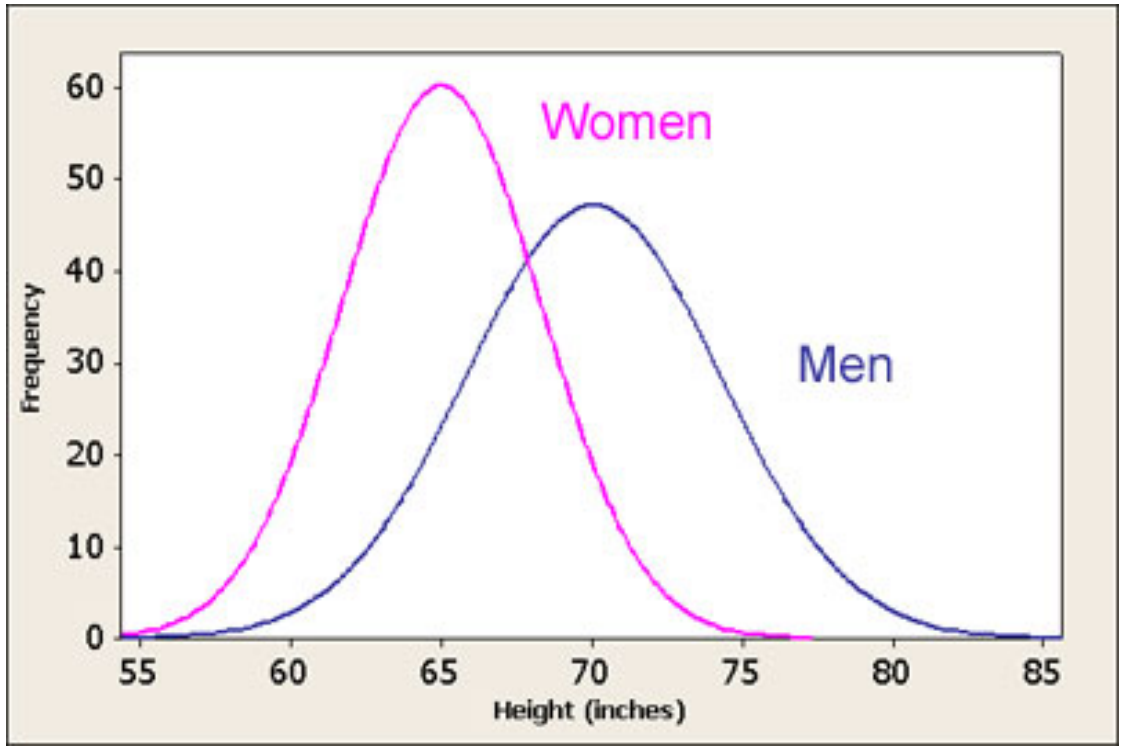


Gene Expression Example (ALL Data)



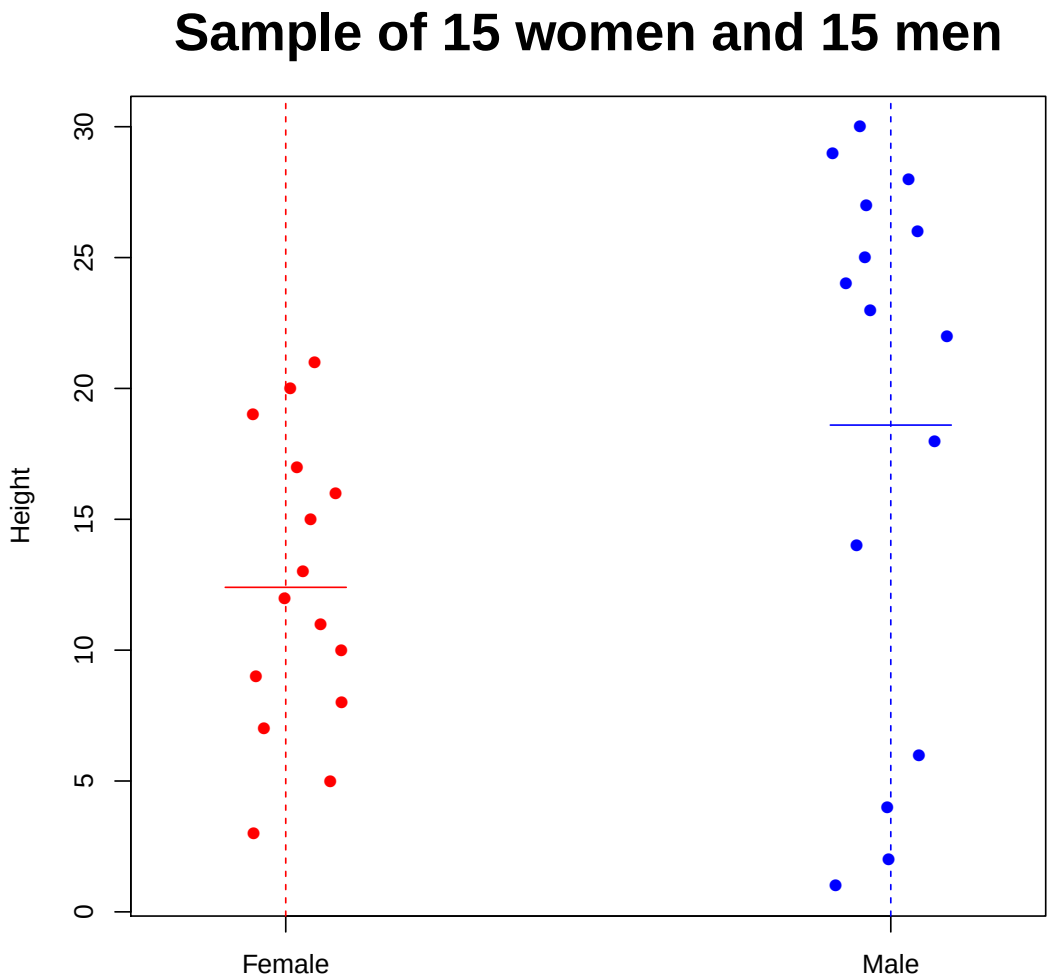
- Is this difference worth reporting?
- Some journal requires statistical significance. What does it mean?

Men are taller than women

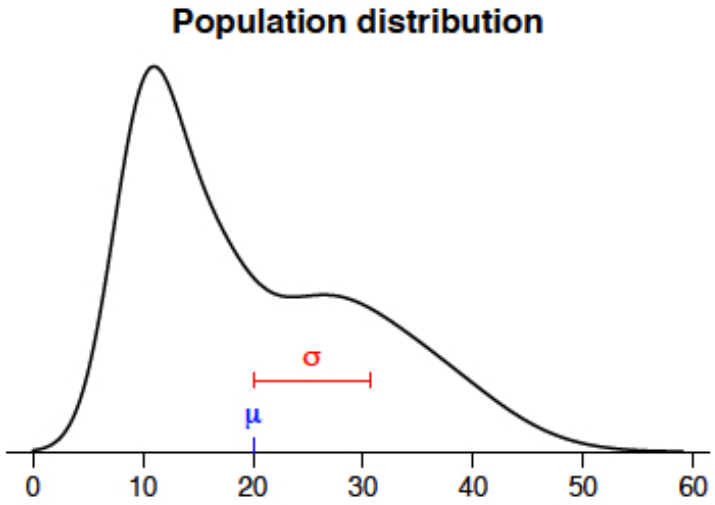


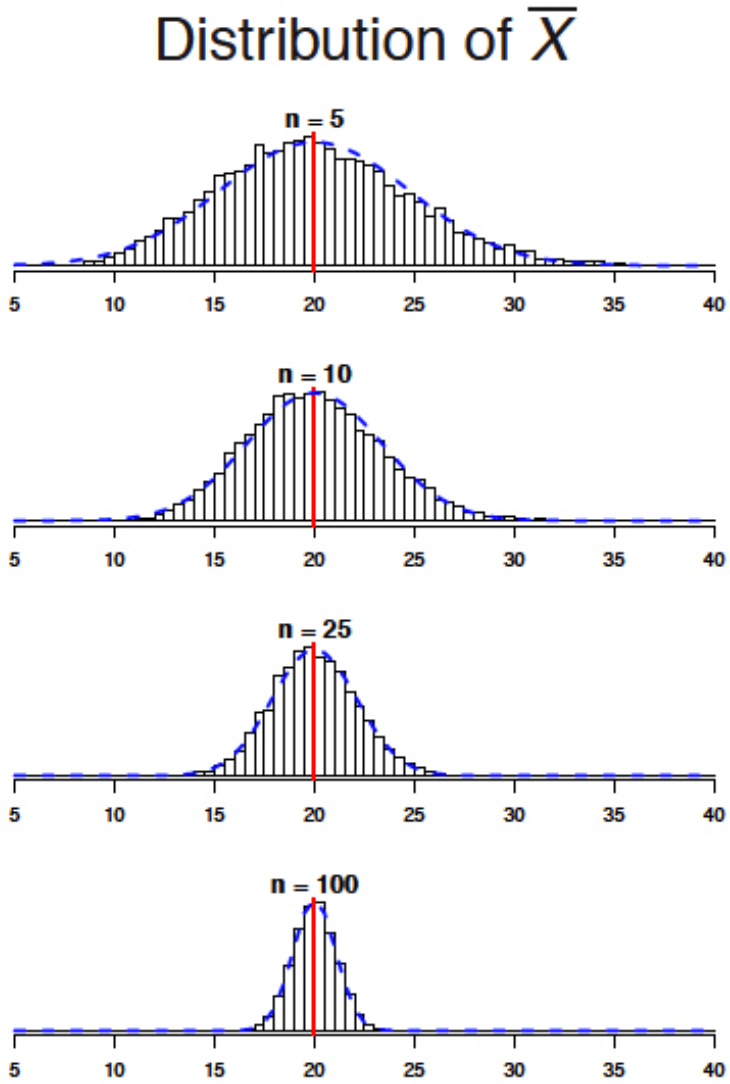
This statement refers to population averages: the population average of men's height is larger than the population average of women

One Data Point


Sample of 15 women and 15 men
Height
0
5
10
15
20
25
30
Female
Male

Sampling Distribution of Means





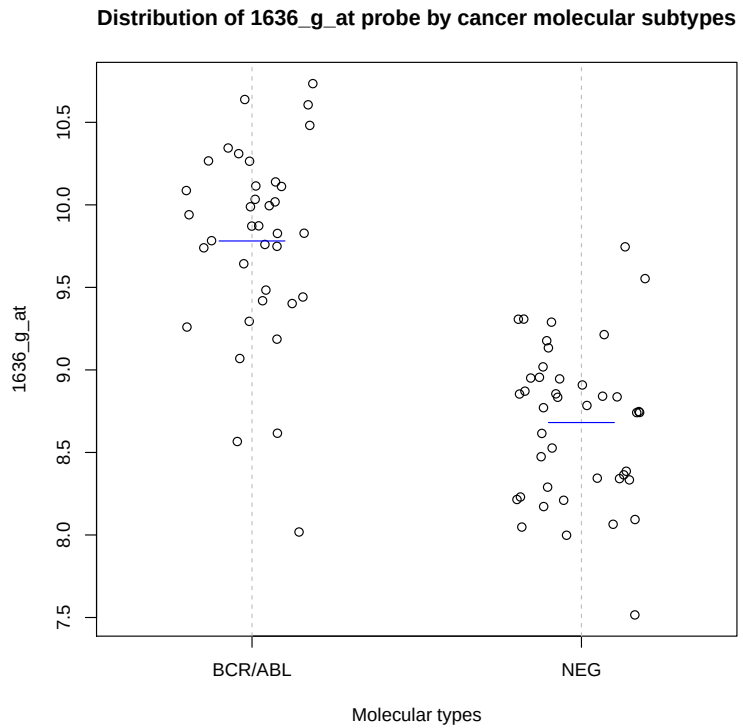
The sampling distribution depends on:

- The type of statistic
- The population distribution
- The sample size

Hypothesis Testing

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

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



Two Sample T-Test

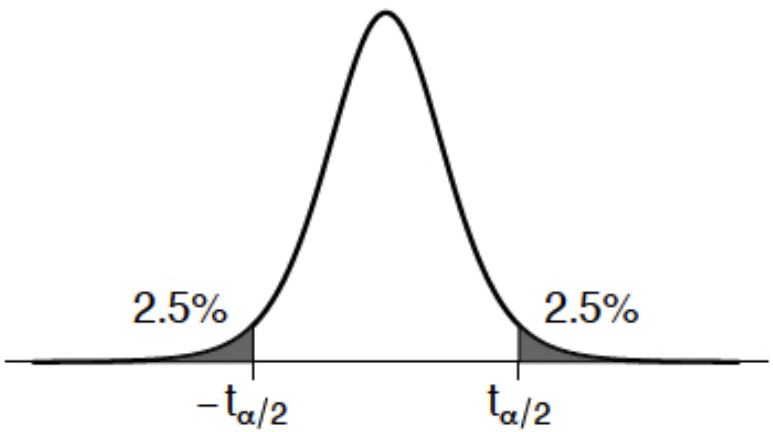
$$H_0 : \mu_1 = \mu_2$$

versus

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

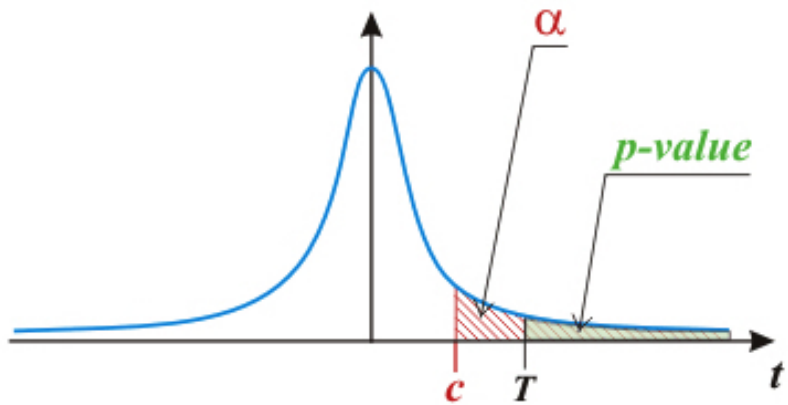
$$\text{Test Statistic: T} = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \text{ (signal to noise ratio)}$$

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)

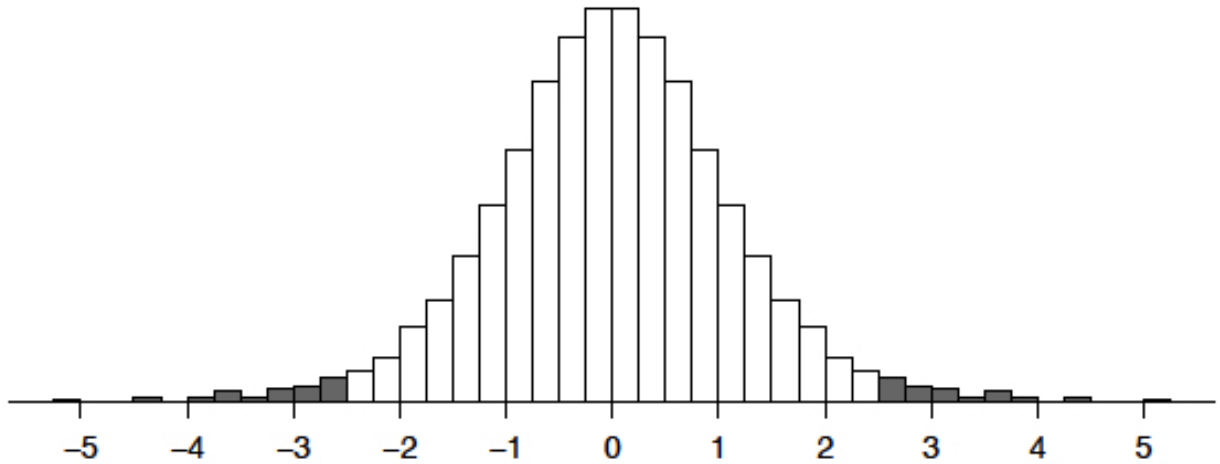
Wilcoxon rank sum test
data: g1 and g2
W = 1432, p-value = 8.306e-13
alternative hypothesis: true location shift is not equal to 0

Permutation

Idea: generate the null distribution by random shuffling group label

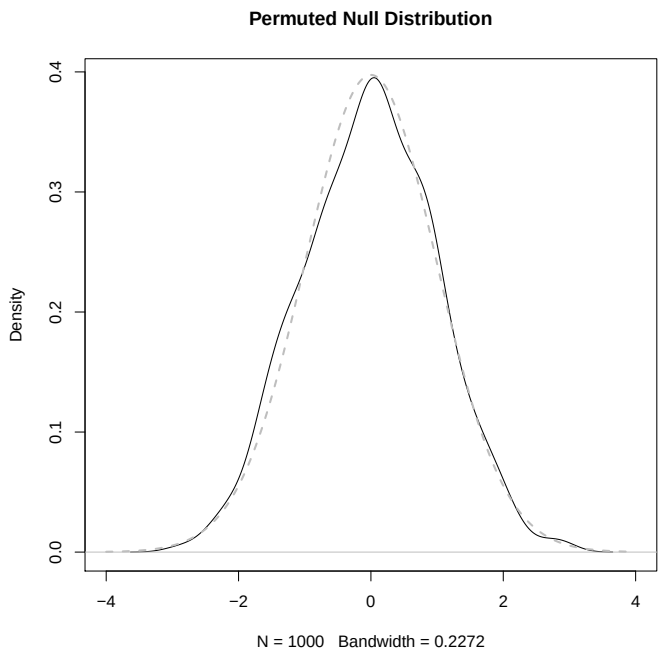
Group 1	Group 2
0.82	-1.19
0.12	-0.84
0.46	1.89

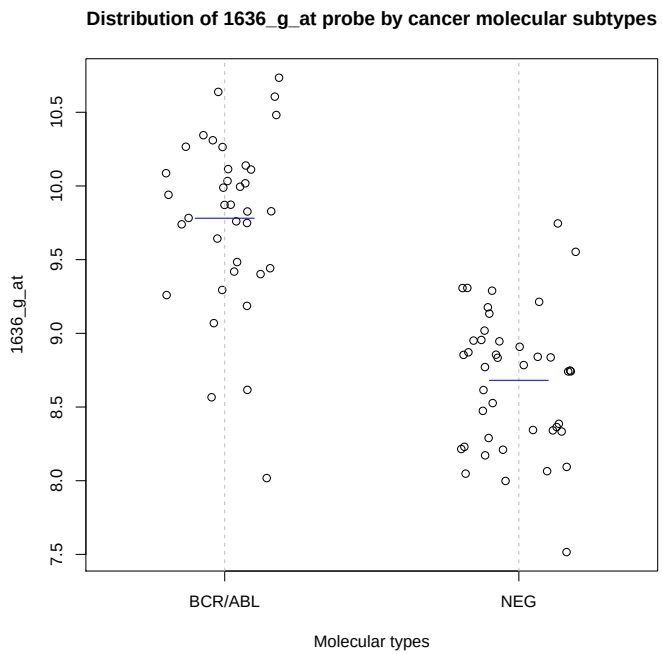
Randomly assign the group labels $\rightarrow T^*$



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

Permutation Test





Permutation Test is A Good Friend

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

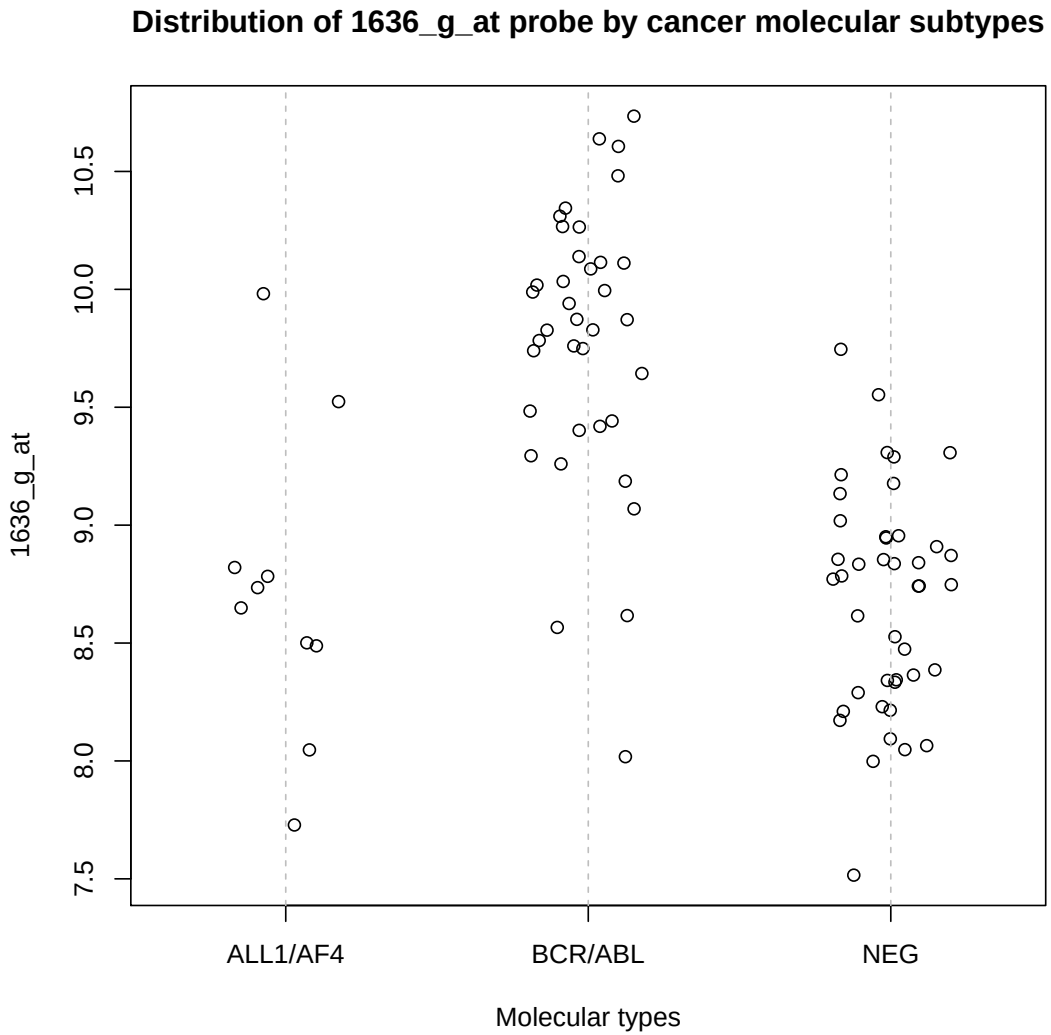
What to Use

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



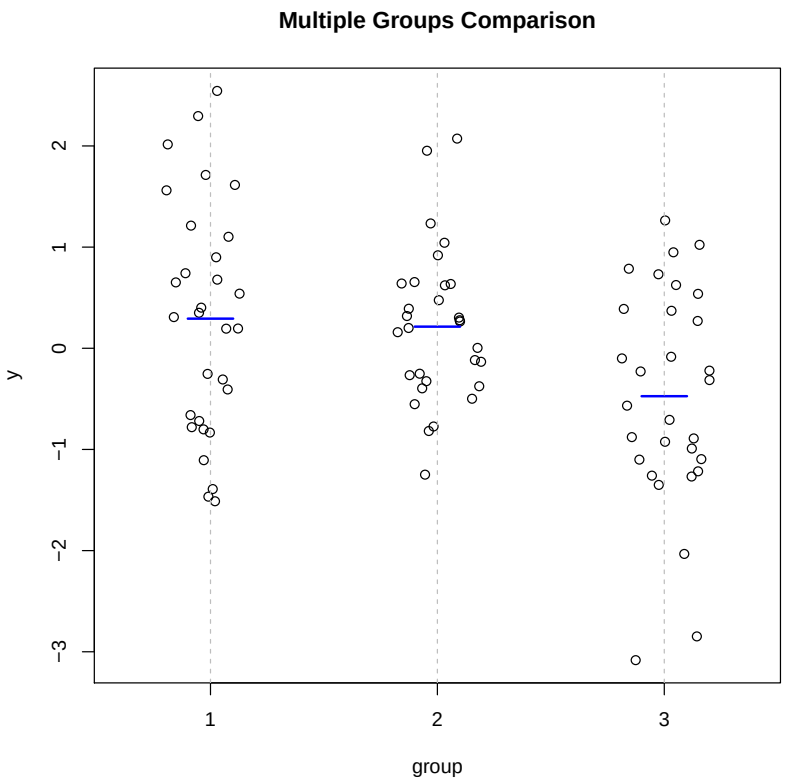
Multiple groups comparison: Hypothesis

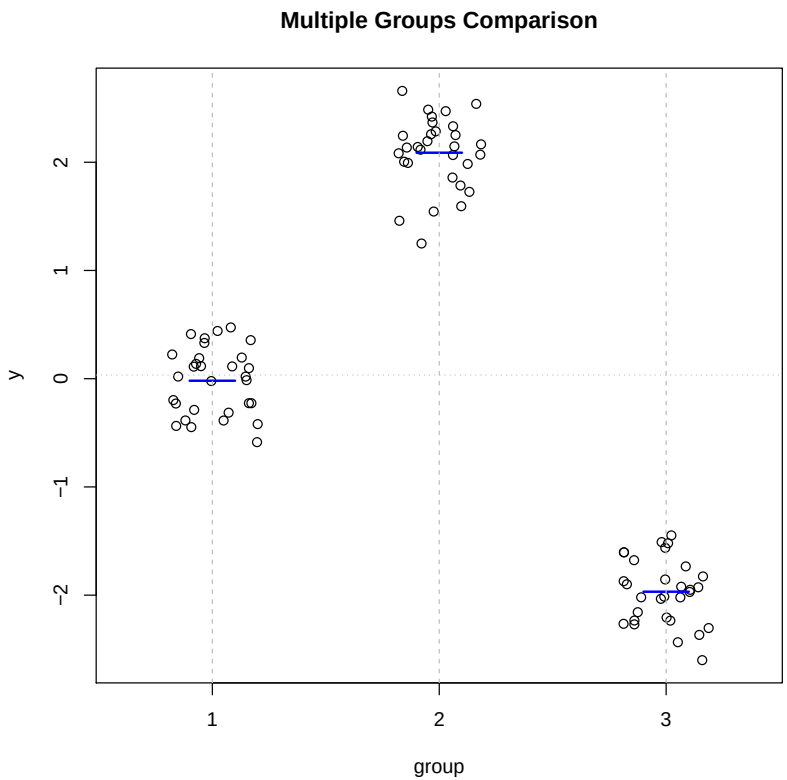
Are there differences in the means of gene expression among the three molecular groups (ALL1/AF4, BCR/ABL, NEG) ?

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

ANOVA

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

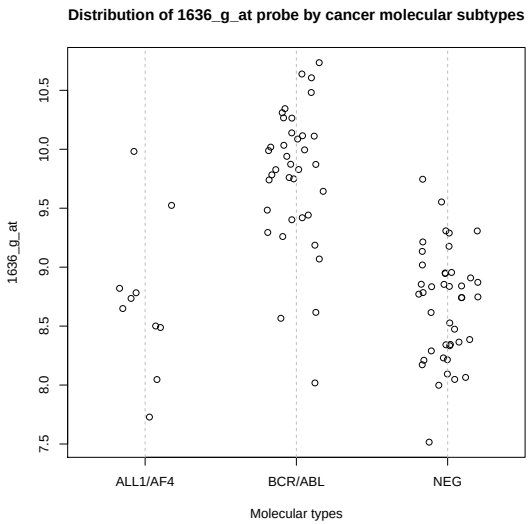




ANOVA: Gene Expression Example

>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		



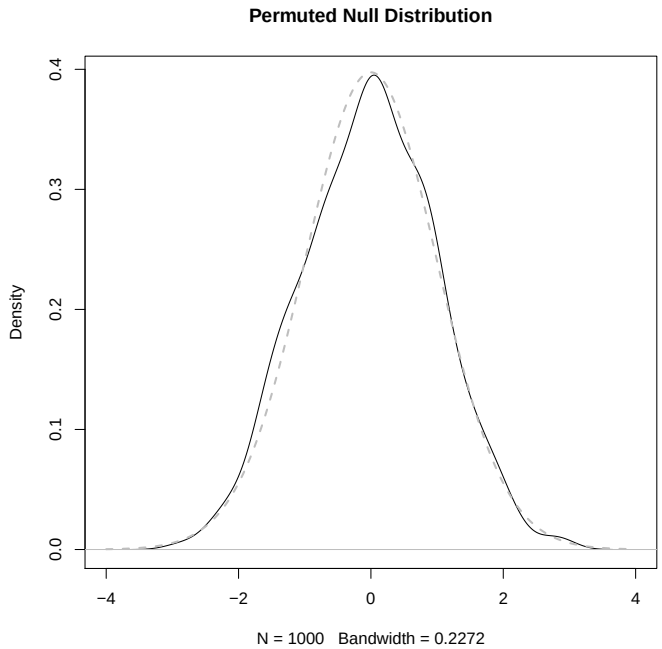
Kruskal-Wallis (K-W) Test

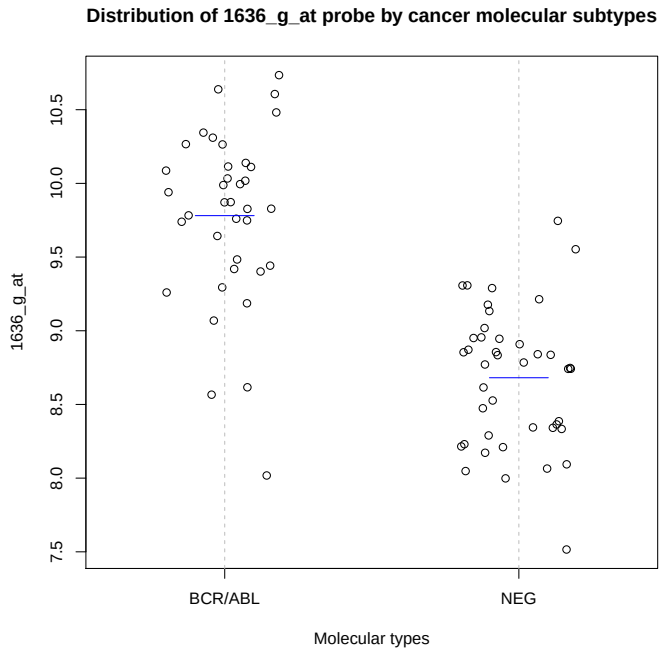
Small sample setting when normality assumption is not reasonable

```
> kruskal.test(all[whs, ], ALL3$mol.biol, na.action=na.exclude)

Kruskal-Wallis rank sum test
data: all[whs, ] and ALL3$mol.biol
Kruskal-Wallis chi-squared = 43.5804, df = 2, p-value = 3.441e-10
```

Permutation Test





Exercise: Your turn, use the ALL data example to generate the permuted null distribution.

Permutation Test

```
>perm <-1000
>tstar<- rep(NA, perm)
> for (i in 1:perm){
        group <- sample(ALL_bcrneg$mol.biol)
        g1<- data[whp, group==‘‘BCR/ABL"]
        g2 <- data[whp, group==‘‘NEG"]
        tstar[i] <- t.test(g1, g2)$statistic
}
> plot(density(tstar), main=‘‘Permuted Null
Distribution")
> pvalue <- mean(abs(tstar) >= abs(tobs))
> pvalue
[1] 0
```