

# Lecture 5 – Two Group Comparisons

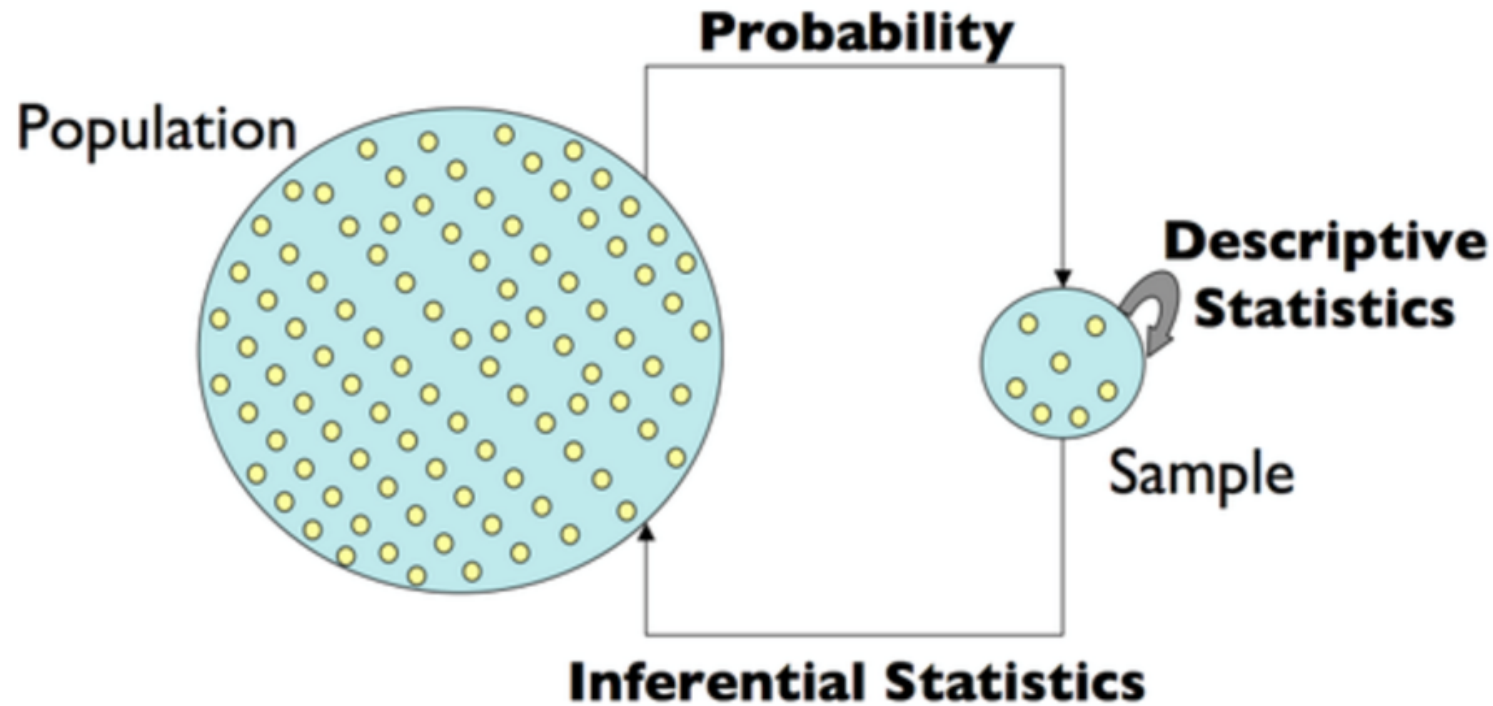
STAT/BIOF/GSAT 540: Statistical Methods for High Dimensional Biology

Keegan Korthauer

2020/01/20

Slides by: Gabriela Cohen Freue with contributions from Jenny Bryan, Sara Mostafavi, and Keegan Korthauer

# Statistical inference



We want to understand a **population** (e.g., gene behaviour) but we can only study a **random sample** from it.  
*[Picture from Dr. Fowler, UW]*

# Book and online resources

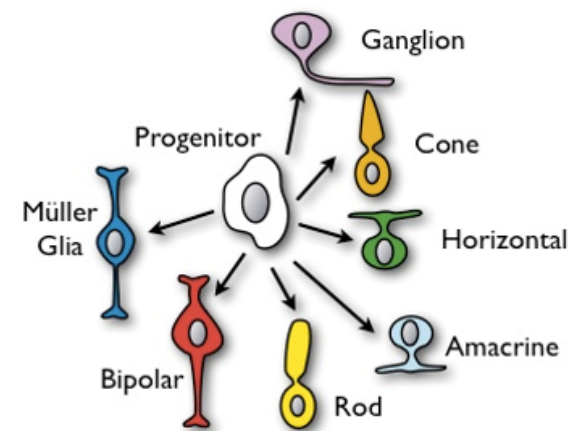
- Linear Models with R by Julian J. Faraway, Chapman & Hall/CRC Texts in Statistical Science, 2004
- <http://www.biostat.jhsph.edu/~iruczins/teaching/jf/faraway.html>
- <https://sites.google.com/a/cs.washington.edu/genome560-spr18/CourseMaterials>
- <http://genomicsclass.github.io/book/>

# Hypothesis Testing in Genomics

## Targeting of GFP to newborn rods by Nrl promoter and temporal expression profiling of flow-sorted photoreceptors

Masayuki Akimoto<sup>\*†</sup>, Hong Cheng<sup>\*</sup>, Dongxiao Zhu<sup>§¶</sup>, Joseph A. Brzezinski<sup>||</sup>, Ritu Khanna<sup>\*</sup>, Elena Filippova<sup>\*</sup>, Edwin C. T. Oh<sup>‡</sup>, Yuezhou Jing<sup>¶</sup>, Jose-Luis Linares<sup>\*</sup>, Matthew Brooks<sup>\*</sup>, Sepideh Zarepari<sup>\*</sup>, Alan J. Mears<sup>\*.\*\*\*</sup>, Alfred Hero<sup>§¶††‡‡</sup>, Tom Glaser<sup>||§§</sup>, and Anand Swaroop<sup>\*‡||¶¶</sup>

- Retina presents a model system for investigating **regulatory networks** underlying neuronal differentiation.
- **Nrl** transcription factor is known to be important for Rod development
- **What happens if you delete Nrl?**



# Why a Hypothesis Test?

From the Akimoto et al. (2006) paper:

"we hypothesized that *Nrl* is the ideal transcription factor to gain insights into gene expression changes ..."

**Biological question:** Is the expression level of gene A affected by knockout of the *Nrl* gene?

We can use **statistical inference** to answer this biological question!

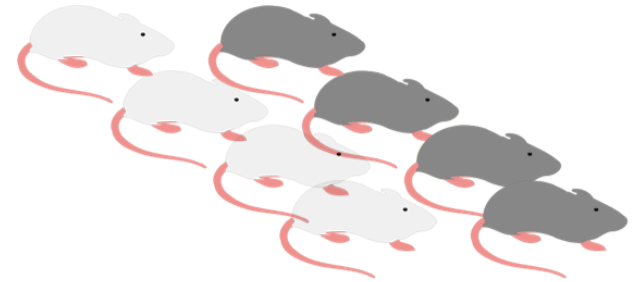
# Statistical inference

## Statistical inference:

We observe and study a **random sample** to make conclusions about a population (e.g., random sample of gene expressions from mice)

## Experimental design:

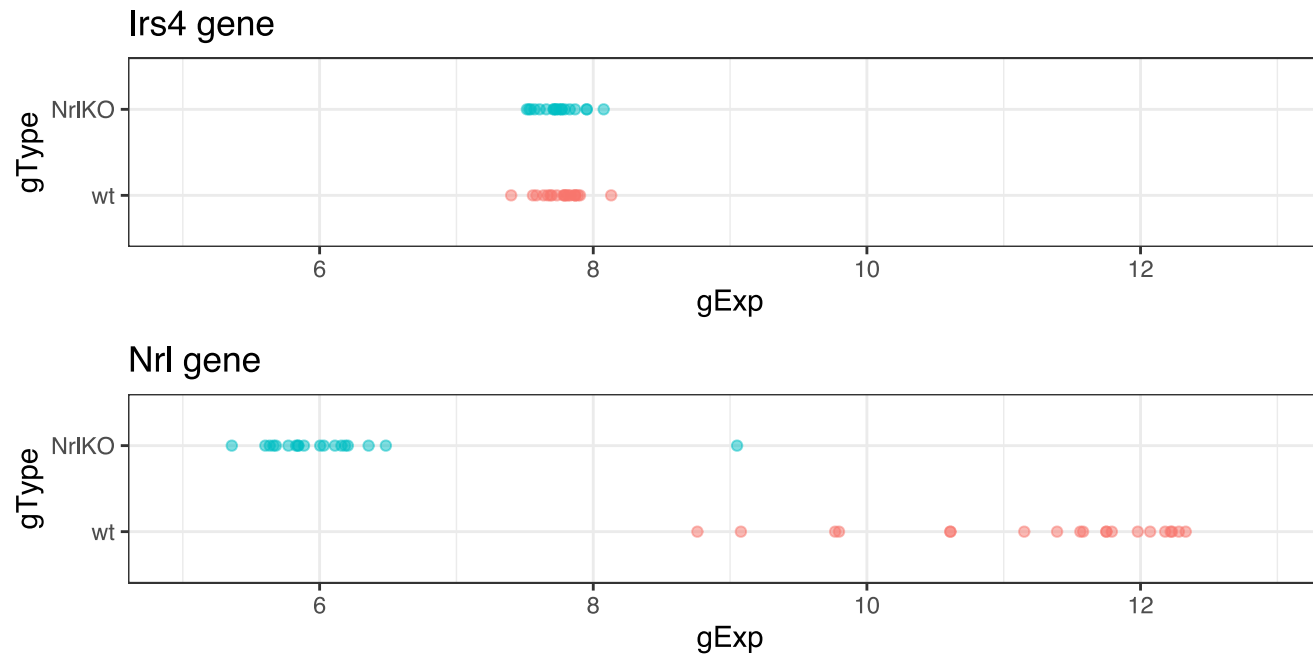
- 4 developmental stages
- 2 genotypes: Wild type (WT), Nrl Knockout (NrlKO)
- 3-4 replicates for each combination



Let's take a look at 2 genes as an example: **Irs4** and **Nrl**

**Biological question:** Are these genes truly different in NrlKO compared to WT?

We can't answer this question in general. We can *only* study these genes in collected data:



**We only observe a random sample of gene expressions.**

# Statistical Hypothesis

## Experimental design:

- 2 conditions: WT vs NrlKO
- random sample: we observe the expression of many genes in all mice

**Biological hypothesis:** for *some* genes, the expression levels are different between conditions.

**Statistical hypotheses:** (for each gene  $g = 1, \dots, G$ )

- $H_0$  (null hypothesis): the expression level of gene  $g$  is the *same* in both conditions.
- $H_A$  (alternative hypothesis): the expression level of gene  $g$  is *different* between conditions.



# Notation

Random variables and estimates (we can observe):

$Y_i$  : expression of gene  $g$  in the WT sample  $i$

$Z_i$ : expression of gene  $g$  in NrlKO sample  $i$

$Y_1, Y_2, \dots, Y_{n_Y}$  : a **random sample** of size  $n_Y$

$\bar{Y} = \frac{\sum_{i=1}^{n_Y} Y_i}{n_Y}$ : sample mean of gene  $g$  expression from WT mice

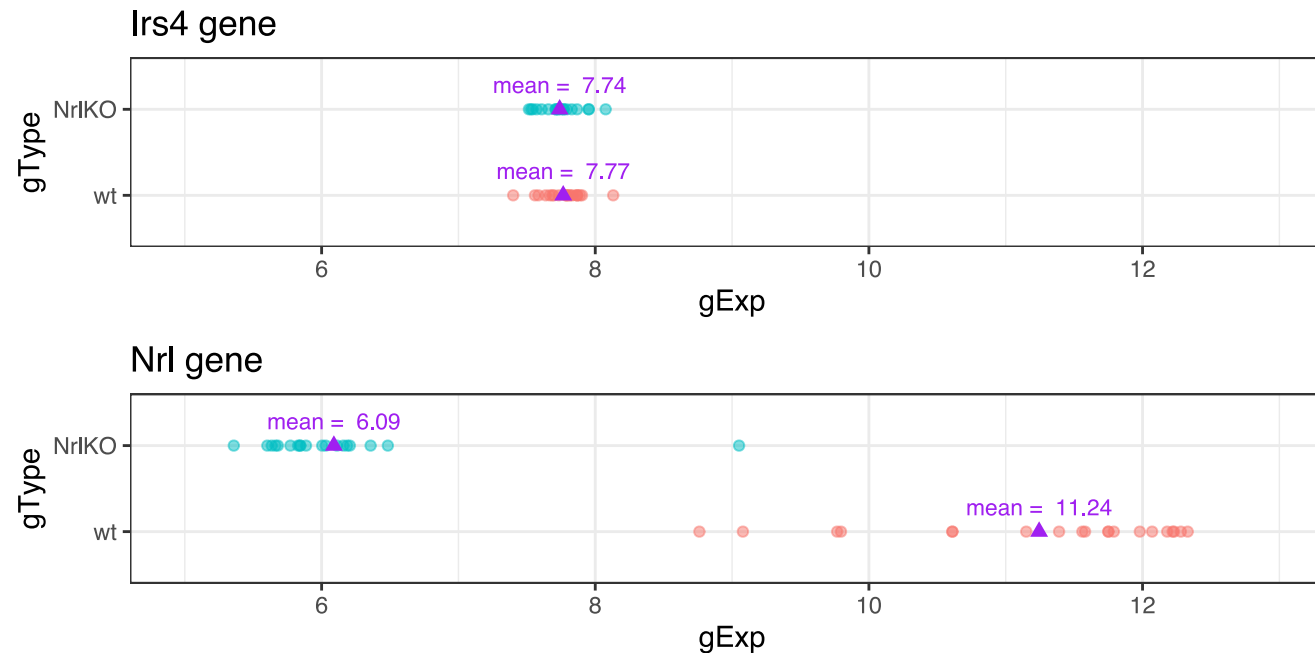
note the different meanings of the word "sample"

Population parameters (unknown/unobservable):

$\mu_Y = E[Y]$  : the (population) expected expression of gene  $g$  in WT mice

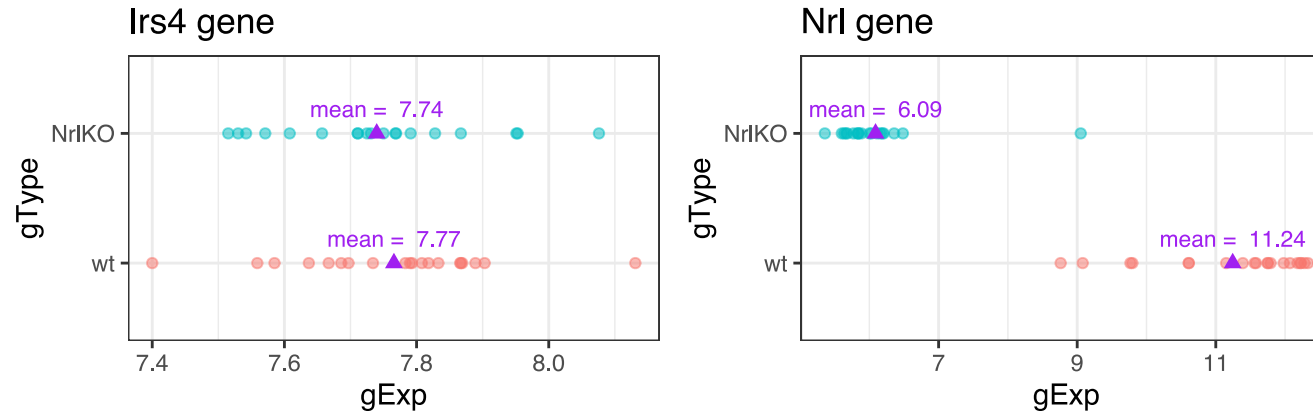
Is there **enough** evidence in the data to reject  $H_0$ ?

$$H_0 : \mu_Y = \mu_Z$$



**Statistical Inference:** random samples are used to learn about the population

# What we observe: the difference between the sample averages: $\bar{Y}$ vs $\bar{Z}$



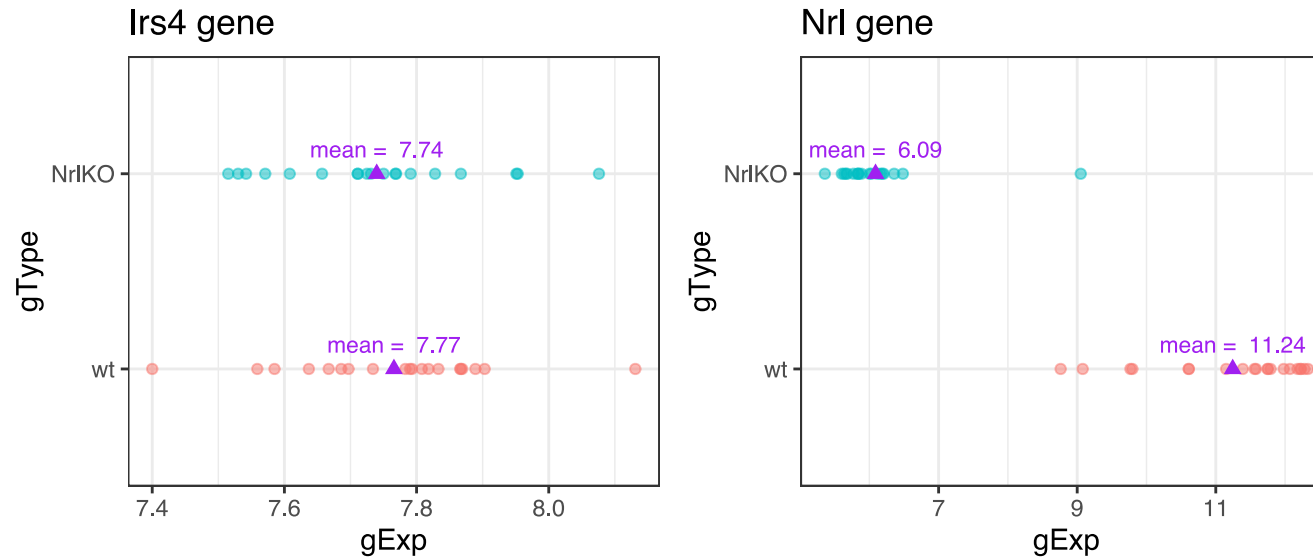
```
theAvg <- with(miniDat,  
               tapply(gExp, list(gType, gene), mean))
```

```
##      Irs4   Nr1  
## wt    7.766 11.244  
## Nr1KO 7.740  6.090
```

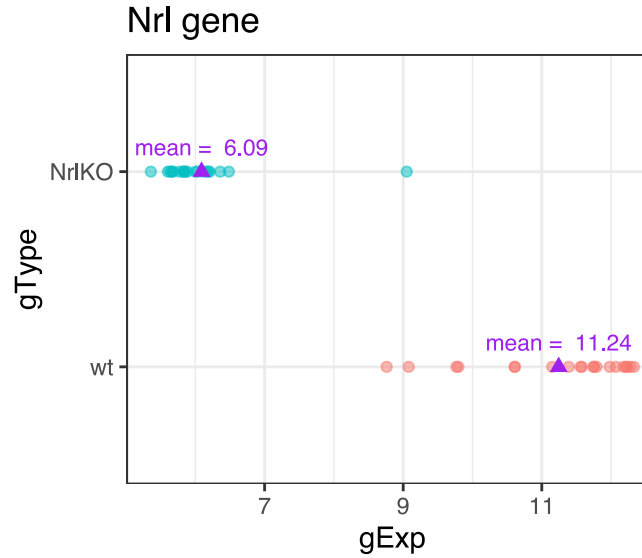
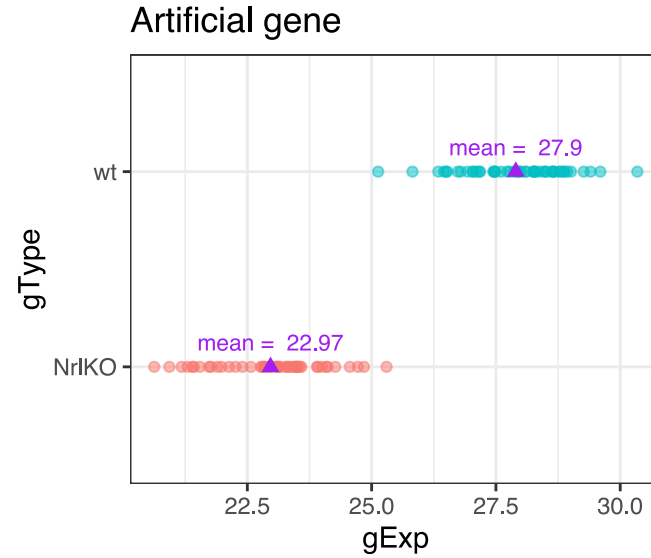
```
theDiff <- theAvg["Nr1KO", ] - theAvg["wt", ]
```

```
##   Irs4   Nr1  
## -0.026 -5.155
```

Is the difference between  $\bar{Y}$  and  $\bar{Z}$  informative to reject  $H_0$ ?



- The sample means,  $\bar{Y}$  vs  $\bar{Z}$ , by themselves are not enough to make conclusions about the population
- What is a "large" difference? "large" relative to what?



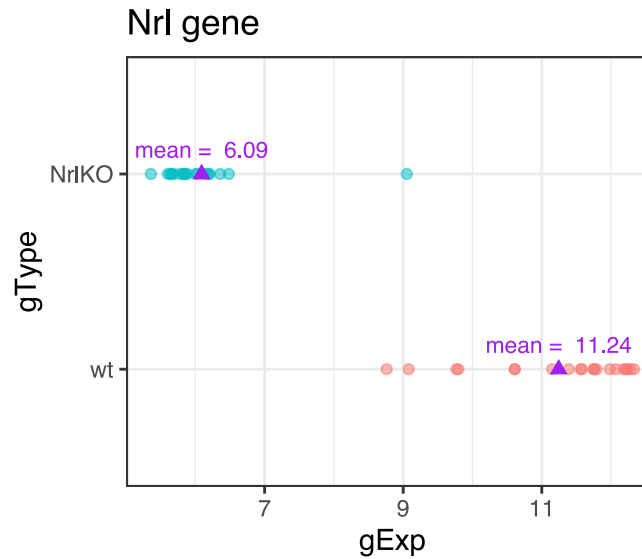
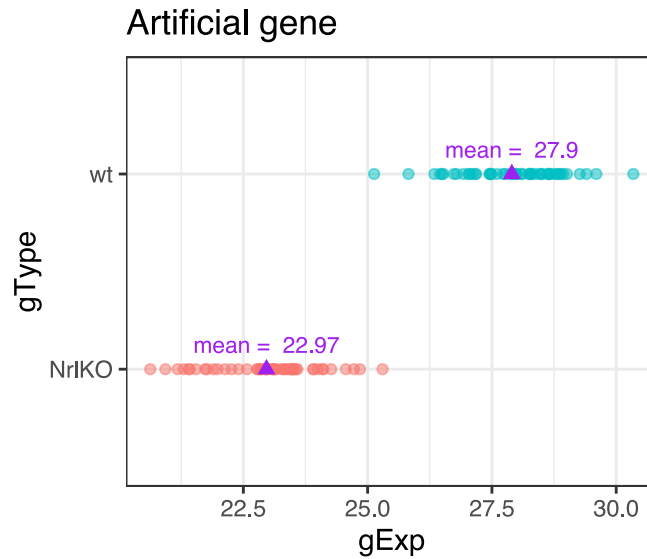
What can we use to interpret the size of the mean difference?

$$\frac{\bar{Y} - \bar{Z}}{??}$$

What can we use to interpret the size of the mean difference?

"large" relative to the observed variation

$$\frac{\bar{Y} - \bar{Z}}{\sqrt{V(\bar{Y} - \bar{Z})}}$$



# Quantifying observed variation

- Recall that if  $V(Y_i) = \sigma_Y^2$ , then  $V(\bar{Y}) = \frac{\sigma_Y^2}{n_Y}$
- Assume that the random variables within each group are *independent and identically distributed* (iid), and that the groups are independent. More specifically, that
  1.  $Y_1, Y_2, \dots, Y_{n_Y}$  are iid,
  2.  $Z_1, Z_2, \dots, Z_{n_Z}$  are iid, and
  3.  $Y_i, Z_j$  are independent. Then, it follows that

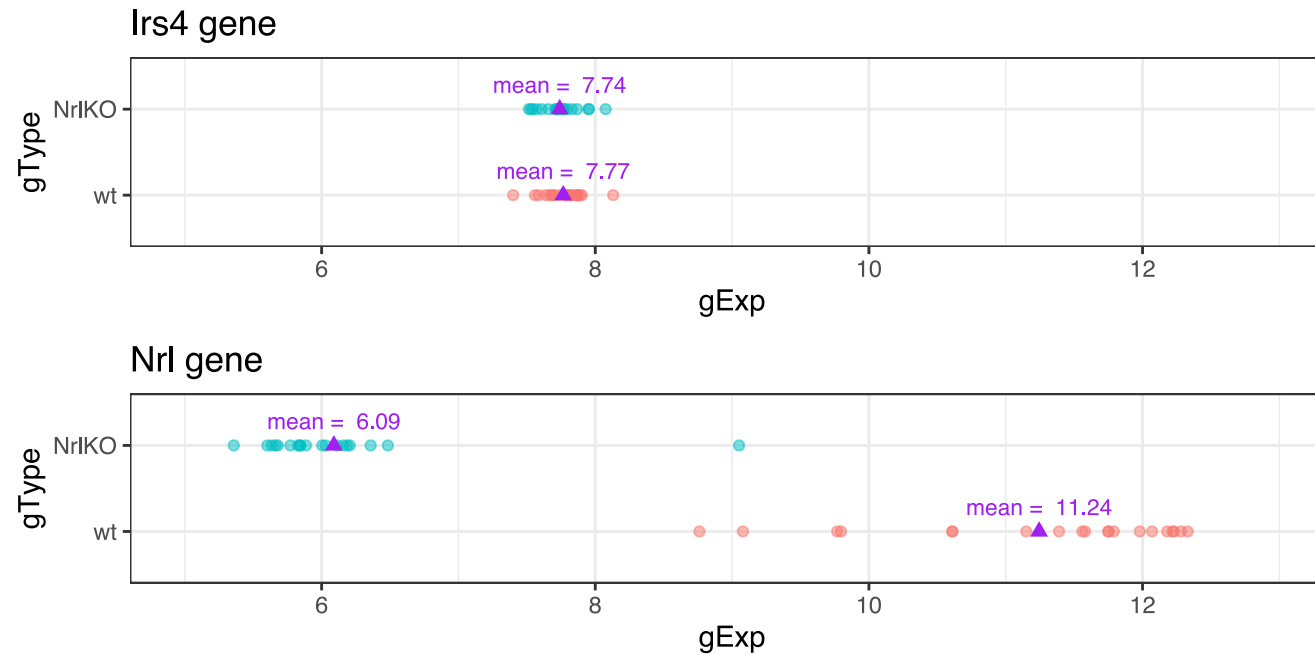
$$V(\bar{Z} - \bar{Y}) = \frac{\sigma_Z^2}{n_Z} + \frac{\sigma_Y^2}{n_Y}$$

- If we also assume equal population variances:  $\sigma_Z^2 = \sigma_Y^2 = \sigma^2$ , then

$$V(\bar{Z} - \bar{Y}) = \frac{\sigma_Z^2}{n_Z} + \frac{\sigma_Y^2}{n_Y} = \sigma^2 \left[ \frac{1}{n_Z} + \frac{1}{n_Y} \right]$$

But how can we calculate population variance  $\sigma$  if it is **unknown**?

...using the sample variances (combined, somehow)!



```
theVars <- with(miniDat,  
  tapply(gExp, list(gType, gene), var))
```

```
##      Irs4   Nrl  
## wt    0.024 1.224  
## NrlKO 0.023 0.594
```

e.g., for Nrl:  $\hat{\sigma}_Y^2 = S_Y^2 = \frac{1}{n_Y} \sum_{i=1}^{n_Y} (Y_i - \bar{Y})^2 = 1.224$



Plug these sample variances into your chosen formula for the variance of the difference of sample means

Assuming **equal** variance of Y's and Z's

$$\hat{\sigma}_{\text{pooled}}^2 = S_Y^2 \frac{n_Y - 1}{n_Y + n_Z - 2} + S_Z^2 \frac{n_Z - 1}{n_Y + n_Z - 2}$$

$$\hat{V}(\bar{Z}_n - \bar{Y}_n) = \hat{\sigma}_{\text{pooled}}^2 \left[ \frac{1}{n_Y} + \frac{1}{n_Z} \right]$$

Assuming **unequal** variance of Y's and Z's

$$\hat{V}(\bar{Z}_n - \bar{Y}_n) = \hat{\sigma}_{\bar{Z}_n - \bar{Y}_n}^2 = \frac{S_Y^2}{n_Y} + \frac{S_Z^2}{n_Z}$$

■ Note: the 'hat' (^) is used to distinguish an 'estimate' from a 'parameter'.

# The Test Statistic: $T = \frac{\bar{Z}_n - \bar{Y}_n}{\hat{\sigma}_{Z_n - Y_n}}$

Assuming equal variances:

```
tstStat <- theDiff / sqrt(s2Diff)
```

```
##      Irs4      Nrl  
## -0.529 -16.795
```

Without assuming equal variances:

```
welchStat <- theDiff / sqrt(s2DiffWelch)
```

```
##      Irs4      Nrl  
## -0.529 -16.949
```

## Can we now say that the observed differences are 'big'?

The difference is about half a standard deviation for Irs4 and ~16 standard deviations for Nrl.

The test statistic  $T$  is a **random variable** because it's based on our **random sample**.

We need a measure of its uncertainty to determine how big  $T$  is:

| If we were to repeat the experiment many times, what's the probability of observing a value of  $T$  **as extreme** as the one we observed?

We need to have a probability distribution!

However, this is unknown to us so we need to **make more assumptions**.

Theory now tells us specific **null distributions** for these test statistics, depending on your assumptions.

--> Willing to assume that F and G are normal distributions?

2-sample *t*-test:

(equal variances)

$$T \sim t_{n_Y + n_Z - 2}$$

Welch test:

(unequal variances)

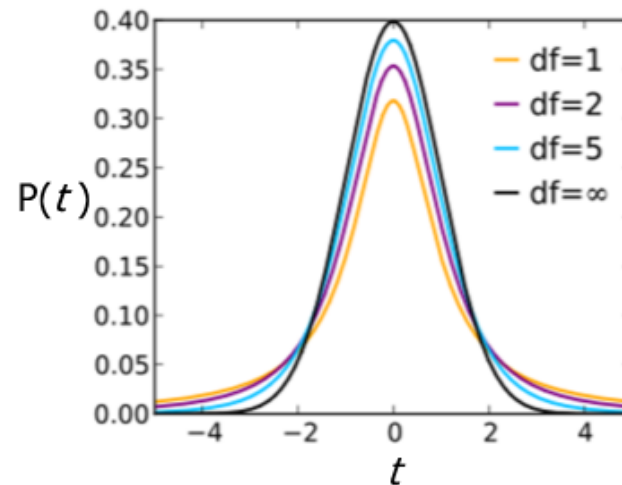
$$T \sim t_{\text{something ugly}}$$

--> Unwilling to assume that F and G are normal distributions? But you feel that  $n_Y$  and  $n_Z$  are large enough?

Then the *t*-distributions above or even a normal distribution are decent approximations.

# Student's $t$ -distribution

Recall that  $T$  is a **random variable**. Under certain assumptions, we can prove that  $T$  follows a  $t$ -distribution.



where  $df$  = degrees of freedom.

# Hypothesis testing

1. Define a **test statistic**: 2-sample  $t$ -test
2. Compute the **observed value** for the test statistic

```
tstStat <- theDiff / sqrt(s2Diff)
```

```
##      Irs4      Nr1  
## -0.529 -16.795
```

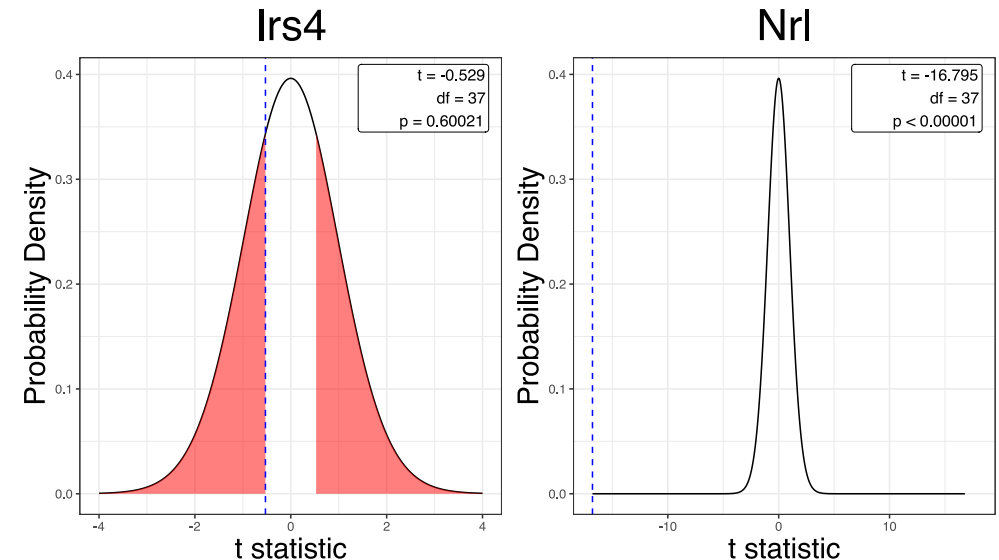
3. Compute the probability of seeing a test statistic at least as extreme as that observed, under the **null sampling distribution** (this is the definition of the p-value)

```
## miniDat$gene: Irs4  
## [1] 0.6002058  
## -----  
## miniDat$gene: Nr1  
## [1] 6.764663e-19
```

In other words, assuming that  $H_0$  is true:

For Irs4, the probability of seeing a test statistic as extreme as that observed ( $t = -0.53$ ) is pretty high ( $p = 0.6$ ).

But for Nr1, the probability of seeing a test statistic as extreme as that observed ( $t = -16.8$ ) is extremely low ( $p = 6.76 \times 10^{-19}$ )



#### 4. Make a decision about significance of results, based on a pre-specified value (alpha, significance level)

The significance level  $\alpha$  is usually set at 0.05. However, this value is arbitrary and usually depends on the study.

Using  $\alpha = 0.05$ , since the p-value for the Irs4 test is greater than 0.05, we conclude that there is not *enough evidence* in the data to claim that Irs4 has a differential expression in WT compared to Nrl models.

We do not reject  $H_0$ !



## Assuming equal variances: *t*-test

```
> by(miniDat, miniDat$gene, function(theDat) {  
+   t.test(gExp ~ gType, theDat, var.equal = TRUE)  
+ })  
miniDat$gene: Irs4
```

Two Sample t-test

```
data: gExp by gType  
t = -0.5286, df = 37, p-value = 0.6002  
<snip, snip>
```

-----  
miniDat\$gene: Nrl

Two Sample t-test

```
data: gExp by gType  
t = -16.7947, df = 37, p-value < 2.2e-16
```

## Without assuming equal variances: Welch *t*-test

```
> by(miniDat, miniDat$gene, function(theDat) {  
+   t.test(gExp ~ gType, theDat)  
+ })
```

miniDat\$gene: Irs4

Welch Two Sample t-test

```
data: gExp by gType  
t = -0.5289, df = 36.948, p-value = 0.6001
```

```
<snip, snip>
```

# What is a p-value?

Likelihood of obtaining a test statistic at least **as extreme as the one observed**, given that the null hypothesis is true (we are making a conditional p-value statement)

## What is a p-value **NOT**?

- Not the probability that the **null hypothesis is true**
- Not the probability that the **finding is a “fluke”**
- Not the probability of **falsely rejecting the null**
- Does not **indicate the size or importance** of observed effects.

*[Credit to Dr. Fowler, UW]*

# "Genome-wide" testing of differential expression

- In genomics, we often perform thousands of statistical tests (e.g., a  $t$ -test per gene)
- The distribution of p-values across all tests provide good diagnostics/insights.
- Is it uniform (should be in most experiments) and if not, is the departure from uniform expected based on biological knowledge?

## Different kinds of $t$ -tests:

- One sample *or* **two samples**
- One-sided *or* **two sided**
- Paired *or* **unpaired**
- **Equal variance** *or* unequal variance

# Types of Errors in Hypothesis Testing

| Actual Situation "Truth" |  |  |  |
|--------------------------|--|--|--|
| Decision                 | $H_0$ True                                     | $H_0$ False                                    |  |
| Do Not Reject $H_0$      | Correct Decision<br>$1-\alpha$                 | Incorrect Decision<br>Type II Error<br>$\beta$ |  |
| Reject $H_0$             | Incorrect Decision<br>Type I Error<br>$\alpha$ | Correct Decision<br>$1-\beta$                  |  |

Dr. Fowler, UW]

[Picture from

What if you don't wish to assume the underlying data is normally distributed AND you aren't sure your samples are large enough to invoke CLT?

## What are alternatives to the $t$ -test?

First, one could use the  $t$  test statistic but use a **bootstrap approach** to compute its p-value. We will cover this later on.

Alternatively, there are *non-parametric* tests that are available here:

- **Wilcoxon rank sum test**, aka Mann Whitney, uses ranks to test differences in population means.
- **Kolmogorov-Smirnov test** uses the empirical CDF to test differences in population cumulative distributions.

# Wilcoxon rank sum test

Rank all data, ignoring the **grouping** variable

**Test statistic** = sum of the ranks for one group (optionally, subtract the minimum possible which is  $\frac{n_Y(n_Y+1)}{2}$ )

(Alternative but equivalent formulation based on the number of  $y_i, z_i$  pairs for which  $y_i \geq z_i$ )

Null distribution of such statistics can be worked out or approximated.

, center

```
miniDat$gene: Irs4
```

Wilcoxon rank sum test with continuity correction

```
data:  gExp by gType
W = 220.5, p-value = 0.3992
alternative hypothesis: true location shift is not equal to 0
```

```
-----
miniDat$gene: Nrl
```

Wilcoxon rank sum test with continuity correction

```
data:  gExp by gType
W = 379, p-value = 1.178e-07
alternative hypothesis: true location shift is not equal to 0
```

```
miniDat$gene: Irs4
```

Welch Two Sample t-test

```
data:  gExp by gType
t = 0.5289, df = 36.948, p-value = 0.6001

<snip, snip>
```

```
-----
miniDat$gene: Nrl
```

Welch Two Sample t-test

```
data:  gExp by gType
t = 16.9486, df = 34.005, p-value < 2.2e-16

<snip, snip>
```

# Kolmogorov-Smirnov test (two sample)

**Null hypothesis:**  $F = G$ , i.e. the distributions are the same

Estimate each CDF with the empirical CDF (ECDF)

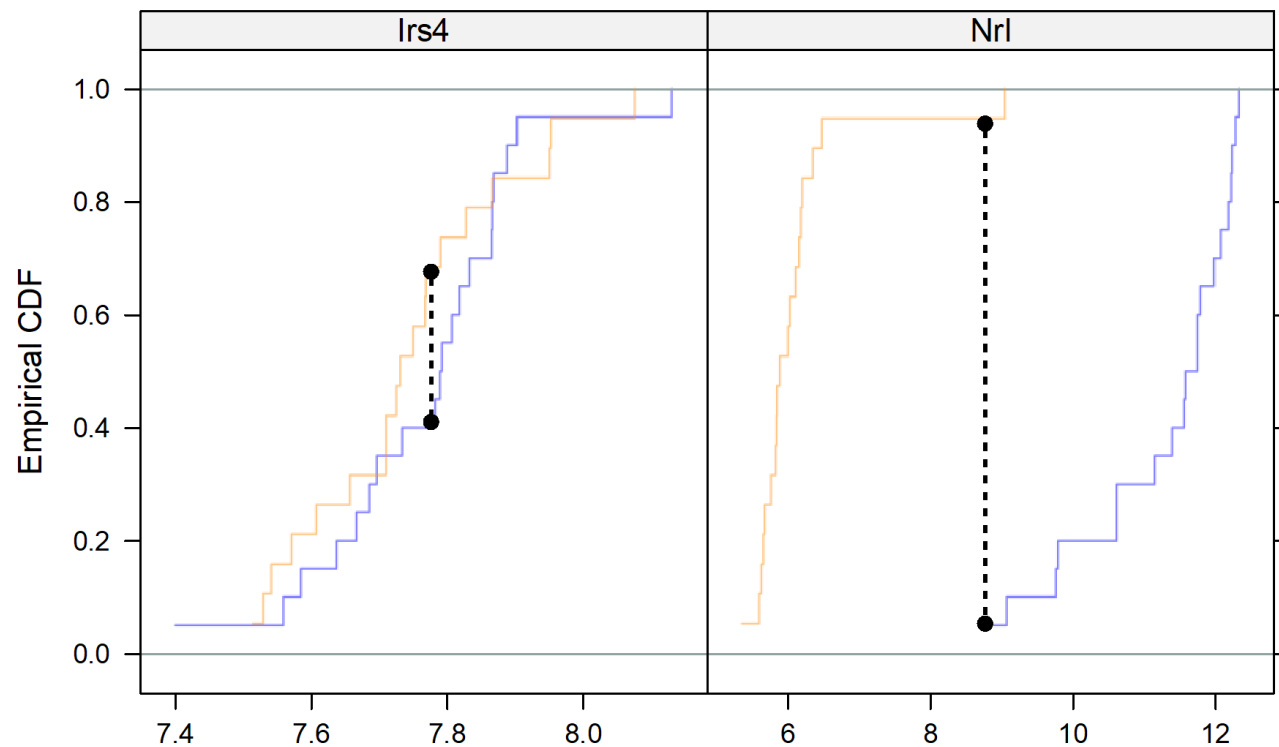
$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^n I[x_i \leq x]$$

**Test statistic** is the maximum of the absolute difference between the ECDFs

$$\max |\hat{F}(x) - \hat{G}(x)|$$

Null distribution does not depend on  $F, G$  (!)  
(I'm suppressing a detail here.)





miniDat\$gene: Irs4

gExp

miniDat\$gene: Irs4

Two-sample Kolmogorov-Smirnov test

data: theDat\$gExp[theDat\$gType == "wt"] and theDat\$gExp[theDat\$gType == "NrlKO"]

D = 0.2842, p-value = 0.4107

alternative hypothesis: two-sided

Welch Two Sample t-test

data: gExp by gType

t = 0.5289, df = 36.948, p-value = 0.6001

<snip, snip>

miniDat\$gene: Nrl

miniDat\$gene: Nrl

# Discussion and questions ...

What if you are unsure whether your sample size is large enough? Outliers with small samples could be problematic

Which test result should one report ... the 2-sample  $t$ -test, the Wilcoxon, or the KS?

Treat p-values as one type of evidence that you should incorporate with others.

It is worrisome when methods that are equally appropriate and defensible give very different answers.