Statistical Methods for High Dimensional Biology

STAT/BIOF/GSAT 540

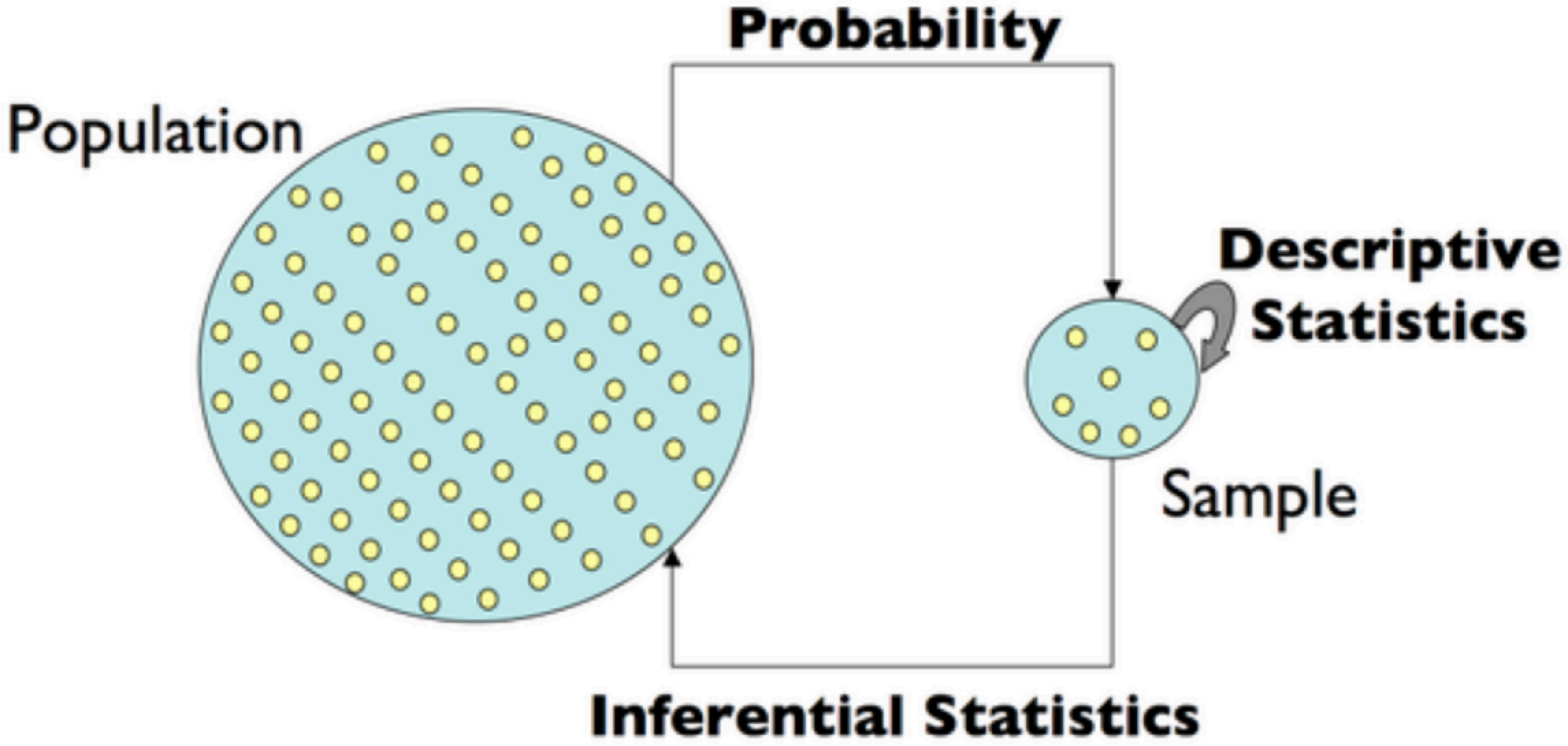
Lecture 5 – Two group comparisons

Gabriela Cohen Freue

January 21 2019

Other contributors: Jenny Bryan, Sara Mostafavi, Su-In Lee

Statistical inference



We want to understand a population (e.g., gene behavior) but we can only study a random sample from it.

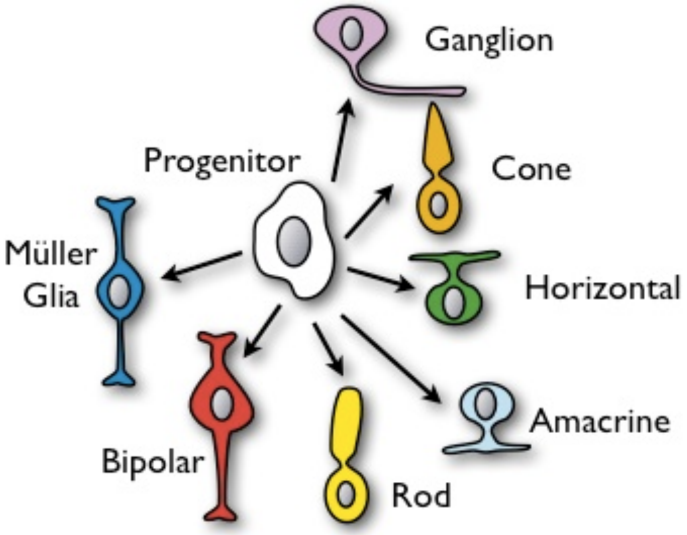
(Picture from Dr Fowler, UW)

Hypothesis Testing in Genomics

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

Masayuki Akimoto[*†], Hong Cheng[‡], Dongxiao Zhu[§¶], Joseph A. Brzezinski[‖], Ritu Khanna[*], Elena Filippova[*], Edwin C. T. Oh[‡], Yuezhou Jing[¶], Jose-Luis Linares[*], Matthew Brooks[*], Sepideh Zareparsi[*], Alan J. Mears[*,**], Alfred Hero[§¶††‡‡], Tom Glaser[‖§§], and Anand Swaroop[*‡‖¶¶]

- 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?

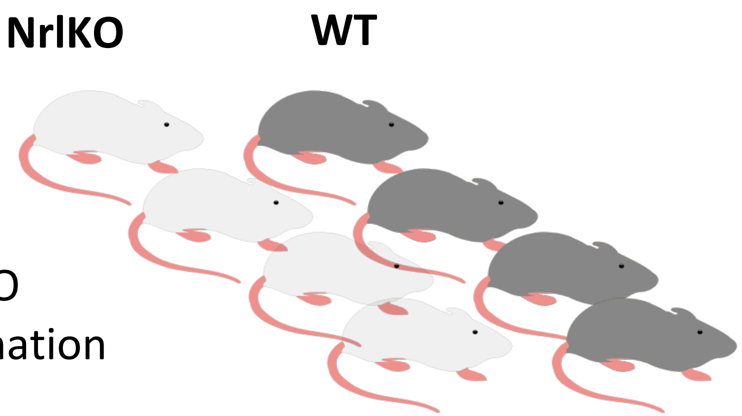


Why a hypothesis test?

From 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 ablation of the *Nrl* gene in mice?

Experimental Design: we observe a random sample
(random sample of gene expressions from our experiment)



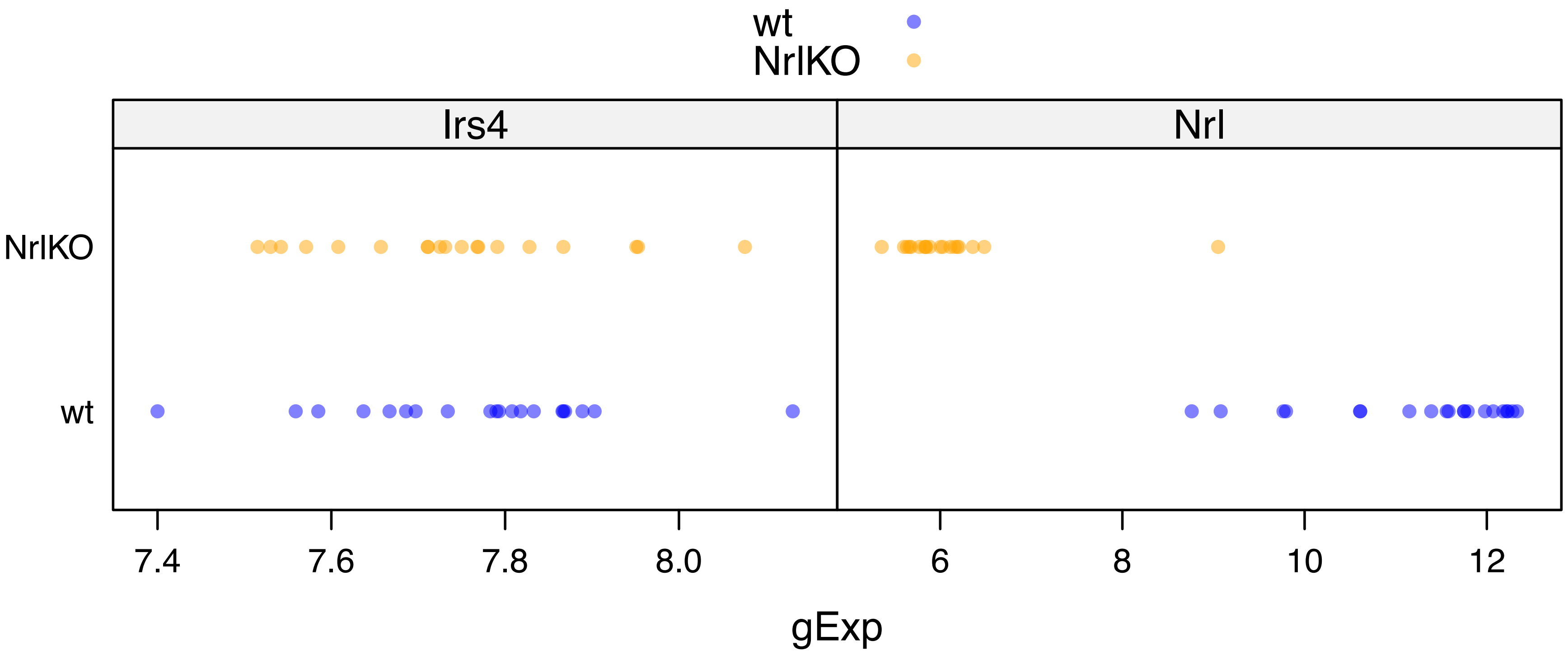
4 developmental stages
2 genotypes: wild-type , Nrl KO
3-4 replicates for each combination

Statistical Inference

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

Are these genes truly different in NrlKO compared to wt?

We only observe a random sample.



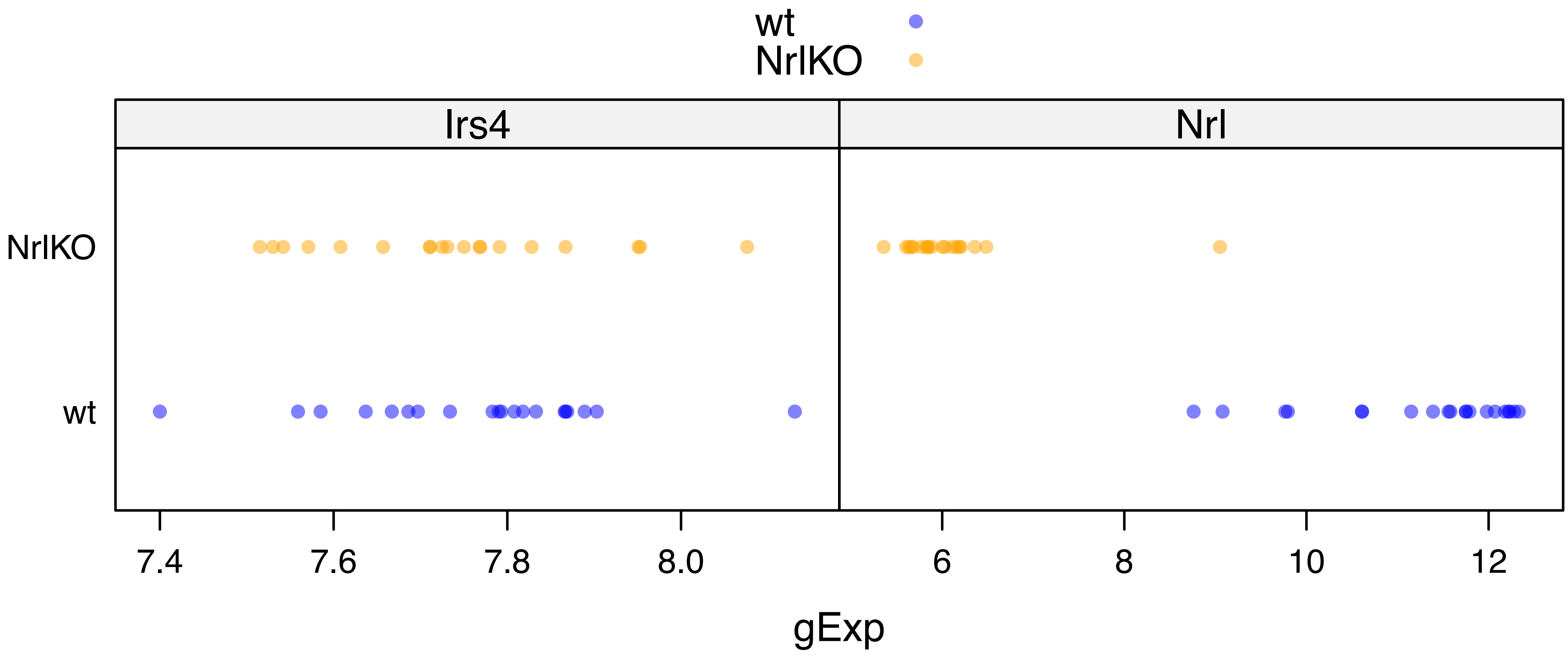
Statistical hypothesis

- **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 in both conditions.
- **Statistical hypotheses:** one gene at a time
  - H_0 (null hypothesis): the expression level of gene A is the same.
  - H_A (alternative hypothesis): the expression level of gene A is different.

Are these genes truly different in NrlKO compared to wt?

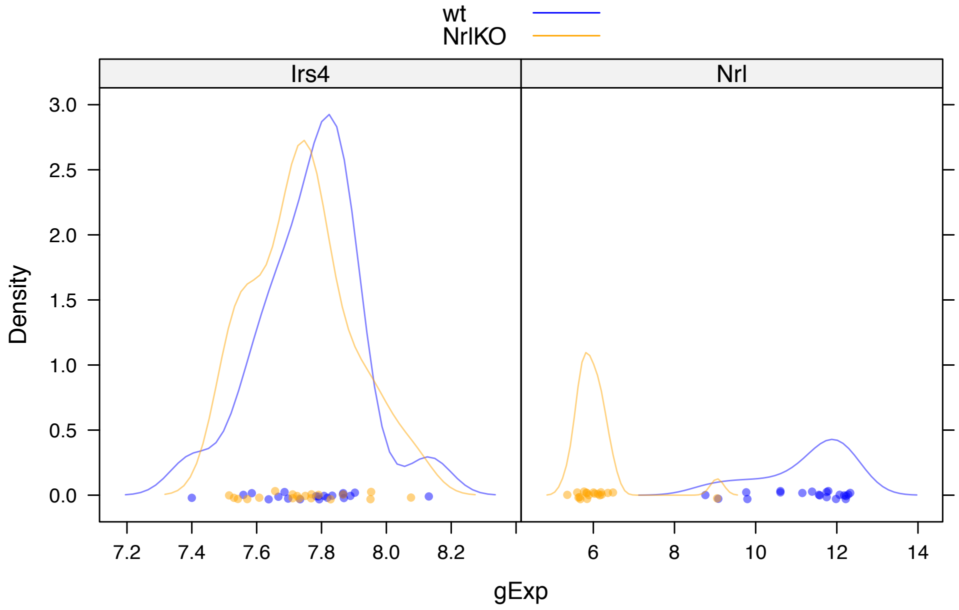
H_0: the expression level of gene A is the same

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

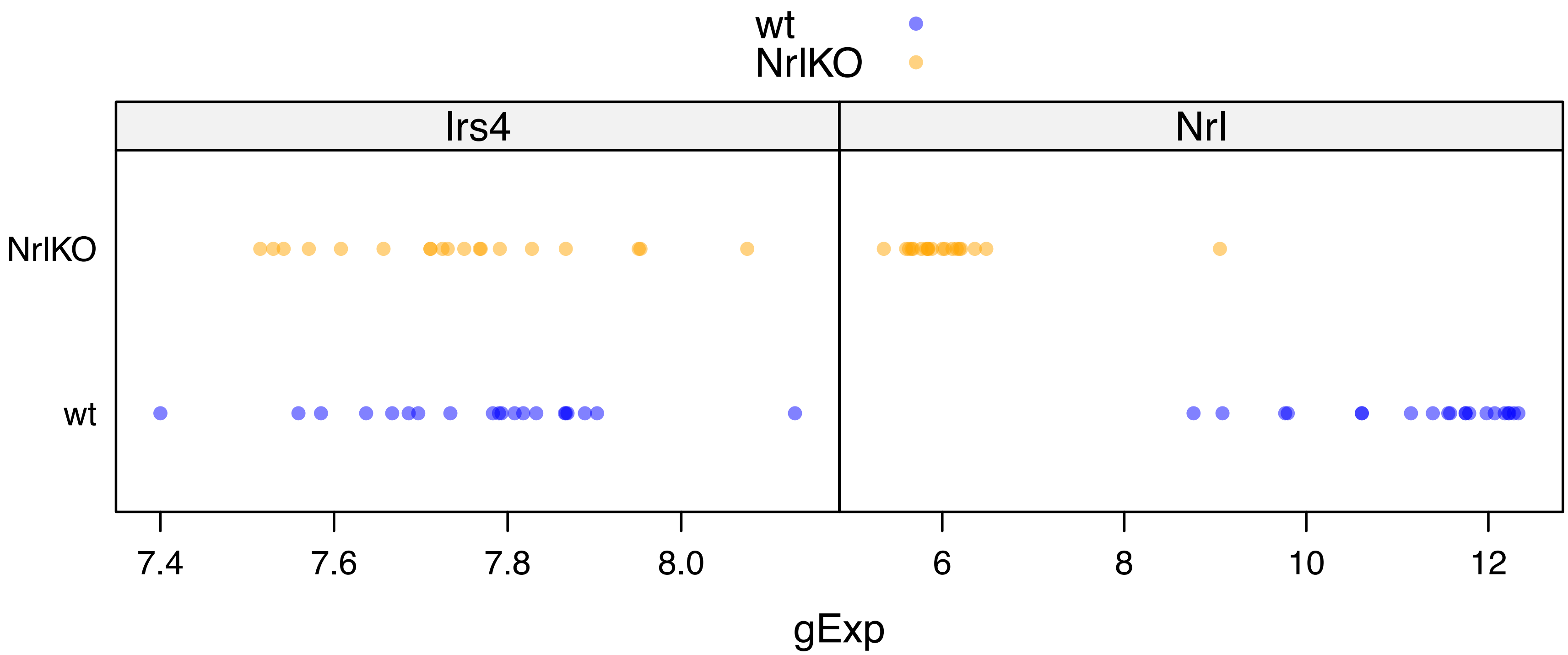


Empirical distributions

The true underlying distribution is ***unknown***

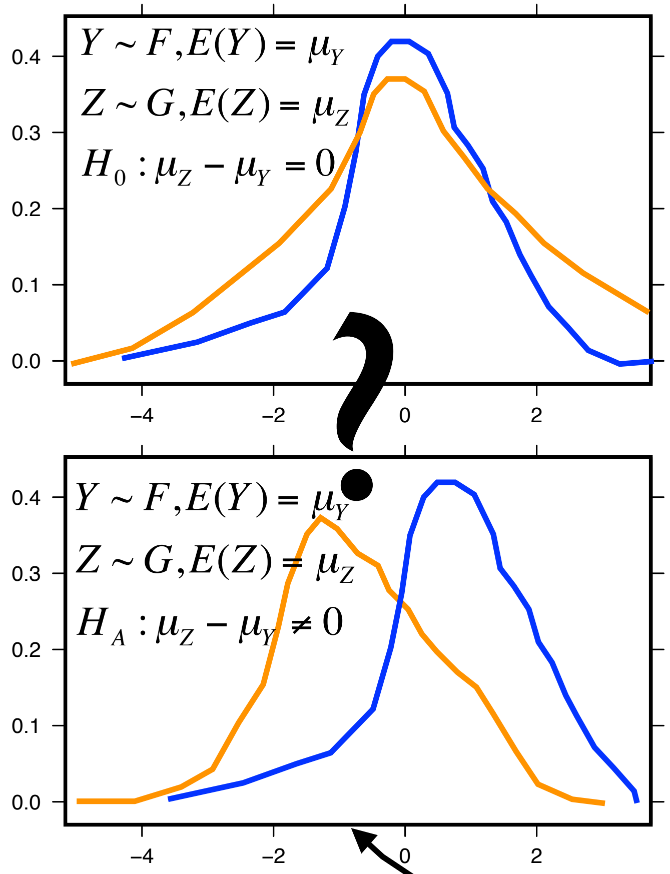


Based on DATA

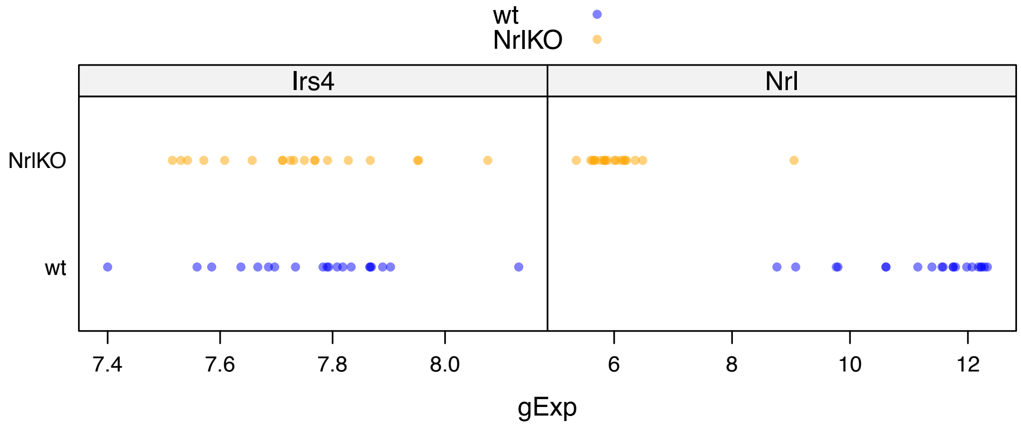


probability

data generating model



observed data



We will use a *statistical test* to study differences in the population based on our observed sample

statistical inference

Notation

- **Random variables (we can observe)**

Y_i : expression of gene A in the wt sample i

Z_i : expression of gene A in NrlKO sample i

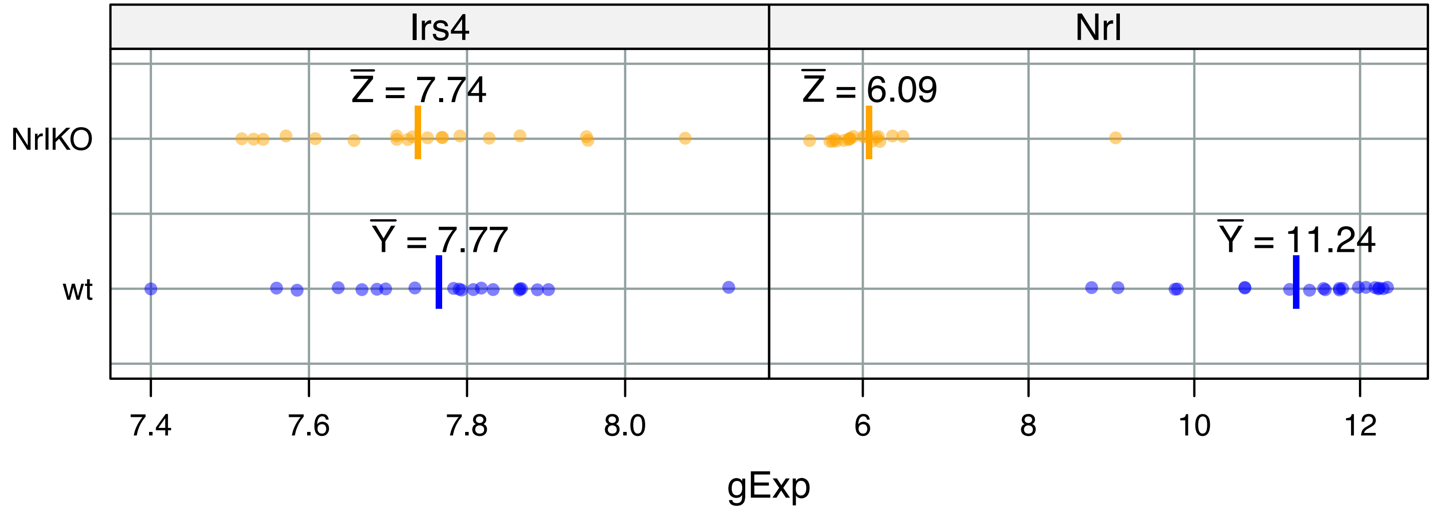
$Y_1,\ Y_2, \ldots, Y_{n_Y}$: a random sample of size n_Y

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

- **Population parameters (unknown)**

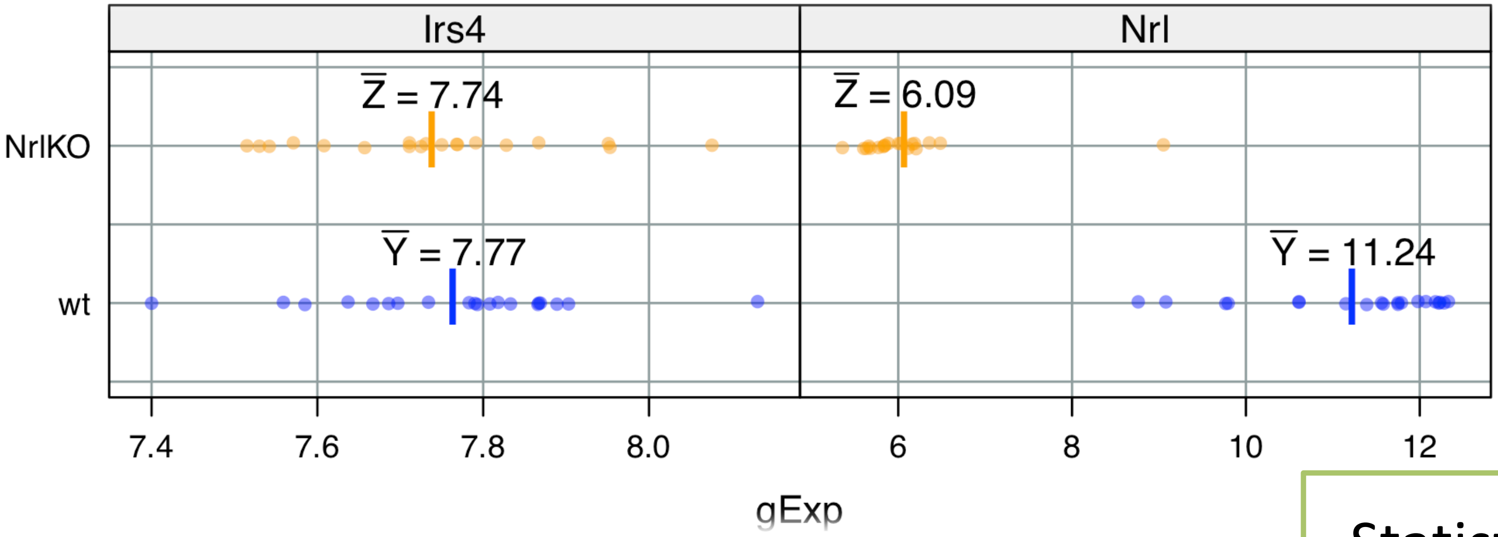
$\mu_Y = E[Y]$: the (population) expected expression level of gene A in wt mice

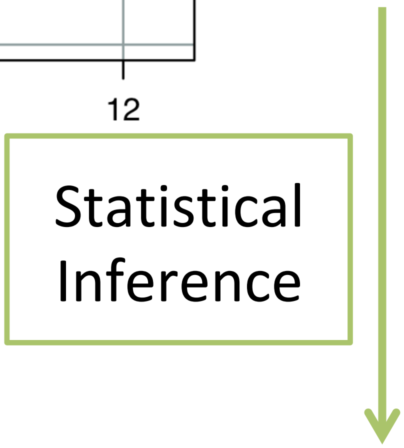
We observe… the difference between the **sample averages**!



```
> (theAvgs <- with(miniDat,
+                  tapply(gExp, list(gType, gene), mean)))
          Irs4       Nrl
wt    7.765750 11.244200
NrlKO 7.739684  6.089632

> (theDiff <- theAvgs["NrlKO", ] - theAvgs["wt", ])
       Irs4         Nrl
-0.02606579 -5.15456842
```

Is this convincing evidence that $\mu_Y \neq \mu_Z$???





Sample estimates

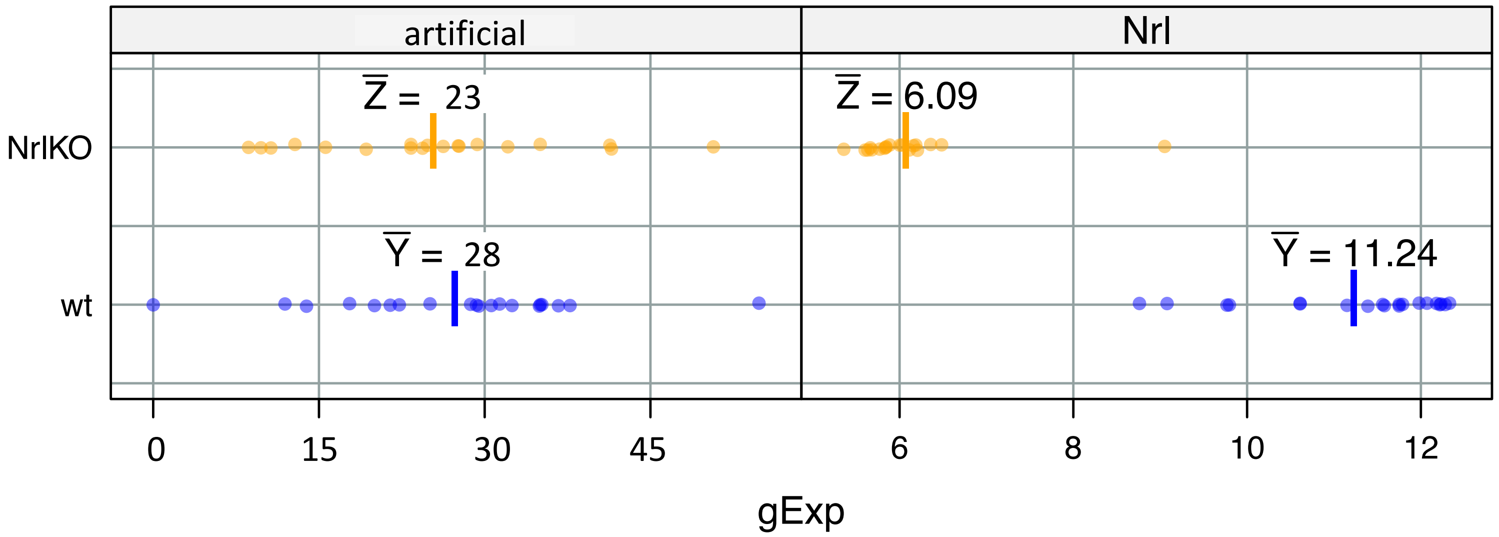
Statistical Inference

Population parameters

$$H_0 : \mu_Y = \mu_Z$$

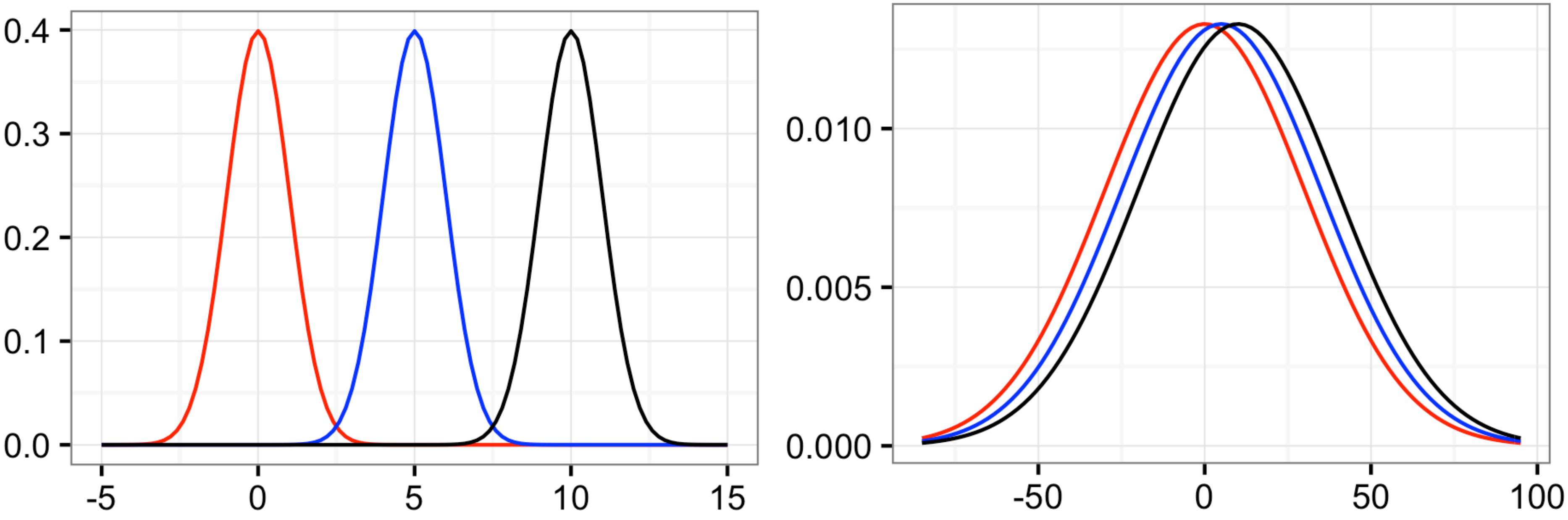
$$H_A : \mu_Y \neq \mu_Z$$

Is this convincing evidence that $\mu_Y \neq \mu_Z$???



- The sample means by themselves are not enough to make conclusions about the population
- What is a “large” difference? “large” relative to what?

Same centers, different variances



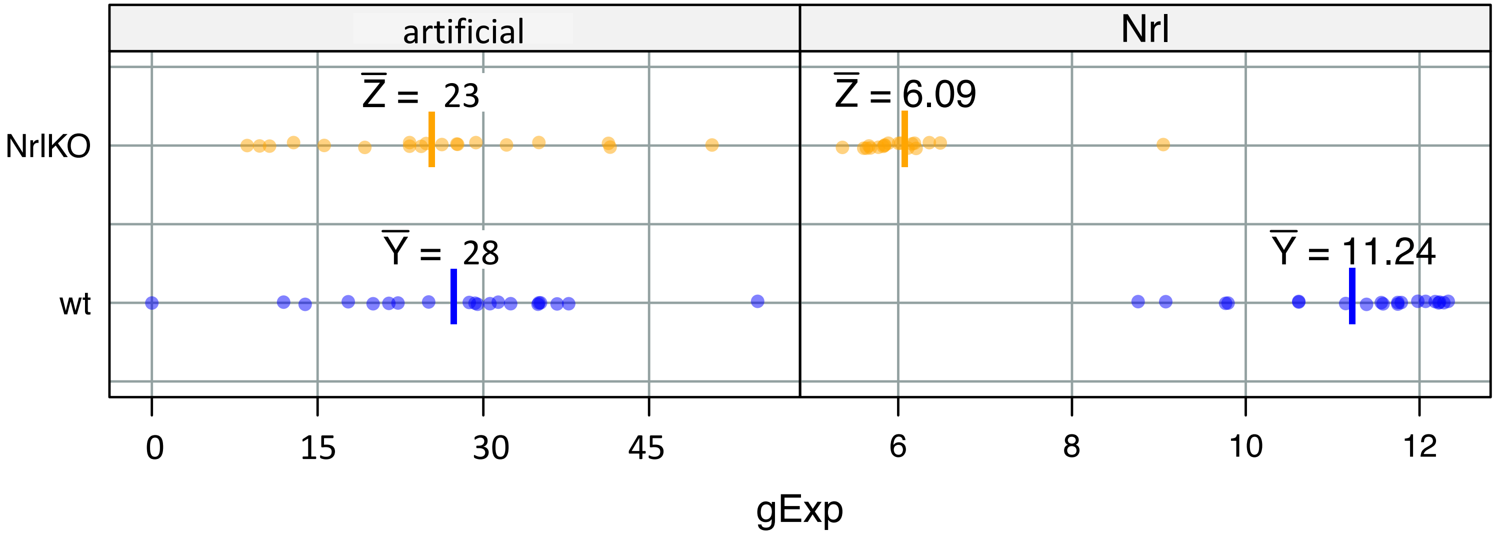
What do we want to know to help us interpret the mean difference?

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

What do we want to know to help us interpret the mean difference?

"Large" relative to the observed variation

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



Assuming that the random variables of each group are independent and identically distributed (iid):

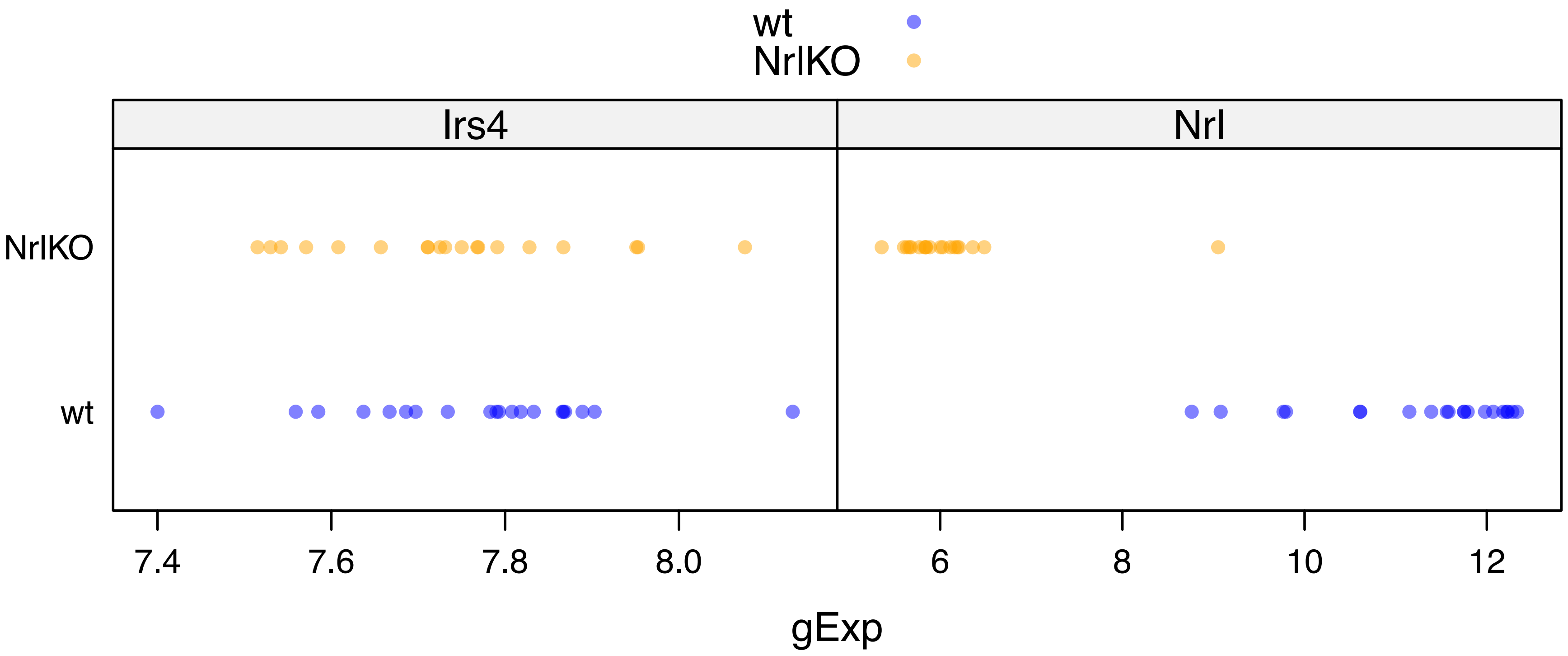
- $Y_1,\ Y_2,\ldots,Y_{n_Y}$ are iid
- $Z_1,\ Z_2,\ldots,Z_{n_Z}$ are iid
- $Y_i,\ Z_j$ are independent

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

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

$$\boxed{\begin{aligned} V(\bar{Z}-\bar{Y}) &= \frac{\sigma_Z^2}{n_Z} + \frac{\sigma_Y^2}{n_Y} \\ &\overset{???}{=} \enclose{circle}{\sigma^2}\left[\frac{1}{n_Z} + \frac{1}{n_Y}\right] \end{aligned}}$$

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



```
> (theVars <- with(miniDat,
+                    tapply(gExp, list(gType, gene), var)))
             Irs4       Nrl
wt     0.02403557 1.2243331
NrlKO  0.02332078 0.5942802
```

$$S_Y^2 = \frac{1}{n_Y}\sum_{i=1}^{n_Y}(Y_i - \bar{Y})^2$$

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

$$\text{"pooled" } \hat{\sigma}^2 = s_Y^2 \frac{n_Y - 1}{n_Y + n_Z - 2} + s_Z^2 \frac{n_Z - 1}{n_Y + n_Z - 2}$$

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

assuming unequal variance of Y’s and Z’s

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

The « hat » means « estimate »

$$T = \frac{\bar{Z}_n - \bar{Y}_n}{\hat{\sigma}_{\bar{Z}_n - \bar{Y}_n}}$$

T is a test statistic

Assuming equal variances

```
> (tstStat <- theDiff / sqrt(s2Diff))
       Irs4         Nrl
 -0.5286494 -16.7947224
```

Without assuming equal variances

```
> (welchStat <- theDiff / sqrt(s2DiffWelch))
       Irs4         Nrl
 -0.5288595 -16.9486146
```

Now can we say the observed differences are “big”?

The difference is about half a standard deviation for Irs4 and 16 or 17 standard deviations for Nrl.

I predict we will conclude that true means are same for Irs4 and different for Nrl.

The test statistic *T* is a *random variable* because is based on our random sample.

We need a measure of its uncertainty to determine how big/small *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!! But this is unknown to us!

We need to make more assumptions!!

Theory now tells us specific null distributions for this test statistic, depending on your assumptions.

Willing to assume that F and G are normal distributions?

eq var

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

uneq var

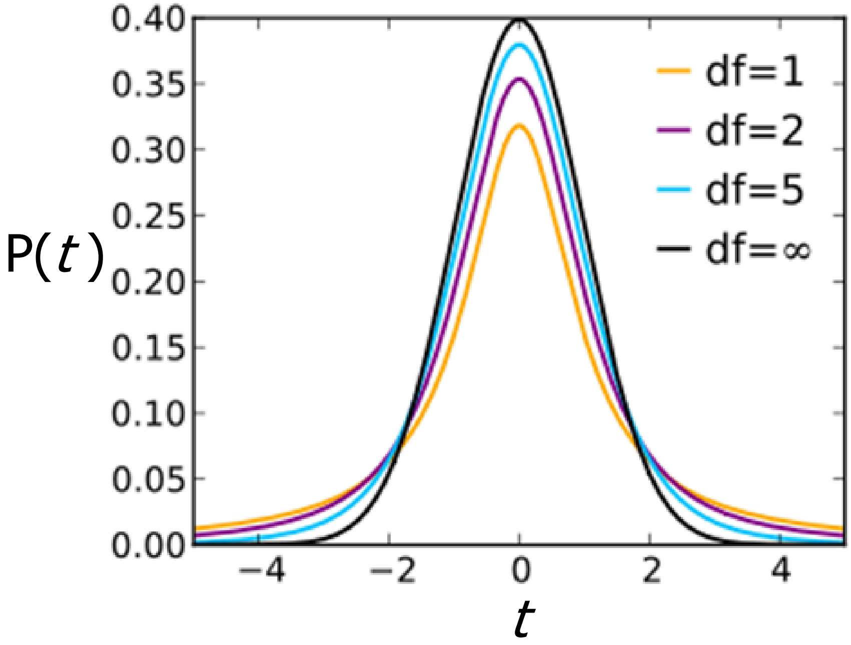
$$T \sim t_{\text{<sthg ugly>}}$$

“Welch’s t test”

Unwilling to assume that F and G are normal distributions? But you feel n_Y and n_Z are “large enough”? Then go right ahead use the t dist’n above or even a normal distribution as a decent approximation.

Student’s t-distribution

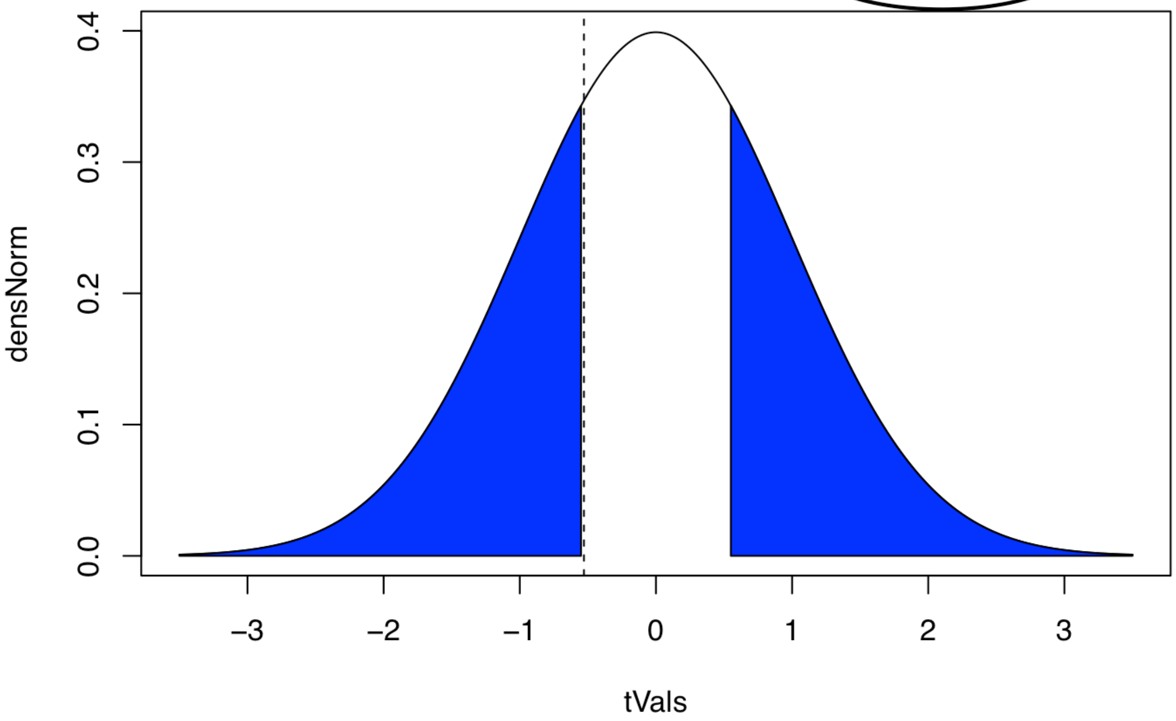
Recall that *T* is a random variable. Under certain assumption, we can prove that *T* follows a t-distribution



df=degrees of freedom

Assuming that H_0 is true:

```
> (tstStat <- theDiff / sqrt(s2Diff))
      Irs4         Nrl
-0.5286494 -16.7947224
```



We see that prob. of seeing a test stat as or more extreme than observed (T = -0.53) is pretty high.

Then, we don’t have *enough* evidence to 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>

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

    Welch Two Sample t-test

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

Hypothesis testing

1. Define a test-statistic (*T*).
2. Compute the observed value for the test statistic based on our random sample.
3. Compute p-value for the observed statistic under its null sampling distribution.
4. Make a decision about significance of results, based on a pre-specified value (alpha, significance level)

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
  - Doe not indicate the size or importance of observed effects.

(From Dr. 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 kind of *t*-tests:

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

Errors in hypothesis testing

	Actual Situation “Truth”	
Decision	**H_0 True**	**H_0 False**
Don Not Reject H_0	Correct Decision 1-α	Incorrect Decision Type II Error β
Reject H_0	Incorrect Decision Type I Error α	Correct Decision 1-β

α = P(Type I Error)   β = P(Type II Error)

Power = 1 - β

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 obtain statistical significance. Later lecture on this.

Alternatively, there are nonparametric tests that are available here:

Wilcoxon rank sum test, aka Mann Whitney, uses ranks

Kolmogorov-Smirnov uses the empirical CDF

Wilcoxon test

Rank all the data, ignoring the grouping variable

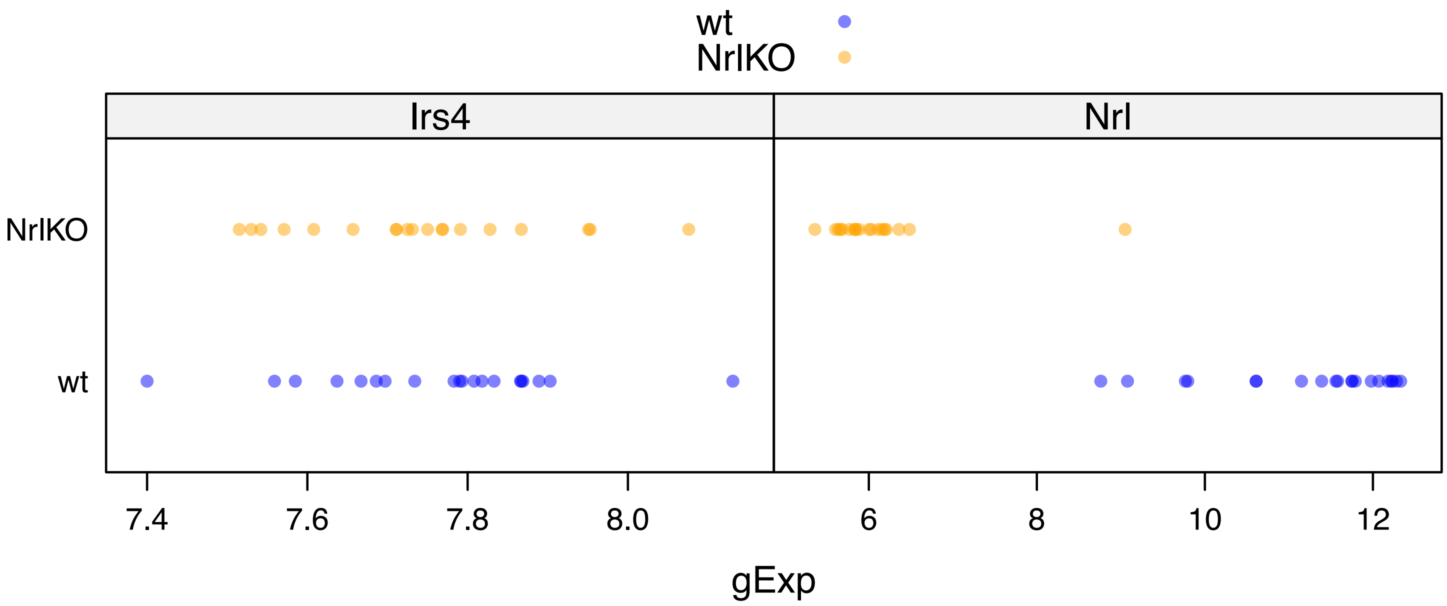
Test stat = sum of the ranks for one group (optionally, subtract the minimum possible which is nY (nY + 1)/2)

(Alternative but equivalent formulation based on the number of yi, zi pairs for which yi >= zi)

Null distribution of such statistics can be worked out or approximated

```
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. distributions are same

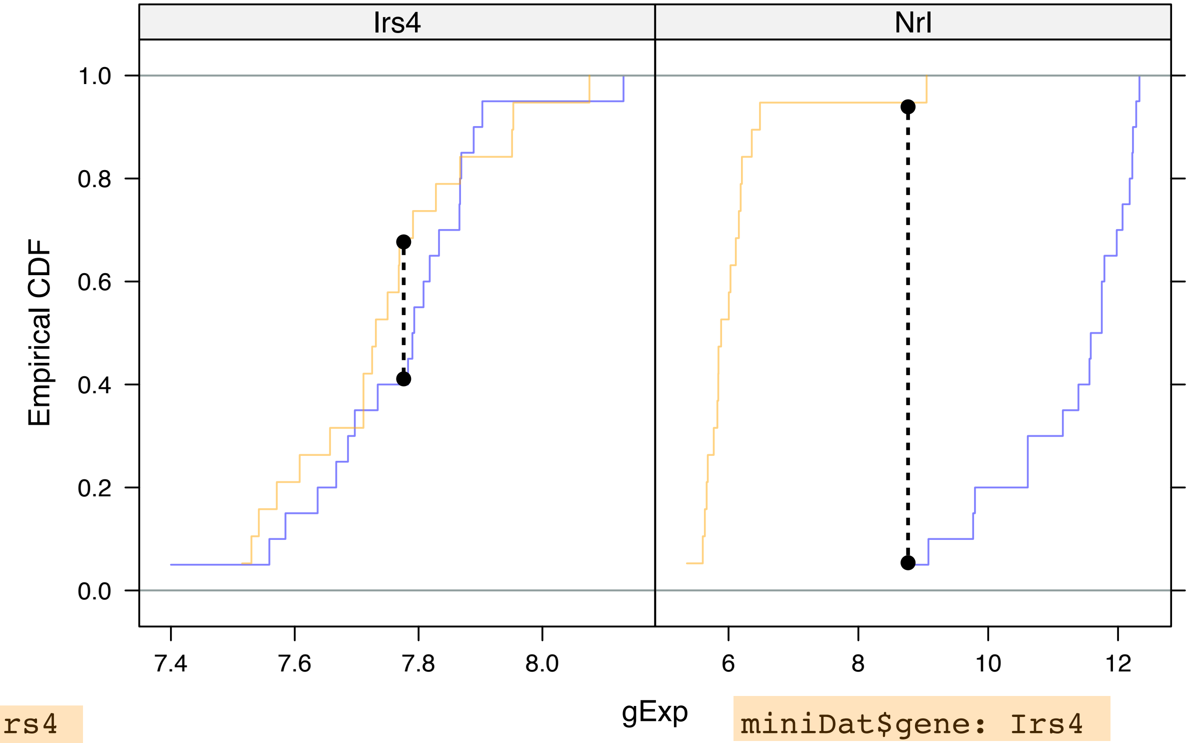
Estimate each CDF with the empirical CDF (ECDF)

$$\hat{F}(x) = \frac{1}{n}\sum_i I[x_i \le x]$$

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

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

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



```
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

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

     Two-sample Kolmogorov-Smirnov test

data:  theDat$gExp[theDat$gType == "wt"] and theDat
$gExp[theDat$gType == "NrlKO"]
D = 0.95, p-value = 4.603e-08
alternative hypothesis: two-sided
```

```
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>
```

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 two 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.