STAT 540
Class meeting 06
Wednesday, January 21, 2015

Dr. Gabriela Cohen Freue
Department of Statistics

Based on previous preparation by Dr. Jennifer (Jenny) Bryan

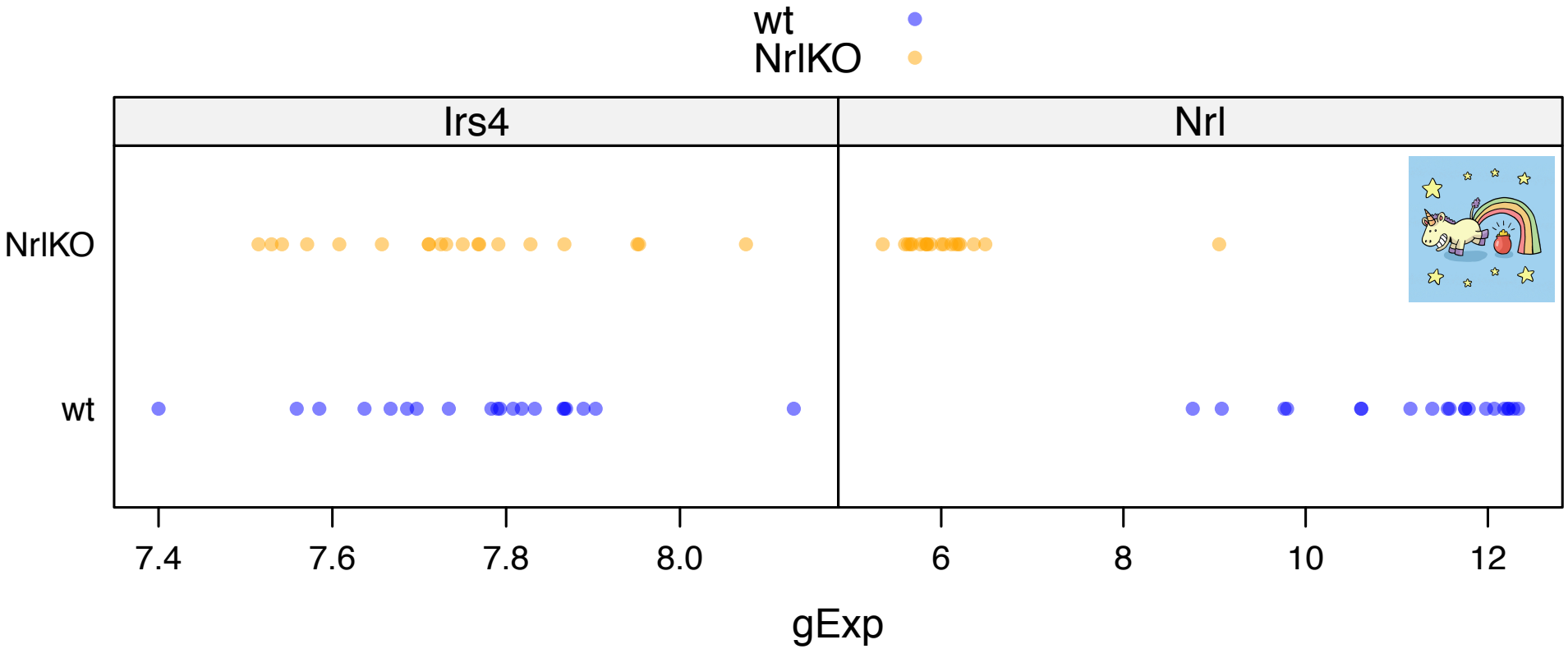
Two group comparisons

CC BY NC

```
> str(prDes)
'data.frame':  39 obs. of  3 variables:
 $ sample  : num  20 21 22 23 16 17 6 24 25 26 ...
 $ devStage: Factor w/ 5 levels "E16","P2","P6",..: 1 1 1 1 1 1 1 2 2 2 ...
 $ gType   : Factor w/ 2 levels "wt","NrlKO": 1 1 1 1 2 2 2 1 1 1 ...

> str(prDat, max.level = 0)
'data.frame':  29949 obs. of  39 variables:

> peek(subset(prDat, select = 1:5))
             Sample_20 Sample_21 Sample_22 Sample_23 Sample_16
1416535_at       8.133     8.143     7.899     8.054     7.867
1437399_at       8.567     8.554     7.931     8.182     6.257
1441587_at       6.134     5.745     6.137     5.953     6.575
1445975_at       6.022     5.960     5.994     6.069     6.418
1446741_at       6.024     6.009     6.073     5.961     7.046
1450103_a_at     8.376     8.902     8.570     8.755     7.991
1452844_at       8.490     8.700     8.288     8.544     7.256

> with(prDes, table(devStage, gType))
          gType
devStage  wt NrlKO
  E16      4     3
  P2       4     4
  P6       4     4
  P10      4     4
  4_weeks  4     4
```

photoRec dataset
mouse photoreceptors
Affy

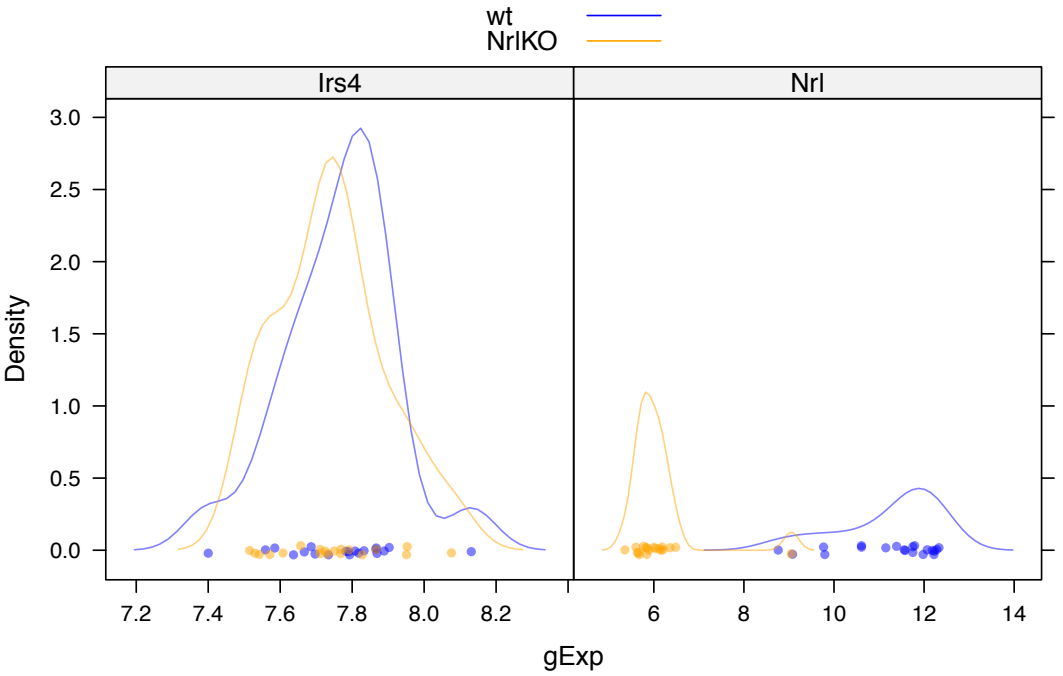
Do we think the orange's and blue's are generated by different underlying distributions?

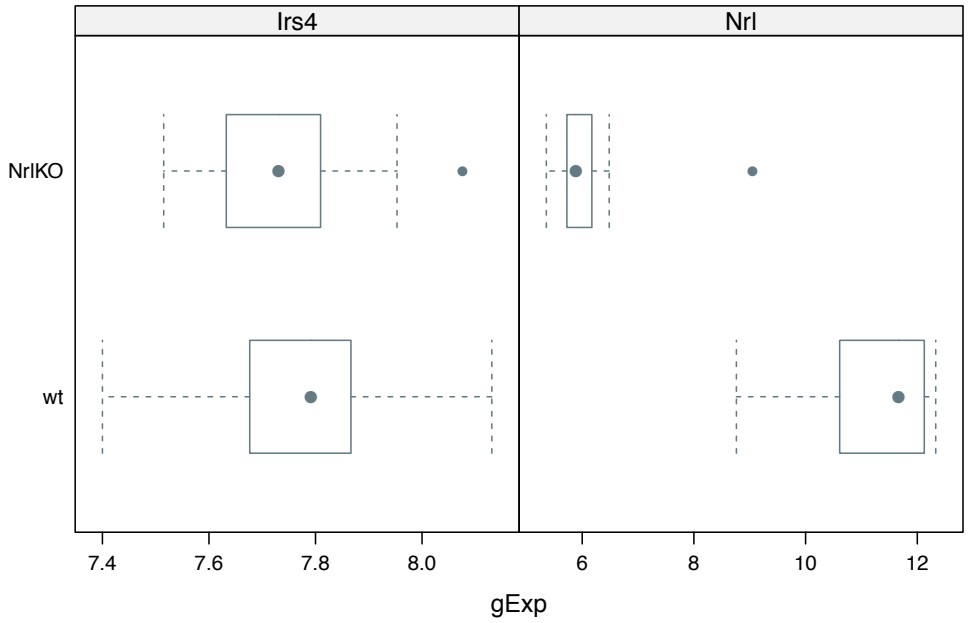


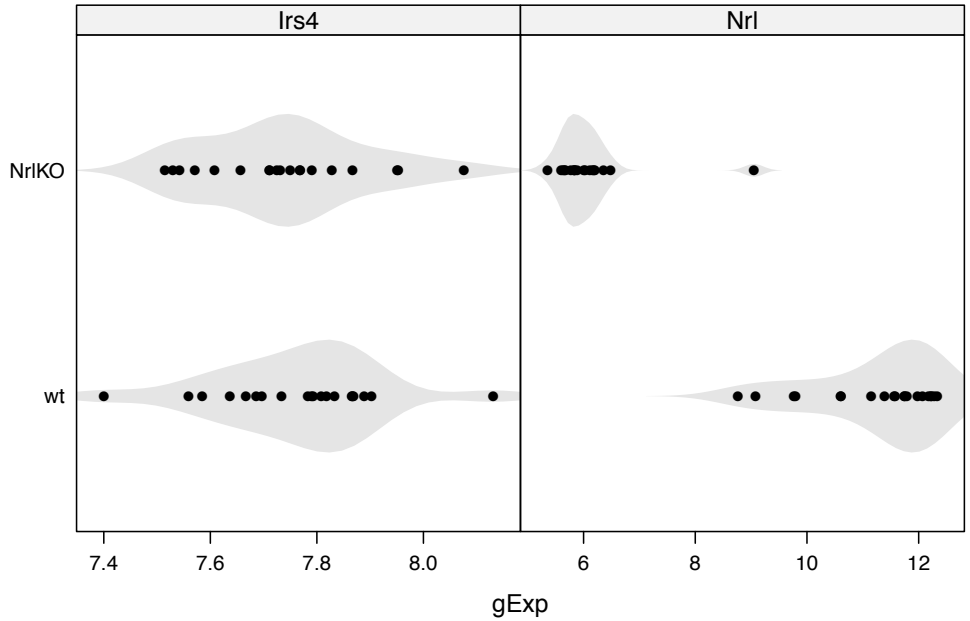
Irs4 (insulin receptor substrate 4) was selected at random as a boring non differentially expressed gene; NrlKO ~= wt

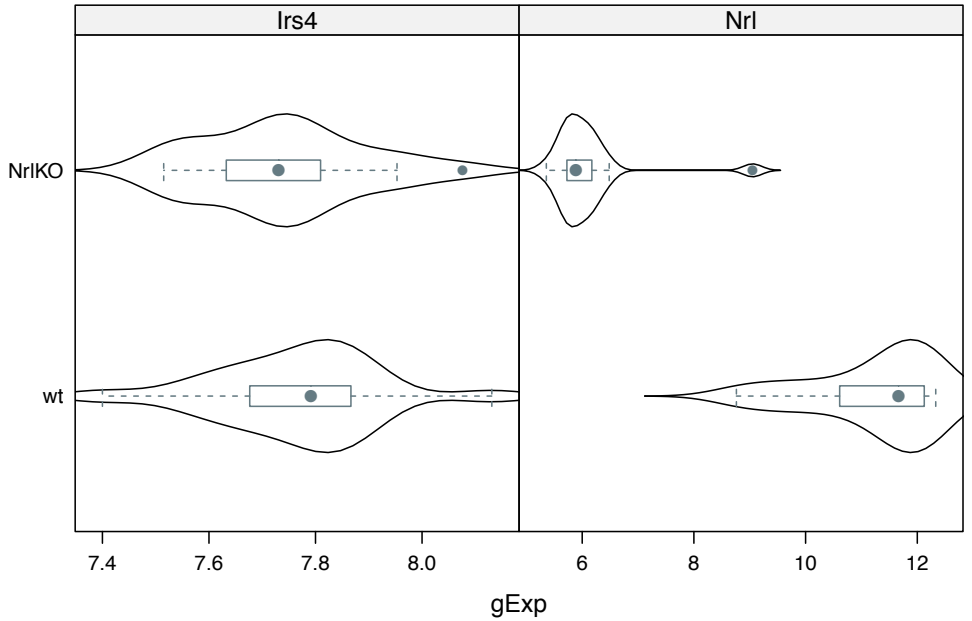
Nrl (neural retina leucine zipper gene) is the gene that was knocked out in half the mice; obviously should be differentially expressed; NrlKO << wt

Do we think expression of gene [Irs4|Nrl] in wild type vs. knockout mice comes from different underlying distributions?





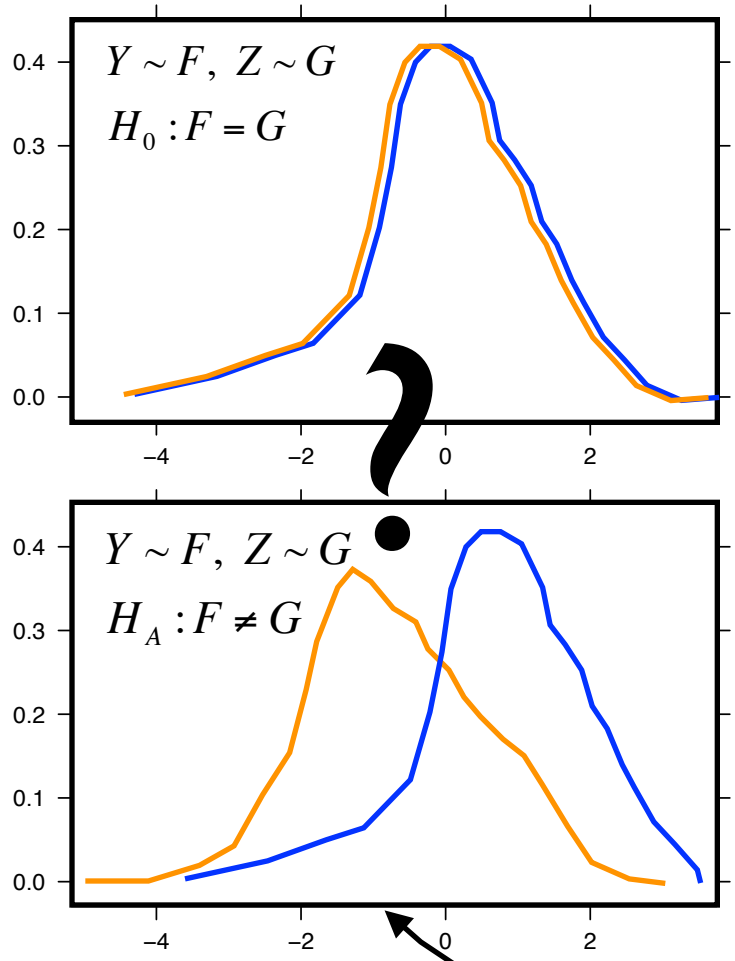




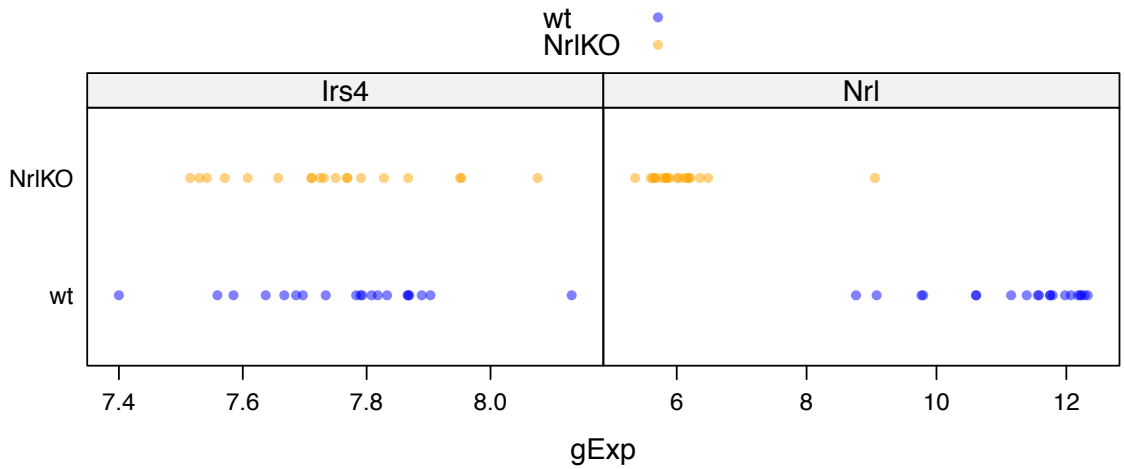
probability

data generating model

observed data



?



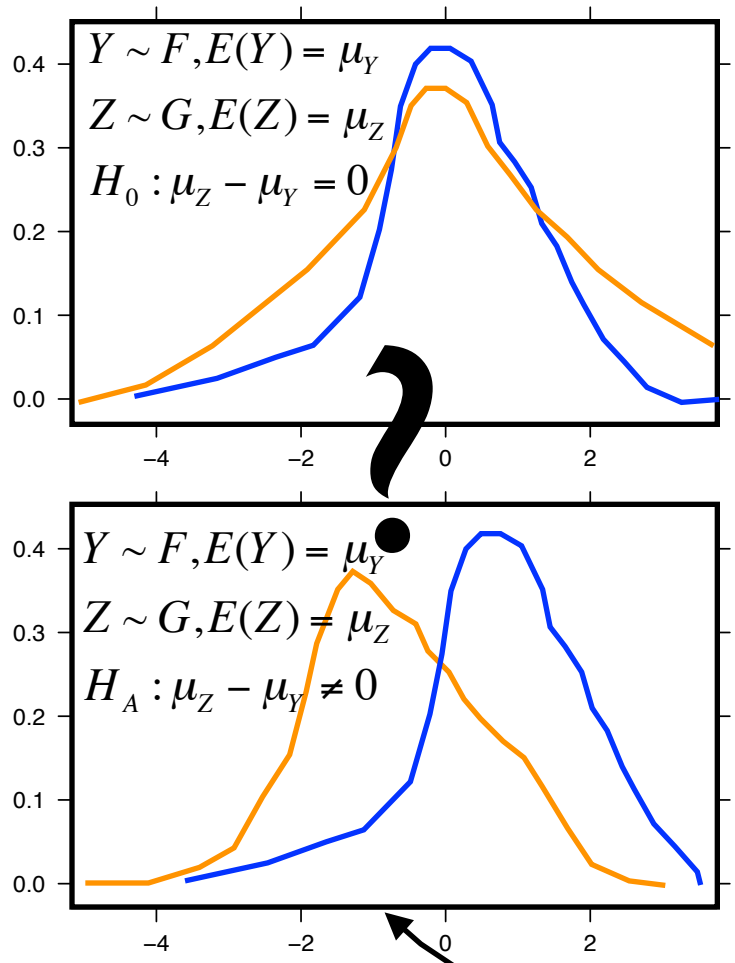
Pick your favorite explanation for the observed data .....

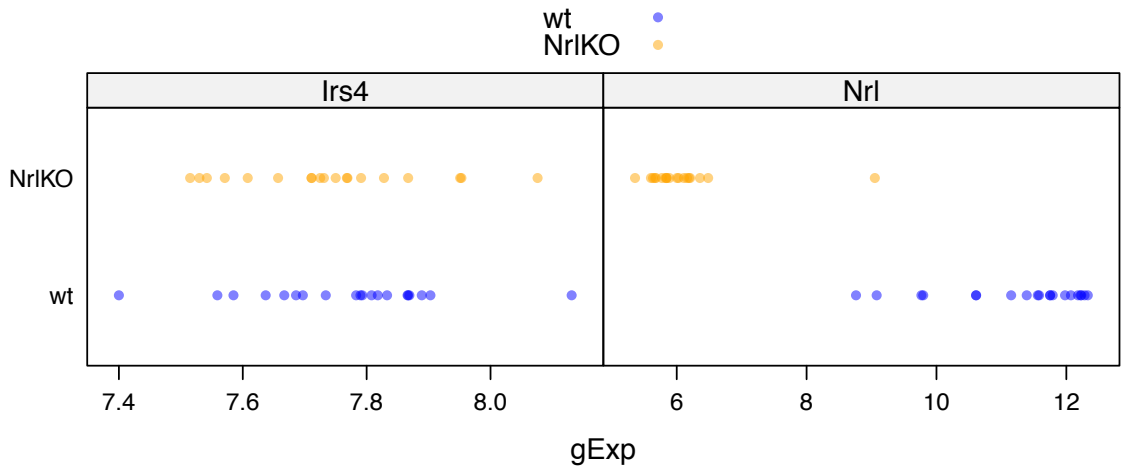
statistical inference

probability

data generating model

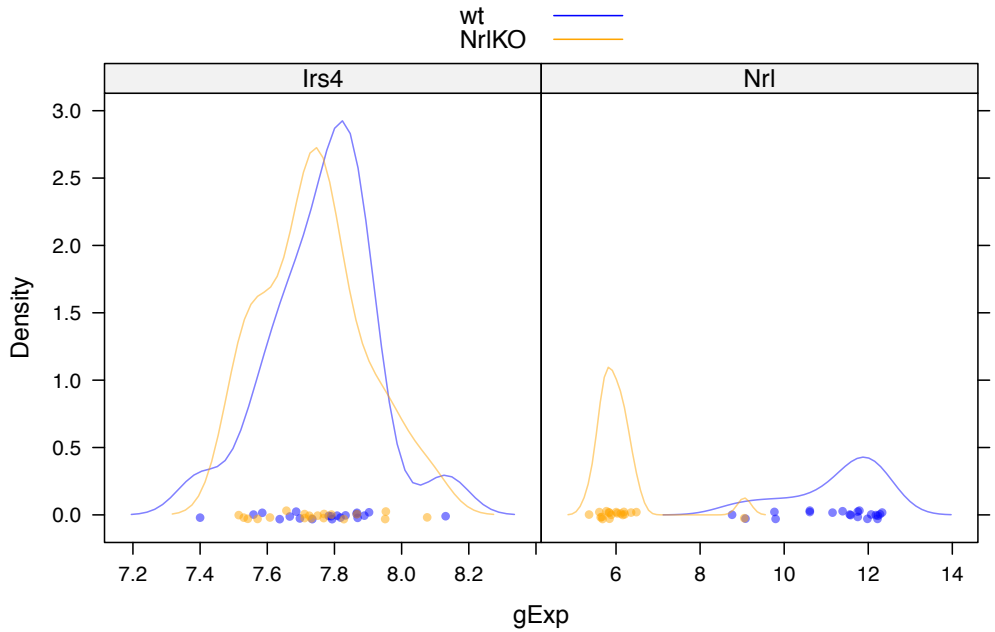
observed data

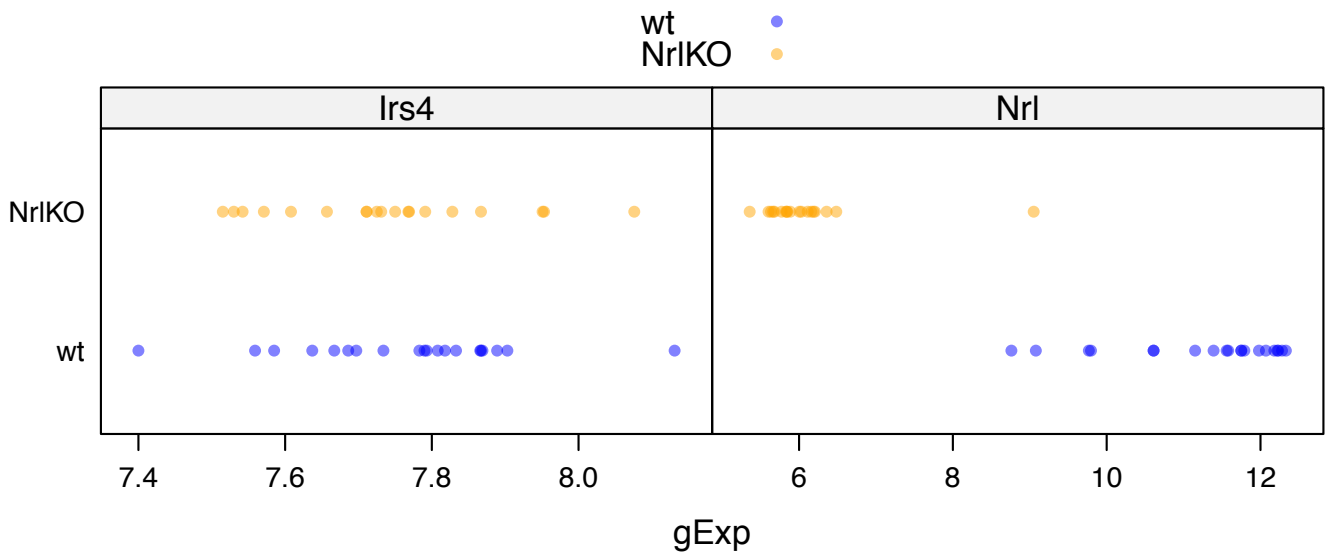




Common simplification: test only for equality of **means**

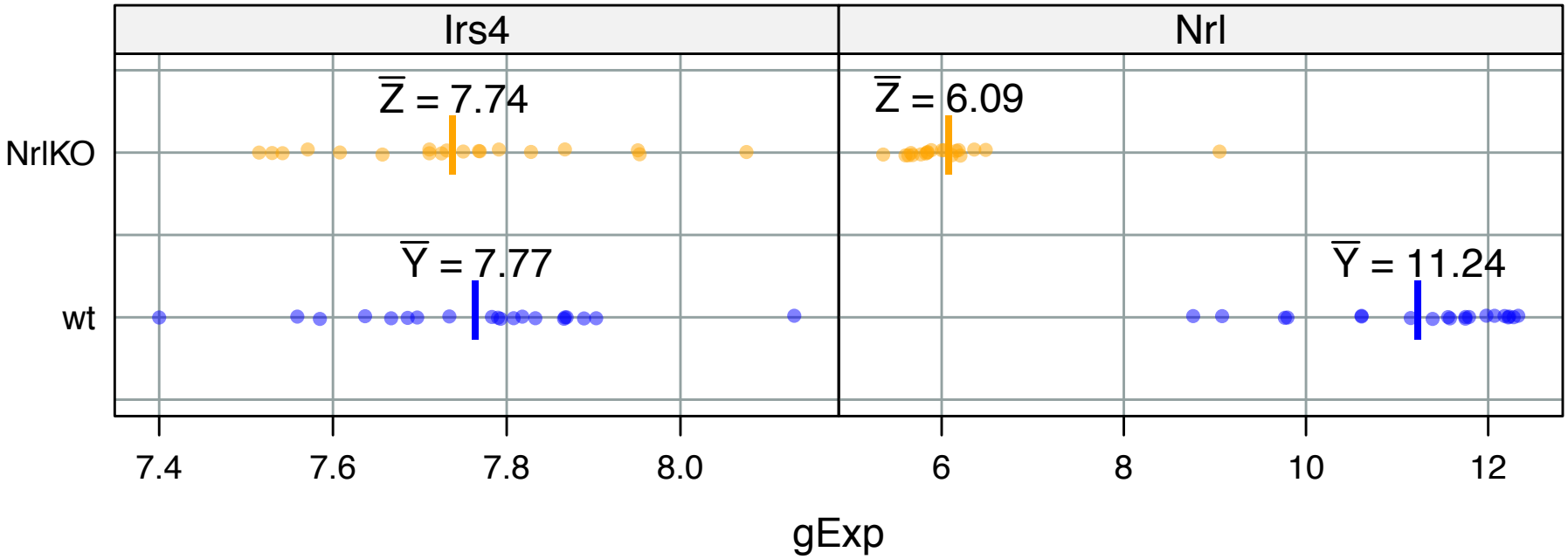
statistical inference





What's your quick-and-dirty best guess at $\mu_Z - \mu_Y$?

... 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
```

Are these observed differences **convincing** evidence that $\mu_Z - \mu_Y \neq 0$?

We need to know the background variability in the difference of sample averages under the null hypothesis that $\mu_Z - \mu_Y = 0$.

Then we can divide by the relevant standard deviation -- also called a standard error, in this setting -- and have a better idea.

$$V(\bar{Z}_n - \bar{Y}_n) = V(\bar{Z}_n) + (-1)^2 V(\bar{Y}_n) + 2(-1)\mathrm{cov}(\bar{Y}_n, \bar{Z}_n) \quad [1]$$

$$= V(\bar{Z}_n) + V(\bar{Y}_n) - 2\,\mathrm{cov}(\bar{Y}_n, \bar{Z}_n)$$

$$= V(\bar{Z}_n) + V(\bar{Y}_n) \quad [2]$$

$$= \frac{\sigma_Z^2}{n_Z} + \frac{\sigma_Y^2}{n_Y} \quad [3]$$

[1] basic probability result about variance of sums of scaled rvs

[2] by assuming the Y's and Z's are independent from each other, we get that covariance is zero

[3] basic result about variance of a mean of an iid sample

* See how independence assumptions are sprinkled everywhere?

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

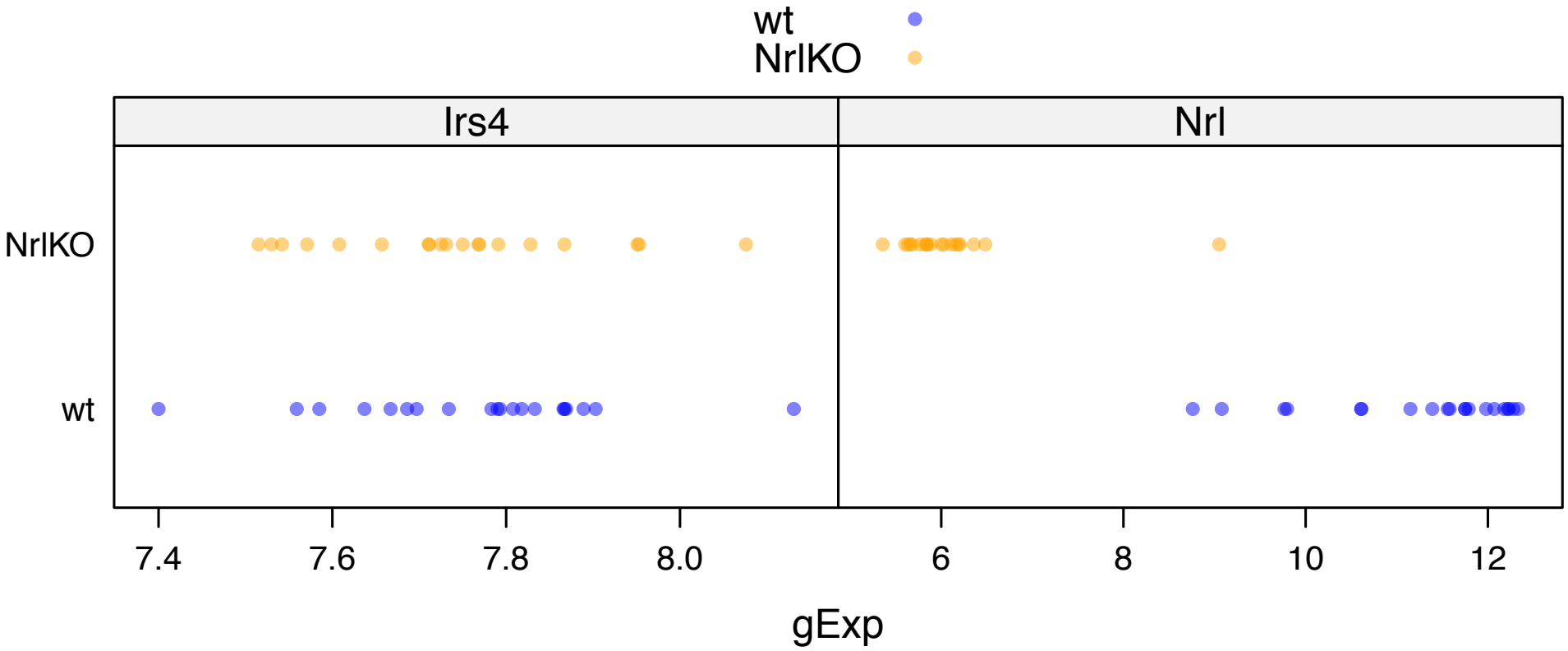
if we assume that $\sigma^2_Z = \sigma^2_Y = \sigma^2$

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

$$= \sigma^2 \left[\frac{1}{n_Z} + \frac{1}{n_Y} \right]$$

What's your quick-and-dirty best guess at σ^2?

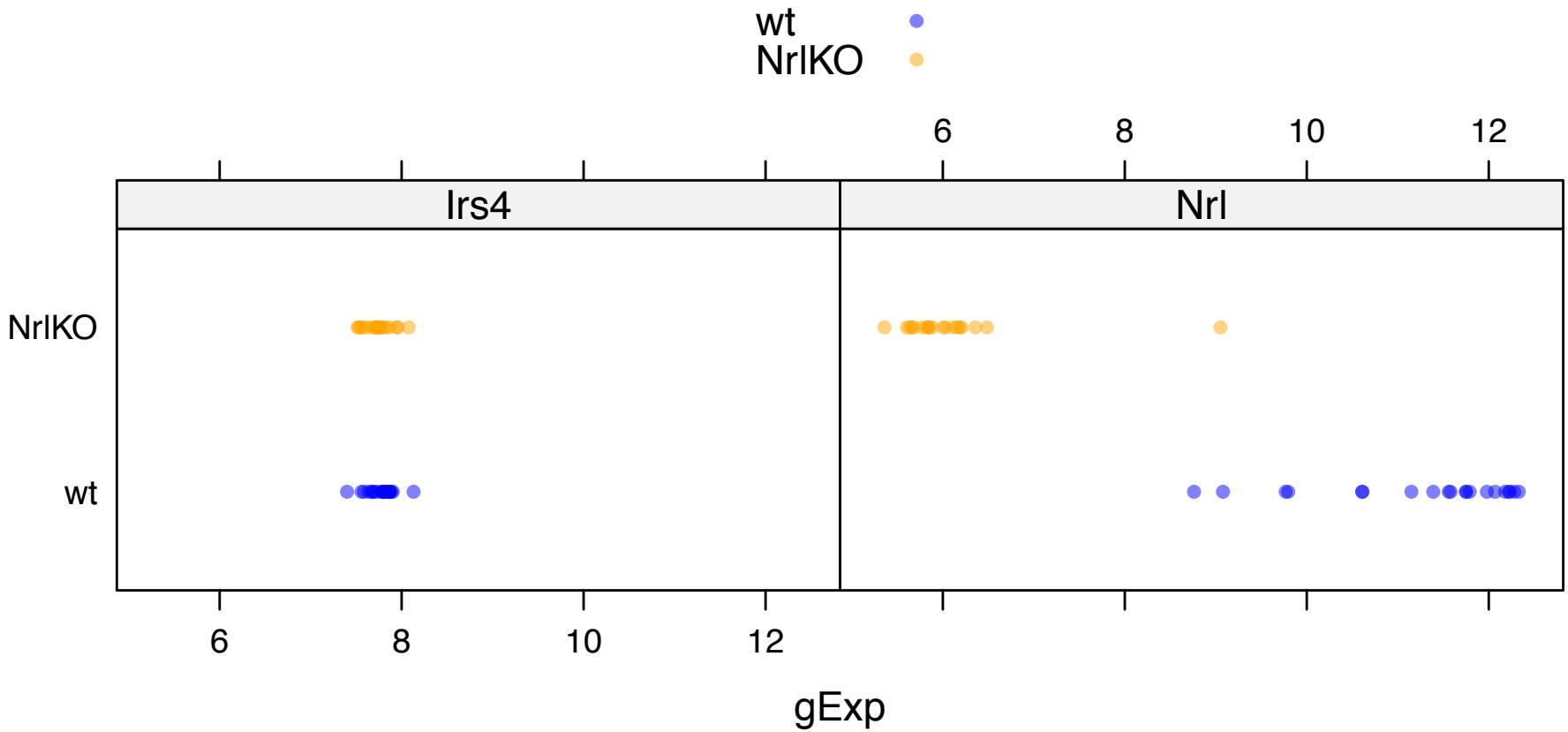
... the sample variances (combined, somehow)!




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

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2$$

... 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
```

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2$$

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

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

```
> (nY <- with(miniDat, sum(gType == "wt" & gene == "Nrl")))
[1] 20
> (nZ <- with(miniDat, sum(gType == "NrlKO" & gene == "Nrl")))
[1] 19
```

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

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

```
> (s2Pooled <- colSums(theVars * c((nY - 1) / (nY + nZ - 2),
+                                  (nZ - 1) / (nY + nZ - 2))))
      Irs4        Nrl
0.02368783 0.91782091

> (s2Diff <- s2Pooled * (1/nY + 1/nZ))
      Irs4        Nrl
0.00243112 0.09419741
```

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

```
> (s2DiffWelch <- colSums(theVars / c(nY, nZ)))
      Irs4        Nrl
0.002429188 0.092494563
```

Now we can compute the observed difference in sample mean divided by our best guess at it's standard deviation under H_0, i.e. we can report the observed difference in appropriate "sd" units.

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

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

R's default is to NOT assume equal variance, i.e. to perform "Welch's Two sample 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

<snip, snip>
```

We have just re-derived the two sample t test statistic.

Now we can compute the observed difference in sample mean divided by our best guess at it's standard deviation under H_0, i.e. we can report the observed difference in appropriate "sd" units.

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

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

It is also easy to do a t-test assuming common variance.

```
> 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
<snip, snip>
```

We have just re-derived the two sample t test statistic.

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

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


> (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.

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

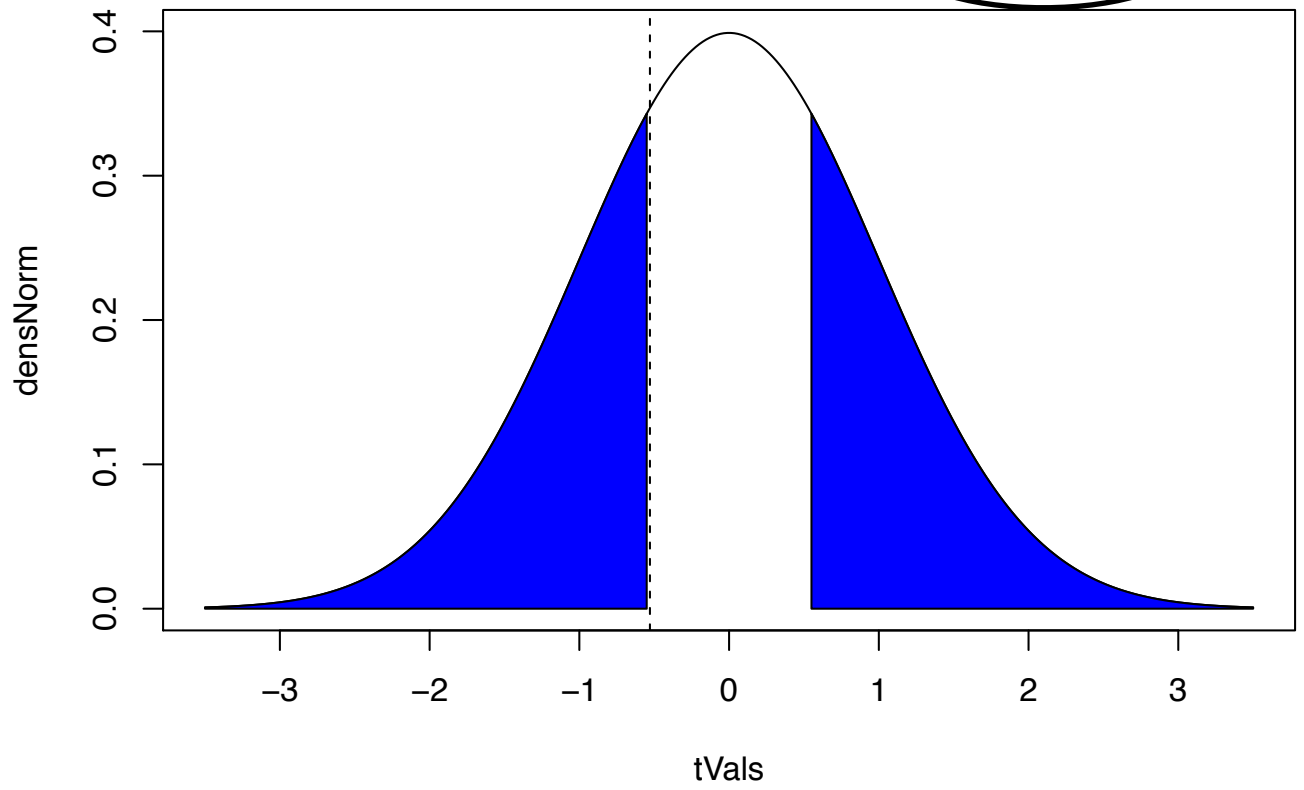
$$T \sim t_{\text{<sthg ugly>}}$$

"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.

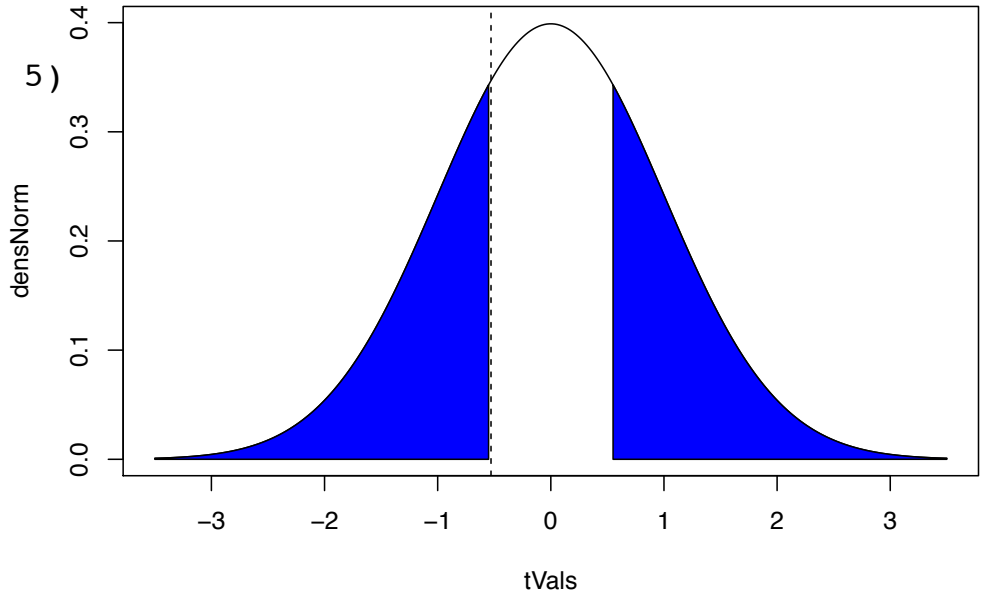
Depicted here is the standard normal distribution (which is visually indistinguishable from t w/ 58 df).

```
> (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.

```
> round(pt(-1 * abs(tstStat), df = nY + nZ - 2) * 2, 5)
   Irs4     Nrl
0.60021 0.00000

> round(pnorm(-1 * abs(tstStat)) * 2, 5)
   Irs4     Nrl
0.59705 0.00000
```




```
miniDat$gene: Irs4

        Two Sample t-test

data:  gExp by gType
t = -0.5286, df = 37, p-value = 0.6002
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.07383844  0.12597002
sample estimates:
   mean in group wt mean in group NrlKO
           7.765750            7.739684
```

```
miniDat$gene: Irs4

        Welch Two Sample t-test

data:  gExp by gType
t = -0.5289, df = 36.948, p-value = 0.6001
alternative hypothesis: true difference in means is
95 percent confidence interval:
 -0.0738035  0.1259351
sample estimates:
   mean in group wt mean in group NrlKO
           7.765750            7.739684
```

we knew we'd see extreme statistical significance for Nrl ... and we do

```
miniDat$gene: Nrl

     Two Sample t-test

data:  gExp by gType
t = -16.7947, df = 37, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 4.532698 5.776439
sample estimates:
   mean in group wt mean in group NrlKO
          11.244200             6.089632


miniDat$gene: Nrl

     Welch Two Sample t-test

data:  gExp by gType
t = -16.9486, df = 34.005, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 4.536507 5.772630
sample estimates:
   mean in group wt mean in group NrlKO
          11.244200             6.089632
```

In real life, working with just one (or two) genes, it's hard to believe in your gut that a difference of sample means or a two sample t statistic has a null *distribution*. It feels like it's just a particular number -- e.g. t stat = 0.53 for Irs4 in our current example.

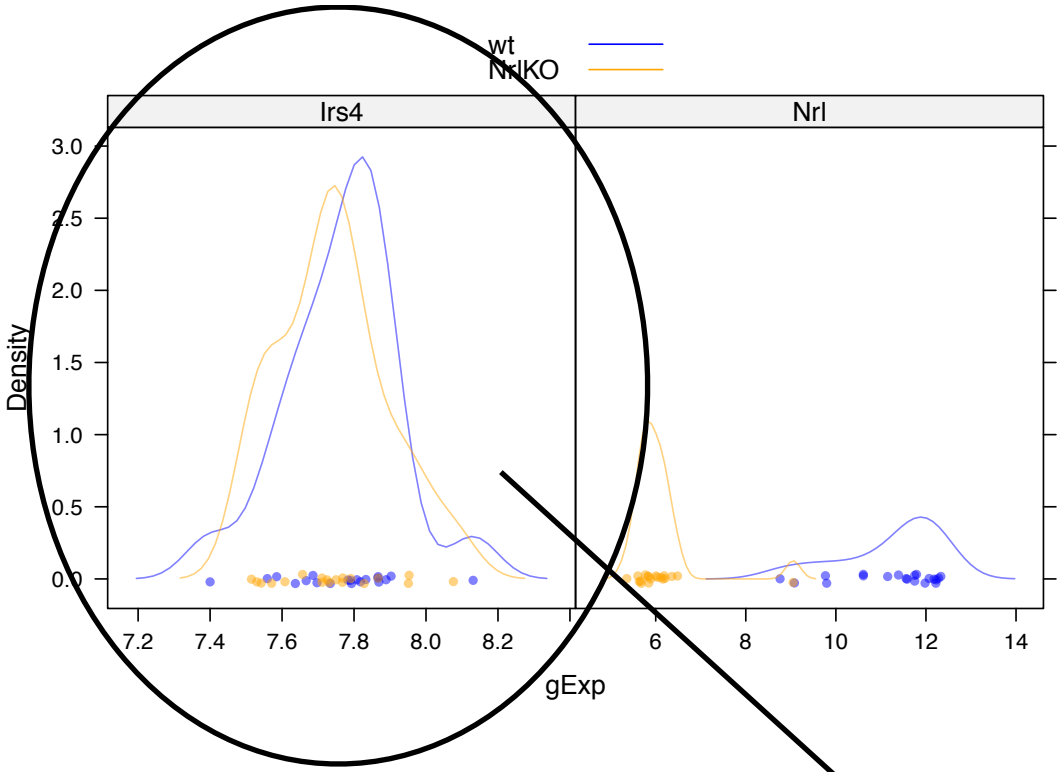
But you must think of it as a fleeting realization of a specific random variable.

You've simply observed one of an infinity of possible values and it's the underlying null distribution that speaks to that and puts your specific observation into context.

I will simulate data -- more blue Y's and more orange Z's -- and compute the observed difference of sample means and the t statistic.

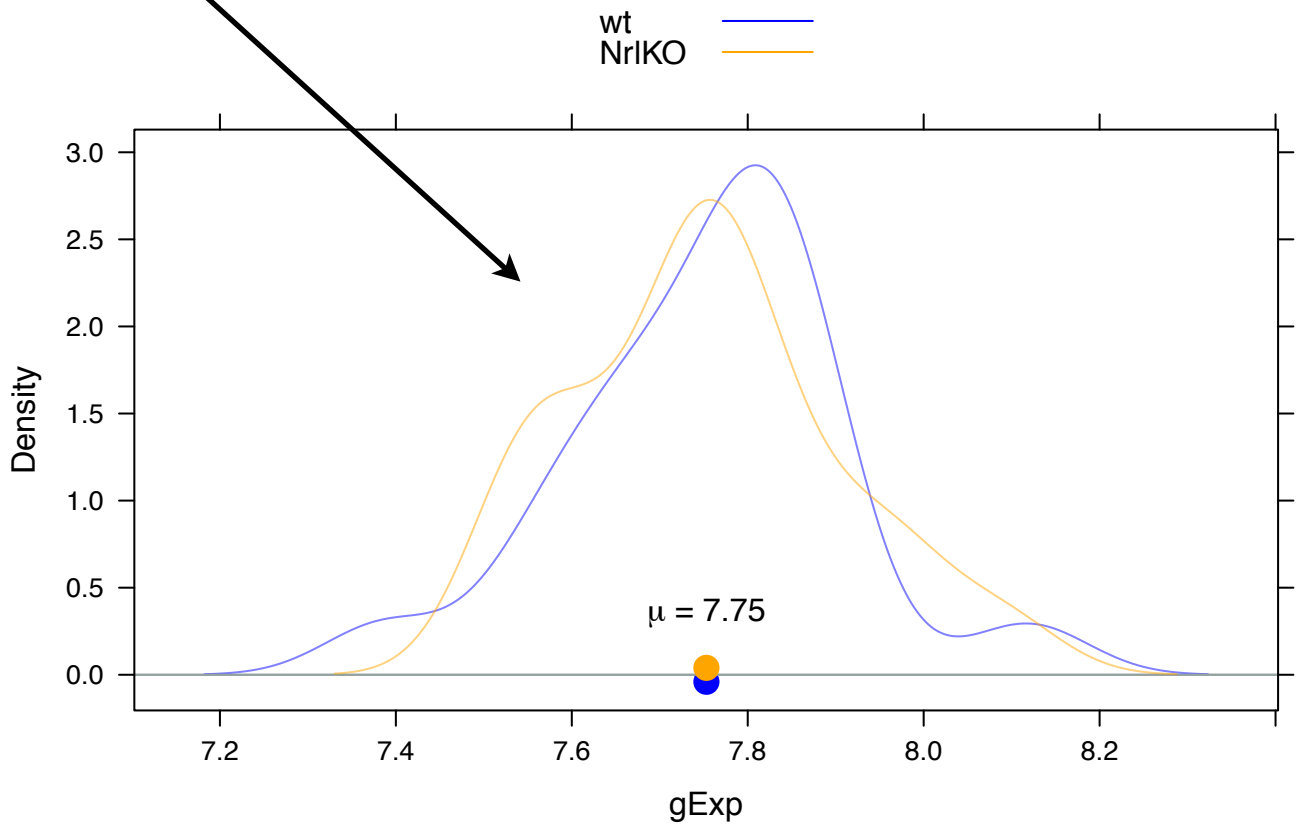
We'll compare the empirical distribution of this larger set of observations to the theoretical distributions just mentioned and used.

We'll feel really good about how this all works, at least when the *assumptions truly hold*.

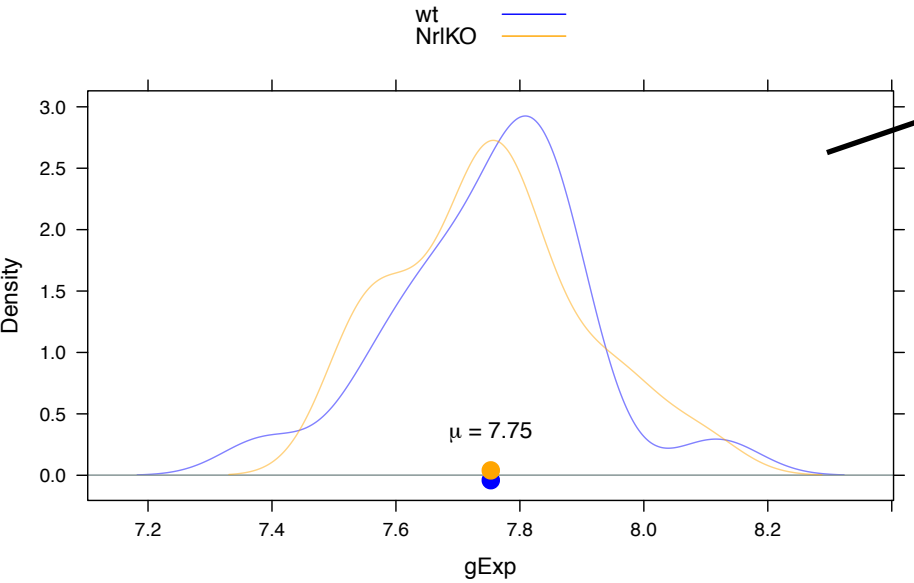


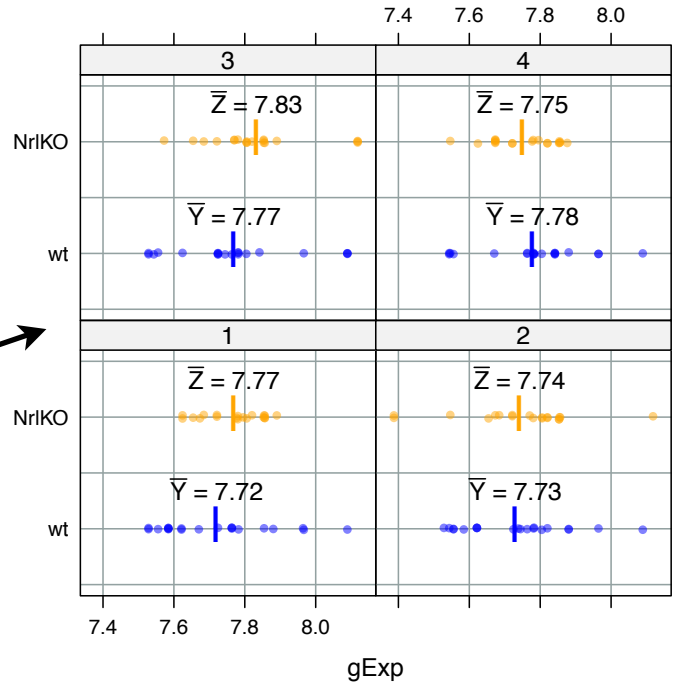
Our data-generating distributions are inspired by the observed data from Irs4.

Exact match except wt and NrlKO groups have common mean.



Underlying true dist'ns, upholding the null hypothesis of equal means

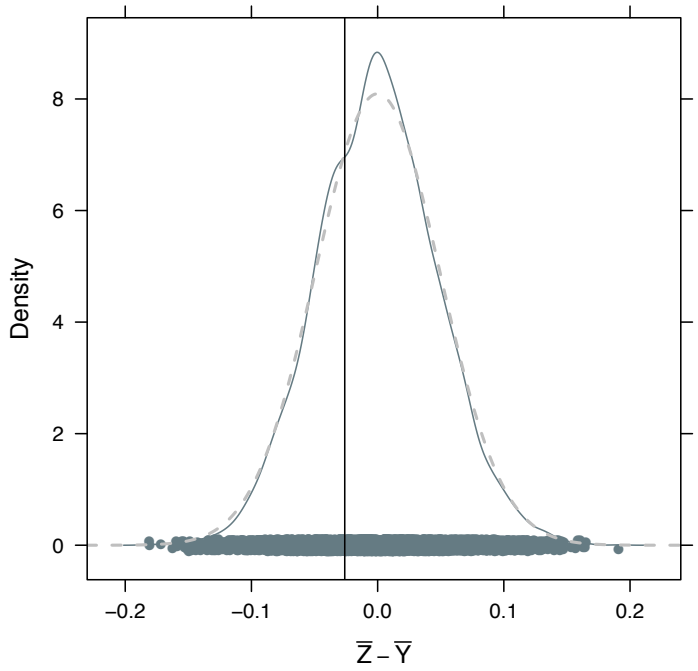


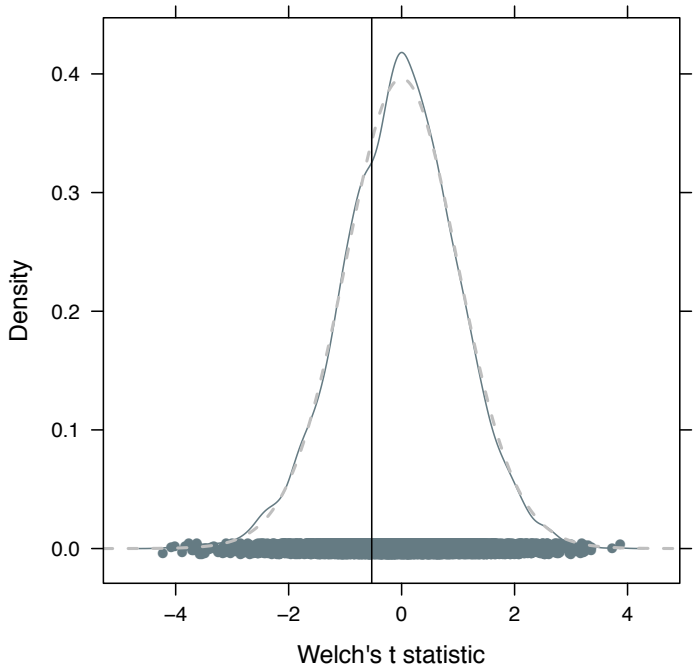


... and many many more *in silico* repeats of this experiment ...

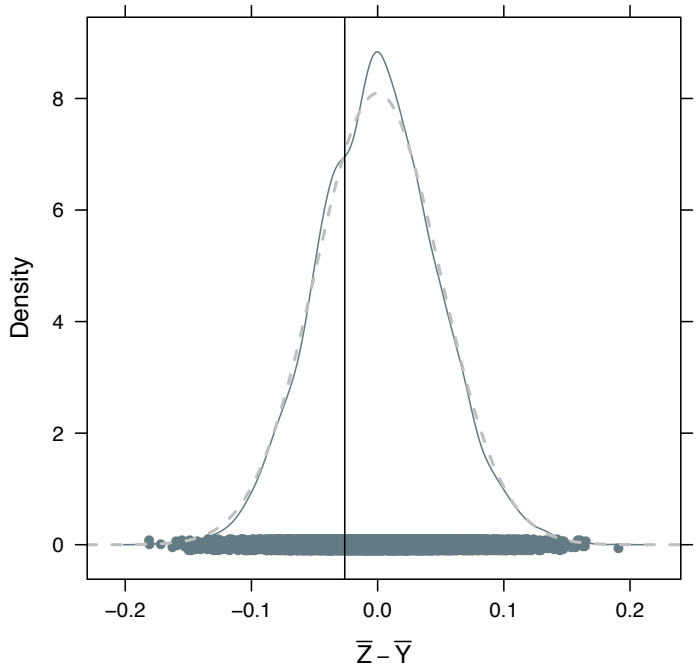
Here's the observed difference in sample means, the Welch's t statistic, and the associated p-value from the first 6 in silico datasets:

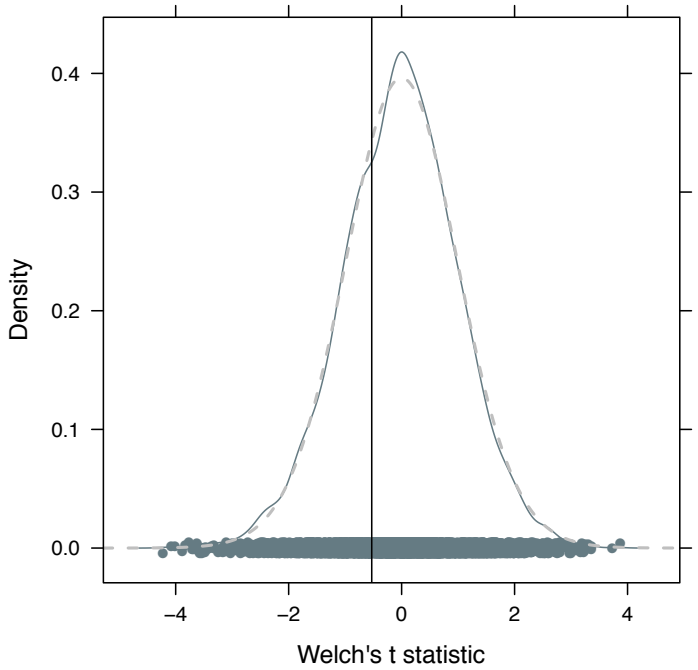
	smDiff	tStat	pVal
1	-0.049219079	-1.1866161	0.2449818
2	-0.012561184	-0.2422272	0.8099760
3	-0.063784868	-1.2212680	0.2298243
4	0.028180921	0.7100104	0.4827649
5	0.008151974	0.1881476	0.8525778
6	0.018928289	0.4349598	0.6661791





Empirical distribution of 10,000 observations, under the null of equal means, of the difference in sample means (left) and Welch's two sample t statistic (right). Overlaid w/ normal / t theoretical distributions (dashed line). Sample mean difference and t statistic from the real Irs4 data showed w/ vertical line.



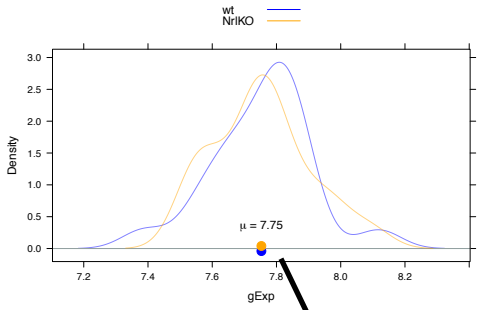


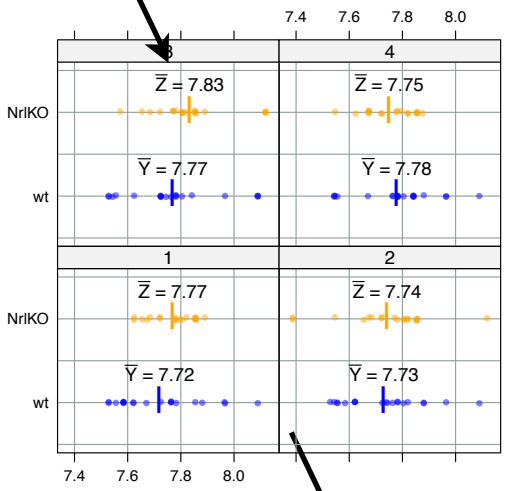
Let's sanity check the canned p-values. What proportion of these sample mean differences or Welch statistics are as or more extreme than what we observed?

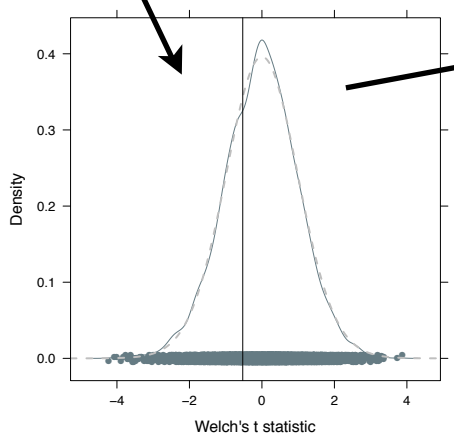
```
miniDat$gene: Irs4

     Welch Two Sample t-test

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

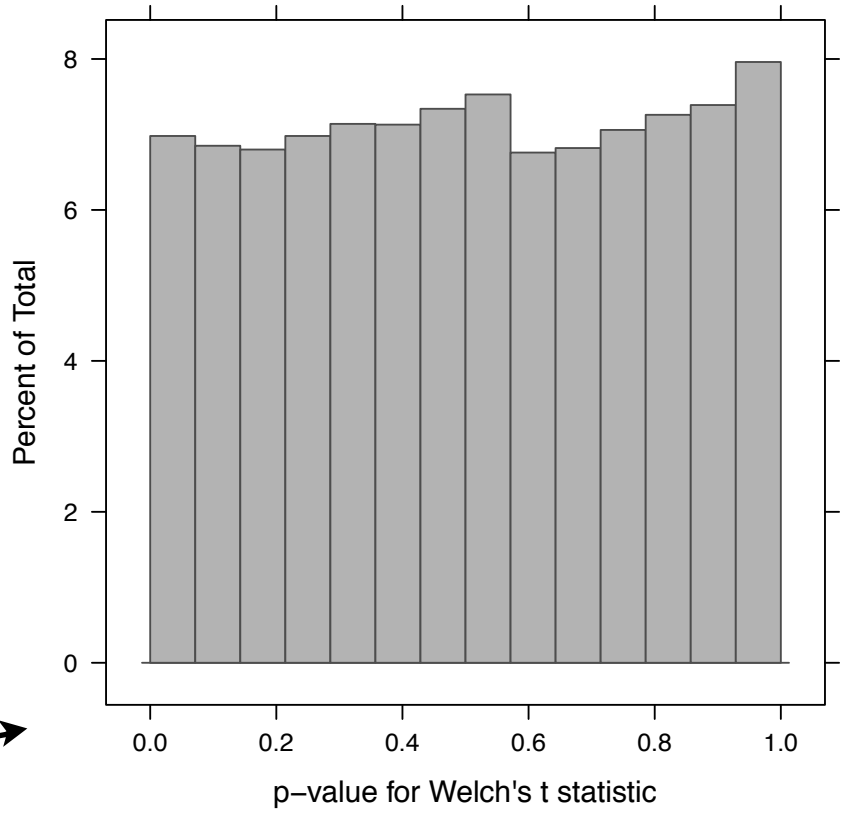
```
> mean(abs(bootTestStats$tStat) >= abs(welchStat))
[1] 0.5942

> mean(abs(bootTestStats$smDiff) >= abs(theDiff))
[1] 0.5818
```

Pretty bang on!



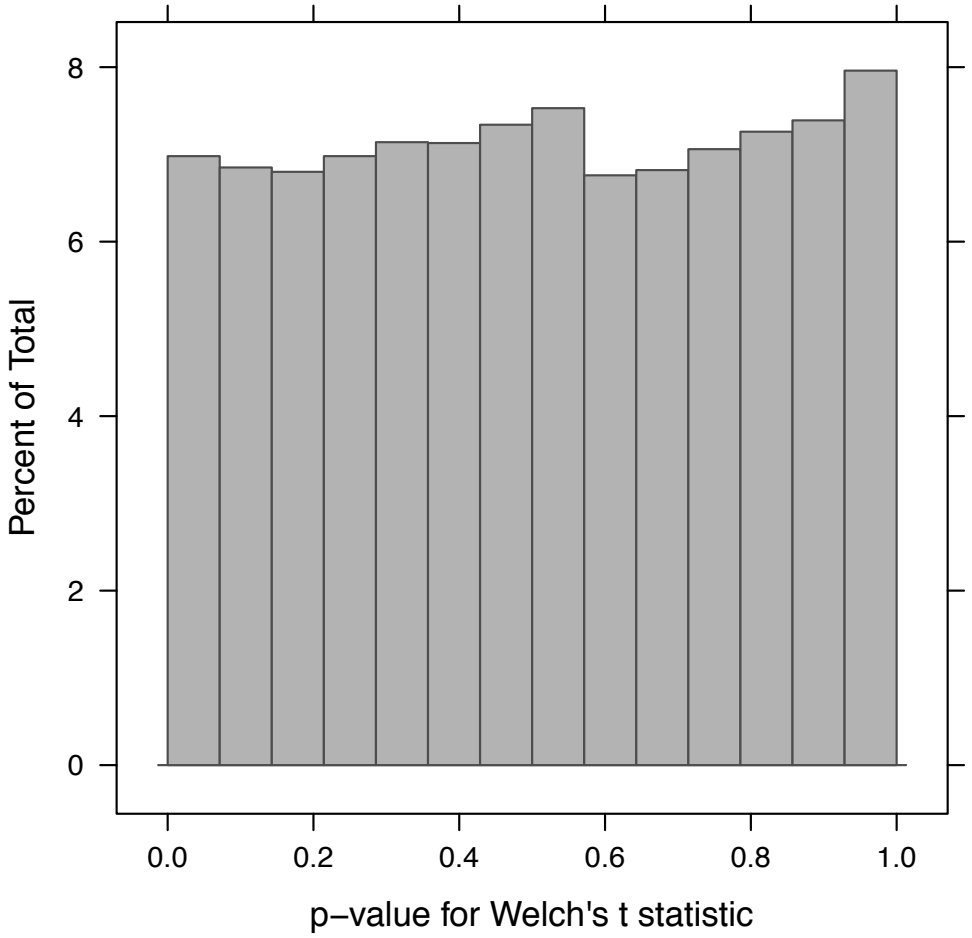






What does the distribution of the p-values look like when the null hypothesis holds?

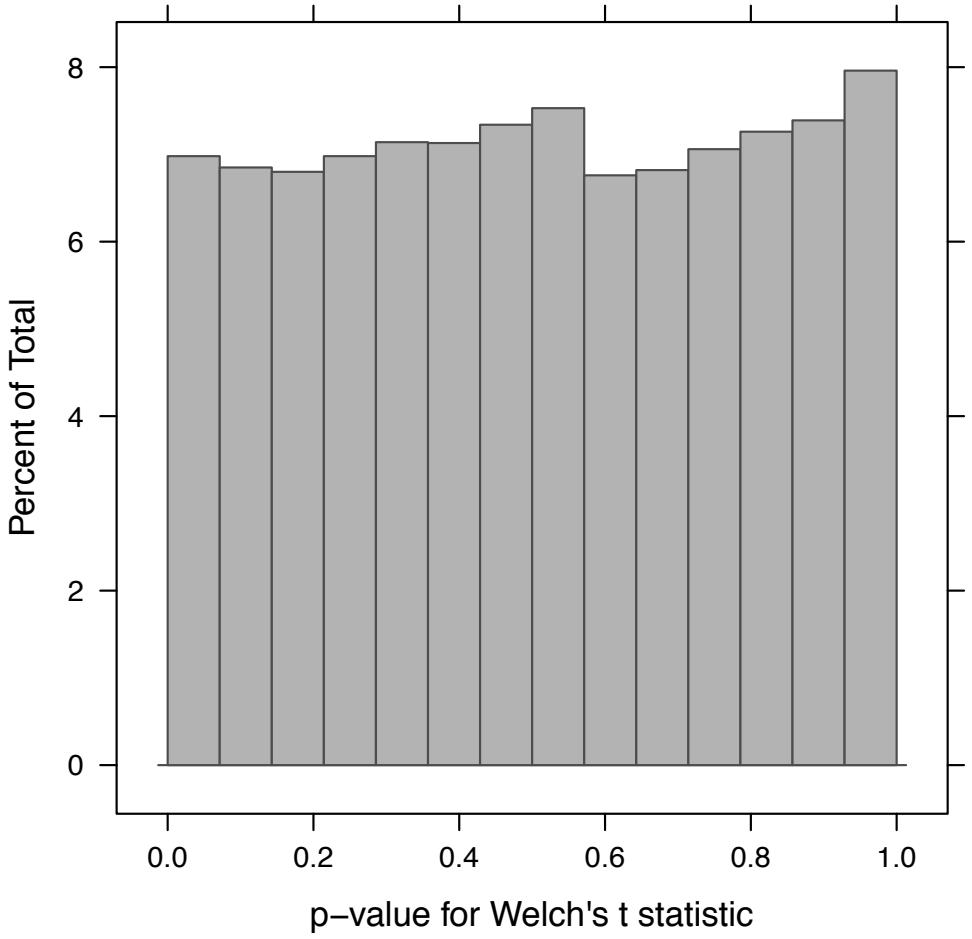
What does the distribution of the p-values look like when the null hypothesis holds?



What does the distribution of the p-values look like when the null hypothesis holds?

It is, by definition, Unif[0,1].

Internalize that fact, because it gets utilized when thinking about doing lots of tests and correcting for that.



$$V(\bar{X}_n - \bar{Y}_n) = \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}$$

True variance of difference of sample means depends on the underlying variance of the data and the sample sizes.

$$\hat{V}(\bar{X}_n - \bar{Y}_n) = \text{"pooled"}\ \hat{\sigma}^2 \left[\frac{1}{n_X} + \frac{1}{n_Y} \right] \quad \text{assuming } \sigma_X^2 = \sigma_Y^2$$

$$\hat{V}(\bar{X}_n - \bar{Y}_n) = \hat{\sigma}^2_{\bar{X}_n - \bar{Y}_n} = \frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y} \quad \text{assuming } \sigma_X^2 \neq \sigma_Y^2$$

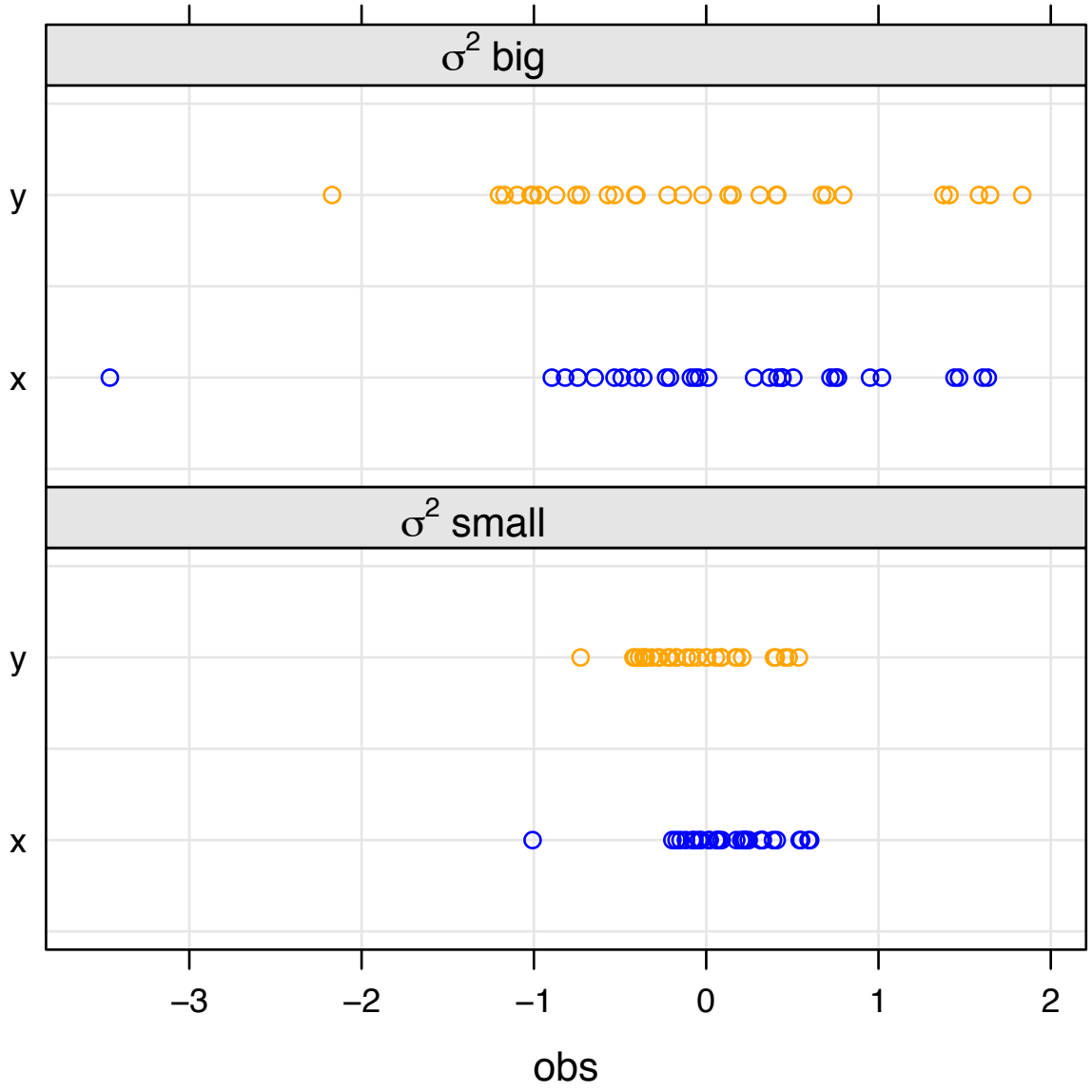
Sample variance is used to estimate it.

$$T = \frac{\bar{X}_n - \bar{Y}_n}{\hat{\sigma}_{\bar{X}_n - \bar{Y}_n}} \qquad \text{Under } H_0,\ T \sim t_{n_X + n_Y - 2} \text{ or } t_{\text{<sthg ugly>}}$$

What if the underlying variance could be reduced dramatically?

Less variance means same apparent effect is much more statistically significant.

```
> with(lDat,
+      by(lDat, sigStat, function(yo) {
+        t.test(obs ~ rv, yo)
+      }))

sigStat: big

     Welch Two Sample t-test

data:  obs by rv
t = 0.7314, df = 58, p-value = 0.4675
<snip, snip>
sample estimates:
mean in group x mean in group y
      0.1269433      -0.0618942

----------------------------------------
sigStat: small

     Welch Two Sample t-test

data:  obs by rv
t = 2.3128, df = 58, p-value = 0.02430
<snip, snip>
sample estimates:
mean in group x mean in group y
      0.1269433      -0.0618942
```



note: using simulated data not seen yet today

$$V(\bar{X}_n - \bar{Y}_n) = \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}$$

True variance of difference of sample means depends on the underlying variance of the data and the sample sizes.

$$\hat{V}(\bar{X}_n - \bar{Y}_n) = \text{"pooled"}\ \hat{\sigma}^2 \left[\frac{1}{n_X} + \frac{1}{n_Y}\right] \quad \text{assuming } \sigma_X^2 = \sigma_Y^2$$

$$\hat{V}(\bar{X}_n - \bar{Y}_n) = \hat{\sigma}^2_{\bar{X}_n - \bar{Y}_n} = \frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y} \quad \text{assuming } \sigma_X^2 \neq \sigma_Y^2$$

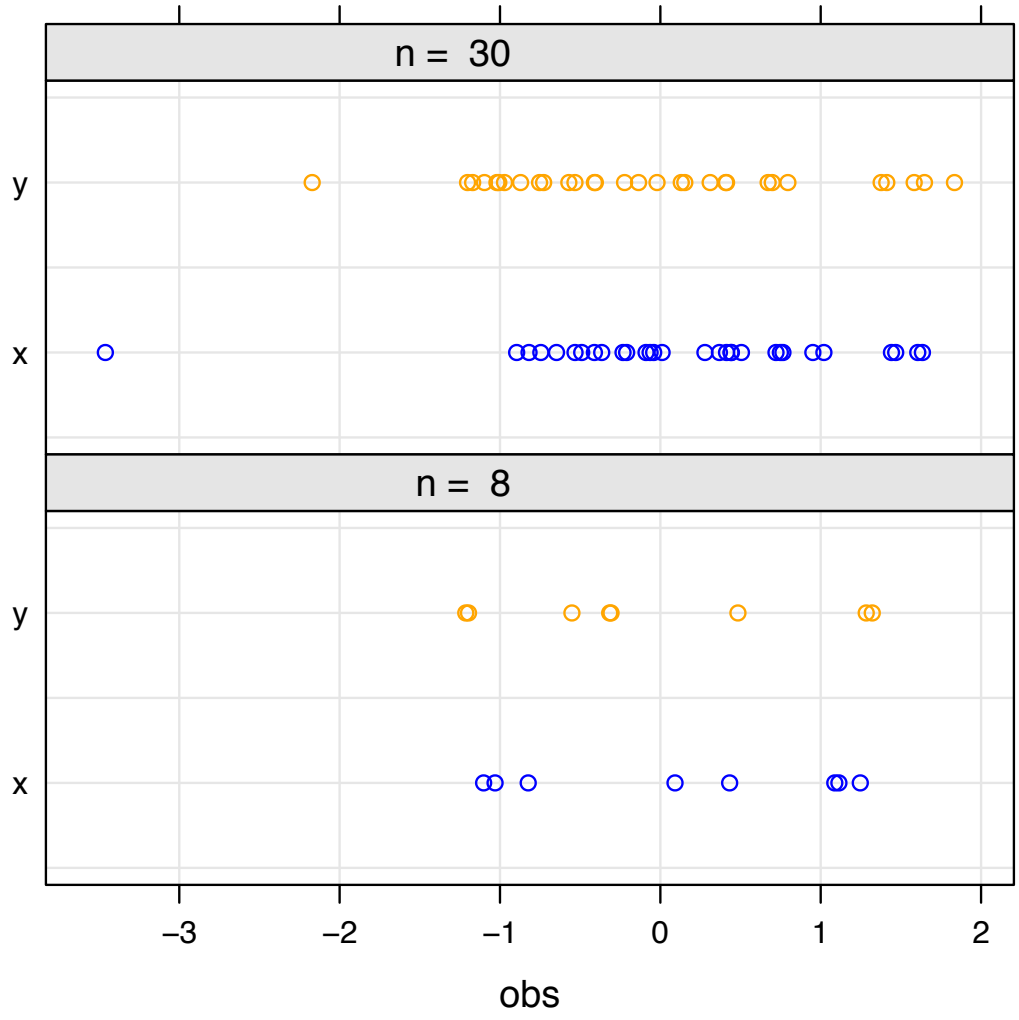
Note that sample sizes appear in denominators throughout.

$$T = \frac{\bar{X}_n - \bar{Y}_n}{\hat{\sigma}_{\bar{X}_n - \bar{Y}_n}} \qquad \text{Under } H_0,\ T \sim t_{n_X + n_Y - 2} \text{ or } t_{\text{<sthg ugly>}}$$

What if the sample size gets cut way down?

Smaller sample means same apparent effect is much less statistically significant.*

```
> with(mDat,
+      by(mDat, n, function(yo) {
+        t.test(obs ~ rv, yo)
+      }))
n: big

     Welch Two Sample t-test

data:  obs by rv
t = 0.7314, df = 58, p-value = 0.4675
<snip, snip>
sample estimates:
mean in group x mean in group y
      0.1269433      -0.0618942

------------------------------------
n: small

     Welch Two Sample t-test

data:  obs by rv
t = 0.3777, df = 14, p-value = 0.7113
<snip, snip>
sample estimates:
mean in group x mean in group y
      0.1269433      -0.0618942
```



* I also held the sample variance constant here.

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. Plus, we basically demonstrated that today.

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

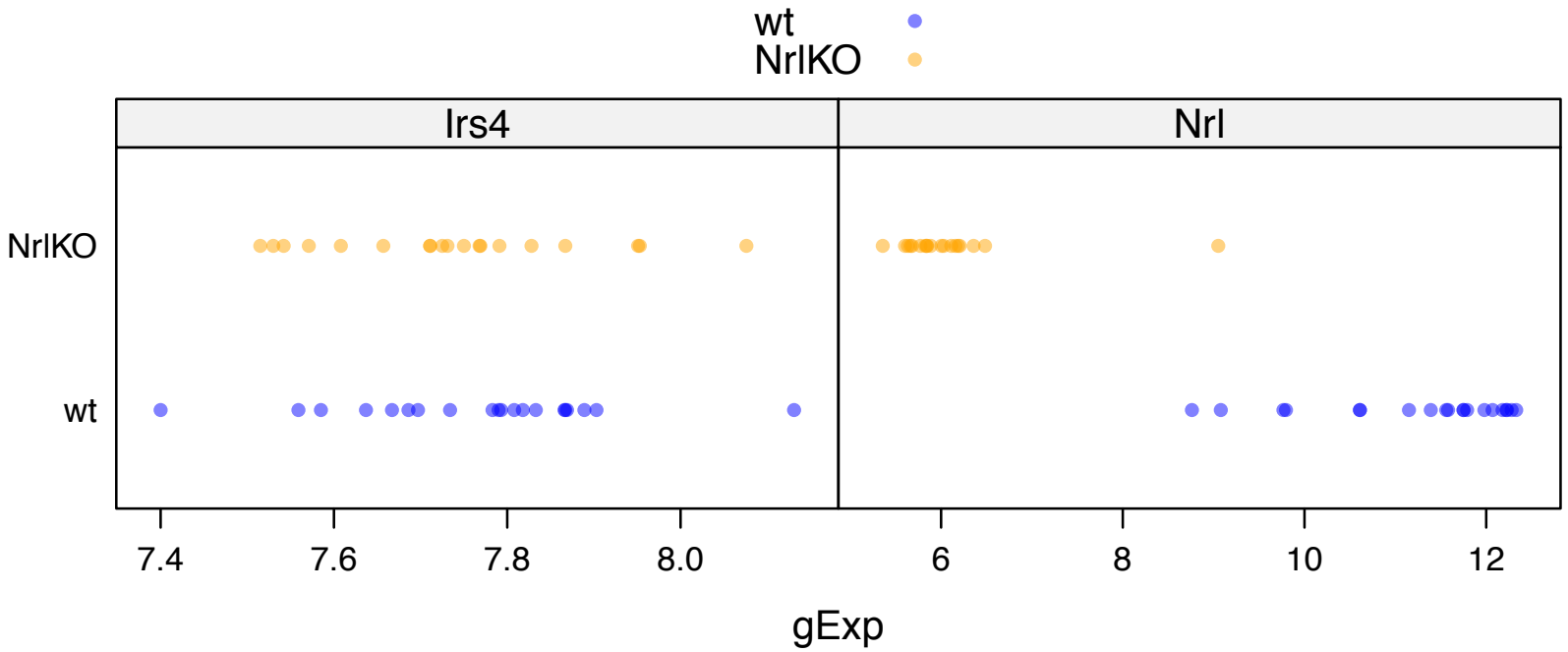
Test stat = sum of the ranks for one group (optionally, subtract the minimum possible which is nY (nY + 1)/2)

(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

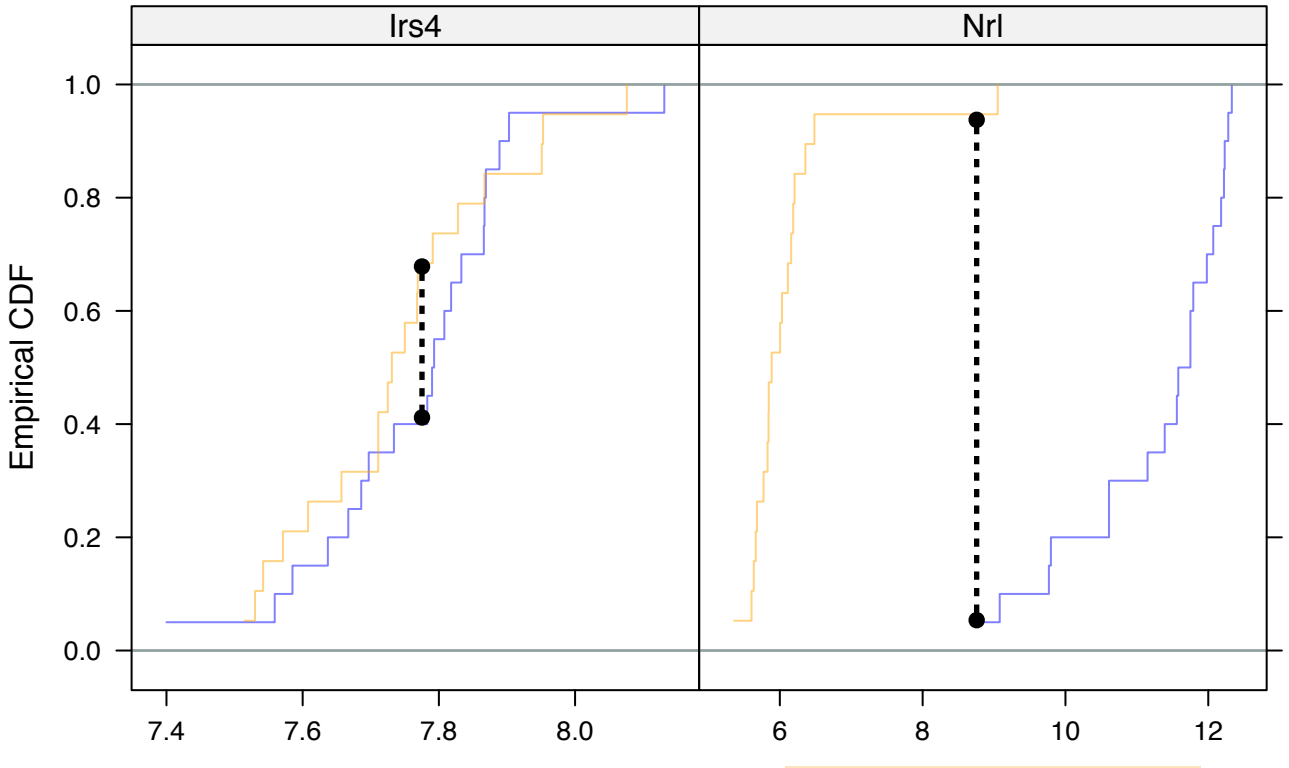
Estimate each CDF with the empirical CDF (ECDF)

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

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: 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

    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: Nrl

    Welch Two Sample t-test

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

<snip, snip>
```