Two group comparisons

Keegan Korthauer

24 January 2022

with slide contributions from Jenny Bryan and Sara Mostafavi

Reminders

- Intro assignment due today
- Project groups posted to Discussion board last week
- Initial project proposals due this Friday (Jan 28)
- Reminder that (free online) resources for review of statistical concepts are listed in the syllabus

Central dogma of statistics

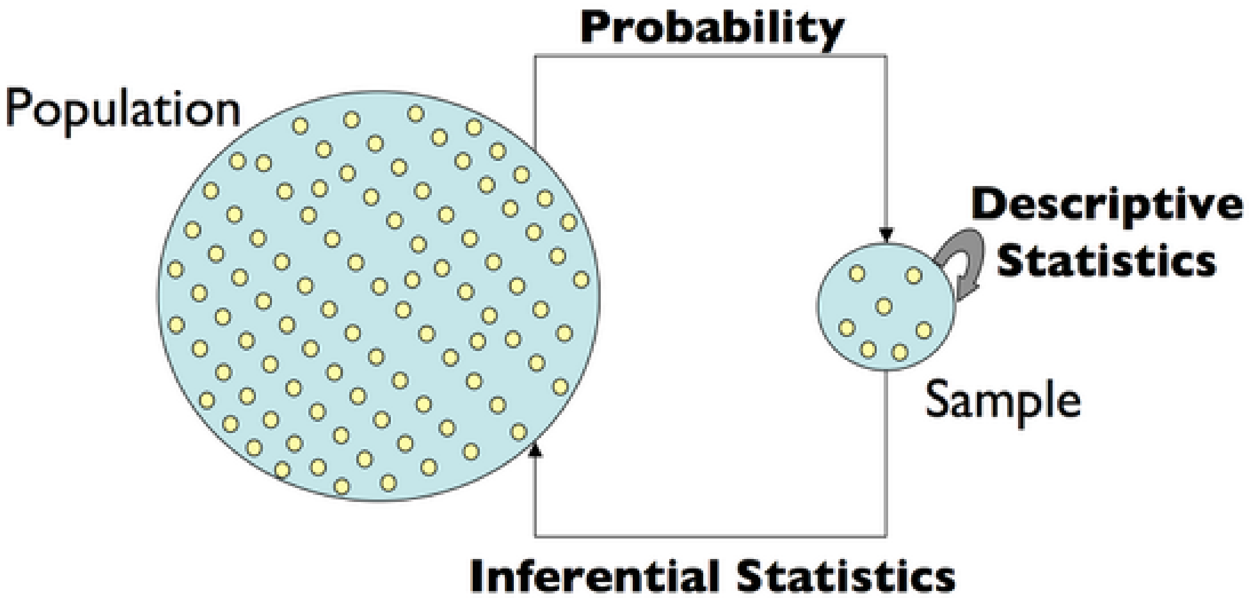
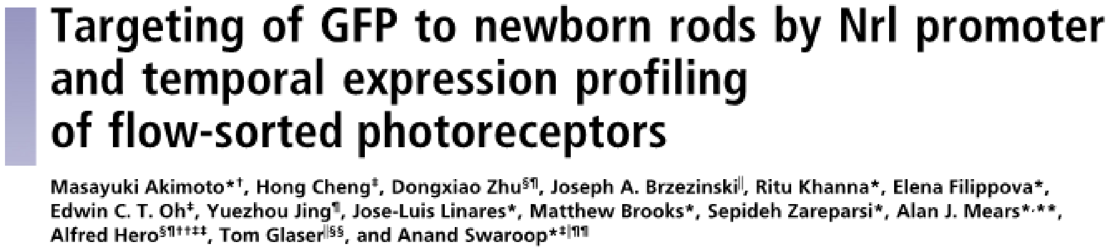


Image source: Josh Akey's Lecture notes

We want to understand a population (e.g. all individuals with a certain disease) but we can only study a random sample from it

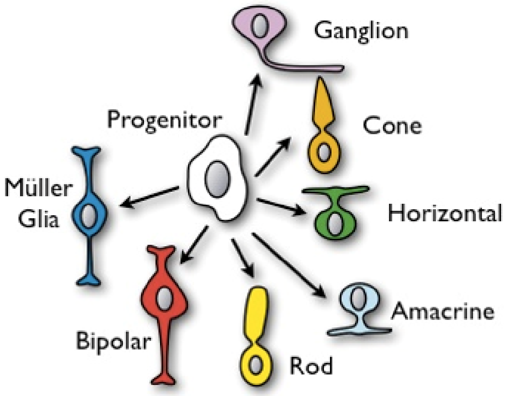
Hypothesis Testing in Genomics


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[*,‡,‖,¶¶]

[Akimoto et al. (2006)](#)

- Retina presents a model system for investigating regulatory networks underlying neuronal differentiation
- Nrl transcription factor is known to be important for Rod development



Hypothesis Testing in Genomics

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

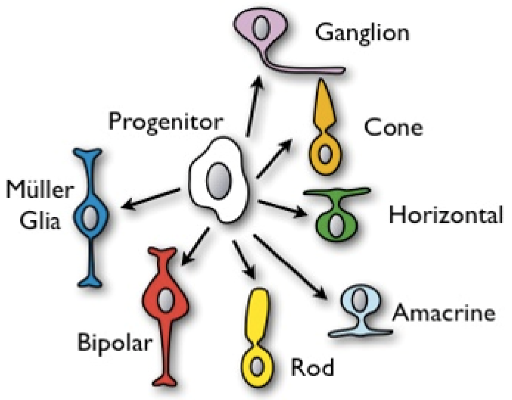
Masayuki Akimoto[*,†], Hong Cheng[‡], Dongxiao Zhu[§,¶], Joseph A. Brzezinski[||], Ritu Khanna[*], Elena Filippova[*], Edwin C. T. Oh[‡], Yuezhou Jing[¶], Jose-Luis Linares[*], Matthew Brooks[*], Sepideh Zareparsi[*], Alan J. Mears[*,**], Alfred Hero[§,¶,††,‡‡], Tom Glaser[||,§§], and Anand Swaroop[*,‡,||,¶¶]

[Akimoto et al. (2006)](#)

- 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?

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!

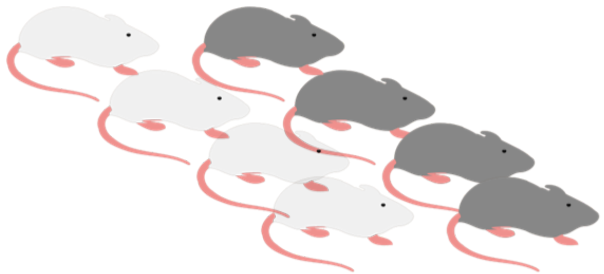
Statistical inference

Statistical inference:

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

Experimental design:

- 5 developmental stages (E16, P2, P6, P10, 4Weeks)
- 2 genotypes: Wild type (WT), Nrl Knockout (NrlKO)
- 3-4 replicates for each combination



Reading in / exploring the data

- Data obtained from the Gene Expression Omnibus (GEO) repository
    - Load directly into R session using `GEOquery` package
- Practice with this in Seminars 4 and 5 (Seminar 5 uses the exact same data set!)
- Review lecture 3 (exploratory data analysis) for general principles

Two example genes: Irs4 and Nrl

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

Two example genes: Irs4 and Nrl

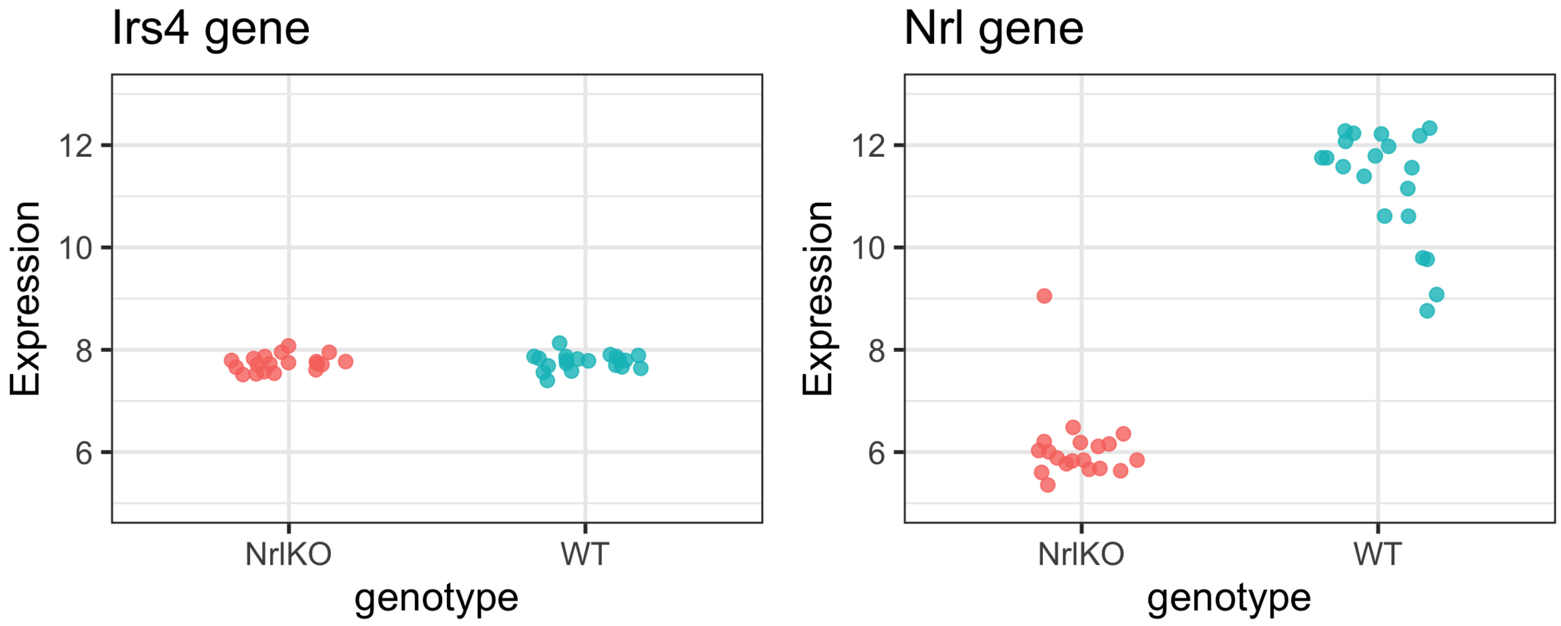
Biological question: Are these genes truly different in NrlKO compared to WT?

We can't answer this question in general; we can *only* study these genes in collected data (gene expression values from a random sample of mice):

```
twoGenes
```

```
## # A tibble: 78 × 5
##    gene  sample_id Expression dev_stage genotype
##    <chr> <chr>          <dbl> <fct>     <fct>
##  1 Irs4  GSM92610        7.71 4_weeks   NrlKO
##  2 Irs4  GSM92611        7.77 4_weeks   NrlKO
##  3 Irs4  GSM92612        7.73 4_weeks   NrlKO
##  4 Irs4  GSM92613        7.57 4_weeks   NrlKO
##  5 Irs4  GSM92614        7.95 E16       NrlKO
##  6 Irs4  GSM92615        7.52 E16       NrlKO
##  7 Irs4  GSM92616        8.08 E16       NrlKO
##  8 Irs4  GSM92617        7.71 P10       NrlKO
##  9 Irs4  GSM92618        7.87 P10       NrlKO
```

Visualizing Irs4 and Nrl genes in our sample



Statistical Hypothesis

Experimental design: (ignoring developmental time for now)

- 2 conditions: WT *vs* NrlKO
- we observe the expression of many genes in a random sample of mice from each condition

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

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

- 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

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

$Z_1, Z_2, \ldots, Z_{n_Z}$: a random sample of size n_Z NrlKO mice

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

$\bar{Z} = \frac{\sum_{i=1}^{n_Z} Z_i}{n_Z}$: sample mean of gene g expression from NrlKO mice

Notation

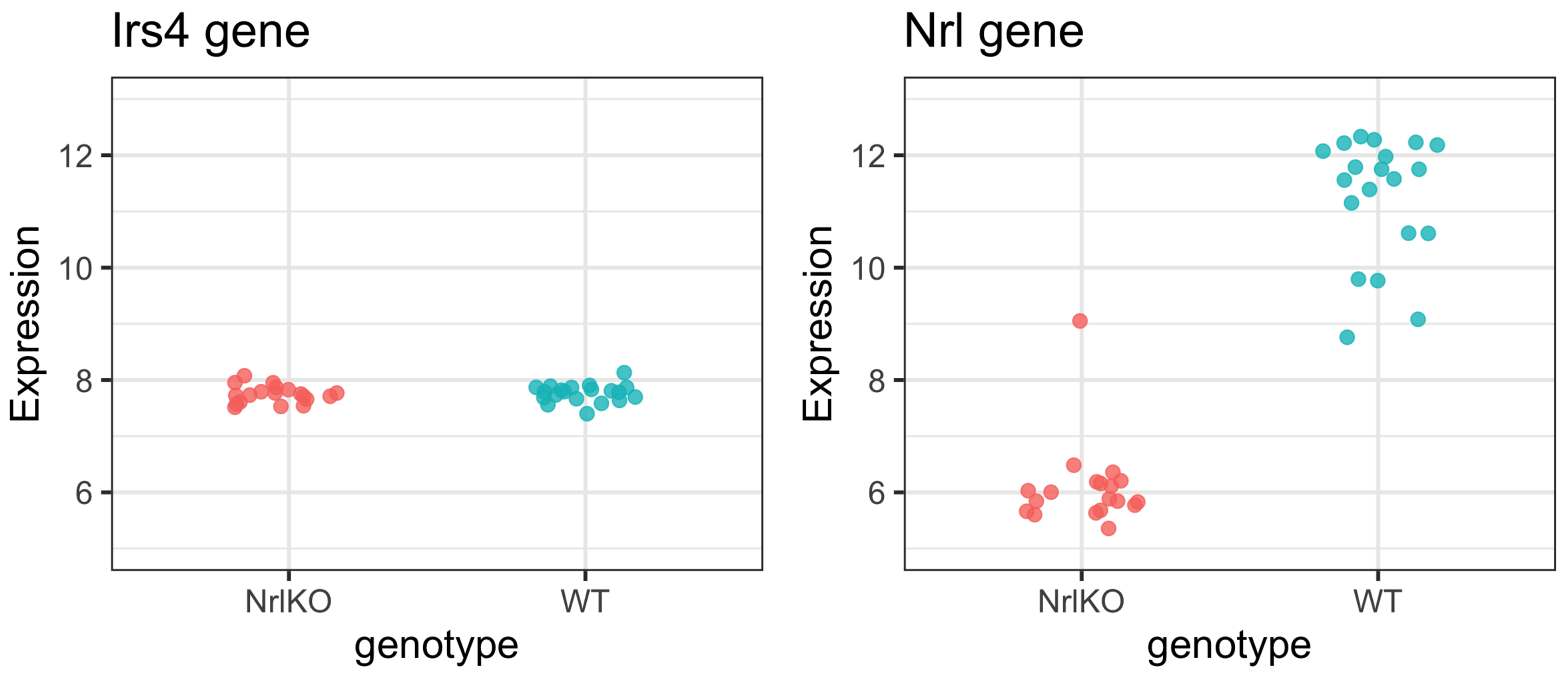
Population parameters (unknown/unobservable):

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

$\mu_Z = E[Z]$: the (population) expected expression of gene g in NrlKO mice

Is there enough evidence to reject H_0?

$$H_0 : \mu_Y = \mu_Z$$



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

What we observe: sample averages: $\bar{Y}$ vs $\bar{Z}$

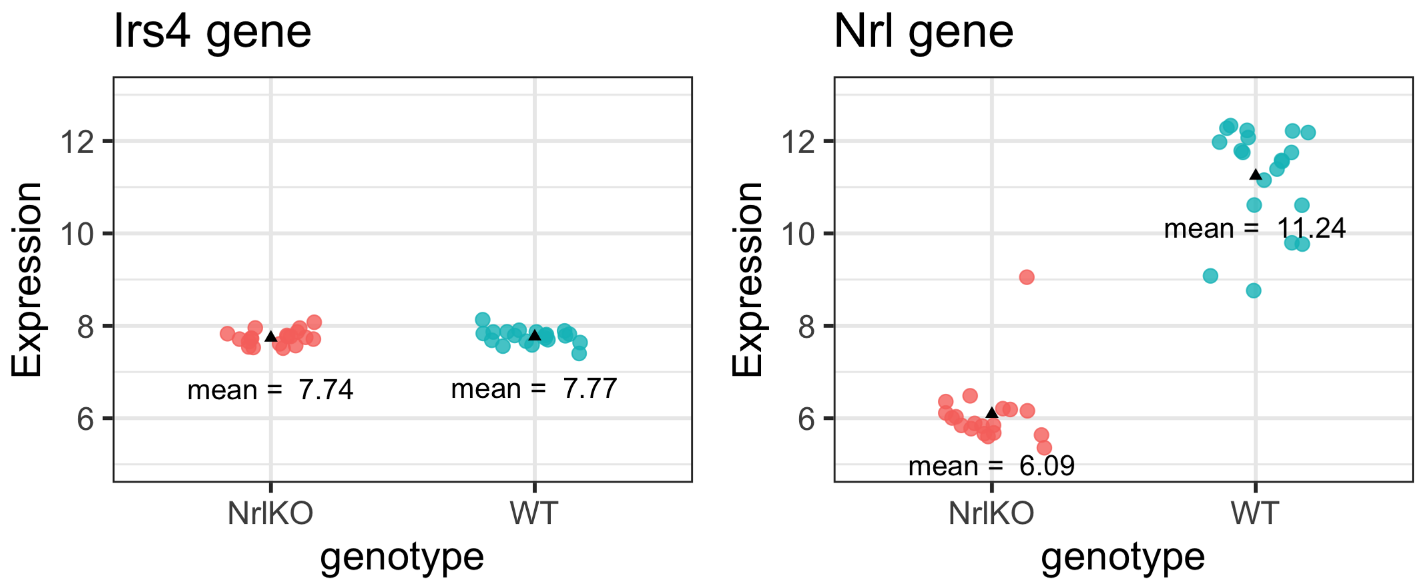
```
meanExp <- twoGenes %>%
  group_by(gene, genotype) %>%
  summarize(meanExpr = mean(Expression)) %>%
  spread(genotype, meanExpr) %>%
  mutate(diffExp = NrlKO - WT)
meanExp
```

```
## # A tibble: 2 × 4
## # Groups:   gene [2]
##   gene  NrlKO    WT diffExp
##   <chr> <dbl> <dbl>   <dbl>
## 1 Irs4   7.74  7.77 -0.0261
## 2 Nrl    6.09 11.2  -5.15
```

This code uses tidy data wrangling functions to calculate:

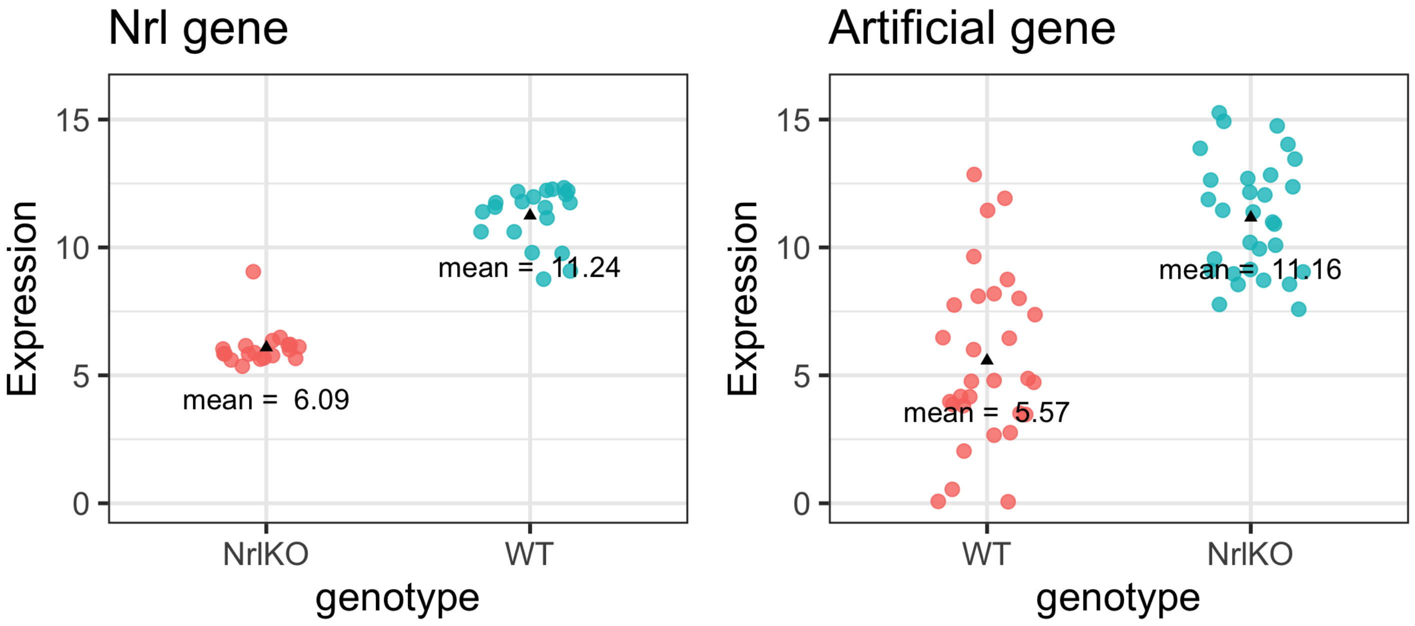
- the mean expression of each gene per genotype group
- the difference in mean expression of each gene in Nrl KO vs WT groups

Is the difference between $\bar{Y}$ and $\bar{Z}$ enough 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?

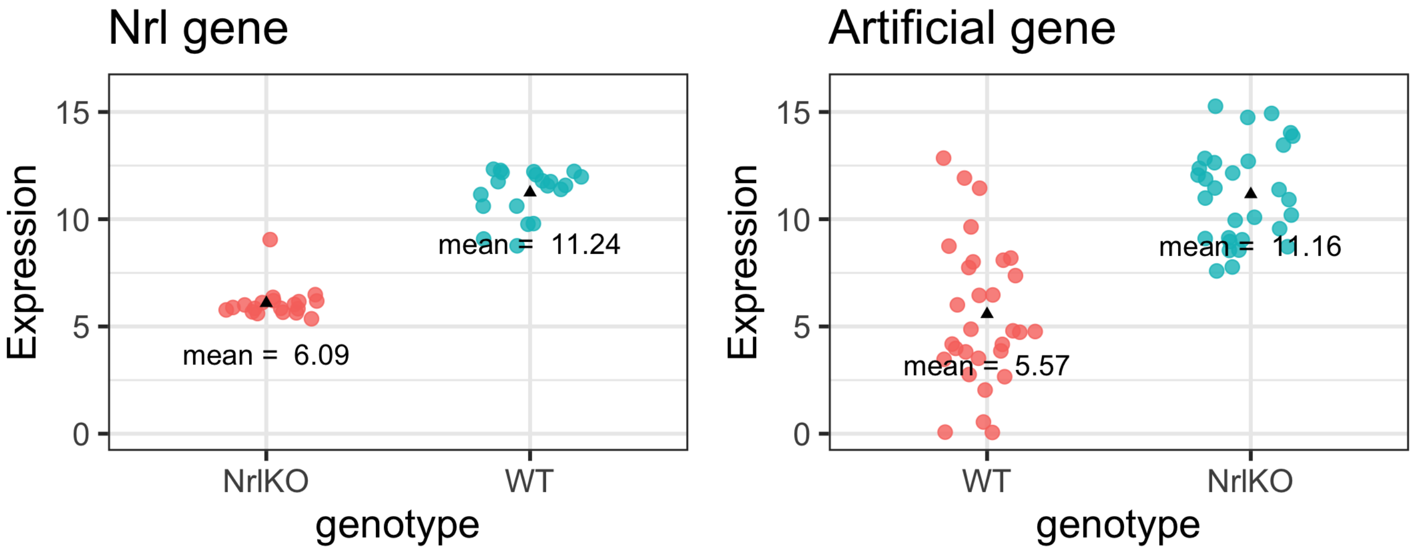


$$\frac{\bar{Y} - \bar{Z}}{??}$$

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

"Large" relative to the observed variation:

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



Quantifying observed variation

- Recall that if $Var(Y_i) = \sigma_Y^2$, then $Var(\bar{Y}) = \frac{\sigma_Y^2}{n_Y}$
- Assume that the random variables within each group are *independent and identically distributed* (iid), and that the groups are independent. More specifically, that

  1. $Y_1, Y_2, \ldots, Y_{n_Y}$ are iid,
  2. $Z_1, Z_2, \ldots, Z_{n_Z}$ are iid, and
  3. Y_i, Z_j are independent. Then, it follows that

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

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

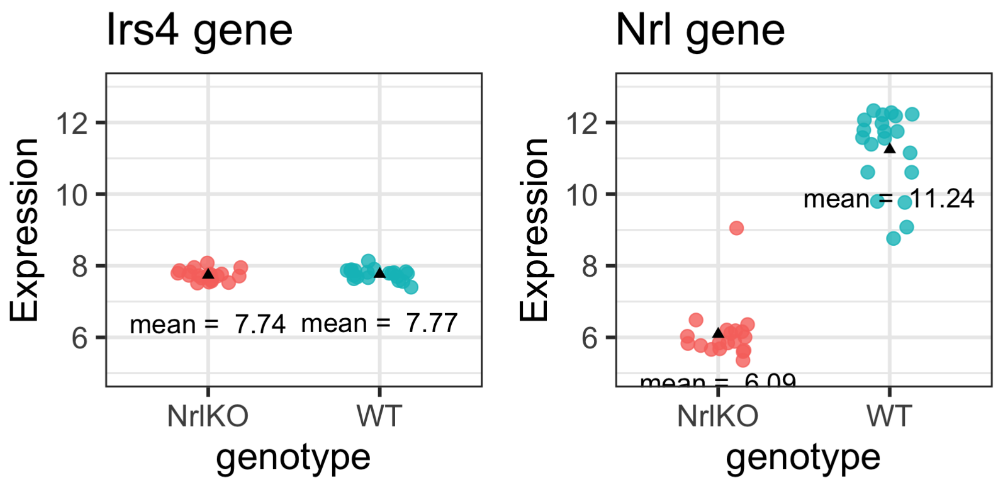
$$Var(\bar{Z} - \bar{Y}) = \frac{\sigma_Z^2}{n_Z} + \frac{\sigma_Y^2}{n_Y} = \sigma^2 \left[\frac{1}{n_Z} + \frac{1}{n_Y}\right]$$

But how can we calculate population variance σ if it is unknown?

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

```
twoGenes %>%
  group_by(gene, genotype) %>%
  summarize(groupVar = var(Expression))
```

```
## # A tibble: 4 × 3
## # Groups:   gene [2]
##   gene  genotype groupVar
##   <chr> <fct>       <dbl>
## 1 Irs4  NrlKO      0.0233
## 2 Irs4  WT         0.0240
## 3 Nrl   NrlKO      0.594
## 4 Nrl   WT         1.22
```



For example, for Nrl in WT:

$$\hat{\sigma}_Y^2 = S_Y^2 = \frac{1}{n_Y} \sum_{i=1}^{n_Y} (Y_i - \bar{Y})^2 = 1.22$$

Combining sample variances

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

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

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

- Assuming unequal variance of Y's and Z's (Welch's t-test)

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

> Recall: the 'hat' (^) is used to distinguish an 'estimate' from a 'parameter'

Test Statistic

$$T = \frac{\bar{Z}_n - \bar{Y}_n}{\sqrt{\hat{Var}(\bar{Z}_n - \bar{Y}_n)}}$$

```
meanExp %>%
    mutate(t = diffExp / sqrt(s2Diff)) %>%
    mutate(tWelch = diffExp / sqrt(s2DiffWelch))
```

```
## # A tibble: 2 × 8
## # Groups:   gene [2]
##   gene  NrlKO    WT diffExp s2DiffWelch  s2Diff  tWelch       t
##   <chr> <dbl> <dbl>   <dbl>       <dbl>   <dbl>   <dbl>   <dbl>
## 1 Irs4   7.74  7.77 -0.0261     0.00243 0.00243  -0.529  -0.529
## 2 Nrl    6.09 11.2  -5.15       0.0925  0.0942  -17.0   -16.8
```

Can we now say whether the observed differences are 'big'?

Test Statistic

$$T = \frac{\bar{Z}_n - \bar{Y}_n}{\sqrt{\hat{Var}(\bar{Z}_n - \bar{Y}_n)}}$$

```
meanExp %>%
    mutate(t = diffExp / sqrt(s2Diff)) %>%
    mutate(tWelch = diffExp / sqrt(s2DiffWelch))
```

```
## # A tibble: 2 × 8
## # Groups:   gene [2]
##   gene  NrlKO    WT diffExp s2DiffWelch  s2Diff  tWelch       t
##   <chr> <dbl> <dbl>   <dbl>       <dbl>   <dbl>   <dbl>   <dbl>
## 1 Irs4   7.74  7.77 -0.0261     0.00243 0.00243  -0.529  -0.529
## 2 Nrl    6.09 11.2  -5.15       0.0925  0.0942  -17.0   -16.8
```

Can we now say whether the observed differences are 'big'?

The difference is about half a standard deviation for Irs4 and ~17 standard deviations for Nrl

What to do with this statistic?

- 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?

What to do with this statistic?

- 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 a probability distribution!
- However, this is unknown to us so we need to make more assumptions

Null distribution assumptions

- If we know how our statistic behaves when the *null hypothesis is true*, then we can evaluate how extreme our observed data is
  - The null distribution is the probability distribution of T under H_0
- Let's assume that Y_i and Z_i follow (unknown) probability distributions called F and G:

$$(Y_i \sim F, \text{ and } Z_i \sim G)$$

- Depending on the assumptions we make about F and G, theory tells us specific null distributions for our test statistic

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's 2-sample *t*-test:

(unequal variances)

$$T \sim t_{<something\,ugly>}$$

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's 2-sample *t*-test:

(unequal variances)

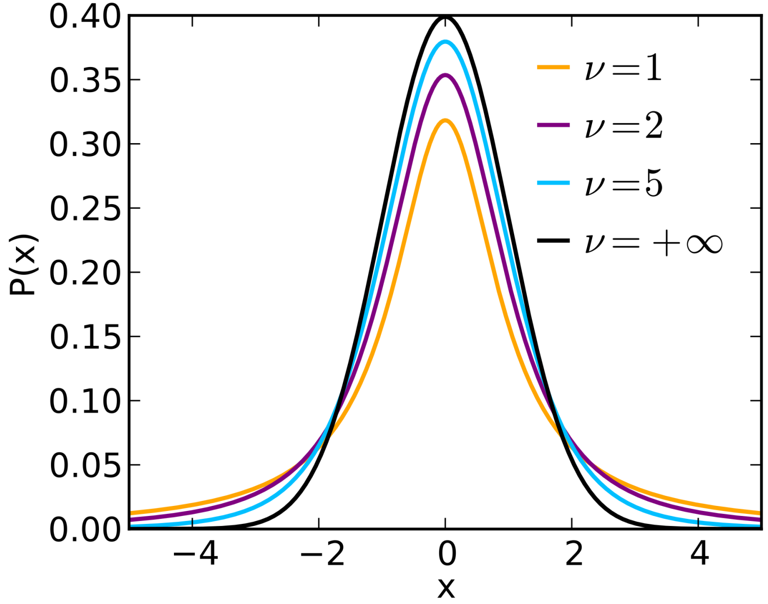
$$T \sim t_{<something\ ugly>}$$

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

Summary: $T = \frac{\bar{Z}_n - \bar{Y}_n}{\sqrt{\hat{Var}(\bar{Z}_n - \bar{Y}_n)}}$ is a random variable, and under certain assumptions, we can prove that T follows a *t*-distribution



Recall that the *t*-distribution has one parameter: df = degrees of freedom

Hypothesis testing: Step 1

1. Formulate your hypothesis as a statistical hypothesis

In our example:

$$H_0 : \mu_Y = \mu_Z \text{ vs } H_A : \mu_Y \neq \mu_Z$$

Hypothesis testing: Step 2

2a. Define a test statistic

In our example: 2-sample *t*-test

2b. Compute the observed value for the test statistic

For our two example genes:

```
twoGenes %>%
  group_by(gene) %>%
  summarize(t = t.test(Expression ~ genotype,
                            var.equal=TRUE)$statistic)
```

```
## # A tibble: 2 × 2
##   gene        t
##   <chr>   <dbl>
## 1 Irs4   -0.529
## 2 Nrl   -16.8
```

Hypothesis testing: Step 3

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)

For our two example genes:

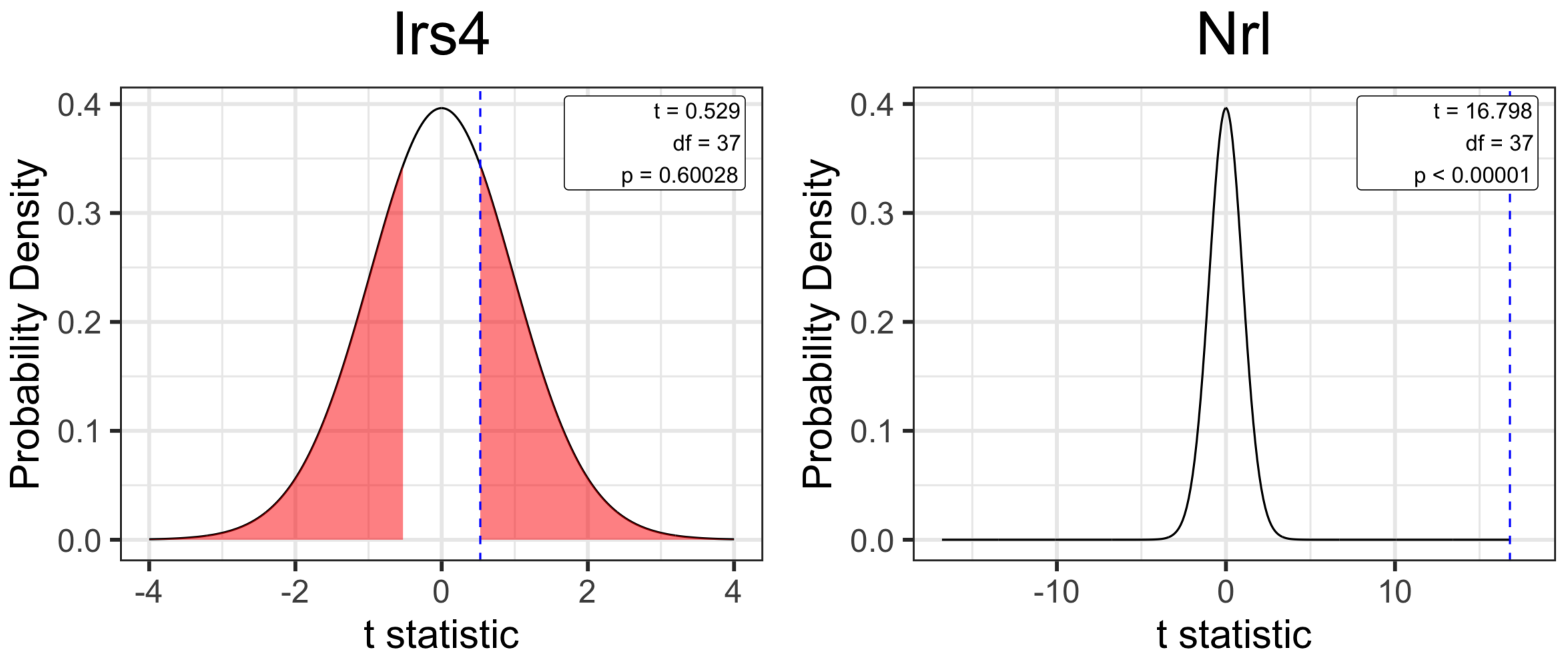
```
twoGenes %>%
  group_by(gene) %>%
  summarize(pvalue = t.test(Expression ~ genotype,
                            var.equal=TRUE)$p.value)
```

```
## # A tibble: 2 × 2
##   gene     pvalue
##   <chr>     <dbl>
## 1 Irs4  6.00e- 1
## 2 Nrl   6.73e-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 Nrl, the probability of seeing a test statistic as extreme as that observed $(t = -16.8)$ is extremely low $(p = 6.76 \times 10^{-19})$



Hypothesis Testing: Step 4

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

The significance level α is often set at 0.05. However, this value is arbitrary and may depend on the study.

Irs4

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 differential expression in WT compared to NrlKO models.

We do not reject H_0!

Nrl

Using $\alpha = 0.05$, since the p-value for the Nrl test is much less than 0.05, we conclude that there is significant evidence in the data to claim that Nrl has differential expression in WT compared to NrlKO models.

We reject H_0!

`t.test` function in R

Assuming equal variances

```
twoGenes %>% filter(gene == "Nrl") %>%
  t.test(Expression ~ genotype,
         var.equal=TRUE, data = .)
```

```
##
##      Two Sample t-test
##
## data:  Expression by genotype
## t = -16.798, df = 37, p-value < 2.2e-16
## alternative hypothesis: true difference in means
between group NrlKO and group WT is not equal to 0
## 95 percent confidence interval:
##  -5.776672 -4.533071
## sample estimates:
## mean in group NrlKO    mean in group WT
##            6.089579           11.244451
```

Assuming equal variances

```
twoGenes %>% filter(gene == "Nrl") %>%
  t.test(Expression ~ genotype,
         var.equal=FALSE, data = .)
```

```
##
##      Welch Two Sample t-test
##
## data:  Expression by genotype
## t = -16.951, df = 34.01, p-value < 2.2e-16
## alternative hypothesis: true difference in means
between group NrlKO and group WT is not equal to 0
## 95 percent confidence interval:
##  -5.772864 -4.536879
## sample estimates:
## mean in group NrlKO    mean in group WT
##            6.089579           11.244451
```

Check out `?t.test` for more options, including how to specify one-sided tests

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 probability* 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

"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 provides good diagnostics/insights
- Is it mostly uniform (flat)? If not, is the departure from uniform expected based on biological knowledge?
- We will revisit these topics in greater detail in later lectures

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

Decision	Actual Situation "Truth": H_0 True	Actual Situation "Truth": H_0 False
Do Not Reject H_0	Correct Decision $1-\alpha$	Incorrect Decision Type II Error β
Reject H_0	Incorrect Decision Type I Error α	Correct Decision $1-\beta$

$$\alpha = P(\text{Type I Error}),\ \beta = P(\text{Type II Error}),\ \text{Power} = 1-\beta$$

H_0: "*Innocent until proven guilty*"

- The default state is $H_0 \rightarrow$ we only reject if we have enough evidence
- If H_0: Innocent and H_A: Guilty, then
  - Type I Error (α): Wrongfully convict innocent (*False Positive*)
  - Type II Error (β): Fail to convict criminal (*False Negative*)

What are alternatives to the *t*-test?

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?

- First, one could use the t test statistic but use a bootstrap approach to compute its p-value; we'll revisit this topic later
- Alternatively, there are *non-parametric* tests that are available:
  - 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 $y_i \geq z_i$)

The null distribution of such statistics can be worked out or approximated

wilcox.test function in R

```
wilcox.test(Expression ~ genotype,
            data = twoGenes  %>% filter(gene == "Irs4"))
```

```
##
##      Wilcoxon rank sum exact test
##
## data:  Expression by genotype
## W = 160, p-value = 0.4115
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(Expression ~ genotype,
            data = twoGenes  %>% filter(gene == "Nrl"))
```

```
##
##      Wilcoxon rank sum exact test
##
## data:  Expression by genotype
## W = 1, p-value = 5.804e-11
## alternative hypothesis: true location shift is not equal to 0
```

Kolmogorov-Smirnov test (two sample)

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

Estimate each CDF with the empirical CDF (ECDF)

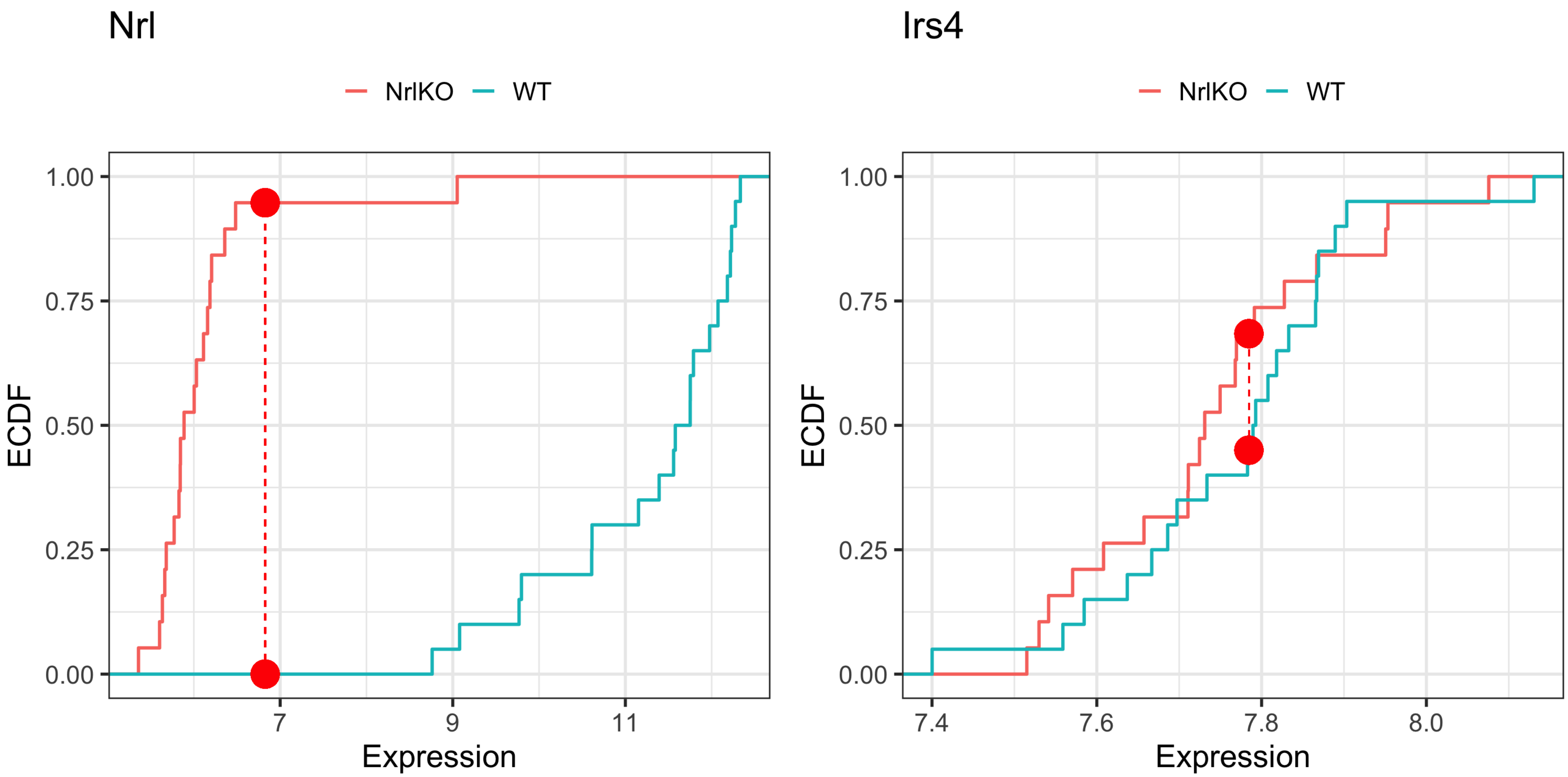
$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} I[x_i \leq x]$$

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 detail here)

Kolmogorov-Smirnov test (two sample)



ks.test function in R

```
Nrlgene <- twoGenes %>% filter(gene == "Nrl")
ks.test(Nrlgene$Expression[Nrlgene$genotype == "WT"],
        Nrlgene$Expression[Nrlgene$genotype == "NrlKO"])
```

```
##
##      Two-sample Kolmogorov-Smirnov test
##
## data:  Nrlgene$Expression[Nrlgene$genotype == "WT"] and Nrlgene$Expression[Nrlgene$genotype == "NrlKO"]
## D = 0.95, p-value = 5.804e-10
## alternative hypothesis: two-sided
```

```
Irs4gene <- twoGenes %>% filter(gene == "Irs4")
ks.test(Irs4gene$Expression[Irs4gene$genotype == "WT"],
        Irs4gene$Expression[Irs4gene$genotype == "NrlKO"])
```

```
##
##      Two-sample Kolmogorov-Smirnov test
##
## data:  Irs4gene$Expression[Irs4gene$genotype == "WT"] and Irs4gene$Expression[Irs4gene$genotype == "NrlKO"]
## D = 0.28421, p-value = 0.3278
## alternative hypothesis: two-sided
```

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