

Statistical Methods for High Dimensional Biology

STAT/BIOF/GSAT 540

Lecture 5 – Two group comparisons

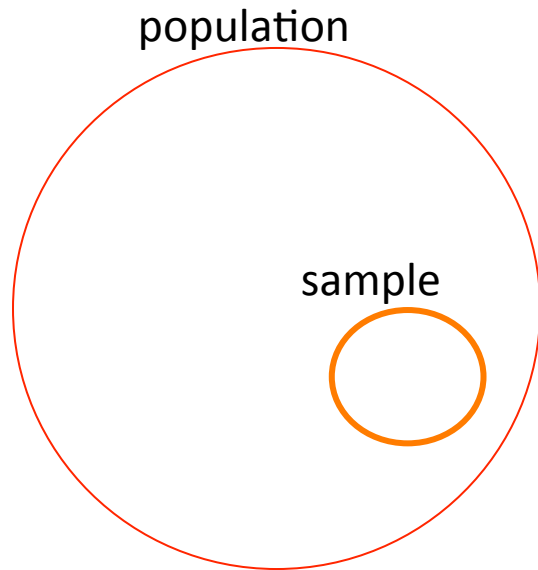
Sara Mostafavi
January 17 2017

Slide credits: Jenny Bryan, Su-In Lee

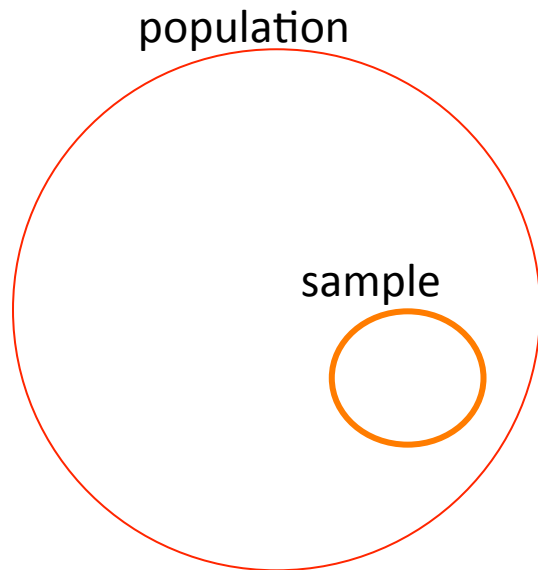
Announcements

- Practice Assignment due Jan 17
- Initial group formation and report due Jan 24
- GitHub issues

Review: population vs. sample



- We are interested in making statements about populations but we can only study a **random** sample from that population.
- We want to **infer** a general conclusion → **statistical inference**
- e.g., what's the average height of males?



- We are interested in making statements about populations but we can only study a random sample from that population.
- We want to *infer* a general conclusion → **statistical inference**

- We model the data with a particular (typically) parametric distribution $Y \sim F$.
- We'd like to estimate the parameters of the model.
- ** We don't know the population parameters ** ; we estimate the parameter from our data (our "estimates")

| family | typical notation | parameter θ |
|-------------|----------------------------|----------------------------|
| <generic> | $Y \sim F_{\theta}$ | θ |
| Bernoulli | $Y \sim \text{Bern}(p)$ | $\theta = p$ |
| binomial | $Y \sim \text{Bin}(n, p)$ | $\theta = (n, p)$ |
| uniform | $Y \sim \text{Unif}[a, b]$ | $\theta = (a, b)$ |
| Normal | $Y \sim N(\mu, \sigma^2)$ | $\theta = (\mu, \sigma^2)$ |
| Student's t | $Y \sim t_{df}$ | $\theta = df$ |

Estimation...

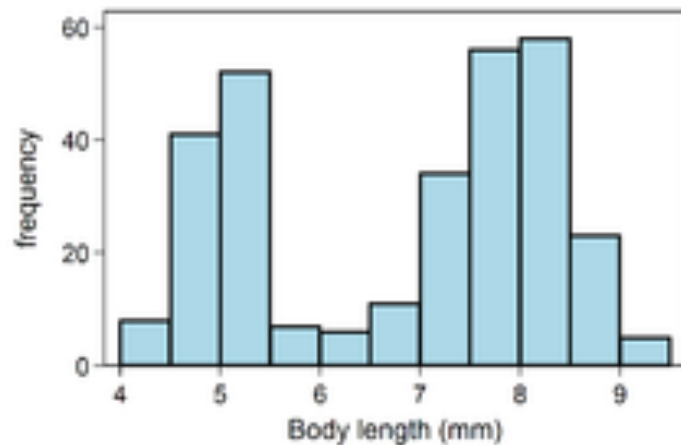
- Our estimate of the true mean is the sample mean

$$\bar{Y}_n = \frac{1}{n} \sum_{i=1}^n Y_i = \hat{\mu}$$

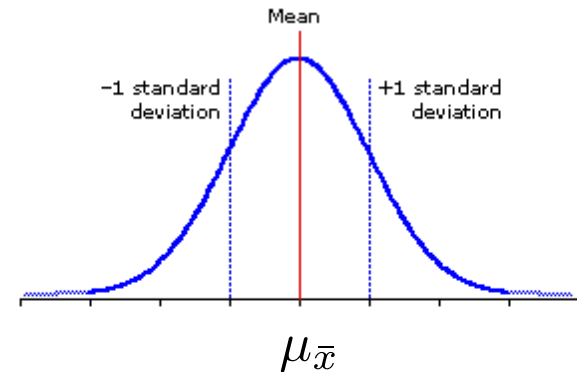
- *Sample mean is an RV and hence has an associated probability distribution function (“sampling distribution”)

Sampling distribution vs data distribution

Data generated by following distribution



Distribution of sample mean
– recall CLT



Variance of the sampling
distribution, depends on
sample size

More generally:

- An RV that is a function of the data (mean, variance) is called a **statistic**.
- Probability distribution (aka distribution) of a statistic is called its sampling distribution – not the same as the distribution of the data.
- Sampling distribution: distribution of the statistic if we ****were**** to repeatedly draw samples from the population and compute the value of the statistic each time.
- Important to think about: we only do our experiments once! From statistics, we have a framework to think about what would happen if we *did* take repeated samples!

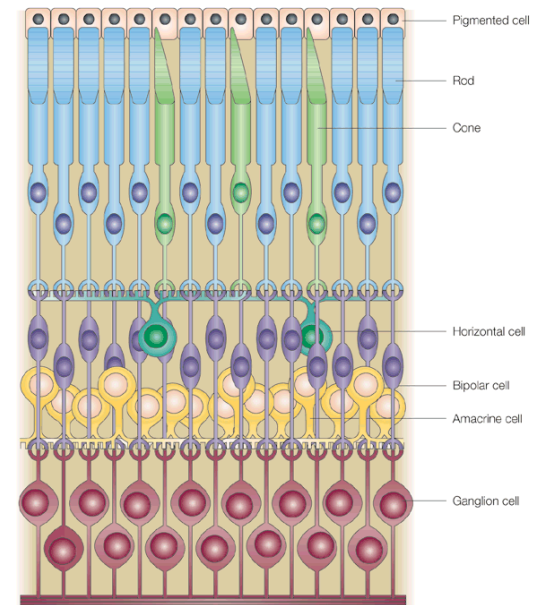
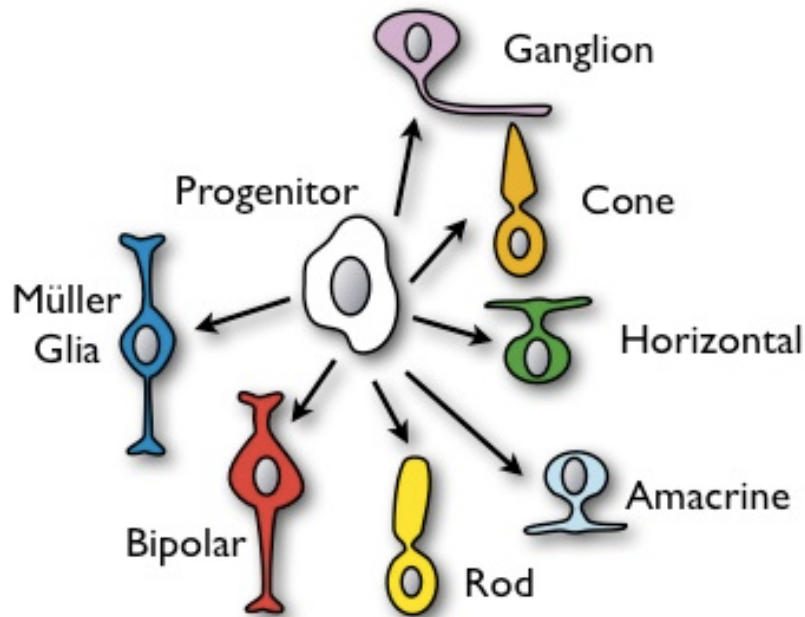
Hypothesis testing

- “A statistical test examines a set of sample data, and on the basis of the expected distribution of the data, leads to a decision about whether to accept hypothesis underlying the expected distribution or to reject the hypothesis and accepts an alternative one.”
- Motivating example: given the expression level of gene A in two different conditions, determine if gene A is differentially expressed.
- **Mutually exclusive** hypotheses:
 - H_0 : the expression level is the same.
 - H_A : the expression level is different.

We will analyze data from this study...

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 Zareparsa^{*}, Alan J. Mears^{*,**}, Alfred Hero^{§¶††††}, Tom Glaser^{||§§}, and Anand Swaroop^{*‡||¶¶}



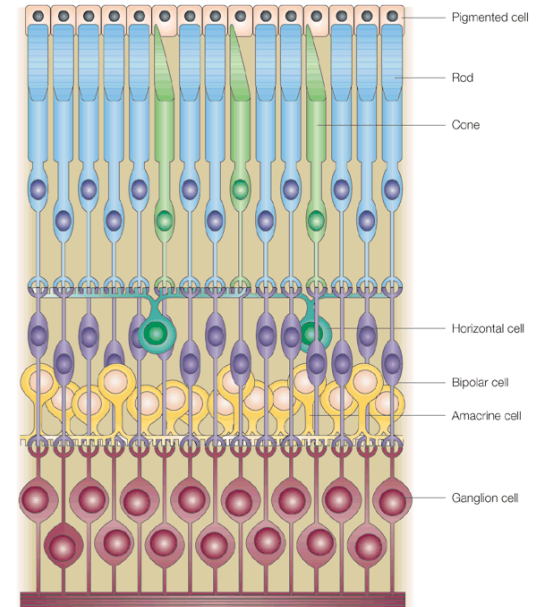
We will analyze data from this study...



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 Zareparsa^{*}, 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 (TF) known to be important for Rod development.
- **What happens if you delete Nrl?**



Developing mouse retina – time course for the experiment

So sample collections:

4 developmental stages

2 genotypes: wild-type , Nrl KO

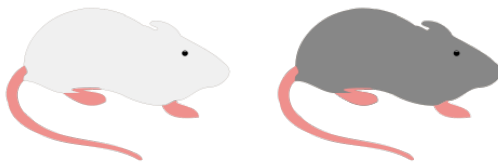
3-4 replicates for each combination

Experimental design

| devStage | wt | NrlKO |
|----------|----|-------|
| E16 | 4 | 3 |
| P2 | 4 | 4 |
| P6 | 4 | 4 |
| P10 | 4 | 4 |
| 4_weeks | 4 | 4 |

NrlKO

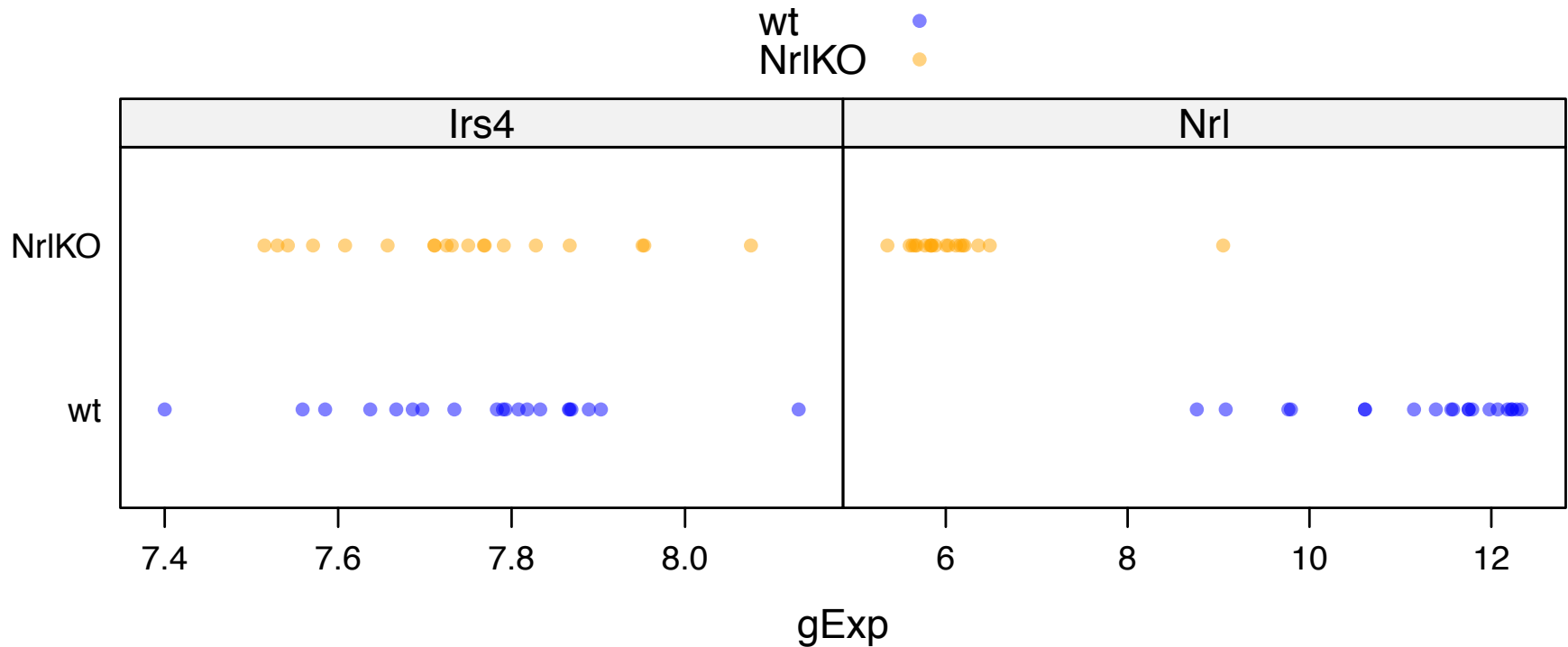
WT



What are the genes that are differentially expressed between WT and Nr1KO?

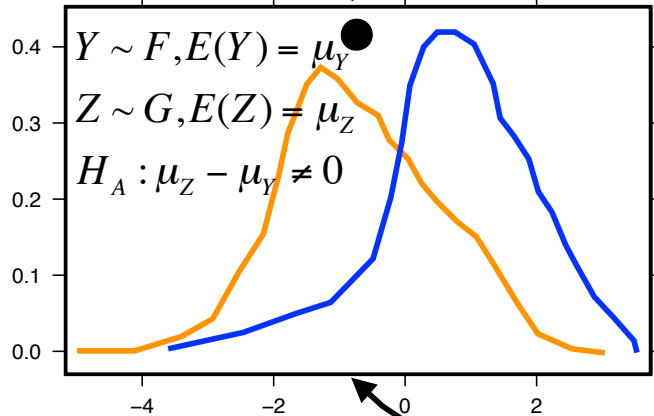
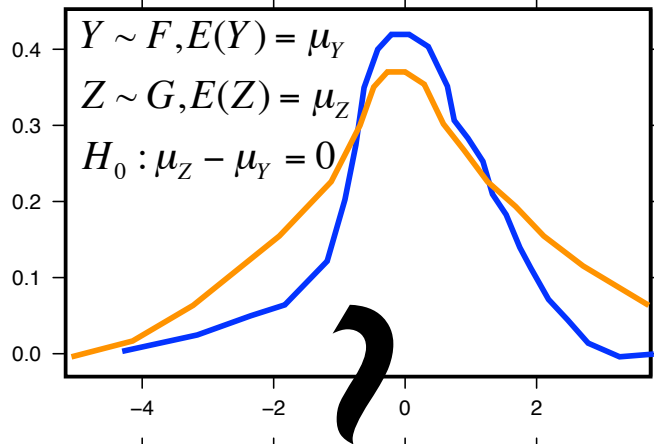
Let's do it for 2 genes ... we can then apply the same procedure to all genes, one at a time

We will use hypothesis testing (specifically t-test) to assess whether *Irs4* or *Nrl* genes are behaving different (i.e., differentially expressed) in WT and KO.

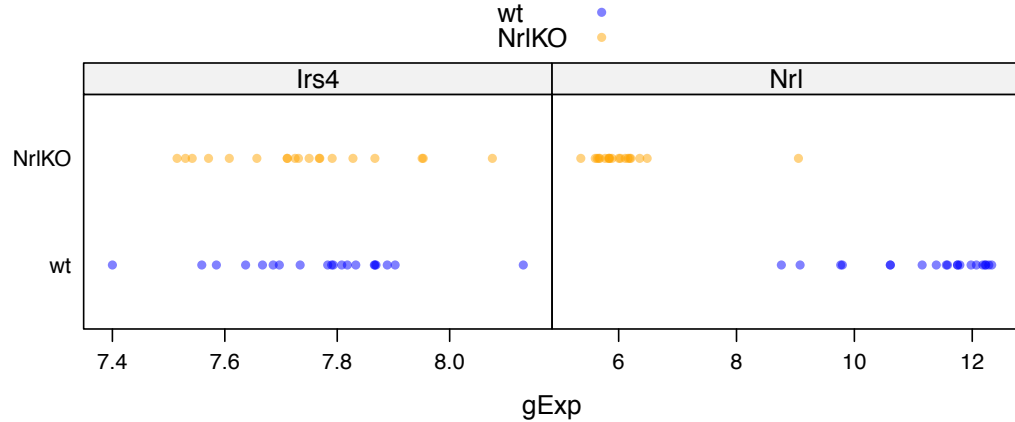


probability

data generating model



observed data

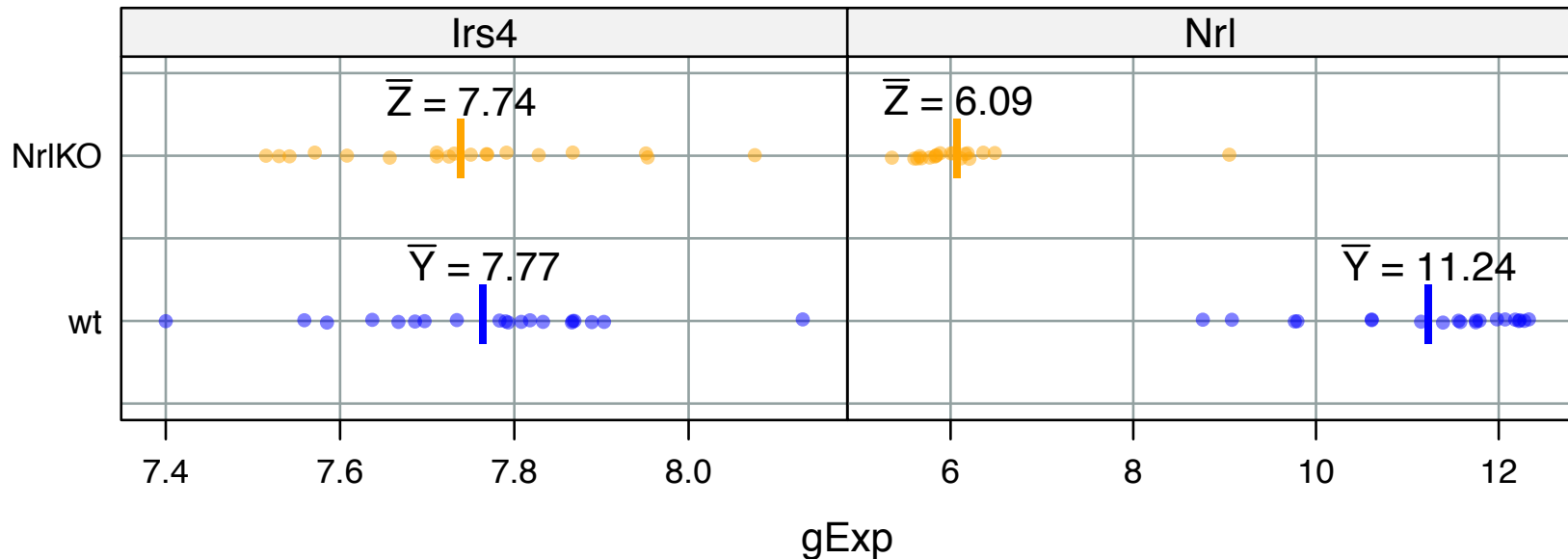


We will use a t-test to test for the differences in mean in two conditions.

statistical
inference

- **The null hypothesis, H_0 :**
 - There is no difference between the expression of Irs4/Nrl gene in WT and KO conditions.
 - $H_0: \mu_{wt} = \mu_{KO}$
- **The alternative hypothesis, H_A :**
 - There is a difference between the expression of Irs4/Nrl gene in WT and KO conditions.
 - $H_A: \mu_{wt} \neq \mu_{KO}$

- **The null hypothesis, H_0 :**
 - There is no difference between the expression of Irs4/Nrl gene in WT and KO conditions.
 - $H_0: \mu_{wt} = \mu_{KO}$
- **The alternative hypothesis, H_A :**
 - There is a difference between the expression of Irs4/Nrl gene in WT and KO conditions.
 - $H_A: \mu_{wt} \neq \mu_{KO}$



- The sample means by themselves are not enough to make robust conclusions.
- 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.

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?

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

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

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

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



variance of sample mean

[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?

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

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

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

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


 variance of sample mean

[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?

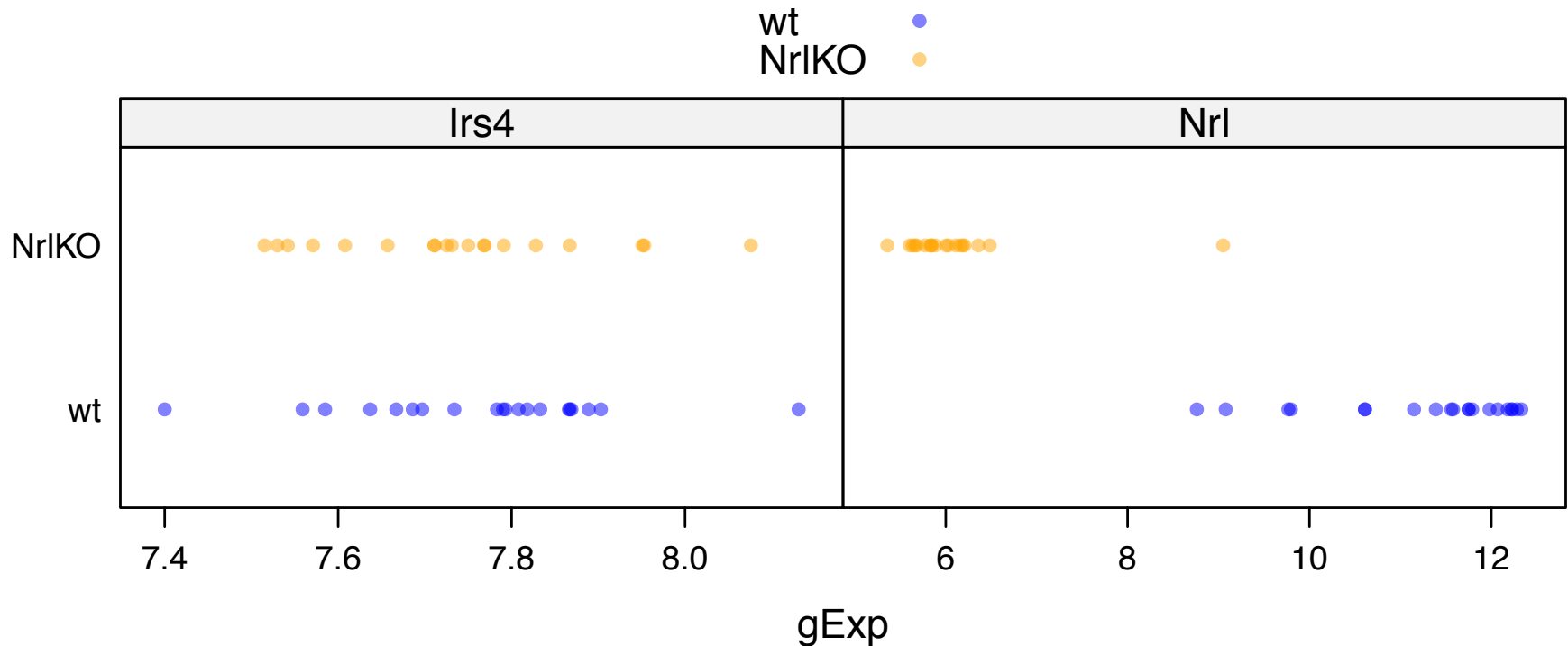
$$V(\bar{Z}_n - \bar{Y}_n) = \frac{\sigma_Z^2}{n_Z} + \frac{\sigma_Y^2}{n_Y}$$

if we assume that $\sigma_Z^2 = \sigma_Y^2 = \sigma^2$

$$\begin{aligned} V(\bar{Z}_n - \bar{Y}_n) &= \frac{\sigma^2}{n_Z} + \frac{\sigma^2}{n_Y} \\ &= \sigma^2 \left[\frac{1}{n_Z} + \frac{1}{n_Y} \right] \end{aligned}$$

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
Nr1KO 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

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


```
> (nY <- with(miniDat, sum(gType == "wt" & gene == "Nr1")))  
[1] 20  
> (nZ <- with(miniDat, sum(gType == "Nr1KO" & gene == "Nr1")))  
[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      Nr1  
0.02368783 0.91782091
```

```
> (s2Diff <- s2Pooled * (1/nY + 1/nZ))
```

```
      Irs4      Nr1  
0.00243112 0.09419741
```

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

```
> (s2DiffWelch <- colSums(theVars / c(nY, nZ)))
```

```
      Irs4      Nr1  
0.002429188 0.092494563
```

Now we can compute the observed difference in sample mean divided by our best guess at its 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.

We quantify how big/small t (or any test statistic) is using probability:

- Interpretation: if I were to repeat the experiment many times, assuming H_0 holds, what's the probability of observing a value of t as extreme as the one we observed.
- That probability will tell you how likely it is to observe a test statistic at least as extreme as the one we observed, given the assumptions we made are satisfied!
- Allows you to quantify the statistical significance of evidence.

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

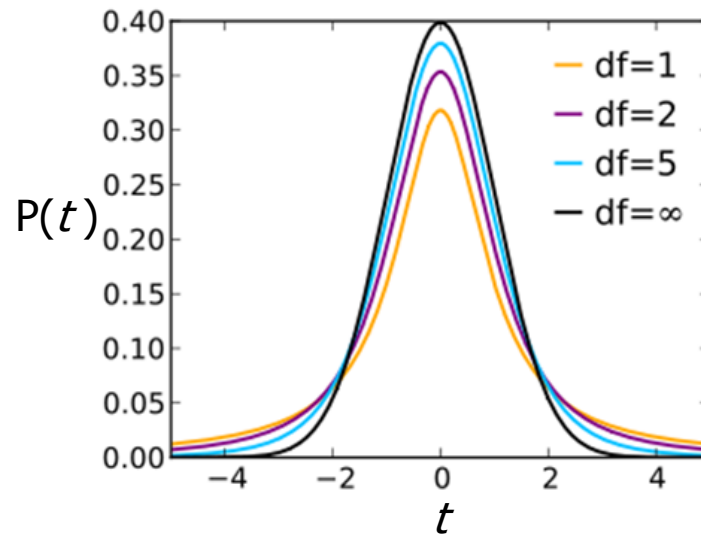
$$T \sim t_{\text{<sthg ugly>}}$$

“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

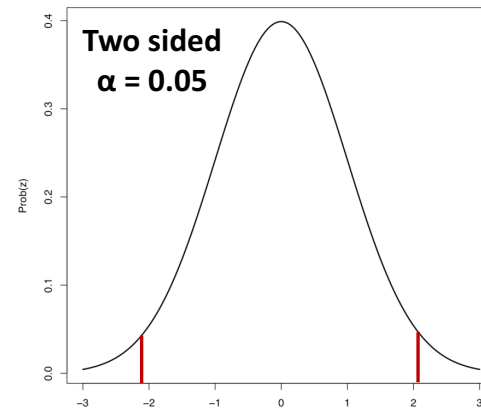
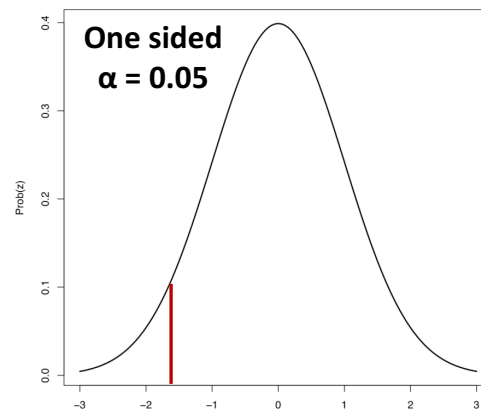
- The t-value follows a t-distribution



df=degrees of freedom

When to reject H_0

- **Level of significance, α :** specified before an experiment to define the rejection region.
- **Rejection region:** set of all test statistics values for which H_0 is rejected.
- **Critical value:** the smallest test value required to reject.



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
```

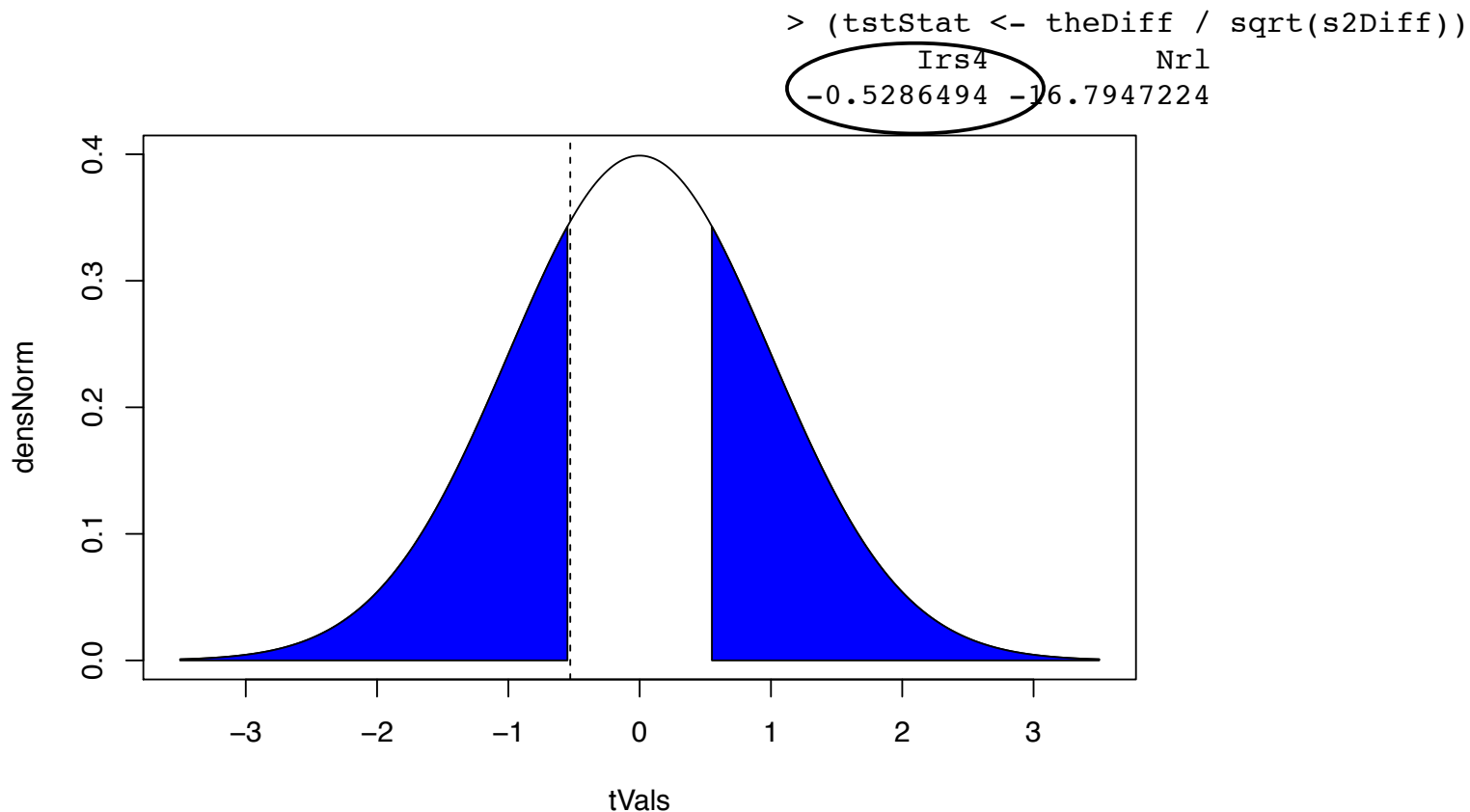
```
sample estimates:
```

```
mean in group wt mean in group NrlKO
```

```
11.244200
```

```
6.089632
```

Depicted here is the standard normal distribution (visually undistinguishable from T with $DF > \sim 58$)



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
```

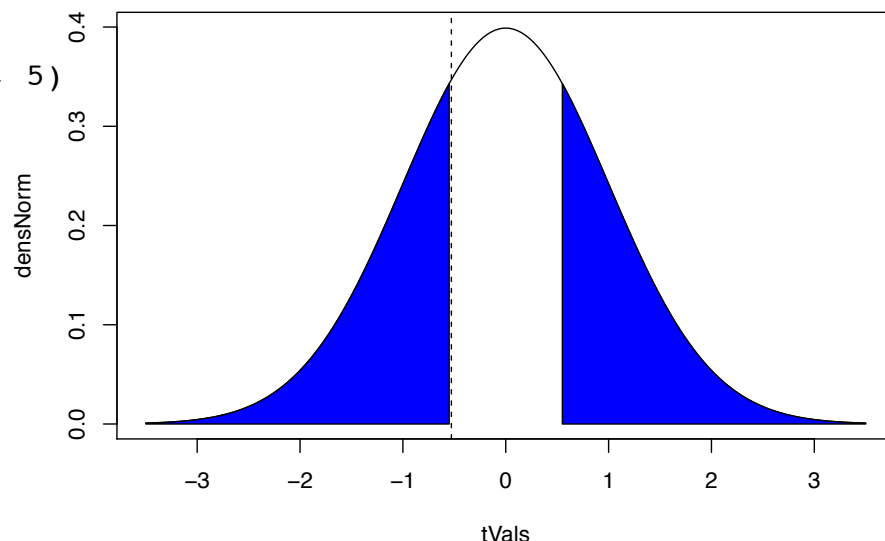
```
95 percent confidence interval:
```

```
-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
```

Hypothesis testing

1. Define a test-statistic (T statistic).
2. Compute the observed value for the test statistic.
3. Compute pvalue for the observed statistic under its null sampling distribution.
4. Make a decision about significance of results, based on a pre-specified critical value (α)

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 pvalue 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
 - Do not indicate the size or importance of observed effects.

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. $t \text{ stat} = 0.53$ for *Irs4* in our current example.

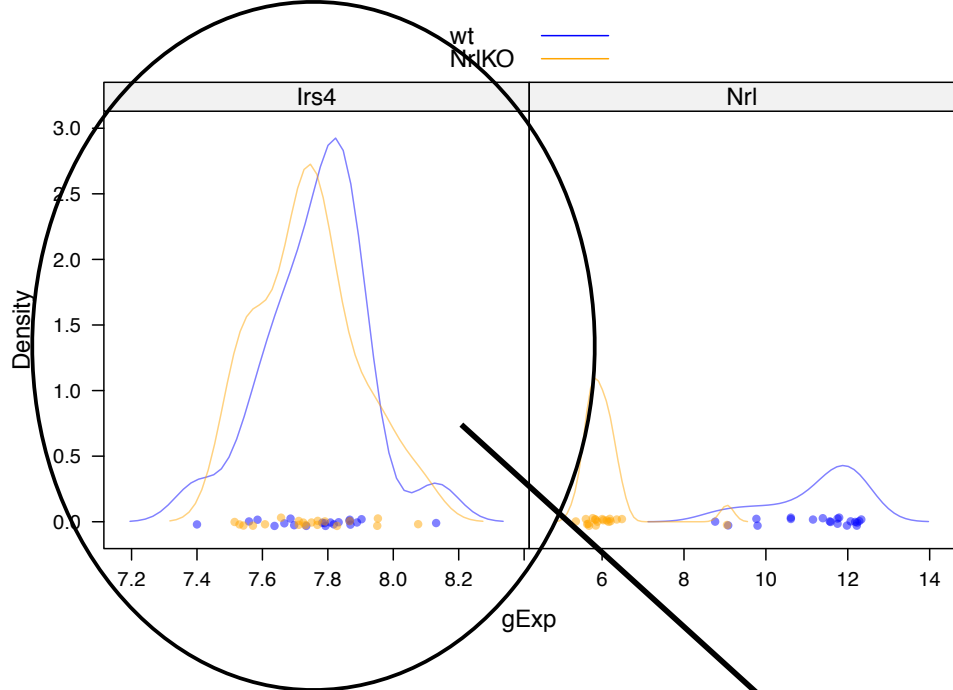
But you must think of it as a fleeting realization of a specific random variable.

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.

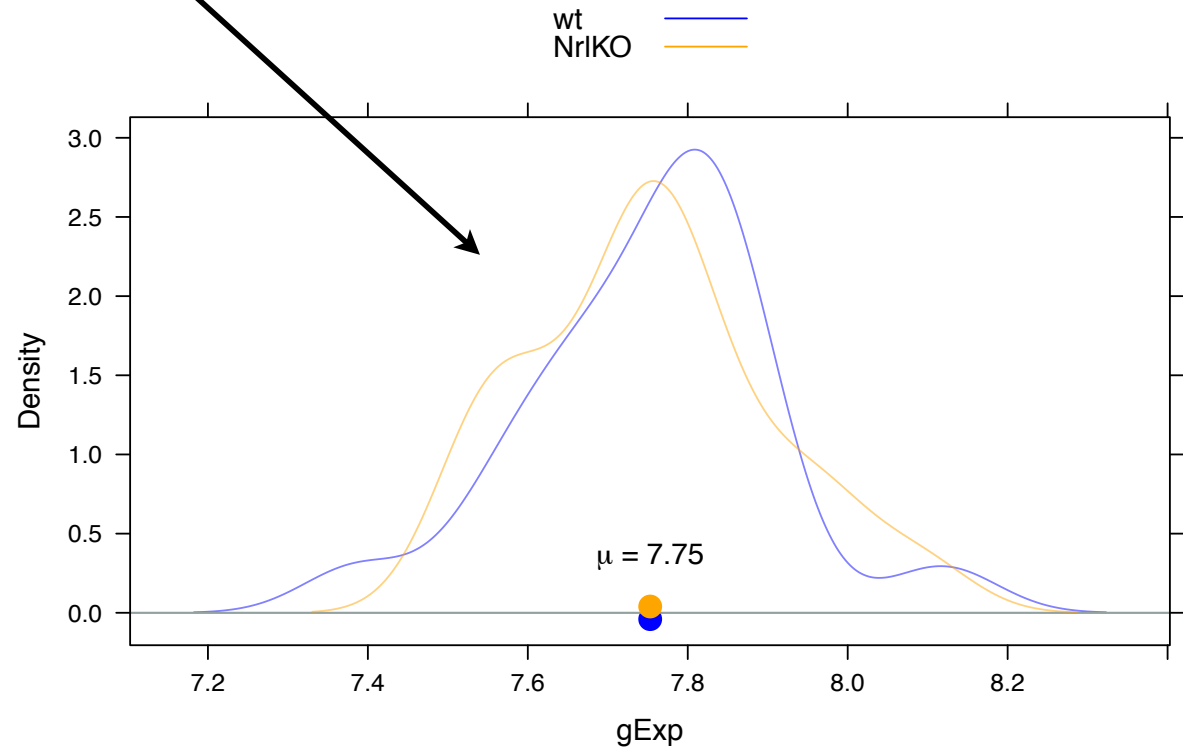
I will simulate data -- more blue Y's and more orange Z's -- and compute the observed difference of sample means and the t statistic.

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*.

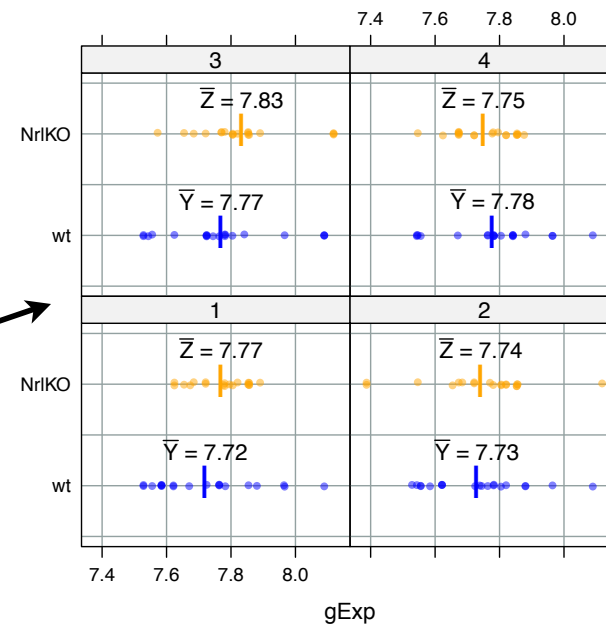
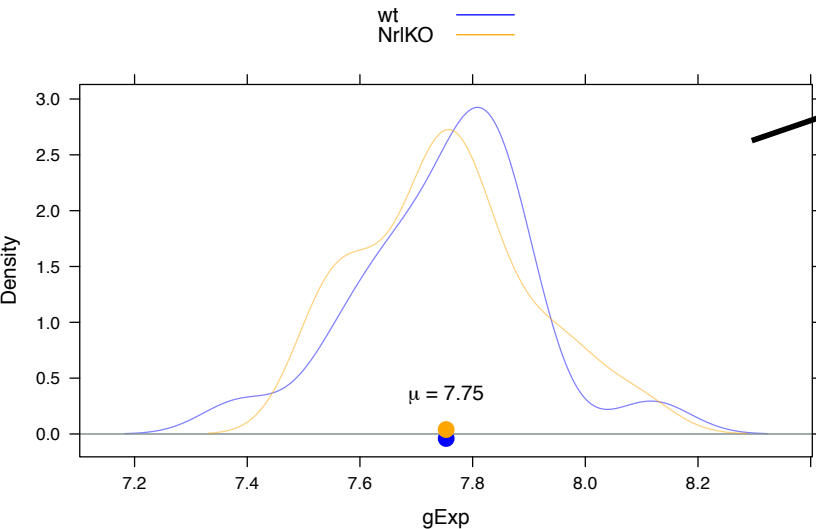


Our data-generating distributions are inspired by the observed data from Irs4.



Exact match
except wt and
NrlKO groups
have common
mean.

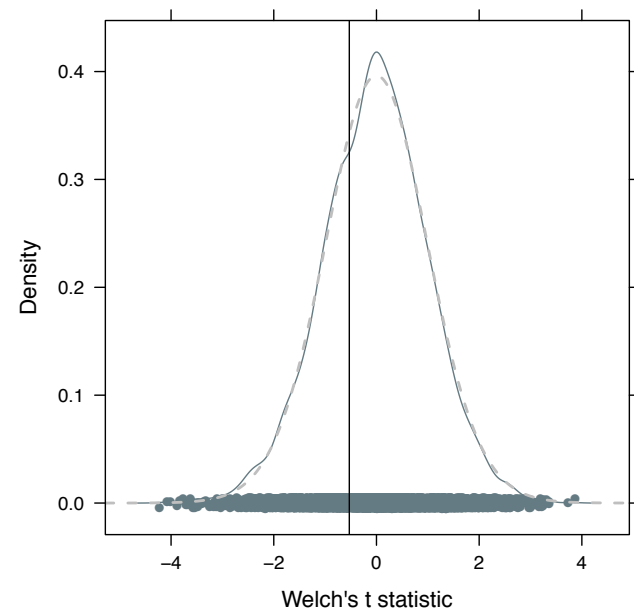
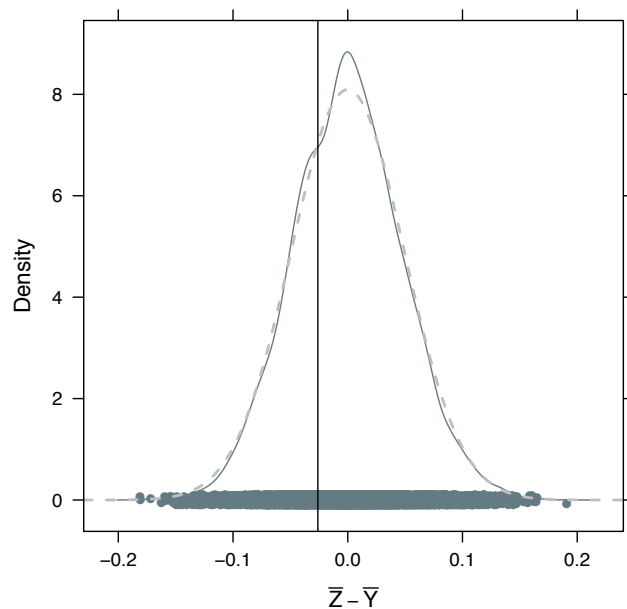
Underlying true dist'ns,
upholding the null
hypothesis of equal means



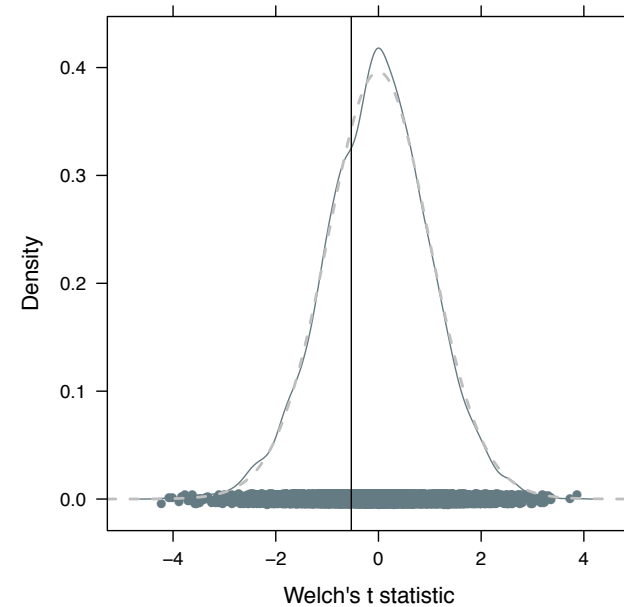
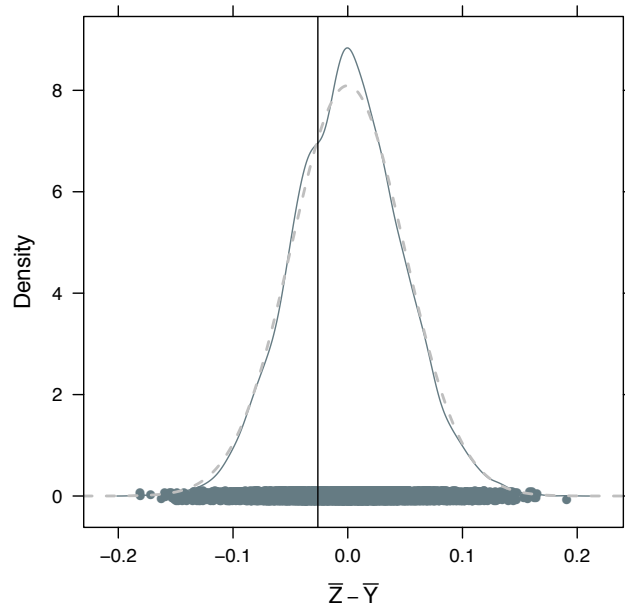
... and many many
more *in silico*
repeats of this
experiment ...

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
```

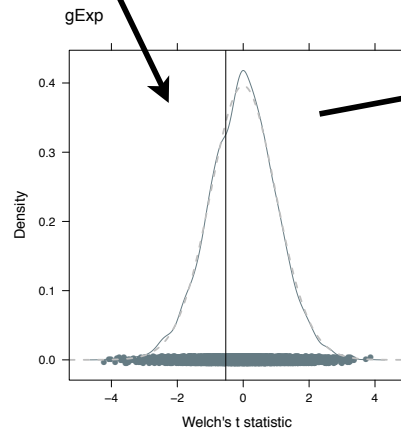
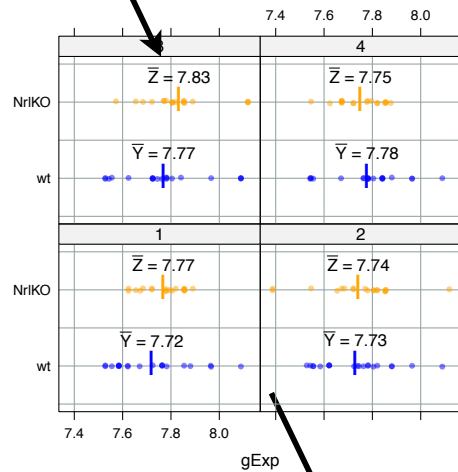
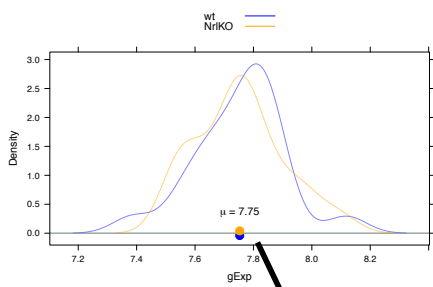
```
data: gExp by gType
```

```
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!

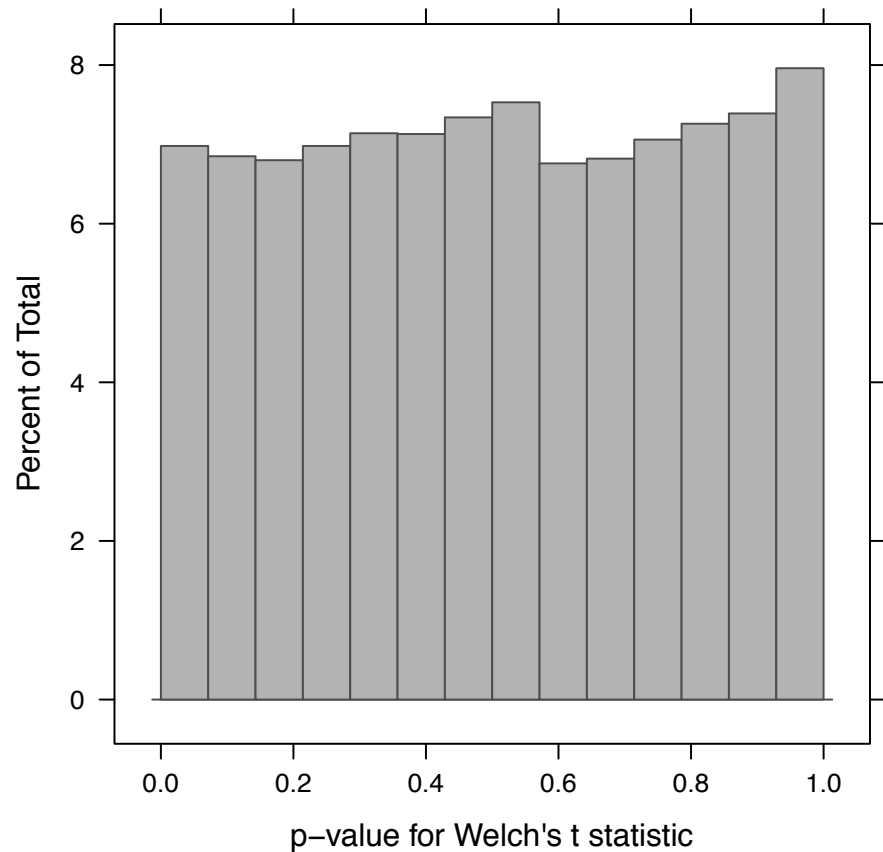


What does the distribution of the p-values look like when the null hypothesis holds?

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

It is, by definition, $\text{Unif}[0, 1]$.

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



“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?

T-Tests:

- One sample vs two samples
- One-sided vs two sided
- Paired vs unpaired
- Equal variance vs unequal variance

T-Tests:

- One sample vs two samples
- One-sided vs two sided
- Paired vs unpaired
- Equal variance vs unequal variance

Errors in hypothesis testing

| | | Actual Situation "Truth" | |
|----------------------|--|--|--|
| Decision | | H_0 True | H_0 False |
| Don 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 $1-\beta$ |

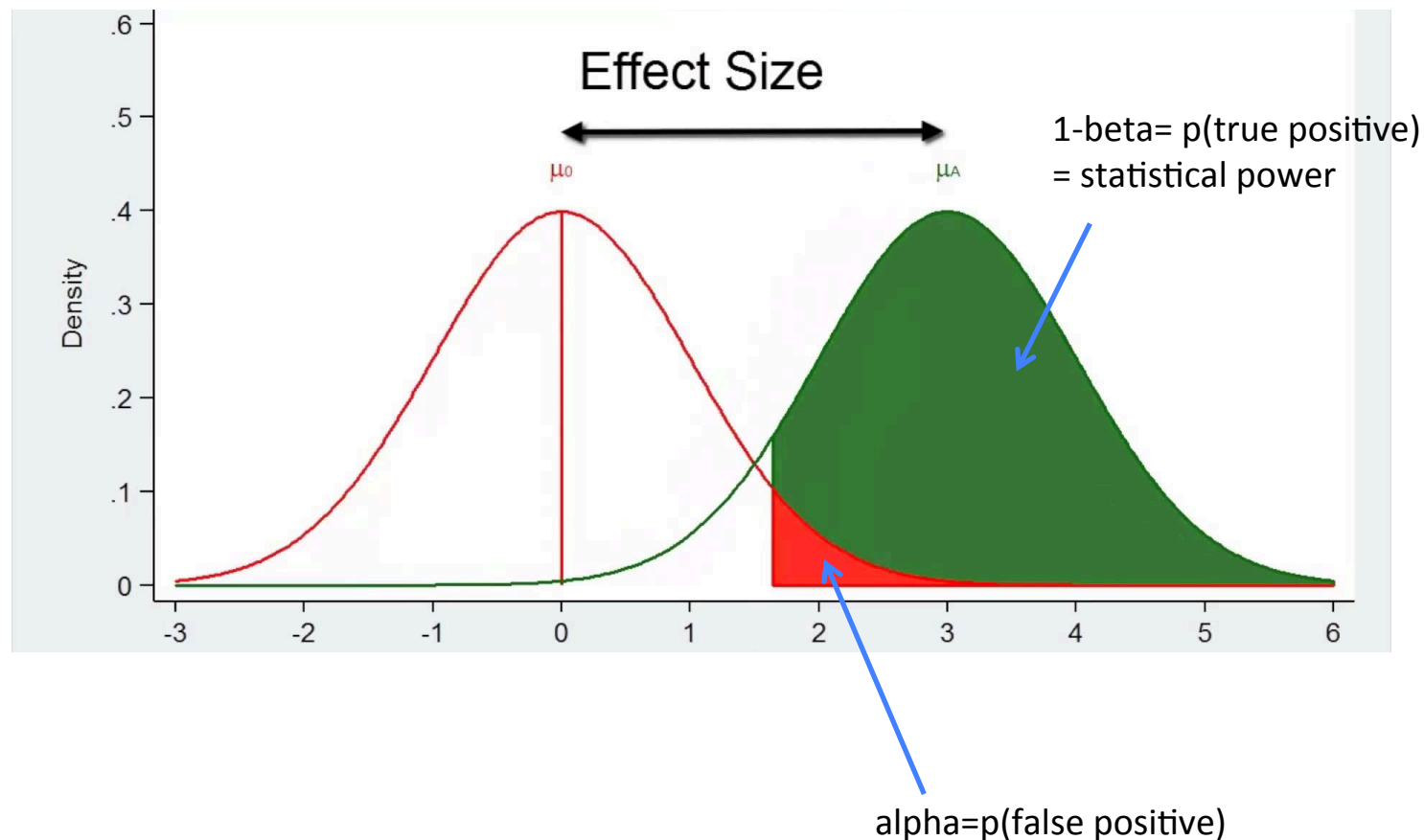
$\alpha = P(\text{Type I Error})$ $\beta = P(\text{Type II Error})$

Power = $1 - \beta$

Statistical power

statistical power is the likelihood that a study will detect an effect when there is an effect there to be detected

$$\text{Power} = p(\text{reject } H_0 \mid H_1)$$



How do we increase statistical power?

- Effect size
- Sample variance
- Sample size
- Significance threshold (i.e., type I error)

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 $n_Y (n_Y + 1)/2$)

(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

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: Nr1

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

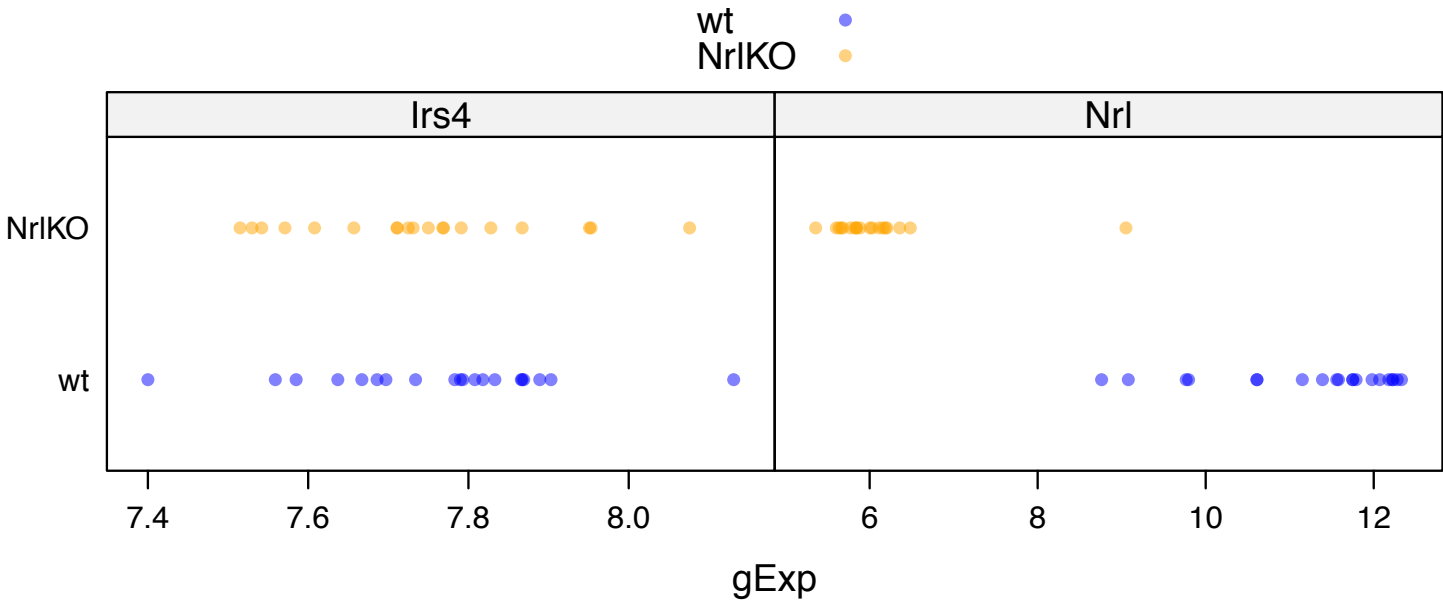
<snip, snip>

miniDat\$gene: Nr1

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

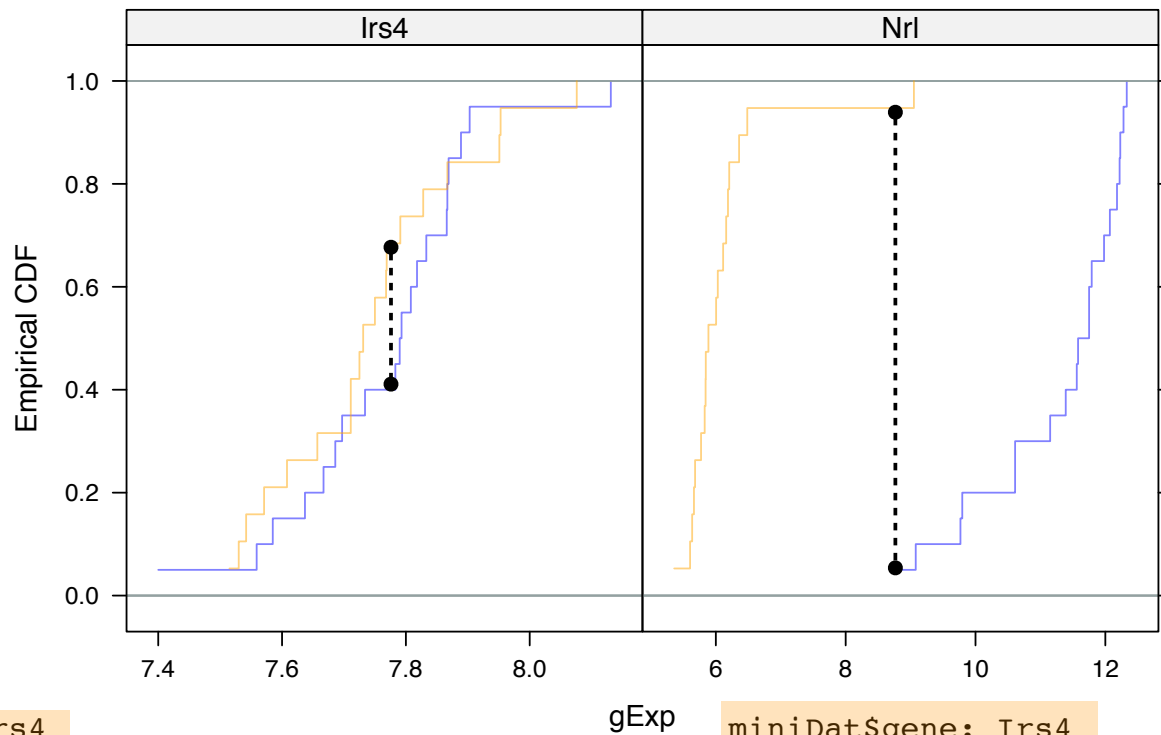
Estimate each CDF with the empirical CDF (ECDF)

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

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

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

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>

Discussion and questions ...

What if you are unsure whether your sample size is large enough?
Outliers with small samples could be problematic (i.e., results way too optimistic)

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

Is it cheating to report KS because it delivers an exciting, “statistically significant” result?

Answer: All part of being a good scientists. Treating pvalue as one type of evidence that you can incorporate with others. I'd probably perform multiple tests and report the most conservative. It is worrisome when methods that are equally appropriate and defensible give very different answers.