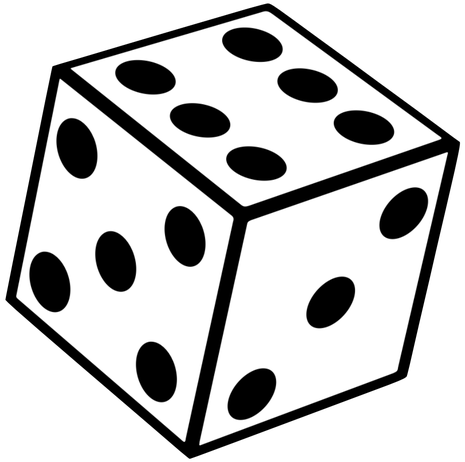
Lecture 10

11.2.21

The problem with multiple testing



1. If you roll a die once, what is the probability of rolling a 6?

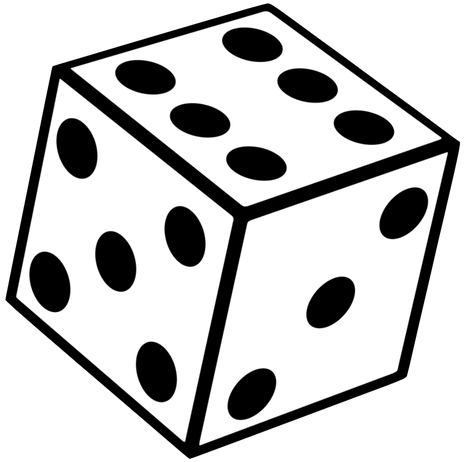
Pr{rolling a six}: 1/6 = 0.1667

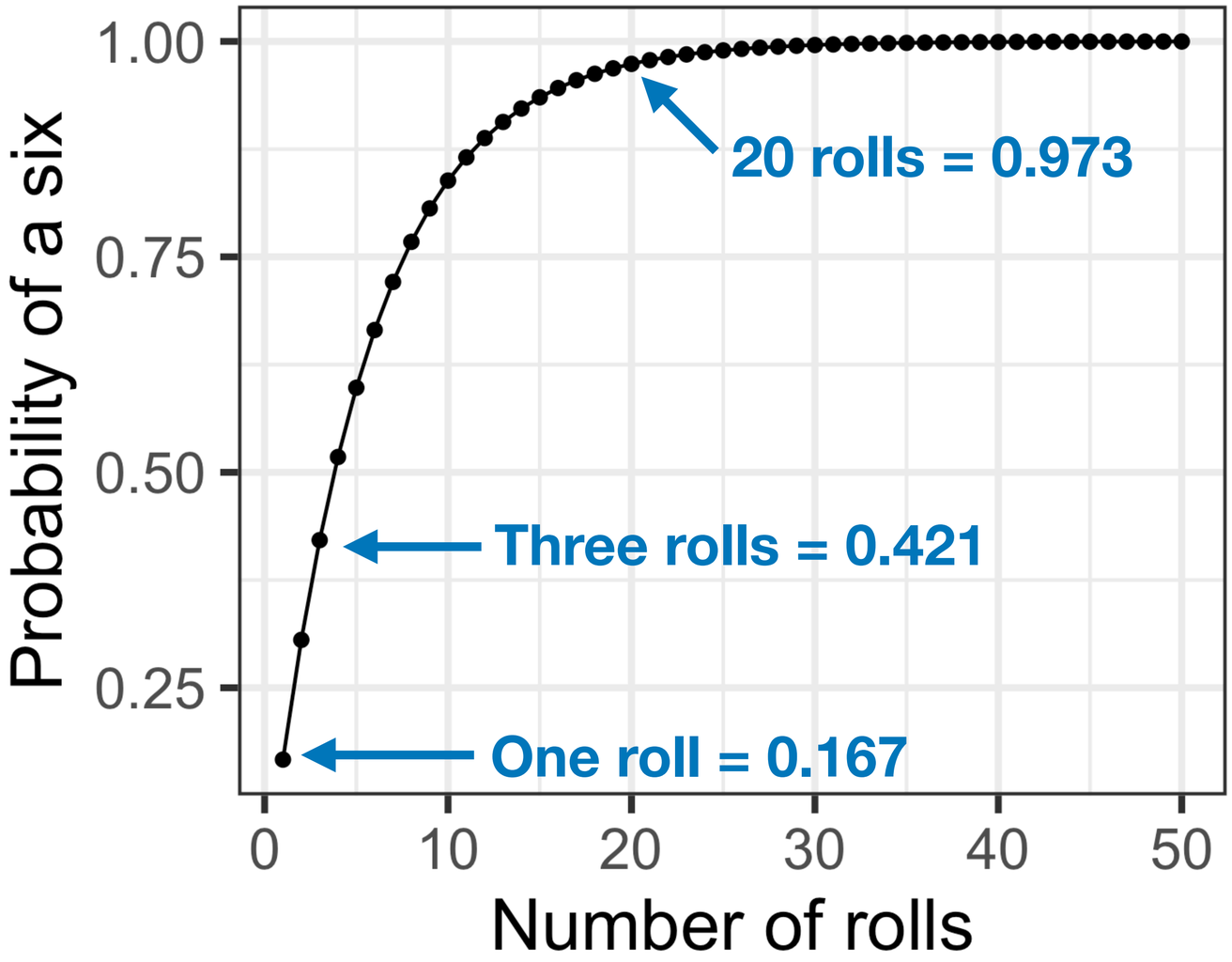
2. If you roll a die three times, what is the probability of rolling a 6 at least once?

Pr{at least one six} = 1 - Pr{no sixes}

Pr{at least one six} = 1 - [(5/6)(5/6)(5/6)] = 0.421

The problem with multiple testing





The problem with multiple testing

t test, $\alpha = 0.05$:

Remember: α is the Type I (false positive) error rate

One test:

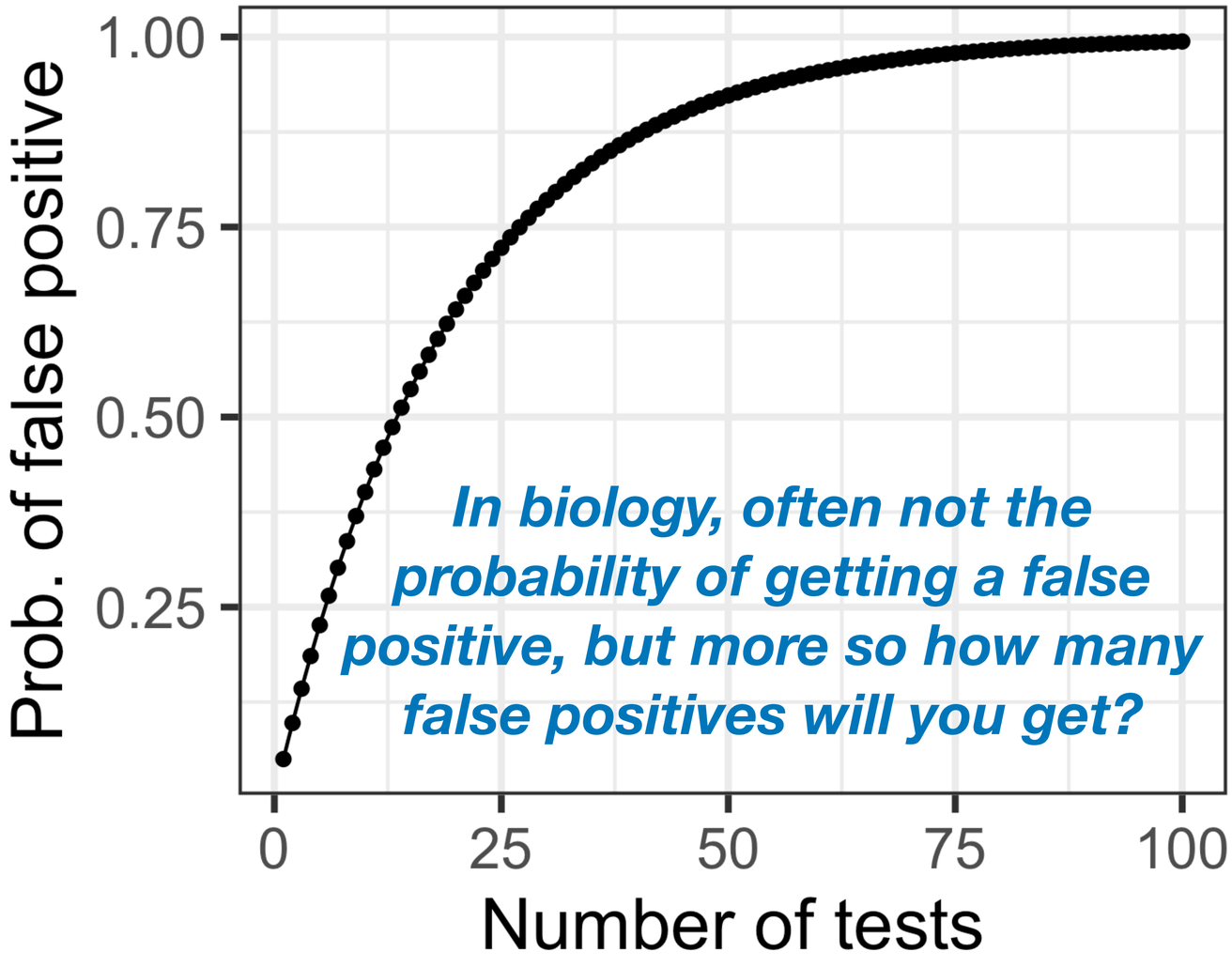
$$1 - (1 - 0.05)^1 = 0.05$$

20 tests:

$$1 - (1 - 0.05)^{20} = 0.641$$

100 tests:

$$1 - (1 - 0.05)^{100} = 0.994$$



In biology, often not the probability of getting a false positive, but more so how many false positives will you get?

Refresher: Type I and type II errors

Null hypothesis is…	**TRUE**	**FALSE**
REJECTED	Type I error *(False positive)*	**Correct!** (True positive)
NOT REJECTED	**Correct!** (True negative)	Type II error *(False negative)*

↓ **Type I Error** → ↑ **Type II Error**

Multiple hypothesis correction

Adjusting α in some way so that the probability of a test being significant by chance is lower (remains below the significance threshold)

Bonferroni correction (FWER)

False Discovery Rate (FDR)

Family-wise error rate (FWER)

FWER = probability of at least one type I error

- Most common FWER is the **Bonferroni correction**
  - Based on the idea that the probability that at least one of several events will occur cannot exceed the sum of the individual probabilities

0.05

$\alpha = 0.05; 3$ tests; H_0 is TRUE

P(at least one test is significant) $\leq 0.05 + 0.05 + 0.05$

P(at least one test is significant) $\leq 3(0.05)$

$$0.05 \leq 3(X)$$

$$\frac{0.05}{3} \leq X$$

Bonferroni: divide α / n and use this value as the new significance threshold α

The Bonferroni correction

Knockout of 10 different genes, measured a phenotype of interest and performed a *t*-test to compare each knockout to the control. The following are the p-values obtained:

Which genes are significant at $\alpha = 0.05$?

A	B	C	D	E	F	G	H	I	J
0.084	0.036	0.063	0.186	0.108	0.042	0.01	0.132	0.175	0.0012

Which genes are significant at $\alpha = 0.05/10 = 0.005$?

The Bonferroni correction

Knockout of 10 different genes, measured a phenotype of interest and performed a *t*-test to compare each knockout to the control. The following are the p-values obtained:

Which genes are significant at $\alpha = 0.05/10 = 0.005$?

A	B	C	D	E	F	G	H	I	J
0.084	0.036	0.063	0.186	0.108	0.042	0.01	0.132	0.175	0.0012

$$1 - (1 - 0.05)^{10} = 0.401 \longrightarrow 1 - (1 - \frac{0.05}{10})^{10} = 0.048$$

Bonferroni correction is very **strict/conservative**—It is based on the fact that the null hypothesis is true, and could lead to **a high percentage of false negatives**

The Bonferroni correction

- Assumes all tests are independent — although they often are not, which can lead to a high percentage of false negatives (i.e. Type II errors)
- Counter-intuitive: the interpretation of findings depends on the number of other tests run simultaneously
- Simple to perform and does reduce the false positives (i.e. Type I errors)

Bonferroni correction is very **strict/conservative—**It is based on the fact that the null hypothesis is true, and could lead to **a high percentage of false negatives**

Multiple hypothesis correction

Bonferroni correction (FWER)

- Easy to perform
- Very strict/conservative
- Set threshold to α/n
- You care more about preventing false positives than false negatives
- Sample size is high
- Effect of interest is very large and/or consistent

False Discovery Rate (FDR)

False discovery rate (FDR)

FDR = expected proportion of false positives among all significant results

Null hypothesis is…	TRUE	FALSE
REJECTED	**Type I error** ***(False positive)***	**Correct!** ***(True positive)***
NOT REJECTED	**Correct!** ***(True negative)***	**Type II error** ***(False negative)***

$$\text{FDR} = \frac{\text{Number of false positives}}{\text{False positives + true positives}}$$

False discovery rate (FDR)

FDR = expected proportion of false positives among all significant results

Null hypothesis is…	**TRUE**	**FALSE**
REJECTED	**Type I error** ***(False positive)***	**Correct!** ***(True positive)***
NOT REJECTED	**Correct!** ***(True negative)***	**Type II error** ***(False negative)***

$$\text{FDR} = \frac{\text{Number of false positives}}{\text{False positives + true positives}}$$

$$\text{FPR}^* = \frac{\text{Number of false positives}}{\text{False positives + true negatives}}$$

**false positive rate*

False discovery rate (FDR)

FDR = expected proportion of false positives among all significant results

- Benjamini-Hochberg method is a popular form of FDR
  - If many variables are significant, then surely there must be a true effect
  - Therefore, the actual chance of a false positive is much lower than FWER suggests
- FDR estimates the rejection region so that FDR < α

$$\text{FDR} = \frac{\text{Number of false positives}}{\text{False positives + true positives}}$$

J. R. Statist. Soc. B (1995)
57, *No.* 1, *pp.* 289–300

Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing

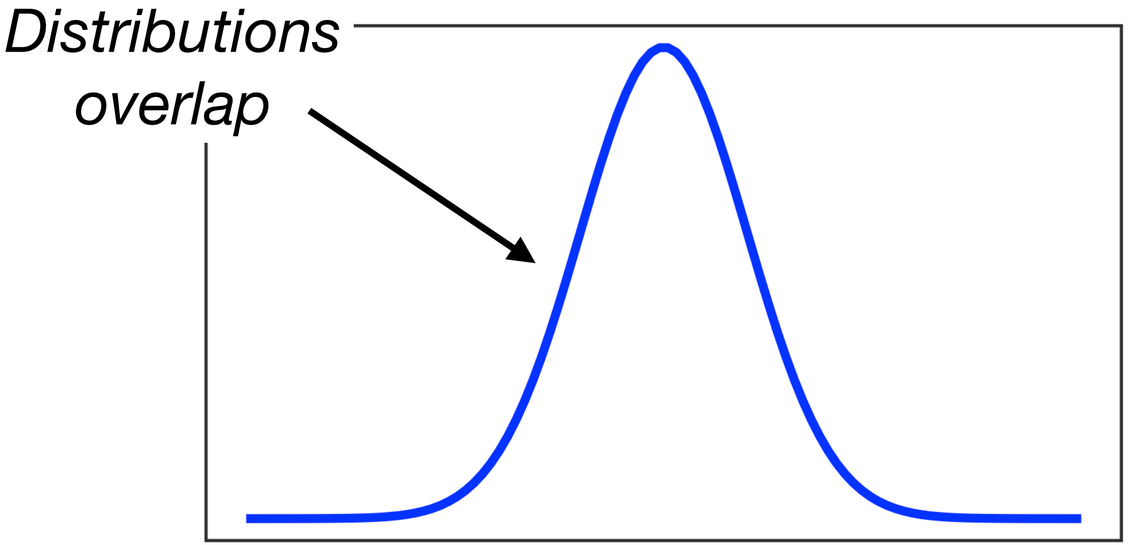
By YOAV BENJAMINI† and YOSEF HOCHBERG

Tel Aviv University, Israel

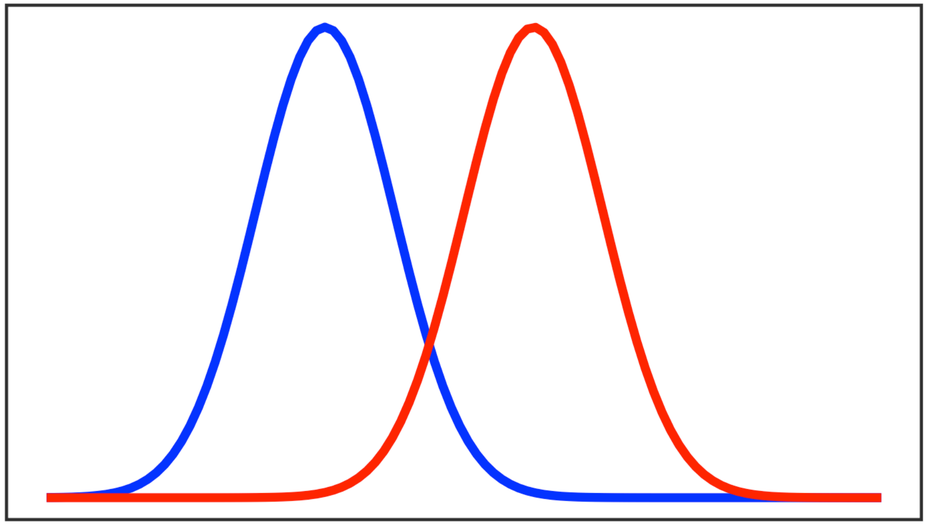
An aside on *p*-value distributions

Suppose we are testing many gene's expressions before and after drug treatment

Most of the time, genes will not be expressed differently



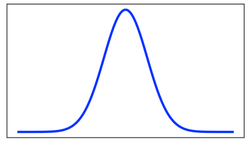
Sometimes, genes **will** be expressed differently



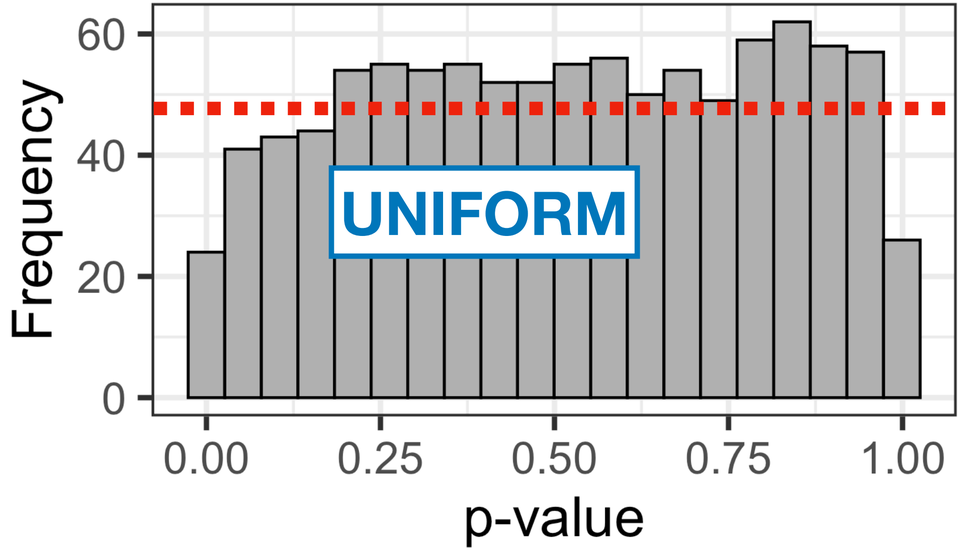
Remember: a *p*-value of 0.05 means 5% of the time you will get this value by chance

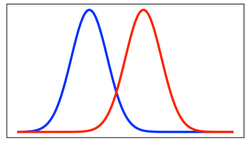
An aside on *p*-value distributions

Suppose we are testing many gene's expressions before and after drug treatment

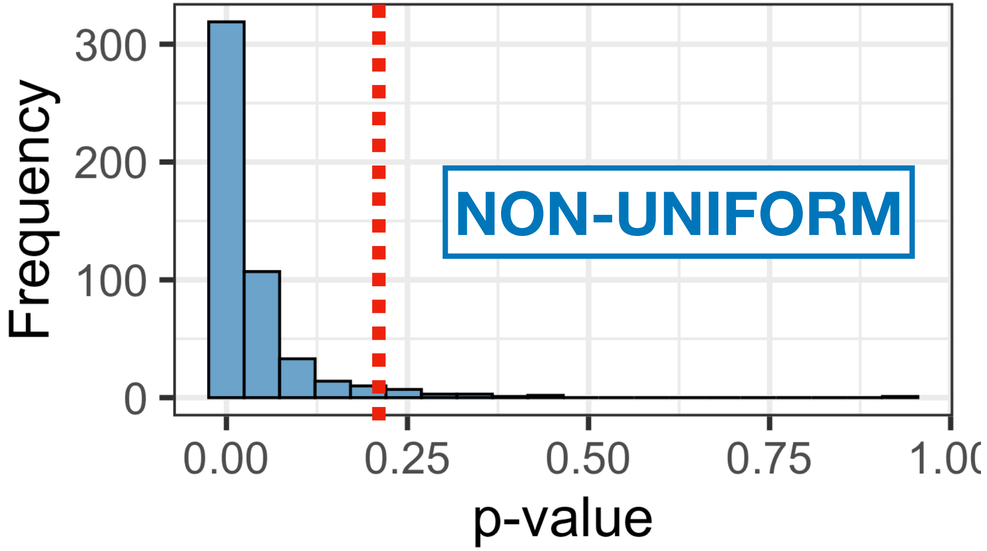


Most of the time, genes will not be expressed differently





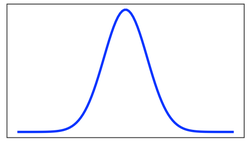
Sometimes, genes **will** be expressed differently



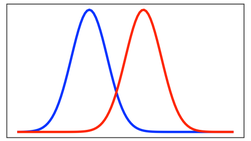
Remember: a *p*-value of 0.05 means 5% of the time you will get this value by chance

An aside on *p*-value distributions

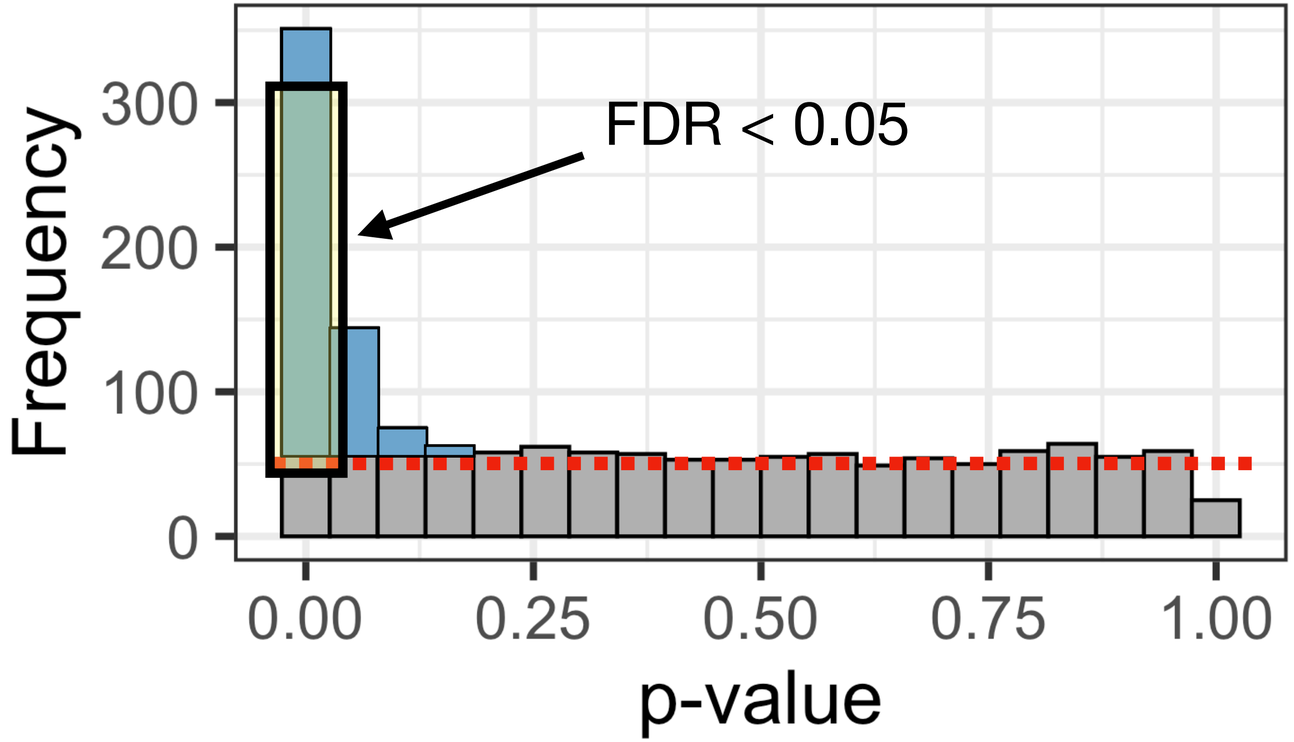
Suppose we are testing many gene's expressions before and after drug treatment



Most of the time, genes will not be expressed differently

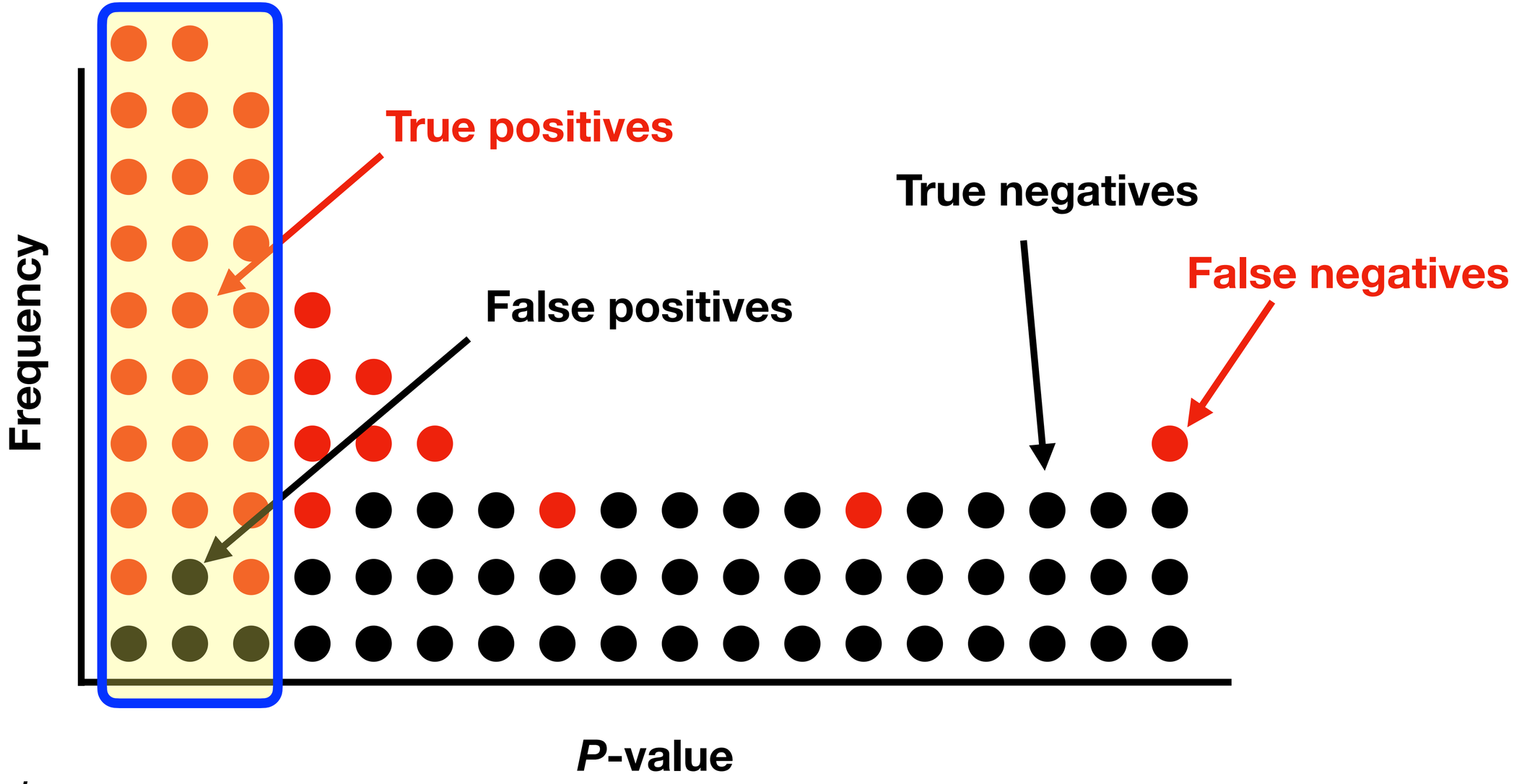


Sometimes, genes **will** be expressed differently



Adjusts p-values (makes them larger) so that **5% of the "significant"** results will be **false positives**.

An aside on *p*-value distributions



**not to scale*

Calculating the FDR

FDR = expected proportion of false positives among all significant results

A	B	C	D	E	F	G	H	I	J
0.084	0.036	0.063	0.186	0.108	0.042	0.024	0.132	0.175	0.0012

1. Order p-values from smallest to largest

Calculating the FDR

FDR = expected proportion of false positives among all significant results

J	G	B	F	C	A	E	H	I	D
0.0012	0.024	0.036	0.042	0.063	0.084	0.108	0.132	0.175	0.186

1. Order p-values from smallest to largest ✓

2. Rank the p-values

Calculating the FDR

FDR = expected proportion of false positives among all significant results

J	G	B	F	C	A	E	H	I	D
0.0012	0.024	0.036	0.042	0.063	0.084	0.108	0.132	0.175	0.186
1	2	3	4	5	6	7	8	9	10

1. Order p-values from smallest to largest ✓

2. Rank the p-values ✓

3. Adjust the p-values (starting with largest):

Current p-value * (total number of p-values)/(rank of p-value)

Calculating the FDR

FDR = expected proportion of false positives among all significant results

J	G	B	F	C	A	E	H	I	D
0.0012	0.024	0.036	0.042	0.063	0.084	0.108	0.132	0.175	0.186
1	2	3	4	5	6	7	8	9	10
									0.186

1. Order p-values from smallest to largest ✓

2. Rank the p-values ✓

3. Adjust the p-values (starting with largest):

Largest p-value is always the same

Current p-value * (total number of p-values)/(rank of p-value)

Calculating the FDR

FDR = expected proportion of false positives among all significant results

J	G	B	F	C	A	E	H	I	D
0.0012	0.024	0.036	0.042	0.063	0.084	0.108	0.132	0.175	0.186
1	2	3	4	5	6	7	8	9	10
								0.186	**0.186**

1. Order p-values from smallest to largest ✓

2. Rank the p-values ✓

0.175(10/9) = 0.194

3. Adjust the p-values (starting with largest):

Previous p-value **OR** Current p-value * (total number of p-values)/(rank of p-value)

Calculating the FDR

FDR = expected proportion of false positives among all significant results

J	G	B	F	C	A	E	H	I	D
0.0012	0.01	0.036	0.042	0.063	0.084	0.108	0.132	0.175	0.186
1	2	3	4	5	6	7	8	9	10
0.012	**0.05**	**0.12**	**0.105**	**0.126**	**0.14**	**0.154**	**0.165**	**0.186**	**0.186**

$\alpha = 0.05$

1. Order p-values from smallest to largest ✓

2. Rank the p-values ✓

3. Adjust the p-values (starting with largest): ✓

Previous p-value **OR** Current p-value * (total number of p-values)/(rank of p-value)

Multiple hypothesis correction

Bonferroni correction (FWER)

- Easy to perform
- Very strict/conservative
- Set threshold to α/n
- You care more about preventing false positives than false negatives
- Sample size is high
- Effect of interest is very large and/or consistent

False Discovery Rate (FDR)

- More complicated to do
- More permissive
- Set threshold so 5% of positives are false
- Both minimize false positives and keep false negatives low
- Limited sample size
- Effect of interest is not very large nor consistent

Comparing Bonferroni vs. FDR

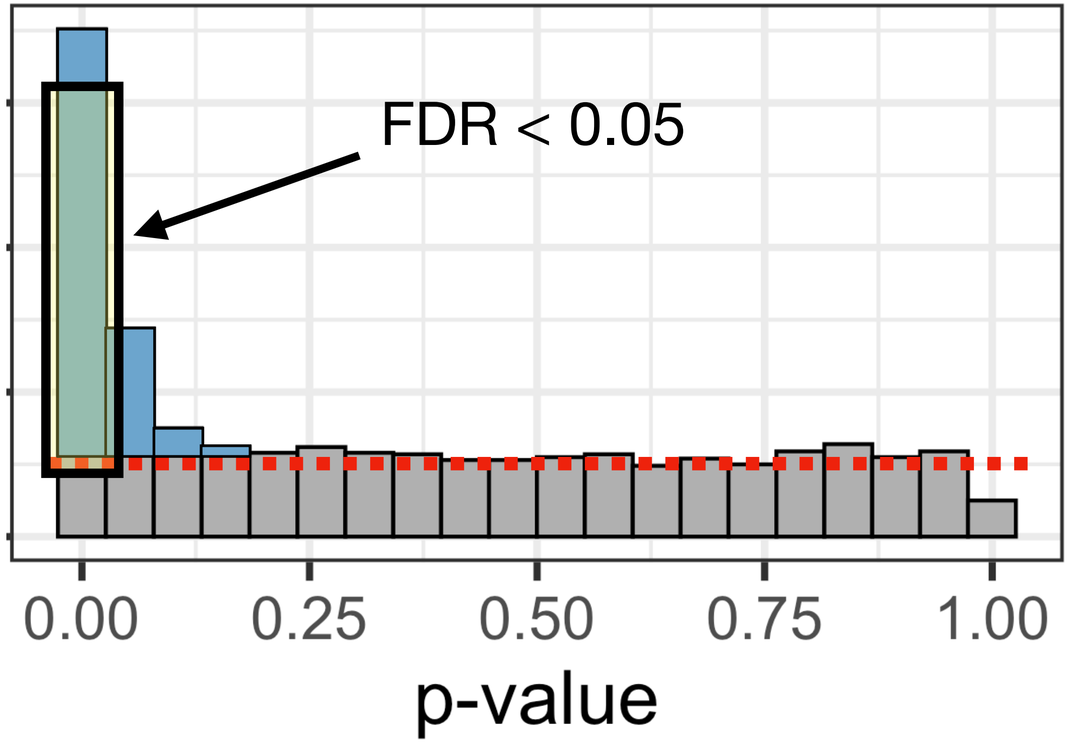
	J	G	B	F	C	A	E	H	I	D
p-value	0.0012	0.01	0.036	0.042	0.063	0.084	0.108	0.132	0.175	0.186
$\alpha < 0.05$	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
BF	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FDR	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

- Very significant cases (i.e. gene “J”) will always be significant
- Border cases are often the most interesting (and common) in biology
- **One may use FDR to explore and Bonferroni/FWER to confirm**

Bonferroni and FDR correction in R

```
# create vector of p-values
> vals <- c(0.0012, 0.01, 0.036, 0.042, 0.063,
0.084, 0.108, 0.132, 0.175, 0.186)

# adjust p-values with bonferroni correction
> p.adjust(vals, method = "bonferroni")

[1] 0.012 0.100 0.360 0.420 0.630 0.840 1.000 1.000
1.000 1.000

# adjust p-values with FDR correction
> p.adjust(vals, method = "fdr")

[1] 0.01200 0.05000 0.10500 0.10500 0.12600 0.14000
0.15428 0.16500 0.18600 0.18600
```

Bonferroni and FDR correction in R



$$\text{FDR} = \frac{\text{Number of false positives}}{\text{False positives + true positives}}$$

```
# Table of unadjusted p-values
that are significant
                        FALSE TRUE
  Two-distributions       114  386
  Uniform                 957   43

> fdr <- 43 / (43 + 386)

# Table of FDR-adjusted p-values
that are significant
                        FALSE TRUE
  Two-distributions       350  150
  Uniform                 995    5

> fdr <- 5 / (5 + 150)
```

Beyond Bonferroni and FDR

- Adjusting the p-value to reduce number of false positives is a very active area of statistics
- There are many more ways, but these two are perhaps the most common, especially in biology
- Others: positive false discovery rate (pFDR), Holm (type of FWER), local false discovery rate (local FDR), permutation/randomization

Beyond Bonferroni and FDR

- Adjusting the p-value to reduce number of false positives is a very active area of statistics
- There are many more ways, but these two are perhaps the most common, especially in biology
- Others: positive false discovery rate (pFDR), Holm (type of FWER), local false discovery rate (local FDR), **permutation/randomization**

Permutation and FWER

- One main disadvantage for Bonferroni FWER correction is that it assumes all tests are independent, when often they are correlated

- Permutation, by definition, tests the null hypothesis that there is no effect and breaks down any correlation structure of the data

- **(1) Randomize the data and conduct hypothesis test ~1000 times**

- **(2) Pick the right threshold value such that (if $\alpha = 0.05$) 5% of tests are significant (by chance)**

- **(3) Compare original p-values to new threshold**

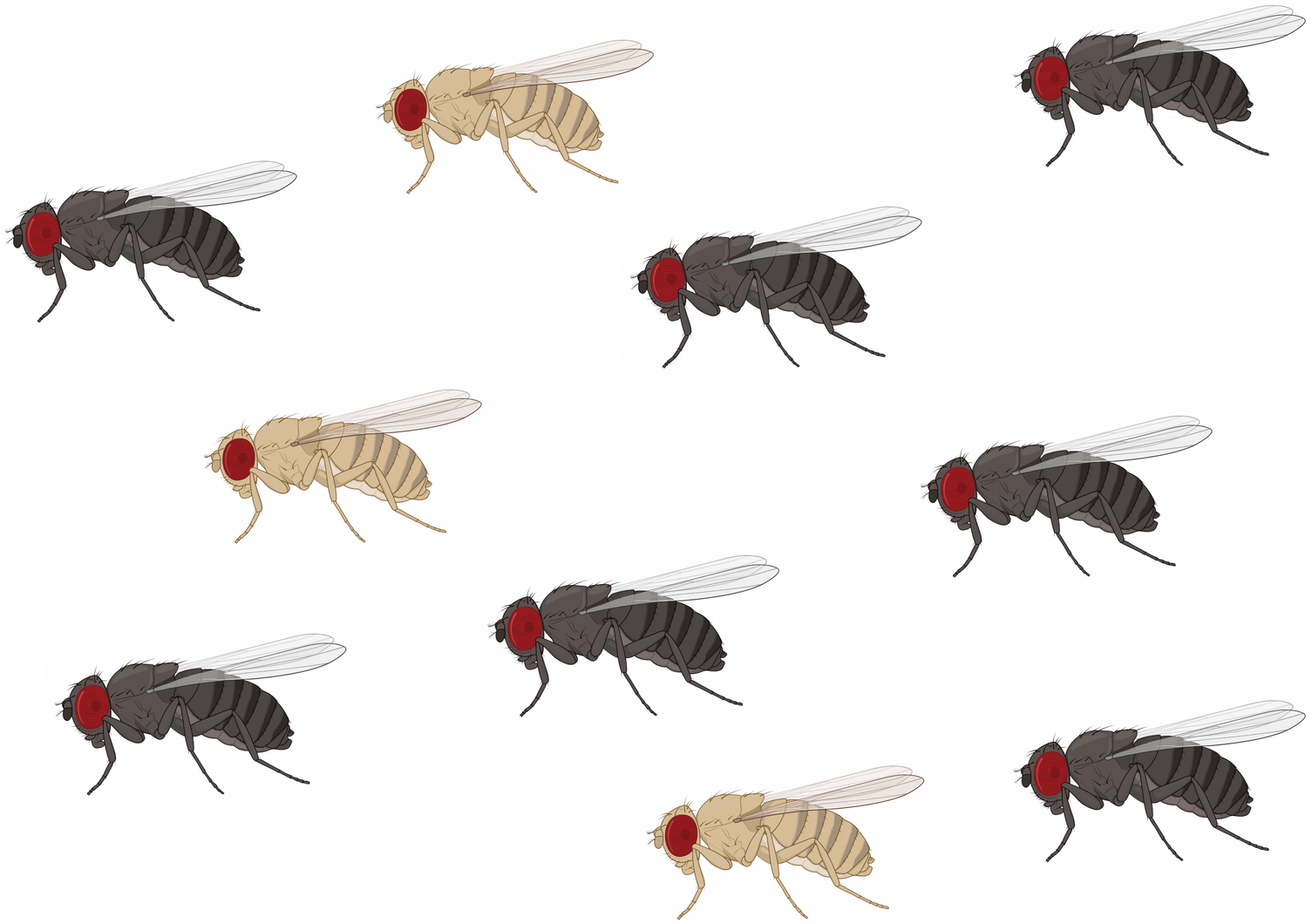
Introduction to enrichment analysis

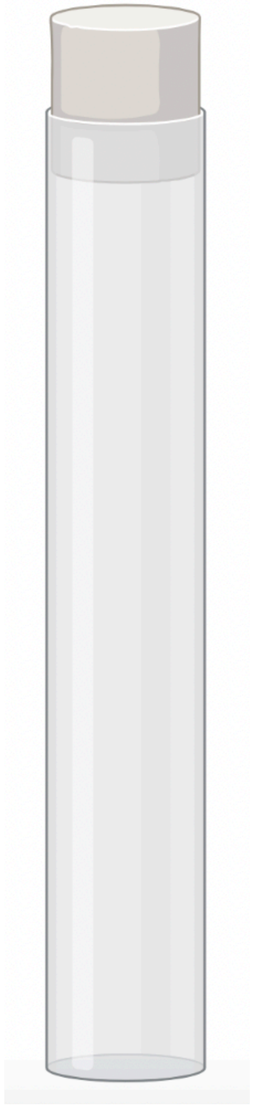
Suppose you measured expression of 10,000 transcripts. After multiple hypothesis correction, you found 200 were significant. Interestingly, many of these 200 significant transcripts looked like they function in the immune system, which would be an exciting discovery!

Should you be excited or skeptical?

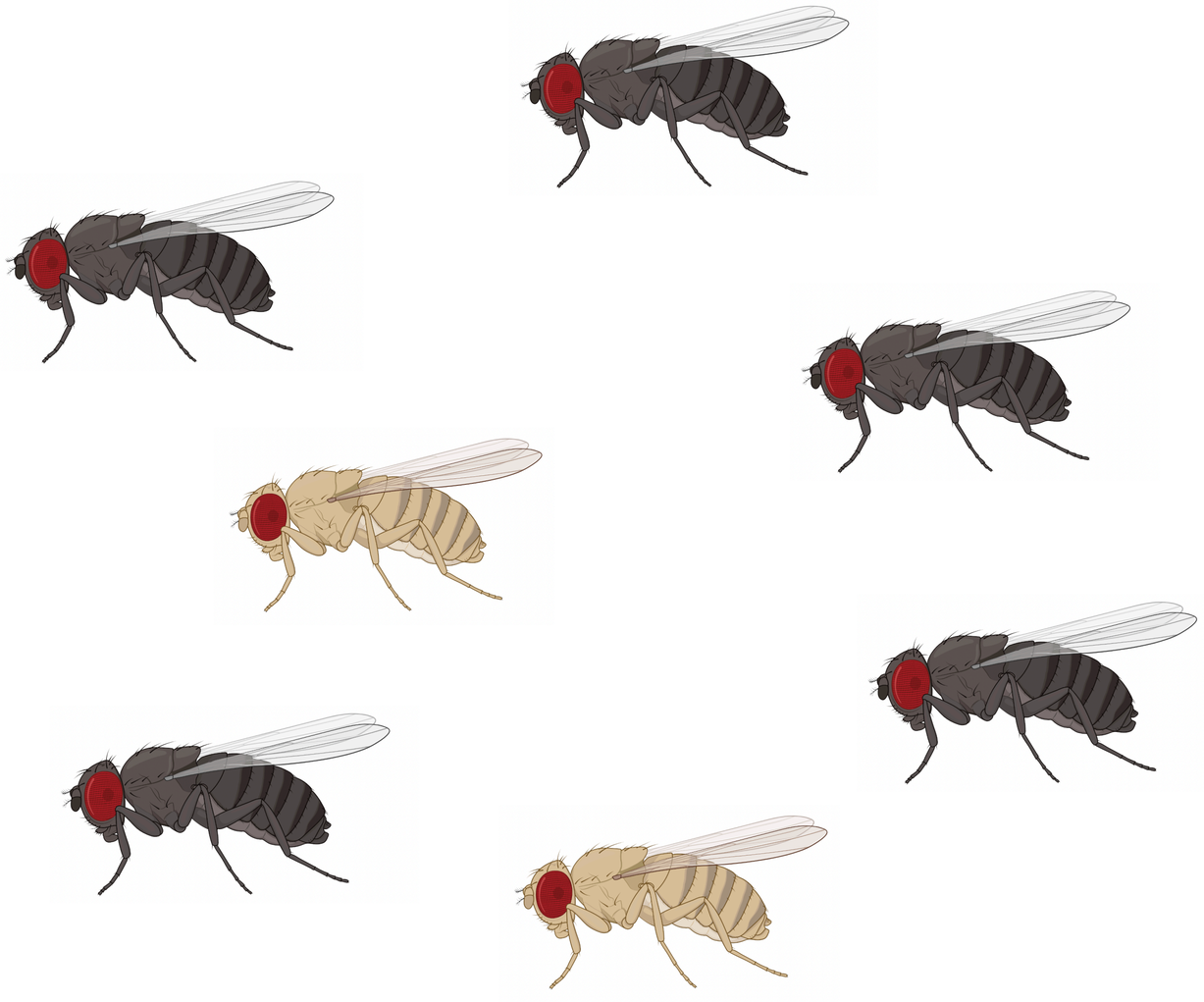
Skeptical. Immune genes are relatively abundant in the genome, we need more information. **Do you have more significant immune genes that you'd expect if all gene types were equally abundant?** (i.e. is my list **enriched** for immune genes?)

Polya urn models and the hypergeometric distribution





Polya urn models and the hypergeometric distribution



What is the probability of choosing 2 black flies and 1 grey fly?

Is this binomial?

No!!! Why?

(Sampling without replacement, not independent, different p-values)

Polya urn models and the **hypergeometric distribution**

Pr{2 black + 1 white} =

7/10 X 6/9 X 3/8 = 0.175

7/10 X 3/9 X 6/8 = 0.175

7/10 X 3/9 X 6/8 = 0.175

= 0.525

7  **;** **3** 

Polya urn models and the **hypergeometric distribution**

$$\Pr\{2 \text{ black} + 1 \text{ white}\} = \frac{(\text{How many ways to get 2 black}) \times (\text{How many ways to get 1 white})}{(\text{How many ways to get 3 flies})}$$

$$= \frac{{}_7C_2 \times {}_3C_1}{{}_{10}C_3}$$

$$= 0.525$$

7  ; 3 

Polya urn models and the **hypergeometric distribution**

$$\Pr\{x \text{ black}\} = \frac{{}_mC_x \times {}_nC_{k-x}}{{}_{m+n}C_k}$$

	Black	White	Total
Chosen	x	k-x	**k**
Not chosen	m-x	n-k+x	**m+n-k**
Total	**m**	**n**	**m+n**

7  **;** **3** 

m = number of black flies
n = number of white flies

k = total number of flies chosen
x = number of black flies chosen

Polya urn models and the **hypergeometric distribution**

$$\Pr\{x \text{ black}\} = \frac{{}_mC_x \times {}_nC_{k-x}}{{}_{m+n}C_k}$$

	Black	White	Total
Chosen	2	1	**3**
Not chosen	5	2	**7**
Total	**7**	**3**	**10**

7  **;** **3** 

m = number of black flies
n = number of white flies

k = total number of flies chosen
x = number of black flies chosen

Polya urn models and the **hypergeometric distribution**

$$\Pr\{2 \text{ black} + 1 \text{ white}\} = \frac{{}_7C_2 \times {}_3C_1}{{}_{10}C_3}$$

```
> dhyper(x, m, n, k)
```

	Black	White	Total
Chosen	2	1	**3**
Not chosen	5	2	**7**
Total	**7**	**3**	**10**

7  **;** **3** 

m = number of black flies
n = number of white flies

k = total number of flies chosen
x = number of black flies chosen

Polya urn models and the **hypergeometric distribution**

Pr{2 black + 1 white} $= \dfrac{{}_{7}C_{2} \times {}_{3}C_{1}}{{}_{10}C_{3}}$

```
> dhyper(x=2, m=7, n=3, k=3)
```

	Black	White	Total
Chosen	2	1	**3**
Not chosen	5	2	**7**
Total	**7**	**3**	**10**

7  **;** **3** 

m = number of black flies
n = number of white flies

k = total number of flies chosen
x = number of black flies chosen

Polya urn models and the **hypergeometric distribution**

What is the probability of choosing at least one white fly (out of 3)?

Pr{at least 1 white} = Pr{1 white} + Pr{2 whites} + Pr{3 whites}

= 1 - Pr{0 whites}

7  **;** **3** 

m = number of **white** flies

n = number of **black** flies

k = total number of flies chosen

x = number of **white** flies chosen

Polya urn models and the **hypergeometric distribution**

What is the probability of choosing at least one white fly (out of 3)?

Pr{at least 1 white} = Pr{1 white} + Pr{2 whites} + Pr{3 whites}

	Black	White	Total
Chosen	3	0	**3**
Not chosen	4	3	**7**
Total	**7**	**3**	**10**

= 1 - Pr{0 whites}

$$= 1 - \frac{{}_{m}C_{x} \times {}_{n}C_{k-x}}{{}_{m+n}C_{k}}$$

7  **;** **3** 

m = number of **white** flies

n = number of **black** flies

k = total number of flies chosen

x = number of **white** flies chosen

Polya urn models and the **hypergeometric distribution**

What is the probability of choosing at least one white fly (out of 3)?

Pr{at least 1 white} = Pr{1 white} + Pr{2 whites} + Pr{3 whites}

	Black	White	Total
Chosen	3	0	**3**
Not chosen	4	3	**7**
Total	**7**	**3**	**10**

= 1 - Pr{0 whites}

$$= 1 - \frac{{}_{3}C_{0} \times {}_{7}C_{3-0}}{{}_{3+7}C_{3}}$$

```
> 1 - dhyper(x=0, m=3, n=7, k=3)
```

= **0.708**

7  **; 3** 

m = number of **white** flies

n = number of **black** flies

k = total number of flies chosen

x = number of **white** flies chosen

Polya urn models and the **hypergeometric distribution**

What is the probability of choosing at least one white fly (out of 3)?

Pr{at least 1 white} = Pr{1 white} + Pr{2 whites} + Pr{3 whites}

	Black	White	Total
Chosen	3	0	**3**
Not chosen	4	3	**7**
Total	**7**	**3**	**10**

```
> 1 - phyper(q=0, m=3, n=7, k=3)
```

= 1 - Pr{0 whites}

$$= 1 - \frac{{}_3C_0 \times {}_7C_{3-0}}{{}_{3+7}C_3}$$

= **0.708**

7  **;** **3** 

m = number of **white** flies
n = number of **black** flies

k = total number of flies chosen
x = number of **white** flies chosen

Polya urn models and the **hypergeometric distribution**

What is the probability of choosing at least one white fly (out of 3)?

Pr{at least 1 white} = Pr{1 white} + Pr{2 whites} + Pr{3 whites}

	Black	White	Total
Chosen	3	0	**3**
Not chosen	4	3	**7**
Total	**7**	**3**	**10**

= 1 - Pr{0 whites}

$$= 1 - \frac{{}_{3}C_{0} \times {}_{7}C_{3-0}}{{}_{3+7}C_{3}}$$

```
> phyper(q=0, m=3, n=7, k=3, lower.tail = F)
```

= **0.708**

7  ; **3** 

m = number of **white** flies

n = number of **black** flies

k = total number of flies chosen

x = number of **white** flies chosen

Introduction to enrichment analysis

Suppose you measured expression of 10,000 transcripts. After multiple hypothesis correction, you found 200 were significant. Interestingly, 80 of these 200 significant transcripts looked like they function in the immune system, which would be an exciting discovery!

Pr{80+ sig. immune genes} = Pr(80) + Pr(81) + Pr(82) + … + P(200)

	Significant	Not sig.	Total
Immune	80		**3,000**
Not immune			
Total	**200**	**9,800**	**10,000**

Introduction to enrichment analysis

Suppose you measured expression of 10,000 transcripts. After multiple hypothesis correction, you found 200 were significant. Interestingly, 80 of these 200 significant transcripts looked like they function in the immune system, which would be an exciting discovery!

Pr{80+ sig. immune genes} = Pr(80) + Pr(81) + Pr(82) + … + P(200)

	Significant	Not sig.	Total
Immune	**80**	2,920	**3,000**
Not immune	120	6,880	**7,000**
Total	**200**	**9,800**	**10,000**

We could do it by hand…

… but it would take a LONG time

Introduction to enrichment analysis

Suppose you measured expression of 10,000 transcripts. After multiple hypothesis correction, you found 200 were significant. Interestingly, 80 of these 200 significant transcripts looked like they function in the immune system, which would be an exciting discovery!

Pr{80+ sig. immune genes} = Pr(80) + Pr(81) + Pr(82) + … + P(200)

	Significant	Not sig.	Total
Immune	**80**	2,920	**3,000**
Not immune	120	6,880	**7,000**
Total	**200**	**9,800**	**10,000**

We could do it by hand…

… but it would take a LONG time

Strictly GREATER THAN (not equal to)

```
> phyper(q=79, m=200, n=9800, k=3000, lower.tail = F)
```

Introduction to enrichment analysis

Suppose you measured expression of 10,000 transcripts. After multiple hypothesis correction, you found 200 were significant. Interestingly, 80 of these 200 significant transcripts looked like they function in the immune system, which would be an exciting discovery!

Pr{80+ sig. immune genes} = Pr(80) + Pr(81) + Pr(82) + … + P(200)

	Significant	Not sig.	Total
Immune	**80**	2,920	**3,000**
Not immune	120	6,880	**7,000**
Total	**200**	**9,800**	**10,000**

[1] 0.001479778

Strictly GREATER THAN (not equal to)

```
> phyper(q=79, m=200, n=9800, k=3000, lower.tail = F)
```

Enrichment and the **Fisher's Exact Test**

Pr{80+ sig. immune genes} = Pr(80) + Pr(81) + Pr(82) + … + P(200)

	Significant	Not sig.	Total
Immune	80	2,920	**3,000**
Not immune	120	6,880	**7,000**
Total	**200**	**9,800**	**10,000**

H_0 : P(Sig. immune) = P(not sig immune)

H_A : P(Sig. immune) > P(not sig immune)

```
# make data frame
> genes <- data.frame(sig = c(80, 120), not_sig = c(2920, 6880))

# run fishers exact test - greater (for enrichment)
> fisher.test(genes, alternative = "greater")
```

Enrichment and the **Fisher's Exact Test**

Pr{80+ sig. immune genes} = Pr(80) + Pr(81) + Pr(82) + … + P(200)

	Significant	Not sig.
Immune	80	2,920
Not immune	120	6,880
Total	**200**	**9,800**

```
        Fisher's Exact Test for Count Data

data:  genes
p-value = 0.00148
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 1.220544      Inf
sample estimates:
odds ratio
    1.5707
```

[1] 0.001479778

```
# make data frame
> genes <- data.frame(sig

# run fishers exact test - greater (for enrichment)
> fisher.test(genes, alternative = "greater")
```

Enrichment and the **Fisher's Exact Test**

Q: Is the Fisher's exact test a parametric or non-parametric test?

A: It is a non-parametric test! (Because there are no assumptions about the underlying population distribution)

Q: Why is it called the Fisher's <u>EXACT</u> test?

A: It calculates the <u>EXACT</u> *p*-value (the probability that we see <u>OUR</u> data (or more extreme) out of all the possible combinations). Unlike a *t*-test, the *p*-value is NOT estimated from a distribution.

Q: Can we have a non-directional Fisher's exact test?

A: Yes! But cumbersome to calculate by hand… almost always want to use R's `fisher.test()`