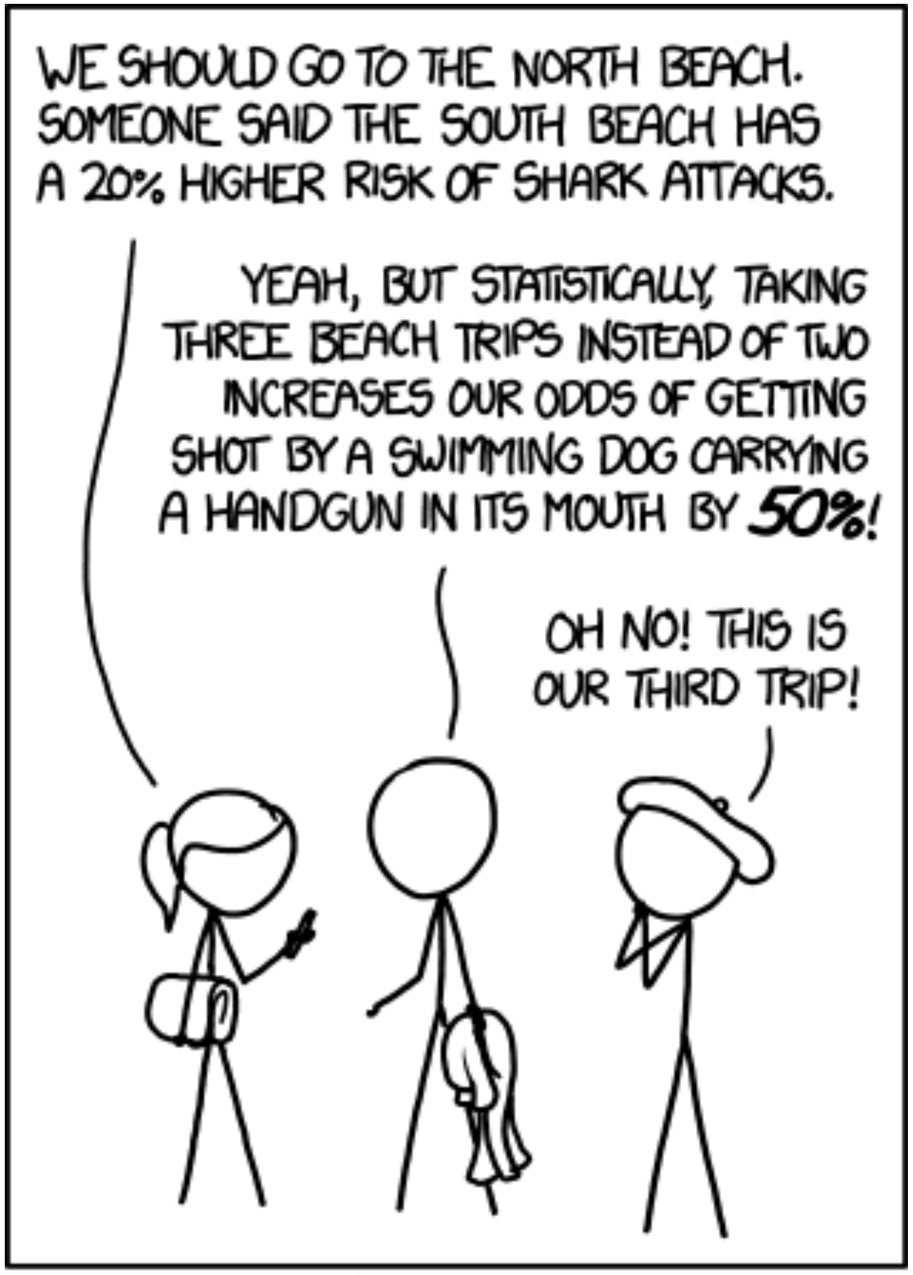
Lecture 11

11.9.21



Refresher Quiz: FDR or BF

A stricter multiple hypothesis correction (will result in fewer significant values)

Bonferroni **FDR**

Balances the number of false positives AND false negatives

Bonferroni **FDR**

Can be calculated by dividing the *p*-value by the number of tests

Bonferroni **FDR**

Sets the significance threshold so that 5% of the significant values are false

Bonferroni **FDR**

Refresher Quiz: FDR or BF

A stricter multiple hypothesis correction (will result in fewer significant values)

Bonferroni **FDR**

Balances the number of false positives AND false negatives

Bonferroni **FDR**

Can be calculated by dividing the *p*-value by the number of tests

Bonferroni **FDR**

Sets the significance threshold so that 5% of the significant values are false

Bonferroni **FDR**

If the code runs in R but is really slow with markdown, you could try:
1. Running the code in R
2. Saving the data to your computer
3. not evaluating the problem chunk in the markdown, but loading your saved data so you can still do the analysis

Example:

```{r, eval = F}

# first, run this chunk of code in R (not by knitting)
# then set eval = F to knit
# load the data in a new chunk (so it IS evaluated)
total <- 0

for(i in 1:100000) {
    total = total + i
}

save(total, file = "~/Desktop/total.Rda")

```

```{r}
load("~/Desktop/total.Rda")

print(total)
```

gene.p.vals

gene_fcs

If you are struggling to knit, at least turn in the .Rmd file with your code and written answers!!!

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

3/10 X 7/9 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 black + 1 white} = $\dfrac{{}_7C_2 \times {}_3C_1}{{}_{10}C_3}$

```
> 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} = $\dfrac{{}_7C_2 \times {}_3C_1}{{}_{10}C_3}$

```
> 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

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(immune | sig) = P(immune | not sig.)

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

80/200 = 0.4 **2920/9800 = 0.3**

```
# 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 EXACT test?

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

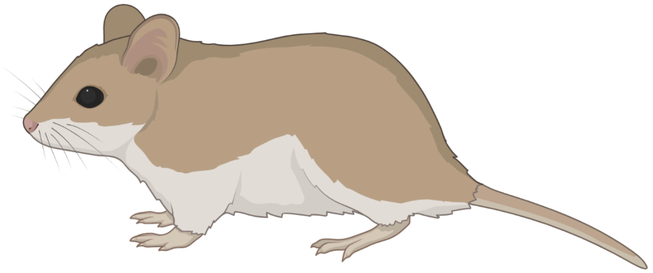
(Similar to randomization)

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()`

You are studying the prevalence of a disease in wild field mice. You catch 30 mice (14 southern grasshopper mice and 16 four-striped grass mice) and check for disease. You find that 5 southern grasshopper and 1 four-striped mice have the disease.

Is there a disease enrichment in the southern grasshopper mice?



Southern grasshopper mouse

Four-striped grass mouse

You are studying the prevalence of a disease in wild field mice. You catch 30 mice (14 southern grasshopper mice and 16 four-striped grass mice) and check for disease. You find that 5 southern grasshopper and 1 four-striped mice have the disease.

Is there a disease enrichment in the southern grasshopper mice?

H_0 : P(disease | southern) = P(disease | striped) H_A : P(disease | southern) > P(disease | striped)

	Southern grasshopper	Striped grass	Total
Disease	5	1	**6**
No disease	9	15	**24**
Total	**14**	**16**	**30**

```
mice <- data.frame(
            southern = c(5, 9),
            striped = c(1, 15)
            )

fisher.test(mice,
      alternative = "greater")
```

p-value = 0.059

You are studying the prevalence of a disease in wild field mice. You catch 30 mice (14 southern grasshopper mice and 16 four-striped grass mice) and check for disease. You find that 5 southern grasshopper and 1 four-striped mice have the disease.

Is there a disease enrichment in the southern grasshopper mice?

P(5+ southern disease) = P(5) + P(6)

	Southern grasshopper	Striped grass	Total
Disease	5	1	**6**
No disease	9	15	**24**
Total	**14**	**16**	**30**

$$P(5) = \frac{{}_{14}C_5 \times {}_{16}C_1}{{}_{30}C_6}$$

```
> dhyper(5, 14, 16, 6)
```

$$P(6) = \frac{{}_{14}C_6 \times {}_{16}C_0}{{}_{30}C_6}$$

```
> 1 - phyper(4, 14, 16, 6)
```

```
> dhyper(6, 14, 16, 6)
```

The Chi-Square Goodness-of-Fit Test

A cross between white and yellow summer squash gave a progeny of the following colors: 155 white, 40 yellow, and 10 green. Are these data consistent with the 12:3:1 ratio predicted by a certain genetic model?

H_0: Pr{white} = 12/16 = 0.75

H_0: Pr{yellow} = 3/16 = 0.1875

H_0: Pr{green} = 1/16 = 0.0625

“Compound hypothesis”

(TWO independent statements)

The Chi-Square Goodness-of-Fit Test

A cross between white and yellow summer squash gave a progeny of the following colors: 155 white, 40 yellow, and 10 green. Are these data consistent with the 12:3:1 ratio predicted by a certain genetic model?

H_0: Pr{white} = 12/16 = 0.75

and

H_0: Pr{yellow} = 3/16 = 0.1875

and

H_0: Pr{green} = 1/16 = 0.0625

H_A: Pr{white} != 12/16 != 0.75

and/or

H_A: Pr{yellow} != 3/16 != 0.1875

and/or

H_A: Pr{green} != 1/16 != 0.0625

The Chi-Square Goodness-of-Fit Test

A cross between white and yellow summer squash gave a progeny of the following colors: 155 white, 40 yellow, and 10 green. Are these data consistent with the 12:3:1 ratio predicted by a certain genetic model?

H_0: Pr{white} = 0.75	**Pr{white} = 155/205 = 0.756**	H_A: Pr{white} != 0.75
and		and/or
H_0: Pr{yellow} = 0.1875	**Pr{yellow} = 40/205 = 0.195**	H_A: Pr{yellow} != 0.1875
and		and/or
H_0: Pr{green} = 0.0625	**Pr{green} = 10/205 = 0.0487**	H_A: Pr{green} != 0.0625

Goodness-of-fit test assesses the compatibility of the data with H_0

Common goodness-of-fit test is the **chi-square test** (χ^2)

The Chi-Square Goodness-of-Fit Test

A cross between white and yellow summer squash gave a progeny of the following colors: 155 white, 40 yellow, and 10 green. Are these data consistent with the 12:3:1 ratio predicted by a certain genetic model?

$$\chi_s^2 = \sum \frac{(o_i - e_i)^2}{e_i}$$

o: observed
e: expected

Goodness-of-fit test assesses the compatibility of the data with H_0

Common goodness-of-fit test is the **chi-square test** (χ^2)

The Chi-Square Goodness-of-Fit Test

A cross between white and yellow summer squash gave a progeny of the following colors: 155 white, 40 yellow, and 10 green. Are these data consistent with the 12:3:1 ratio predicted by a certain genetic model?

$$\chi_s^2 = \sum \frac{(o_i - e_i)^2}{e_i}$$

o: observed
e: expected

	White	Yellow	Green	Total
Observed				
Expected				
Total				

(Absolute, NOT relative frequencies)

The Chi-Square Goodness-of-Fit Test

A cross between white and yellow summer squash gave a progeny of the following colors: 155 white, 40 yellow, and 10 green. Are these data consistent with the 12:3:1 ratio predicted by a certain genetic model?

$$\chi_s^2 = \sum \frac{(o_i - e_i)^2}{e_i}$$

o: observed
e: expected

	White	Yellow	Green	Total
Observed	155	40	10	205
Expected				
Total				

(Absolute, NOT relative frequencies)

The Chi-Square Goodness-of-Fit Test

A cross between white and yellow summer squash gave a progeny of the following colors: 155 white, 40 yellow, and 10 green. Are these data consistent with the 12:3:1 ratio predicted by a certain genetic model?

$$\chi_s^2 = \sum \frac{(o_i - e_i)^2}{e_i}$$

o: observed
e: expected

	White	Yellow	Green	Total
Observed	155	40	10	205
Expected	(12/16)*205	(3/16)*205	(1/16)*205	205
Total				

(Absolute, NOT relative frequencies)

The Chi-Square Goodness-of-Fit Test

A cross between white and yellow summer squash gave a progeny of the following colors: 155 white, 40 yellow, and 10 green. Are these data consistent with the 12:3:1 ratio predicted by a certain genetic model?

$$\chi_s^2 = \sum \frac{(o_i - e_i)^2}{e_i}$$

o: observed
e: expected

	White	Yellow	Green	Total
Observed	155	40	10	205
Expected	153.75	38.4375	12.8125	205
Total				

(Absolute, NOT relative frequencies)

The Chi-Square Goodness-of-Fit Test

A cross between white and yellow summer squash gave a progeny of the following colors: 155 white, 40 yellow, and 10 green. Are these data consistent with the 12:3:1 ratio predicted by a certain genetic model?

$$\chi_s^2 = \sum \frac{(o_i - e_i)^2}{e_i}$$

o: observed
e: expected

(Absolute, NOT relative frequencies)

	White	Yellow	Green	Total
Observed	155	40	10	205
Expected	153.75	38.4375	12.8125	205
O - E	155-153.75	40-38.4375	10-12.8125	
$(O-E)^2$				
$(O-E)^2/E$				

The Chi-Square Goodness-of-Fit Test

A cross between white and yellow summer squash gave a progeny of the following colors: 155 white, 40 yellow, and 10 green. Are these data consistent with the 12:3:1 ratio predicted by a certain genetic model?

$$\chi_s^2 = \sum \frac{(o_i - e_i)^2}{e_i}$$

o: observed
e: expected

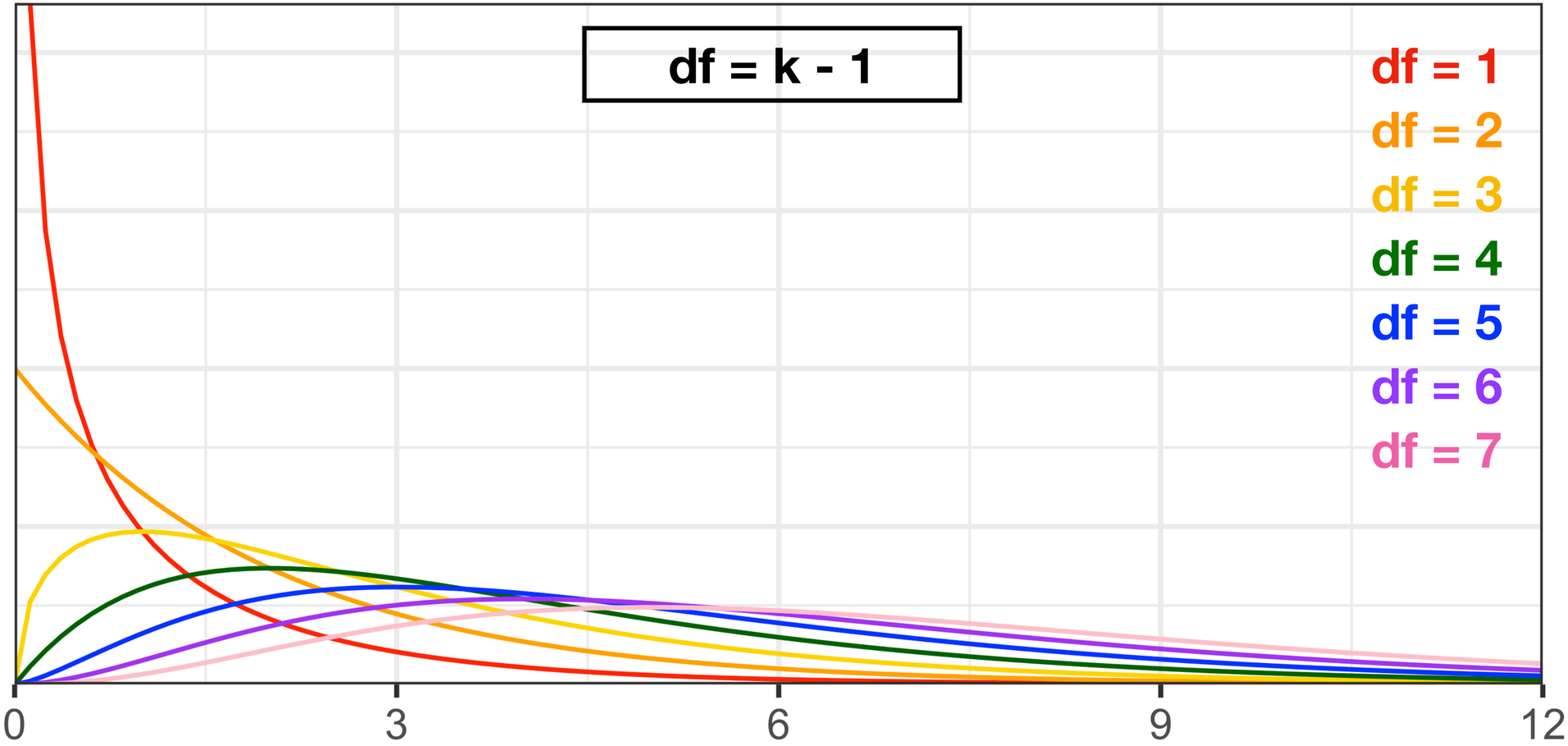
(Absolute, NOT relative frequencies)

	White	Yellow	Green	Total
Observed	155	40	10	205
Expected	153.75	38.4375	12.8125	205
O - E	1.25	1.5625	-2.8125	
$(O-E)^2$	1.5625	2.4412	7.91	
$(O-E)^2/E$	0.01	0.06	0.61	

$$\chi_s^2 = 0.01 + 0.06 + 0.61 = 0.69$$

The Chi-Square Goodness-of-Fit Test

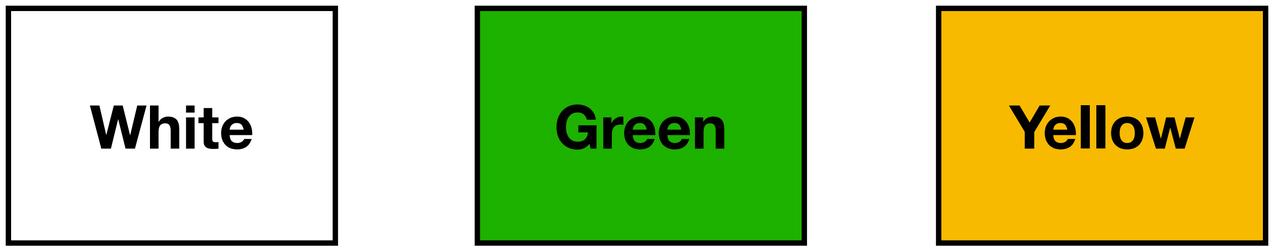
Chi-square distribution: depends on “degrees of freedom”



The Chi-Square Goodness-of-Fit Test

Chi-square distribution: depends on “degrees of freedom”

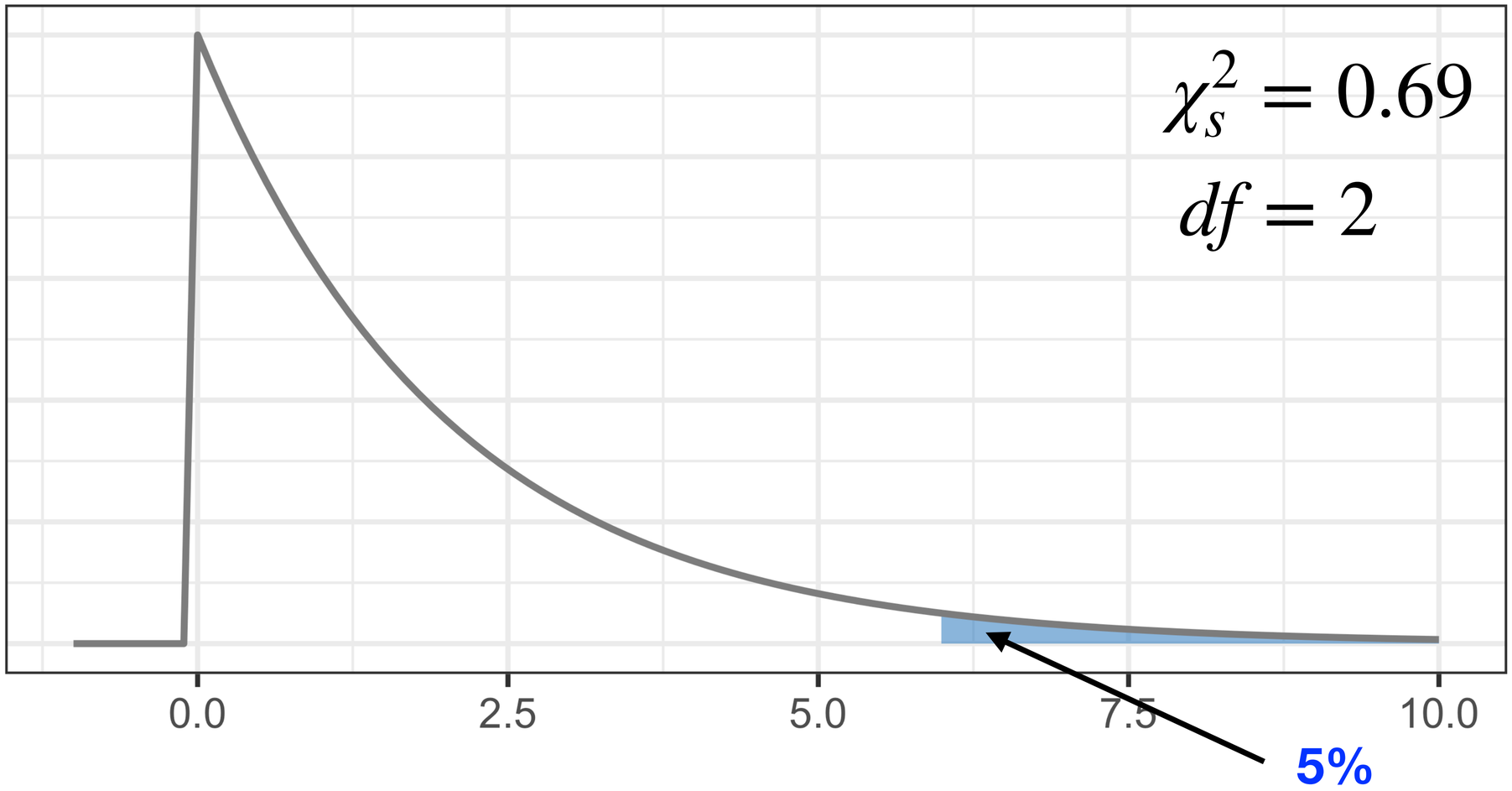
df = k - 1



$$df = 2$$

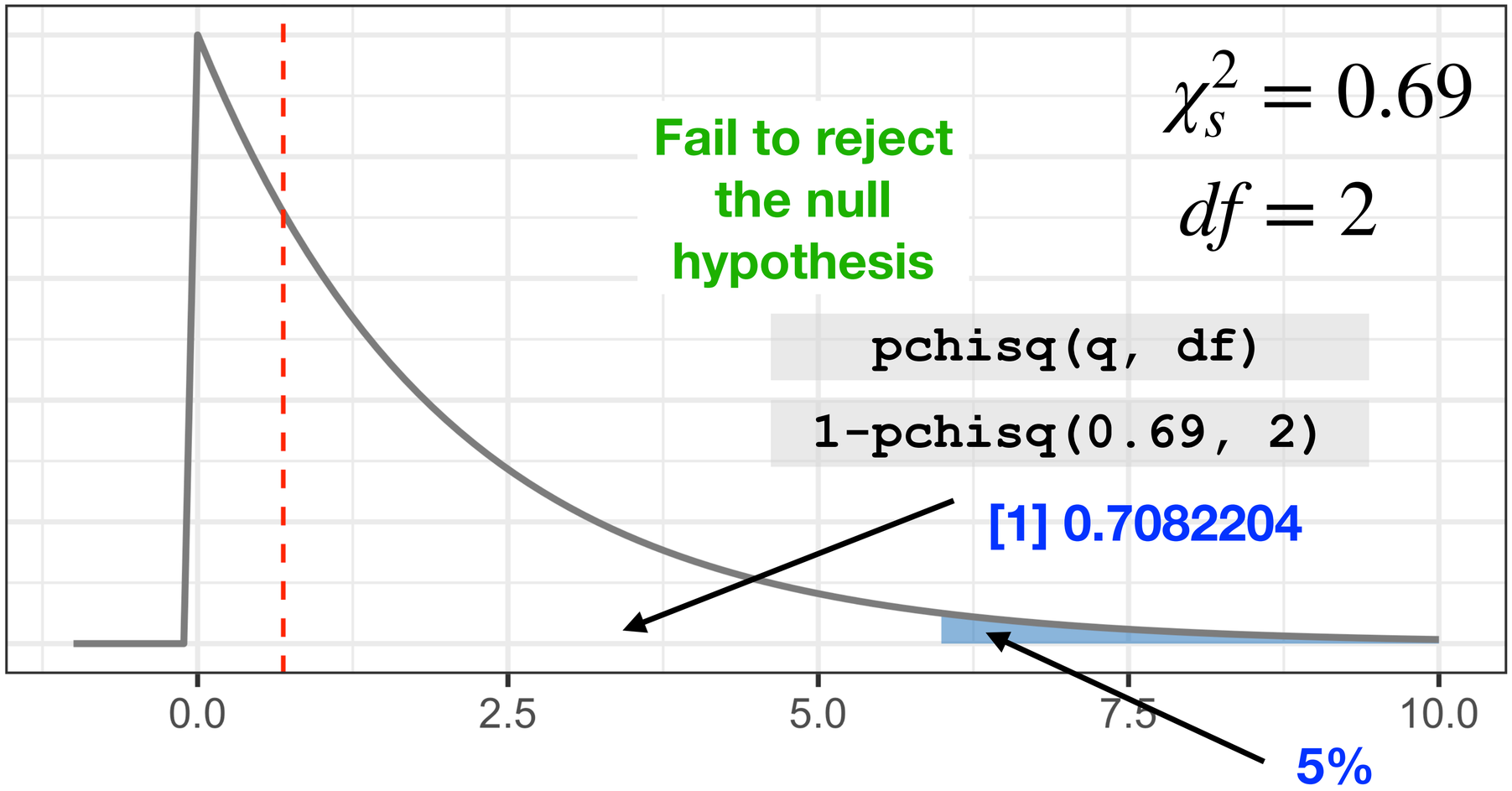
The Chi-Square Goodness-of-Fit Test

Chi-square distribution: depends on “degrees of freedom”



The Chi-Square Goodness-of-Fit Test

Chi-square distribution: depends on “degrees of freedom”



At a Midwestern hospital there were a total of 932 births in 20 consecutive weeks. Of these births, 216 occurred on weekends (i.e. Saturday and Sunday). Do these data reveal more than chance deviation from random timing of the births?

Calculate the χ^2 test statistic.

	Weekday	Weekend
Observed	932-216	216
Expected		
O - E		
$(O-E)^2$		
$(O-E)^2/E$		

At a Midwestern hospital there were a total of 932 births in 20 consecutive weeks. Of these births, 216 occurred on weekends (i.e. Saturday and Sunday). Do these data reveal more than chance deviation from random timing of the births?

Calculate the χ^2 test statistic.

	Weekday	Weekend
Observed	716	216
Expected	(5/7)(932)	(2/7)(932)
O - E		
$(O-E)^2$		
$(O-E)^2/E$		

Expected: each day should have 1/7 of all births

Weekend: (1/7) + (1/7) = 2(1/7)

Weekday: 5(1/7)

At a Midwestern hospital there were a total of 932 births in 20 consecutive weeks. Of these births, 216 occurred on weekends (i.e. Saturday and Sunday). Do these data reveal more than chance deviation from random timing of the births?

Calculate the χ^2 test statistic.

	Weekday	Weekend
Observed	716	216
Expected	665.7	266.3
O - E	50.3	-50.3
$(O-E)^2$	2530.09	2530.09
$(O-E)^2/E$	3.8	9.5

Expected: each day should have 1/7 of all births

Weekend: (1/7) + (1/7) = 2(1/7)

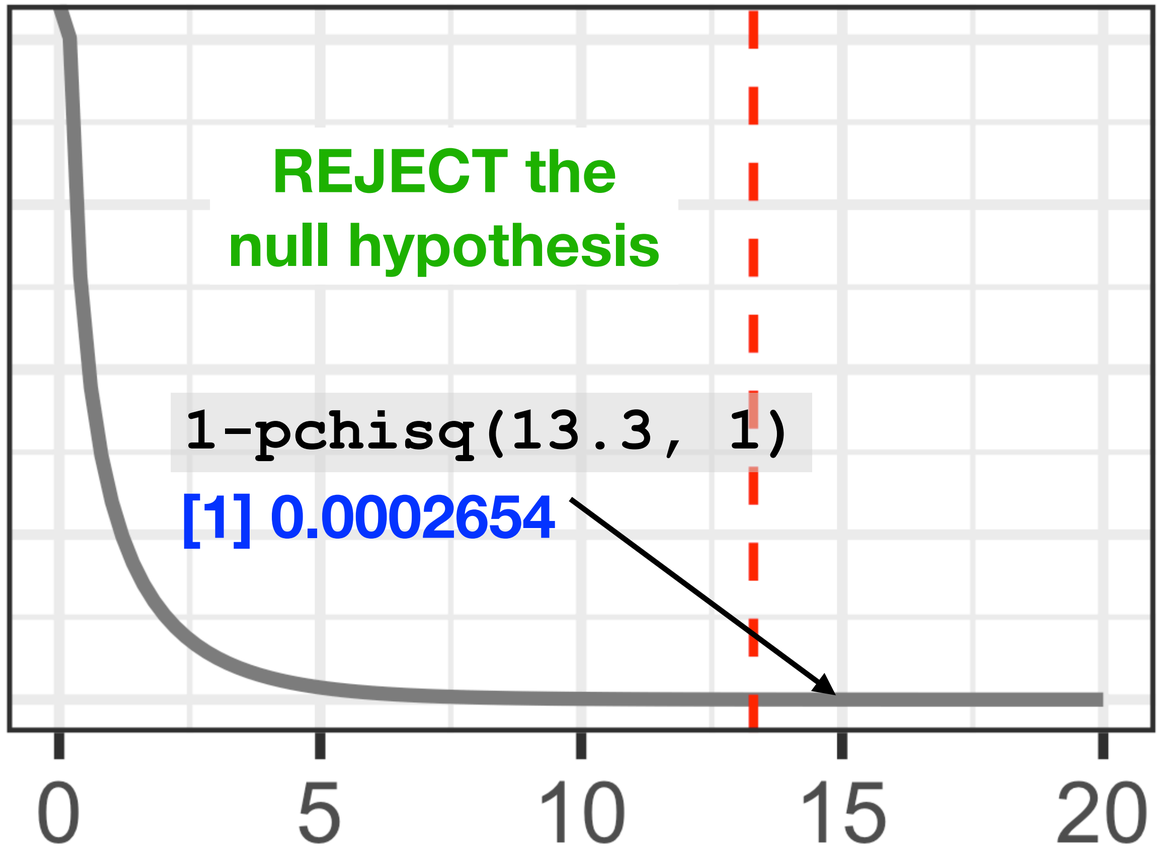
Weekday: 5(1/7)

$$\chi^2 = 3.8 + 9.5 = 13.3$$

$$df = k - 1 = 2 - 1 = 1$$

At a Midwestern hospital there were a total of 932 births in 20 consecutive weeks. Of these births, 216 occurred on weekends (i.e. Saturday and Sunday). Do these data reveal more than chance deviation from random timing of the births?

Calculate the χ^2 test statistic.



Expected: each day should have 1/7 of all births

Weekend: (1/7) + (1/7) = 2(1/7)

Weekday: 5(1/7)

$$\chi^2 = 3.8 + 9.5 = 13.3$$

$$df = k - 1 = 2 - 1 = 1$$

A note on sample size and Chi-Square

- Critical values for chi-square test **do not depend on the sample size** (*degrees of freedom depends on the number of categories, not the number of observations)*
- However, **the test procedure is still affected by sample size, *n***

REJECT the null hypothesis

n = 932	Weekday	Weekend
Observed	716	216
Expected	665.7	266.3
O - E	50.3	-50.3
$(O-E)^2$	2530.09	2530.09
$(O-E)^2/E$	3.8	9.5

$$\chi^2 = 3.8 + 9.5 = 13.3$$

FAIL to reject the null hypothesis

n = 416	Weekday	Weekend
Observed	316.5	95.5
Expected	294.3	117.7
O - E	22.2	-22.2
$(O-E)^2$	492.84	492.84
$(O-E)^2/E$	1.67	4.19

$$\chi^2 = 1.67 + 4.19 = 5.86$$

A note on sample size and Chi-Square

- Critical values for chi-square test **do not depend on the sample size** (*degrees of freedom depends on the number of categories, not the number of observations)*
- However, **the test procedure is still affected by sample size, *n***

REJECT the null hypothesis

n = 932	Weekday	Weekend
Observed	716	216
Expected	665.7	266.3
O - E	50.3	-50.3
$(O-E)^2$	2530.09	2530.09
$(O-E)^2/E$	3.8	9.5

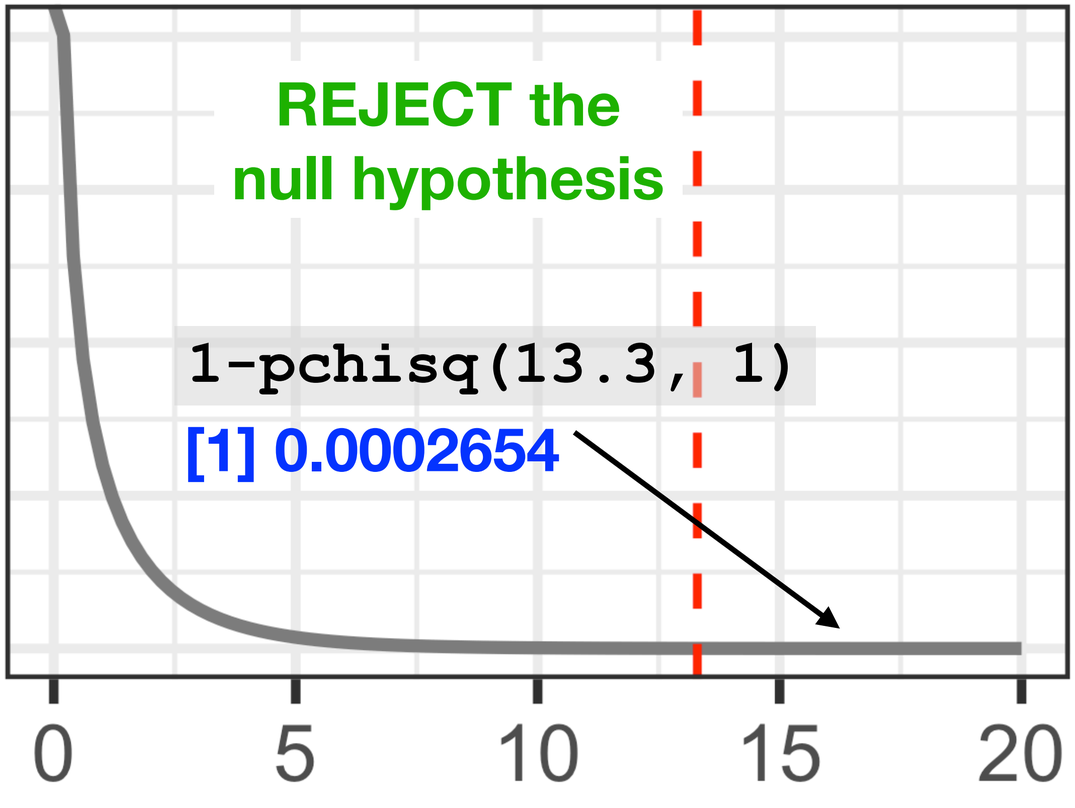
FAIL to reject the null hypothesis

n = 416	Weekday	Weekend
Observed	**316.5**	**95.5**
Expected	294.3	117.7
O - E	22.2	-22.2
$(O-E)^2$	492.84	492.84
$(O-E)^2/E$	1.67	4.19

Increased sample size magnifies any discrepancy between observed and expected

Directional Chi-Square test

At a Midwestern hospital there were a total of 932 births in 20 consecutive weeks. Of these births, 216 occurred on weekends (i.e. Saturday and Sunday). Do these data reveal more than chance deviation from random timing of the births?



H_0 : Weekday births = Weekend births

H_A : Weekday births ≠ Weekend births

Even though the χ^2 distribution only evaluates the upper tail area, it is still considered a two-tailed (non-directional) test

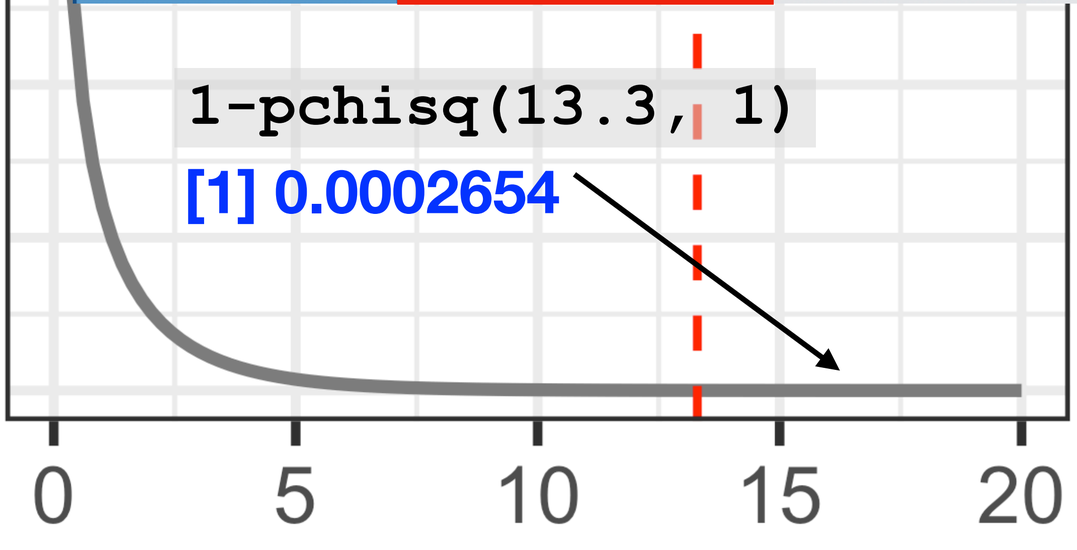
H_0 : **Weekday births = (5/7)(total births)**

H_A : **Weekday births > (5/7)(total births)**

Directional Chi-Square test

At a Midwestern hospital there were a total of 932 births in 20 consecutive weeks. Of these births, 216 occurred on weekends (i.e. Saturday and Sunday). Do these data reveal more than chance deviation from random timing of the births?

	Weekday	Weekend
Observed	716	216
Expected	665.7	266.3



H_0 : Weekday births = (5/7)(total births)

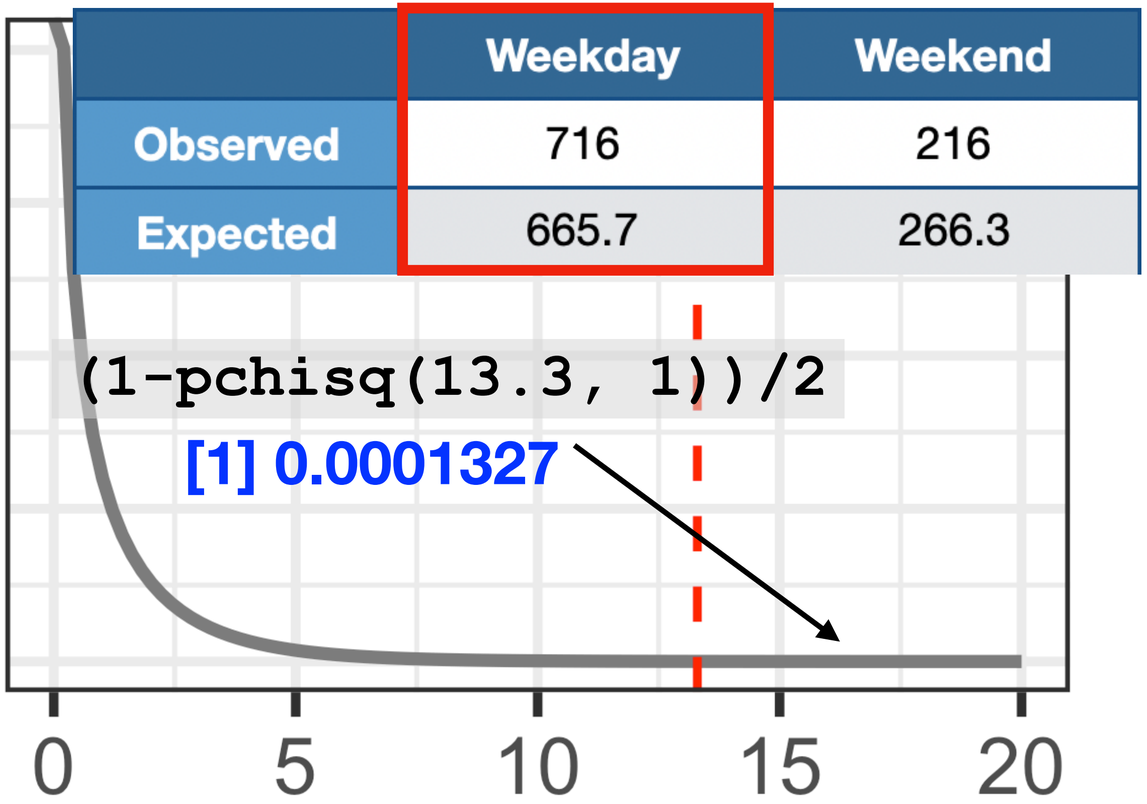
H_A : Weekday births > (5/7)(total births)

1. **Check directionality**
   1. **If data deviates from H_0 in the wrong direction, *p*-value = 0.5**
   2. **If data deviates from H_0 in the right direction, *p*-value is 1/2 of the non-directional *p*-value**

Directional Chi-Square test

At a Midwestern hospital there were a total of 932 births in 20 consecutive weeks. Of these births, 216 occurred on weekends (i.e. Saturday and Sunday). Do these data reveal more than chance deviation from random timing of the births?

	Weekday	Weekend
Observed	716	216
Expected	665.7	266.3



H_0 : Weekday births = (5/7)(total births)

H_A : Weekday births > (5/7)(total births)

1. **Check directionality**
   1. **If data deviates from H_0 in the wrong direction, *p*-value = 0.5**
   2. **If data deviates from H_0 in the right direction, *p*-value is 1/2 of the non-directional *p*-value**

Directional Chi-Square test

	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Observed	94	115	101	81	130	109	70
Expected	100	100	100	100	100	100	100

H_0 : Equal number of births on each day of the week

H_0 : Mon. = Tues H_0 : Tues. = Wed. H_0 : Tues. = Thurs.

Compound null hypothesis

H_A : Non-equal number of births on each day of the week

Compound hypotheses are non-directional by necessity.

Introduction to categorical relationships

“Contingency table” : assesses association between row and column variable

	Real surgery	Sham surgery	Total
Success	41	15	**56**
No success	8	11	**19**
Total	**49**	**26**	**75**

Conditional probability

41/49 = 83.7%
Real surgeries successful

15/26 = 57.7%
Sham surgeries successful

Introduction to categorical relationships

“Contingency table” : assesses association between row and column variable

	Real surgery	Sham surgery	Total
Success	41	15	**56**
No success	8	11	**19**
Total	**49**	**26**	**75**

Conditional probability

41/49 = 83.7%
Pr{Success | Real}

15/26 = 57.7%
Pr{Success | Sham}

A randomization test for categorical data

“Contingency table” : assesses association between row and column variable

	Real surgery	Sham surgery	Total
Success	~~41~~ 42	~~15~~ 14	**56**
No success	~~8~~ 7	~~11~~ 12	**19**
Total	**49**	**26**	**75**

Conditional probability

42/49 = 85.7%
Pr{Success | Real}

14/26 = 53.8%
Pr{Success | Sham}

A randomization test for categorical data

“Contingency table” : assesses association between row and column variable

	Real surgery	Sham surgery	Total
Success	~~41~~ 37	~~15~~ 19	**56**
No success	~~8~~ 12	~~11~~ 7	**19**
Total	**49**	**26**	**75**

Conditional probability

37/49 = 75.5%
Pr{Success | Real}

19/26 = 73.1%
Pr{Success | Sham}

Randomize 1000x…

The Chi-Square Test for contingency tables

$$\chi_s^2 = \sum \frac{(o_i - e_i)^2}{e_i}$$

o: observed
e: expected

H_0: Pr{Success | Real} = Pr{Success | Sham}

H_A: Pr{Success | Real} > Pr{Success | Sham}

	Real surgery	Sham surgery	Total
Success	41	15	**56**
No success	8	11	**19**
Total	**49**	**26**	**75**

Null hypothesis:

41/49 = 83.7%
Pr{Success|Real}

?
=

15/26 = 57.7%
Pr{Success|Sham}

The Chi-Square Test for contingency tables

$$\chi_s^2 = \sum \frac{(o_i - e_i)^2}{e_i}$$

o: observed; e: expected

	Real surgery	Sham surgery	Total
Success	41	15	**56**
No success	8	11	**19**
Total	**49**	**26**	75

1. Calculate the expected values using marginal frequencies

Real: (56/75)*49
= 36.59 successful outcomes expected

Sham: (56/75)*26
= 19.41 successful outcomes expected

$$e = \frac{(row_T)(column_T)}{Grand_T}$$

Null hypothesis:

41/49 = 83.7%
Pr{Success|Real}

?
=

15/26 = 57.7%
Pr{Success|Sham}

The Chi-Square Test for contingency tables

$$\chi_s^2 = \sum \frac{(o_i - e_i)^2}{e_i} = \frac{(41 - 36.59)^2}{36.59} + \frac{(15 - 19.41)^2}{19.41} + \frac{(8 - 12.41)^2}{12.41} + \frac{(11 - 6.59)^2}{6.59} = 6.06$$

df = 1

	Real surgery	Sham surgery	Total
Success	41 (36.59)	15 (19.41)	**56**
No success	8 (12.41)	11 (6.59)	**19**
Total	**49**	**26**	**75**

O (E)

Null hypothesis:

41/49 = 83.7% Pr{Success|Real} **?** **=** 15/26 = 57.7% Pr{Success|Sham}

The Chi-Square Test for contingency tables

$$\chi_s^2 = \sum \frac{(o_i - e_i)^2}{e_i} = \frac{(41 - 36.59)^2}{36.59} + \frac{(15 - 19.41)^2}{19.41} + \frac{(8 - 12.41)^2}{12.41} + \frac{(11 - 6.59)^2}{6.59} = 6.06$$

df = 1

	Real surgery	Sham surgery	Total
Success	49-19-10 ✗	26-10 ✗	56
No success	19-10 ✗	10 ✓	19
Total	49	26	75

Null hypothesis: 41/49 = 83.7% Pr{Success|Real} ? = 15/26 = 57.7% Pr{Success|Sham}

The Chi-Square Test for contingency tables

$$\chi_s^2 = \sum \frac{(o_i - e_i)^2}{e_i} = \frac{(41 - 36.59)^2}{36.59} + \frac{(15 - 19.41)^2}{19.41} + \frac{(8 - 12.41)^2}{12.41} + \frac{(11 - 6.59)^2}{6.59} = 6.06$$

df = 1

	Real surgery	Sham surgery	Total
Success	49-19 ✗	26-19-19 ✗	56
No success	19 ✓	19-19 ✗	19
Total	49	26	75

Null hypothesis: 41/49 = 83.7% Pr{Success|Real} ? = 15/26 = 57.7% Pr{Success|Sham}

The Chi-Square Test for contingency tables

$$\chi_s^2 = \sum \frac{(o_i - e_i)^2}{e_i} = \frac{(41 - 36.59)^2}{36.59} + \frac{(15 - 19.41)^2}{19.41} + \frac{(8 - 12.41)^2}{12.41} + \frac{(11 - 6.59)^2}{6.59} = 6.06$$

df = 1

	Real surgery	Sham surgery	Total
Success	41 (36.59)	15 (19.41)	**56**
No success	8 (12.41)	11 (6.59)	**19**
Total	**49**	**26**	**75**

```
> 1 - pchisq(6.06, 1)
```

```
> 0.0138
```

Because HA is directional…

```
> 0.0138 / 2 = 0.0069
```

REJECT null hypothesis

Null hypothesis: 41/49 = 83.7% Pr{Success|Real} **?=** 15/26 = 57.7% Pr{Success|Sham}

Chi-Square (χ^2) Distribution

Degrees of Freedom	Area to the Right of Critical Value									
	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	—	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.299
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928

The Chi-Square Test for contingency tables

$$\chi_s^2 = \sum \frac{(o_i - e_i)^2}{e_i} = \frac{(35 - 36.59)^2}{36.59} + \frac{(19 - 19.41)^2}{19.41} + \frac{(14 - 12.41)^2}{12.41} + \frac{(5 - 6.59)^2}{6.59} = 0.501$$

df = 1

	Real surgery	Sham surgery	Total
Success	35 (36.59)	19 (19.41)	**56**
No success	14 (12.41)	5 (6.59)	**19**
Total	**49**	**26**	**75**

```
> 1 - pchisq(0.501, 1)
```

```
> 0.479
```

```
> chisq.test(data,
             correct = F)
```

Null hypothesis:

35/49 = 71.4% Pr{Success|Real} **?** **=** 19/26 = 73.0% Pr{Success|Sham}

The Chi-Square Test for contingency tables

- Note: although the formula for χ^2 contingency tables is the **same** as for goodness-of-fit tests, the method of calculating **e's** is quite different because **the null hypothesis is different**
- Degrees of freedom in a 2x2 contingency table **is always one**
- χ^2 measures discrepancy between data and null **indirectly**
    - If sample conditional probabilities **are equal,** χ^2 **= 0**
    - χ^2 also depends on **directly** on the **sample size**: a given percentage deviation from H_0 is **less likely to occur by chance with large sample size**

A study was conducted to observe the effects of prolonged use of cell phones. A group of patients with brain tumors were asked whether they routinely help the cell phone to the right or left ear. The 88 responses are shown below. Do the data provide sufficient evidence to conclude that use of cell phones leads to an increase in brain tumors on that side of the head?

	Phone - Left	Phone - Right
Tumor - left	14	28
Tumor - right	19	27

1. State the null hypothesis in words

2. State the alternative hypothesis in words

3. Compute the expected values and test statistic

A study was conducted to observe the effects of prolonged use of cell phones. A group of patients with brain tumors were asked whether they routinely help the cell phone to the right or left ear. The 88 responses are shown below. Do the data provide sufficient evidence to conclude that use of cell phones leads to an increase in brain tumors on that side of the head?

	Phone - Left	Phone - Right
Tumor - left	14	28
Tumor - right	19	27

1. State the null hypothesis in words

> The side you hold your phone does not affect the side of brain tumor

2. State the alternative hypothesis in words

> The side you hold your phone DOES not affect the side of brain tumor — Left = Left

3. Compute the expected values and test statistic

A study was conducted to observe the effects of prolonged use of cell phones. A group of patients with brain tumors were asked whether they routinely help the cell phone to the right or left ear. The 88 responses are shown below. Do the data provide sufficient evidence to conclude that use of cell phones leads to an increase in brain tumors on that side of the head?

	Phone - Left	Phone - Right	Total
Tumor - left	14	28	**42**
Tumor - right	19	27	**46**
Total	**33**	**55**	**88**

$$e = \frac{(row_T)(column_T)}{Grand_T} = \frac{(42)(33)}{88} = 15.75$$

A study was conducted to observe the effects of prolonged use of cell phones. A group of patients with brain tumors were asked whether they routinely help the cell phone to the right or left ear. The 88 responses are shown below. Do the data provide sufficient evidence to conclude that use of cell phones leads to an increase in brain tumors on that side of the head?

	Phone - Left	Phone - Right	Total
Tumor - left	14 (15.75)	28 (26.25)	**42**
Tumor - right	19 (17.25)	27 (28.75)	**46**
Total	**33**	**55**	**88**

$$e = \frac{(row_T)(column_T)}{Grand_T} = \frac{(42)(33)}{88} = 15.75$$

A study was conducted to observe the effects of prolonged use of cell phones. A group of patients with brain tumors were asked whether they routinely help the cell phone to the right or left ear. The 88 responses are shown below. Do the data provide sufficient evidence to conclude that use of cell phones leads to an increase in brain tumors on that side of the head?

	Phone - Left	Phone - Right	Total
Tumor - left	14 (15.75)	28 (26.25)	**42**
Tumor - right	19 (17.25)	27 (28.75)	**46**
Total	**33**	**55**	**88**

$$\chi_s^2 = \sum \frac{(o_i - e_i)^2}{e_i} = \frac{(14 - 15.75)^2}{15.75} + \frac{(28 - 26.25)^2}{26.25} + \frac{(19 - 17.25)^2}{17.25} + \frac{(27 - 28.75)^2}{28.75} = 0.595$$

A study was conducted to observe the effects of prolonged use of cell phones. A group of patients with brain tumors were asked whether they routinely help the cell phone to the right or left ear. The 88 responses are shown below. Do the data provide sufficient evidence to conclude that use of cell phones leads to an increase in brain tumors on that side of the head?

df = 1

Directional!

Left tumors from left phones > left tumors from right phones

	Phone - Left	Phone - Right	Total
Tumor - left	14 (15.75)	28 (26.25)	**42**
Tumor - right	19 (17.25)	27 (28.75)	**46**
Total	**33**	**55**	**88**

$$\chi_s^2 = \sum \frac{(o_i - e_i)^2}{e_i} = \frac{(14-15.75)^2}{15.75} + \frac{(28-26.25)^2}{26.25} + \frac{(19-17.25)^2}{17.25} + \frac{(27-28.75)^2}{28.75} = 0.595$$

```
> pchisq(0.595, 1, lower.tail = F)/2
```

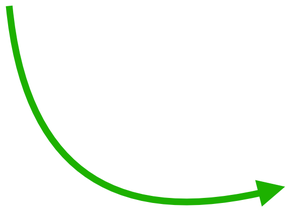
[1] 0.2202461

Fail to reject

Two contexts for contingency tables

(Migraine example)

- **Two independent samples** with **a dichotomous** observed variable
- **One sample** with **two dichotomous** observed variables



	Dark hair	Light hair	Total
Dark eyes	726	131	**857**
Light eyes	3,129	2,814	**5,943**
Total	**3,855**	**2,945**	**6,800**

- Sample of 6,800 people
- Some had dark hair, some had light hair

Math is same in both contexts, but conclusions and interpretation of hypothesis can be different

Dichotomous = two possible values

Independence and association

Responses

	Real surgery	Sham surgery	Total
Success	41 (36.59)	15 (19.41)	**56**
No success	8 (12.41)	11 (6.59)	**19**
Total	**49**	**26**	**75**

Treatments

Variable 2

	Dark hair	Light hair	Total
Dark eyes	726	131	**857**
Light eyes	3,129	2,814	**5,943**
Total	**3,855**	**2,945**	**6,800**

Variable 1

Pr{Success | Real} ?= Pr{Success | Sham}

Pr{Real | Success} ?= Pr{Sham | Success}

Pr{D.eyes | D.hair} ?= Pr{D.eyes | L.hair}

Pr{D.hair| D.eyes} ?= Pr{L.hair | D.eyes}

Chi-square “test of independence”

When the data is viewed as a single sample with two observed variables, the relationship expressed by H_0 is called **statistical independence**

Variable 2

	Dark hair	Light hair	Total
Dark eyes	726	131	**857**
Light eyes	3,129	2,814	**5,943**
Total	**3,855**	**2,945**	**6,800**

Variable 1

H_0: Hair color and eye color are independent.

These hypotheses are the same

Pr{Dark hair | Dark eyes} ?= Pr{Dark hair | light eyes}

Pr{Dark eyes | Dark hair} ?= Pr{Dark eyes | light hair}

Language of “association”

“… you should say what you mean,” the March hare went on.

“I do,” Alice hastily replied; “at least - at least I mean what I say - that’s the same thing you know.”

“Not the same thing a bit!” Said the Hatter. “Why, you might just as well say that ‘I see what I eat’ is the same thing as ‘I eat what I see’!”

Alice in Wonderland (Lewis Carroll)

Language of “association”

	Dark hair	Light hair	Total
Dark eyes	726	131	**857**
Light eyes	3,129	2,814	**5,943**
Total	**3,855**	**2,945**	**6,800**

$$\chi^2 = 314; p \approx 0$$

P(Dark eyes | Dark hair) > Pr(Dark eyes | Light hair)

There is sufficient evidence to conclude that dark-haired men have a greater tendency to be dark-eyed than do light-haired men.

Language of “association”

	Dark hair	Light hair	Total
Dark eyes	726	131	**857**
Light eyes	3,129	2,814	**5,943**
Total	**3,855**	**2,945**	**6,800**

$$\chi^2 = 314; p \approx 0$$

P(Dark eyes | Dark hair) > Pr(Dark eyes | Light hair)

There is sufficient evidence to conclude that dark-haired men have a greater tendency to be dark-eyed than do light-haired men.

than to be light-eyed.

Can be misleading…

P(Dark eyes | Dark hair) > Pr(Light eyes | Dark hair)

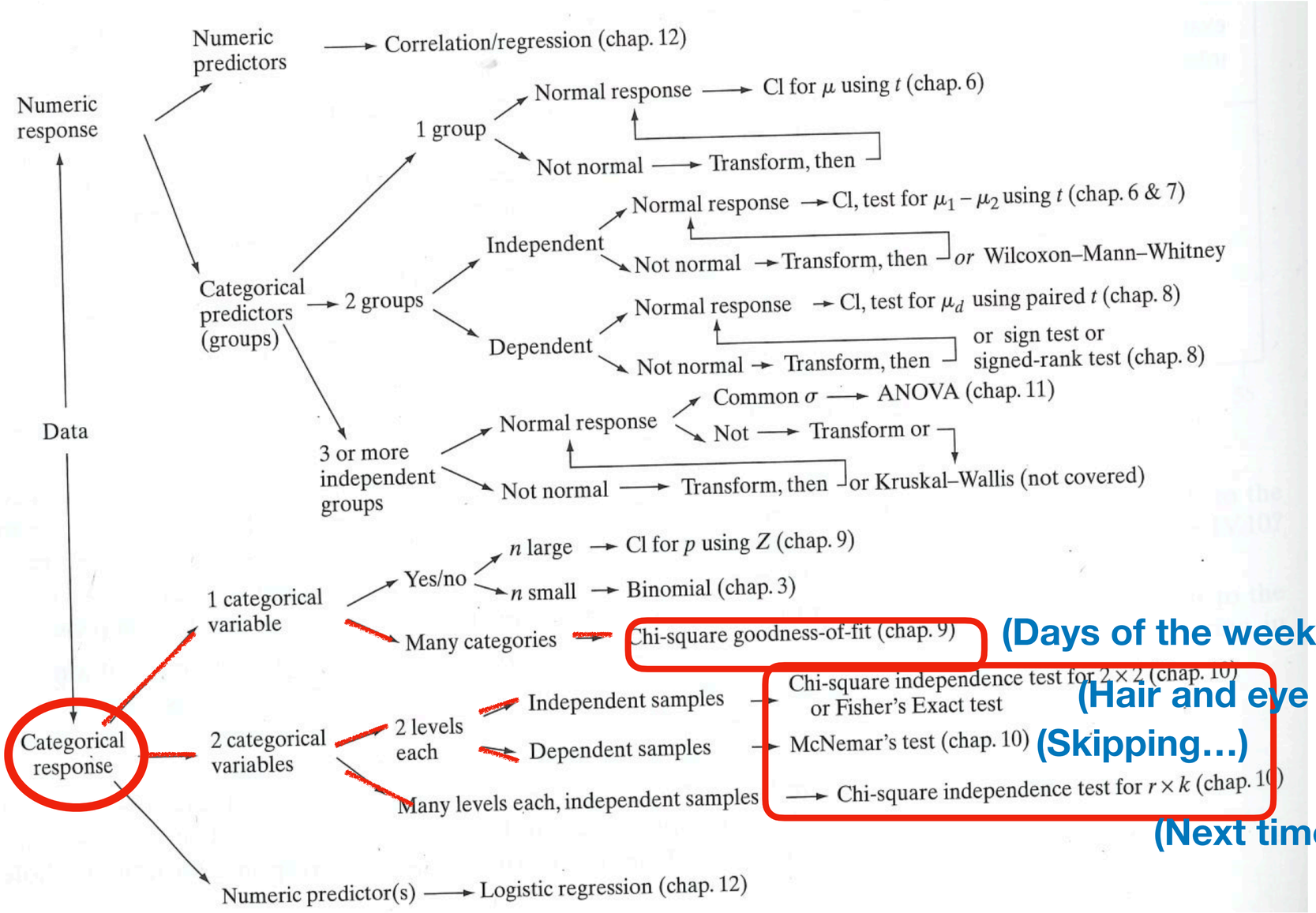
There is sufficient evidence to conclude that dark hair is associated with dark eyes

There is sufficient evidence to conclude that most dark-haired men are dark-eyed.

Conditions for chi-square test

- **Design conditions**: Data must be either (1) two or more independent random samples observed with respect to a categorical variable OR (2) one random sample observed with respect to two categorical variables
- **Sample size conditions**: “large enough” sample size. Rule of thumb — each expected frequency (e) must be at least 5
- **Form of H_0**: the row variable and the columns variable are independent
- **Scope of inference**: if data arise from experiments we can draw a causal inference; if the data arise from observational study we can only infer that the observed association is not due to chance but cannot rule out other explanations

When to use Chi-square test



(Days of the week)

(Hair and eye color)

(Skipping…)

(Next time)

Announcements

- **Thursday's class will be in Daniel Hale Williams, McGaw 2-320**
- Midterm regrades due by **Thursday after class** (key will be posted tonight)
- Final project proposal due **Friday @ 5pm**