

Module 4A : Hypothesis Testing

Applied Epistemology: A Framework for how we know things scientifically

Agenda:

- Overflow lecture (there were serious concepts in 3A/3B)
- Working through examples of hypothesis testing
 - Binomial Example
 - χ^2 Goodness of fit tests

Does wearing a red shirt help win during a wrestling match?



16 out of **20 rounds** had more red-shirted than blue-shirted winners in the 2004 Olympics in wrestling, taekwondo and boxing.

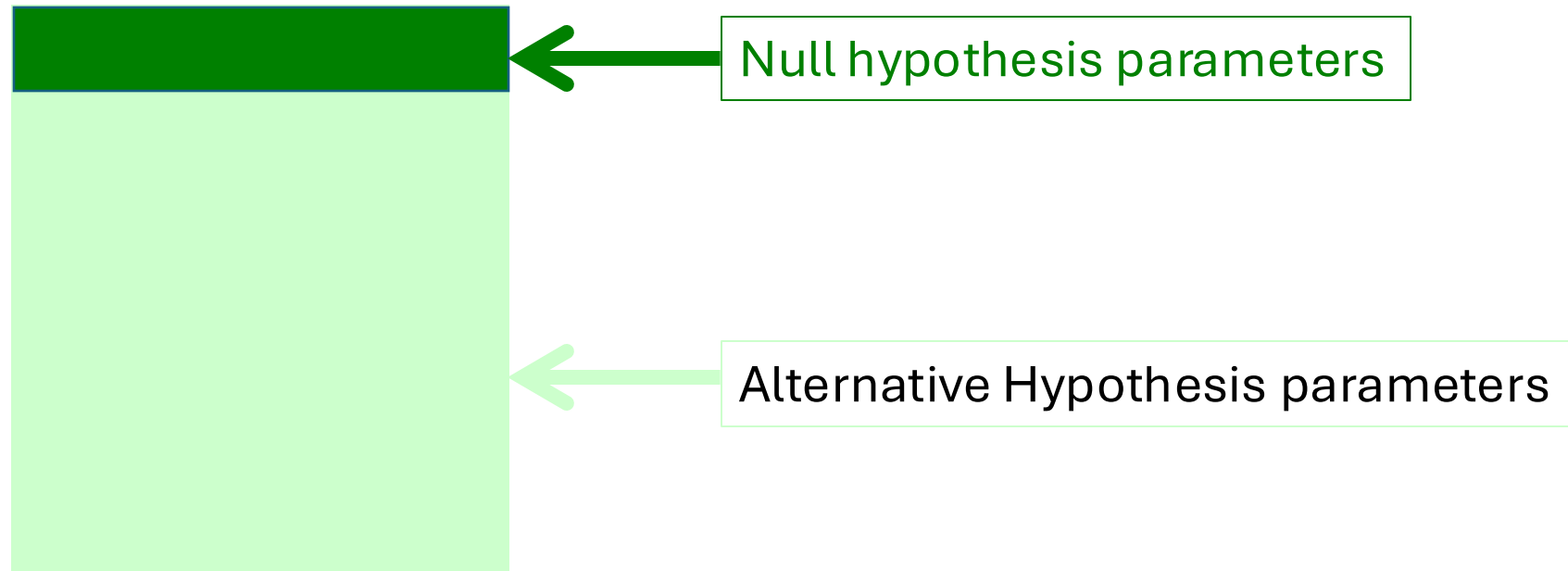
Does wearing a red shirt help win in combat sports?

Step 1: Formulate Hypothesis

Four steps in hypothesis testing:

1. Formulate Hypothesis

- o Most of the mental effort
- o Quantifies how unusual data is *if you assume that the null hypothesis is true*
- o H_0 and H_A - mutually exclusive

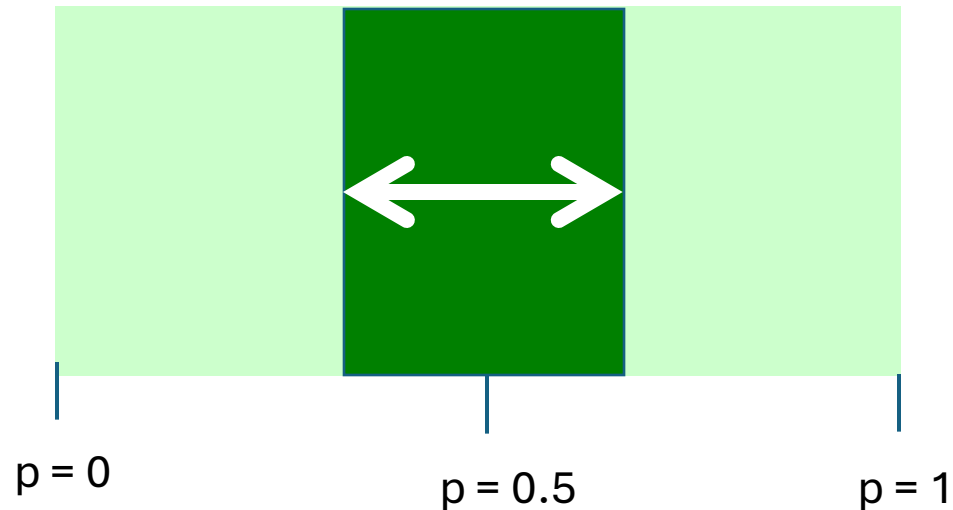


Does wearing a red shirt help win in combat sports?

Step 1: Formulate Hypothesis

H_0 : Red and blue shirted athletes are equally likely to win
(proportion = 0.5)

H_A : Red and blue shirted athletes are not equally likely to win
(proportion $\neq 0.5$)



Does wearing a red shirt help win in combat sports?

Step 1: Formulate Hypothesis

H_0 : Red and blue shirted athletes are equally likely to win (proportion = 0.5)

H_A : Red and blue shirted athletes are not equally likely to win (proportion $\neq 0.5$)

Step 2: Identify test statistic

Does wearing a red shirt help win in combat sports?

Step 1: Formulate Hypothesis

H_0 : Red and blue shirted athletes are equally likely to win (proportion = 0.5)

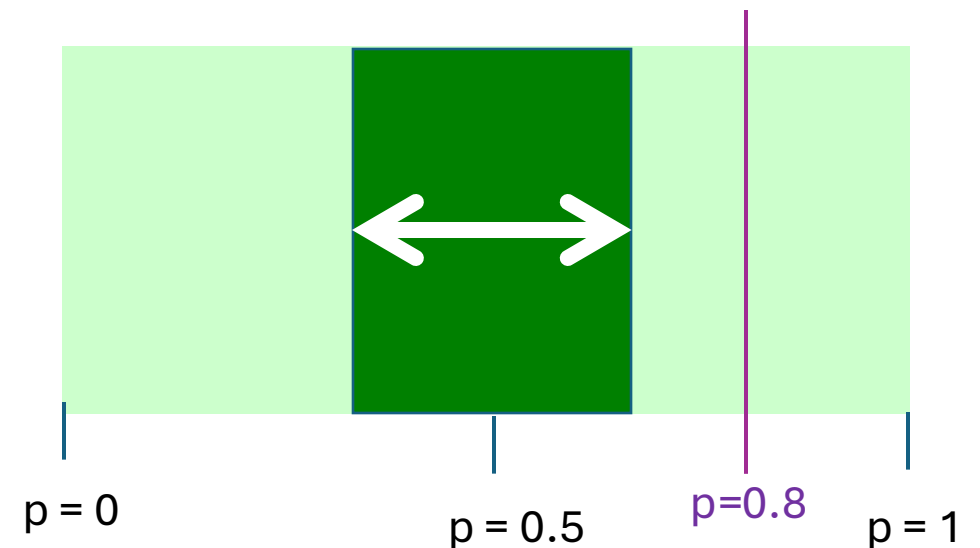
H_A : Red and blue shirted athletes are not equally likely to win (proportion $\neq 0.5$)

Step 2: Identify test statistic

16 out of 20 red shirted winners

--> proportion = **0.8**

This is a discrepancy of **0.3** from H_0 . Can it be due to chance alone?



Does wearing a red shirt help win in combat sports?

Step 1: Formulate Hypothesis

H_0 : Red and blue shirted athletes are equally likely to win (proportion = 0.5)

H_A : Red and blue shirted athletes are not equally likely to win (proportion \neq 0.5)

Step 2: Identify test statistic

16 out of 20 red shirted winners --> proportion = 0.8

Step 3: Calculate the P-Value/Compare to critical values or fixed Significance

Does wearing a red shirt help win in combat sports?

Step 1: Formulate Hypothesis

H_0 : Red and blue shirted athletes are equally likely to win (proportion = 0.5)

H_A : Red and blue shirted athletes are not equally likely to win (proportion \neq 0.5)

Step 2: Identify test statistic

16 out of 20 red shirted winners --> proportion = 0.8

Step 3: Calculate the P-Value/Compare to critical values or fixed significance

If H_0 is true, what is the chance of observing a test statistic with a value at least as extreme as the one we have observed? <-p-value

Does wearing a red shirt help win in combat sports?

Step 1: Formulate Hypothesis

H_0 : Red and blue shirted athletes are equally likely to win (proportion = 0.5)

H_A : Red and blue shirted athletes are not equally likely to win (proportion $\neq 0.5$)

Step 2: Identify test statistic

16 out of 20 red shirted winners --> proportion = 0.8

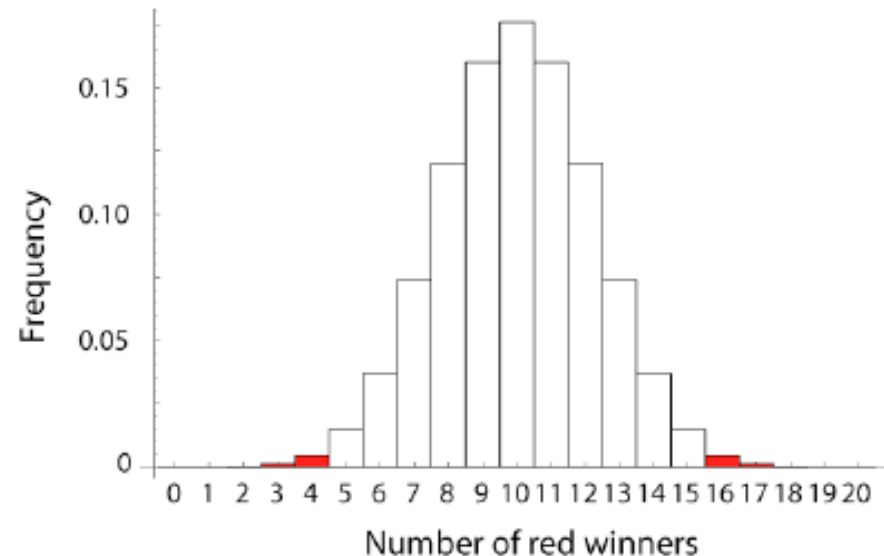
Step 3: Calculate the P-Value/Compare to critical values or fixed significance

Null Distribution of the sample proportion

The Binomial Distribution

explains this type of proportion data

If H_0 is true, what is the chance of observing a test statistic value **at least as extreme** as the one we have observed?



Does wearing a red shirt help win in combat sports?

Step 1: Formulate Hypothesis

H_0 : Red and blue shirted athletes are equally likely to win (proportion = 0.5)

H_A : Red and blue shirted athletes are not equally likely to win (proportion $\neq 0.5$)

Step 2: Identify test statistic

16 out of 20 red shirted winners --> proportion = 0.8

Step 3: Calculate the P-Value/Compare to critical values or fixed significance

If H_0 is true, what is the chance of observing a test statistic value at least as extreme as the one we have observed?

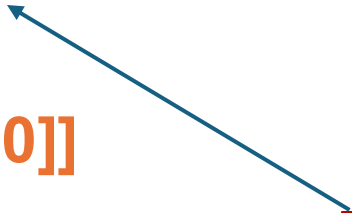
The P-value from the null distribution of the proportion is calculated as:

$$P = [P[0]+P[1]+P[2]+P[3]+P[4]+P[16]+P[17]+P[18]+P[19]+P[20]]$$

= due to symmetry

$$= 2 \times P[16] + P[17] + P[18] + P[19] + P[20]$$

$$= 0.012$$


$$P\left(\frac{20}{16}\right) = \frac{20!}{16!4!} 0.5^{16} (1 - 0.5)^4$$

Does wearing a red shirt help win in combat sports?

Step 1: Formulate Hypothesis

H_0 : Red and blue shirted athletes are equally likely to win (proportion = 0.5)

H_A : Red and blue shirted athletes are not equally likely to win (proportion $\neq 0.5$)

Step 2: Identify test statistic

16 out of 20 red shirted winners --> proportion = 0.8

Step 3: Calculate the P-Value/Compare to critical values or fixed significance

$$P = 2 \times [P[16] + P[17] + P[18] + P[19] + P[20]] = 0.012$$

Does wearing a red shirt help win in combat sports?

Step 1: Formulate Hypothesis

H_0 : Red and blue shirted athletes are equally likely to win (proportion = 0.5)

H_A : Red and blue shirted athletes are not equally likely to win (proportion $\neq 0.5$)

Step 2: Identify test statistic

16 out of 20 red shirted winners --> proportion = 0.8

Step 3: Calculate the P-Value/Compare to critical values or fixed significance

$$P = 2 \times [P[16] + P[17] + P[18] + P[19] + P[20]] = 0.012$$

What is alpha?

$$\alpha = 0.05 \quad \text{and} \quad P\text{-value} = 0.012$$

$$P < \alpha \quad \text{so we can reject } H_0$$

Does wearing a red shirt help win in combat sports?

Step 1: Formulate Hypothesis

H_0 : Red and blue shirted athletes are equally likely to win (proportion = 0.5)

H_A : Red and blue shirted athletes are not equally likely to win (proportion $\neq 0.5$)

Step 2: Identify test statistic

16 out of 20 red shirted winners --> proportion = 0.8

Step 3(a): Calculate the P-Value

$$P = 2 \times [P[16] + P[17] + P[18] + P[19] + P[20]] = 0.012$$

Step 3(b): Compare to a fixed significance

$$\alpha = 0.05 \text{ and } P\text{-value} = 0.012$$

$P < \alpha$ so we can reject H_0

Step 4: ALWAYS CONCLUDE

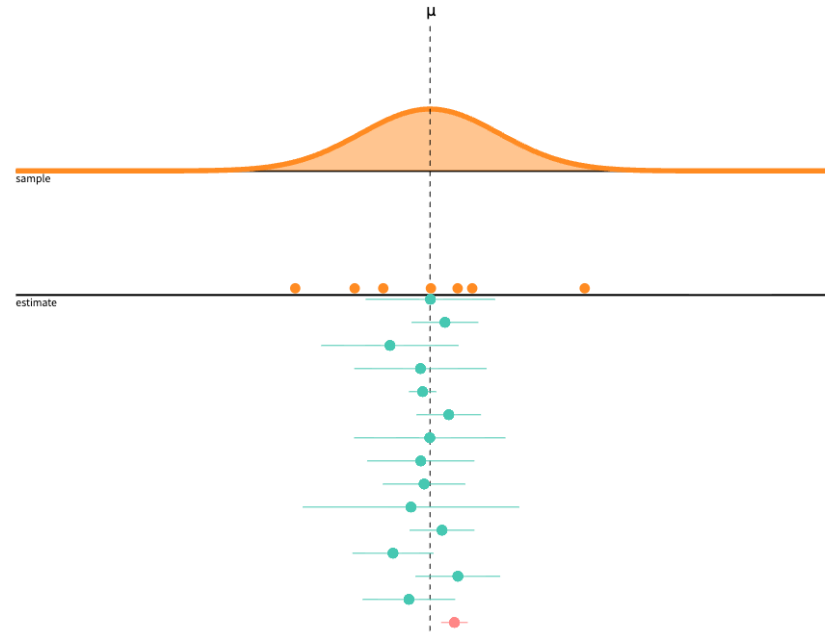
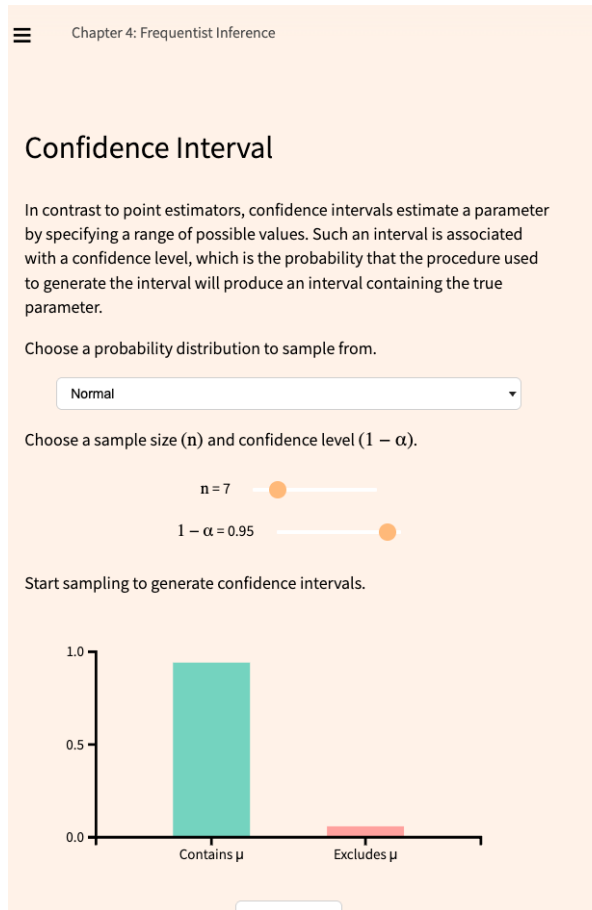
Athletes in red and blue shirts are not equally likely to win

(normally, we also put a confidence interval or any additional information to support our conclusion here, such as confidence interval, effect size calculation or whatever additional evidence is appropriate for your model.)

$$\hat{p} - 1.96 * \sqrt{\frac{\hat{p}(1 - \hat{p})}{20}} < p < \hat{p} + 1.96 * \sqrt{\frac{\hat{p}(1 - \hat{p})}{20}}$$

For Confidence Interval 95%: $0.625 < p < 0.975$

Confidence Intervals



<https://seeing-theory.brown.edu/frequentist-inference/index.html#section2>

95% CI means that if we repeated the study many times, 95% of those intervals would capture the true μ , not that there's a 95% chance *this* one does

Research Claim	H ₀ (Null Hypothesis)	H ₁ (Alternative Hypothesis)	Discussion
1. A new cholesterol drug reduces LDL levels compared to placebo.			
2. Sleep duration is associated with fasting glucose levels.			
3. CRISPR editing increases the rate of successful gene knock-in events.			
4. Cancer cells express Gene X more than normal cells.			

Research Claim	H_0 (Null Hypothesis)	H_1 (Alternative Hypothesis)	Discussion
1. A new cholesterol drug reduces LDL levels compared to placebo.	$\mu_1 = \mu_2$	$\mu_1 < \mu_2$	Should this be one-tailed or two-tailed?
2. Sleep duration is associated with fasting glucose levels.	$\rho = 0$	$\rho \neq 0$	What kind of data and test (correlation, regression)?
3. CRISPR editing increases the rate of successful gene knock-in events.	$p_1 = p_2$	$p_1 > p_2$	What if the baseline knock-in rate is already high?
4. Cancer cells express Gene X more than normal cells.	$\mu_{\text{cancer}} = \mu_{\text{normal}}$	$\mu_{\text{cancer}} > \mu_{\text{normal}}$	How does normalization affect interpretation?

χ^2 Goodness of fit test:

- **Compares observed counts to those predicted by a discrete probability distribution**
- Non-parametric (it does not require a normal probability distribution)

Assumptions:

Expected counts should ≥ 5 in $\geq 80\%$ categories

No category should have an EXPECTED value of < 1

We have just (implicitly) seen a specific category of χ^2 Goodness of fit test that gives us EXACT probabilities: The **Binomial Test**. This is a Goodness-of-Fit test that is limited to categorical variables with two outcomes only!

Fitting Discrete Models:

- A goodness-of-fit test compares observed counts to a discrete probability distribution

Example: Days of the week when babies born

- Discrete distribution is a probability distribution which describes discrete numerical random variables

Example:

Number of heads (10 flips of a coin)

Number of flowers in a square meter

Number of disease outbreaks in a year

Hypotheses for the χ^2 test:

H₀: *The data come from a particular discrete probability distribution*

H_A: *The data do not come from that distribution*

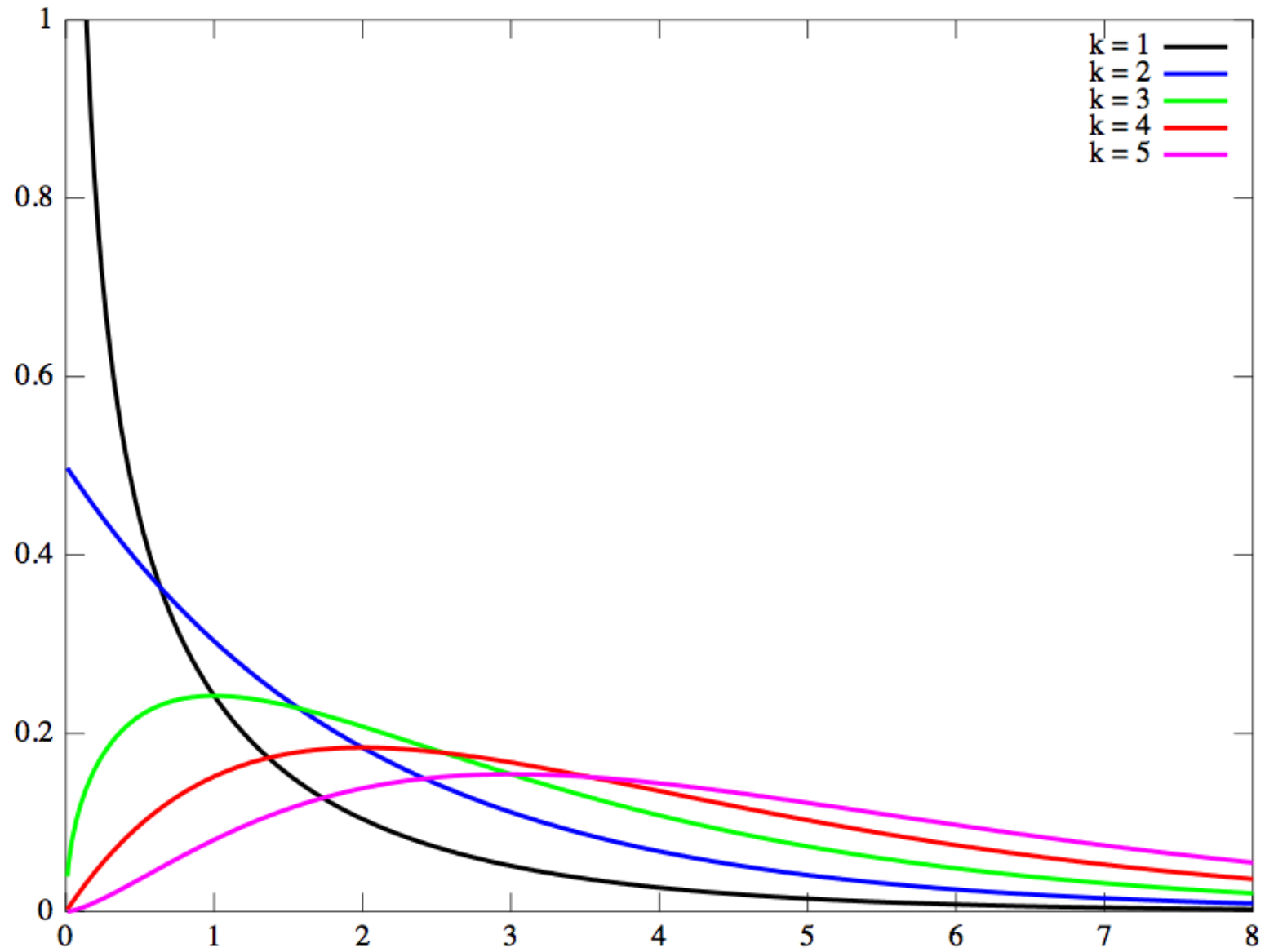
Test statistic for the χ^2 test:

$$\chi^2_{df} = \sum_i \frac{(\text{Observed}_i - \text{Expected}_i)^2}{\text{Expected}_i}$$

Degrees of Freedom:

d.o.f. = # categories - 1 - # est. parameters

χ^2 Distribution:



Each of the following accurately represents characteristics of the Chi-Square distribution except for:

A. As the degrees of freedom increase, the critical value of the Chi-Square distribution becomes larger

B. The region of rejection is always in the left-tail of the Chi-Square distribution

C. It is a positively skewed distribution

D. Its shape depends on the number of degrees of freedom

Finding the P-value of χ^2 distrⁿ:

- P-value of χ^2 test uses only right-hand tail of distribution
- χ^2 distribution is continuous, so probability is measured by area under curve ---> either use a computer statistical package or get really, really good at calculus

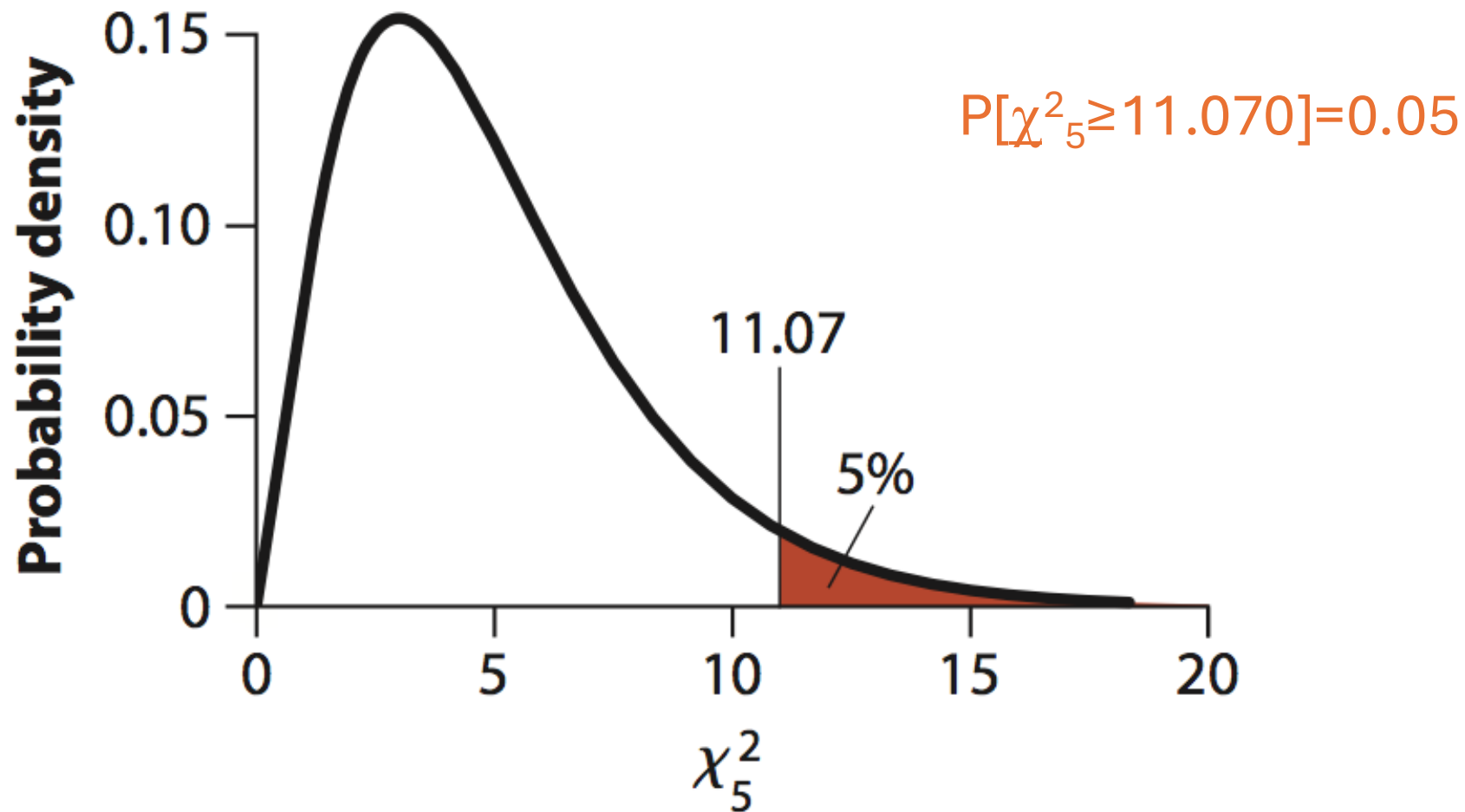
Finding critical values for χ^2 distrⁿ:

- **Critical value:** values of the test statistic that marks the boundary of a specified area of the sampling distribution
- Statistical Table:

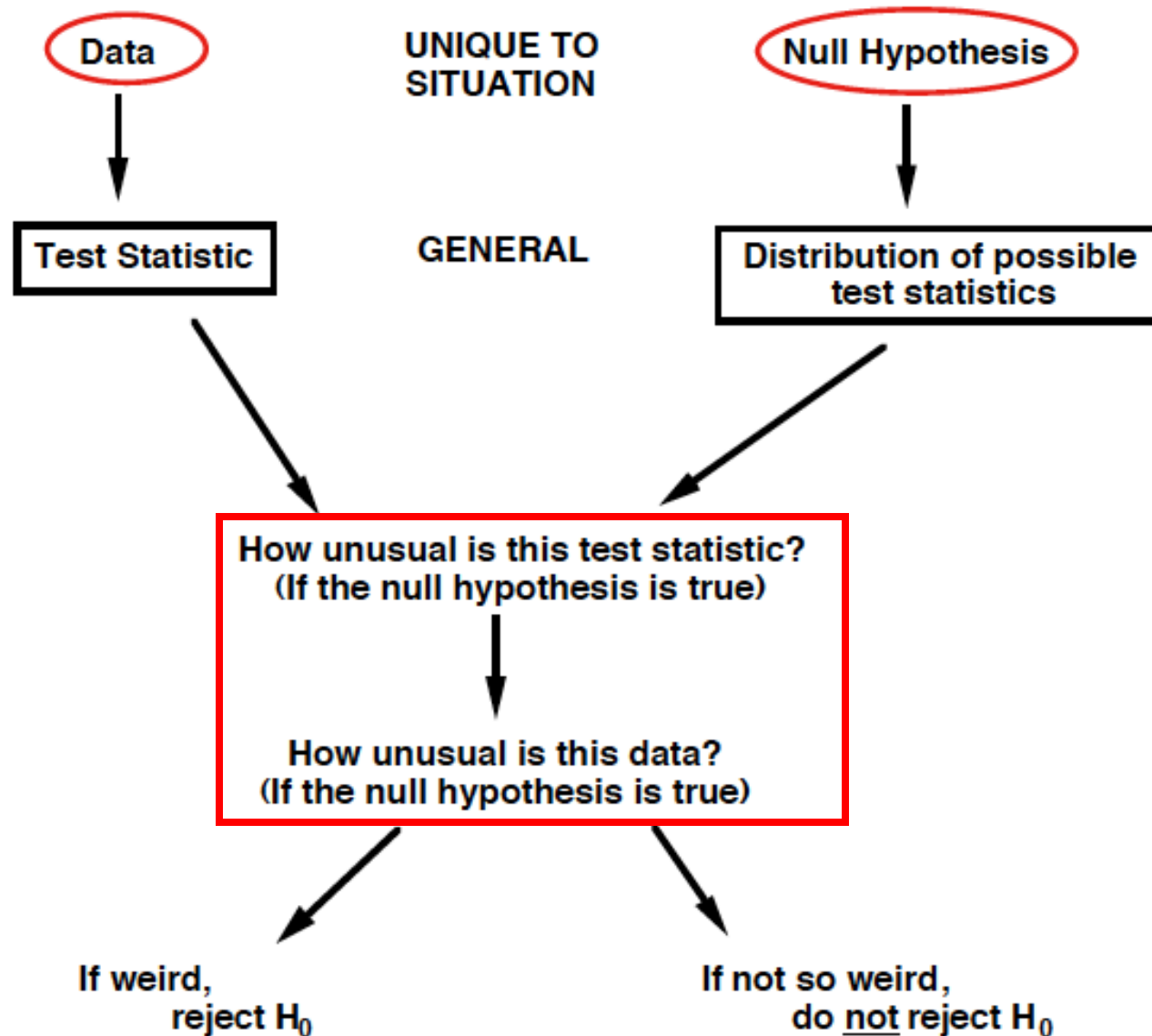
<https://www.math.arizona.edu/~jwatkins/chi-square-table.pdf>

df	0.995	0.950	0.05
1	0.000		0.00393	3.841
...				
5	0.412		1.145	11.070
6	0.676		1.635	12.592

Finding critical values for χ^2 distrⁿ:



Test Statistics and Hypothesis Testing



“A model is a mathematical tool that mimics how we *think* a natural process works..”

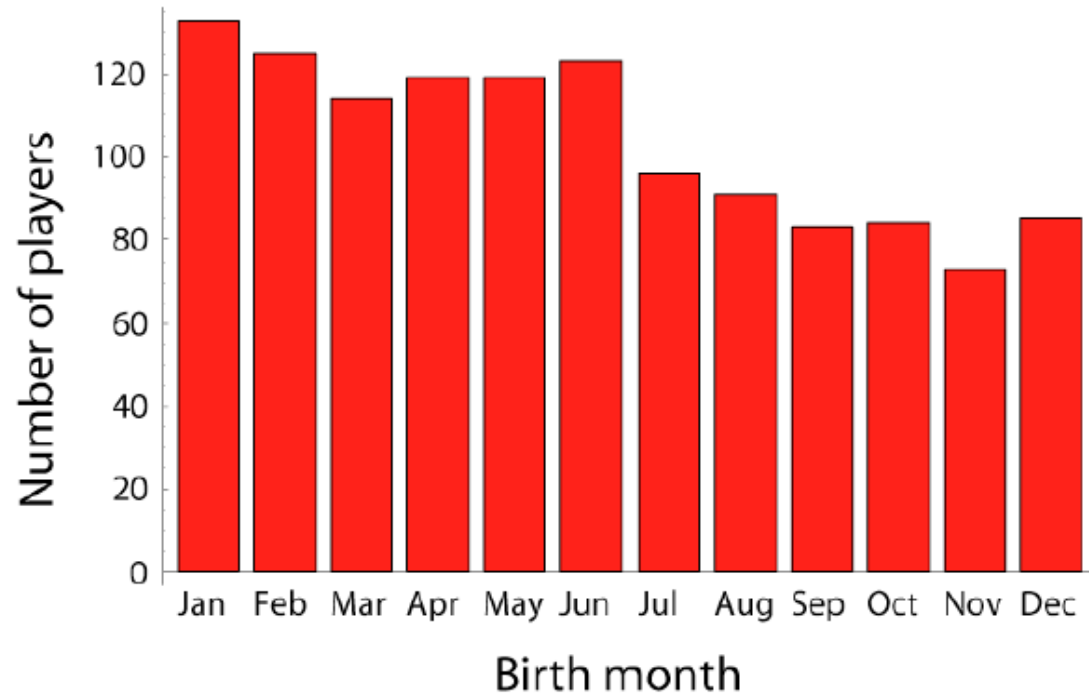
Life is interesting when a model doesn't fit the data because it suggests that at least one of the major assumption about how we think about the process is wrong

All models are wrong, but some are useful
- George E.P. Box

More of the χ^2 test:

Assumptions:

- No more than 20% of categories have **expected** frequencies < 5
- No category with **expected** frequencies < 1
 - You can sometimes work around these assumptions by chopping up your categories in a different manner so that they fulfill the criteria



Month	Num of Players
January	133
February	126
March	114
April	119
May	119
June	123
July	96
August	91
September	83
October	84
November	73
December	85

The birth month of Canadian hockey players

This information comes from Malcolm Gladwell “Outliers” (about 2006 players). Additional work, finding opposite:

<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0182827>
(analyzing 2010 Olympic teams)

H_0 :

The probability of an NHL birth occurring in any given month is equal to national birth proportions

H_A :

*The probability of an NHL birth occurring in any given month is **NOT** equal to national birth proportions*

Month	# of Players	Expected %
January	133	7.94
February	126	7.63
March	114	8.72
April	119	8.63
May	119	8.95
June	123	8.57
July	96	8.76
August	91	8.50
September	83	8.54
October	84	8.19
November	73	7.70
<u>December</u>	85	7.85
TOTAL	1245	100.0

Month	# of Players	Expected (%)	Expected (of 1245)
January	133	7.94	99
February	126	7.63	95
March	114	8.72	109
April	119	8.63	107
May	119	8.95	111
June	123	8.57	107
July	96	8.76	109
August	91	8.5	106
September	83	8.54	106
October	84	8.19	102
November	73	7.70	96
<u>December</u>	85	7.85	98
TOTAL	1245	100.0	1245

We'll go through the calculation for January:

$$\frac{(Observed_i - Expected_i)^2}{Expected_i} = \frac{(133 - 99)^2}{99} = \frac{1156}{99}$$

= January + February + March + April + May + June +.....+ November + December

$$= \frac{1156}{99} + \frac{900}{95} + \frac{25}{109} + \frac{144}{107} + \frac{64}{111} + \frac{256}{107} + \frac{169}{109} + \frac{225}{106} + \frac{529}{106} + \frac{324}{102} + \frac{529}{96} + \frac{169}{98}$$

$$= 44.77$$

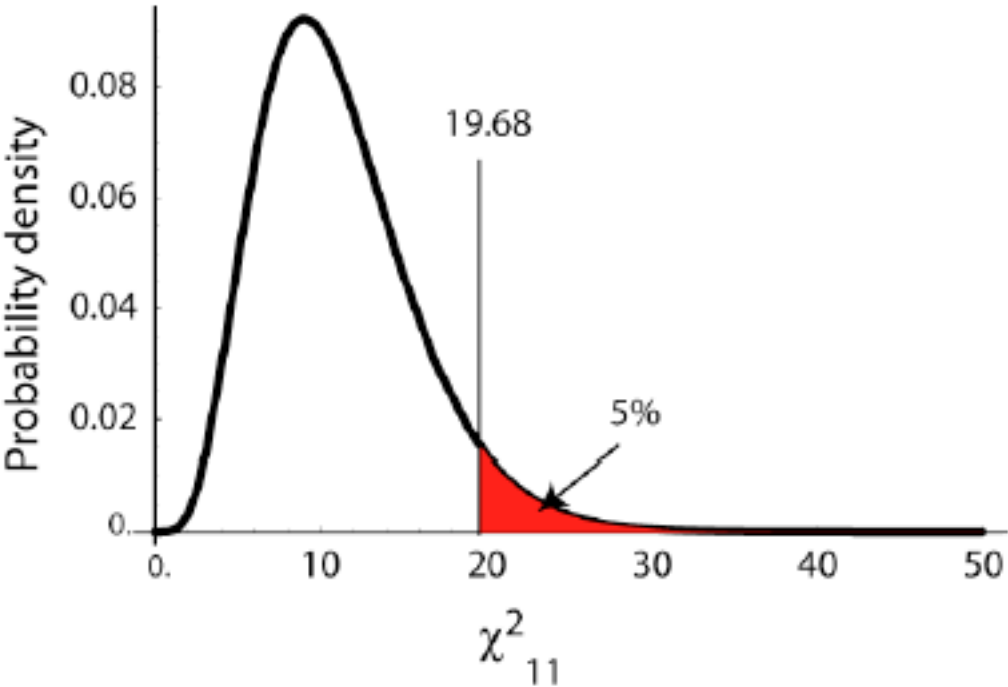
Step 3: There are 12 categories so: **dof = 12 - 1 - 0 = 11**

What is true about p-value in terms of χ^2 goodness-of-fit test?

- a. The p-value is the probability of getting a χ^2 value less than the observed χ^2 value calculated from the data
- b. The p-value is the probability of getting a χ^2 value equal to the observed χ^2 value calculated from the data
- c. The p-value is the probability of getting a χ^2 value greater than or equal to the observed χ^2 value calculated from the data
- d. The p-value does not make any significant impact

Find Critical Value using Table:

df	0.999	0.995	0.99	0.975	0.95	0.05	0.025	0.01	0.005	0.001
1	0.0000016	0.000039	0.00016	0.00098	0.00393	3.84	5.02	6.63	7.88	10.83
2	0.002	0.01	0.02	0.05	0.10	5.99	7.38	9.21	10.60	13.82
3	0.02	0.07	0.11	0.22	0.35	7.81	9.35	11.34	12.84	16.27
4	0.09	0.21	0.30	0.48	0.71	9.49	11.14	13.28	14.86	18.47
5	0.21	0.41	0.55	0.83	1.15	11.07	12.83	15.09	16.75	20.52
6	0.38	0.68	0.87	1.24	1.64	12.59	14.45	16.81	18.55	22.46
7	0.60	0.99	1.24	1.69	2.17	14.07	16.01	18.48	20.28	24.32
8	0.86	1.34	1.65	2.18	2.73	15.51	17.53	20.09	21.95	26.12
9	1.15	1.73	2.09	2.70	3.33	16.92	19.02	21.67	23.59	27.88
10	1.48	2.16	2.56	3.25	3.94	18.31	20.48	23.21	25.19	29.59
11	1.83	2.60	3.05	3.82	4.57	19.68	21.92	24.72	26.75	31.41
12	2.21	3.07	3.57	4.40	5.23	21.03	23.34	26.22	28.30	33.41



ALWAYS CONCLUDE:

We can reject the null hypothesis: NHL players are *not* born in the same proportions per month as the population at large with a P-value $\leq \alpha$ (=0.05).

Was Mendel's data 'too good'?

- RA Fisher accused Mendel of having data that fit too well (approx. 3/100,000 experiments should have data that fit as well as Mendel's)
- A raging debate ever since:
 - <http://www.istics.net/stat/>
 - Go to "Not random enough" in the sidebar
 - <http://arxiv.org/pdf/1104.2975.pdf>

Example: the results of a **Monohybrid cross** between a (heterozygous) yellow pea plant (**Yy**) and a green pea (**yy**) plant are as follows: **14 yellow** and **6 green**. Are these results consistent with Mendel's first law (**segregation**) which should a 1:1 ratio in this case? (alpha = 0.05). Punnet square shown:

Yellow (Hetero)\Green (Homo)	y	y	
Y	Yy	Yy	← 14
y	yy	yy	← 6

- A. Yes – the results FTR the null hypothesis
- B. No – the results Reject the null hypothesis
- C. Yes – the results reject the null hypothesis
- D. No – the results FTR the null hypothesis

Example: the results of a **Monohybrid cross** between a (heterozygous) yellow pea plant (Yy) and a green pea (yy) plant are as follows: **14 yellow** and **6 green**. Are these results consistent with Mendel's first law (**segregation**) which should a 1:1 ratio in this case?

step 1

H_0 :

*If the hypothesis (Yy is yellow and yy is green) is true then we would expect a 1:1 ratio in progeny (or, to put it into counts: 10 **Yellow** and 10 green)*

H_A :

*The genotypes that result in **Yellow** and green are not simply Yy and yy (there may be something else going on)*

Example: the results of a **Monohybrid cross** between a (heterozygous) yellow pea plant (Yy) and a green pea (yy) plant are as follows: **14 yellow** and **6 green**. Are these results consistent with Mendel's first law (**segregation**) which should a 1:1 ratio in this case?

Step 2 Test Statistic:

assumptions: no category with expected freq < 1; no more than 20% categories have expected freq < 5

$$X^2 = (14-10)^2/10 + (6-10)^2/10 = 1.6 + 1.6 = 3.2$$

step 3 dof = # categories – 1 = 1

critical value that corresponds to 0.05 = 3.84

step 4 conclusion: since the critical value > value calculated from the **X²** test statistic; the ratio of progeny obtained is FTR Ho; **consistent** with a 1:1 ratio.