Module 4B: Hypothesis Testing

Applied Epistemology: A Framework for how we know things scientifically

Agenda:

- Working through examples of hypothesis testing
  - Binomial Example
  - **χ^2 Goodness of fit tests**

Your pipeline for hypothesis testing in statistics

Step 1

Formulate your null **hypothesis**

- Null hypothesis is ***only*** *hypothesis that is tested*
- Falsification: *want* to reject your null

Step 2

Identify appropriate **test statistic**

- Assumptions of your test

Step 3

Quantify the results of your test

- **P value** or comparison to **critical values**
- How *unusual* is your data?

Step 4

Conclude: reject or fail to reject

- based on alpha value
- if appropriate, confidence interval of the parameter

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)
- *There is a second related test called the χ^2 Contingency test that is a type of χ^2 GOF test but has different degrees of freedom and a specific Ho*

Assumptions:

- Expected counts should >=5 in >=80% categories
- No category should an EXPECTED value of <1

We have just (implicitly) seen a specific category of X^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 GOF test:

H$_0$: *The data come from a particular discrete probability distribution*

H$_A$: *The data do not come from that distribution*

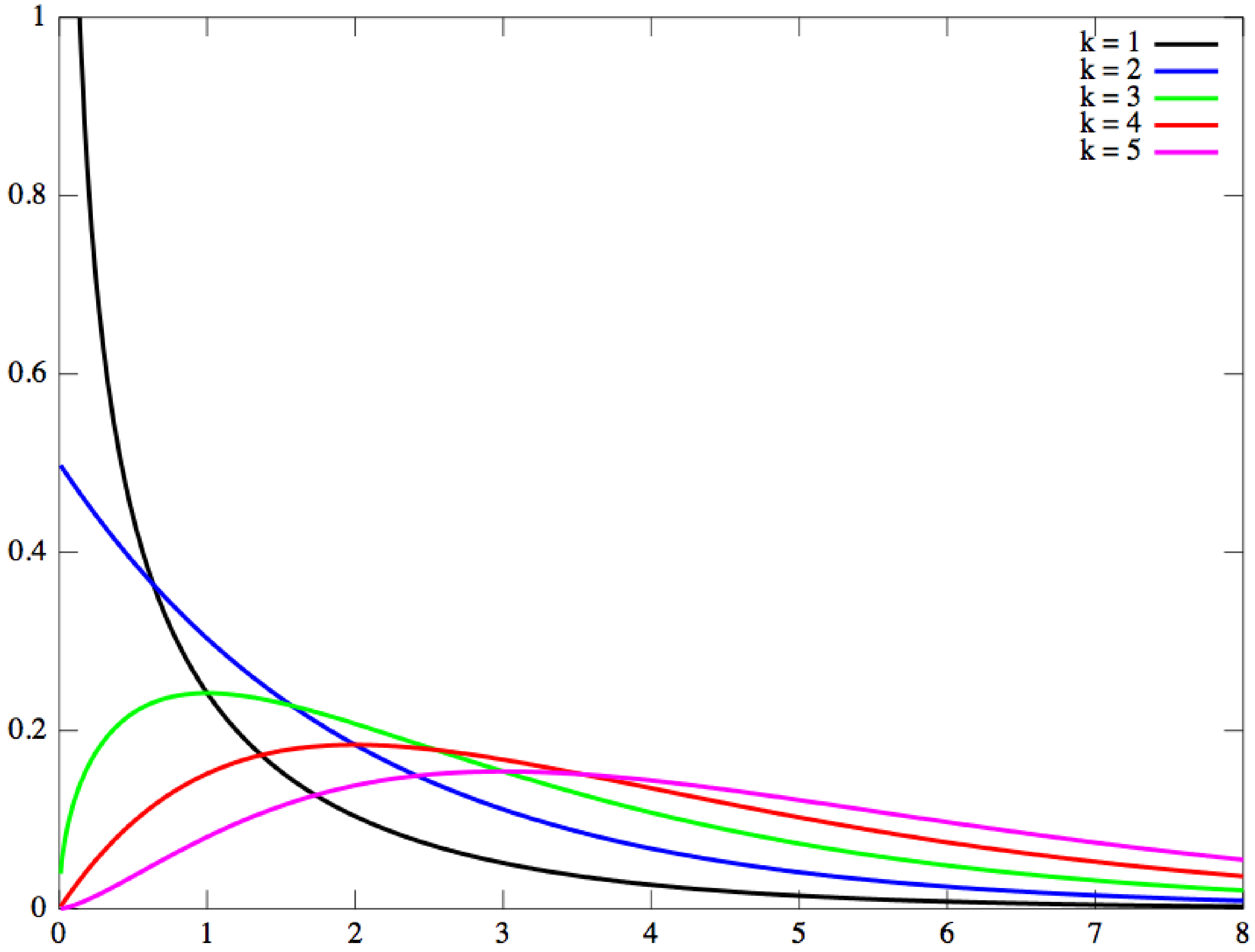
Test statistic for the χ^2 test:

$$\boxed{\chi^2_{df} = \sum_i \frac{(Observed_i - Expected_i)^2}{Expected_i}}$$

Degrees of Freedom:

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

χ^2 Distribution:



Finding the P-value of χ^2 distr^n:

- 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 distr^n:

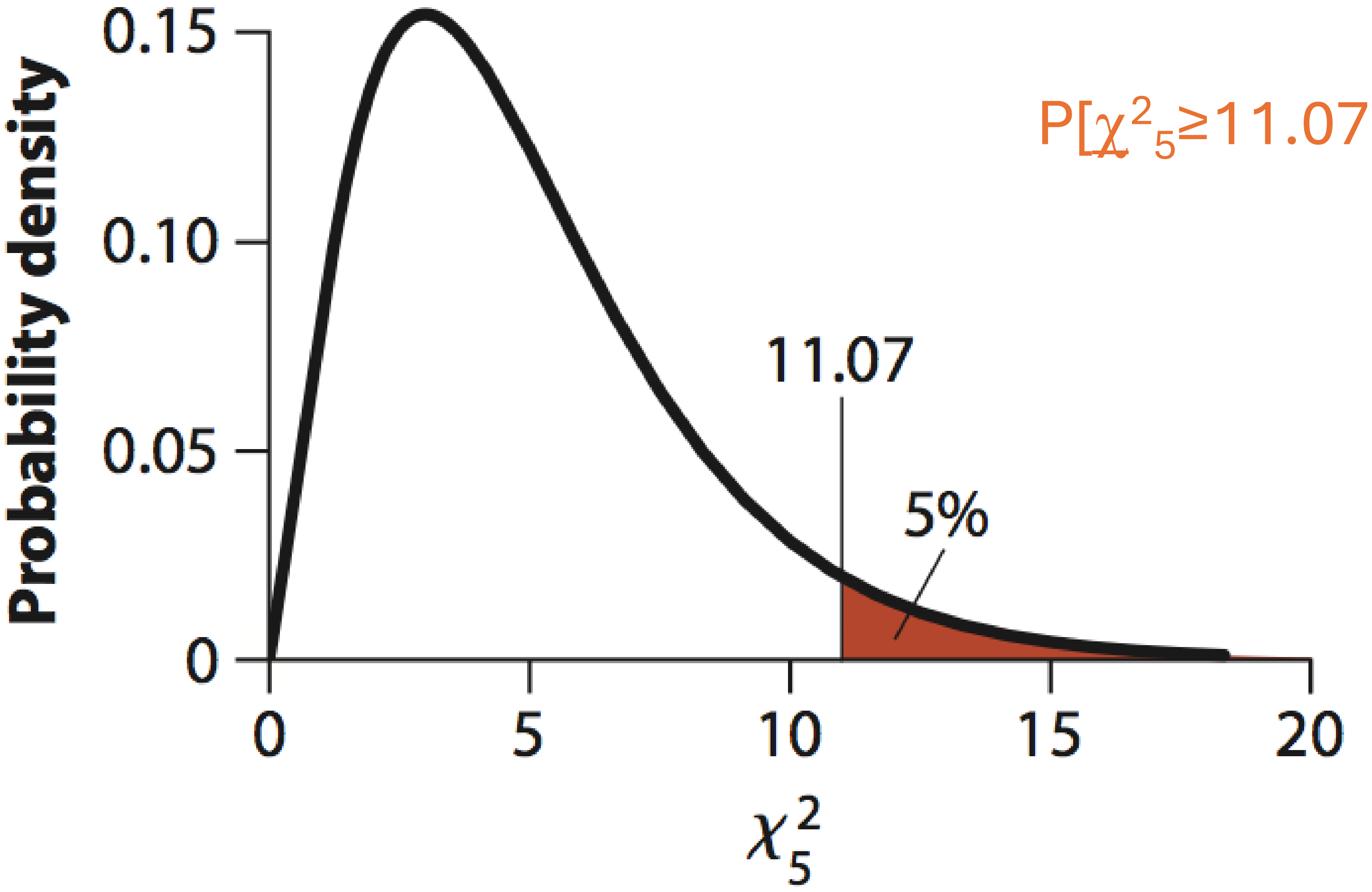
- **Critical value**: values of the test statistic that marks the boundary of a specified area of the sampling distribution

- Statistical Table:

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

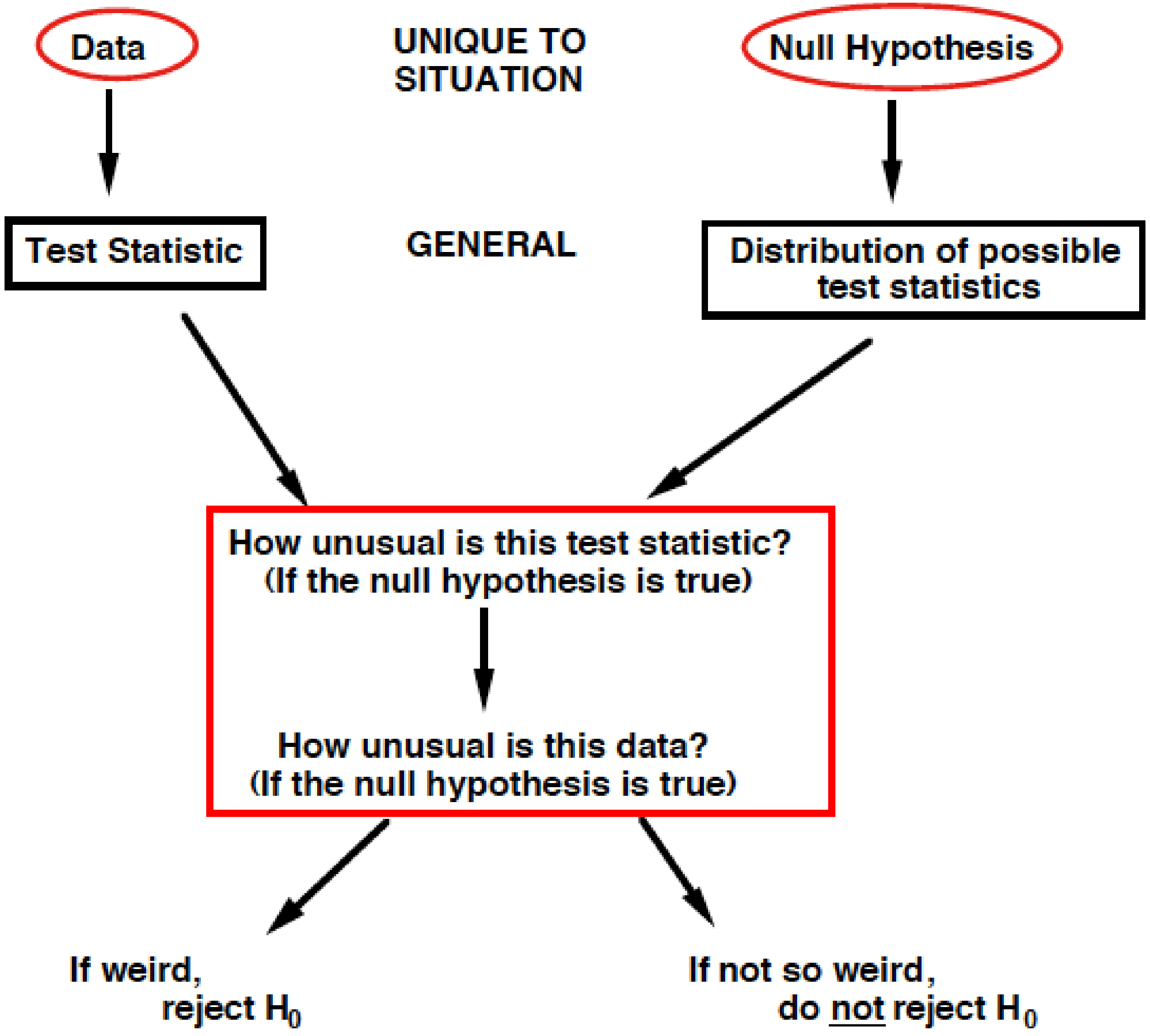
df	**0.995**	**....**	**0.950**	**0.05**
1	0.000		0.00393	3.841
…				
5	0.412		1.145	**11.070**
6	0.676		1.635	12.592

Finding **critical values for** χ^2 distr^n:



$P[\chi^2_5 \geq 11.070] = 0.05$

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

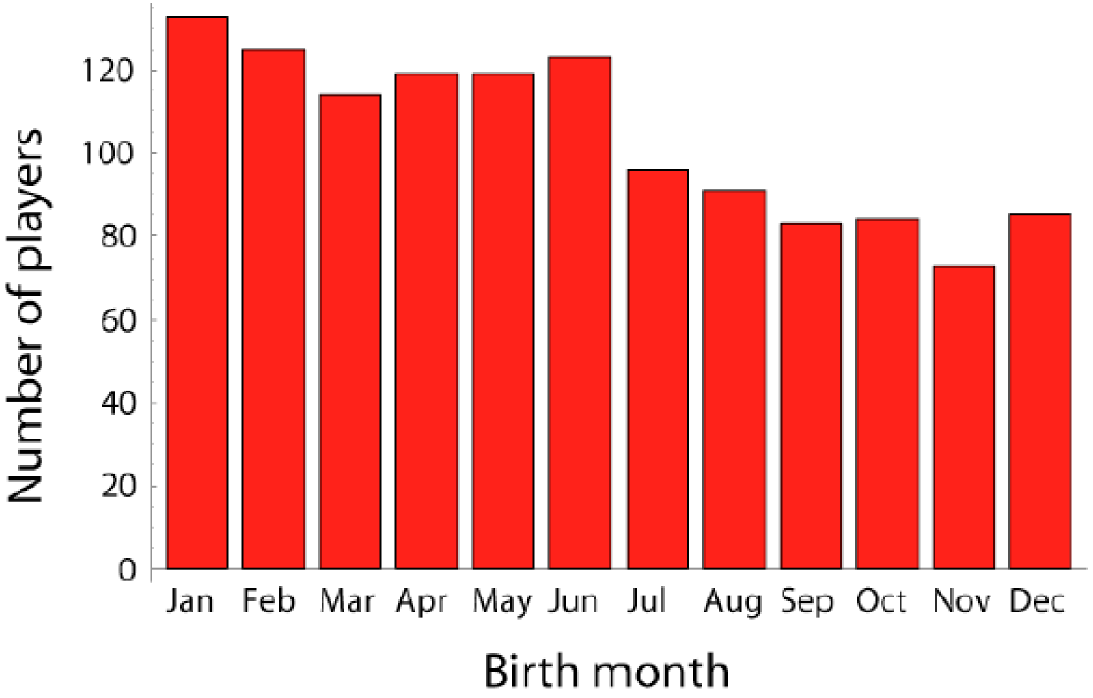
More of the χ^2 test:

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
December	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
December	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

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

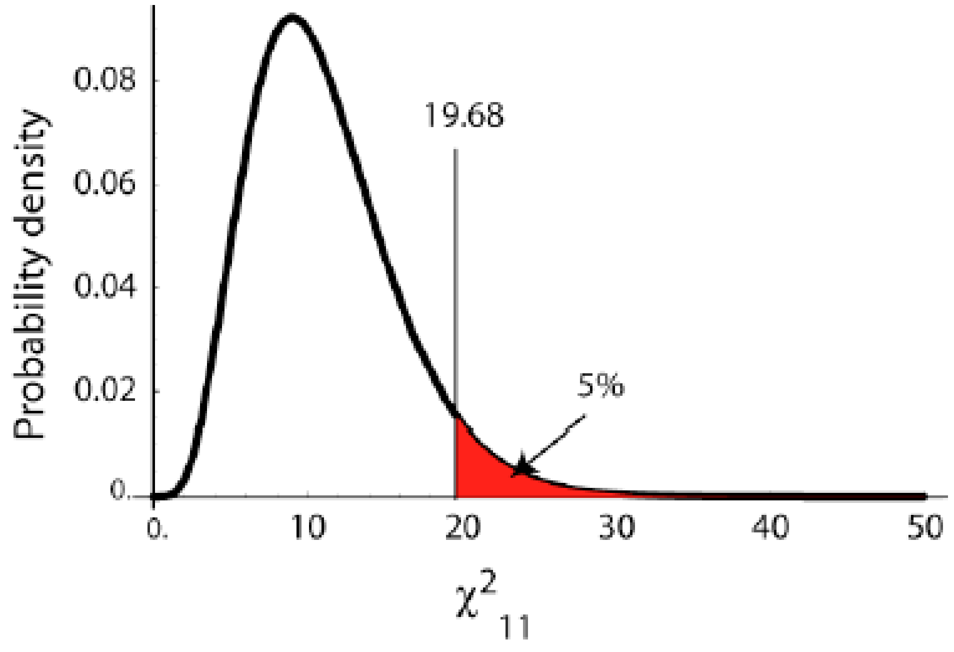
= 44.77

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

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

<u>Find Critical Value using Table:</u>

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	$\boxed{19.68}$	21.92	24.72	26.76	31.26
12	2.21	3.07	3.57	4.40	5.23	21.03	23.34	26.22	28.30	32.91



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, =0.001).