Fitting probability models to frequency data

χ^2 Goodness-of-fit test

Compares counts to a discrete probability distribution

Discrete distribution

A probability distribution describing a discrete numerical random variable

For example,

- Number of heads from 10 flips of a coin
- Number of flowers in a square meter
- Number of disease outbreaks in a year

Hypotheses for χ^2 test

H_0: The data come from a particular discrete probability distribution.

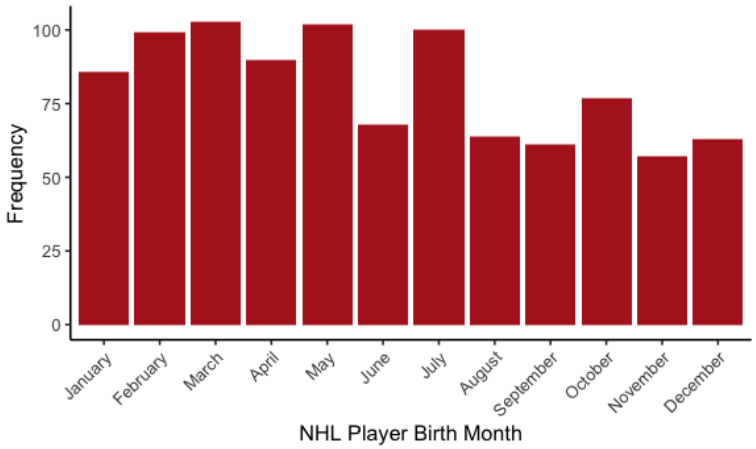
H_A: The data do not come from that distribution.

Test statistic for χ^2 test

$$\chi^2 = \sum_{all\ classes} \frac{\left(Observed_i - Expected_i\right)^2}{Expected_i}$$

Month	Number of NHL players
January	86
February	99
March	103
April	90
May	102
June	68
July	100
August	64
September	61
October	77
November	57
December	63
Sum	970

Data from https://www.quanthockey.com/nhl/birth-month-totals for 2019-2020



A *Goodness-of-Fit test* compares count data to a model of the expected frequencies of a set of categories.

Hypotheses for birth month example

H_0: The probability of a NHL birth occurring on any given month is equal to national proportions.

H_A: The probability of a NHL birth occurring on any given month is *not* equal to national proportions.

NHL compared to all Canadians

Month	Number of NHL players	Proportion Canadian births
January	86	0.081
February	99	0.077
March	103	0.087
April	90	0.086
May	102	0.09
June	68	0.086
July	100	0.088
August	64	0.085
September	61	0.085
October	77	0.082
November	57	0.076
December	63	0.077
Sum	970	1

Computing Expected values

Month	Number of NHL players	Proportion Canadian births	Expected
January	86	0.081	78.57
February	99	0.077	74.69
March	103	0.087	84.39
April	90	0.086	83.42
May	102	0.09	87.3
June	68	0.086	83.42
July	100	0.088	85.36
August	64	0.085	82.45
September	61	0.085	82.45
October	77	0.082	79.54
November	57	0.076	73.72
December	63	0.077	74.69
Sum	970	1	970

The calculation for January

$$\frac{(Observed\ -Expected)^2}{Expected} = \frac{(86-78.57)^2}{78.57} = 0.7026$$

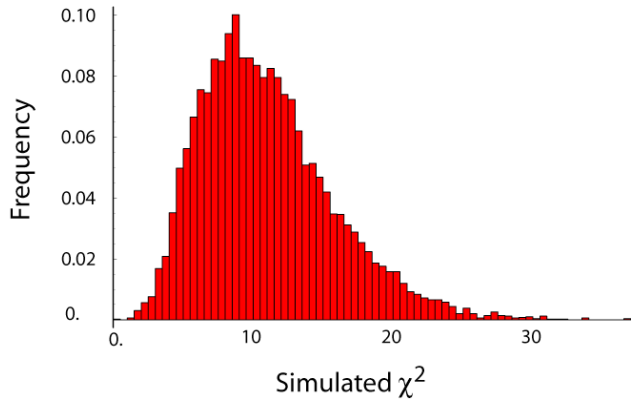
Calculating χ^2

$$\chi^2 = \sum_{all\ classes} \frac{(Observed\ - Expected)^2}{Expected}$$

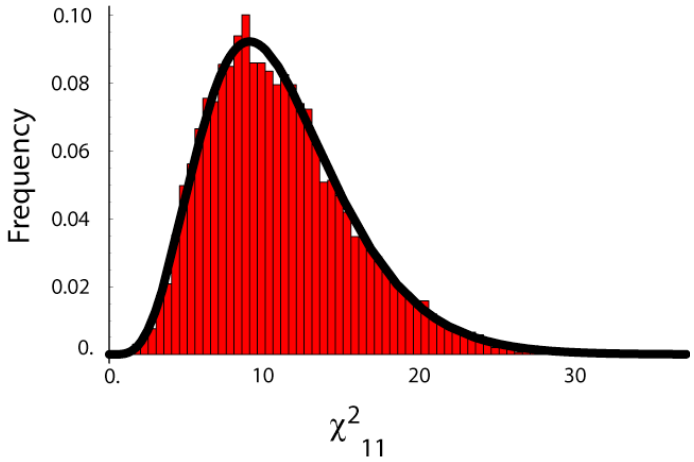
$$= \begin{pmatrix} 0.703 + 7.912 + 4.104 + 0.519 + 2.475 + 2.850 + \\ 2511 + 4.129 + 5.580 + 0.081 + 3.792 + 1.830 \end{pmatrix}$$

$$= 36.5$$

The sampling distribution of χ^2 by simulation



Sampling distribution of χ^2 by the χ^2 distribution



Degrees of freedom

The number of degrees of freedom of a test specifies which of a family of distributions to use.

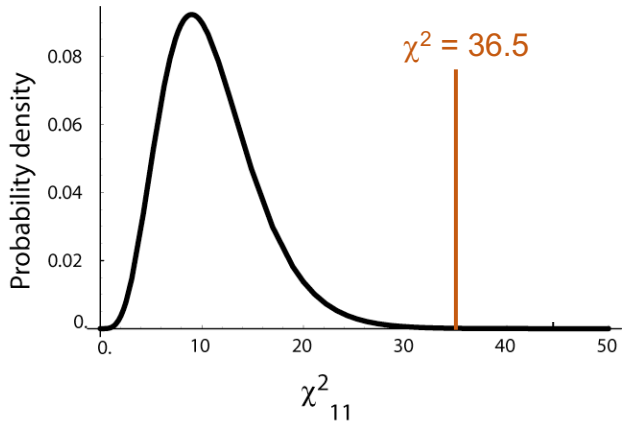
Degrees of freedom for χ^2 test

df = (Number of categories)
– (Number of parameters estimated from the data)
– 1

Degrees of freedom for NHL month of birth

$df = 12 - 0 - 1 = 11$

Finding the *P*-value



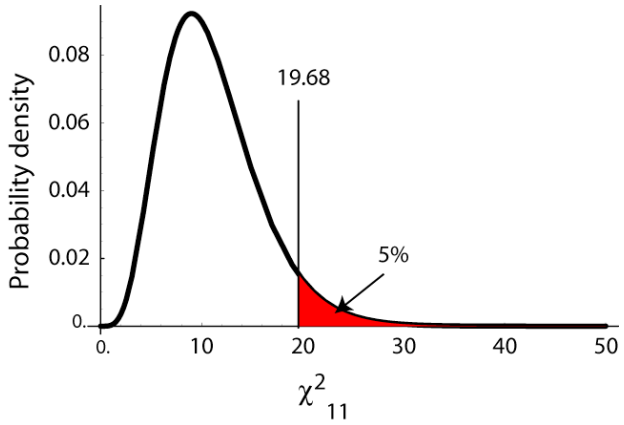
Critical value

The value of the test statistic where $P = \alpha$.

Table A - χ^2 distribution

	α									
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	26.76	31.26
12	2.21	3.07	3.57	4.40	5.23	21.03	23.34	26.22	28.30	32.91

The 5% critical value



P<0.05,
so we can reject the null hypothesis

NHL players are not born in the same proportions per month as the population at large.

Goodness of fit in R

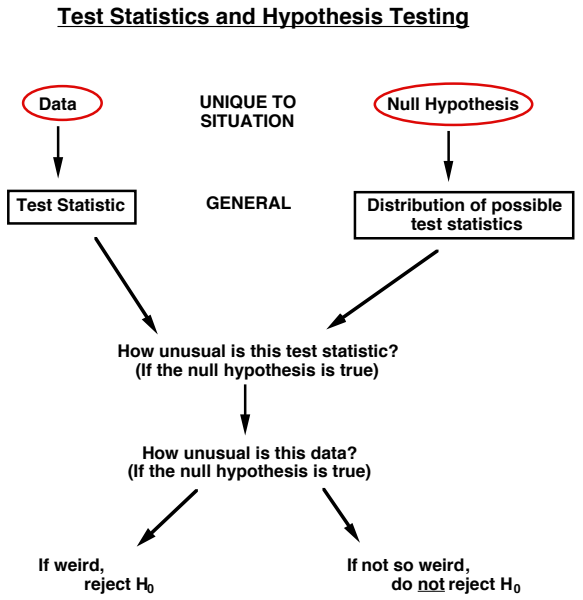
```
NHLBirthMonthTable = table(NHLPlayerData$BirthMonth)

CanadianBirthMonthProportions =
      c(0.081, 0.077, 0.087, 0.086, 0.09, 0.086, 0.088,
      0.085, 0.085, 0.082, 0.076, 0.077)

chisq.test(NHLBirthMonthTable,
      p = CanadianBirthMonthProportions)
```

Test statistics

A *test statistic* is a number calculated from the data and the null hypothesis that can be compared to a standard distribution to find the *P*-value of the test.



χ^2 test as approximation of binomial test

χ^2 goodness-of-fit test works even when there are only two categories, so it can be used as a substitute for the binomial test.

Very useful if the number of data points is large.

See text for an example.

Assumptions of χ^2 test

- No more than 20% of categories have *Expected*<5
- No category with *Expected* ≤ 1

Estimating parameters from data

HUU (Hyperuricosuria and hyperuricemia) caused by a mutation in the SLC2A9 gene

Zierath, S. 2017. Frequency of five disease-causing genetic mutations in a large mixed-breed dog population (2011–2012). PLoS ONE 12(11): e0188543.

Estimating parameters from data

34,118 mixed breed dogs tested:

57 Homozygous for mutation
1517 Heterozygotes
32,544 Homozygous for wild type

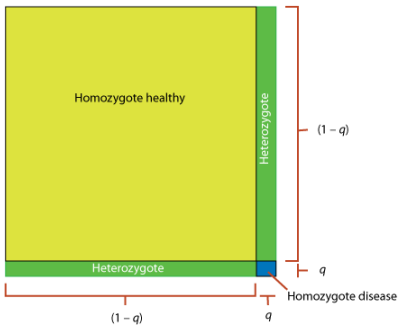
Do these genotypes appear in frequencies predicted by random pairing of alleles?

Zierath, S. 2017. Frequency of five disease-causing genetic mutations in a large mixed-breed dog population (2011–2012). PLoS ONE 12(11): e0188543.

Estimating parameters from data

The expectation for these frequencies is

Homozygous mutant: q^2
Heterozygote: $2\,q\,(1-q)$
Homozygous healthy: $(1-q)^2$



Estimating parameters from data

H_0: Genotype frequencies follow predictions of random association of alleles:

$$q^2 : 2\,q\,(1-q) : (1-q)^2$$

But what is the value of q?

Estimating parameters from data

But what is the value of *q*?

$$q = \frac{Freq.homozygote + \frac{1}{2} Freq.heterozygote}{Total\ number}$$

$$\hat{q} = \frac{57 + 1517/2}{34118} = 0.024$$

Estimating parameters from data

Expected values:

Homozygous mutant:	$\hat{q}^2\, n$	= 19.5
Heterozygote:	$2\,\hat{q}\,(1-\hat{q})\, n$	= 1592.0
Homozygous healthy:	$(1-\hat{q})^2\, n$	= 32506.5

But remember – we had to estimate one parameter ($\hat{q}$) from the data.

Estimating parameters from data

$$\chi^2 = \frac{(57-19.5)^2}{19.5} + \frac{(1517-1592.0)^2}{1592.0} + \frac{(32544-32506.5)^2}{32506.5} = 75.8$$

df = Number of classes – number of parameters estimated from data – 1

= 3 – 1 – 1 = 1

We had to estimate one parameter ($\hat{q}$) from the data.

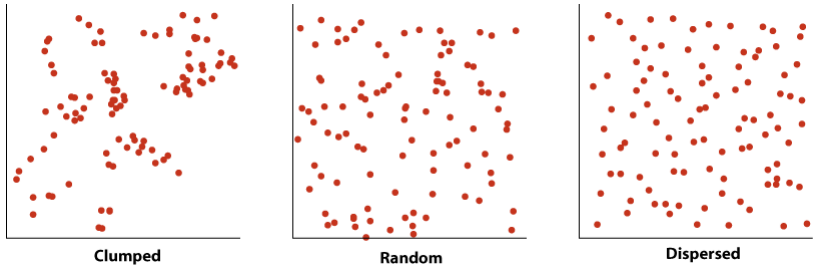
Estimating parameters from data

```
pchisq(sum(chiParts),df = 1, lower.tail=FALSE)
[1] 3.217808 e-18
```

Therefore P = 3.2 x 10^{-18}, and we reject the null hypothesis. These genotypes do not occur as we would expect by random combinations of alleles.

Fitting other distributions: the Poisson distribution

The Poisson distribution describes the probability that a certain number of events occur in a block of time or space, when those events happen independently of each other and occur with equal probability at every point in time or space.



Poisson distribution

$$\Pr[X] = \frac{e^{-\mu}\mu^{X}}{X!}$$

Example: Number of goals per side in World Cup Soccer

Q: Is the outcome of a soccer game (at this level) random?

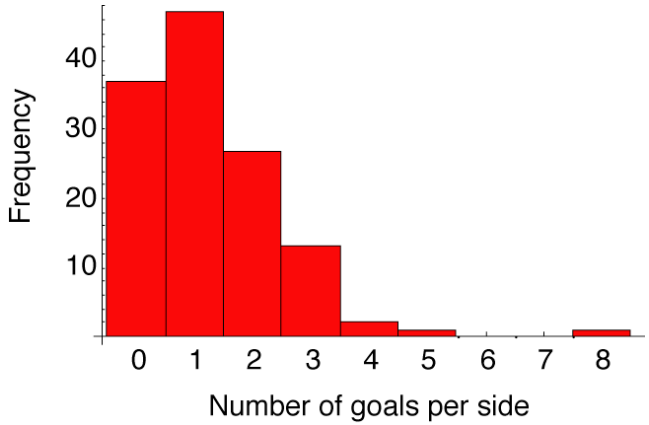
In other words, is the number of goals per team distributed as expected by pure chance?

Hypotheses

H_0: Number of goals per side follows a Poisson distribution.

H_A: Number of goals per side does not follow a Poisson distribution.

World Cup 2002 scores



Number of goals for a team (World Cup 2002)

Number of goals	Frequency
0	37
1	47
2	27
3	13
4	2
5	1
6	0
7	0
8	1
Total	128

What's the mean, μ?

$$\bar{x} = \frac{37(0) + 47(1) + 27(2) + 13(3) + 2(4) + 1(5) + 1(8)}{128}$$

$$= \frac{161}{128}$$

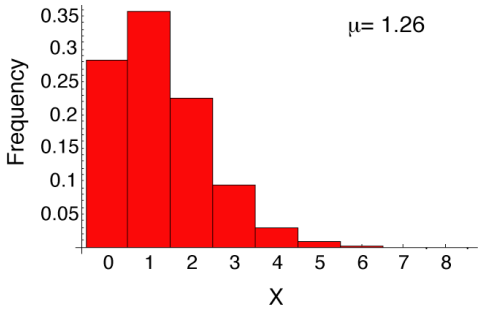
$$= 1.26$$

Poisson with $\mu = 1.26$

Example:

$$\Pr[2] = \frac{e^{-\mu}\mu^{X}}{X!} = \frac{e^{-1.26}(1.26)^{2}}{2!} = \frac{(0.284)1.59}{2} = 0.225$$

Poisson with $\mu = 1.26$



X	Pr[X]
0	0.284
1	0.357
2	0.225
3	0.095
4	0.030
5	0.008
6	0.002
7	0
≥8	0

Finding the *Expected*

X	Pr[X]	*Expected*
0	0.284	36.3
1	0.357	45.7
2	0.225	28.8
3	0.095	12.1
4	0.030	3.8
5	0.008	1.0
6	0.002	0.2
7	0	0.04
≥8	0	0.007

} Too small!

Calculating χ^2

X	*Expected*	*Observed*	$\frac{(Observed_i - Expected_i)^2}{Expected_i}$
0	36.3	37	0.013
1	45.7	47	0.037
2	28.8	27	0.113
3	12.1	13	0.067
≥ 4	5.0	4	0.200

$$\chi^2 = \sum_{all\ classes} \frac{(Observed_i - Expected_i)^2}{Expected_i} = 0.429$$

Degrees of freedom

df = (Number of categories)

– (Number of parameters estimated from the data)

– 1

= 5 – 1 – 1 = 3

Critical value

	α									
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	26.76	31.26
12	2.21	3.07	3.57	4.40	5.23	21.03	23.34	26.22	28.30	32.91

Comparing χ^2 to the critical value

$$\chi^2 = 0.429$$

$$\chi^2_3 = 7.81$$

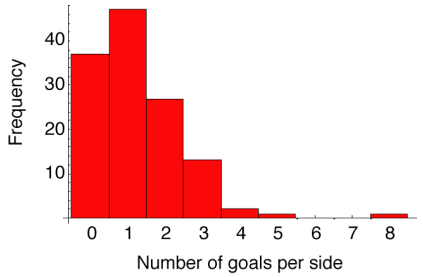
$$0.429 < 7.81$$

```
pchisq(0.429,df = 3, lower.tail=FALSE)
[1] 0.9341887
```

So we cannot reject the null hypothesis.

There is no evidence that the score of a World Cup Soccer game is not Poisson distributed.

World Cup 2002 scores



Poisson distribution