Chi-Square Tests

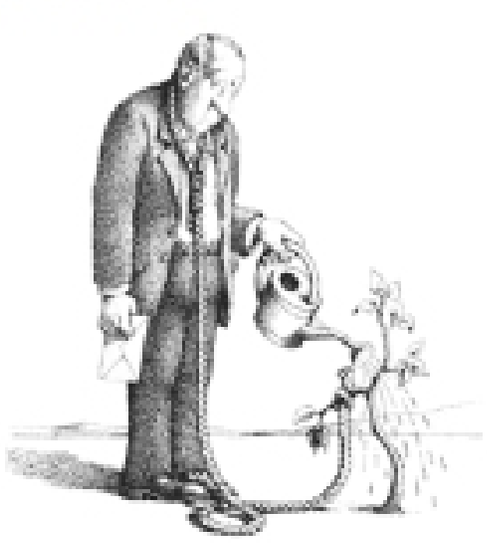
Testing for Goodness of Fit and Independence

Part I: Testing Goodness of Fit

- There is a chance model
- There are observed frequency counts
- Wish to see whether the counts are consistent with the chance model (whether it fits the data well)

Example: Counts of Suicides by Month in US in 1970

Jan	1867
Feb	1789
Mar	1944
Apr	2094
May	2097
Jun	1981
Jul	1887
Aug	2024
Sept	1928
Oct	2032
Nov	1978
Dec	1859
Total	23,480



Are all months equally likely? Compare observed frequencies to those expected from a box model:

- Tickets: labeled 1-365 for days of the year
- Draws: 23,480 with replacement
- Group the results into months

According to this chance model, a June ticket has a probability of 31/365. The expected number of June suicides is

$$23480 \times (31/365) = 1929.86$$

	Days	Observed	Expected
Jan	31	1867	1994.19
Feb	28	1789	1801.21
Mar	31	1944	1994.19
Apr	30	2094	1929.86
May	31	2097	1994.19
Jun	30	1981	1929.86
Jul	31	1887	1994.19
Aug	31	2024	1994.19
Sep	30	1928	1929.86
Oct	31	2032	1994.19
Nov	30	1978	1929.86
Dec	31	1859	1994.19

	Days	Observed	Expected	O - E	$(O - E)^2/E$
Jan	31	1867	1994.19	-127.19	8.11
Feb	28	1789	1801.21	-12.21	0.08
Mar	31	1944	1994.19	-50.19	1.26
Apr	30	2094	1929.86	164.14	13.96
May	31	2097	1994.19	102.81	5.30
Jun	30	1981	1929.86	51.14	1.36
Jul	31	1887	1994.19	-107.19	5.76
Aug	31	2024	1994.19	29.81	0.45
Sep	30	1928	1929.86	-1.86	0.00
Oct	31	2032	1994.19	37.81	0.72
Nov	30	1978	1929.86	48.14	1.20
Dec	31	1859	1994.19	-135.19	9.17

$$\text{Total} = \chi^2 = 47.37$$

"chi-square"

The chi-square statistic measures how closely the observed and expected counts agree.

Even if the chance model from which the expected counts are derived holds exactly, the two will not agree perfectly, just because of chance.

In order to judge how big is unusual, we need to know the probability law of the chi-square statistic when the chance model is true.

This is similar to the case of the z-statistic: it's numerator will generally be different from 0 even when the null hypothesis is true.

Null hypothesis: the chance model generated the data

Alternative hypothesis: it didn't, there is something else going on.

In our example:

Null hypothesis: suicides are equally likely on any day.

Alternative hypothesis: There is something else going on, like seasons have an effect.

Chi-square distribution

If the null hypothesis is true, the probability histogram of the chi-square statistic is approximately equal to the chi-square distribution with "degrees of freedom" equal to the number of cells minus one.