§9.2 Goodness of fit

Take a sample of size n of a categorical variable with k possible outcomes:

outcome	1	2	3	...	k	total
actual count	n_1	n_2	n_3	...	n_k	n
expected count	$p_1 n$	$p_2 n$	np_3	...	np_k	n

Do the follow a particular distribution $p_1 \dots p_k$?

expected count: np_i

χ^2-square statistic measures the difference from the expected values:

$$\chi^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$ ← distributed as χ^2 with $k-1$ d.f.

Note: want expected counts ≥ 5 in all cells.

Example (9.13) Calcium study

Number of study participants in sub sample:

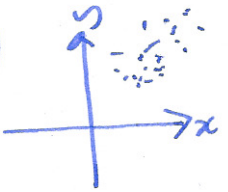
	AZ	CA	HI	IN	NV	OH	total
	167	257	297	~~107~~ 297	107	482	1567
p_i	0.105	0.172	0.164	0.188	0.070	0.301	1
np_i	164.4	269	256	294	109	471	1567

null hypothesis: number of participants should be proportional to state population in original sample.

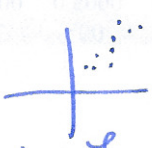
χ^2 statistic is 0.93. df = 5. P-value about 0.25, not significantly different.

§2.2 Scatterplots

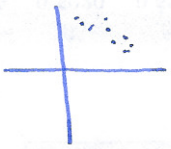
data with two variables, e.g. (height, weight) can plot as (x, y)



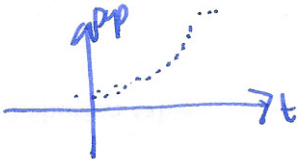
positive correlation



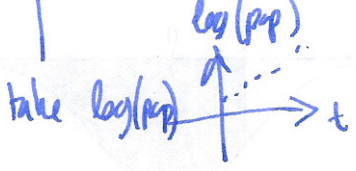
negative correlation



can change coordinates if useful



take log(pop)



§2.3 Correlation

correlation coefficient is $r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$.

(n data points)