5. Goodness of fit

Are the model assumptions plausible?

Have seen: plots, histograms, Q-Q plots

Are the model assumptions necessary?

Often the central limit theorem suffices

Are there outliers or unusual values in the data?

Check the data collection technique

Test: Discrepancy between the data and the model?

discrete: Chisquare test for goodness of fit: Pearson's Chisquare

Recall: Multinomial distribution: n balls, m cells, probability p_i for cell i, $i = 1, \ldots, m$, balls thrown independently; x_i is the count in cell i, $i = 1, \ldots, m$

$$f(\mathbf{x}, (p_1, \ldots, p_n; m)) = \binom{n}{x_1, \ldots, x_m} p_1^{x_1} \cdots p_m^{x_m}$$

$H_0 : p_i = p_i(\theta)$, $i = 1, \ldots, m$; $H_1 : not$

$$X^2 = \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i}$$

degrees of freedom: $dim\Theta - dim\Theta_0$

Example Compare demographic characteristics of jurors with the general population in Alameda County, California

Age	County perc.	no. jurors
21-40	42	5
41-50	23	9
51-60	16	19
61 and over	19	33
Total	100	66

Were the 66 jurors selected at random from the population?

Let p_i be the probability to select from cell i, $i = 1, 2, 3, 4$,

$H_0 : p_1 = .42, p_2 = .23, p_3 = .16, p_4 = .19$

Calculate expected: for example, expected in cell 1 are $66 \times 42/100 = 27.7$

Age	obs.	exp.
21-40	5	27.7
41-50	9	15.2
51-60	19	10.6
61 and over	33	12.5
Total	66	66

Calculate $X^2 \approx 61$;

degrees of freedom $= 4 - 1 = 3$,

p-value ≈ 0: reject H_0.

Example: Testing independence

Array: m rows, n columns,

p_{ij} probability for row i, column j

$p_{i\cdot}$ probability for row i, $p_{\cdot j}$ probability for column j

Let N_{ij} number of observations in cell (i, j)

$N_{i\cdot}$ number of observations in row i

$N_{\cdot j}$ number of observations in column j

N total number of observations

m.l.e.'s are

$\widehat{p}_{i\cdot} = \frac{N_{i\cdot}}{N}$

$\widehat{p}_{\cdot j} = \frac{N_{\cdot j}}{N}$

$\widehat{p}_{ij} = \frac{N_{ij}}{N}$

Under H_0 : rows and columns are independent:

$p_{ij} = p_{i\cdot}p_{\cdot j}$ for $i = 1, \ldots, m, j = 1, \ldots, n$

Θ_0 has dimension $(m-1) + (n-1)$

For the general alternative: as $\sum_{i,j} p_{ij} = 1$ have dimension $mn - 1$

So Chisquare test has degrees of freedom

$$mn - 1 - (m-1) - (n-1) = (m-1)(n-1)$$

Example

Are handedness and sex independent?

	Men	Women	Total
right-handed	934	1,070,	2,004
left-handed	113	92	205
ambidextrous	20	8	28
Total	1,067	1,170	2,237

Calculate expected for cell $(1, 1)$, under H_0:

percentage right-handed in sample is

$2,004/2,237 \times 100\% = 89.6\%$

Total number men in sample is 1,067

so expect $1,067 \times 89.6\% = 956$

other cells similar, gives

Expected counts:

	Men	Women	Total
right-handed	956	1,048	2,004
left-handed	98	107	205
ambidextrous	13	15	28
Total	1,067	1,170	2,237

Calculate $X^2 = 12$, have $(2-1) \times (3-1) = 2$ degrees of freedom

p-value ≈ 0.002, so reject H_0

Can create discrete data by binning continuous data

Rule of thumb: bin the data such that the expected number of observations in each bin is at least 5 *continuous: Kolmogorov-Smirnov Statistic*

i.i.d. observations from unknown c.d.f. F

Compare empirical cumulative distribution function to hypothethized cumulative distribution function F_0

H_0: $F = F_0$

H_1: not

test statistic

$$D_n = \sup_x |F_n(x) - F_0(x)|$$

Fact: The distribution of D_n does not depend on F_0; it is tabulated

What if the fit is not good?

1. Outliers

- investigate data collection process

- repeat analysis with outliers left out

- or: leave the lowest 100 α% and the highest 100 α% of the data out: *trimming*, leads to *robust statistics*

2. Transformations

coefficient of skewness is $\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^3}{S^3}$

should be close to 0 if data are normal

Tukey's ladder of powers lists transformatinos of the form

$$\ldots, x^{-2}, x^{-1}, x^{-1/2}, \log x, x^{1/2}, x, x^2, x^3, \ldots$$

For example, the transformations $y_i = \sqrt{x_i}$

$y_i = x_i^{1/3}$

$y_i = \log x_i$

$x^2, x^3, \ldots$ reduce negative skewness,

$\ldots, x^{-2}, x^{-1}, x^{-1/2}, \log x, x^{1/2}$ reduce positive skewness

Note: do not have to stick to integers for powers

Example: wages in India (see lecture)

Deciding optimal power:

- trial and error

- Box-Cox transformation (minimizes so-called *Beck score variance*)

3. Use tests without model assumptions

Tests just based on order statistics, such as *Wilcoxon's signed rank test*

leads to nonparametric procedures