

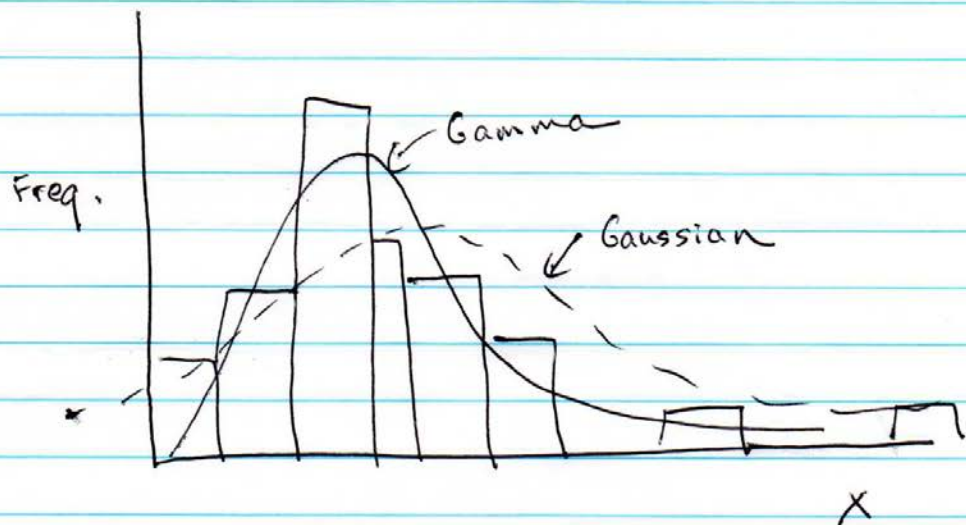
Goodness of fit

How good is a given distribution assumption to a set of data?

Ways to evaluate

- 1) Superimpose theoretical PDFs onto the histogram of observed data.

Fig. 4.15 From Wilks:



Not explicitly a quantitative way to evaluate goodness of fit.

2) More quantitative tests

χ^2 test

$$\chi^2 = \sum_{\text{bins}} \frac{(\# \text{ observed} - \# \text{ expected})^2}{\# \text{ expected}}$$

observed = observed value in histogram bin

expected = expected value from whatever theoretical distribution (gaussian, gamma, etc)

Read significant χ^2 statistic ~~off~~ off of Table (B.3 in Wilks)

Null hypothesis = ~~the~~ theoretical distribution fits observed.

Alternative = theoretical dist. does not fit observed.

Degrees of freedom

$$= \# \text{ of classes (bins)} - \# \text{ parameters} - 1$$

Another test which can be used to quantitatively evaluate goodness of fit is the Kolmogorov - Smirnov (K-S) test.

→ Compares empirical and theoretical CDFs.

Described in detail in Wilks p.p. 148 - 153.

K-S test is slightly more "elegant" test.

Non-parametric tests

So far we've discussed types of statistical tests which assume some theoretical distribution for the data (e.g. normal or t-distribution)

Test not requiring a distribution assumption are called non-parametric.

Conditions for use (Wilks)

- 1) Know or suspect that the parametric assumptions required for a test are not met (e.g. data are grossly non Gaussian)
- 2) Test statistic is complicated function of the data, and its sampling

distribution is unknown and/or cannot be derived analytically

~~Wil~~ Wilcoxon - Mann - Whitney Rank Sum test.

Assumption: Given 2 batches of indep. (no rel. in time and space), aim is to test whether 2 batches have the same "location" (or the analog of a different mean as compared to a standard test).

Null hypothesis: Two data batches are from the same distribution, so labeling of each value as belonging to one or another is arbitrary.

Alternative: Two data batches are not from the same distribution (because they rank differently).

Steps

1) Rank the members of each batch in the pooled distribution

Batch 1 = a_1, a_2, a_3, a_4, a_5

Batch 2 = b_1, b_2, b_3, b_4, b_5

Put the numbers in order (from lowest to highest)

	low $\longrightarrow$ High									
Pooled batch	a_1	b_5	a_3	b_3	b_4	a_2	b_1	a_4	a_5	b_2
	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$
Rank	1	2	3	4	5	6	7	8	9	10

2) Define the sum of the ranks for each distribution (R_1 , R_2)

~~R_1~~

$$R_1 = \text{Rank } a_1 + \text{Rank } a_3 + \text{Rank } a_2 + \text{Rank } a_4 + \text{Rank } a_5$$

$$R_1 = 1 + 3 + 6 + 8 + 9$$

$$R_2 = \text{Rank } b_5 + \text{Rank } b_3 + \text{Rank } b_4 + \text{Rank } b_1 + \text{Rank } b_2$$

$$R_2 = 2 + 4 + 5 + 7 + 10$$

If the null hypothesis is satisfied, then R_1/n_1 will not be that much different than R_2/n_2

Mann-Whitney U Statistic.

$$U_1 = R_1 - \frac{n_1}{2}(n_1 + 1)$$

$$U_2 = R_2 - \frac{n_2}{2}(n_2 + 1)$$

$$\mu_u = \frac{n_1 n_2}{2} \rightarrow \text{Analogous to "Mean value"}$$

$$\sigma_u = \left[\frac{n_1 n_2 (n_1 + n_2 + 1)}{12} \right]^{1/2}$$

To compute the statistic analogous to the z-statistic

$$z_{u_1} = \frac{u_1 - \mu_u}{\sigma_u}$$

$$z_{u_2} = \frac{u_2 - \mu_u}{\sigma_u}$$

→ Evaluate the statistic on a normal distribution.

Wilks gives the example (very pertinent to a lot of students in this class) of lightning strikes from thunderstorms (p. 158)

Problem: Want to determine if there is a statistically significant difference in lightning from storms that are seeded vs. not seeded.

A non-parametric test would be appropriate here because lightning strikes would probably not follow a normal distribution.