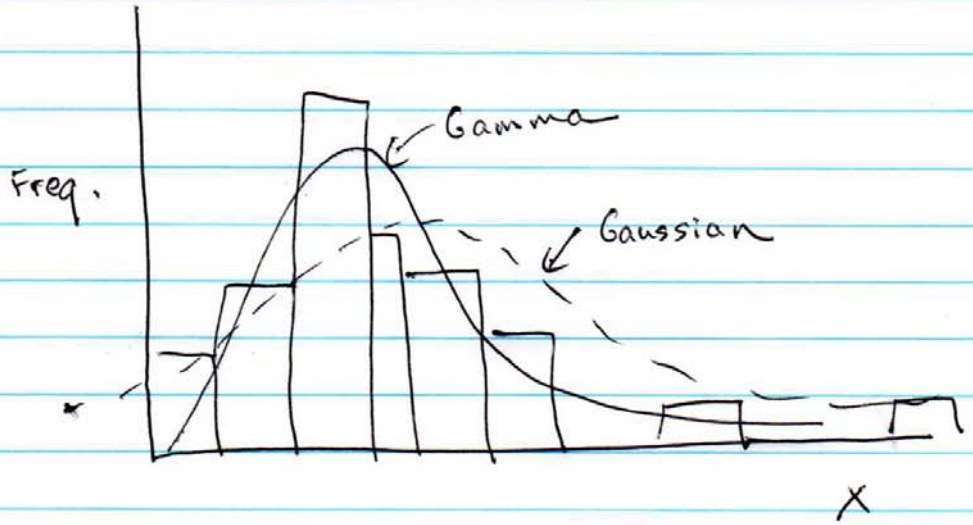
Goodness of fit

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 of Table (B.3 in Wilks)

Null hypothesis = theoretical distribution fits observed.

Alternative = theoretical dist. does not fit observed.

Degrees of freedom
= # of classes (bins) − # 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

* 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 ⟶ High

Pooled batch =	a_1,	b_5,	a_3,	b_3,	b_4,	a_2,	b_1,	a_4,	a_5,	b_2
	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
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)

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

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

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

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

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.

$$u_1 = R_1 - \frac{n_1}{2}(n_1+1)$$

$$u_2 = R_2 - \frac{n_2}{2}(n_2+1)$$

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

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

To compute the statistic analogous to the z-statistic

$$z_{u_1} = \frac{u_1 - \mu_u}{\sigma_u} \qquad z_{u_2} = \frac{u_2 - \mu_u}{\sigma_u}$$

→ 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.

Wilcoxon signed rank test is similiar, except it ranks the differences between paired values.

Resampling or Monte-Carlo approach

Idea: Construct artificial datasets from a collection of real data, by resampling the obervations in a manner consistent with the null hypothesis.

This is used to construct a null distribution, against which the alternative (interesting) hypothesis can be evaluated

You necessarily need a computer to do this, which can randomize the data.

Permutation: Samples are drawn from a distribution without replacement so each individual observation is represented only once.

Bootstrap: Samples are drawn from a distribution with replacement so each individual observation may be represented multiple times

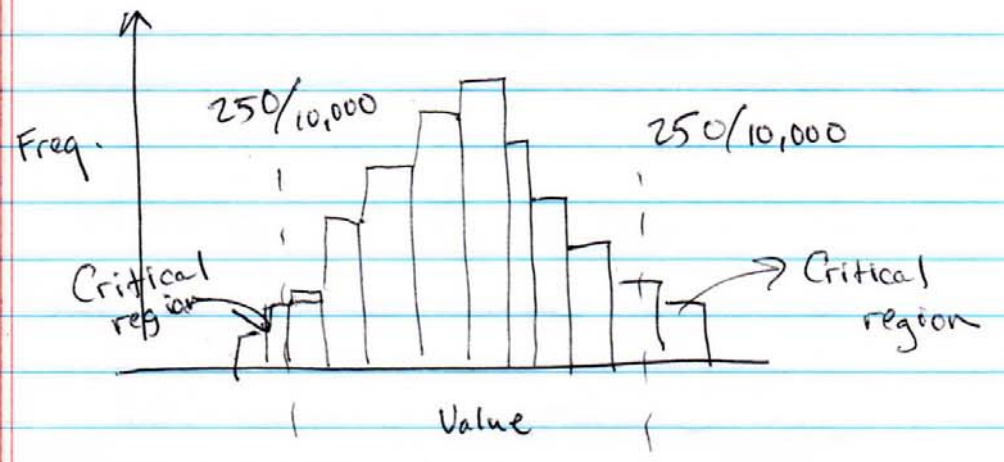
Say we want to sample slips of paper from a hat. Total of n slips of paper. To construct one sample (n, draws) (where $n_1 < n$)

Permutation approach: Each slip of paper is drawn without replacement to construct a sample size of n_1 draws.

Bootstrap approach: After the draw of each slip, it is put back into the hat. Draw again till reach n_1 draws. → less restrictive null hypothesis.

Whatever approach is used, the idea is to generate a distribution from a given amount of estimates. Usually the number of estimates is fairly large to create a robust distribution.

Say 10,000 estimates



In a estimate with 10,000 the critical region occurs near the tail of the distribution. In the critical region, if the original sample falls there it is statistically significant.

Caveat: When using these methods, first need to make sure that the data are not autocorrelated in time (i.e. samples are independent).

Returning to the lightning example discussed earlier with respect to the Mann-Whitney U statistic. (p. 165 Wilks)

Lightning counts from
$n_1 = 12$ seeded storms
$n_2 = 11$ unseeded storms

Compute L-scale statistic, basically gives a measure of spread in a given sample:

$$\lambda_2 = \frac{(n-2)!}{n!} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} |x_i - x_j|$$

"x" = * ≙ strikes

$x_i - x_j$ summation term gives an overall measure of the difference between all possible pairs in the sample size. Equivalent to variance (F-test, χ^2 test)

Compute this statistic for seeded and unseeded storms.

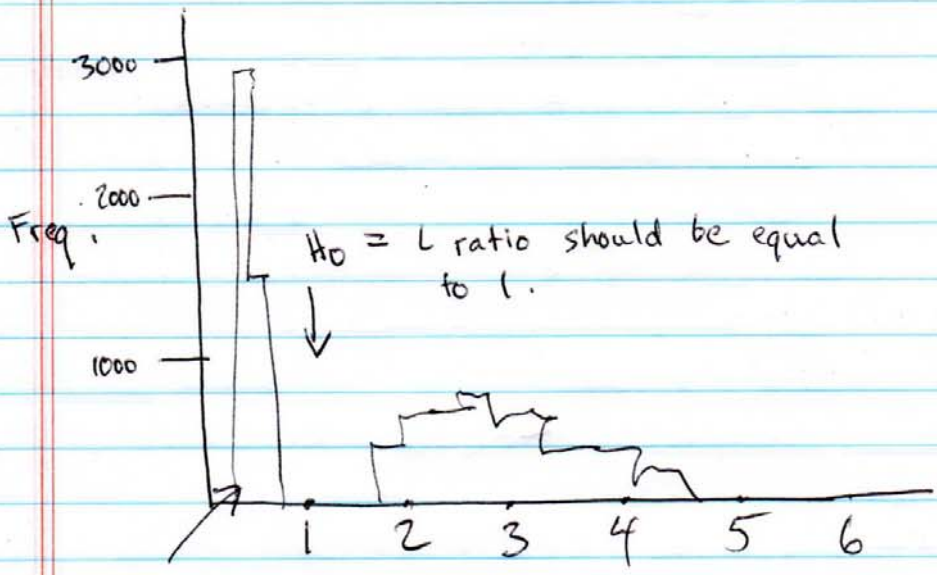
$$L_{scale} = \frac{\lambda_2(\text{seeded})}{\lambda_2(\text{unseeded})} \rightarrow \text{test-statistic}$$

- Seeded storms more variable in lightning if greater than one
- Less variable in lightning if less than one.

To construct null distribution:

- Randomize the entire sample of n (23) storms, pick n_1 and n_2.
  (permutation → because not replacing after draw)
- Compute the test statistic
- Do this a ~~gi~~ sufficient number of times to generate a frequency distribution (like 10,000 times)

Get Fig 5.7.



Observed ratio falls here (so it is significant)

Smaller than all but 49 of 10,000 permutation tests.