Comparison of Distributions

Week 4

Some quotes from astrophysicists

- “Understanding data better is always an unsolved problem in Astrophysics”
  Bill Press, astro-ph/9604126

- “We identify the probability of the data given the parameters as the *likelihood* of the parameters given the data. This identification is entirely based on intuition. It has no formal mathematical basis in and of itself; as we already remarked, statistics is *not* a branch of mathematics.”
  Bill Press, Numerical Recipes 1992 edition, Chapter 15 on maximum likelihood estimation.

- You may ask “What can a hard headed statistician offer to a starry eyed astronomer?”. The answer is “Plenty”. One normally associates statistics with large numbers, and astronomy is full of large numbers…I have every reason to believe that increased interaction between statistics and astronomy will be to the benefit of both subjects.

  Jayant Narlikar to C.R. Rao

Introduction

- Two samples are drawn from the same distribution or whether two sets of measurements imply a difference in measured quantity.
- A sample is consistent as been drawn from a known distribution.
- Two sets of measurements with different measurement errors are drawn have the same mean.

Use data to compute an approximate statistic and compare the data-based value to its expected distribution. The expected distribution is evaluated

by assuming that the *null hypothesis is true.*

We now discuss various non-parametric methods for comparing distributions.

Kolmogorov-Smirnov (K-S) tests

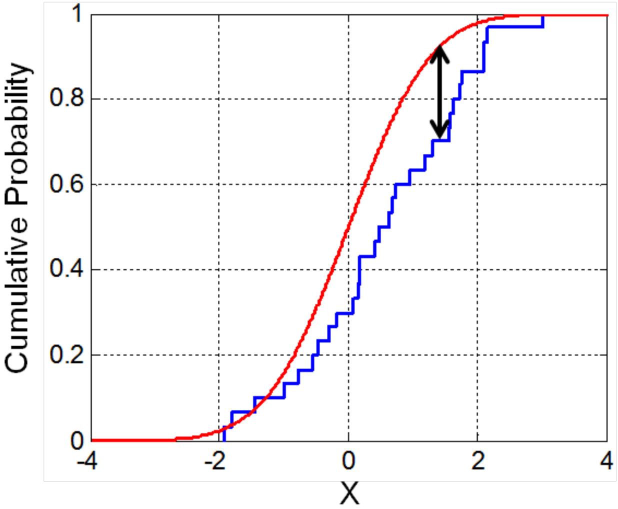
- K-S test compares cumulative distribution function of two samples.
- Consider two samples $\{x1_i\}$ where $i=1,\ldots.N_1$ and $\{x2_i\}$ where $i=1,\ldots.N_2$

Sort the sample and divide the rank of x_i by the sample size to get $F(x_i)$

$F(x)$ is a step function that increases by 1/N at each data point $0 \leq F(x) \leq 1$.

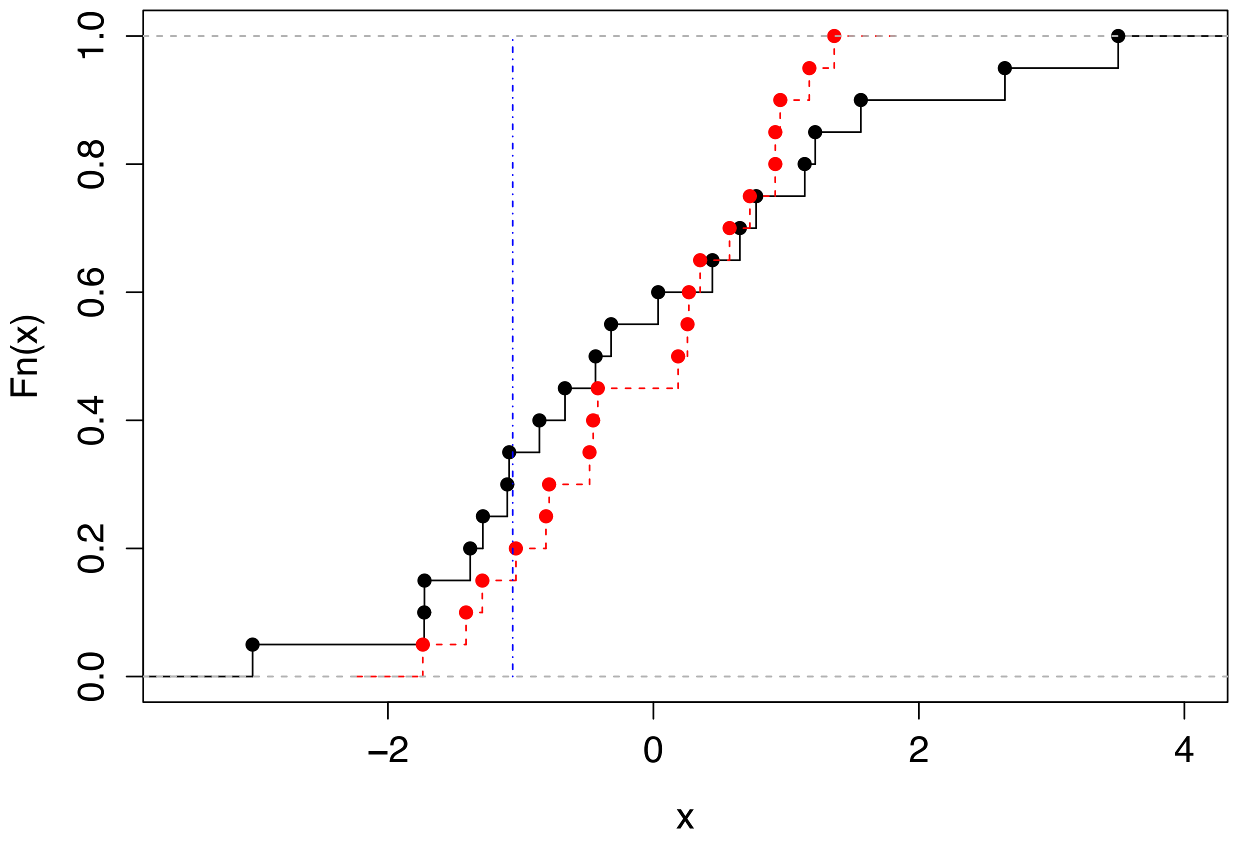
K-S test is based on the maximum distance of the cumulative distribution function $F_1(x_1)$ and $F_2(x_2)$.

$D = \max | F_1(x1) - F_2(x2) |$



Source: wikipedia

Example of 2-sample KS test



Jessica Cisewski notes at 2016 PSU astrostatistics summer school

Null Hypothesis of KS-test

Qt : How often would the value D computed from the data arise by chance if the two samples were drawn from the *same* distribution (null hypothesis in this case).

The probability of obtaining by chance a value D larger than the observed value is given by the function

$$Q_{KS}(\lambda) = 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2\lambda^2}$$

$$\lambda = (0.12 + \sqrt{n_e} + \frac{0.11}{\sqrt{n_e}})\text{D}$$

$$n_e = \frac{N_1 N_2}{N_1 + N_2}$$

One-Sample K-S test

- Qt : Is the measured f(x) consistent with a known reference distribution function h(x)

This is called “one-sample” K-S test and is special case of “two-sample” K-S test discussed previously. In this case $N_2=\infty$

A small value of Q_{ks} indicates that it is unlikely at given confidence level α that the data summarized by f(x) are drawn from h(x).

K-S test is sensitive to the location, scale and shape of the underlying distribution. It is insensitive to reparametrization of the data points.

K-S tests for histograms can be found in arXiv:0804.0380

K-S test in Python

- Lookup `kstest, ks_2samp, ksone in scipy.stats`

```
import numpy as np
from scipy import stats
vals = np.random.normal(loc=0,scale=1,size=1000)
stats.kstest(vals,”norm”)
```

```
0.025, 0.529
D-value = 0.025
p-value = 0.529
```

Critic of usage of K-S test in astrophysics

- See

https://asaip.psu.edu/Articles/beware-the-kolmogorov-smirnov-test

Other non-parametric tests:

Anderson-Darling test

Cramer-Von Mises Test

Watson Test

Gini Coefficient (based on cumulative distribution function) : measures the deviation of a given cumulative distribution function from that expected for a normal distribution.

$$G = 1 - 2\int_{x_{min}}^{x_{max}} F(x)dx$$

Examples of KS test usage in astrophysics

- Too many examples (homework assignment for you all)

Kuiper test

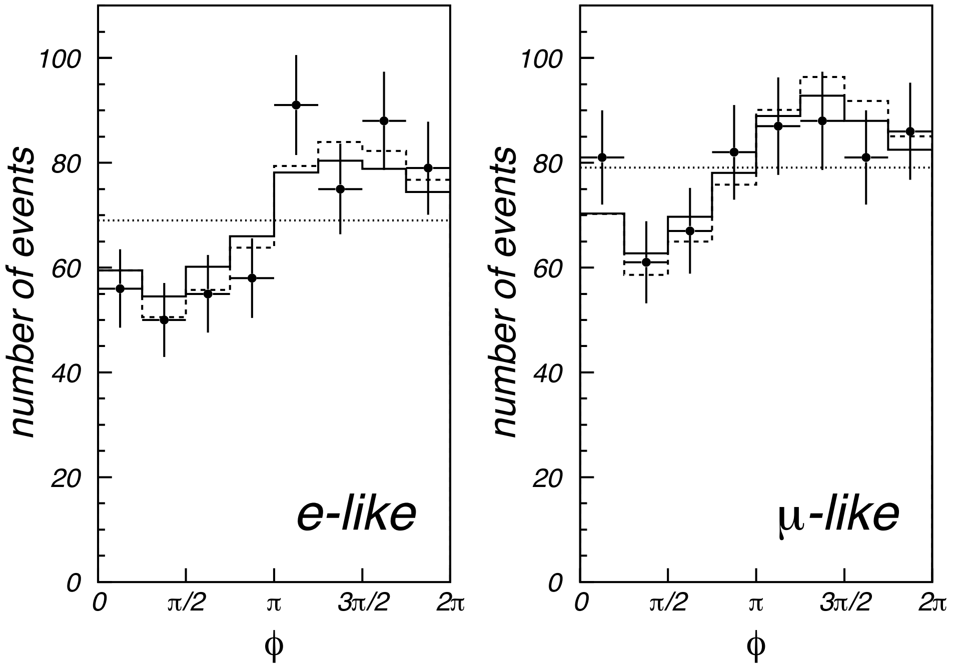
- Simple variant of K-S test to treat distributions defined on a circle. It is based on the statistic

$$D^* = \max\{F_1(x1) – F_2(x2)\} + \max\{F_2(x1) – F_1(x2)\}$$

For distributions on a circle D* is invariant to location of origin. Good test for comparing longitude distribution of two astronomical samples.

Null hypothesis probability calculated in the same way as K-S test.

Example of Kuiper Test usage



astro-ph/9901139

East-West asymmetry of atmospheric neutrino flux in Super-Kamiokande

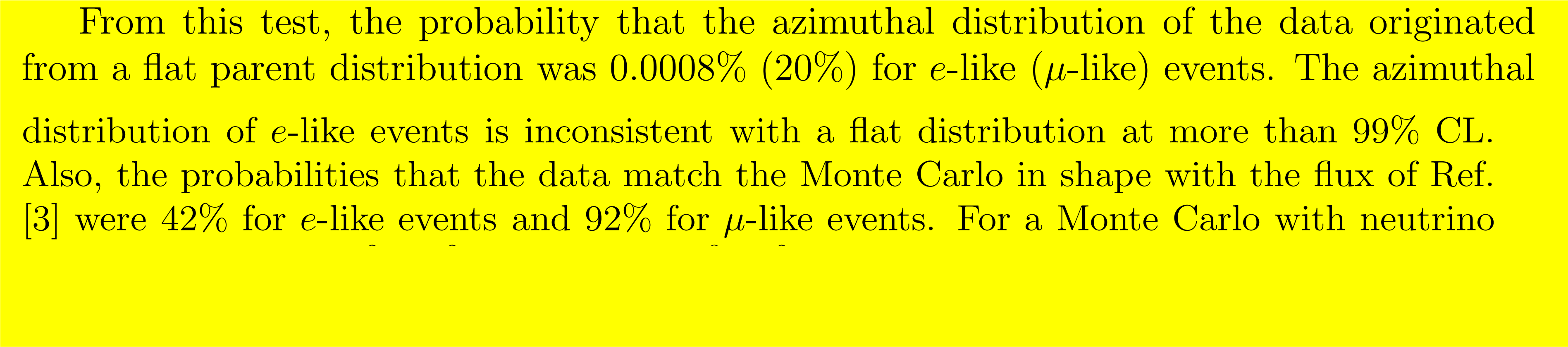
A quantitative comparison between the data and Monte Carlo was performed using a Kuiper test [8,9]. This test calculates a binning and starting-point free probability that observed data are the result of an assumed distribution. The Kuiper statistic V is defined as:

$$V = \max_{0<\phi<2\pi}[S_N(\phi) - P(\phi)] + \max_{0<\phi<2\pi}[P(\phi) - S_N(\phi)],$$

where ϕ is the azimuthal angle, $S_N(\phi)$ is a cumulative probability function from data and $P(\phi)$ is the one from Monte Carlo. The significance is obtained from the statistic $V^* = V(\sqrt{n} + 0.155 + 0.24/\sqrt{n})$, and defined as

$$Prob = 2\sum_{j=1}^{\infty}(4j^2V^{*2} - 1)\exp(-2j^2V^{*2}),$$

where n is number of events.



From this test, the probability that the azimuthal distribution of the data originated from a flat parent distribution was 0.0008% (20%) for e-like (μ-like) events. The azimuthal distribution of e-like events is inconsistent with a flat distribution at more than 99% CL. Also, the probabilities that the data match the Monte Carlo in shape with the flux of Ref. [3] were 42% for e-like events and 92% for μ-like events. For a Monte Carlo with neutrino

Anderson-Darling Test for Gaussianity.

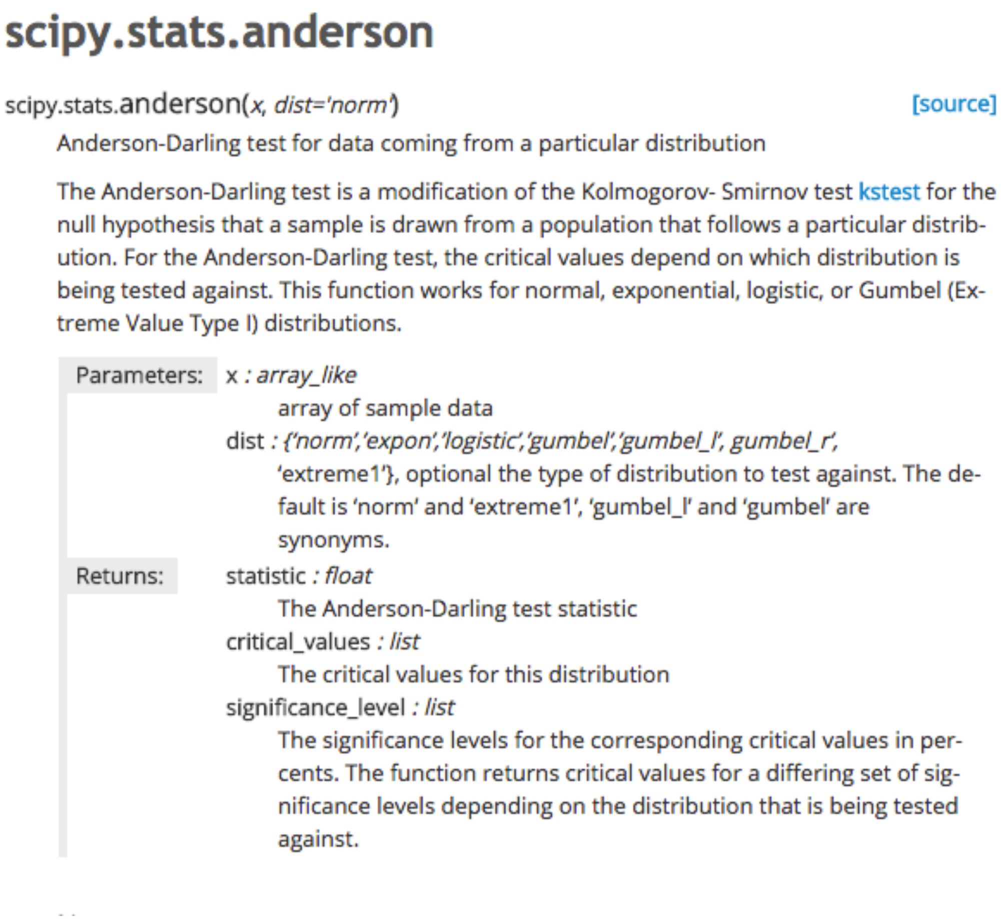
Qt: Is the *shape* of measured f(x) consistent with a Gaussian distribution?
Aim is to predict distribution of test statistic when null hypothesis is true.

$$A^2 = -N - \frac{1}{N}\sum_{i=1}^{N}[(2i-1)\ln(F_i) + (2N-2i+1)\ln(1-F_i)]$$

Where F_i is the ith value of the cumulative distribution function of z_i defined as :

$$z_i = \frac{x_i - \mu}{\sigma}$$

Anderson-Darling test in SciPy



scipy.stats.anderson

scipy.stats.**anderson**(*x*, *dist='norm'*) [source]

Anderson-Darling test for data coming from a particular distribution

The Anderson-Darling test is a modification of the Kolmogorov- Smirnov test kstest for the null hypothesis that a sample is drawn from a population that follows a particular distribution. For the Anderson-Darling test, the critical values depend on which distribution is being tested against. This function works for normal, exponential, logistic, or Gumbel (Extreme Value Type I) distributions.

Parameters:

x : *array_like*

array of sample data

dist : *{‘norm’,’expon’,’logistic’,’gumbel’,’gumbel_l’, gumbel_r’,* ‘extreme1’}, optional the type of distribution to test against. The default is ‘norm’ and ‘extreme1’, ‘gumbel_l’ and ‘gumbel’ are synonyms.

Returns:

statistic : *float*

The Anderson-Darling test statistic

critical_values : *list*

The critical values for this distribution

significance_level : *list*

The significance levels for the corresponding critical values in percents. The function returns critical values for a differing set of significance levels depending on the distribution that is being tested against.

This is also coded in statsmodels

Lookup
statsmodels.stats.diagnostic.normal_ad

Anderson-Darling Test for Gaussianity

```
From scipy import stats
X = np.random.normal(0,1,size=1000)
A,crit,sig = stats.anderson(X,'norm')
```

A-D test in scipy.stats can also be used to test for exponential, logistic, Gumbel distribution

Note that there is also a k-sample -sample A-D test called `anderson_ksamp` in scipy

Shapiro-Wilk test.

$$W = \frac{\left(\sum_{i=1}^{N} a_i R_i\right)^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2}$$

R_i is the rank of the sample and a_i is the expected values of the order statistics for random variables sampled from the normal distribution

Lookup `stats.shapiro` in scipy . A value of W close to 1 indicates that distribution is entirely gaussian

Outliers in Gaussian Distributions

- Compare sample standard deviation and σ_G
- For a Gaussian distribution when N > 100, s/σ_G follows a Gaussian distribution with μ~1 and $\boldsymbol{\sigma} = 0.92/\sqrt{N}$
- Difference between mean and median drawn from a Gaussian distribution follows a Gaussian distribution with μ~0 and σ~0.76/ $\sqrt{N}$

Define $Z_1 = 1.3\ |\mu - q_{50}|\ \sqrt{N}\ /s$ and $Z_2 = 1.1 \left|\frac{s}{\sigma_G} - 1\right| \sqrt{N}$

obey Gaussian distribution with mean $= Z_1$ and std. deviation $= Z_2$

Cheat-Sheet of Tests for Gaussianity

- For data which depart from Gaussian distributions :

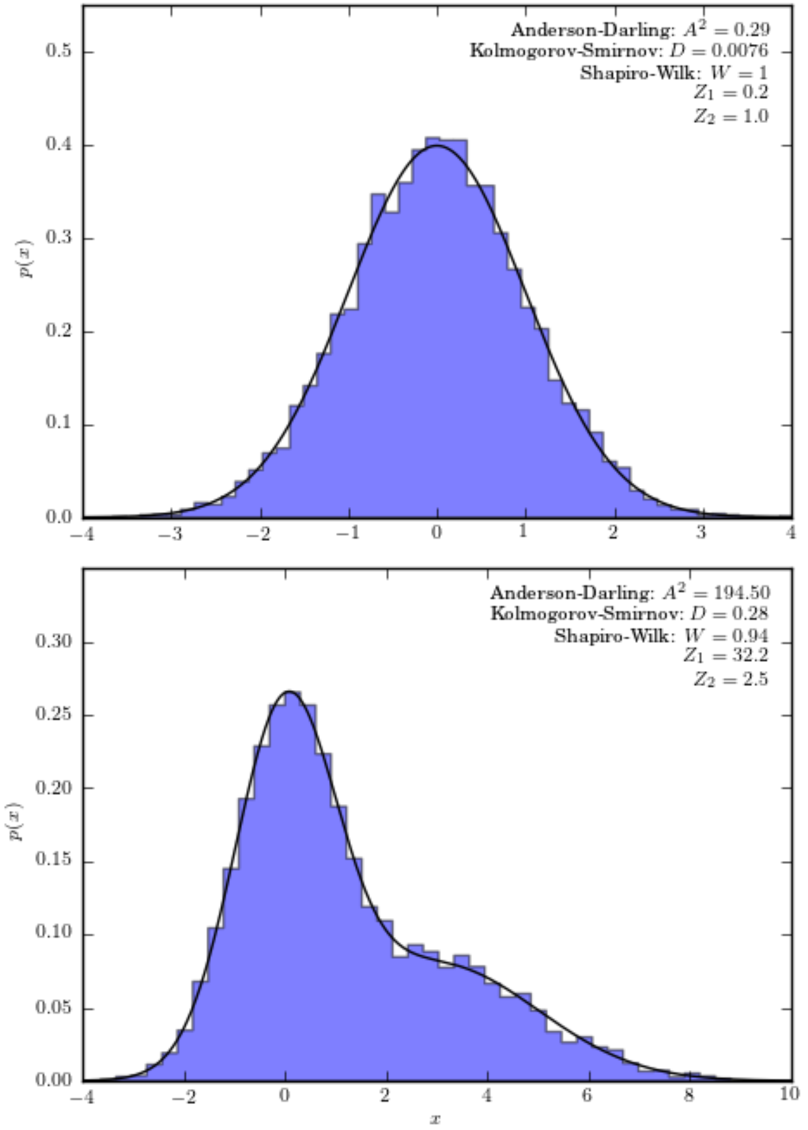
  A-D test (A^2) >1

  K-S test (D) > 1/sqrt(N)

  Shapiro-Wilk’s test (W) < 1

  Z_1, Z_2 > (many times) $\boldsymbol{\sigma}$

- For an example from astrophysics (dynamical state of galaxy groups) see arXiv:0908.0938



Normal Distribution

Combination of two normal distributions

Astroml figure 4.7

Modeling Non-Gaussianity

- In case our empirical distributions fail tests for Gaussianity, but if there is no strong motivation for choosing a specific alternative distribution.
- Model Non-Gaussian distribution by Gram-Charlier series

$$h(x) = \mathcal{N}(\mu, \sigma) \sum_{k=0}^{\infty} a_k H_k(z)$$

where

$$z = \frac{x - \mu}{\sigma}$$

$H_k(z)$ are Hermite polynomials, which form an orthonormal set

and

$$a_k = \frac{1}{k!} \int_{-\infty}^{\infty} f(x) H_k(x) dx$$

RC Boston 1971, IEEE

Mann-Whitney Test/U Test/Wilcoxon rank sum tests

○ Used to check if two datasets are drawn from distributions with different location parameters (or different means). (In case the two datasets are drawn from Gaussian distributions, you can use t-test).

○ Calculation of U is obtained from the ranks of the two samples after concatenating them (will skip details of calculation. consult wikipedia)

```
>>> import numpy as np
>>> from scipy import stats
>>> x,y=np.random.normal(0,1,size=(2,1000))
>>> stats.mannwhitneyu(x,y)          (See also stats.ranksums)
MannwhitneyuResult(statistic=488708.0, p-value=0.19094553275490711)
```

U-test expected result is close to the expected N1N2/2 if they are drawn from the same distribution with Same mean. See https://data.library.virginia.edu/the-wilcoxon-rank-sum-test/ for an illustrative example

Wilcoxon signed-rank test

Special case of comparing the means of two data sets is when the data sets have the same size ($N_1 = N_2 = N$) and data points are paired. Non-parametric test used to compare means of two distributions is the Wilcoxon signed-rank test. This test cannot be used when data points are not equal

- Calculate $y_i = x1_i - x2_i$ and exclude all values with $y_i=0$
- Order the new sample by $|y_i|$ resulting in the rank R_i for each pair
- Each pair is assigned $\Phi_i=1$ if $x1_i > x2_i$ and 0 otherwise.
- Calculate W_+ as follows

$$W_+ = \sum_i^m \Phi_i R_i$$

All ranks with $y_i > 0$ are summed

Similarly calculate W_- and Wilcoxian signed rank test (T) is smaller of the two.

p-value of Wilcoxon signed-rank test

For value of n > 20 , behaviour of T can be approximated by Gaussian distribution $\mathcal{N}(\mu, \sigma)$

where :

$$z = \frac{T - \mu}{\sigma}$$

$$\mu = \frac{N(2N + 1)}{2}$$

$$\sigma = N\sqrt{\frac{2N + 1}{12}}$$

Wilcoxon signed-rank test in scipy

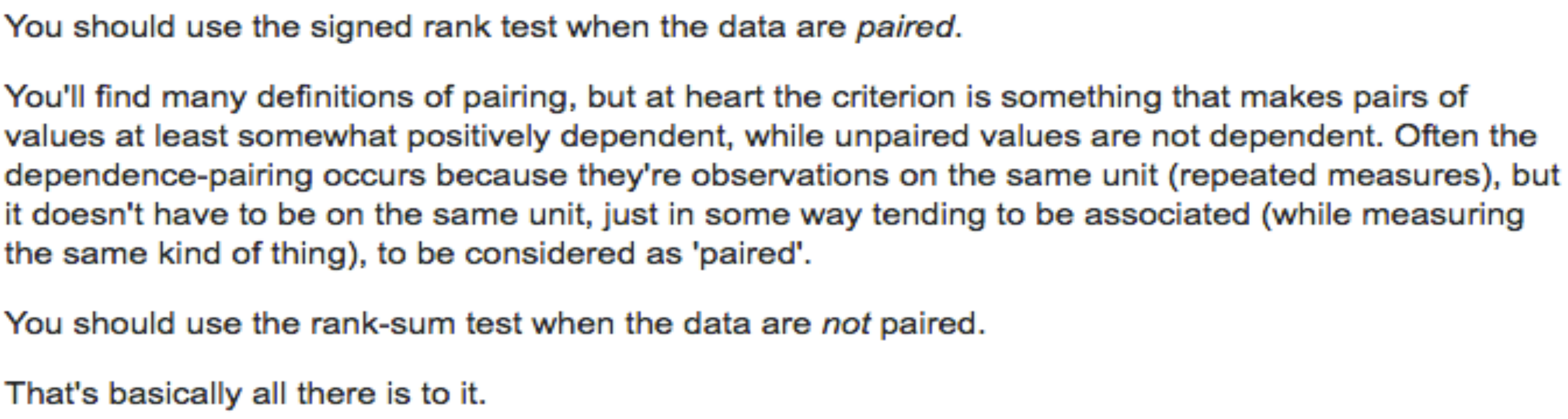
```
>>> import numpy as np
>>> from scipy import stats
>>> x,y = np.random.normal(0,1,size=(2,1000))
>>> T,p= stats.wilcoxon(x,y)
>>> T,p  # T is the value of smaller value of W+ and W–
(243583.0, 0.46552103910817078)
```

Difference between Wilcoxon signed-rank test and Wilcoxon rank-sum test

19

12

I was wondering what the theoretical difference is between the Wilcoxon Rank-Sum Test and the Wilcoxon Signed-Rank Test using paired observations. I know that the Wilcoxon Rank-Sum Test allows for different amount of observations in two different samples, whereas the Signed-Rank test for paired samples does not allow that, however they both seem to test the same in my opinion. Can someone give me some more background / theoretical information when one should use the Wilcoxon Rank-Sum Test and when one should use the Wilcoxon Signed-Rank Test using paired observations?



You should use the signed rank test when the data are *paired*.

You'll find many definitions of pairing, but at heart the criterion is something that makes pairs of values at least somewhat positively dependent, while unpaired values are not dependent. Often the dependence-pairing occurs because they're observations on the same unit (repeated measures), but it doesn't have to be on the same unit, just in some way tending to be associated (while measuring the same kind of thing), to be considered as 'paired'.

You should use the rank-sum test when the data are *not* paired.

That's basically all there is to it.

More details available in http://tinyurl.com/j37sts5

Parametric Methods for comparing distributions (t-test)

- If two datasets are drawn from Gaussian distributions with different means, but same sigma and sigma is known , you can calculate p-value of the difference in means using error propagation.

- If we do not know sigma, calculate $\Delta = \overline{x1} - \overline{x2}$ and $M_s = \dfrac{\Delta}{s_\Delta}$ where

$s_\Delta = \sqrt{s_1^2/N_1 + s_2^2/N_2}$ follows Student's t-distribution. DOF = N_1+N_2-2

For large samples, this asymptotes to the case of distributions with known sigma (or Gaussian distribution)

If the two datasets are drawn from two underlying distributions with different variances, then appropriate test is called Welch's t-test.

t-test in Python

- Different variants of t-test in Python called `ttest_rel, ttest_ind,  and ttest_1samp`
  See documentation for details.

```
>>> import  numpy as np
>>> from scipy import stats
>>> x, y = np.random.normal(size=(2,1000))
>>> t,p = stats.ttest_ind(x,y)
>>> t,p
(0.93039549523650522, 0.35227874770046763)
```

Comparison of Gaussian variances using F-test

- F test is used to compare variances of two unspecified distributions. The null hypothesis is that variances of the two samples are equal and the statistics is based on the ratio of the sample variances

$F = s_1^2/s_2^2$ follows Fisher F distribution with $d_1 = N_1-1$ and $d_2=N_2-1$

```
>>> import  numpy as np
>>> from scipy import stats
>>> x, y = np.random.normal(size=(2,10000))
>>> F,p = stats.f_oneway(x,y)
```

- p
- 0.85157563615482457
- F
- 0.035010462229159989

Efficacy of these tests

In astrophysics a comparative study of these tests for detecting non-gaussianity was done in https://arxiv.org/abs/0908.0938

Choosing histogram bin width

- Three different rules for choosing bin width:
  - ➢Scott Rule
  - ➢ Freedman rule
  - ➢ Knuth Rule

```
 from astroML.plotting import hist
  hist(x,bins=‘freedman’) # can also choose knuth or
scott
```