Adolfo Carvalho
ASTR 600

Non-Parametric Statistics

Problem 1: Kolmogorov-Smirnov tests and noise

In astronomy, we often use the Kolmogorov-Smirnov test to determine how closely related the distributions underlying our data are. However, data tend to be messy, especially astronomical data, containing various kinds of noise. This begs the question: how does this noise affect the efficacy of the K-S test?

The two data sets below have been sampled randomly from the same Poisson distribution ($\lambda = 1000$). Then, I added some normally distributed noise ($\mu = 0,\ \sigma = 10$) to each of them.

Perform a single sample K-S test on the two raw data sets and determine the confidence level to which one can say they were in fact both sampled from this distribution.

Now, perform a two-sample K-S test on the two noisy data sets, determine the confidence level to which one can say they were in fact both sampled from this distribution. Compare this now to the two single sample tests, and a two-sample test of the two raw data sets.

Does the added noise "break" our K-S test? How much noise is necessary to break the test? In order to generate more noise, simply increase the σ value in the normal distribution centered on 0, sample that distribution, and add the vector to the raw data. Also, perform a K-S test to compare the noisy data set to the original distribution. Consider that all of this noise-adding procedure amounts to changing the underlying distribution from a pure Poisson to a Poisson + Gaussian distribution. How might all of this change if we changed our sample number?

Data Set 1	Data Set 2	Data Set 1 with Noise	Data Set 2 with Noise
1004	977	1008.0275	970.5421
1038	1015	1040.833333	1021.0279
1020	1019	1023.4055	1007.6187
1071	991	1080.7255	994.7194
1046	1022	1040.1561	1019.7227
1000	1012	1002.4139	1001.0680
1005	1076	1000.1803	1081.0662
972	982	973.8036	995.3939
898	1027	906.3602	1033.7239
1021	967	1020.7095	979.6040
1054	1039	1056.5591	1048.2999
1009	981	1023.3009	987.6447
1003	971	1006.2810	975.7149
946	970	961.2034	966.8071
979	1037	970.5180	1030.0379

(Note: All values can be easily imported into R by first copying the tables into excel and then importing the excel file)

Adolfo Carvalho
ASTR 600

Problem 2: KS test: drawing parameters from the data

We have been discussed many times in class and been told repeatedly that when performing a single sample K-S test on a sample set, you cannot derive the parameters for your proposed distribution from the sample you are testing. But doing so is usually a very convenient way to determine the necessary parameters for the distribution we want to compare to. So how bad is it really?

Use a single sample K-S test to compare the data below to a gaussian distribution with $\mu = 0,\ \sigma = 1$, and determine the confidence level to which you can say the data come from this gaussian. Then, use the mean of the data, $\bar{x}$ and the variance, s, as estimates of the μ and σ, perform a single sample K-S test comparing the data to this new distribution, and determine the confidence level to which the data might have come from a distribution like that.

How do the confidence levels compare? Is it generally safe, then, to use the data to approximate our desired model parameters?

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

Trial	Value
1	0.0851
2	-0.2958
3	0.6849
4	0.5964
5	0.2936
6	-0.3275
7	0.0394
8	0.4627
9	0.8396
10	-1.5447
11	-0.3034
12	1.4083
13	-2.553
14	-0.9737
15	-1.524

(Note: All values can be easily imported into R by first copying the tables into excel and then importing the excel file)

Adolfo Carvalho
ASTR 600

Problem 3: The Weather and the Houston Marathon

The marathon is often considered one of the most grueling physical challenges one can undertake. It seems logical then that warmer temperatures would increase a marathoner's chances of overheating and dropping out. Is this really the case, though? How sensitive are marathoners to the weather? And how might running a half marathon in similar temperatures compare?

Below are the numbers of finishers for the Houston Marathon and Houston Half Marathon each year from 2008 to 2017, and the temperature on race day. The columns you should use are the "Corrected" columns, as the finisher numbers in those columns have been corrected for the generally increasing finisher number due to increased popularity of the event.

Use the Spearman rank test and the Kendall Tau test to determine if the numbers of finishers and the race day temperatures are significantly correlated.

Year	Marathon Finishers	Corrected Marathon Finishers	Half Marathon Finishers	Corrected Half Marathon Finishers	Temperature
2008	5519	5519	8226	8226	65
2009	5349	5139	8334	7965	72
2010	6287	5867	9918	9180	62
2011	6852	6222	9313	8206	74
2012	7614	6774	9374	7898	68
2013	6530	5480	10065	8220	48
2014	6945	5685	10500	8286	70
2015	7004	5534	11664	9081	66
2016	7802	6122	11079	8127	57
2017	7154	5264	11416	8095	75

Data courtesy of the Houston Marathon Media Booklet, with corrections performed by simple linear fit and subtraction

(Note: All values can be easily imported into R by first copying the tables into excel and then importing the excel file)