Statistics for Astronomers: Lecture 15, 2020.12.08

Prof. Sundar Srinivasan

IRyA/UNAM

Review

Hypothesis testing – F-test and one application.

Nonparametric statistics – the Kolmogorov-Smirnov test.

Jupyter demos

1. ▸ Download this Jupyter notebook.
2. ▸ Navigate to Colaboratory.
3. Sign in
4. Click on "Upload" and upload the notebook you downloaded in step 1.

Kolmogorov-Smirnov test

One-sample test (implementation: `scipy.stats.kstest`):

Given: sample of size N. Question: Is the sample drawn from a particular distribution?

H_0: yes. H_A: No (two-sided) or *greater*/*less* than CDF of distribution (one-sided).

Two-sample test (implementation: `scipy.stats.ks_2samp`):

Given: samples of sizes N_1, N_2. Question: Are samples drawn from the same distribution?

H_0: yes. H_A: No (two-sided).

KS statistic: maximum distance between CDFs that are being compared. One-sample case:

$D_{\mathrm{KS}} = \max\left|CDF_{\mathrm{model}}(x) - ECDF(x)\right|$ (two-sided). $D_{\mathrm{KS}} = \max\left(CDF_{\mathrm{model}}(x) - ECDF(x)\right)$ (one-sided).

Let's go to the Jupyter notebook again...

Advantages: No binning! Small samples! More powerful for intermediate-size samples! Can also work as a one-tailed test (see `sscipy.stats.kstest`).

Disadvantages: Not sensitive to differences in the tails. Doesn't care about $\#dof$.

Beware the KS test!

The Anderson-Darling test: an alternative to the KS test

KS test less sensitive to relative deviations in the distribution tails ($F(x) \approx 0$ or $F(x) \approx 1$, where $F(x)$ is usually slowly-varying).

One alternative – consider deviation over entire range instead of just maximum deviation.

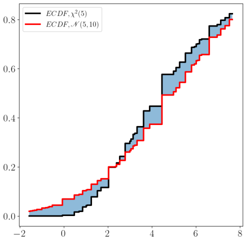
Anderson-Darling statistic: $$A^2 \equiv \int_0^1 dF(x) \left[\overbrace{\frac{\hat{F}_N(x) - F(x)}{\sqrt{\mathrm{Var}[\hat{F}_N(x)]}}}^{\text{standardised}} \right]^2$$ (A quadratic ECDF test).

Recall: $\mathbb{E}[\hat{F}_N(x)] = F(x)$; $\quad \mathrm{Var}[\hat{F}_N(x)] = \dfrac{F(x)(1 - F(x))}{N}$

$F(x)(1 - F(x)) \approx 0$ when $F(x) \approx 0$ or $F(x) \approx 1$.

$\implies$ more weight on observations in tails of the distribution.

In most cases when applying the KS test, also use the AD test.



Wilcoxon-Mann-Whitney U-test

KS and AD tests: nonparametric version of F-test (tests whether scale parameters are similar).

The U-test is a nonparametric version of the two-sample (independent) t-test.

Tests for whether the samples have similar location parameters.

Ideal when the samples are drawn from a general unspecified distribution.

H_0: samples are drawn from distribution with the same location parameter.

Choose H_A carefully:

sample 1 has different location parameter from sample 2 (two-sided).

sample 1 has location parameter smaller than sample 2 ("less").

sample 1 has location parameter larger than sample 2 ("greater").

Python implementation: `scipy.stats.mannwhitneyu`. Interpret carefully!

```
ustat, pvalue = mannwhitneyu(sample1, sample2, alternative = 'less')
```

If `pvalue` $> \alpha$, since `alternative = 'less'`,

interpretation: not enough evidence that the mean of sample 1 $<$ mean of sample 2.

Visualising data

Five-number summary

Five number summary of a dataset of size N: $x_{(1)}, q_{25}, q_{50}, q_{75}, x_{(N)}$.

q_{50} = median, a robust location measure.

Interquartile range, $\mathrm{IQR} = q_{75} - q_{50}$ is a robust scale measure

Encloses the central 50% of the sample.

(for a normal distribution, $\mathrm{IQR} \approx 1.349\sigma$).

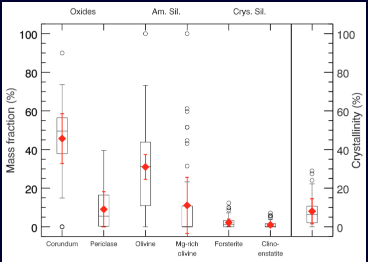
Compare $\bar{x}$ to q_{50} to check for asymmetric/skewness.

One way of visualising data, the box plot, uses the five-number summary.

Box plot, Box-and-whisker plot, or candlestick plot

A non-parametric way to visualise the data distribution, without binning.

1. Identify median with a horizonal line. In addition, can show the mean with a dotted line. Compare the two $\rightarrow$ skewness.
2. Draw a box enclosing the central 50% of the data (the box edges are q_{25} and q_{75}).
3. From each box edge, extend a "whisker" of length $\frac{3}{2}\mathrm{IQR}$. These whiskers display the tails of the distribution.
4. Any data outside the box-and-whisker region are outliers and can be displayed with individual symbols.
5. Mild $\left(\frac{3}{2} \leq \frac{|x - q_{50}|}{\mathrm{IQR}} < 3\right)$ and extreme $\left(\frac{|x - q_{50}|}{\mathrm{IQR}} \geq 3\right)$ outliers can be also distinguished.



Comparing relative locations and sizes of boxes $\rightarrow$ comparing distributions.

Variation on box plot: violin plot

https://en.wikipedia.org/wiki/Violin_plot