Statistics for Astronomers: Lecture 11, 2019.03.21

Prof. Sundar Srinivasan

IRyA/UNAM

Recall: One-sided confidence intervals

Needed when either dealing with asymmetric distributions (*e.g.*, the IMF), or when dealing with phenomena whose effects that are not symmetric about the mean (*e.g.*, non-detections and saturated sources, which lead to upper and lower limits, respectively).

Example: for some observational setup, the flux limit happens to be $F_{\mathrm{lim}} = 10\ \mu$Jy, with standard deviation $\sigma = 2\ \mu$Jy. A source is targeted but not detected above this threshold.

Then, the probability that the source's flux is below 2σ of the flux limit is

$$P(F <= F_{\mathrm{lim}} - 2\sigma) = P\left(\frac{F - F_{\mathrm{lim}}}{\sigma} <= -2\right) = \frac{1}{2}\left(1 - \mathrm{erf}(\sqrt{2})\right) \approx 2.3\%$$

Therefore, the 2σ limit in this case corresponds to about a 97.7% CI. In the two-sided case, $2\sigma \equiv$ a 95% CI.

Since $10 - 1.96 \times 2 = 6.08\ \mu$Jy, the 97.7% CI in this case is $(-\infty, 6.08]$.

Recall: Basic procedure behind hypothesis testing

1. Design a test statistic T which can be computed from the data.
2. Assuming H_0 is true, obtain the distribution of T.
3. Using the data, compute the value of T.
4. Given H_0 is true, compute the probability of observing a value of T equal to that computed. This is the p-value.
5. If the p-value is lower than the significance α, reject H_0.

$\alpha =$ level of tolerance to wrongly rejecting $H_0 =$ tolerance for false positives (Type-I error rate).

Statistical power: $P(\mathrm{reject}\ H_0|H_1\mathrm{true}) = 1 - P(\mathrm{reject}\ H_0|H_1\mathrm{true}) = 1-$ false-negative rate. (False-negative rate $=$ Type-II error rate).

Example: Tossing a coin 10 times, we observe 9 heads.

Statistic in this case: S_{10}, the total number of heads in 10 tosses.
Null hypothesis: fair coin. S_{10} then has a binomial distribution. Significance chosen: $\alpha = 0.05$.

p-value: $P(S_{10} = 9) = \binom{10}{9}\frac{1}{2}^{10} \approx 0.009$.

Since the p-value $(= 0.009)$ is much lower than the significance, we reject the null hypothesis at significance level $\alpha = 0.05$.

Recall: The t test is an example of a parametric test

Parametric: tests in which either H_0 or the test statistic corresponding to it assumes a distribution with associated parameters.

The t test is used to determine if two samples (or one sample and H_0) have the same mean. It is assumed that they have the same variance. There are other tests for unequal variances.

If comparing two populations, the means of the two populations must be normally distributed.

One-sample test: My data consists of $N = 10$ values with $\bar{x} = -0.47$ and $s = 0.94$. Are these data consistent with being drawn from a population with mean $\mu = 1.0$?

Note: for $N = 10$, the sample mean $\bar{x}$ is approximately normally distributed about μ (actually, t-distributed). So the t-test can be applied to this problem.

Choose $\alpha = 0.05$. $H_0 : \mu = 1$. $t = \dfrac{\bar{x} - \mu}{s/\sqrt{N}}$. Observed value: $t_{\mathrm{data}} = \dfrac{-0.47}{0.296} \approx -1.59$.
Using `scipy.stats.t.pdf with df = 9`, p value $= P(|t| > 1.59) = 0.11$.

Therefore, H_0 cannot be rejected; the data is drawn from a population with mean 1.

Recall: The χ^2 test is a nonparametric tests

Nonparametric tests do not make assumptions about the underlying distribution, and can be applied to small samples. BUT usually need binning, which is bad.

The χ^2 test works for any data (including classifications) that can be binned.
In each bin, predict the population using Poisson statistics (model). Compare to observations.

Model prediction for bin $i = M_i$. Associated variance $= M_i$ (Poissonian). Observed: D_i. The statistic is

$\chi^2 = \sum_{i=1}^{k} \frac{(D_i - M_i)^2}{M_i}$ $\sim \chi^2(k-1)$, with mean $k-1$ and variance $2(k-1)$.

We can define a χ^2 per dof, $\chi^2_{dof} = \chi^2/(k-1)$, which should be ≈ 1 if the model is a good representation of the data. (variance of $\chi^2_{dof} = 2$).

(One dof is lost because of normalisation: $\sum\limits_{i=1}^{k} D_i = \sum\limits_{i=1}^{k} M_i$.)

For the statistic to be $\sim \chi^2(k-1)$, each term in the summation must be $\sim \mathcal{N}(0,1)$.
This is approximately true only if there are enough data points in each bin.
Rule of thumb: $>$ 80% of bins need $>$ 5 points.

Before we move on to the next nonparametric test, we need to learn another concept...

The empirical distribution function (CDF from sample)

Given a dataset of N points $X_i (i = 1, 2, \cdots, N)$ drawn from an underlying CDF $F(x)$, the empirical distribution is defined as

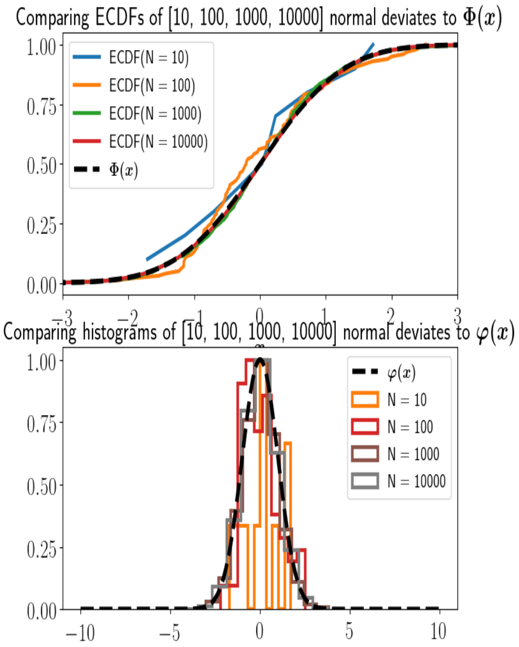
$\hat{F}_N(x) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}_{X_i \le x}(x)$. Here, $\mathbb{I}$ is the indicator function: $\mathbb{I}_{X_i \le x}(x) = \begin{cases} 1 & X_i \le x \\ 0 & \text{otherwise} \end{cases}$

For fixed x, the indicator function is a Bernoulli variable.
$P(\text{“success”}) = P(X_i \le x) = F(x)$. Variance: $F(x)(1 - F(x))$.

$\hat{F}_N(x)$ is the mean of N such Bernoulli variables $\Longrightarrow \hat{F}_N(x)$ is a binomial variable.
Expectation: $F(x) \Longrightarrow \hat{F}_N(x)$ is an unbiased estimator of $F(x)$.
Variance: $\frac{F(x)(1 - F(x))}{N} \Longrightarrow \hat{F}_N(x)$ is a consistent estimator of $F(x)$.

Empirical distribution function: effect of sample size

Draw $N = [10, 100, 1000, 10000]$ values from the standard normal. Use `statsmodels.distributions..empirical_distribution.ECDF` to compute $\hat{F}_N(x)$. Compare it to $\Phi(x)$ on a plot.



Code for plot available here.

CDF version smoothes faster than PDF version. An effect of summing/averaging – relative noise reduced because standard deviations add in quadrature. Opposite effect if differentiating!

If D is a measure of departure of the ECDF from the CDF, D is a function of N.

Glivenko-Cantelli Theorem:
If sample is drawn from the same distribution as the model, then
$N \to \infty \implies D \to 0$.

Back to nonparametric tests

Kolmogorov-Smirnov test

Comparing the distribution of a sample to that of a model (or another sample). In this case, we compare the ECDF of the sample to the CDF of the model (or ECDF of the other sample).

Statistic: $D = \max\left|CDF_{\rm model}(x) - ECDF(x)\right|$. Compute p-value using (*e.g.*) `Python` implementation.

Advantages: No binning! Only alternative for small samples! More powerful for intermediate-size samples! Can also be modified to work as a one-tailed test (see `Python` implementation).

Disadvantages: χ^2 doesn't require a numerical scale, χ^2 can be adapted to incorporate the dof.

Draw 40 standard normal deviates. Find the associated D statistic and p-value that these are drawn from a standard normal using the KS test.

Use `scipy.stats.kstest` – you don't even have to generate the 40 numbers yourself. Read the documentation.

Example: `normD, normp = kstest(norm.rvs(size = 40), 'norm')` results in $normD = 0.130154400472554 33$, $normp = 0.473978590846319$.

Investigate whether `scipy.stats.kstest` or an alternative can be used for two samples. The KS test is also applicable for 2D distributions! Look into this!

Order statistics

Order statistics

Let $X_1, X_2, ..., X_N$ be a finite sample drawn from a distribution $p(x)$ with CDF $F(x)$.
(Note: dropping all "X" subscripts on distributions.)
The r^{th} order statistic is defined as the r^{th}-smallest number among the X_i. It is denoted $X_{(r)}$.

What is the distribution of $X_{(r)}$? (If N numbers are drawn multiple times and ordered, what are the values of the r^{th}-smallest number?)

For $h > 0$, divide real line into three pieces: $(-\infty, x_r], (x_r, x_r + h]$, and $(x_r + h, \infty)$.
In order for only x_r to be in the second bin, we need the first $r - 1$ order statistics to be in bin 1 and the last $N - r$ statistics in bin 2. "Trinomial" distribution!
(The theorem can also be proved for more than one point inside the second bin).

$$p(x_r) = \overbrace{\binom{N}{1}}^{x_r \text{ from the } X_i} \times \underbrace{\binom{N-1}{r-1}}_{r-1 \text{ from the rest}} \left[\int_{-\infty}^{x_r} p(x)dx\right]^{r-1} \left[\int_{x_r}^{x_r+h} p(x)dx\right] \left[\int_{x_r+h}^{\infty} p(x)dx\right]^{N-r}$$

$$h \to 0 \implies p(x_r) = \binom{N}{r-1} p(x_r) \Big[F(x_r)\Big]^{r-1} \Big[1 - F(x_r)\Big]^{N-r}.$$

Sample distribution of the median

The population median $\mu_{1/2}$ is defined such that $F(\mu_{1/2}) = 1/2$.

For a given sample, once the location of the sample median is known, we can use the equation for the sample distribution of order statistics to get the distribution of the sample median.

For large N, it turns out that $x_{1/2}$ is normally distributed about $\mu_{1/2}$ (one proof is available here). Specifically,

$$x_{1/2} \sim \mathcal{N}(\mu_{1/2}, \sigma^2_{1/2}), \qquad \text{with} \qquad \sigma_{1/2} = \frac{1}{2a_N p(\mu_{1/2})}, \qquad a_N = \begin{cases} \sqrt{N} & N \text{ even} \\ \sqrt{N-1} & N \text{ odd} \end{cases}$$

In the above equation, $p(\mu_{1/2})$ is the value of the underlying pdf at the population median, which is usually unknown. Workaround: replace $\sigma_{1/2}$ with $s_{1/2}$, the standard deviation of the bootstrap estimate of the sample median! See Babu et al. 1986 for proof of consistency (pdf).

The original sample must first be Winsorized (one way for compensating for the presence of outliers; see Homework 3) for this procedure.

Review: symmetric two-sided confidence intervals

Let $\theta =$ parameter to be estimated and $\hat{\theta} =$ estimator of θ. Given a distribution of $\hat{\theta}$ values, we can compute the sample mean $\bar{\hat{\theta}}$ and sample standard deviation $s_{\hat{\theta}}$.

Then, a symmetric $100(1-\alpha)$ CI for θ, centered at $\bar{\hat{\theta}}$, is such that $P(\theta$ outside the CI$) \leq \alpha$.
(Note: in the frequentist interpretation, this means that if you generate 100 CIs, the parameter will be recovered $\approx 100(1-\alpha)$ times.)

A symmetric CI means that $P(\theta \text{ outside CI, to its left}) = P(\theta \text{ outside CI, to its right}) = \alpha/2$.
To compute the CI, we first devise a test statistic and compute the value it must take for a given significance α. If the estimator is normally (or t-) distributed, we can standardise (studentise) the estimator.

Our test statistic will either be $z = \dfrac{\hat{\theta} - \bar{\hat{\theta}}}{\sigma_{\hat{\theta}}}$ or $t = \dfrac{\hat{\theta} - \bar{\hat{\theta}}}{s_{\hat{\theta}}}$, depending on whether $\sigma_{\hat{\theta}}$ was provided to us. Let's suppose our statistic is z.
The value of z corresponding to a significance of α is called its critical value and is written as $z_{\alpha/2}$ (the 1/2 alludes to the fact that a symmetric CI splits the probability α into two pieces on either side of it).

The CI is, therefore, $[\bar{\hat{\theta}} - z_{\alpha/2}\ \sigma_{\hat{\theta}}, \bar{\hat{\theta}} + z_{\alpha/2}\ \sigma_{\hat{\theta}}]$.
The critical value $z_{\alpha/2}$ can be interpreted as the distance from the mean $\bar{\hat{\theta}}$ to one edge of the interval, in units of the number of standard deviations.

Asymmetric two-sided confidence intervals

It is not straightforward to write down a two-sided interval for a general asymmetric distribution.

We first compute the ECDF, which allows us to find ranges with specified probabilities on either side of a location parameter. A few ways to define CIs:

1. Compute the sample mean and then use the ECDF to define regions in either direction whose areas correspond to a probability of $(1-\alpha)/2$ each (because of the asymmetry, the two areas will be unequal). The interval is "symmetric" in the sense that the probabilities are the same in either direction.
2. Compute the sample median and use the ECDF to define regions with equal probabilities as in the previous case. This is called an equal-tailed interval (only used if the location parameter is the median).
3. Compute the sample mode and use the ECDF to define regions with equal probabilities. This is called the highest posterior density (HPD) interval, as the mode is the peak of the pdf. The HPD interval minimises the N-dimensional volume and is therefore connected to optimisation.

These terms are usually encountered in the Bayesian context, where the term "confidence interval" is replaced by "credible interval".

Confidence bands

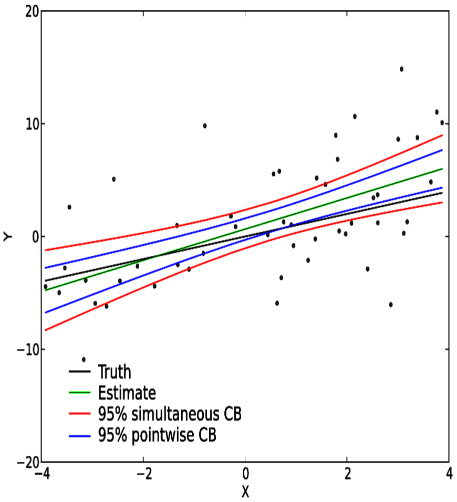
Confidence bands (CBs) are how we translate the variance associated with each observation in a sample into the variance associated with the distribution of these points, or with a regression relationship.

Two ways to generate CBs:

1. "Pointwise" CB: Compute a CI for each point (histogram bin) along the y-axis and combine (logical addition, OR, UNION).
2. "Simultaneous" CB: Combine the individual CIs via Cartesian product (logical multiplication, AND, INTERSECTION).

For a given significance α, simultaneous CBs are narrower (intersection of individual CIs).

CBs are very important for distributions as well as in regression.



Credit: en:User:Skbkekas CC BY 3.0, via Wikimedia Commons.