Statistics for Astronomers: Lecture 09, 2019.03.14

Prof. Sundar Srinivasan

IRyA/UNAM

Recall: The bootstrap algorithm

How do we estimate parameter errors (variances) and confidence intervals when the underlying distribution is unknown?
Instead of making assumptions about the population, we could treat the observations as a hypothetical population and simulate multiple datasets from it. One such resampling technique is bootstrapping.

Given a sample of N points,

1. draw N points with replacement from this dataset $\longrightarrow$ N^N ways to do this.
   If the original data are X_i with $i = 1, 2, \cdots, N\}$, then randomly select N integers $j \in \{1, \cdots, N\}$ with repetition and generate X^*_j.
2. compute the statistic/parameter estimate.
3. repeat M times.

The distribution of the recomputed statistic can be used to estimate the parameter uncertainty and also to generate confidence intervals.

No assumptions about the underlying distribution! Preserves characteristics of original data, including selection effects such as truncation/censoring.

Parametric bootstrap: fit model to data, perform bootstrap by drawing samples from the model distribution.

Recall: Hypothesis testing

Hypothesis: assertion/statement that can be tested using observations.

Typically, a null hypothesis H_0 expresses no correlation between observations and the model suggested (*i.e.*, the data are not significantly different from noise), and the alternate hypothesis H_a suggests a relationship.

If the probability of the data occurring by chance is below a threshold (significance), then we reject the null hypothesis.

Frequentist inference: probability that a given hypothesis is correct is either 0 or 1.
Just because we reject H_0 on the basis of one set of data does not mean H_0 is wrong or H_a is correct.

Convention: "We were [un]able to reject H_0 with significance α", never "We were able to accept H_a!!!!1!!ONE!!11!"

However, misused very often (not just in sociology, but also in astronomy).

Recall: The p-value

The p-value expresses the probability that the observed data occurred by chance.
Associated with a pre-specified significance level (same as α used in constructing CIs), typically taken to be 0.05.

However, this usually assumes that H_0 has a normal distribution. There are many cases where the "standard" threshold is an inadequate description of reality.
Type I error: rejecting the null hypothesis when it is true. Leads us to infer that H_a might be true (false positive).

Example: roll two dice, note the sum of the numbers displayed. $H_0 =$ 'Dice are fair'.
Suppose we observe that the sum is 12.
$P(\mathrm{sum} = 12) = 1/36 \approx 0.028 < 0.05 \longrightarrow H_0$ "is rejected at significance level $p < 0.05$"!
The significance should be tailored to the problem being analysed.

Problem of multiple comparisons: testing many hypothesis with a single dataset increases the probability of "outliers".
Data dredging – multidimensional dataset, plot a bunch of CMDs and look for any correlations... the probability of finding a correlation increases as the number of tests increase!
Can be remedied (*e.g.*, Bonferroni Correction).

One- and two-sided confidence intervals

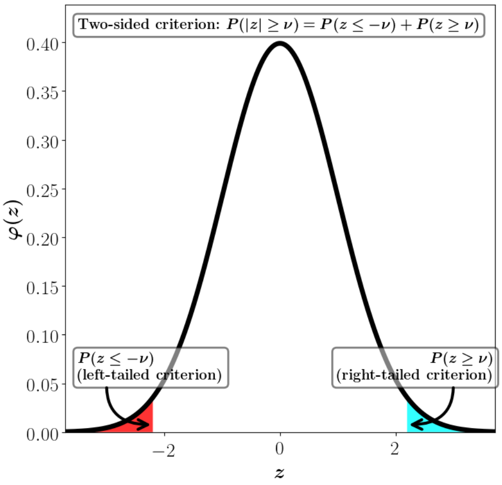
Two-sided CI: place a constraint only on the distance (and not the direction) of the outlier from the distribution mean. One-sided CI: If we also specify the direction (positive or rightward vs. negative or leftward) – left-tailed and right-tailed intervals, respectively.

Example:
Consider a normal distribution.
If $X \sim \mathscr{N}(\mu, \sigma^2)$, standardise: $Z = \dfrac{X - \mu}{\sigma}$.
Now, we can ask for $P(|z| \leq \nu)$, where $\nu > 0$.

Here, we are asking for the probability of an outlier being at distances longer than ν times the standard deviation away from the mean, regardless of the direction w.r.t. the mean. This results in two shaded areas, one for positive distances (cyan) and one for negative distances (red).

For a symmetric distribution (e.g., the Gaussian), the two areas are equal.

$P(|z| \geq \nu)$, $P(z \geq \nu)$, and $P(z \leq -\nu)$ for $\nu = 2.2$
for a standard normal distribution



Code for plot available here.

For a fixed distance from the mean ν, the probability (and, therefore, the significance α) associated with a 2-sided CI is higher than that associated with a 1-sided CI (for a symmetric distribution, it is twice as high). Conversely, for a given α, a two-sided CI enforces a larger distance from the mean (more σs).

One- and two-sided confidence intervals (contd.)

For asymmetric distributions, a given distance away from the mean corresponds to different probabilities depending on the direction, so the resulting "error bars" are not symmetric. As an example, let us consider the χ^2 distribution with k dof.

The scaled version $Q = \dfrac{\chi^2(k) - 1}{2k}$ has $\mu = 0$ and $\sigma = 1$.

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 $(-\infty, 6.08]$.

One-sided confidence intervals (contd.)

Another example: The flux from a very bright source was observed to saturate a detector whose saturation limit was 100 kJy, with a standard deviation $\sigma = 2kJy$ on that particular night. What is the 95% CI on the flux of the source? (use `scipy.special.erfinv`).

$Z = \dfrac{F - F_{\mathrm{lim}}}{\sigma}$ is such that $P(Z > a) = 0.05$ (we need to solve for a).

Since $P(Z > a) = \dfrac{1}{2}\left(1 - \mathrm{erf}\left(\dfrac{a}{\sqrt{2}}\right)\right)$, $\qquad a = \sqrt{2}\,\mathrm{erf}^{-1}(1 - 2 \times 0.05) \approx 1.64.$

Therefore, $F = 1.64\sigma + F_{\mathrm{lim}} \approx 103.3$ kJy, and the 95% CI is $[103.3, \infty)$.

Note here that if this was a two-sided interval, a 95% CI corresponds to about 2σ instead of 1.64σ.

Hypothesis testing

Hypothesis testing and statistical power

Given a null hypothesis, once a test is performed, there are four possible events:

1. We reject H_0 given H_0 is false ("true positive" because we're rightfully favouring any alternate hypotheses).
2. We reject H_0 given H_0 is true ("false positive" because we're erroneously ignoring any alternate hypotheses). This is also called a Type I error.
3. We accept H_0 given H_0 is true ("true negative" because we're correctly ignoring any alternate hypotheses).
4. We accept H_0 given H_0 is false ("false negative" because we're erroneously favouring any alternate hypotheses). This is also called a Type II error.

The significance level α used for confidence intervals and hypothesis testing is the level of tolerance for wrongly rejecting H_0.
This is also the tolerance for false positives and, therefore, α is the Type-I error rate.
We define statistical power as the probability of true positives (the probability of rejecting H_0 given H_0 is false).

$$\begin{aligned}\text{Statistical power } &= P(\text{reject } H_0|H_0 \text{ false}) = 1 - P(\text{reject } H_0|H_0 \text{ true})\\ &= 1 - \text{ false-negative rate} = 1 - \text{Type-II error rate.}\end{aligned}$$

Hypothesis testing: basic procedure

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.

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 $\alpha = 0.05$.

Parametric tests

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

Given a test sample and a distribution corresponding to H_0 (or two test samples),
Do they have the same mean? Use the t test.
Do they have the same variance? Use the F test.
Both these tests compare data to normal distributions. There are other tests for non-normal distributions.

t test

Whenever, under H_0, the test statistic follows the t distribution.

The 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.

t test examples

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: $t_{\rm data} = \dfrac{-1.47}{0.296} \approx 4.97$.
Using `scipy.stats.t.pdf` with df $= 9$, p value $= P(|t| > 1.59) = 0.00053$.

Therefore, H_0 is rejected ($p \ll \alpha = 0.05$); the population mean of the data is not 1.

Two(independent, equal-sized)-sample test: Two $N = 20$ datasets have sample means 80 and 180 with sample standard deviations 18 and 28. Are they drawn from distributions with the same mean?
Note: If $Y_j \sim \mathcal{N}(\mu_1, \sigma^2)$ for $j = 1, 2$, then $Y_1 - Y_2 \sim \mathcal{N}(\mu_1 - \mu_2, 2\sigma^2)$.

Choose $\alpha = 0.05$. $H_0 : \mu_1 = \mu_2 = \mu$ (say).
$t = \dfrac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{s/\sqrt{N}} = \dfrac{\bar{x}_1 - \bar{x}_2}{s/\sqrt{N}}$, with $s = \sqrt{s_1^2 + s_2^2}$. Observed: $t = 13.44$.
Using `scipy.stats.t.pdf` with df $= 19$, p value $= P(|t| > 13.44) = 2.4 \times 10^{-11}$.

Therefore, H_0 is readily rejected; the samples have significantly different means.

Nonparametric tests

(From Wall & Jenkins)

Why use nonparametric tests?:

1. Very little assumed about the data and underlying distributions. Great for when we don't know!
2. Small sample sizes!
3. Also applicable when data is non-numeric (*e.g.*, classifications).
4. Can work with samples drawn from several different populations.

Major issue with nonparametric tests:
Some tests require binning of data, and binning's bad, mmkay?

The χ^2 test

Works for data 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: $\sum_{i=1}^{k} D_i = \sum_{i=1}^{k} M_i$.)

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.

$$P(\text{“success”}) = P(X_i \le x) = F(x). \text{ Variance: } F(x)(1 - F(x)).$$

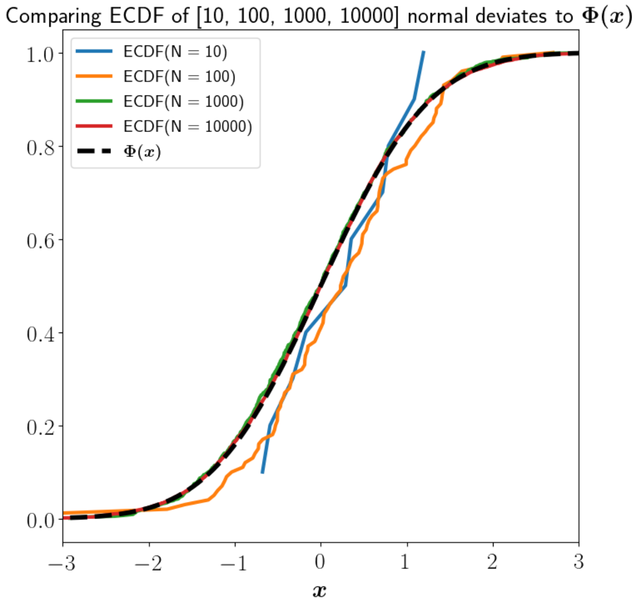
$\hat{F}_N(x)$ is the mean of N such Bernoulli variables $\implies \hat{F}_N(x)$ is a binomial variable.

Expectation: $F(x) \implies \hat{F}_N(x)$ is an unbiased estimator of $F(x)$.

Variance: $\dfrac{F(x)(1 - F(x))}{N} \implies \hat{F}_N(x)$ is a consistent estimator of $F(x)$.

Empirical distribution function (contd.)

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.



$D = f(N)$.

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