Statistics for Astronomers: Lecture 09, 2019.03.11

Prof. Sundar Srinivasan

IRyA/UNAM

Recall: the χ^2 distribution and distributions of some statistics

- The sum of squares of N independent standard normal deviates is the χ^2 distribution for N degrees of freedom. The mean and variance of a $\chi^2(N)$ distribution are N and $2N$ respectively.
- Due to the CLT, the sample mean of N samples from **any population** is $\sim \mathcal{N}(\mu, \sigma^2/N)$.
- For normally distributed iid variables, the sample variance $s^2 \sim \frac{\sigma^2}{N-1}\chi^2(N-1)$, with mean σ^2 (s is unbiased) and variance $\frac{2\sigma^4}{N-1}$ (s^2 is consistent).
- Due to the nonlinear relationship, while the variance is well-behaved, the standard deviation isn't. $s \sim \frac{\sigma}{\sqrt{N-1}}\chi(N-1)$ is not an unbiased estimator of σ.

Recall: the z- and $t-$scores

Location+scale distribution $\longrightarrow$ location-only distribution:

$$Z = \underbrace{\frac{X-\mu}{\sigma}}_{\text{"standardisation"}} \sim \mathcal{N}(0,1), \qquad T = \underbrace{\frac{\bar{X}-\mu}{S/\sqrt{N}}}_{\text{"studentisation"}} \sim \left(1+\frac{t^2}{\nu}\right)^{-(\nu+1)/2}.$$

$$P(Z \leq a) = \frac{1}{2}\left[1+\text{erf}\left(\frac{a}{\sqrt{2}}\right)\right]$$

$P(|Z| < a)$ $1\sigma : 0.68, 2\sigma : 0.95, 3\sigma : 0.997$

$P(|Z| > z_{\alpha/2}) \Longrightarrow z_{\alpha/2} = \sqrt{2}\,\text{erf}^{-1}\left(1-\alpha\right)$.

Usually $1-\alpha = 95\%$

For $\nu = 4$:

$P(T_{\nu=4} = 1) \approx 0.81; P(Z = 1) \approx 0.84$

$P(T_{\nu=4} = 2) \approx 0.94; P(Z = 2) \approx 0.98$

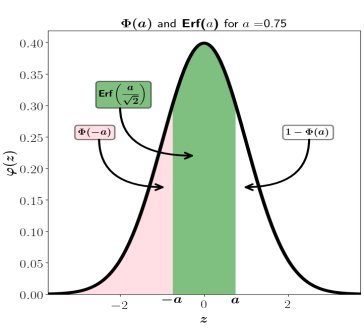
$P(T_{\nu=4} = 3) \approx 0.98; P(Z = 3) \approx 0.999$

$\alpha = 0.1$: $t_{\nu=4,\alpha/2} \approx 2.13; z_{\alpha/2} \approx 1.64$
(`print(scipy.stats.t.ppf(1 − α/2))`)

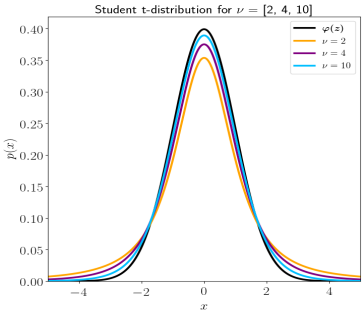
$\alpha = 0.05$: $t_{\nu=4,\alpha/2} \approx 2.78; z_{\alpha/2} \approx 1.96$

$\alpha = 0.003$: $t_{\nu=4,\alpha/2} \approx 6.44; z_{\alpha/2} \approx 2.97$

T and Z scores are very different because of behaviour in the tails!



Code for plot available here.



Code for plot available here.

Recall: confidence sets

$1-\alpha$ confidence interval for a parameter θ: $P_\theta(\theta \in (a,b)) \geq 1-\alpha$, where $a, b : (X_1, X_2, \cdots, X_N) \longrightarrow \mathbb{R}$. (a,b) is the called a $100(1-\alpha)\%$ confidence interval for θ.

$\theta \longrightarrow \vec{\theta}$: confidence interval $\longrightarrow$ confidence set.

In the frequentist paradigm? The probability that a single interval traps the true parameter value is either 0 or 1!

The CI expresses uncertainty about the process of interval estimation, not about the true parameter. If the procedure is repeated a large number of times, the resulting intervals will trap the true parameter value $100(1-\alpha)\%$ of the time.

Perform an experiment each day, trap a parameter θ_j in a 95% CI on the j^{th} day. As long as you use the same procedure to construct the CI, it doesn't even have to be the same experiment!!.

In the long run, 95% of the intervals you constructed would have trapped the true value of whatever parameter you were exploring.

BUT P(parameter trapped in today's CI) $\in \{0, 1\}$.

Confidence interval: Example 1

Flip a coin $N = 100$ times. Observe: 60 heads, 40 tails.
What is the probability of getting a head on a single flip of the coin? What is the 95% confidence interval for this estimate?

$i^{\rm th}$ flip = Bernoulli variable X_i. Final outcome: sum of $N \gg 1$ Bernoulli trials:
#Heads $X_{\rm tot} = \sum_{i=1}^{N=100} X_i \implies X_{\rm tot} \sim \mathcal{N}(\mu, \sigma^2)$ (CLT).

Let p be the probability of getting a head on a single flip.
Observe: 60 heads $\implies \hat{\mu} = 100\hat{p} = 60, \hat{p} = 0.6, \hat{\sigma} = \sqrt{100\hat{p}(1-\hat{p})} = 4.90$.

95% confidence interval on the true mean μ:
For a normal distribution, $1 - \alpha = 0.95 \implies z_{\alpha/2} = 1.96$.
95% CI for $\mu = [100\hat{p} - 1.96\sqrt{100\hat{p}(1-\hat{p})}, 100\hat{p} + 1.96\sqrt{100\hat{p}(1-\hat{p})}] = [55.1, 64.9]$.
$\implies$ 95% CI for $p = [0.551, 0.649]$.

Confidence interval: Example 2

A sample of 10 draws from a standard normal has a mean $\bar{x} = 0.6073$ and standard deviation $s = 0.6417$. Construct a 95% CI on the true mean of the distribution.

For $N < 30$, use the t distribution instead of the normal. #dof $= \nu = N - 1 = 9$.
95% CI $\implies \alpha = 0.05, t_{\nu=9,\alpha/2=0.025} = 2.262$ (`print(scipy.stats.t.ppf(0.95+0.05/2, 9))`).

Standard error on the mean: $\sigma_{\bar{x}} = \dfrac{s}{\sqrt{N}} = 0.2029$.
95% CI on true mean $\mu = [\bar{x} - \sigma_{\bar{x}} t_{\nu,\alpha/2}, \bar{x} + \sigma_{\bar{x}} t_{\nu,\alpha/2}] = [0.1483, 1.066]$.

In this example, the CI does not trap the true mean $\mu = 0$. However, if this procedure is repeated a large number of times, about 95% of the intervals will trap the true mean.

In-class assignment: repeat this procedure 1000 times and compute the fraction of the 1000 intervals that trap the true mean (zero). You can use `scipy.stats.norm.rvs` or `numpy.random.normal` to draw from the normal distribution.

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.

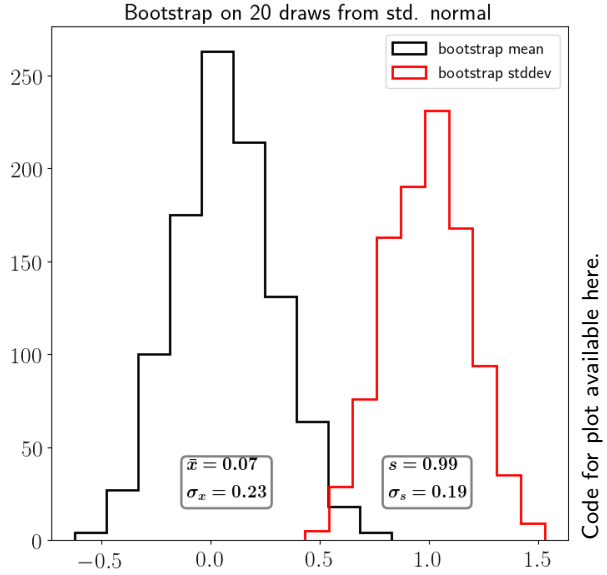
The distribution of the recomputed statistic can be used to estimate the parameter uncertainty and also to generate confidence intervals.

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.

Bootstrap example

Data: 20 draws from a standard normal.
Mission: Generate CIs for μ and σ.



Does the CI for μ trap $\mu = 0$?
Does the CI for σ trap $\sigma = 1$?

Note:
From CLT, $\bar{x} \sim \mathcal{N}(\mu, \sigma^2/N) \Longrightarrow$
$\text{mean}(\bar{x}) = \mu$ and $\sigma_{\bar{x}} = \sigma/\sqrt{N}$.

Similarly, using the theoretical mean and variance for the χ distribution, we can estimate s and σ_s:

$$s = \frac{\sigma}{\sqrt{N-1}}\chi(N-1) \Longrightarrow$$
$$\bar{s} \approx 0.99\sigma \text{ and } \sigma_s \approx 0.16\sigma.$$

Results from the simulation are consistent with the above theoretical estimates.

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 significant α", never "We were able to accept H_a!!!!1!!ONE!!11!"

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

The p-value

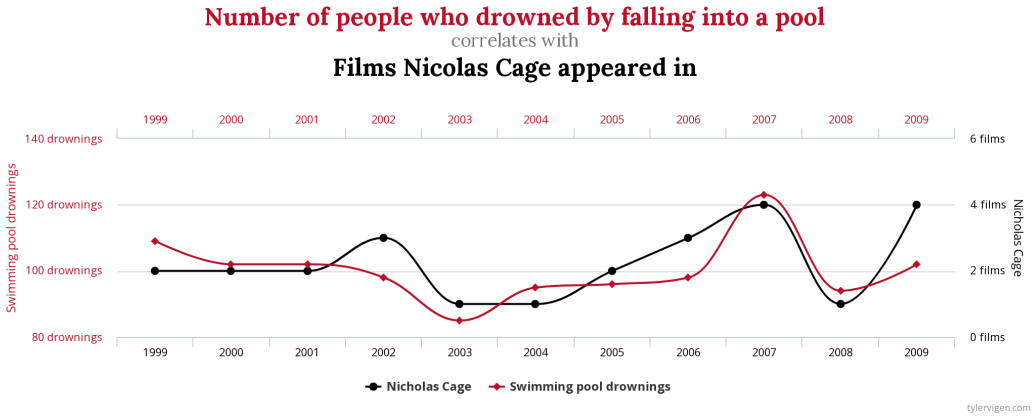
The p-value expresses the probability that the observed data occurred by chance.
Associated with a pre-specified significance level α, typically taken to be 0.05.

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 0.05"!
The significance should be tailored to the problem being analysed.

P-values (contd.)

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!



For more: http://www.tylervigen.com/spurious-correlations.

Can be remedied (*e.g.*, Bonferroni Correction).