# Statistics for Astronomers: Lecture 13, 2019.04.04

Prof. Sundar Srinivasan

IRyA/UNAM

## Recall: The ECDF

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

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

For fixed $x$, the indicator function is a Bernoulli variable.

$P(\text{“success”}) = P(X_i \leq x) = F(x)$. 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 $F(x)$.

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

# Recall: 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. BUT not always connected.

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

# Recall: Confidence bands

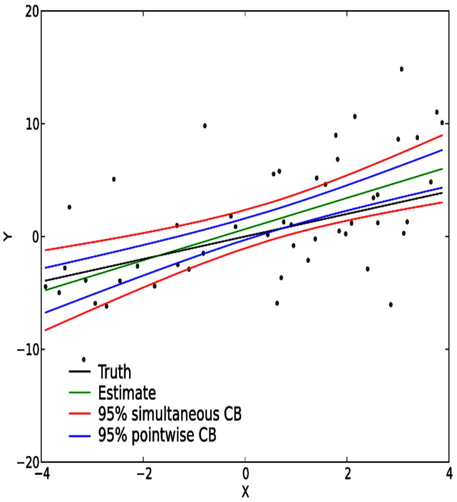
Confidence bands (CBs) are how we translate the variance associated with each observation in a sample into the variance associated with the distribution of these points, or with a regression relationship.

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 $\alpha$, 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.

# Recall: The 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:
$\chi^2$ doesn't require a numerical scale, $\chi^2$ can be adapted to incorporate the dof.
Can't "see" discrepancies near the tails of the EDF.
Not very sensitive if the distance between the two functions changes sign in the central region of the distribution.

---

`scipy.stats.ks_2samp` can be used to compare the EDF of one sample to that of another.

While it is used for higher-dimensional distributions quite often in the literature, experts do not recommend this. See "Beware the Kolmogorov-Smirnov test!" (https://asaip.psu.edu/Articles/beware-the-kolmogorov-smirnov-test).

# Alternatives to the KS test

The KS test is less sensitive to relative deviations at either end of the distribution ($F(x) \approx 0$ or $F(x) \approx 1$, where $F(x)$ is usually slowly-varying). In addition, the KS statistic (maximum deviation) is reduced if the ECDFs cross each other multiple times.
There are two important alternatives to the KS test. They are both quadratic ECDF tests.

In quadratic ECDF tests, the distance is computed as $N \int_{-\infty}^{\infty} dF(x) \left[\hat{F}_n(x) - F(x)\right]^2 w(x)$,

where $w(x)$ is a weight function to emphasize different regions of the distribution.

The Cramér-von Mises test uses $w(x) = 1 \; \forall \; x$.
The Anderson-Darling test uses $w(x) = \dfrac{1}{F(x)(1-F(x))} = \dfrac{1}{Var(\hat{F}_n(x))}$, placing more weight on observations near the tails of the distribution.

In most cases where you're tempted to use the KS test, use the AD test instead.

# AD test vs KS test: 1-sample version

Draw 100 standard normal variates.
Remove any that are $\leq -0.5$ (truncated normal).
Using scipy.stats.kstest, run the KS test to see if the truncated data are drawn from a normal.
Using scipy.stats.anderson, run the AD test to see if the truncated data are drawn from a normal.

Sample result:

```
KS test produces p = 0.3569360869959737.  The null hypothesis is accepted at 95%
significance.
AD test produces A2 = 0.8562595199770016, which is > the critical value for 95%
significance.  Null hypothesis rejected!
```

Comparing relative efficiencies of the two tests:
Repeat the entire procedure 1000 times $\cdots$

```
KS test null hypothesis rejection fraction:  13.6%
AD test null hypothesis rejection fraction:  21.8%
```

AD test better at rejecting null hypothesis for this truncated normal example.

# Spearman's rank correlation test

Given a sequence $X_i$ of size $N$, the rank $\widetilde{X}_i$ of the data point $X_i$ is its location in the ordered sequence. For example, if $X = [5.2, 2.1, 1.0, -1.1, 4.3]$, then $\widetilde{X} = [5, 3, 2, 1, 4]$.

The relationship between two variables $X$ and $Y$ can be quantified using Pearson's correlation coefficient: $\rho_{XY} = \dfrac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$.

Perfect correlation: $\rho_{XY} = 1$. Also, $\widetilde{X} = \widetilde{Y}$.
Perfect anti-correlation $\rho_{XY} = -1$. Also, $\widetilde{X} = \text{reverse}(\widetilde{Y})$.

Use Pearson's correlation coefficient, except with $\widetilde{X}, \widetilde{Y}$ instead of $X, Y$.

This is Spearman's rank correlation coefficient, $r_S = \dfrac{Cov(\widetilde{X}, \widetilde{Y})}{\sqrt{Var(\widetilde{X})Var(\widetilde{Y})}}$.

If (and only if) the $\widetilde{X}$ and $\widetilde{Y}$ are distinct integers (no $X$ or $Y$ values are repeated),

$$r_S = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)}, \text{ where } d_i = \widetilde{X}_i - \widetilde{Y}_i.$$

# Spearman's rank correlation test (contd.)

$\rho_{XY}$ measures the extent of linear relationship between $X$ and $Y$. $r_S$, on the other hand, measures the extent of any monotonic relationship (think of the definition of $d_i$).

Try this example:

```
>>> x = np.linspace(-3, 3, 1001); y = norm.cdf(x)
>>> print(pearsonr(x, y)); print(spearmanr(x, y))
>>> plt.plot(x, y); plt.show()
```

$Y = \Phi(X)$ is a monotonic function, so $X$ and $Y$ in this case must be perfectly correlated. However, since $\rho_{XY}$ is built to detect linear relationships, it is not exactly 1.

Try this example which adds an outlier:

```
>>> x = np.linspace(-3, 3, 1001); y = norm.cdf(x); y[800] = 1e5
>>> print(pearsonr(x, y)); print(spearmanr(x, y))
```

$r_S$ is much more resistant to outliers!!

# Anscombe's Quartet and Triassic reptiles

Quantities like the mean, standard deviation, and correlation coefficient are data summaries.

**Always** visualise your data in addition to computing summaries. The summaries alone are not enough to reconstruct the dataset in your head.

To illustrate this concept, statistician Francis Anscombe designed four datasets, all of which have the same mean and variance in each direction, and the same correlation coefficient.

More recently, this has been extended to a dozen (or more) datasets... The Datasaurus Dozen.

# Bayesian inference

# Bayes' Theorem, reframed

Given some prior information $I$, we can select a model $M$ that includes parameters $\vec{\theta}$. Bayes' Theorem is then

$$\overset{\text{posterior prob. distrib.}}{\underbrace{p(M,\vec{\theta}|D,I)}_{\text{for model and model parameters}}} = \frac{\overbrace{p(D|M,\vec{\theta},I)}^{\text{likelihood}}\ \overbrace{p(M,\vec{\theta}|I)}^{\text{prior}}}{\underbrace{p(D|I)}_{\text{prior predictive prob.}}}$$

This form is appropriate for parameter estimation.
Frequentist: parameters are fixed!
Bayesian: this is our degree of belief in a given value of the parameter.

For model selection, we expand the prior: $p(M,\vec{\theta}|I) = p(\vec{\theta}|M,I)\ p(M|I)$
(second term on RHS $\neq 1$ for model selection).

# Illustration of prior selection: waiting for a bus

from Ivezic et al. AstroML book

The bus arrives $t$ min after you arrive at the stop. What is the mean interval $\tau$ between successive buses?

Method 1: The wait time $t$ is distributed uniformly in the interval $0 \leq t \leq \tau$.
$E[t] = \tau/2 \implies \tau = 2t$.

Method 2: The likelihood of wait time $t$ given interval $\tau$ is

$$p(t|\tau) = \begin{cases} 1/\tau & 0 \leq t \leq \tau \\ 0 & \text{otherwise} \end{cases}$$

The likelihood is maximised for the smallest value of $\tau$ such that $t \leq \tau \implies \tau = t$.

# Illustration of prior selection: waiting for a bus: Bayesian method

Method 3, Bayesian:
$p(\tau|t, I) \propto p(t|\tau, I)\ p(\tau|I)$.
$p(t|\tau, I)$ as in Method 2.
Prior:Choose the least informative prior $p(\tau|I) \propto 1/\tau, \text{ for } t \leq \tau \leq \infty$.
$\implies p(\tau|t, I) = C/\tau^2 \text{ for } t \leq \tau \leq \infty$, where $C$ is a normalisation constant.
Find $C$.
$\implies p(\tau|t) = t/\tau^2, t \leq \tau \leq \infty$.
Find the CDF of this distribution. Use this to find the median and the 95% CI.
$F(\tau) = 1 - t/\tau$; median: $\tau = 2t$, as in method 1.
95% CI for $\tau$: $1 - t/\tau = 0.025$ and $1 - t/\tau = 1 - 0.025 = [1.03t, 40t]$.

The median computed here is one example of an estimator based on the posterior.
Instead of maximising the likelihood, in the Bayesian interpretation, we maximise the posterior to get the *maximum a posteriori* (MAP) estimate.