Homework 4

Due Fri, 16 Feb, at the beginning of class.

1 Sample Mean and Sample Variance as Statistical Estimators

The goal of this problem is to derive some basic properties of the sample mean and sample variance, and to inform your physical intuition on the expected uncertainty or scatter in these estimators, due to the fact that they are random variables.

Given: A set of N independent samples $\{x_i\}$ from a parent PDF $f_X(x)$ with mean $\mathrm{E}[x] = \mu_x$ and variance $\mathrm{V}[x] = \sigma_x^2$.

a) Show that the sample mean is unbiased, i.e.,

$$\mathrm{E}[\hat{\mu}_x] = \mu_x$$

where

$$\hat{\mu}_x = \frac{1}{N}\sum_{i=1}^{N} x_i$$

is the sample mean.

b) Show that the sample variance with a *known* mean is unbiased, i.e.,

$$\mathrm{E}\left[\hat{\sigma}_x^2\right] = \sigma_x^2$$

where

$$\hat{\sigma}_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)^2$$

is the sample variance using the known mean of the underlying distribution.

c) Show that the sample variance for a distribution with an *unknown* mean is unbiased, i.e.,

$$\mathrm{E}\left[\hat{\sigma}_x^2\right] = \sigma_x^2$$

where

$$\hat{\sigma}_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \hat{\mu}_x)^2$$

is the sample variance estimated using the *sample* mean.

d) The sample mean is itself a random variable, and in a) you derived its mean value. Now, derive the variance of the sample mean, $\mathrm{V}[\hat{\mu}_x]$.

e) The sample variance is b) is itself a random variable, and in b) you derived its mean value. Now, derive the variance of the sample variance with a known mean, $\mathrm{V}\left[\hat{\sigma}_x^2\right]$. You may leave your solution in terms of the variance (second central moment), σ_x^4, and the fourth central moment, $\mathrm{E}\left[(x_i - \mu_x)^4\right]$, since these cannot be further reduced without knowing the specific PDF.

2 Naïve Error Estimation

The goal of this problem is to numerically test some results from Problem 1, and examine some pitfalls of "naïve" error analysis. Please turn in your source code, and be sure to comment it well.

a) Using a unit Gaussian random number generator in Python/IDL/matlab (Gaussian random number with $\mu_x = 0, \sigma_x^2 = 1$), draw $N = 10$ samples from the distribution. Find the sample mean $\hat{\mu}_x$, the sample variance $\hat{\sigma}_x^2$ assuming the mean $\mu_x = 0$ is known, and the sample variance $\hat{\sigma}_x^2$ assuming the mean is unknown.

b) Find the (true) variance for each of the first two statistical estimators you calculated in a), skipping the sample variance with unknown mean, since you did not calculate this in Prob. 1. (It is close to the variance of the sample variance with known mean for large N). Do your numerical values for the estimators in a) agree with the true values to within one or two sigma?

Given the outcome of your experiment, (where, let's say the mean and variance of the parent Gaussian distribution are both unknown) you may be tempted to report the result

$$\hat{\mu}_x \pm \hat{\sigma}_{\hat{\mu}_x}$$

where

$$\hat{\sigma}_{\hat{\mu}_x} = \sqrt{\frac{\hat{\sigma}_x^2}{N}}$$

is the estimated uncertainty of the sample mean, and $\hat{\sigma}_x^2$ is the sample variance assuming (of course) the mean is unknown. The sample mean and error estimate above has the (frequentist) interpretation that the interval

$$[\hat{\mu}_x - \hat{\sigma}_{\hat{\mu}_x}, \hat{\mu}_x + \hat{\sigma}_{\hat{\mu}_x}]$$

contains the "true" mean μ_x 68% of the time. Let's test to see if this interpretation is indeed correct.

c) Repeat your "experiment" $M = 1000$ times. For each of your $M = 1000$ experiments, construct the confidence interval

$$[\hat{\mu}_x - \hat{\sigma}_{\hat{\mu}_x}, \hat{\mu}_x + \hat{\sigma}_{\hat{\mu}_x}]$$

where $\hat{\mu}_x$ is the sample mean and $\hat{\sigma}_{\hat{\mu}_x}$ is the estimator of the uncertainty of the sample mean constructed from the $N = 10$ samples in the experiment. Find the number of times the constructed confidence interval contains the true mean. Is it 68.3% of the time (683/1000), as you might expect? Increase the number of experiments M until you are satisfied your fractional result has converged to two significant digits. Is your result 68%, as expected? If not, explain conceptually why not.

3 Simple Error Propagation

Consider two independent random variables, x, y, with means μ_x, μ_y and variances σ_x^2, σ_y^2, respectively.

a) Derive the well-known expression for the variance σ_z^2 of the difference of two independent variables,

$$z = x - y$$

using the formalism for (approximate) multivariate error propagation discussed in class.

b) Derive the exact expression for the variance σ_z^2 of the difference of two independent variables, starting with the definition of the variance,

$$V[z] = E[(z - \mu_z)^2].$$

Does your expression agree with the approximate expression you derived in a)?

c) Derive the well-known expression for the variance σ_z^2 of the product of two independent variables,

$$z = xy$$

using the formalism for (approximate) multivariate error propagation discussed in class.

d) Derive the exact expression for the variance σ_z^2 of the product of two independent variables, starting with the definition of the variance,

$$V[z] = E[(z - \mu_z)^2].$$

Does your expression agree with the approximate expression you derived in c)? If not, what regimes of means μ_x, μ_y and variances σ_x^2, σ_y^2 is your expression in c) a good approximation?