This is an introduction to the statistical tools needed to analyze data, especially as generated by Markov Chain Monte Carlo simulations. Parts of this presentation follow selected parts of Chapters 1 and 2 of Ref. [1].

1 Preliminaries

The set of all possible outcomes Ω of an experiment is called the *sample space*. *Events* A are aggregates of points in the sample space. If Ω is a measurable set, then a *random variable* is a function $X : \Omega \to \mathbb{R}$. In this book we will try to always denote random variables with capital letters.

Next we need a way to define probabilities. In the case that $|\Omega| < \infty$, we might say that the probability of event A is $\mathrm{P}(A) = |A|/|\Omega|$. In this instance, we say that the probability is *uniform* since every point in the sample space is equally likely. However this need not always be the case; one outcome might be more likely than the other ones. In general for some integrable function $f : \mathbb{R} \to \mathbb{R}$, we assign a probability that X lies in the interval $[a, b]$ by

$$\mathrm{P}(X \in [a, b]) = \int_a^b dx\, f(x). \tag{1}$$

Often statisticians say X is a *continuous* random variable, which is to be contrasted with *discrete* random variables. In this book we will only be concerned with continuous random variables, and we will simply call them random variables. The function f is called the *probability distribution function* (PDF), and it must satisfy

$$1 = \int_{-\infty}^{\infty} dx\, f(x). \tag{2}$$

Meanwhile the *cumulative distribution function* (CDF) is the function $F(x)$ given by

$$F(x) \equiv \mathrm{P}(X < x) = \int_{-\infty}^{x} dt\, f(t). \tag{3}$$

Example. Two examples of important probability distributions include the *Gaussian* or *normal* distribution,

$$\mathrm{gau}(x, \hat{x}, \sigma) \equiv \frac{1}{\sigma\sqrt{2\pi}} \exp\left(- \frac{(x - \hat{x})^2}{2\sigma^2} \right) \tag{4}$$

where σ is the standard deviation of the distribution and $\hat{x}$ is the mean, and the *Cauchy* distribution,

$$\text{cau}(x, \alpha) \equiv \frac{\alpha}{\pi\left(\alpha^2 + x^2\right)}. \tag{5}$$

I will refer to these special PDFs later, particularly the normal distribution. I'll call their CDFs Gau and Cau, respectively.

Now that we know what probabilities and PDFs are, we can start thinking about ways to characterize them. For example we can think about typical values taken by a random variable from some distribution. We can get some information from the mean and variance of a distribution. These are both special cases of a more general concept. In particular, let $n \in \mathbb{N}$. The n^{th} *moment* of the distribution $f(x)$ is

$$\langle X^n \rangle = \int_{-\infty}^{\infty} dx\, x^n f(x). \tag{6}$$

The mean and variance are the special cases $\hat{x} = \langle X \rangle$ and $\sigma^2 = \langle (X - \hat{x})^2 \rangle$. Sometimes we call the mean the *expected value* and sometimes we denote the variance var. Note that not all probability distributions have well-defined moments. The Cauchy distribution is very ill-behaved in this regard, since its n^{th} moment diverges $\forall\, n \in \mathbb{N}$. Generally in the lab, one draws random variables from distributions about which one has no a priori knowledge. Therefore in principle one doesn't know the true moments these distributions. The definition (6) suggests a way to estimate them. Suppose you draw a sample $X_1, ..., X_N$: An *estimator* of the n^{th} moment is

$$\bar{X}^n \equiv \frac{1}{N} \sum_{i=1}^{N} X_i^n. \tag{7}$$

In the case $n = 1$ we obtain the ordinary arithmetic average. We use the hat to distinguish true values from estimators, which will generally be denoted with a bar. For estimators of moments besides the mean, we must be more careful; this is discussed in Section 4.

Consider two intervals $[a, b]$ and $[c, d]$ and two random variables X and Y drawn from PDFs f and g, respectively. Then X and Y are said to be *independent* if

$$\text{P}\left(X \in [a, b] \text{ and } Y \in [c, d]\right) = \int_a^b \int_c^d dx\, dy f(x)\, g(y) \tag{8}$$

Hence we see that the *joint PDF* of X and Y is $f(x)g(y)$. If X and Y are not independent, then we say they are *dependent*. On the other hand, we say X and Y are *uncorrelated* if

$$\langle XY \rangle = \langle X \rangle \langle Y \rangle , \tag{9}$$

and we say they are *correlated* otherwise. The *covariance*

$$\mathrm{cov}[X, Y] \equiv \langle XY \rangle - \langle X \rangle \langle Y \rangle \tag{10}$$

can be used to give a measure of how correlated X and Y are, or one can use the *correlation*

$$\rho(X, Y) = \frac{\mathrm{cov}[X, Y]}{\sqrt{\sigma_X^2 \sigma_Y^2}}. \tag{11}$$

So equivalently we say X and Y are correlated if $\rho(X, Y) = 0$. It's worth emphasizing that if X and Y are independent, it follows that they are uncorrelated. This can be seen by applying definition (6) to the random variable XY, then using definition (8). However if X and Y are uncorrelated, *they can still be dependent.*

Example. Here's an extreme example by Cosma Shalizi [2]. Let X be uniformly distributed on [-1,1] and let $Y = |X|$. Then clearly Y depends on X. However it is easy to see that Y is uniform on [0,1] and $\langle XY \rangle = 0 = \langle X \rangle \langle Y \rangle$. Hence X and Y are not correlated.

The next two propositions show us how to add expectation values and random variables. Let X and Y be independent random variables drawn from PDFs f and g, respectively.

Proposition 1. *Let $a, b \in \mathbb{R}$ be constants. Then*

$$\langle aX + bY \rangle = a \langle X \rangle + b \langle Y \rangle .$$

Proof. Since X and Y are independent, their joint PDF is fg. Then

$$\begin{aligned} \langle aX + bY \rangle &= \int dxdy\, (ax + by) f(x)g(y) \\ &= a \int dxdy\, x\, f(x)g(y) + b \int dxdy\, y\, f(x)g(y) \\ &= a \int dx\, x\, f(x) + b \int dy\, y\, g(y) \\ &= a \langle X \rangle + b \langle Y \rangle . \end{aligned}$$

□

Proposition 2. *The PDF of the random variable $Z = X + Y$ is given by the convolution*

$$h(z) = \int_{-\infty}^{\infty} dx\, f(x)g(z-x)$$

Proof. The CDF of Y is, according to eq. (8),

$$G(y) = \int_{x+y\leq z} dx\, dy\, f(x)g(y) = \int_{-\infty}^{\infty} dx\, f(x) \int_{-\infty}^{z-x} dy\, g(y).$$

The PDF h follows from the Fundamental Theorem of Calculus:

$$h(z) = \frac{dH}{dz} = \frac{dH}{d(z-x)} = \int_{-\infty}^{\infty} dx f(x)g(z-x).$$

□

A sequence $\{X_N\}$ of random variables *converges in probability* toward the random variable X if $\forall \epsilon > 0$

$$\lim_{N\to\infty} \mathrm{P}\left(|X_N - X| > \epsilon\right) = 0, \tag{12}$$

and we write

$$X_N \xrightarrow{\mathrm{P}} X. \tag{13}$$

The sequence converges to X *almost surely* if

$$\lim_{N\to\infty} \mathrm{P}\left(X_N = X\right) = 1, \tag{14}$$

and in this case we write

$$X_N \xrightarrow{\mathrm{AS}} X. \tag{15}$$

Theorem 1 (Chebyshev's Inequality)**.** *Let X be drawn from a PDF with mean $\hat{x}$ and variance σ^2 and let $a > 0$. Then*

$$\mathrm{P}\left(|X - \hat{x}| > a\sigma\right) < a^{-2}.$$

Proof. Let $T = (X - \hat{x})^2$ be a new random variable with PDF g. Then

$$\mathrm{P}\left(|X - \hat{x}| > a\sigma\right) = \mathrm{P}\left(T > a^2\sigma^2\right) = \int_{a^2\sigma^2}^{\infty} dt\, g(t)$$

But

$$\sigma^2 = \int_0^\infty dt\, t\, g(t) = \left(\int_0^{a^2\sigma^2} + \int_{a^2\sigma^2}^\infty \right) dt\, t\, g(t)$$
$$\geq \int_{a^2\sigma^2}^\infty dt\, t\, g(t) > a^2\sigma^2 \int_{a^2\sigma^2}^\infty dt\, g(t) = a^2\sigma^2\, \mathrm{P}\left(T > a^2\sigma^2\right).$$

Dividing through by $a^2\sigma^2$ completes the proof. □

Chebyshev's inequality tells you that large deviations from the mean are unlikely. Intuitively you expect that as the number of measurements increases, the sample average tends toward the true mean. This is called the *Law of Large Numbers* (LLN). To prove it, we set up as follows: Let $X_1, ..., X_N$ be a sequence of random variables drawn from a PDF with mean $\hat{x}$ and variance σ^2.

Theorem 2 (Weak LLN)**.**

$$\bar{X} \xrightarrow{P} \hat{x}.$$

Proof. Our proof will rely on Chebyshev's inequality, so we will first need to compute the mean and variance of the distribution of $\bar{X}$. All the X_i are drawn from the same PDF, so

$$\left\langle \bar{X} \right\rangle = \frac{1}{N} \sum_{i=1}^N \langle X_i \rangle = \frac{N\hat{x}}{N} = \hat{x}.$$

Meanwhile the variance of the distribution of $\bar{X}$ is

$$\sigma_{\bar{X}}^2 = \mathrm{var} \sum_{i=1}^N \frac{X_i}{N} = \sum_{i=1}^N \frac{\sigma^2}{N^2} = \frac{\sigma^2}{N}.$$

Now let $\epsilon > 0$. Then $\exists\, a > 0$ with $\epsilon = a\, \sigma_{\bar{X}}$. Hence by Chebyshev's inequality we have

$$\lim_{N\to\infty} \mathrm{P}\left(|\bar{X} - \hat{x}| > \epsilon\right) \leq \lim_{N\to\infty} \frac{\sigma_{\bar{X}}^2}{\epsilon^2} = \lim_{N\to\infty} \frac{\sigma^2}{N\epsilon^2} = 0.$$

The probability can't be less than 0, so we're done. □

The above proof relies on the PDF having a finite variance. As it turns out, the Weak LLN is true even when the variance is infinite! This can be proved using characteristic functions. But since we don't introduce characteristic functions until Section 3, and since we assume in practice that our data are drawn from PDFs with finite variance anyway, we direct the reader elsewhere. For example, a proof can be found on Wikipedia [3].

For completeness we also list the Strong LLN, but without proof. Like the Weak LLN, the Strong LLN is true even when the PDF variance is infinite.

Theorem 3 (Strong LLN)**.**

$$\bar{X} \xrightarrow{AS} \hat{x}.$$

2 The normal distribution

Now we're going to focus on results about the normal distribution specifically. This first proposition will aid us in some of the calculations.

Proposition 3. *Let $\alpha > 0$. Then*

$$\int_{-\infty}^{\infty} dx\, e^{-\alpha x^2} = \sqrt{\frac{\pi}{\alpha}}.$$

Proof. The trick is to just square the LHS:

$$\left(\int_{-\infty}^{\infty} dx\, e^{-\alpha x^2}\right)^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} dx\, dy\, e^{-\alpha(x^2+y^2)} = \int_0^{\infty} r\, dr \int_0^{2\pi} d\theta\, e^{-\alpha r^2} = \frac{\pi}{\alpha}.$$

□

For the next result let X_1 and X_2 be two independent random variables drawn from normal distributions with respective means $\hat{x}_1$ and $\hat{x}_2$ and standard deviations σ_1 and σ_2.

Proposition 4. *The random variable $Y = X_1 + X_2$ is normally distributed with mean $\hat{x}_1 + \hat{x}_2$ and variance $\sigma_1^2 + \sigma_2^2$.*

Proof. By Proposition 2, the sum Y has the distribution

$$g(y) = \frac{1}{2\pi\sigma_1\sigma_2}\int_{-\infty}^{\infty} dx\, \exp\left[-\frac{(x-\hat{x}_1)^2}{2\sigma_1^2} - \frac{(y-x-\hat{x}_2)^2}{2\sigma_2^2}\right].$$

Pull everything out of the integral that doesn't depend on x, then complete the square with what's left over. One obtains

$$g(y) = \frac{1}{2\pi\sigma_1\sigma_2} \exp\left[-\frac{(y-\hat{x}_1-\hat{x}_2)^2}{2(\sigma_1^2+\sigma_2^2)}\right] \int_{-\infty}^{\infty} dx \, \exp\left[-\left(\frac{\sigma_1^2+\sigma_2^2}{2\sigma_1^2\sigma_2^2}\right)(x+C)^2\right]$$

where C is just a bunch of stuff that doesn't depend on x. Therefore you can make the substitution $u = x + C$ with $du = dx$ and carry out the new integral using Proposition 3. The result is

$$g(y) = \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}} \exp\left[-\frac{(y-\hat{x}_1-\hat{x}_2)^2}{2(\sigma_1^2+\sigma_2^2)}\right].$$

□

Since the normal distribution is so important, so must be its CDF. Unfortunately the integral of the normal PDF is *nonelementary*; that is, it can't be expressed in terms of polynomials or standard functions like sin, cos, or exp. Therefore we give a name to this special function. The *error function* is

$$\operatorname{erf}(x) \equiv \frac{2}{\sqrt{\pi}} \int_0^x dt \, e^{-t^2}. \tag{16}$$

Then we can write the Gaussian CDF with mean 0 as

$$\operatorname{Gau}(x, 0, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} dt \, e^{-t^2/2\sigma^2} = \frac{1}{2} + \frac{1}{2} \operatorname{erf}\left(\frac{x}{\sqrt{2}\sigma}\right). \tag{17}$$

Now we can list some pretty powerful applications of the normal distribution. For instance one often must compare two empirical estimates of some mean. Usually these estimates are different, and one might wonder whether this disparity is real or just plain unlucky. More precisely:

Theorem 4. *Suppose $\bar{X}$ and $\bar{Y}$ are correct estimates of some expectation value, i.e. they are normally distributed with the same mean, and call their respective standard deviations σ_X and σ_Y. Then the probability that $\bar{X}$ and $\bar{Y}$ differ by at least D is*

$$\mathrm{P}\left(|\,\bar{X} - \bar{Y}| > D\right) = 1 - \operatorname{erf}\left(\frac{D}{\sqrt{2(\sigma_X^2+\sigma_Y^2)}}\right).$$

Proof. From Proposition 4, the random variable $\bar{X} - \bar{Y}$ is normally distributed with mean 0 and variance $\sigma_D^2 = \sigma_X^2 + \sigma_Y^2$. Therefore by eq. (17), the probability that $\bar{X}$ and $\bar{Y}$ are at most D apart is

$$\begin{aligned} \mathrm{P}\left(|\,\bar{X} - \bar{Y}| < D\right) &= \mathrm{P}\left(-D < \bar{X} - \bar{Y} < D\right) \\ &= \mathrm{Gau}(D, 0, \sigma_D) - \mathrm{Gau}(-D, 0, \sigma_D) \\ &= 1 - 2\,\mathrm{Gau}(-D, 0, \sigma_D) \\ &= \mathrm{erf}\left(\frac{D}{\sqrt{2}\sigma_D}\right). \end{aligned}$$

And of course, $\mathrm{P}\left(|\,\bar{X} - \bar{Y}| > D\right) = 1 - \mathrm{P}\left(|\,\bar{X} - \bar{Y}| < D\right)$. □

In other words, the above theorem gives the probability that the observed difference $|\bar{X} - \bar{Y}|$ is due to chance. This probability is called the *q-value*. In practice one sets some threshold on q below which one investigates further whether they underlying distrubutions of the estimates are different. Often one takes the threshold as 0.05.

3 The central limit theorem

Let X and Y be real random variables. Then we can construct a complex random variable $F = X + iY$, and its expectation value will be

$$\langle F \rangle = \langle X \rangle + i \langle Y \rangle . \tag{18}$$

This allows us to speak sensibly about Fourier transformations of PDFs. In particular let X be drawn from the PDF f. The *characteristic function* of X is

$$\phi(t) \equiv \left\langle e^{itX} \right\rangle = \int_{-\infty}^{\infty} dx\, e^{itx} f(x). \tag{19}$$

Knowing the characteristic function X is equivalent to knowing its PDF; this is because we can take the inverse Fourier transformation

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} dt\, e^{-itx} \phi(t), \tag{20}$$

which follows from the Dirac δ-function. The derivatives of the characteristic function are easily calculated to be

$$\phi^{(n)}(t) = i^n \int_{-\infty}^{\infty} dx\, x^n e^{itx} f(x); \tag{21}$$

therefore

$$\phi^{(n)}(0) = i^n \langle X^n \rangle . \tag{22}$$

If $|f(x)|$ falls off faster than x^m for any $m \in \mathbb{Z}$, then it follows from the above equation that all moments exist, and the characteristic function is analytic in t about $t = 0$.

These are some neat properties of characteristic functions; however our main use for them is summarized in the next proposition.

Proposition 5. *The characteristic function of a sum of independent random variables equals the product of their characteristic functions.*

Proof. Let X_1,...,X_N be drawn from PDFs f_1,...,f_N with corresponding characteristic functions $\phi_1, ..., \phi_N$, and let $Y = \sum_j X_j$. Then using the definition of the characteristic function we obtain

$$\phi_Y(t) = \left\langle e^{it\sum_j X_j} \right\rangle = \left\langle \prod_{j=1}^{N} e^{itX_j} \right\rangle = \prod_{j=1}^{N} \left\langle e^{itX_j} \right\rangle = \prod_{j=1}^{N} \phi_j(t),$$

where we used independence for the third equality. □

Now suppose you're an experimenter taking independent measurements of some observable. Furthermore suppose you don't know anything about the observable, except that it comes from some distribution with finite variance. The central limit theorem (CLT) says that armed with this information alone, you know that the sample mean will be normally distributed about the true mean. Here is the precise statement.

Theorem 5 (Central limit theorem)**.** *Let $X_1, ..., X_N$ be N independent random variables drawn from PDF f. Suppose further that f has mean $\hat{x}$ and variance σ^2. Then the PDF of the estimator $\bar{X}$ converges to* $\mathrm{gau}(\bar{x}, \hat{x}, \sigma/\sqrt{N})$.

Proof. What we're going to do is look at the characteristic function ϕ_S of the random variable

$$S \equiv \bar{X} - \hat{x} = \frac{X_1 + ... + X_N - N\hat{x}}{N}.$$

If we can show that ϕ_S converges to the characteristic function corresponding to $\mathrm{gau}(s, 0, \sigma/\sqrt{N})$, then we are finished.

In order to show this, we first need the characteristic function for the distribution $\text{gau}(s, 0, \sigma/\sqrt{N})$. By completing the square and using Proposition 3, we find

$$\begin{aligned}\phi_{\text{gau}} &= \frac{1}{\sigma}\sqrt{\frac{N}{2\pi}}\int_{-\infty}^{\infty} ds\, e^{its} \exp\left[-\frac{s^2 N}{2\sigma^2}\right] \\ &= \frac{1}{\sigma}\sqrt{\frac{N}{2\pi}} \exp\left[-\frac{\sigma^2 t^2}{2N}\right] \int_{-\infty}^{\infty} ds\, \exp\left[-\frac{N}{2\sigma^2}(s-C)^2\right] \\ &= \exp\left[-\frac{\sigma^2 t^2}{2N}\right],\end{aligned}$$

where C is a number that doesn't depend on s. It remains to show $\phi_S = \phi_{\text{gau}}$. By Proposition 5 we have

$$\phi_S(t) = \phi_{\frac{1}{N}\sum X_i - \hat{x}}(t) = \left[\phi_{X-\hat{x}}\left(\frac{t}{N}\right)\right]^N,$$

where $\phi_{X-\hat{x}}$ is the characteristic function corresponding to the random variable $X - \hat{x}$. Call its PDF g. From the properties of f, we know that g has mean 0 and variance σ^2. Therefore by expanding ϕ_S about $t = 0$ and using the definition (6), we find

$$\phi_S(t) = \left[1 - \frac{\sigma^2 t^2}{2N^2} + \mathcal{O}\left(\frac{t^3}{N^3}\right)\right]^N = \exp\left[-\frac{\sigma^2 t^2}{2N}\right] + \mathcal{O}\left(\frac{t^3}{N^2}\right),$$

as desired. □

Since the variance of the estimator $\bar{X}$ tends to 0 for large N, it follows that the sample mean converges to the true mean $\hat{x}$. In particular for large N, we expect the true mean to be within $\sigma/\sqrt{N}$ of the estimator roughly 68% of the time. Table 1 gives the area under a Gaussian curve for different numbers of standard deviations away from the mean.

4 Bias

For this section consider independent random variables $X_1, ..., X_N$ drawn from a distribution with mean $\hat{x}$ and variance σ^2. Earlier we recovered the

familiar estimator for the mean, which was just the ordinary arithmetic average. But what about an estimator for the variance? Naively one might write

$$\bar{\sigma}^2_{\text{biased}} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2; \tag{23}$$

While we expect this estimator to converge to the exact result in the limit $N \to \infty$, it disagrees with σ^2 for small N. Most glaringly when $N = 1$, the estimator is zero, regardless of the exact result. An estimator is said to be *biased* when its expectation value does not agree with the exact result. The difference between the expectation value of the estimator and the exact result is correspondingly called the *bias*. When they agree, we say the estimator is *unbiased*.

Proposition 6. *An unbiased estimator of the variance is*

$$\bar{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2.$$

Proof. To construct an unbiased estimator of the variance, we'll determine the bias of the estimator, then remove it. Note

$$\left\langle \bar{\sigma}^2_{\text{biased}} \right\rangle = \frac{1}{N}\sum_{i=1}^{N}\left(\left\langle X_i^2 \right\rangle - 2\left\langle X_i \bar{X} \right\rangle + \left\langle \bar{X}^2 \right\rangle\right).$$

Let us analyze the above equation term by term. Since the random variables X_i are drawn from the same distribution, the first term is an unbiased

Table 1: Table of areas under the curve for the normal distribution. The last column gives the probability that a random variable drawn from the distribution falls at least the given number of error bars away from the mean.

Number of σ from $\hat{x}$	Area under curve	About 1 in ...
1	0.682 689 49	3
2	0.954 499 74	22
3	0.997 300 20	370
4	0.999 936 66	15 787
5	0.999 999 43	1 744 278

estimator of $\langle X^2 \rangle$ for each i. Next the second term can be rewritten as

$$\begin{aligned}\left\langle X_i \bar{X} \right\rangle &= \frac{1}{N}\left(\left\langle X_i^2 \right\rangle + \sum_{j\,|\,j\neq i} \langle X_i X_j \rangle\right) \\ &= \frac{1}{N}\left(\left\langle X^2 \right\rangle + (N-1)\langle X \rangle^2\right) \\ &= \frac{1}{N}\left(\left\langle X^2 \right\rangle - \langle X \rangle^2\right) + \langle X \rangle^2 \\ &= \frac{\sigma^2}{N} + \hat{x}^2,\end{aligned}$$

where in the second line we used the independence of the X_i. Finally for the last term we have

$$\left\langle \bar{X}^2 \right\rangle = \left\langle \frac{1}{N^2}\sum_{i,j} X_i X_j \right\rangle = \frac{1}{N^2}\left(N\left\langle X^2 \right\rangle + \sum_{i,j\,|\,i\neq j} \hat{x}^2\right) = \frac{\sigma^2}{N} + \hat{x}^2,$$

where we again used independence in the second equality. Plugging everything into $\langle \bar{\sigma}^2_{\text{biased}} \rangle$ gives

$$\left\langle \bar{\sigma}^2_{\text{biased}} \right\rangle = \frac{1}{N}\sum_{i=1}^{N}\left(\left\langle X^2 \right\rangle - \frac{\sigma^2}{N} - \hat{x}^2\right) = \left(\frac{N-1}{N}\right)\sigma^2.$$

This equation shows us the bias is $-\sigma^2/N$. Therefore according to this equation, an unbiased estimator of the variance is

$$\bar{\sigma}^2 = \left(\frac{N}{N-1}\right)\bar{\sigma}^2_{\text{biased}} = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

as we wished to show. □

We saw that the bias of the naive variance estimator goes like $1/N$. So one might wonder: How much bias does one typically expect to encounter? Bias problems often appear whenever one wants to estimate some function of the mean $\hat{f} = f(\hat{x})$ that is not necessarily linear near the mean. One might be tempted to take the estimator

$$\bar{f}_{\text{bad}} = \frac{1}{N}\sum_{i=1}^{N} f_i, \tag{24}$$

where $f_i \equiv f(X_i)$. However it turns out that

$$\lim_{N\to\infty} \bar{f}_{\text{bad}} \neq \hat{f}. \tag{25}$$

An estimator that never converges to its true value is called *inconsistent*; otherwise it is *consistent*. So this bad estimator is not a consistent estimator. Note that a biased estimator is not necessarily inconsistent; for instance the biased estimator of the variance eq. (23) is consistent. A consistent estimator of $\hat{f}$ is

$$\bar{f} = f(\bar{X}). \tag{26}$$

We can prove the consistency of $\bar{f}$ for a wide class of functions.

Proposition 7. *Suppose $f : \mathbb{R} \to \mathbb{R}$ has a convergent Taylor series in a region about $\hat{x}$. If $\bar{X}$ maps to this region, then $\bar{f}$ has bias of order $1/N$.*

Proof. If we consider f as a function of the ordinary variable x, we can expand it about $\hat{x}$ as

$$f(x) = f(\hat{x}) + f'(\hat{x})(x-\hat{x}) + \frac{1}{2}f''(\hat{x})(x-\hat{x})^2 + \mathcal{O}\left((x-\hat{x})^3\right).$$

Since $\bar{X}$ maps to the region in which this expansion is valid, we can plug it into the above formula and find its expected value. This gives

$$\left\langle \bar{f} \right\rangle - \hat{f} = f'(\hat{x})\left\langle \bar{X} - \hat{x} \right\rangle + \frac{1}{2}f''(\hat{x})\left\langle (\bar{X}-\hat{x})^2 \right\rangle + \mathcal{O}\left((\bar{X}-\hat{x})^3\right).$$

The LHS of this equation is the bias of $\bar{f}$. To simplify the RHS, note that by the CLT $\left\langle \bar{X} - \hat{x} \right\rangle = 0$ and $\left\langle (\bar{X}-\hat{x})^2 \right\rangle = \sigma^2/N$. Therefore

$$\left\langle \bar{f} \right\rangle - \hat{f} = \frac{1}{2}f''(\hat{x})\frac{\sigma^2}{N} + \mathcal{O}\left(\frac{1}{N^2}\right).$$

□

According to the above proposition, the bias vanishes as $N \to \infty$, which shows that $\bar{f}$ is consistent. For large N, $\bar{X}$ is very likely to be close to $\hat{x}$ by the CLT, so Proposition 7 will essentially hold whenever N is large and f is a nice enough function. There is another important consequence to this proposition: the bias decreases faster than the statistical error bar, which you will recall goes like $1/\sqrt{N}$. Hence when N becomes large enough, the bias can be ignored.

5 Error propagation and covariance

We will now reproduce the commonly used error propagation formula. We know that if we have a smooth function of N variables $f : \mathbb{R}^N \to \mathbb{R}$ we can Taylor expand

$$f(x) = f(\hat{x}) + \sum_{j=1}^{N} \left.\frac{\partial f}{\partial x_j}\right|_{x=\hat{x}} (x_j - \hat{x}_j) + \mathcal{O}\left(x^2\right) \tag{27}$$

where $x = (x_1, ..., x_N)$ and $\hat{x} = (\hat{x}_1, ..., \hat{x}_N)$. So if $|x - \hat{x}|$ is small enough, f is a linear function of the components of x. This motivates the following situation: Suppose we have a set of M random variables Y_i, each of which is a linear function of N random variables X_j; i.e.

$$Y_i = a_{i0} + \sum_{j=1}^{N} a_{ij} X_j. \tag{28}$$

Then the mean is given by

$$\hat{y}_i = \langle Y_i \rangle = a_{i\,0} + \sum_{j=1}^{N} a_{ij} \langle x_j \rangle = a_{i\,0} + \sum_{j=1}^{N} a_{ij} \hat{x}_j. \tag{29}$$

Meanwhile the variance of Y_i is given by

$$\begin{aligned}
\sigma_{Y_i}^2 = \left\langle (Y_i - \hat{y}_i)^2 \right\rangle &= \left\langle \sum_{j=1}^{N} a_{ij}(X_j - \hat{x}_j) \sum_{k=1}^{N} a_{ik}(X_k - \hat{x}_k) \right\rangle \\
&= \sum_{j=1}^{N} a_{ij}^2 \left\langle (X_j - \hat{x}_j)^2 \right\rangle + \sum_{j \neq k} a_{ij} a_{ik} \left\langle (X_j - \hat{x}_j)(X_k - \hat{x}_k) \right\rangle \\
&= \sum_{j=1}^{N} a_{ij}^2 \sigma_{X_j}^2 + \sum_{j \neq k} a_{ij} a_{ik} \operatorname{cov}(X_j, X_k).
\end{aligned} \tag{30}$$

In the case that the X_j and X_k are independent, $\operatorname{cov}(X_j, X_k) = 0$. Furthermore if Y_i is a linear function of the X_j, we can associate the a_{ij} with partial derivatives. Then eq. (30) becomes

$$\sigma_{Y_i}^2 = \sum_{j=1}^{N} \frac{\partial\, Y_i}{\partial X_j} \sigma_{X_j}^2. \tag{31}$$

This is the *error propagation formula* as it's usually stated. Berg [1] emphasizes that this relation is mnemonic because it doesn't make sense to take derivatives with respect to random variables. In practice we can apply eq. (31) when we

1. know how some function f depends on some variables x_i (then taking derivatives with respect to these new variables is well-defined);
2. take measurements of the variables;
3. the measurements are independent; and
4. either the function is exactly linear in the x_i or f is approximately linear in a region close to the mean.

Oftentimes in physics you'll find yourself in a situation where you want to calculate several functions of the same random variables X_j. If the X_j are close enough to their means, or if the Y_i are linear functions of the X_j, this is a situation in which (28) applies. Intuitively one might expect the Y_i to be correlated; this turns out to be the case. In particular

$$\begin{aligned}\operatorname{cov}(Y_i, Y_j) &= \langle (Y_i - \hat{y}_i)(Y_j - \hat{y}_j) \rangle \\ &= \sum_{k,l=1}^{N} a_{ik} a_{jl} \left\langle (X_k - \hat{x}_k)(X_l - \hat{X}_l) \right\rangle \\ &= \sum_{k=1}^{N} a_{ik} a_{jk} \sigma_{X_k}^2 + \sum_{k \neq l} a_{ik} a_{jl} \operatorname{cov}(X_k, X_l),\end{aligned} \tag{32}$$

which shows the Y_i are correlated even if the X_j are not.

6 Jackknife resampling

Let us consider a sample of independent measurements $X_1, ..., X_N$ from some distribution with mean $\hat{x}$ and variance σ^2 and a function f that has a Taylor series expansion near $\hat{x}$, but isn't necessarily linear. From Section 4 we know that $\bar{f} = f(\bar{X})$ is a consistent estimator of $\hat{f} = f(\hat{x})$.

The discussion of Section 5 gives us a way to propagate uncertainty from the random variables to f, but it is effectively unusable whenever f becomes sufficiently complicated. Even when f is simple, if the original data are

correlated, the error propagation formula eq. (31) is still unwieldy. These are some motivations for using the jackknife. The jackknife method is pretty simple to implement, and jackknife error bars agree with usual error bars when there is no bias. Therefore it makes sense to use the jackknife method generally.

Here's how the jackknife method works: We throw away the first measurement from our sample, leaving a data set of $N-1$ resampled values. Statistical analysis is done on this smaller sample. Then we resample again, this time throwing out the second point, and so on. The *jackknife bins* are defined by

$$X_{J,i} \equiv \frac{1}{N-1} \sum_{j \neq i} X_j. \tag{33}$$

They allow us to construct a *jackknife estimator* for the mean $\bar{f}_J$ by

$$\bar{f}_J \equiv \frac{1}{N} \sum_{i=1}^{N} f_{J,i}, \tag{34}$$

where $f_{J,i} \equiv f(X_{J,i})$. The jackknife estimator for the variance of $\bar{f}_J$ is

$$\bar{\sigma}^2_{f_J} = \frac{N-1}{N} \sum_{i=1}^{N} (f_{J,i} - \bar{f}_J)^2. \tag{35}$$

Example. Consider the common problem of calculating the mean of the data and the variance of the mean. Using the unbiased estimator for the variance along with the CLT yields

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i \quad \text{and} \quad \bar{\sigma}^2_{\bar{X}} = \frac{1}{N(N-1)} \sum_{i=1}^{N} (X_i - \bar{X})^2. \tag{36}$$

Meanwhile the jackknife estimator for the variance of $\bar{X}$ gives

$$\bar{\sigma}^2_{\bar{X}_J} = \frac{N-1}{N} \sum_{i=1}^{N} (X_{J,i} - \bar{X}_J)^2. \tag{37}$$

Some simple algebra shows that $(N-1)(X_{J,i} - \bar{X}_J) = \bar{X} - X_i$. Therefore

$$\bar{\sigma}^2_{\bar{X}_J} = \bar{\sigma}^2_{\bar{X}}. \tag{38}$$

Next let's see how the jackknife lets us estimate bias. From Proposition 7 we know the bias of the estimator $\bar{f}$ is of order $1/N$, which we will write

$$\text{bias } \bar{f} = \frac{A}{N} + \mathcal{O}\left(\frac{1}{N^2}\right) \tag{39}$$

for some constant A. Let us determine the bias of $\bar{f}_J$.

Proposition 8. *If the measurements X_i are distributed relatively close to $\hat{x}$, then $\bar{f}_J$ has a bias of order $1/(N-1)$.*

Proof. The assumption on the measurements is that they roughly fall within the series' radius of convergence. We rewrite

$$X_{J,i} = \hat{x} + \frac{1}{N-1}\sum_{j\neq i}(X_j - \hat{x}).$$

Then our strategy is the same as before: We expand f in the same sense as before, and take the average value of $f_{J,i}$. We obtain

$$\begin{aligned}
\langle f_{J,i}\rangle &= \langle f(X_{J,i})\rangle \\
&= \left\langle f\left(\hat{x} + \frac{1}{N-1}\sum_{j\neq i}(X_j - \hat{x})\right)\right\rangle \\
&= \hat{f} + \frac{1}{2}f''(\hat{x})\frac{1}{(N-1)^2}\sum_{\substack{j\neq i\\ k\neq i}}\langle (X_j-\hat{x})(X_k-\hat{x})\rangle + \mathcal{O}\left(\frac{1}{N^2}\right) \\
&= \hat{f} + \frac{1}{2}f''(\hat{x})\frac{1}{(N-1)^2}\left(\sum_{j\neq i}\sigma^2 + \sum_{j\neq k}\operatorname{cov}(X_j, X_k)\right) + \mathcal{O}\left(\frac{1}{N^2}\right) \\
&= \hat{f} + \frac{1}{2}f''(\hat{x})\frac{1}{N-1}\sigma^2 + \mathcal{O}\left(\frac{1}{N^2}\right),
\end{aligned}$$

where in third equality we used $\langle X_j - \hat{x}\rangle = 0$ and in the last equality we used the independence of the measurements. Since the RHS is independent of i, it follows that

$$\left\langle \bar{f}_J\right\rangle - \hat{f} = \frac{1}{2}f''(\hat{x})\frac{\sigma^2}{N-1} + \mathcal{O}\left(\frac{1}{N^2}\right).$$

□

Comparing the final steps of Propositions 7 and 8, we see that they have the same lowest order contribution, except that N is replaced by $N-1$. Therefore we can write

$$\text{bias } \bar{f}_J = \frac{A}{N-1} + \mathcal{O}\left(\frac{1}{N^2}\right) \tag{40}$$

with the same constant A as with eq. (39). Combining both of these equations, we conclude

$$A = N(N-1)\left(\left\langle \bar{f} \right\rangle - \left\langle \bar{f}_J \right\rangle\right) + \mathcal{O}\left(\frac{1}{N}\right), \tag{41}$$

which means that

$$\overline{\text{bias}} = (N-1)(\bar{f} - \bar{f}_J) \tag{42}$$

gives an estimator for the bias of $\bar{f}$, at least up to $\mathcal{O}(1/N^2)$.

Equation (39) shows that the bias decreases faster than the error bar. However it can happen that if A is large, and if N is relatively small, the bias is non-negligible. In practice one should be concerned if the bias is of the same order as the error bar. In such a case one should attempt to correct for this bias. Using eq. (42) it follows that

$$\bar{f}_C \equiv \bar{f} - \overline{\text{bias}} \tag{43}$$

is bias-corrected, in that it has $\mathcal{O}(1/N^2)$ disagreement with the true mean. To build an estimator for the variance of the bias corrected mean, we must do another level of jackknifing. Let $i \neq j$. The *second-level jackknife bins* are

$$X_{J,ij} \equiv \frac{1}{N-2} \sum_{k \neq i,j} X_k. \tag{44}$$

They allow us to construct *second-level jackknife estimators* by

$$f_{J,ij} \equiv f(X_{J,ij}). \tag{45}$$

We can use these second-level estimators to create jackknife samples of bias estimators and bias-corrected estimators with

$$\text{bias}_{J,i} = \frac{1}{N-1} \sum_{k \neq i} (f_{J,i} - f_{J,ik}) \quad \text{and} \quad f_{CJ,i} = f_{J,i} - \text{bias}_{J,i}. \tag{46}$$

With these definitions, we can then calculate estimators for the mean and variance of the mean using the same machinery as the original jackknife, i.e. from eq. (34) and (35). We get

$$\bar{f}_{CJ} = \frac{1}{N}\sum_{i=1}^{N} f_{CJ,i} \quad \text{and} \quad \bar{\sigma}^2_{f_{CJ}} = \frac{N-1}{N}\sum_{i=1}^{N}(f_{CJ,i} - \bar{f}_{CJ})^2, \tag{47}$$

which takes correlations between $f_{J,i}$ and $\text{bias}_{J,i}$ automatically into account. This completes the tool set needed to estimate average values of functions of data, correcting for bias and correlation.

7 The χ^2 distribution and fitting data

Consider a sample of N Gaussian, independent data points (X_i, Y_i), where the Y_i have standard deviations σ_i. For now we will assume the X_i have no error. We will consider a situation where we believe the Y_i are measurements of some real function y of x. Abstractly we model these data with a fit that depends on some set of M parameters

$$y = y(x; a), \tag{48}$$

where $a = (a_1, ..., a_M)$ is the vector of these parameters. Our goal is to estimate the a_j and their error bars, and then determine whether this fit is consistent with the data.

Assuming that $y(x, a)$ is the exact law for the data, the joint PDF of the measurements Y_i is given by eq. (8) to be

$$f(y_1, ..., y_N) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[\frac{-(y_i - y(x_i; a))^2}{2\sigma_i^2}\right]. \tag{49}$$

The PDF given by eq. (49) is an example of the *non-central* χ^2 *distribution*. Generally this distribution has random variable

$$X^2 = \sum_{i=1}^{N} \frac{(Y_i - \hat{y}_i)^2}{\sigma_i^2}, \tag{50}$$

where the random variables Y_i are drawn from $\text{gau}(y, \hat{y}_i, \sigma_i)$. In the special case that the Y_i are drawn from $\text{gau}(y, 0, 1)$ we obtain the random variable

$$X^2 = \sum_{i=1}^{N} Y_i^2. \tag{51}$$

In this case the PDF of X^2 is called the χ^2 *distribution*. It simplifies to

$$f(y_1, ..., y_N) = \frac{1}{(2\pi)^{N/2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{N} y_i^2\right]. \tag{52}$$

Let's think about a general, non-central χ^2 PDF. The likelihood that the data fall within a region near what was observed is

$$\mathrm{P} = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[\frac{-(y_i - y(x_i; a))^2}{2\sigma_i^2}\right] dy_i. \tag{53}$$

Our strategy for determining the correct fit will be to find the vector a that maximizes the above probability. The happens when the argument of the exponential is closest to zero; i.e. when

$$\chi^2 \equiv \sum_{i=1}^{N} \frac{(y_i - y(x_i; a))^2}{2\sigma_i^2} \tag{54}$$

is minimized. This is an example of a *maximum likelihood method*. Once the parameters are found, one can then ask: What is the probability that the discrepancy between the data and the fit is due to chance?

To answer this question, let us begin with the simpler case using the χ^2 CDF (52). It is given by

$$F(\chi^2) = \mathrm{P}\left(X^2 \leq \chi^2\right) = \frac{1}{(2\pi)^{N/2}} \int_{\sum y_i^2 \leq \chi^2} \prod dy_i \; e^{-y_i^2/2}. \tag{55}$$

Switching to hyperspherical coordinates, this becomes

$$F(\chi^2) = \frac{1}{(2\pi)^{N/2}} \int d\Omega \int_0^{\chi} dr \; r^{N-1} e^{-r^2/2}. \tag{56}$$

The RHS looks similar to the gamma function. With this in mind, we can make the substitution $t = r^2/2$ and use Proposition **??** to obtain

$$F(\chi^2) = \frac{1}{\Gamma(N/2)} \int_0^{\chi^2/2} dr \; t^{N/2-1} e^{-t}. \tag{57}$$

The integral

$$\Gamma(s, z) \equiv \frac{1}{\Gamma(s)} \int_0^{z} dt \; t^{s-1} e^{-t} \tag{58}$$

with $\mathrm{Re}\, s > 0$ is called the *incomplete gamma function*. The CDF in the form (57) is well-suited for numerical calculation because it is straightforward to compute the incomplete gamma function.

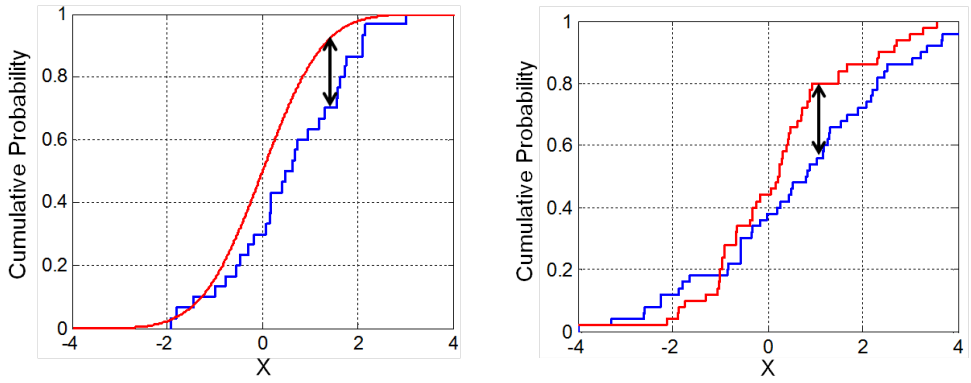

Figure 1: An example Kolmogorov statistic comparing an empirical CDF with a known, exact CDF (left) and a Kolmogorov statistic comparing two empirical CDFs (right). The statistic is indicated by the black, double-sided arrow. Images taken from Wikipedia [4].

8 The Kolmogorov test

Let's say we perform an experiment and extract a CDF $\bar{F}$ from the data. How can we tell whether the data are consistent with some true CDF F? Or if we extract another CDF $\bar{G}$, how can we tell whether $\bar{F}$ and $\bar{G}$ are consistent with each other? These questions can be answered using the *Kolmogorov test*. The statistic we will use to determine consistency is the largest difference between the CDFs, shown in Figure 1.

We start with some preliminary definitions. The *indicator function* of a subset A of a set B is the function $\mathbf{1}_A : B \to \{0, 1\}$ given by

$$\mathbf{1}_A(x) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{otherwise.} \end{cases} \tag{59}$$

Given some measurements $X_1, ..., X_N$, we construct the *empirical CDF* as

$$\bar{F}(x) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{[X_i, \infty]}(x), \tag{60}$$

where each term counts the number of data less than or equal to x. The

measurements have the same CDF, so at each x we have by the LLN

$$\begin{aligned}\bar{F}(x) \xrightarrow{\mathrm{P}} \left\langle \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{[X_i,\infty]}(x) \right\rangle &= \frac{1}{N}\sum_{i=1}^{N} \left\langle \mathbf{1}_{[X_i,\infty]}(x) \right\rangle \\ &= \frac{1}{N}\sum_{i=1}^{N} \mathrm{P}\left(X_i \leq x\right) \\ &= \frac{1}{N}\sum_{i=1}^{N} F(x) \\ &= F(x),\end{aligned} \tag{61}$$

i.e. $\bar{F}$ is an unbiased, consistent estimator.

The *Kolmogorov statistic* is

$$\Delta \equiv \max_{x\in\mathbb{R}} \left|\bar{F}(x) - F(x)\right|. \tag{62}$$

Since $\bar{F}(x) \xrightarrow{\mathrm{P}} F(x)$ for all x, it follows that $\Delta \xrightarrow{\mathrm{P}} 0$. An important but surprising fact about Δ is that it is *distribution free*. The proof is relatively straightforward and given in Theorem 6. Theorem 7 gives the probability that the difference between the empirical and true CDFs is at least as extreme as what we calculate. This proof is tedious and not particularly enlightening, so it has been omitted. If you want to see a proof you can read Berg [1], who attributes it to Birnbaum and Tingey [5] and Smirnov [6].

Theorem 6. *All continuous F have the same Δ.*

Proof. We will start with the slightly easier case where F is monotonically increasing. In this case, F^{-1} exists and is also monotonically increasing. Then by making the variable change $y = F(x)$ we find

$$\begin{aligned}\Delta &= \max_{x\in\mathbb{R}} \left|\bar{F}(x) - F(x)\right| \\ &= \max_{y\in[0,1]} \left|\bar{F}\left(F^{-1}(y)\right) - F\left(F^{-1}(y)\right)\right| \\ &= \max_{y\in[0,1]} \left|\bar{F}\left(F^{-1}(y)\right) - y\right|.\end{aligned}$$

Let's focus on the $\bar{F}(F^{-1}(y)$ term. We can recast this as

$$\bar{F}\left(F^{-1}(y)\right) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{[X_i,\infty]}\left(F^{-1}(y)\right) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{[F(X_i),\infty]}(y),$$

where in the second step we used $F(X) \leq y \Leftrightarrow X \leq F^{-1}(y)$. This shows that the empirical CDF of $F^{-1}(y)$ is none other than the empirical CDF of the sample $F(X_1), ..., F(X_N)$. But

$$\mathrm{P}\left(F(X) \leq t\right) = \mathrm{P}\left(X \leq F^{-1}(t)\right) = F\left(F^{-1}(t)\right) = t,$$

i.e. the sample is drawn from the uniform distribution in the interval $[0, 1]$, regardless of what F is. It follows that $\bar{F}\left(F(y)\right)$ is independent of F, and hence so is Δ.

In the case F is not monotonic, the inverse is not guaranteed. We define

$$F^{+}(y) \equiv \min\{x \,|\, F(x) \geq y\}.$$

The important thing is that $F^{+}(y) \leq z \Leftrightarrow y \leq F(z)$. To see this note that

$$F^{+}(y) \leq z \Rightarrow y \leq F\left(F^{+}(y)\right) \leq F(z)$$

and

$$y \leq F(z) \Rightarrow F^{+}(y) \leq F^{+}\left(F(z)\right) \leq z.$$

The theorem follows by replacing F^{-1} with F^{+} in the first paragraph. □

Theorem 7. *Let $D > 0$. Then*

$$\mathrm{P}\left(\Delta > D\right) = \sum_{k=0}^{K} \binom{N}{k} D \left(D + \frac{k}{N}\right)^{k-1} \left(1 - D - \frac{k}{N}\right)^{N-k},$$

where K is determined by the condition that $1 - D - k/N$ cannot be negative. If N is large enough, this probability can be approximated as

$$\mathrm{P}\left(\Delta > D\right) \approx e^{-2ND^2}.$$

We now have all the ingredients we need to carry out a Kolmogorov test, and by Theorem 6 we are guaranteed it will work, regardless of the underlying distribution, under the modest assumption that its CDF is continuous. In practice one can proceed as follows:

1. Sort the measurements, then place them into an array $\{X_i\}$.

2. The X_i are indexed by i, so the corresponding empirical CDF is just the array of fractions $\{i/n\}$. For example $X_1 \leq X_1$, so $\bar{F}(X_1) = 1/N$. Since X_1, $X_2 \leq X_2$ we have $\bar{F}(X_2) = 2/N$.

3. Compute the exact PDF. For instance if we think the data come from $\mathrm{Gau}(x, 0, \sigma)$, we can compute the exact CDF using eq. (17); or if we think the data come from the uniform distribution on $[0, 1]$, we can just heapsort them.

4. Determine Δ.

5. Calculate $\mathrm{P}(\Delta > D)$ using Theorem 7. If the calculation of the exact probability is slow, you can use the approximation, but only if N is big enough, say larger than 100 or so.

6. The underlying assumption is that empirical CDF is an estimator for true CDF, so if the probability is below some threshold, say 0.05, then this assumption becomes suspect.

9 Statistical analysis of Markov chains

Suppose we have computed using MCMC a time series of N measurements $\{X_1, ..., X_N\}$. In principle each element of this sample is drawn from a PDF with mean $\langle X_i \rangle = \langle X \rangle = \hat{x}$ and variance $\sigma^2 = \langle (X_i - \hat{x})^2 \rangle$, i.e. they all have the same mean and variance. Unbiased estimators for the mean and variance are

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i \quad \text{and} \quad \bar{\sigma}^2 = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \bar{X})^2. \tag{63}$$

The variance of the random variable $\bar{X}$ is

$$\sigma_{\bar{X}}^2 = \left\langle (\bar{X} - \hat{x})^2 \right\rangle = \frac{1}{N^2} \left(\sum_{i \neq j} \langle X_i X_j \rangle + N \left\langle X^2 \right\rangle \right) - \hat{x}^2. \tag{64}$$

In the case that the measurements are uncorrelated, the expected values factorize, and we obtain

$$\sigma_{\bar{X}}^2 = \sigma^2 / N \tag{65}$$

in agreement with the CLT. But in practice measurement $i + 1$ is often correlated with measurement $i+t$ because they are from the same time series. To measure this we draw inspiration from definition (11). The *autocovariance* between measurements X_i and X_{i+t} is

$$c(X_i, X_{i+t}) \equiv \langle (X_i - \hat{x})(X_{i+t} - \hat{x}) \rangle = \langle X_i X_{i+t} \rangle - \langle X_i \rangle \langle X_{i+t} \rangle, \tag{66}$$

For a Markov process in equilibrium, the autocorrelation depends only on the separation t, so we define $c(t) \equiv c(X_i, X_{i+t})$. Finally note that $c(0) = \sigma^2$, which motivates the definition of the *autocorrelation*

$$\gamma(t) \equiv \frac{c(t)}{\sigma^2}. \tag{67}$$

The autocorrelation decays in t as a sum of exponentials. I don't know why this is true, and I couldn't find a reference, but this is what everybody says. Assuming this is the case we can write

$$\gamma(t) = A_{\text{exp}}\, e^{-t/\tau_{\text{exp}}} + \sum_{i=1}^{\infty} A_i\, e^{-t/\tau_i}, \tag{68}$$

where the As are constants and we have picked out the leading exponential behavior; i.e. for all i

$$\tau_{\text{exp}} > \tau_i. \tag{69}$$

τ_{exp} is called the *exponential autocorrelation time*.

Plugging definition (66) into eq. (64) we have

$$\sigma^2_{\bar{X}} = \frac{1}{N^2} \sum_{i,j} c(X_i, X_j). \tag{70}$$

In the last sum, $|i - j| = 0$ occurs N times, and $|i - j| = t$ occurs $2(N - t)$ times. Note $1 \leq t \leq N - 1$. Therefore

$$\sigma^2_{\bar{X}} = \frac{1}{N^2} \left(N\, c(0) + 2 \sum_{t=1}^{N-1} (N - t) c(t) \right). \tag{71}$$

Finally we use $c(0) = \sigma^2$ to find

$$\sigma^2_{\bar{X}} = \frac{\sigma^2}{N} \left(1 + 2 \sum_{t=1}^{N-1} \left(1 - \frac{t}{N} \right) \gamma(t) \right) \equiv \frac{\sigma^2}{N} \tau_{\text{int}}. \tag{72}$$

The quantity

$$\tau_{\text{int}} = \left(1 + 2 \sum_{t=1}^{N-1} \left(1 - \frac{t}{N} \right) \gamma(t) \right) \tag{73}$$

is called the *integrated autocorrelation time*.

From eq. (72) we see that τ_{int} is just the ratio between the estimated variance of the sample mean and what this variance would have been if the data were uncorrelated.

In practice, we often don't know the true mean $\hat{x}$ of the time series. Therefore along the lines of eq. (63), we construct an unbiased estimator of the autocovariance

$$\bar{c}(t) = \frac{N}{(N-1)(N-t)} \sum_{i=1}^{N-t} (X_i - \bar{X})(X_{i+t} - \bar{X}), \tag{74}$$

where it is the factor $N/(N-1)$ that removes the bias, just as before. Also in most situations we work in the limit where N is large. In this limit, we can construct an estimator for τ_{int} by

$$\bar{\tau}_{\text{int}}(n) = 1 + 2 \sum_{t=1}^{n} \bar{\gamma}(t), \tag{75}$$

where $n < N$. To understand the above estimator look at definition (73). When t is small, $1 - t/N \approx 1$. Large t terms are doubly suppressed by the exponential decay of $\gamma(t)$ and by $1 - t/N \approx 0$. If the estimator still makes you uncomfortable, note that in the overly simplistic case where $\gamma(t)$ has only one exponential term, one can prove

$$\lim_{N \to \infty} \tau_{\text{int}} = 1 + 2 \sum_{t=1}^{\infty} \gamma(t), \tag{76}$$

which parallels eq. (75) more closely. To construct a final estimator for τ_{int}, one looks for a window in n for which eq. (75) becomes roughly independent of n. This serves as the final $\bar{\tau}_{\text{int}}$.

References

[1] B. A. Berg. *Markov Chain Monte Carlo Simulations and Their Statistical Analysis*. World Scientific, Singapore, 2004.

[2] C. Shalizi. Reminder no. 1: Uncorrelated vs. independent, 2013. URL http://www.stat.cmu.edu/~cshalizi/uADA/13/reminders/uncorrelated-vs-independent.pdf. [Online; accessed 25-May-2017].

[3] Wikipedia contributors. Law of large numbers — Wikipedia, the free encyclopedia, 2017. URL https://en.wikipedia.org/w/index.php?title=Law_of_large_numbers&oldid=817455983. [Online; accessed 17-February-2018].

[4] Wikipedia contributors. Kolmogorov Smirnov test — Wikipedia, the free encyclopedia, 2017. URL https://en.wikipedia.org/w/index.php?title=Kolmogorov%E2%80%93Smirnov_test&oldid=816607137. [Online; accessed 17-February-2018].

[5] Z. W. Birnbaum and F. H. Tingey. One-sided confidence contours for probability distribution functions. *Ann. Math. Statist.*, 22(4):592–596, 1951.

[6] N. Smirnoff. Sur les écarts de la courbe de distribution empirique. *Mat. Sbornik*, 6:3–26, 1939.