Equilibration and Uncertainty

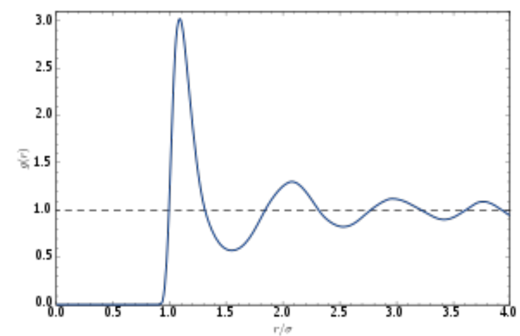
Or, what do my simulations actually mean?

Prof. Michael Shirts
University of Colorado, Boulder
7th i-CoMSE Workshop: Molecular Dynamics
Boise State University, July 8, 2024

What comes out of an MD simulation?

- Time series of the energies and forces (+ other derived quantities like pressure, etc).
- Time series of the coordinates/velocities

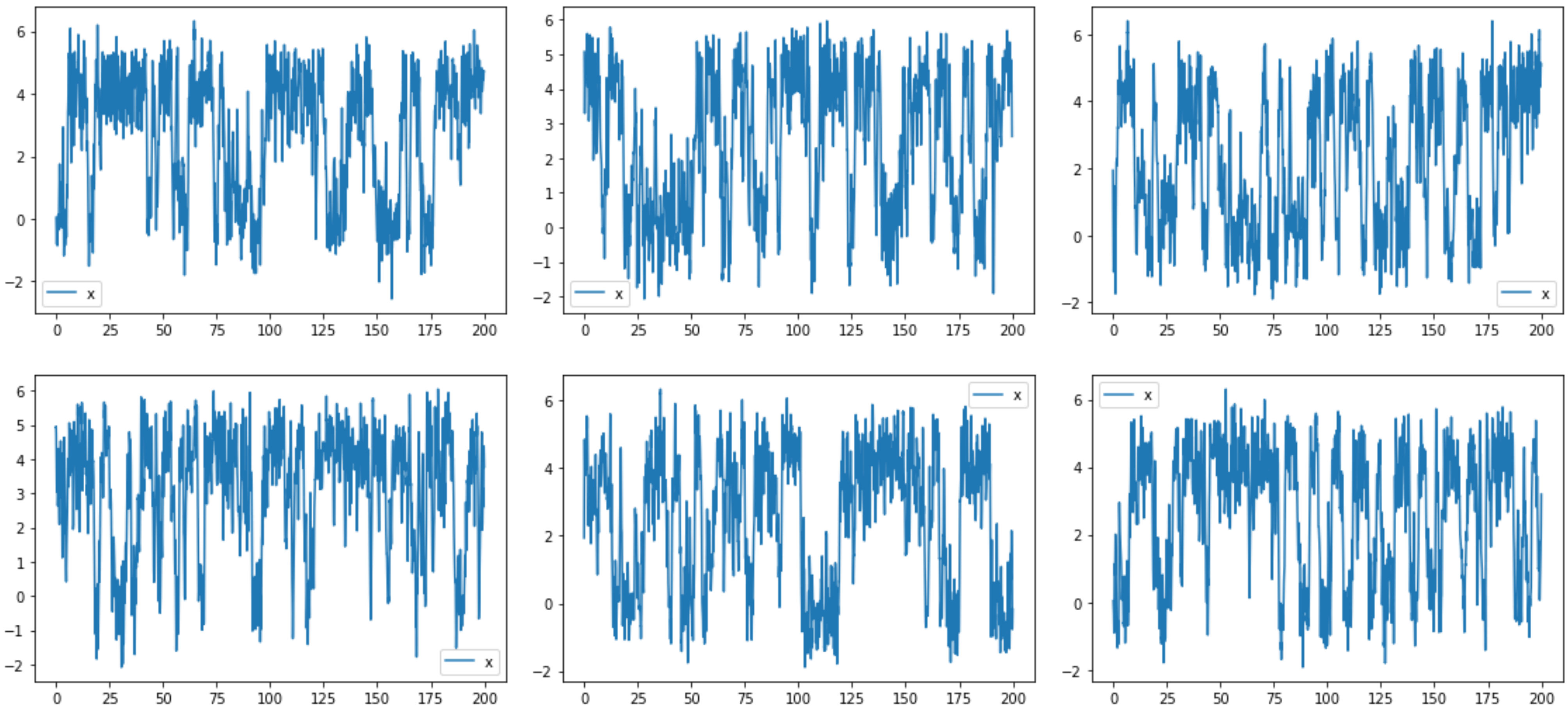
- The **key problem of MD analysis** is **how to extract data of interest** when you have 3N x number of frames data points!
  - How do you take these time series of data points and turn them into useful data?
  - A problem of **data reduction**: remove the noise, extract the meaning
  - Radial distribution functions are one example



i-CoMSE

Problem: NVT or NPT simulations are stochastic

- If we run out simulation multiple times for a finite length, we will get different answers



Uncertainties in averages

- Average in the mean is:
  - $\langle U \rangle = \int p(x) U(x) dx$
  - But if we sample from the distribution $p(x)$, we can replace **integral** with a **sum over observations**
  - $\langle U \rangle = \frac{1}{N} \sum_i U(x_i)$ with x_i sampled from $p(x)$

How do we find the uncertainty of an an estimate of an average?

$$\langle U \rangle = \frac{1}{N} \sum_i U(x_i)$$

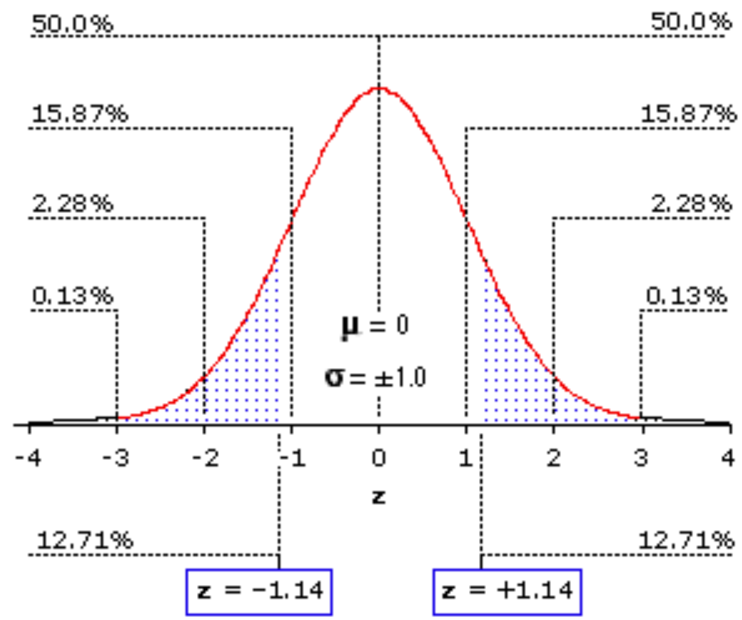
- What do we **mean** by the uncertainty?
- What we generally mean:
  - If we did the same experiment again and again, how different would the results of each experiment be?

- Note: above analysis is for a single variable, but you can extend the same ideas to more complex observables, like an RDF

How do we REPORT the uncertainty?

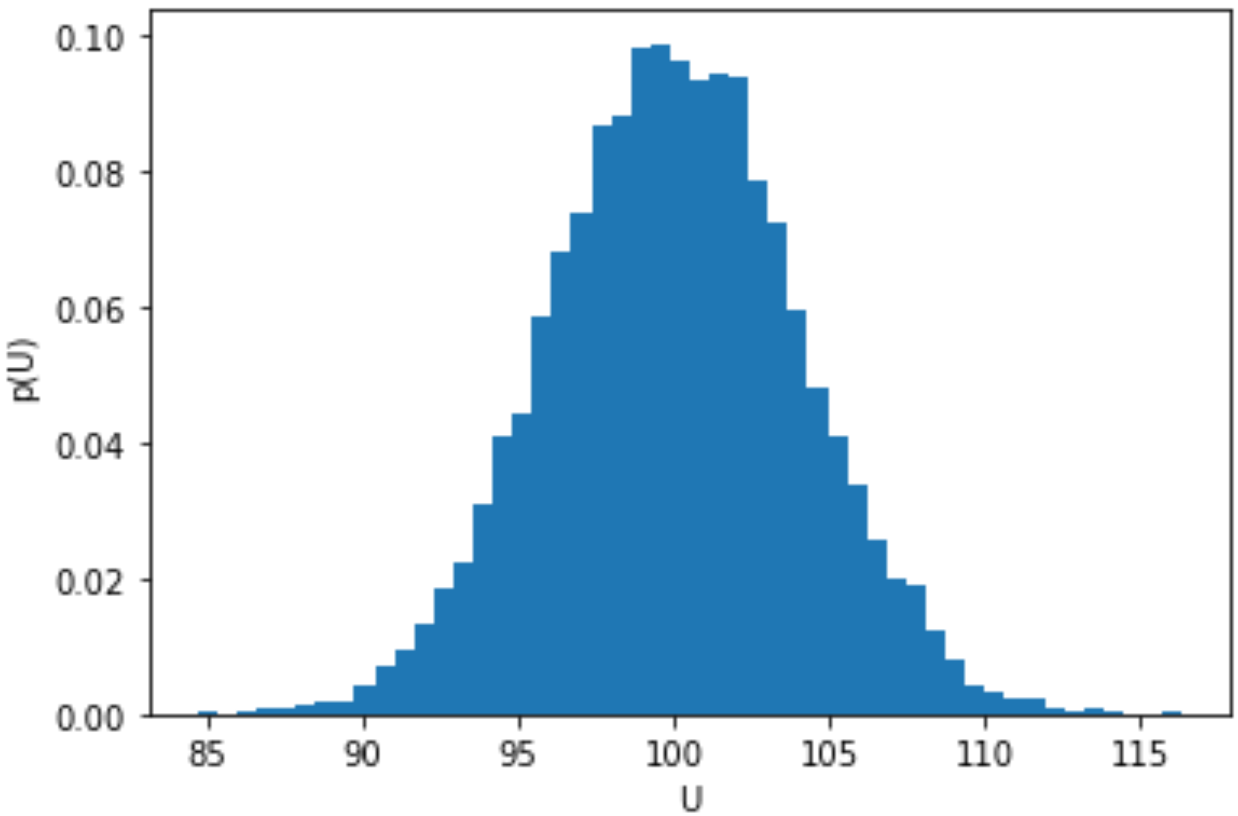
$$\langle U \rangle = \frac{1}{N} \sum_i U(x_i)$$

- We will get a <u>distribution</u> of answers.
- We usually report +/- something. What is that something?
- We usually report a "standard error of the mean".
- What does that mean?



Part 1: Assume we have a distribution for U

This is the distribution of a single sample from U

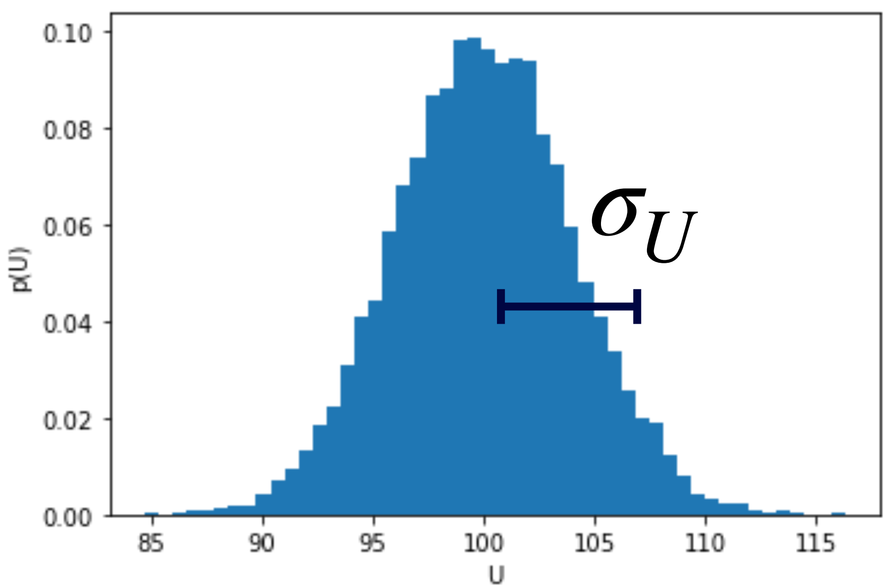


Part 2: What is the standard deviation of this distribution?

- Sample standard deviation
  - An estimate of the standard deviation, computed from samples

$$\sigma_U = \sqrt{\frac{\sum_{i=1}^{N} (U - \langle U \rangle)^2}{N-1}}$$

N is the number of independent samples you are calculating this from

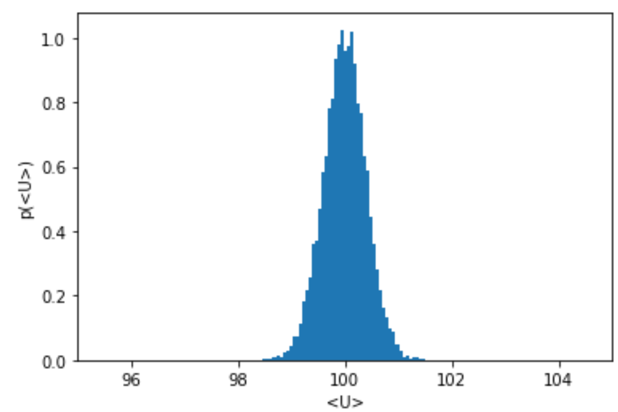


- σ_U does not change in magnitude as you collect more samples, just gets more precise

Part 3: What is the standard deviation in of the MEAN of these samples?

- We don't want the error in 1 sample from U
- We want the errors in N samples from U **averaged together:** $\langle U \rangle = \frac{1}{N}\left(U_1 + U_2 + \ldots + U_N\right)$
- We know that errors add in the square, WHEN they are independent
- $\sigma^2_{\langle U \rangle} = \frac{1}{N^2}\sum_i^N \sigma^2_U$
- $\sigma^2_{\langle U \rangle} = \frac{N}{N^2}\sigma^2_U = \frac{1}{N}\sigma^2_U$
- $\sigma_{\langle U \rangle} = \frac{1}{\sqrt{N}}\sigma_U$

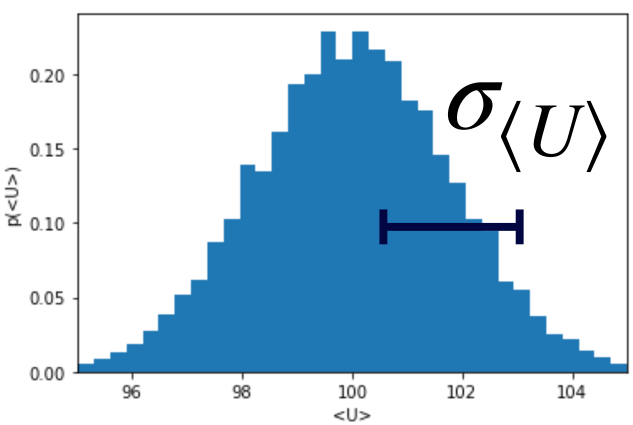
N is the number of independent samples you are averaging together



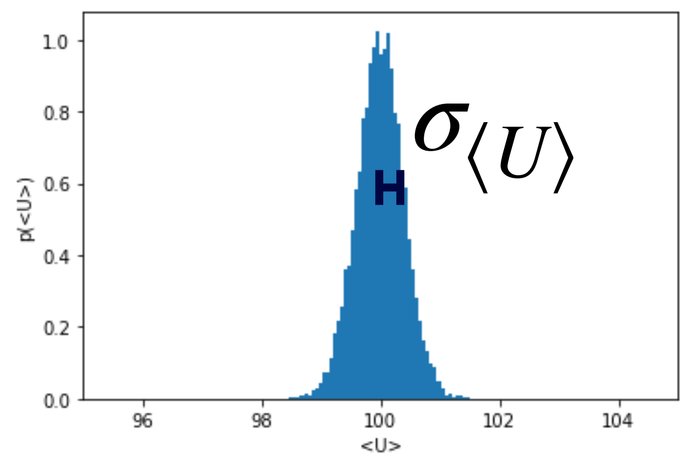
Part 3: What is the standard deviation in of the MEAN of these samples?

$$\sigma_{\langle U \rangle} = \frac{1}{\sqrt{N}} \sigma_U$$

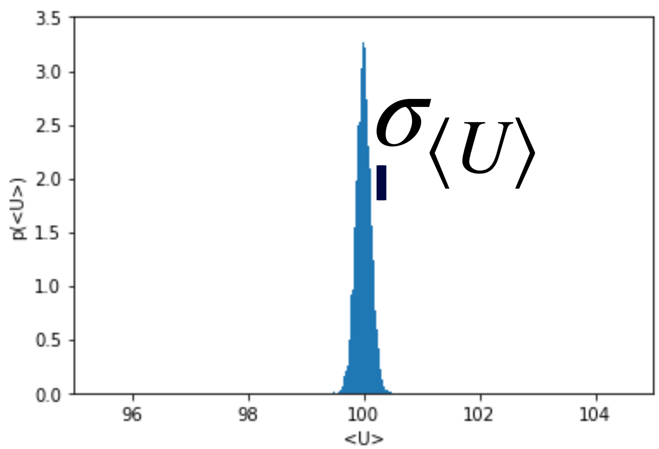
- $\sigma_{\langle U \rangle}$ **does** change as the number of samples **included in the average** increases
- It gets smaller!



N=5



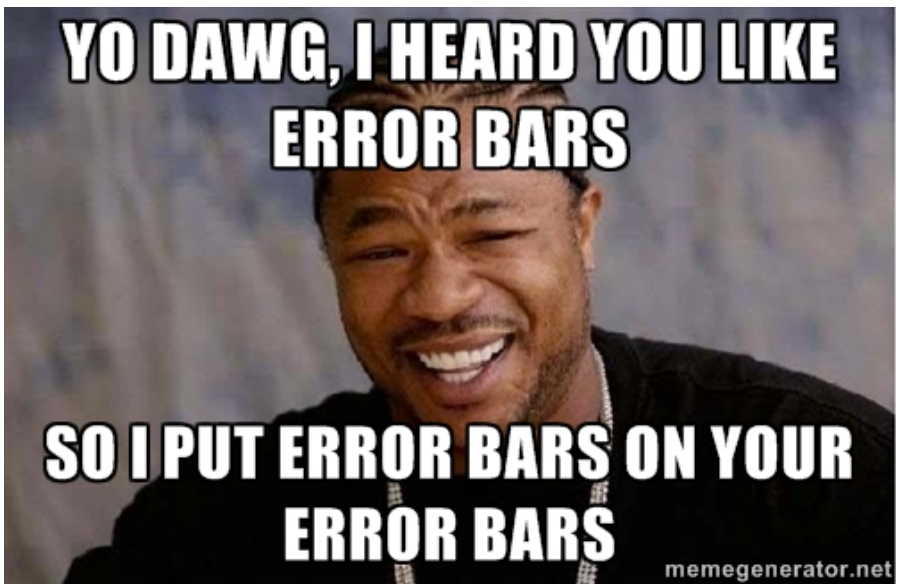
N=100



N=1000

What are the error bars in my error bars?

$$\sigma_{\langle U \rangle} \pm \sigma_{\sigma \langle U \rangle}?$$



- Rough rule of thumb:
- If you have 40 independent samples, then the error in the standard error is <5%
- If you have 5 independent samples, use student t-distribution.

Before we can average: two main tasks

- Identifying when my simulations have become *stationary*.
- Identifying how many simulation points are *independent*.

Go to notebook!

What about distributions that are not normal?

- How to report confidence intervals in those distributions?
- How to report confidence intervals of the <u>means</u> of those distributions?

Go to notebook!

Identifying time correlation in simulation data

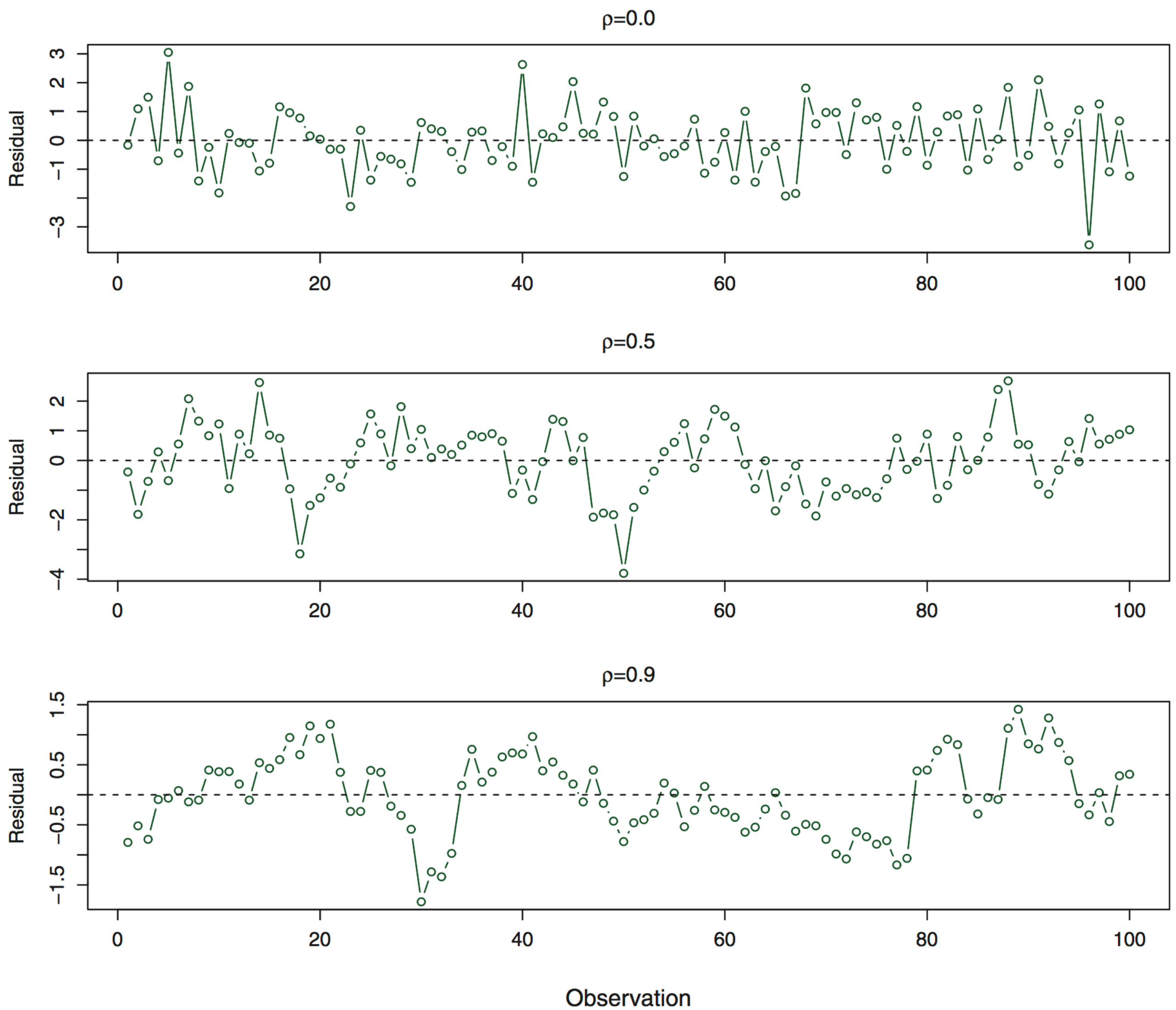


FIGURE 3.10. *Plots of residuals from simulated time series data sets generated with differing levels of correlation ρ between error terms for adjacent time points.*

Go to notebook!

How many *independent* points do I have?

- The autocorrelation function is defined as:

$$A(\tau) = \frac{1}{\mathrm{Var}(A(t))} \int_0^{\infty} A(t)A(t+\tau)dt$$

Where: $\mathrm{Var}(A(t)) = \left\langle (A - \langle A \rangle)^2 \right\rangle$

- We subtract mean from A so that <A> = 0
- We divide by *Var(A(t))* to make sure that the autocorrelation functions starts at 1 when t=0.
- We usually have discrete samples, so we use:

$$A(\tau) = \frac{1}{\mathrm{Var}(A(t))} \frac{1}{N} \sum_{t=0}^{N} A(t)A(t+\tau)$$

Only meaningful for stationary time series!

Three ways to estimate when samples are independent

- Three ways to determine if they are not correlated anymore:
  - Determine when the autocorrelation function crosses to zero
  - Integrate the time under the autocorrelation function, use that as the correlation time.
  - Fit the autocorrelation to an exponential, estimate the characteristic time from $\exp(-t/\tau)$

Go to notebook!