13 – The Chisq Distribution

Physics 281 – Class 13

Grant Wilson

SERIAL

serialpodcast.org

Can you account for what you were doing during the 21 minutes following class 6 weeks ago?

Class 12 Exercises

```
from matplotlib import pyplot as plt
import numpy as np
plt.ion()

class Walker:
   def __init__(self,x=0.,y=0.):
      self.x = x
      self.y = y

   def takeSteps(self,N):
      for i in range(N):
         v = np.random.choice([-1.,1.])
         if(v<0):
            u = np.random.choice([-1.,1.])
            self.x += u
         else:
            u = np.random.choice([-1.,1.])
            self.y += u

   def calcDisplacement(self):
      return np.sqrt(self.x**2 + self.y**2)
```

How the steps should be taken was not well-defined in the problem so we will accept just about any scheme you implemented.

E12.1 continued.

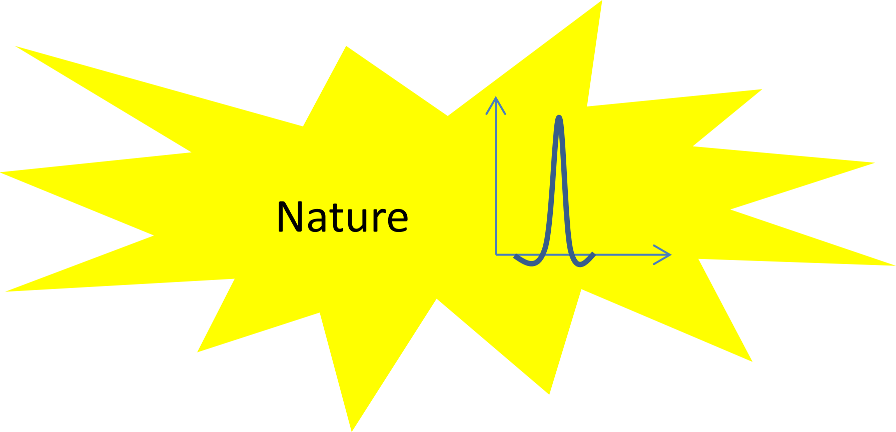
```
#create a list of walkers
zombie = []
nZombies = 10000
for i in range(nZombies):
   zombie.append(Walker())

#now make them walk
nSteps = 100
for i in range(nZombies):
   zombie[i].takeSteps(nSteps)


#plot up the histogram of where they end up
displacement = np.zeros(nZombies)
for i in range(nZombies): displacement[i] = zombie[i].calcDisplacement()
plt.clf()
plt.hist(displacement,bins=40)
plt.xlabel("displacement")
plt.ylabel("counts")
plt.show()
```

Back to the Sample Mean

- Let’s take a step back from the end of last class and talk about what an “experiment” is.

Back to the Sample Mean

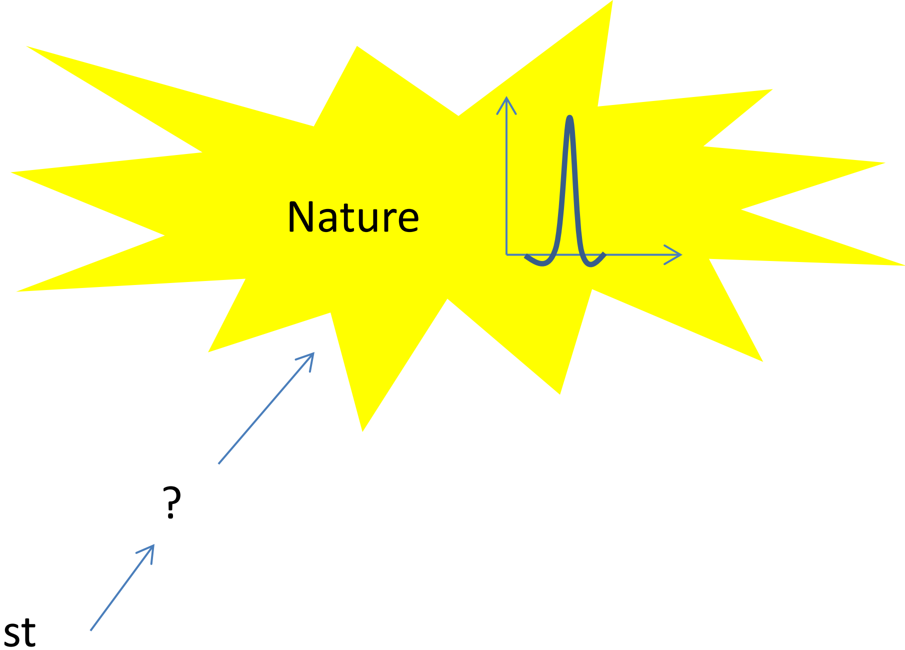
- Let’s take a step back from the end of last class and talk about what an “experiment” is.



Nature

Back to the Sample Mean

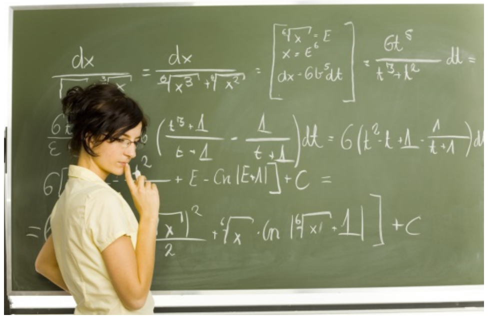
- Let’s take a step back from the end of last class and talk about what an “experiment” is.



Nature

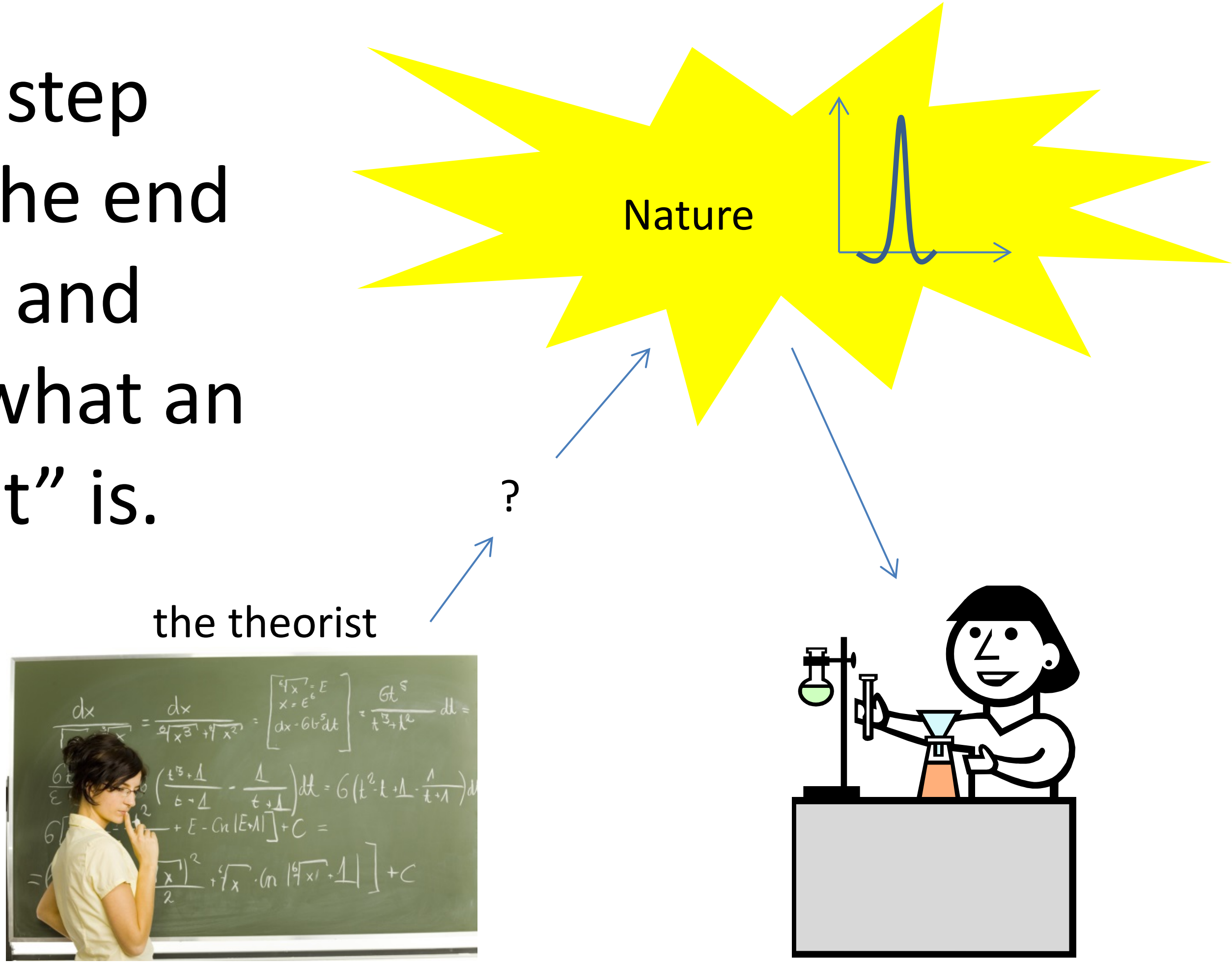
?

the theorist



Back to the Sample Mean

- Let’s take a step back from the end of last class and talk about what an “experiment” is.



Nature

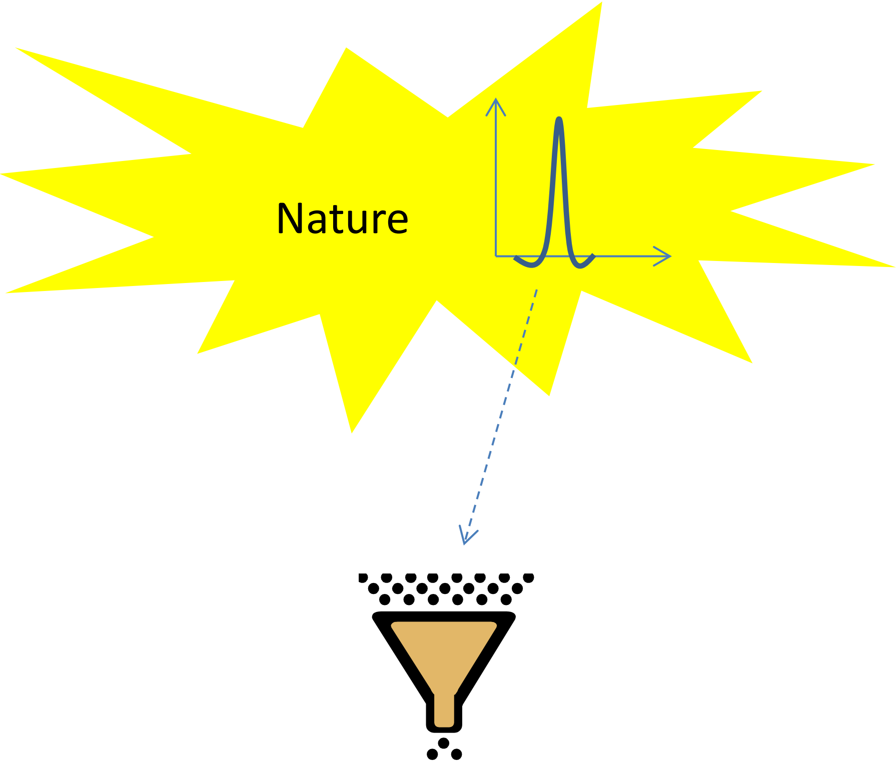
?

the theorist

the experimentalist

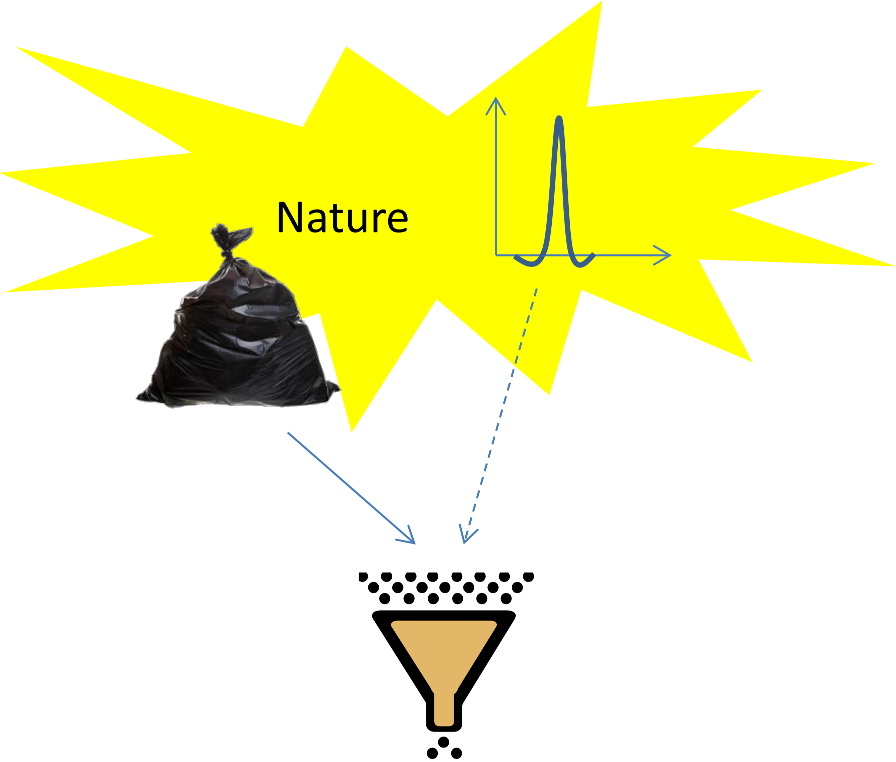
Back to the Sample Mean

- Let’s take a step back from the end of last class and talk about what an “experiment” is.



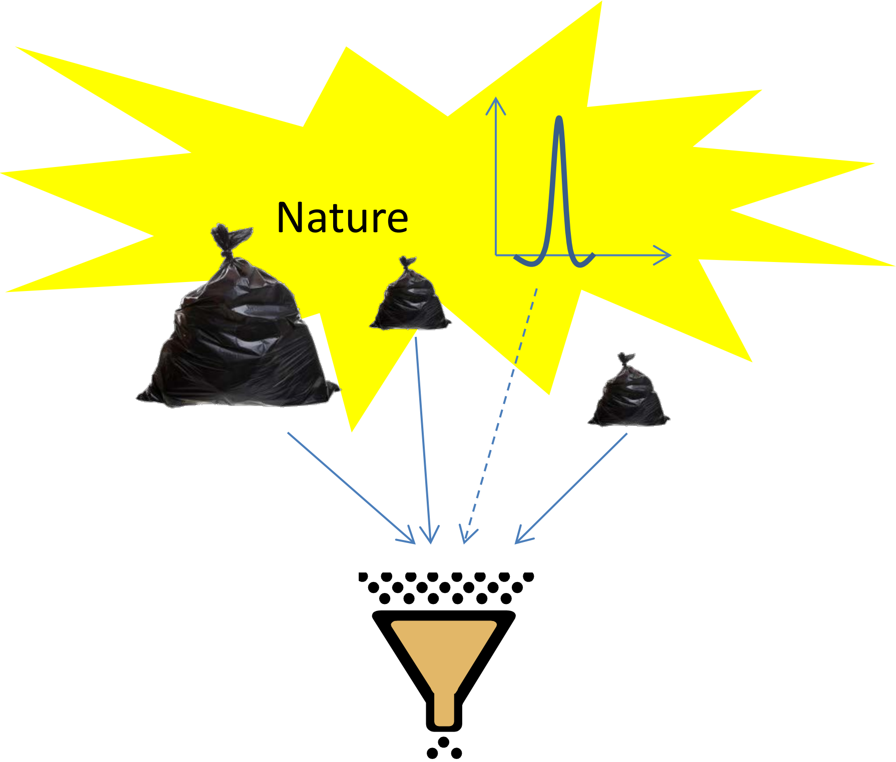
Back to the Sample Mean

- Let’s take a step back from the end of last class and talk about what an “experiment” is.



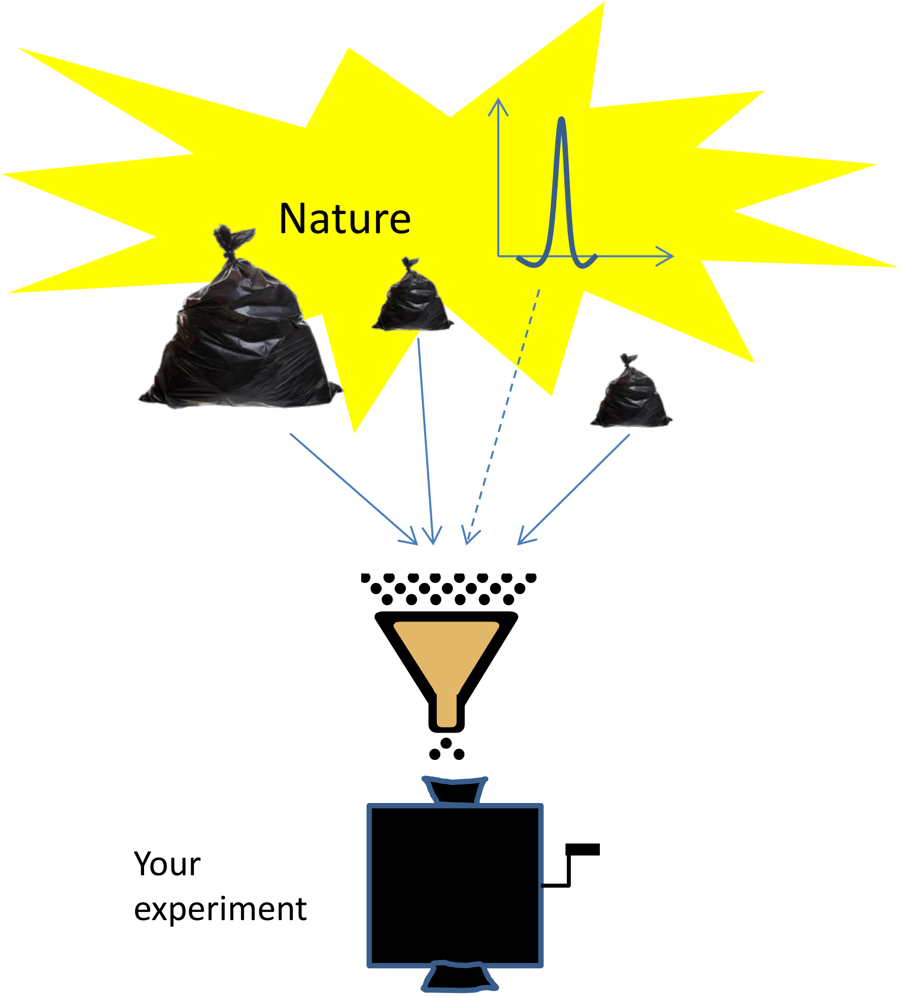
Back to the Sample Mean

- Let’s take a step back from the end of last class and talk about what an “experiment” is.



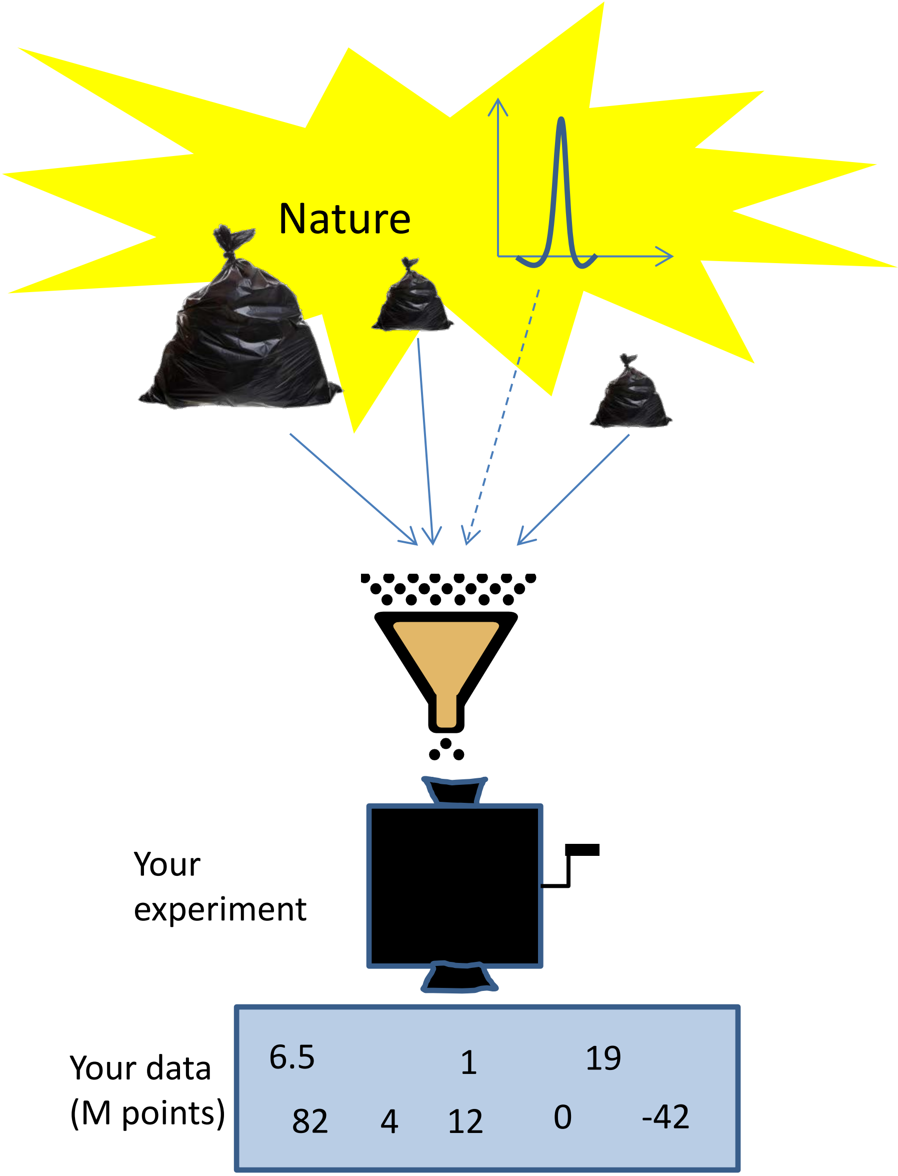
Back to the Sample Mean

- Let’s take a step back from the end of last class and talk about what an “experiment” is.



Back to the Sample Mean

- Let’s take a step back from the end of last class and talk about what an “experiment” is.



Nature

Your experiment

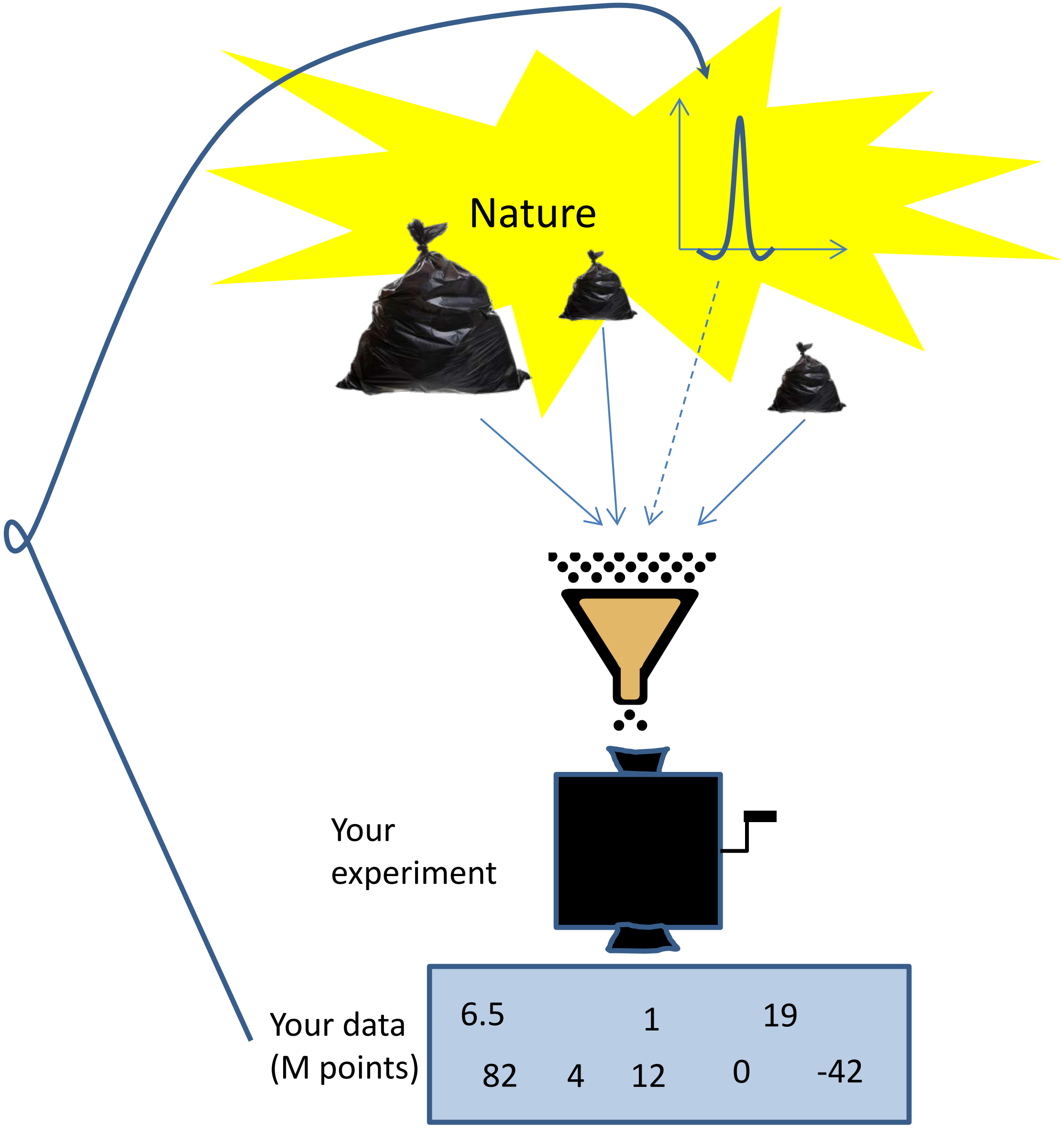
Your data (M points)

6.5 1 19

82 4 12 0 -42

Back to the Sample Mean

- The goal is to try to infer something about some aspect of nature from the data we've collected.



Nature

Your experiment

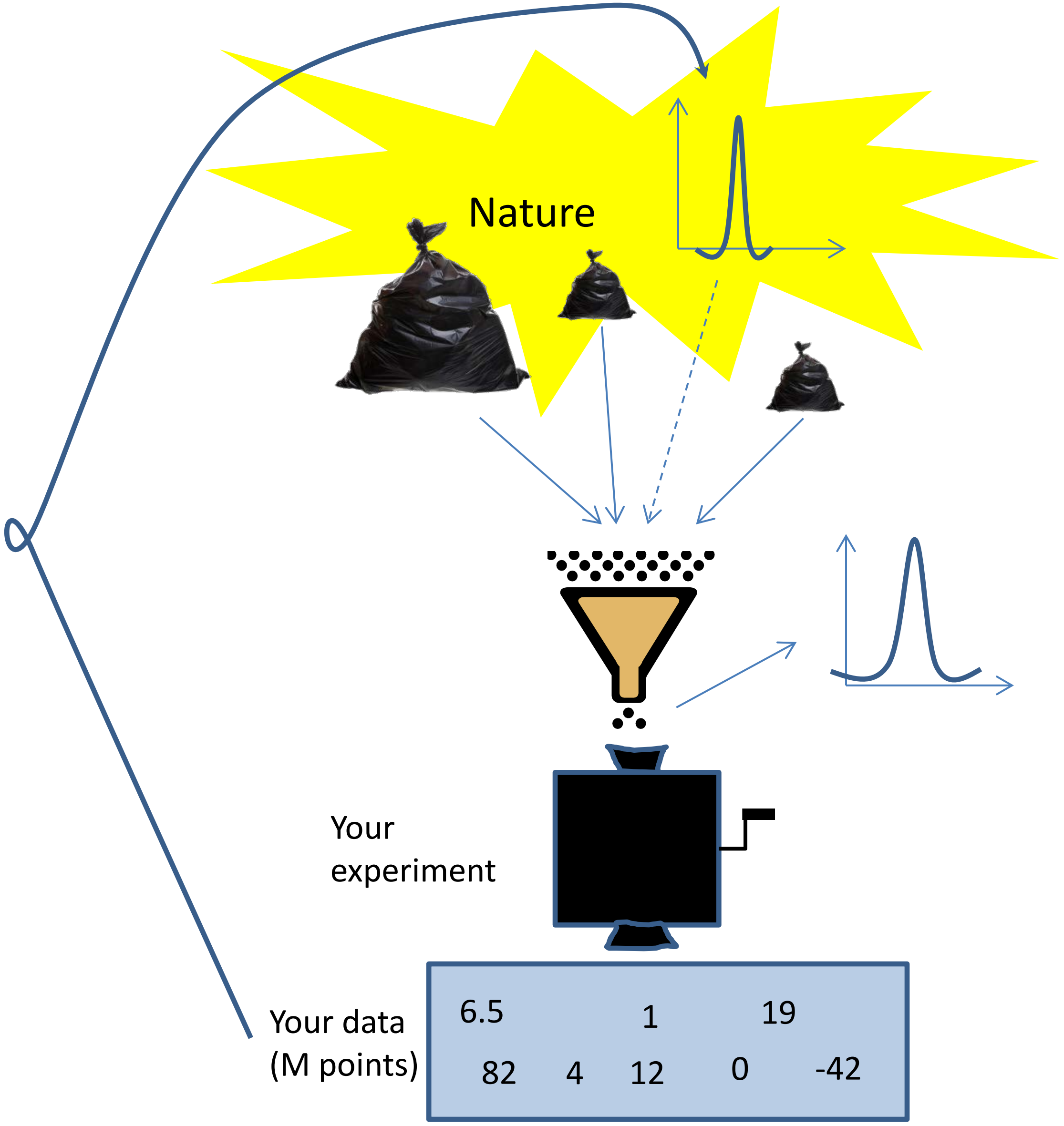
Your data (M points)

6.5 1 19

82 4 12 0 -42

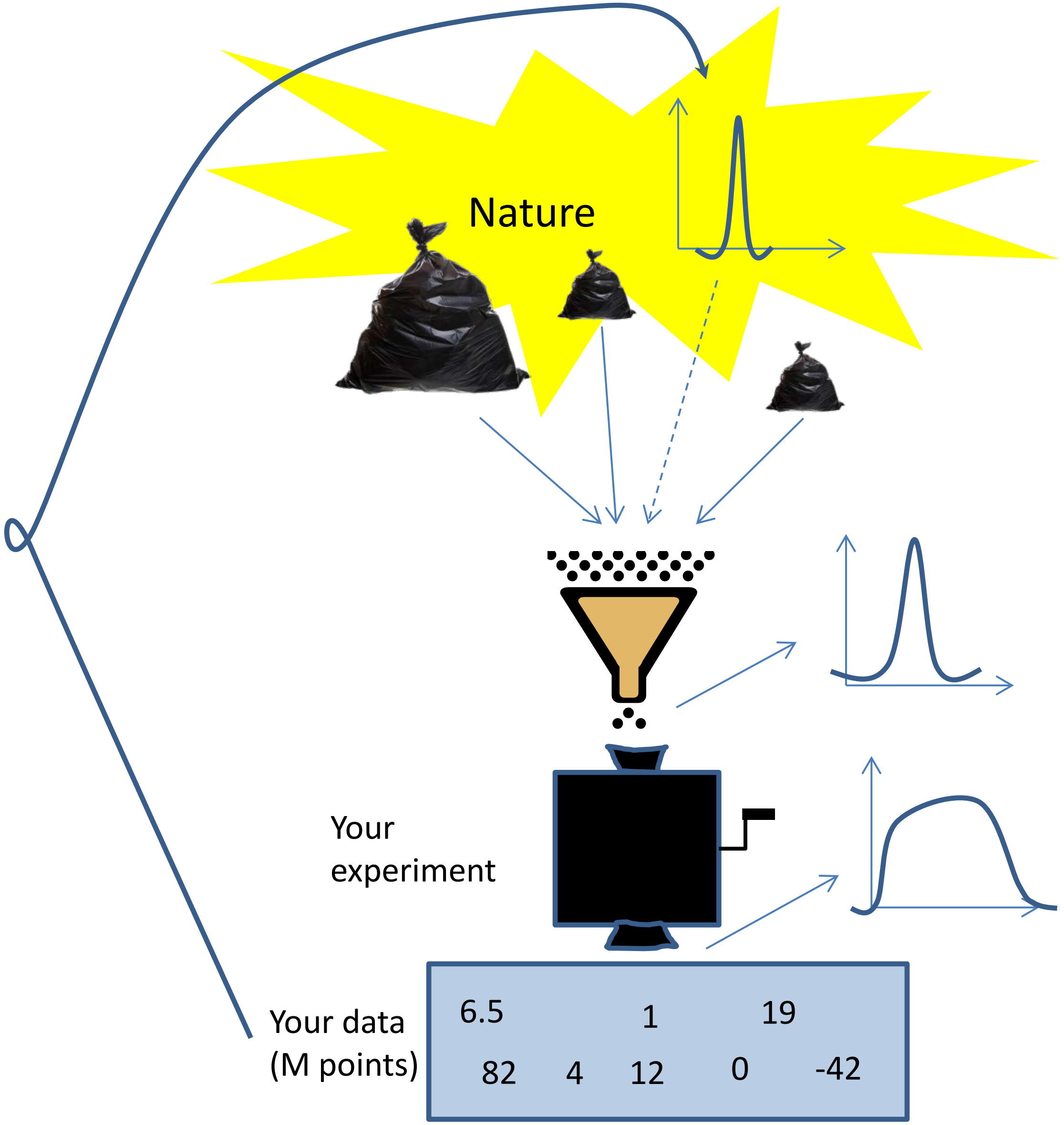
Back to the Sample Mean

- The goal is to try to infer something about some aspect of nature from the data we've collected.



Back to the Sample Mean

- The goal is to try to infer something about some aspect of nature from the data we've collected.



Nature

Your experiment

Your data (M points)

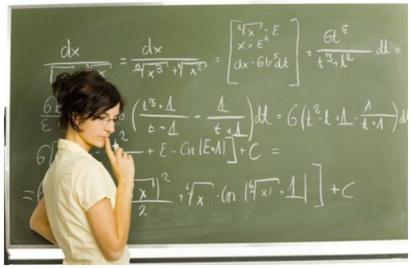
6.5 1 19

82 4 12 0 -42

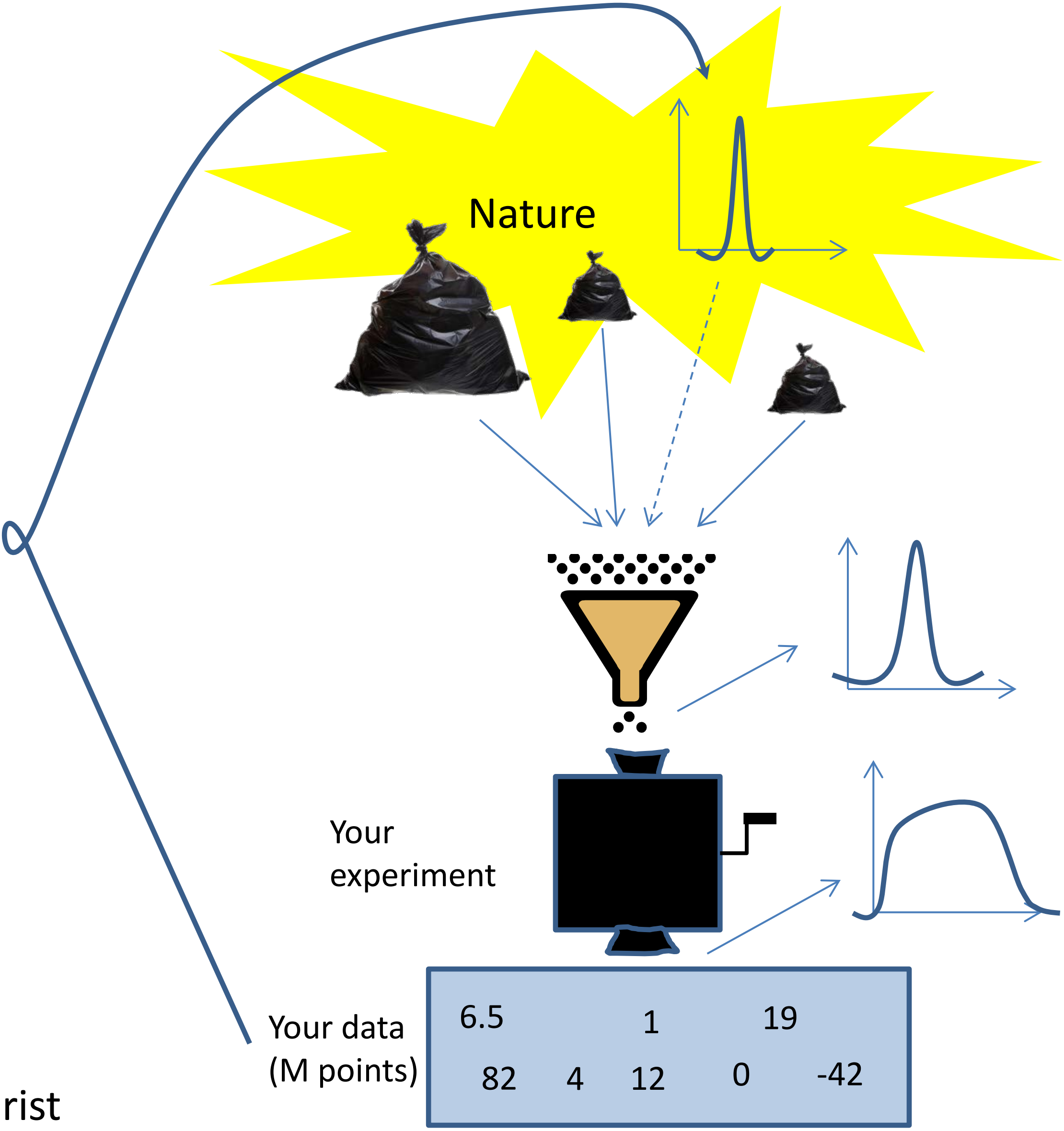
Back to the Sample Mean

- The goal is to try to infer something about some aspect of nature from the data we've collected.

The theorist can help us (or mislead us) but first let's ask what we can do without her for the moment. That is, what does the data say on its own?



the theorist



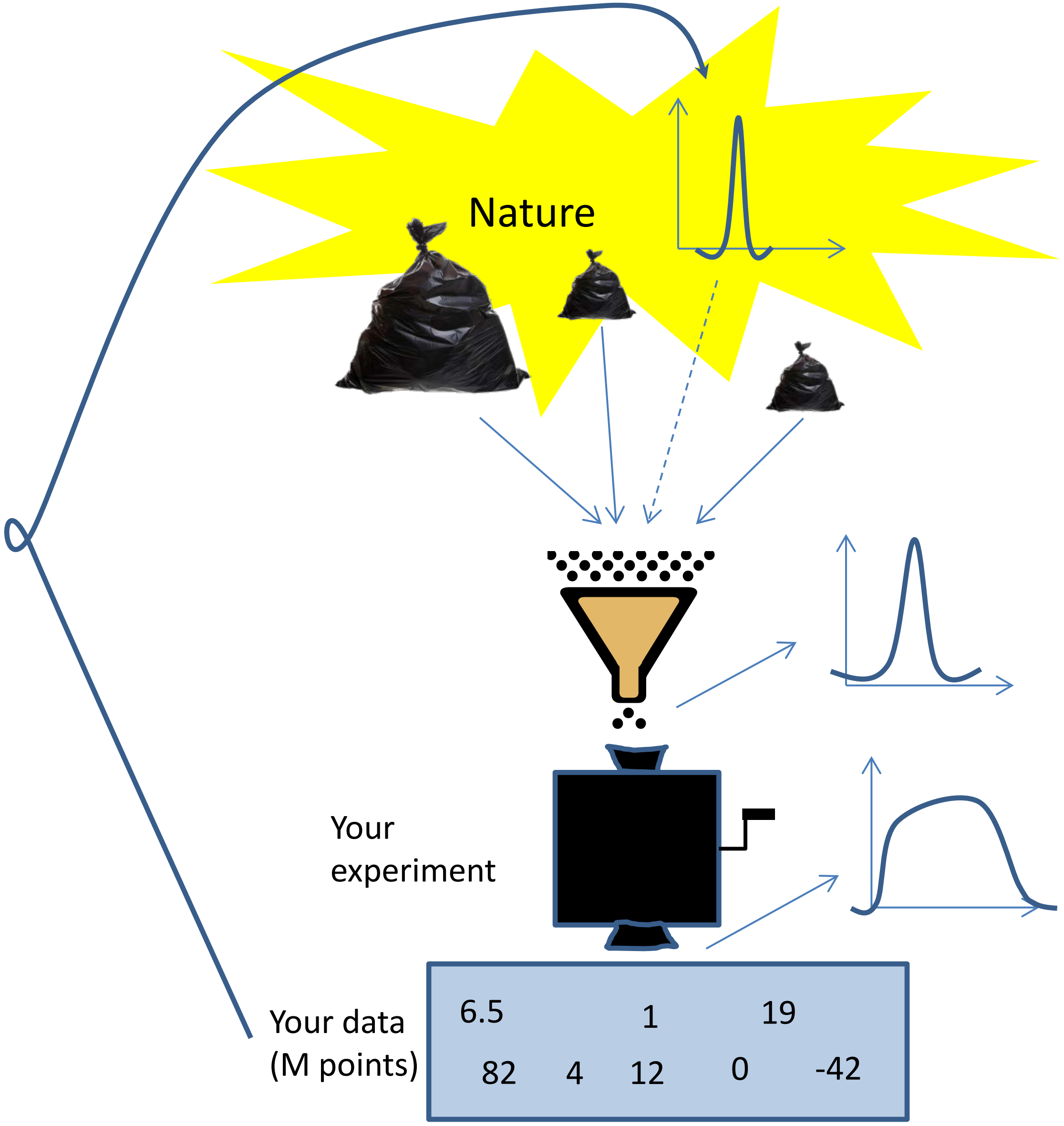
Your data (M points)

6.5 1 19

82 4 12 0 -42

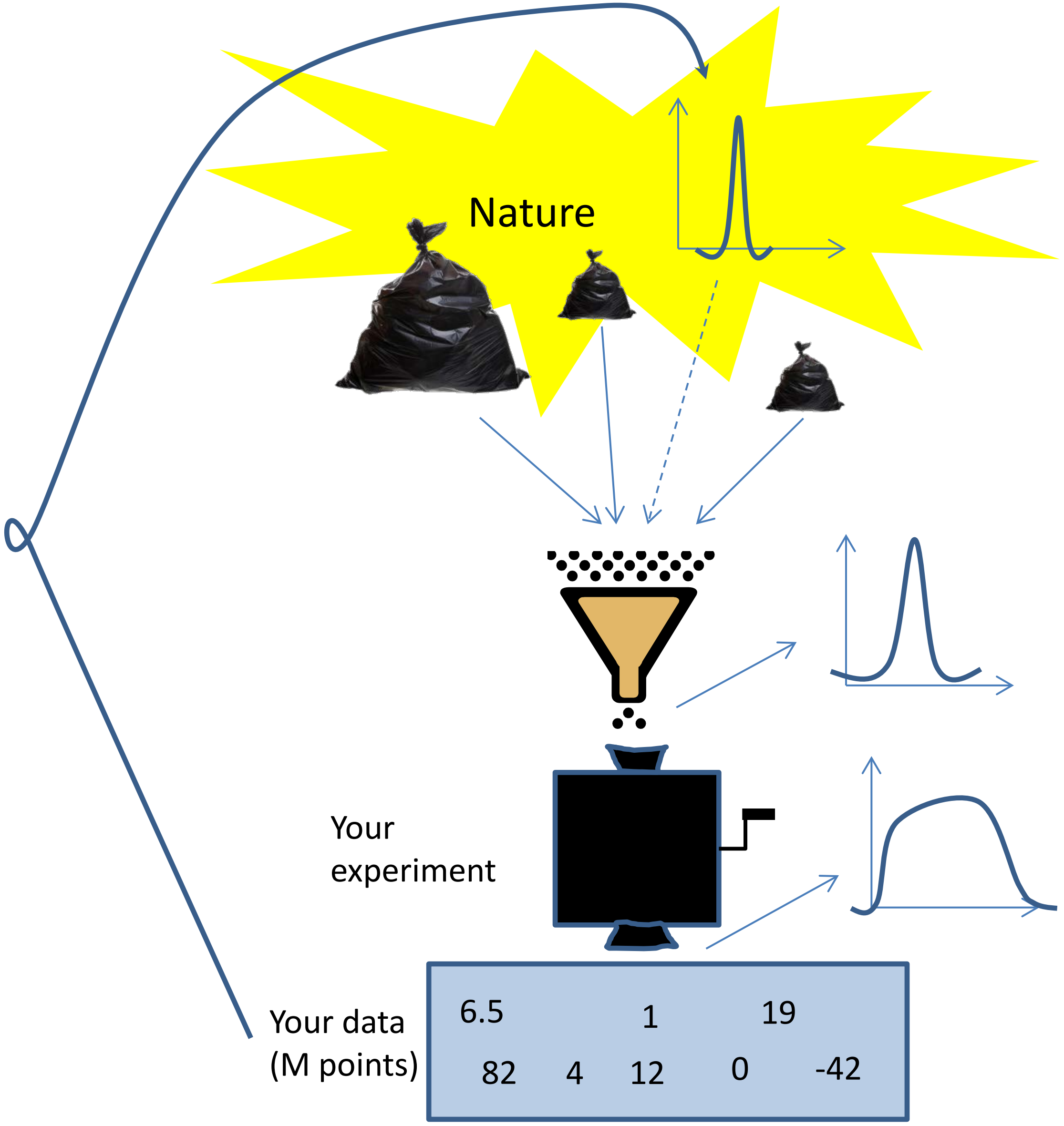
Back to the Sample Mean

- What can the data tell us on its own?
- Two statistics to consider:
  - the mean
  - the standard deviation



Back to the Sample Mean

- What can the data tell us on its own?
- Last time we saw that the "sample mean" is the "most likely estimate" of the "parent mean"



Nature

Your experiment

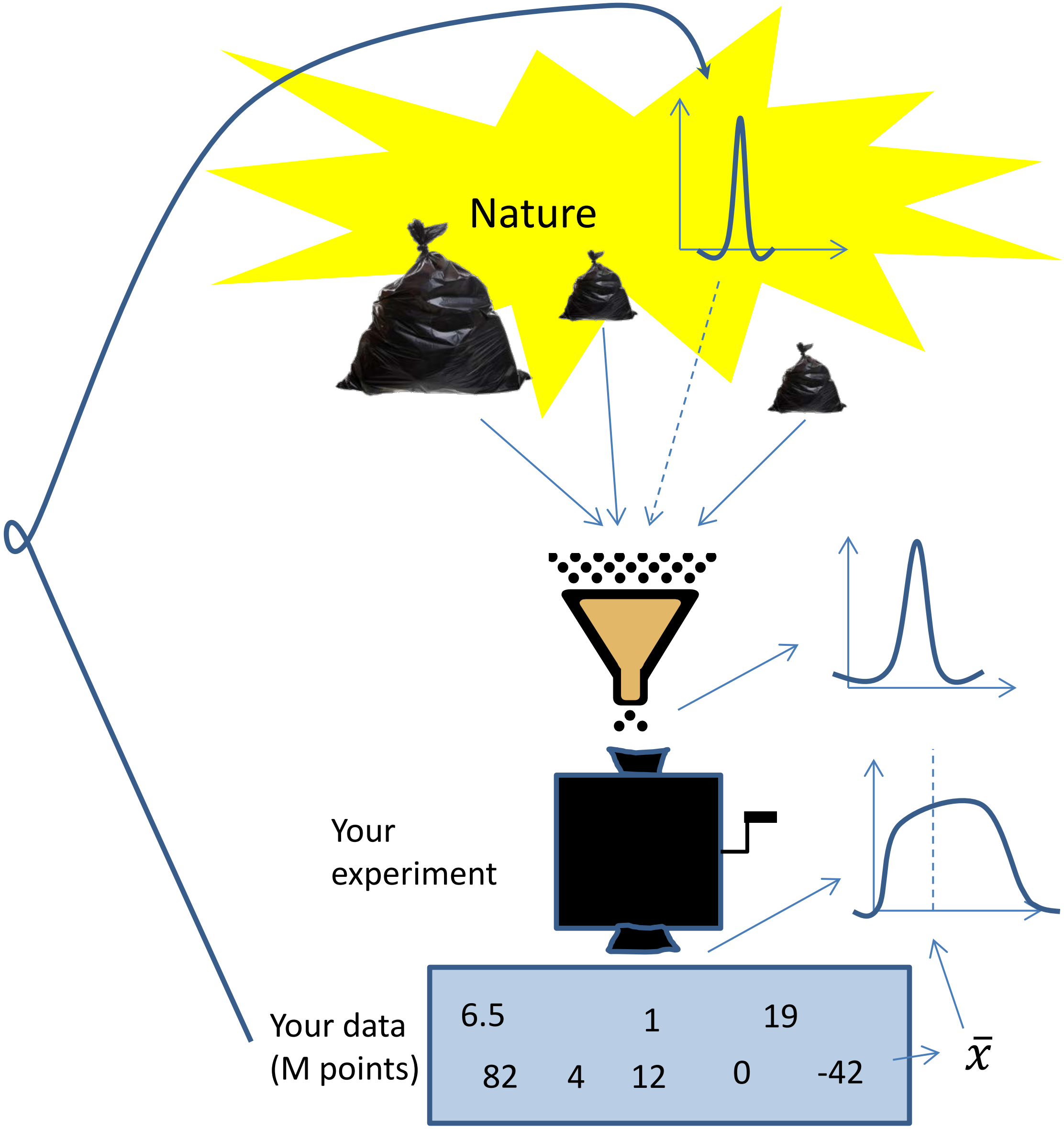
Your data (M points)

6.5 1 19

82 4 12 0 -42

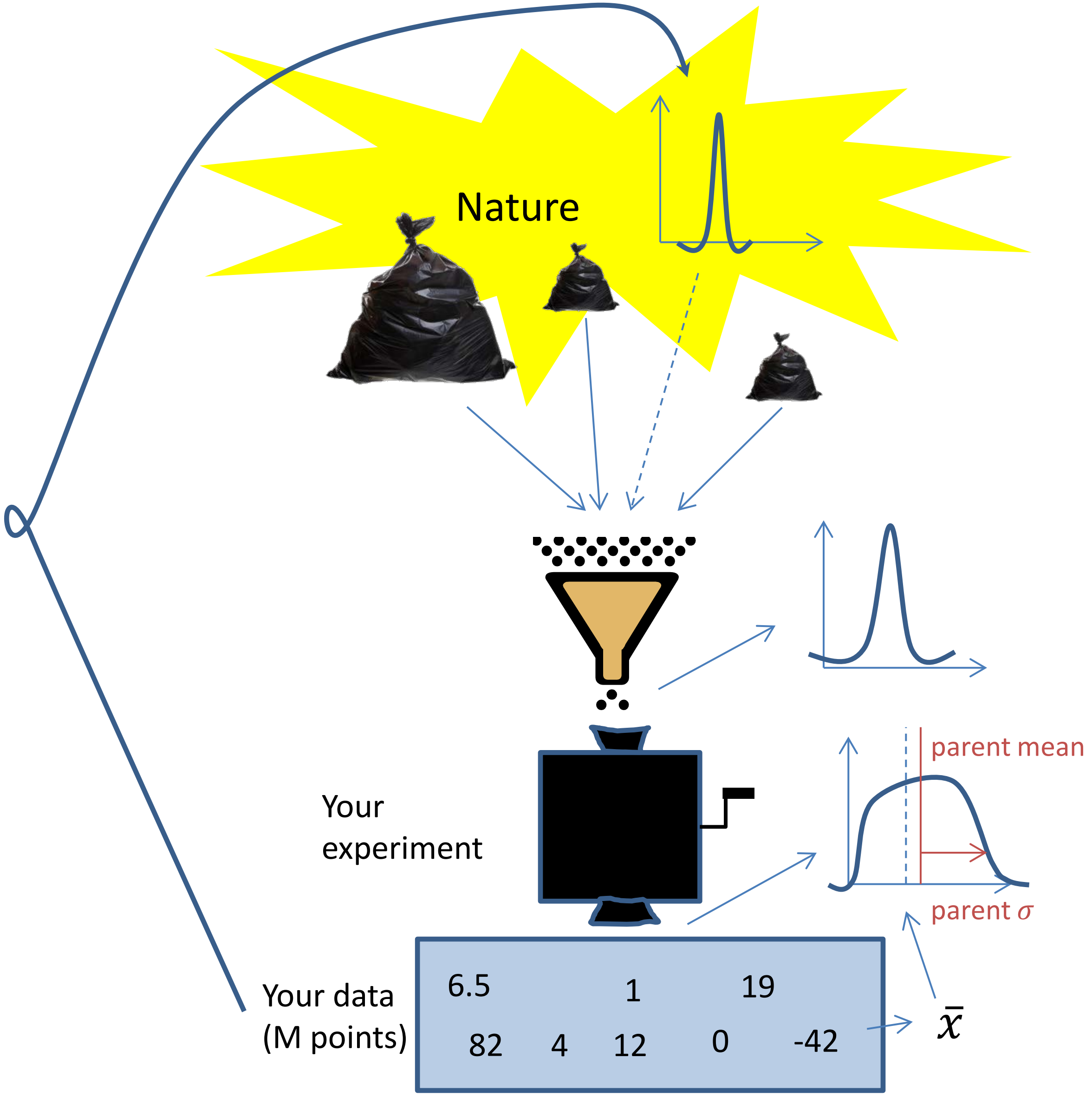
Back to the Sample Mean

- What can the data tell us on its own?
- Last time we saw that the "sample mean" is the "most likely estimate" of the "parent mean"



Back to the Sample Mean

- What can the data tell us on its own?
- Last time we saw that the "sample mean" is the "most likely estimate" of the "parent mean"
- But given that we only have a handful of measurements, how accurate is that estimate of the parent mean?



Nature

Your experiment

Your data (M points)

6.5 1 19

82 4 12 0 -42

parent mean

parent σ

$\bar{x}$

The Central Limit Theorem to the Rescue

- Remember: The CLT says that the sum of many random deviates will itself be a random deviate drawn from a normal distribution.

- So the sample mean (which is a sum of random data points) is drawn from a normal distribution with standard deviation σ_μ

- As you showed in the homework: $\sigma_\mu^2 = \frac{\sigma^2}{M}$ where M is the number of data points in your sample and σ^2 is the variance in the parent distribution.

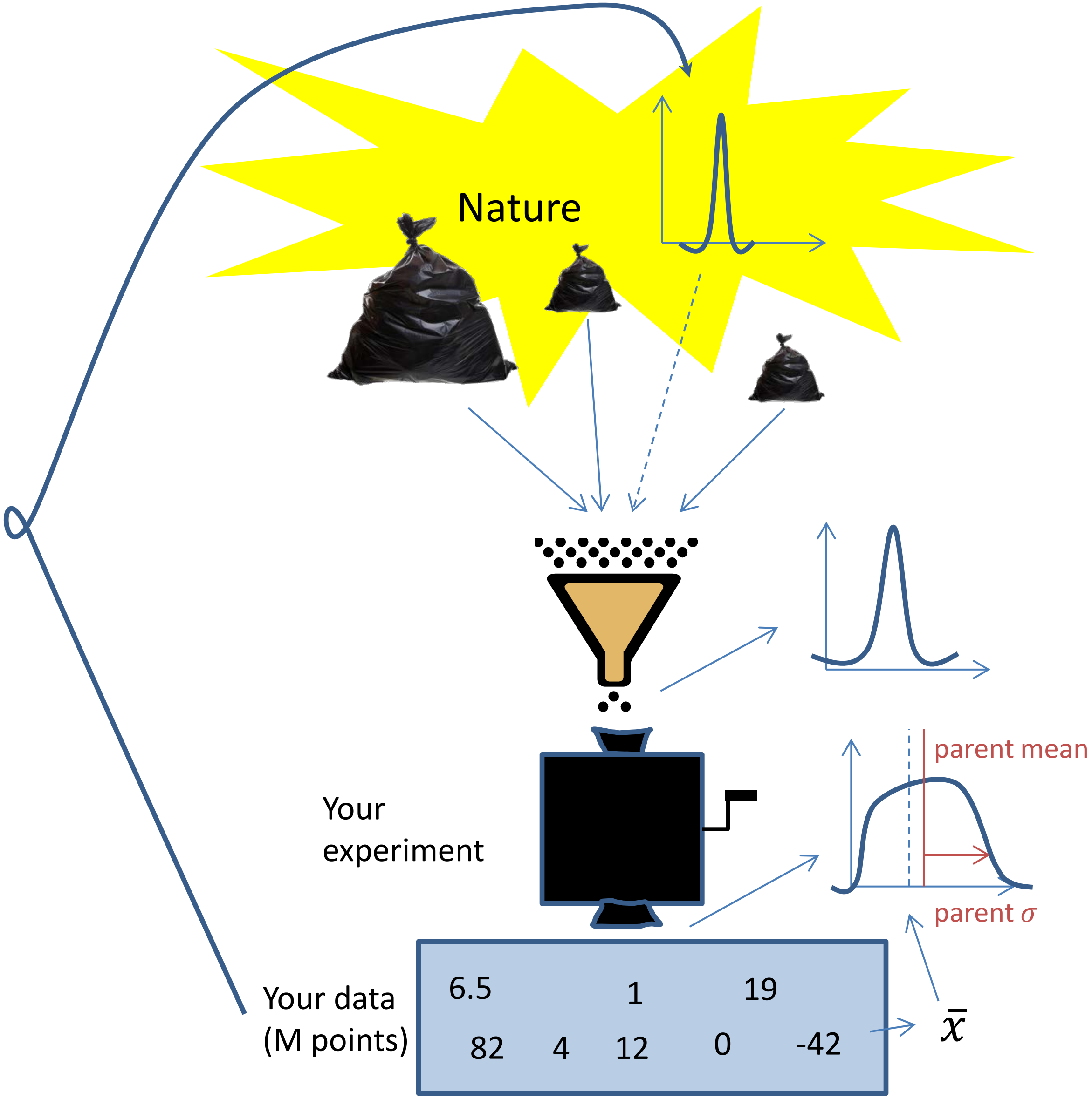
Back to the Sample Mean

Bottom Line:

- $\bar{x}$ - the sample average – is our best guess at the parent mean.

- The error in that guess depends on the number of samples we collect, M, and the standard deviation of the parent distribution, σ, as:

$$\sigma_\mu^2 = \frac{\sigma^2}{M}$$



6.5 1 19

82 4 12 0 -42

Class 12 Exercises

```
#this is my function to draw M samples from the parent distribution
  def parent(mu,sigma,M):
     return np.random.normal(mu,sigma,M)

  #I want to simulate an experiment where nTrials times I draw M samples from the parent distribution.
  #I then calculate the standard deviation, sigma_M, of the nTrials means of my M samples
  M = 20
  nTrials = 1000
  my_means = np.zeros(nTrials)
  parent_mean = 4.
  parent_sigma = 0.2

  #here is the simulation
  for i in range(nTrials):
     data = parent(parent_mean,parent_sigma,M)
     my_means[i] = data.mean()

  #the standard deviation, sigma_M, of my_means is just
  sigma_M = my_means.std()

  #compare this to the expected value
  print 'sigma_M = ', sigma_M
  print 'sigma_theor = ', np.sqrt(parent_sigma**2/M)
```

13 – The Chisq Distribution

Physics 281 – Class 13

Grant Wilson

Let’s Pose Two Questions

1. Given a set of data, how likely is it that our data is a random draw from a normal distribution with mean μ and variance σ^2?

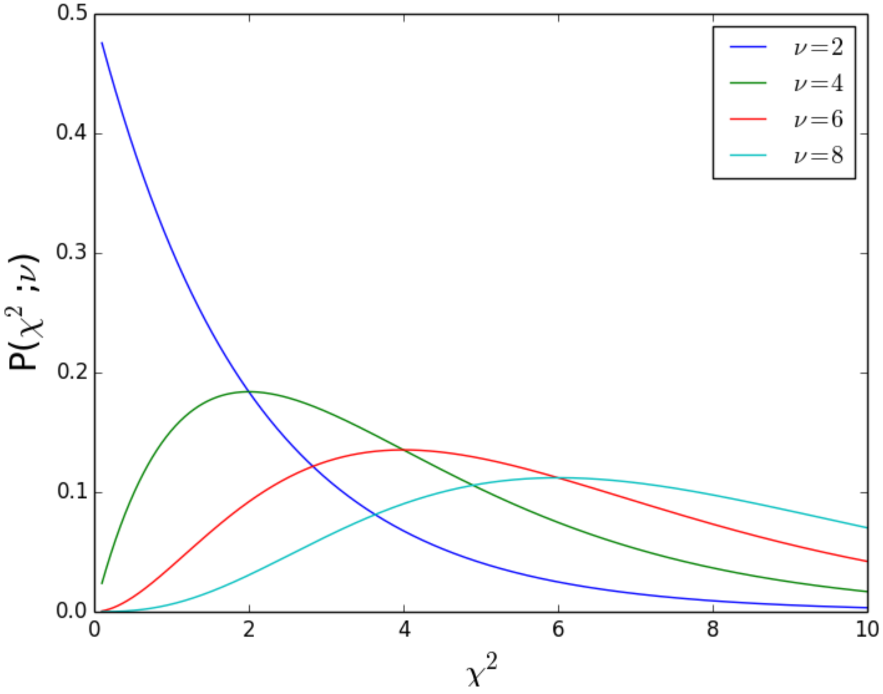
2. How might we estimate the variance, σ^2, of the parent distribution?

Comparing Our Data to a Model

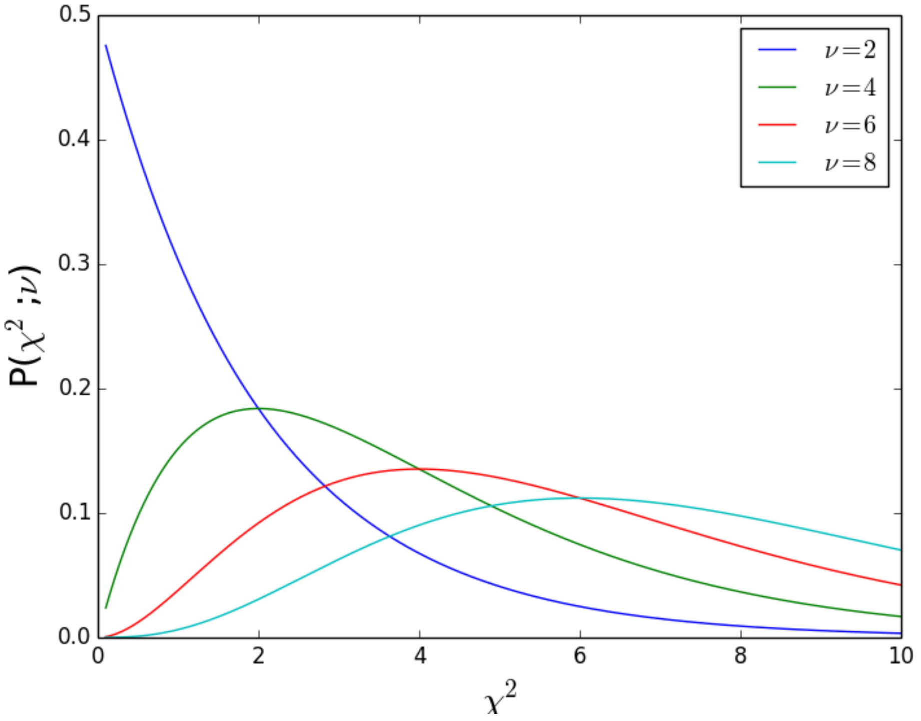
- Define the chi-square (or chisq) statistic.

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

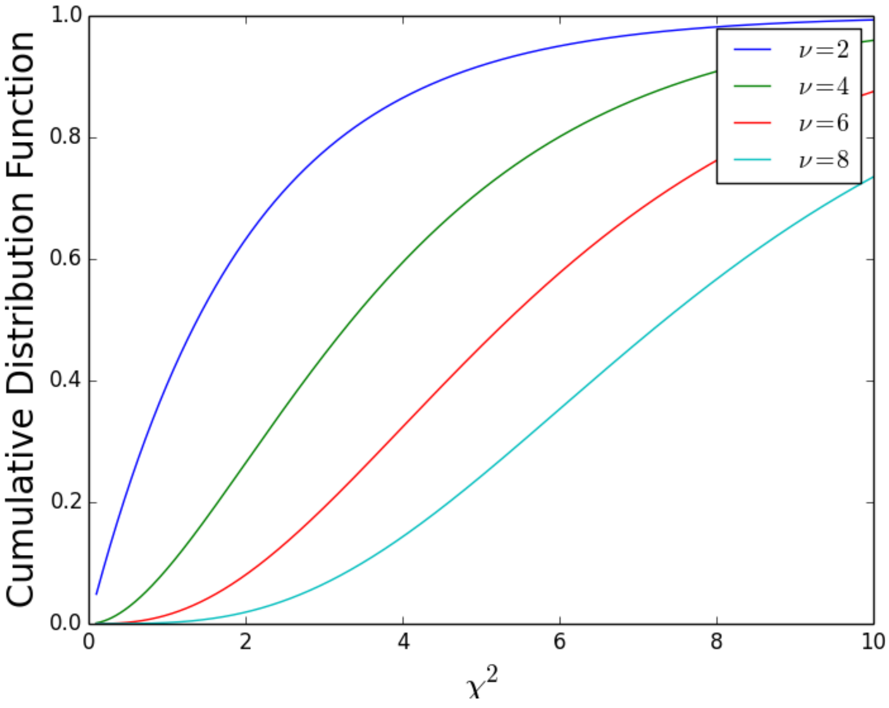
- If the x_i's are ν random and independent samples from a normal distribution, then χ^2 is drawn from a chisq-distribution with ν degrees of freedom.



Using the Chisq Distribution



Probability that a Chi-square taken from a random sample with ν- degrees of freedom has a particular value.



Probability that a Chi-square taken from a random sample with ν- degrees of freedom is *less than* a particular value.

The Cumulative Distribution Function

- Every probability distribution has an associated cumulative distribution function.

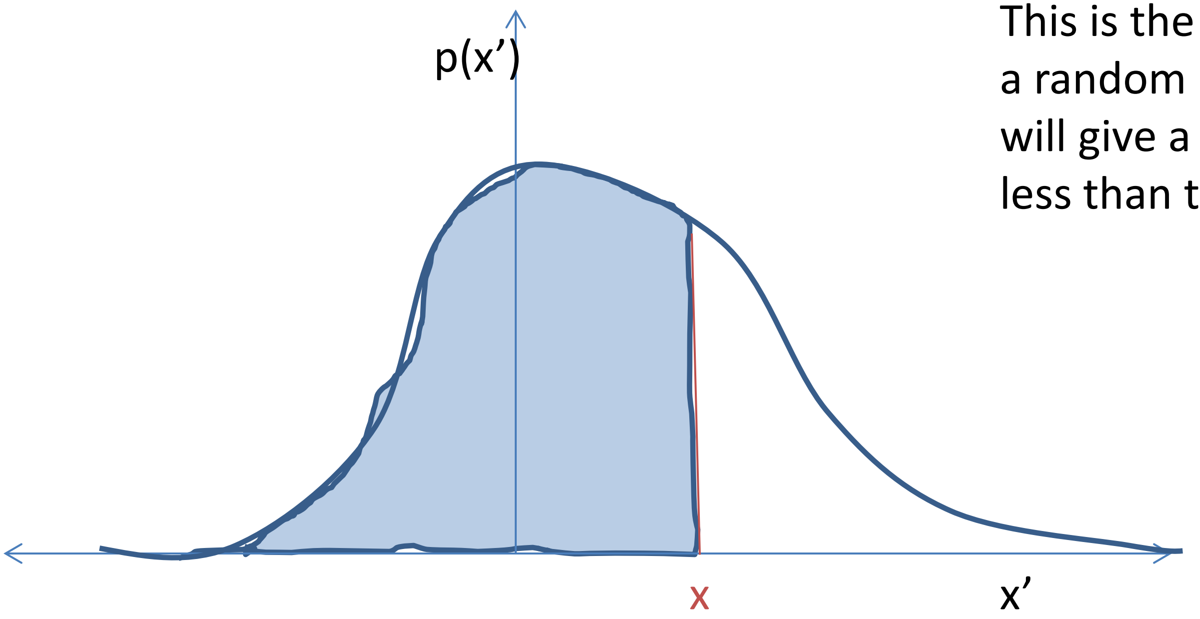
$$CDF(x) = \int_{-\infty}^{x} p(x')dx$$

What is this?

The Cumulative Distribution Function

- Every probability distribution has an associated cumulative distribution function.

$$CDF(x) = \int_{-\infty}^{x} p(x')dx$$



This is the probability that a random draw from p(x') will give a random variate less than the value x.

Why might that be useful?

Exercise

- I perform an experiment with 40 degrees of freedom. I calculate the chisq statistic for my data and get a value of 82. What is the probability of getting a chisq value lower than this? What is the probability of getting a chisq value higher than this?

- Hint, use:
  - from scipy.stats import chi2
  - chi2 has methods chi2.pdf() and chi2.cdf()

This brings us to a fundamental point about science.

Question: What value of χ^2 is required to prove that the data was drawn from a particular probability distribution?

This brings us to a fundamental point about science.

Question: What value of χ^2 is required to prove that the data was drawn from a particular probability distribution?

Another way to ask this:

What total duration of time for the reporter to drive from the high school to the Best Buy parking lot would prove that Adnan murdered Hae?

This brings us to a fundamental point about science.

Question: What value of χ^2 is required to prove that the data was drawn from a particular probability distribution?

Another way to ask this:

What total duration of time for the reporter to drive from the high school to the Best Buy parking lot would prove that Adnan murdered Hae?

Both of these questions are nonsensical. This is not how science works.

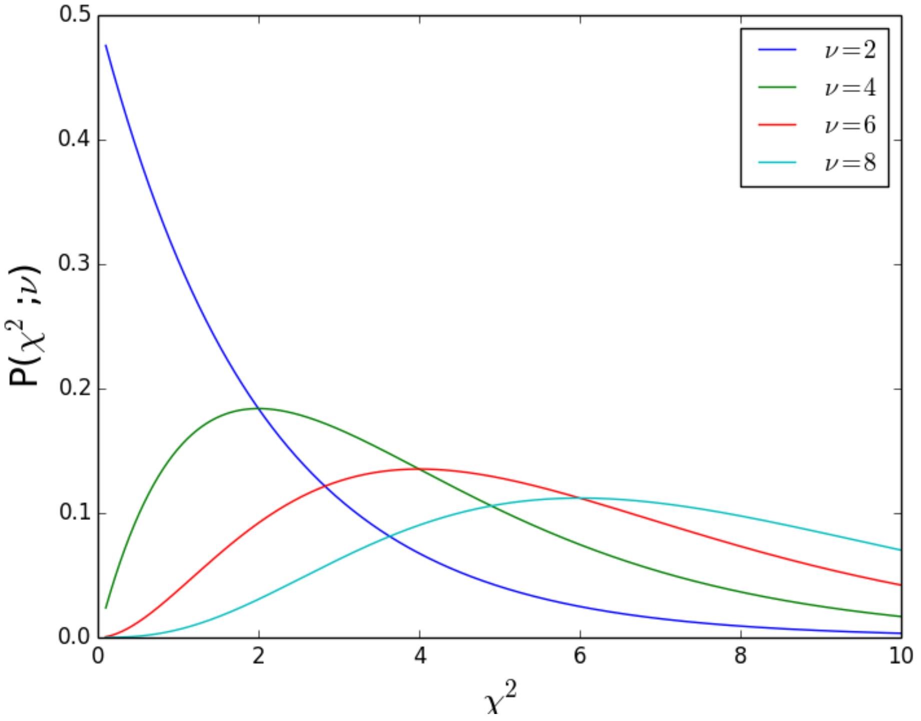
This brings us to a fundamental point about science.

The only kinds of questions we are allowed to (supposed to) ask in science are of the form:

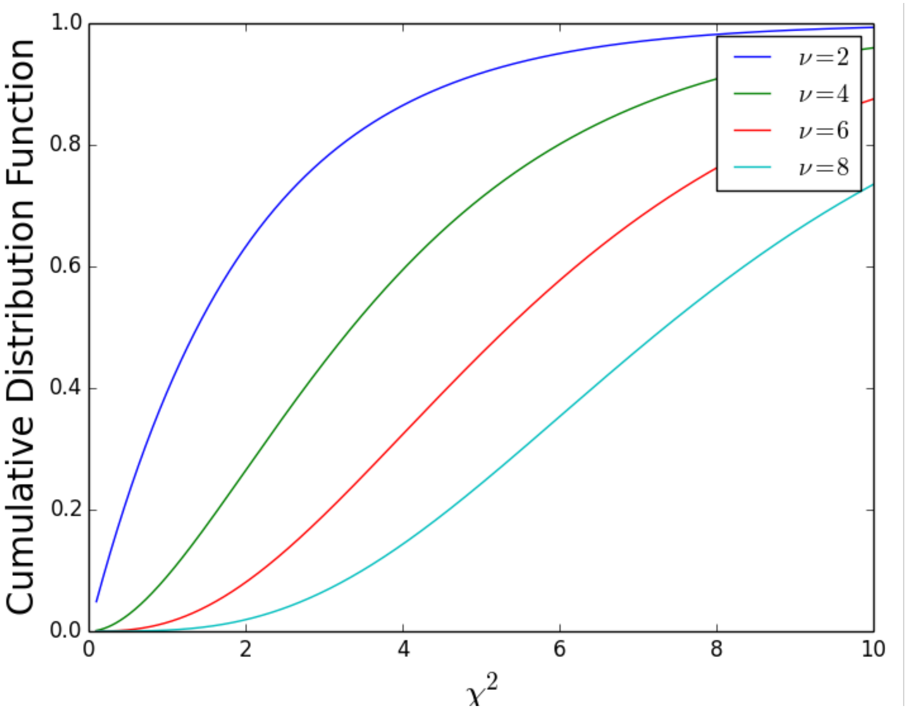
What value of χ^2 (or insert your favorite statistic here) makes it extremely unlikely that our data was drawn from a particular probability distribution?

The cumulative distribution function helps us answer this.

Using the Chisq Distribution

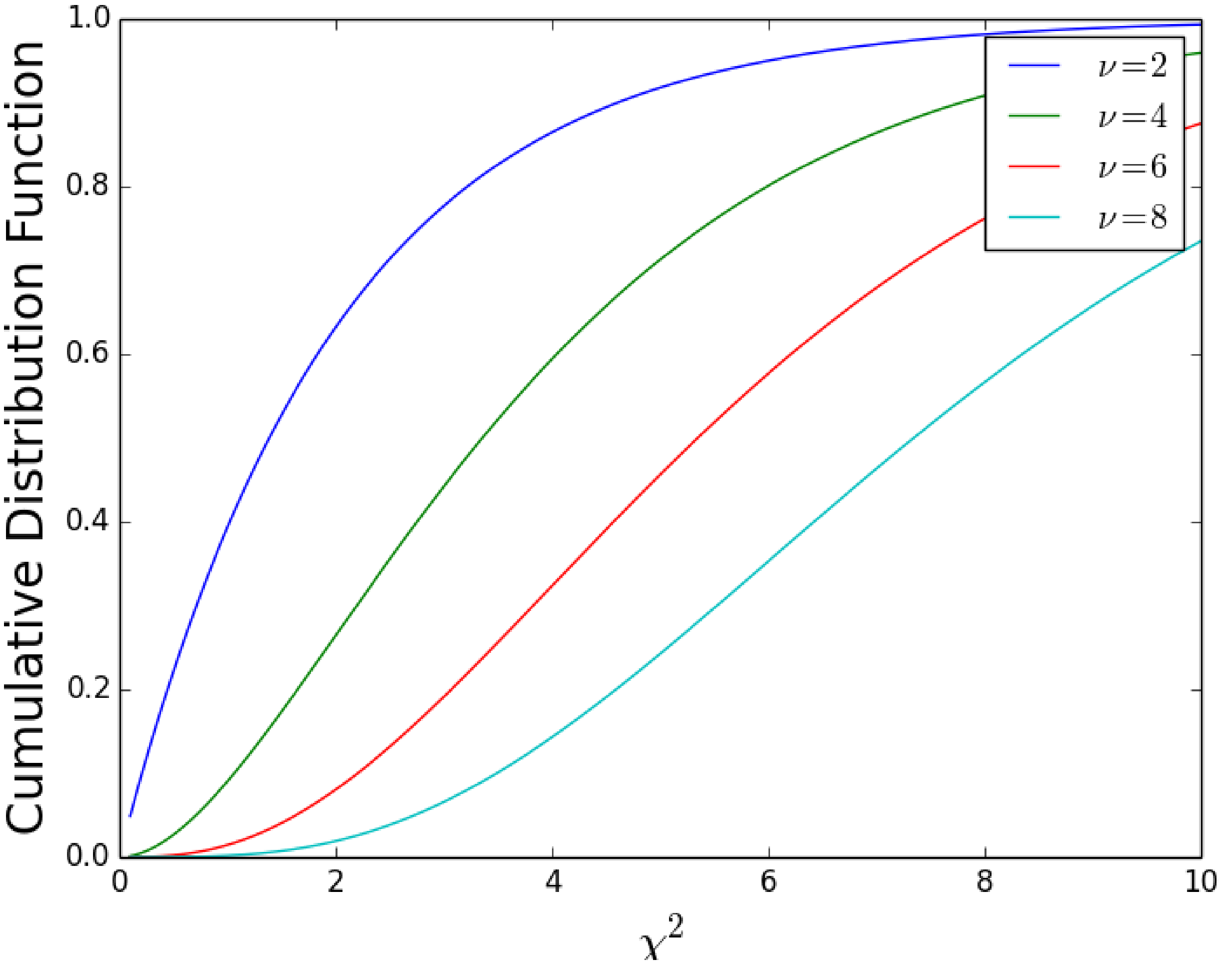


Probability that a Chi-square taken from a random sample with ν- degrees of freedom has a particular value.



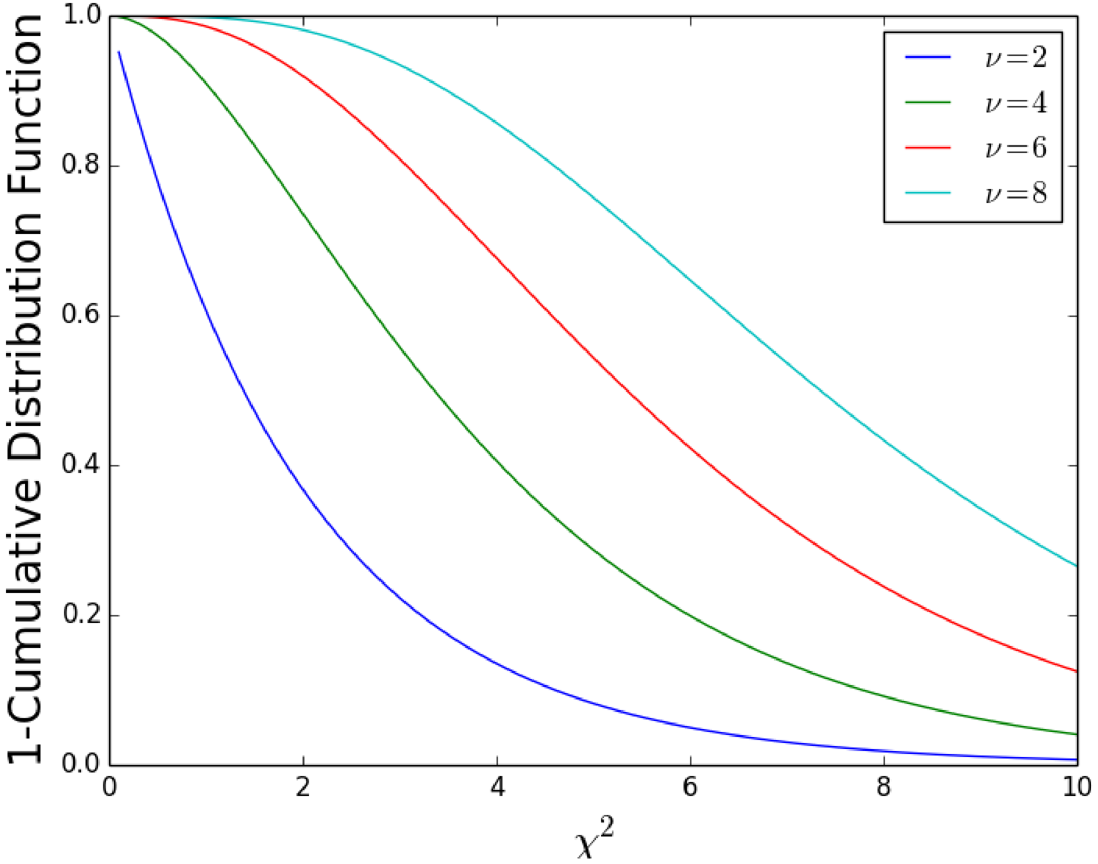
Probability that a Chi-square taken from a random sample with ν- degrees of freedom is *less than* a particular value.

Use the cdf() to get how unlikely a low-value is.



cdf(χ_i^2, ν) is the probability of getting a value of χ^2 lower than what you got in your experiment, χ_i^2. If that probability is very low then it is unlikely that your value of chisq is actually drawn from a chisq-distribution with ν-degrees of freedom as you suspected.

Take 1-cdf() to get how unlikely a high-value is.



1-cdf(χ_i^2, ν) is the probability of getting a value of χ^2 higher than what you got in your experiment, χ_i^2. If that probability is very low then it is unlikely that your value of chisq is actually drawn from a chisq-distribution with ν-degrees of freedom as you suspected.

Here’s the recipe

1. Use your data to calculate your preferred statistic. The requirements on the statistic are:
    1. You know what the underlying probability distribution of the statistic is. (see next slide)
    2. You know what limitations the statistic has (is it biased, is it an MLE, is it consistent, is it sufficient, etc.) In this class we’ll work with statistics for which none of these are a concern.
2. Calculate the CDF of the statistic’s probability distribution at the value of the statistic that you measured.
3. Take 1 – the value you calculated in step 2.
4. Decide how reasonable/unreasonable the value from steps 2 and 3 are.

Two questions raised by the recipe

1. How can we know the underlying probability distribution of the statistic?

2. What’s reasonable in step 4?

Two questions raised by the recipe

1. How can we know the underlying probability distribution of the statistic?

    1. some statistics have well defined probability distributions (like χ^2)

    2. sometimes you have to simulate the probability distribution (like we did in problem #2 of last night’s homework).

Two questions raised by the recipe

1. How can we know the underlying probability distribution of the statistic?

2. What's reasonable in step 4?

There is no "correct" answer to this as it is a matter of personal preference. Often scientists use the corresponding Gaussian confidence intervals (reproduced in the table to the right) as a guide. This leads to talk like "How many sigma is the result?"

x_n	$P(\mu - x_n < x < \mu + x_n)$
1σ	68.3%
2σ	95.4%
3σ	99.7%
4σ	99.993%
5σ	99.99994%

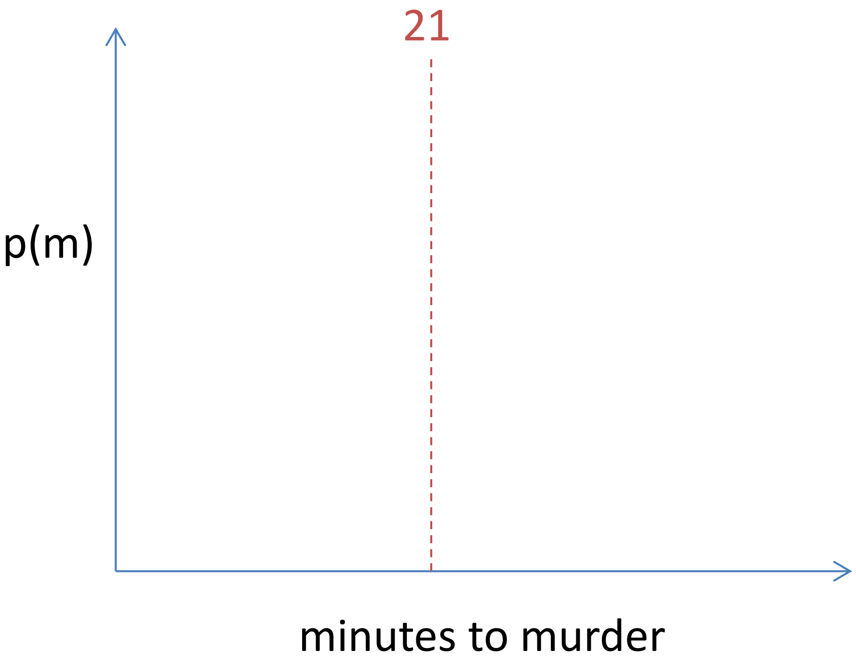
This Idea Works for *Any* Probability Distribution

Can we prove Adnan innocent or guilty based on the timeline of the murder?

This Idea Works for *Any* Probability Distribution

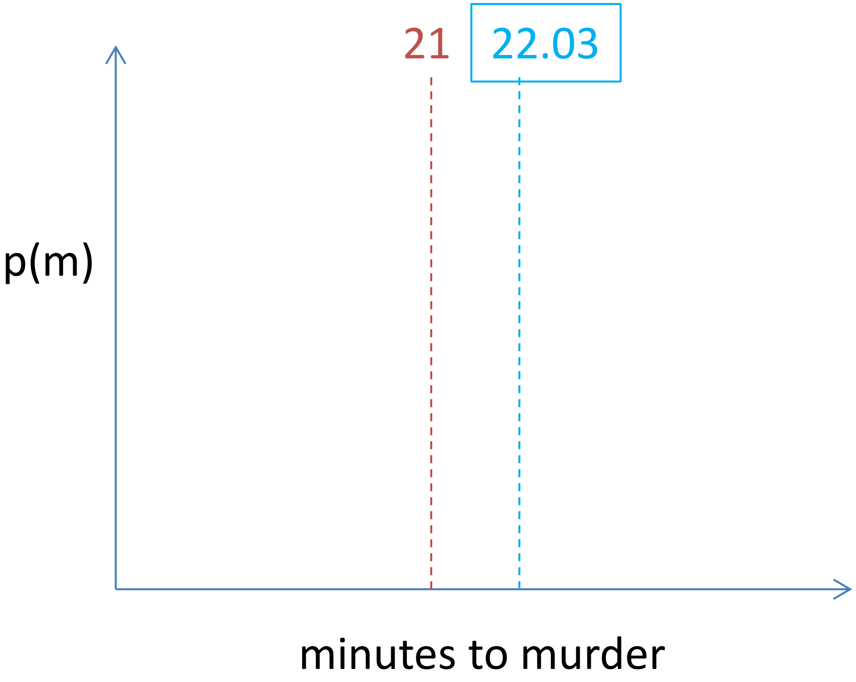
Can we prove Adnan innocent or guilty based on the timeline of the murder?

21 minutes – the time between the end of school and the “I did it” phone call (as contended by the police)



This Idea Works for *Any* Probability Distribution

Can we prove Adnan innocent or guilty based on the timeline of the murder?

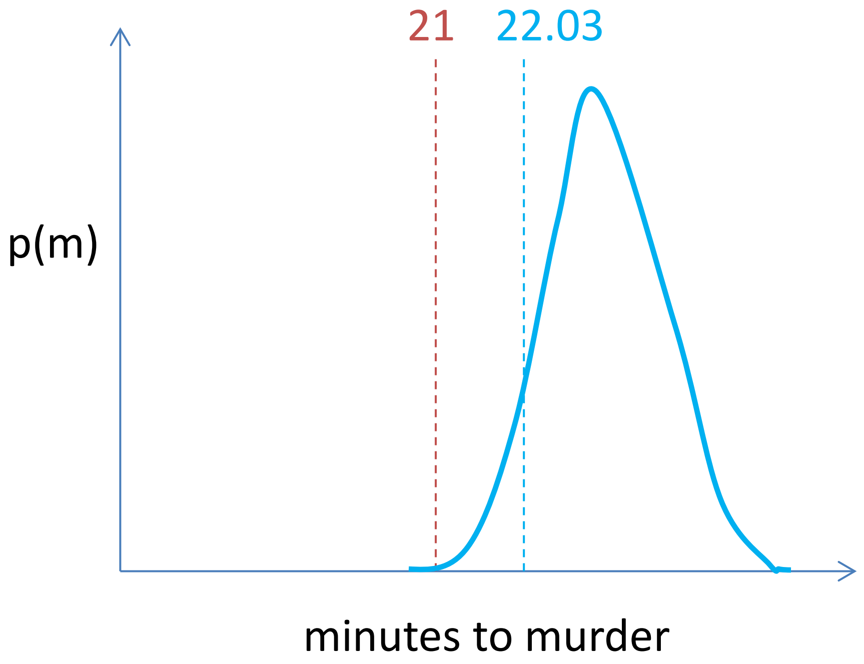


21 minutes – the time between the end of school and the “I did it” phone call (as contended by the police)

22.03 minutes – the time it takes the reporters to recreate the chain of events

This Idea Works for *Any* Probability Distribution

Can we prove Adnan innocent or guilty based on the timeline of the murder?



21 minutes – the time between the end of school and the "I did it" phone call (as contended by the police)

22.03 minutes – the time it takes the reporters to recreate the chain of events

Let's imagine that the reporters redo the drive 100 times and find a probability distribution that is Gaussian with mean, $\mu = 23$, and standard deviation, $\sigma = 0.75$. What's the probability that the value of 21 minutes is pulled from that same distribution?

Exercise

- The police claim that the murder happened within 21 minutes of the end of school.

- Reporters, by simulating the chain of events 100 times, determine that the probability distribution of the duration of the chain of events is Gaussian with mean, $\mu = 23$, and standard deviation, $\sigma = 0.75$.

- With what confidence can the reporters claim that Adnon did not commit the murder? Would a scientist be content with setting him free based on this result?

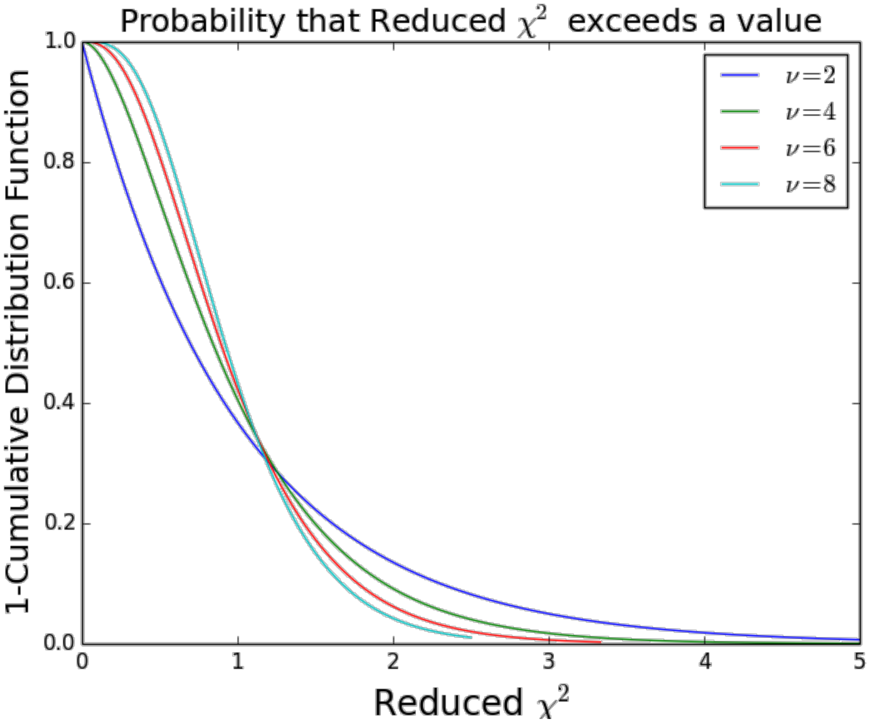
The Reduced-Chisq

- The mean of a chisq-distribution with ν-degrees of freedom is ν

- The Reduced Chisq statistic is simply the chisq normalized by ν

$$Reduced\ \chi^2 = \frac{1}{\nu}\sum_{i=1}^{N}\frac{(x_i-\mu)^2}{\sigma^2}$$

- The mean of the reduced chisq is 1



Estimating Parent Distribution's Variance from the Data

- Up until now, we've seen no estimate for the variance of the parent probability distribution to our data.

- The definition of the variance is an infinite sum and requires knowing the true mean of the parent distribution.

$$\sigma^2 = \lim_{N\to\infty} \sum_{i=1}^{N} (x_i - \mu)^2$$

- We have a finite number of data points and only an estimate of the mean – the sample mean, μ'. So let's do the obvious thing first:

$$\sigma'^2_{guess} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu')^2$$

This is an underestimate of the mean since we've used some information to calculate μ'. In other words, we've used one degree of freedom in our data. So instead:

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

This is the estimate of the variance of the parent distribution from our sample.