Data Science Methods for Clean Energy Research

Week 3, Lecture 1: Distributions

January 4, 2017

Please start a new Jupyter notebook and load the same software stack we have been using.

W3L1 files:
http://faculty.washington.edu/jpfaendt/DIRECTfiles/W3L1Files.zip

UNIVERSITY of WASHINGTON

There is no sadder sight in the world, than to see a beautiful theory killed by brutal fact.

– Thomas Huxley, biologist (1825–1895)

Outline

- **Updated reading and refresh from last time**
- **Warm up**
- **Representing the likelihood of our data**
- **Distributions of** discrete **variables**
  - **The binomial distribution**
  - **The Poisson distribution**
- **Distributions of** continuous **variables**
  - **The normal distribution**
  - **The gamma distribution**
- **Sampling from distributions and the central limit theorem**
- **Getting ready for** hypothesis testing **– why do we care about distributions?**

Using some mathematics defined @ Wolfram alpha today…

Reading to support your learning

- Nature Methods has an excellent series of introductory statistics lectures
- Downloaded several good ones to support our work
- Available from our class GitHub repo or directly here: https://github.com/UWDIRECT/UWDIRECT.github.io/tree/master/DSMCER_content/Reading/Stats

```
/Users/jpfaendt/Dropbox/Teaching/DIRECT/UWDIRECT.github.io/DSMCER_content/Reading/Stats
D-69-91-134-47:Stats jpfaendt$ ls
2013-1-Uncertainty.pdf              2013-4-PowerSampleSize.pdf          2015-Regression.pdf
2013-2-ErrorBars.pdf                2014-ComparingSamples - I.pdf       2016-ModelSelectionOverFitting.pdf
2013-3-P-Values.pdf                 2014-ComparingSamples.pdf           2017-PValues.pdf
D-69-91-134-47:Stats jpfaendt$
```

Key concepts from last time

- Population vs. sample
- Continuous vs. discrete variables

Warm up

- **Inside W3L1Files/ please open "gettingwarm.ipynb"**
- **10 min to work through the notebook and follow instructions**
  - *If needed I will give more time*
- **5 min discussion on observations / findings**

Refresher: histogram (1/3)

- **A histogram is created by doing the following:**
  - **Determine the** minimum **and** maximum **value of your data (the** range**)**
  - **Divide the range into** bins **(usually of equal width, but not strictly required!)**
  - **Loop over the data and count how many entries there are in each bin**
  - **Make a bar plot to visually represent this count:** this is the histogram

```
minirand=np.random.uniform(low=0,high=10,size=10) #  0 <= rand < 10
print(minirand)
[hist,bins]=np.histogram(minirand,range=(0,10),bins=5)
print bins, hist
```

```
[ 7.32338063  5.29793759  0.15800026  1.32764593  3.45961847  3.28937848
  4.07095232  2.47070121  1.23698677  1.56834074]
[  0.   2.   4.   6.   8.  10.] [4 3 2 1 0]
```

Ref

> Th
th
> Le

now

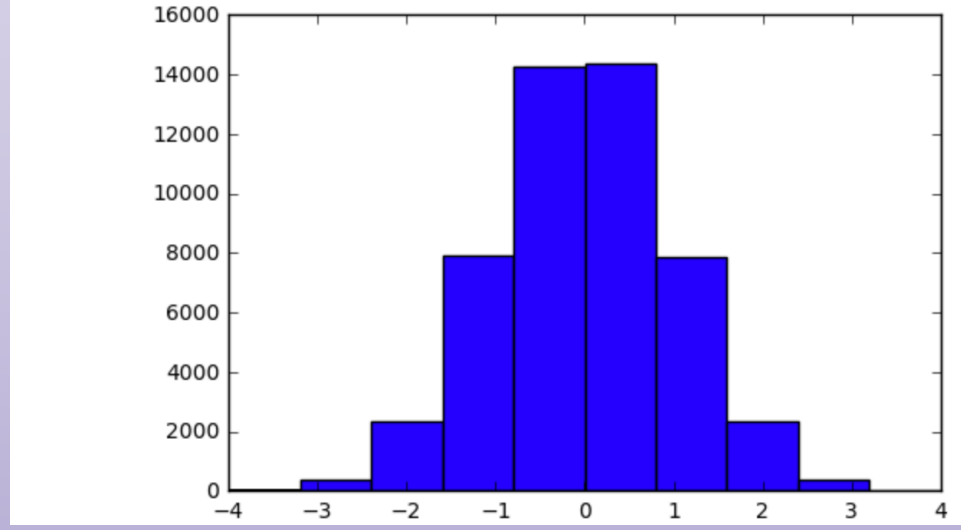
0

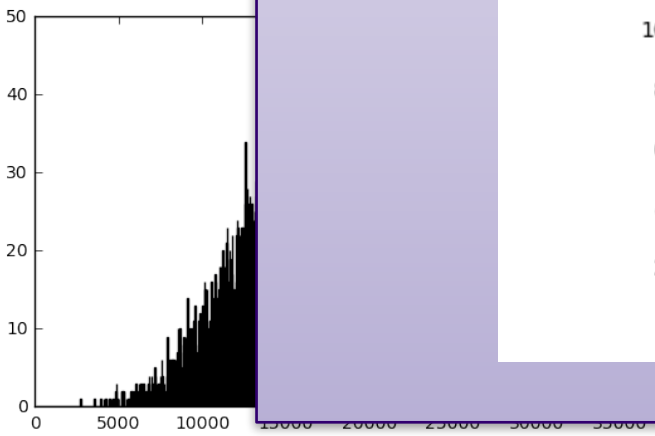
If you do not make any selection of the number of bins when you call the histogram subroutines in numpy, it will make a selection for you.

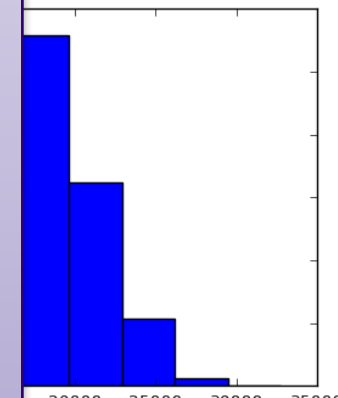
In general it is a ***terrible*** idea to let the computer pick parameters for you w/o any input.

Remember what happened when we let Excel design our plots?

```
In [24]: plt.hist(mydata);
```



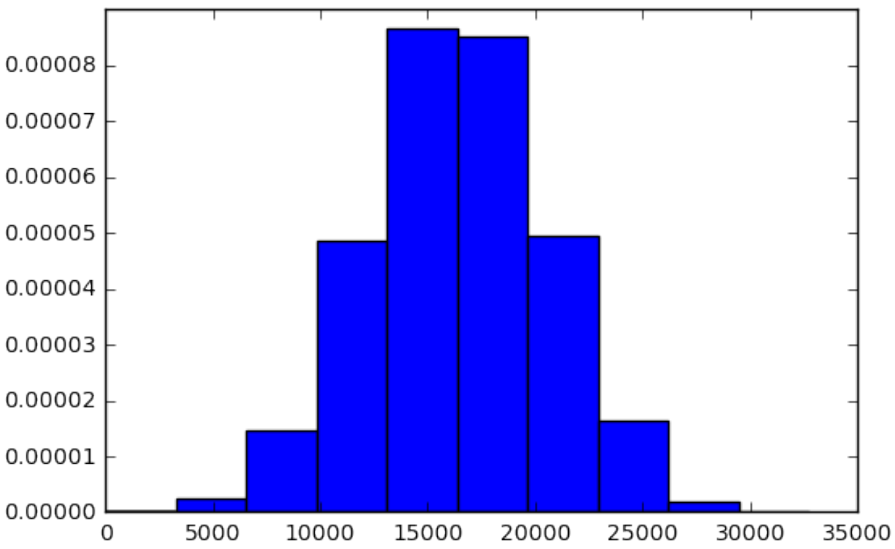


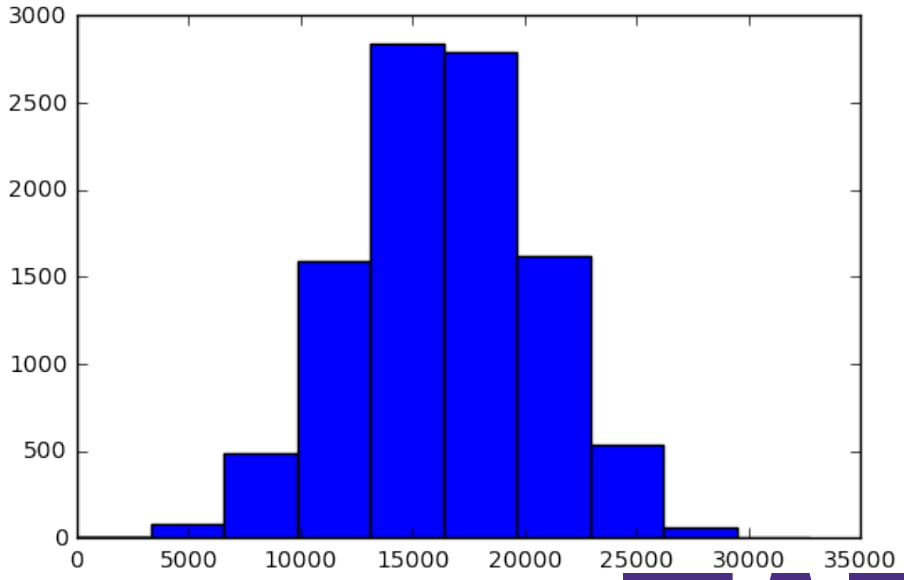


There are some heuristics (e.g., see Wikipedia page subsection "number of bins and width") but no "right answer"

Refresher: histogram (3)

- **The choice of** normalization **determines the unit and magnitude of the** y-axis
- **What is the integral of the** normalized **histogram?**






```
plt.hist(mydata,bins=100,normed=True);
plt.hist(mydata,bins=50,normed=False);
```

Histograms and probability density/mass

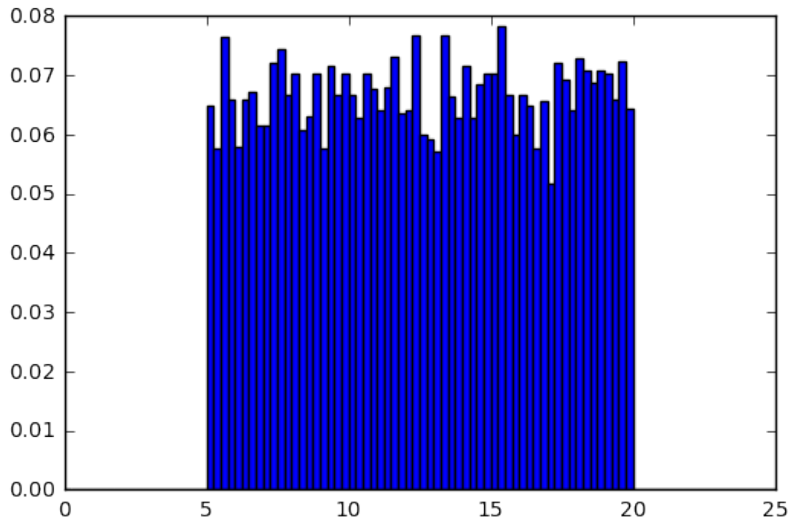
- **For continuous data, the probability of choosing any individual value is zero (at maximum precision)**
  - *Contrast to our 'warmup' when we looked at integers!*
- **The integral of the** probability density function **(PDF) of a** continuous **variable describes the likelihood of randomly drawing a variable between some values a, b**
- **What is the relationship between the PDF and the histogram?**

Histograms and probability density

- **In general, a** PDF**, is a representation of the** population **for** continuous **data**
  - **Consider that if you know** all of the data **it is possible to determine an analytical function that represents how likely it is to obtain any piece of your data**
  - **A** probability mass function (PMF) **is the analogous quantity for** discrete **data**
- **In contrast, a** histogram**, is usually used to represent your** sample.
- **There is no law that says you can't represent all your data as a histogram, or estimate a PDF with a sample**

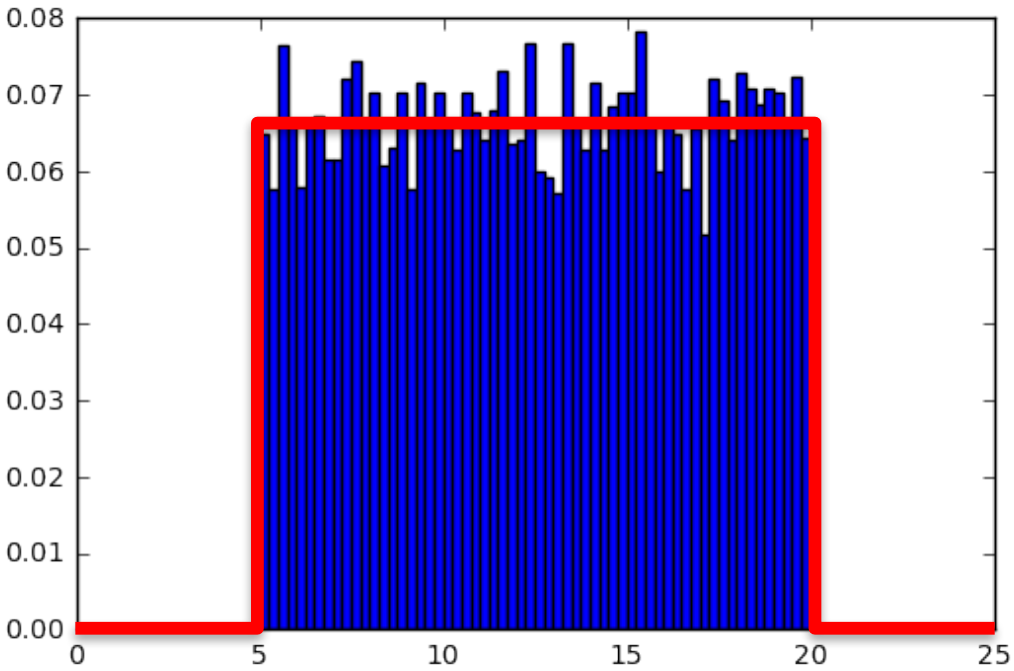
Histograms and probability density

- **For the** uniform **distribution we have been playing with today:**
  - **Normalized histogram of 1000 continuous data drawn between 5 and 20**
  - **What should the maximum value be in the PDF?**
    - PDFs are usually represented as *P(x) or F(x)*



Histograms and probability density

- **For the** uniform **distribution we have been playing with today:**



Specific to Python number draw:

$$P(x) = \begin{cases} 0 & x < 5 \\ \dfrac{1}{20 - 5} & 5 \le x < 20 \\ 0 & x \ge 20 \end{cases}$$

In general we write:

$$P(x) = \begin{cases} 0 & x < a \\ \dfrac{1}{b - a} & a \le x \le b \\ 0 & x > b \end{cases}$$

Probability density and cumulative density

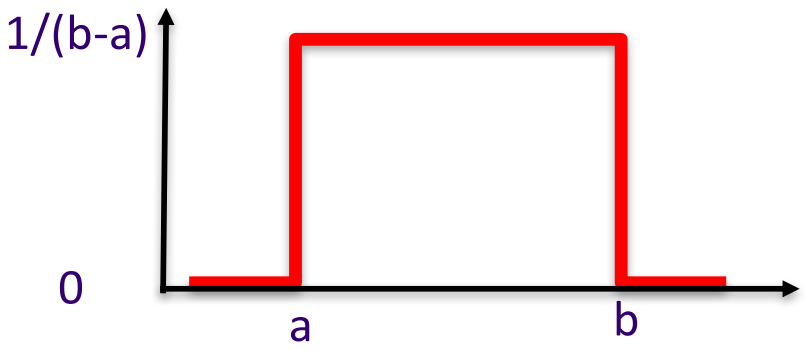
- **We often display a running integral of the** PDF **as a means to answer the question: what is the probability the value is "***less than x***" → This is the** cumulative distribution function (CDF)

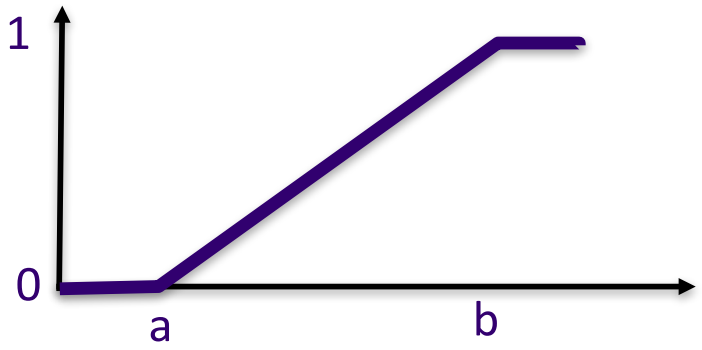
For the uniform PDF

$$P(x) = \begin{cases} 0 & x < a \\ \frac{1}{b-a} & a \le x \le b \\ 0 & x > b \end{cases}$$

The CDF becomes:

$$CDF(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \le x \le b \\ 1 & x > b \end{cases}$$





The binomial distribution (discrete data)

- Imagine a scenario in which you classify hypothetical battery anodes as “success” or “failure” on the basis of achieving a certain level of sustained performance.
- Some properties of the experiment/classification:
  - There are n repeat trials of the experiment, which are all completely independent
  - Every trial has two possible outcomes: success or failure
  - The probability (p) of success is the same on each trial

The binomial distribution PDF

Two ways to write the PDF for a binomial experiment and how to read this statement properly

$$P(n \mid N) = \binom{N}{n} p^n q^{N-n} = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}$$

"The probability of n successes given N trials"
"is equal to"
"N choose n with replacement"
"times the probability of of success of n trials"
"times the probability of failure of (N-n) trials"

This is more convenient for doing the maths

Python gives us a trivial software stack for exploring a huge range of PDFs!

The Poisson distribution (discrete data over continuous intervals)

- In contrast to the binomial experiment where we count attributes of the data (e.g., number of successes). There are many experiments in which we count the rate of success over a fixed interval
  - "Over a 60 minute interval, how many Uber drivers get called to Denny Hall?"
- Some properties of this experiment/classification:
  - Still classifying successes (e.g., *yes 5 drivers got here in 60 min*)
  - Take care with failures and the definition of your question
    - some failures don't make sense "how many Uber drivers do not arrive at Denny Hall" a useless question
  - Probability of success is proportional to size of region (ask the same question using 600,000 minutes)
  - Probability of success in vanishingly small region is zero

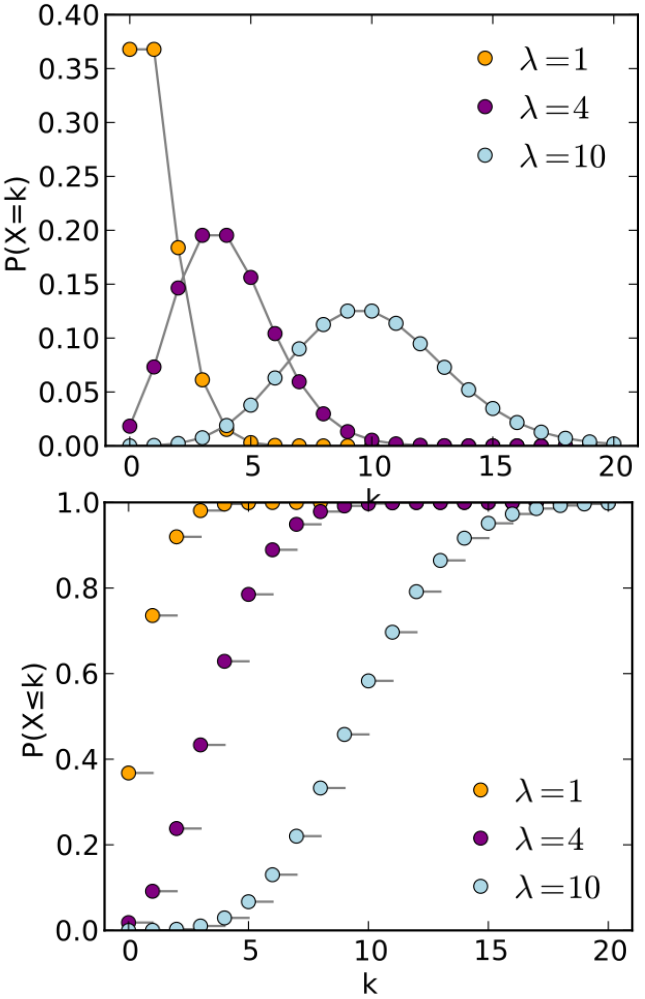
The Poisson distribution

- **Big picture: Poisson distributions tell us about statistics of** *Poisson processes* **→ this is the basis for much of kinetics, reaction dynamics, stochastic processes**
  - **Analogy from chemistry: Poisson statistics can be related to the** turnover number **and its likelihood in catalysis**
  - **Analogy from materials science: Poisson statistics can be related to the** frequency of defects **in materials processing**
- **The Poisson distribution is the limit of the binomial distribution as N → infinity.**
  - **Variables we care about** ν **(**expected number of events in an interval**) and** n **(**our interval**)**

$$P_\nu(n) = \frac{\nu^n e^{-n}}{n!}$$

Teaching you to fish (worst pun)

- Look at the syntax of how we got access to cool stuff from the binomial distribution
- With a partner do the following:
  - Examine the plot of Poisson PDF / CDF on the Wikipedia page (also right)
  - Recreate these plots, stacked, in a single Python notebook cell
    - With available time keep making it look exact like the page
    - **If you make a perfect replica I will give you:** A HUGE HIGH FIVE
- 10 min – go



Distributions for all your data!

- Discrete variables
  - Binomial
  - Poisson
- Continuous variables
  - Normal
    - Gaussian
  - Gamma

The normal distribution

- You are probably already really familiar with the normal distribution
- What are some properties of it?
- What is the difference between normal and Gaussian distributions?

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{(x-\mu)^2}{2\sigma^2}\right)}$$

Given some population mean (μ) and variance (σ^2), what is the probability of choosing some specified value x?

$$P(x) = \frac{1}{\sqrt{2\pi}} e^{-\left(\frac{x^2}{2}\right)}$$

The **standard normal** has a mean ($\mu=0$) and variance ($\sigma^2=1$)

The gamma distribution

- **The gamma distribution, in essence, is a continuous version of the Poisson**
  - **The variable we are collecting / working with is the wait time between subsequent Poisson processes**
  - **Consider a spectroscopic experiment that collects a set of single molecule data: individual times between chemical reactions (**i.e., the inverse of a reaction rate**)**
  - **Often just interested in the mean of such data, however we may wish to know the PDF…**

Given some rate ($1/\theta$) or "scale" (θ) and a shape factor (α), what is the probability of choosing some specified value x?

$$P(x) = \frac{x^{\alpha-1} e^{-x/\theta}}{\Gamma(\alpha)\theta^{\alpha}}$$

Getting some practice with data

- Lots of things we have done so far:
  - Plot time series or correlated data
  - Plot PDFs using Python tools
  - Made histograms of data
- Lets practice a bit more with HCEPD and look how some of our data are distributed
- **Important:** You also need to understand what kind of data you have before you can ask the question “what is the likelihood my data are different?”

Distributions of data in the HCEPD

- Load the HCEPD file (1st 100k)
- Make histograms of
  - PCE , mass , and e_gap_alpha
  - Note observations and/or expectations
- If you complete fast, take the exercise we did in *gettingwarm* and adapt it for the ID data frame and see your sampling of the IDs becomes normal? (plot the histogram of IDs first)

Sampling from distributions and the central limit theorem

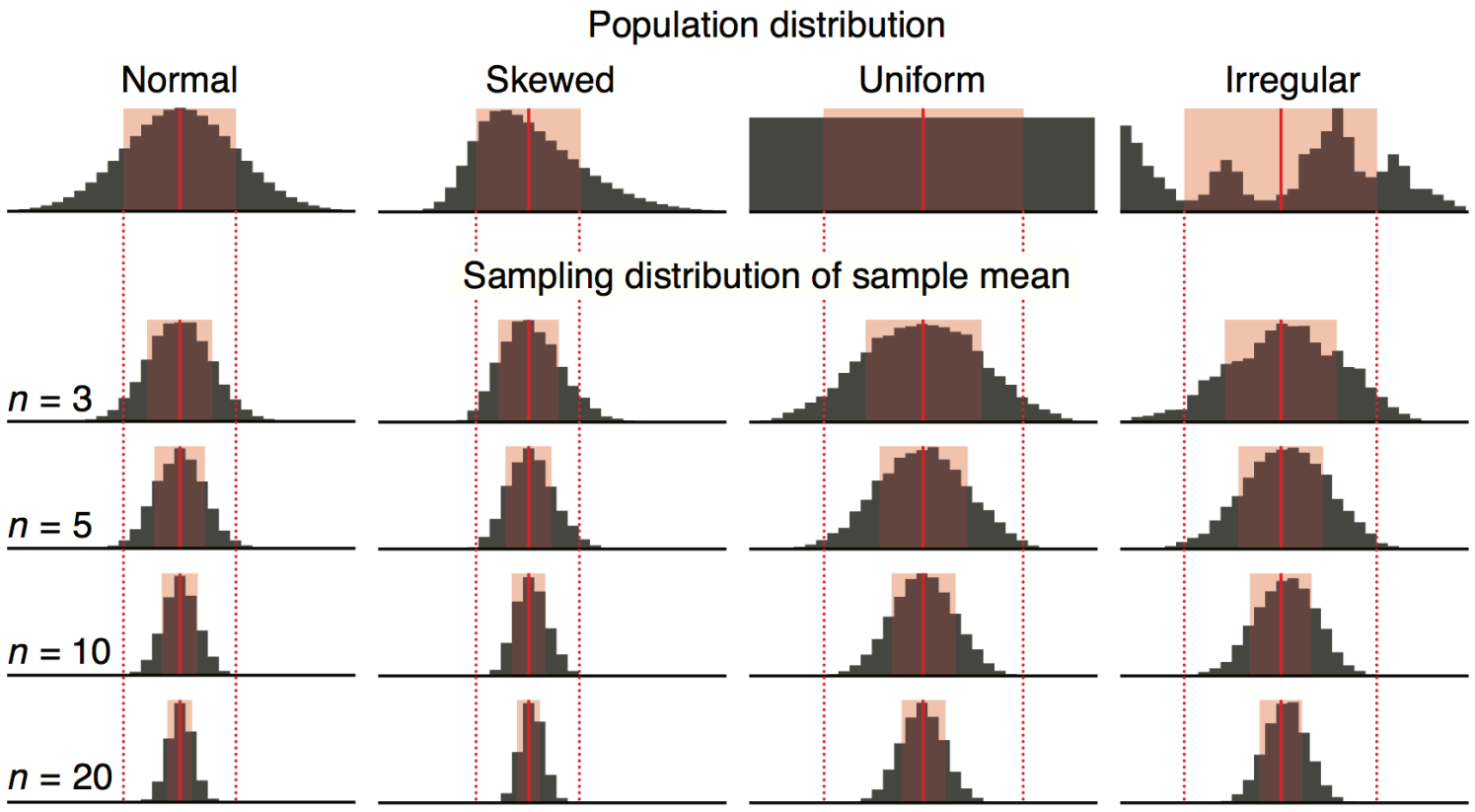


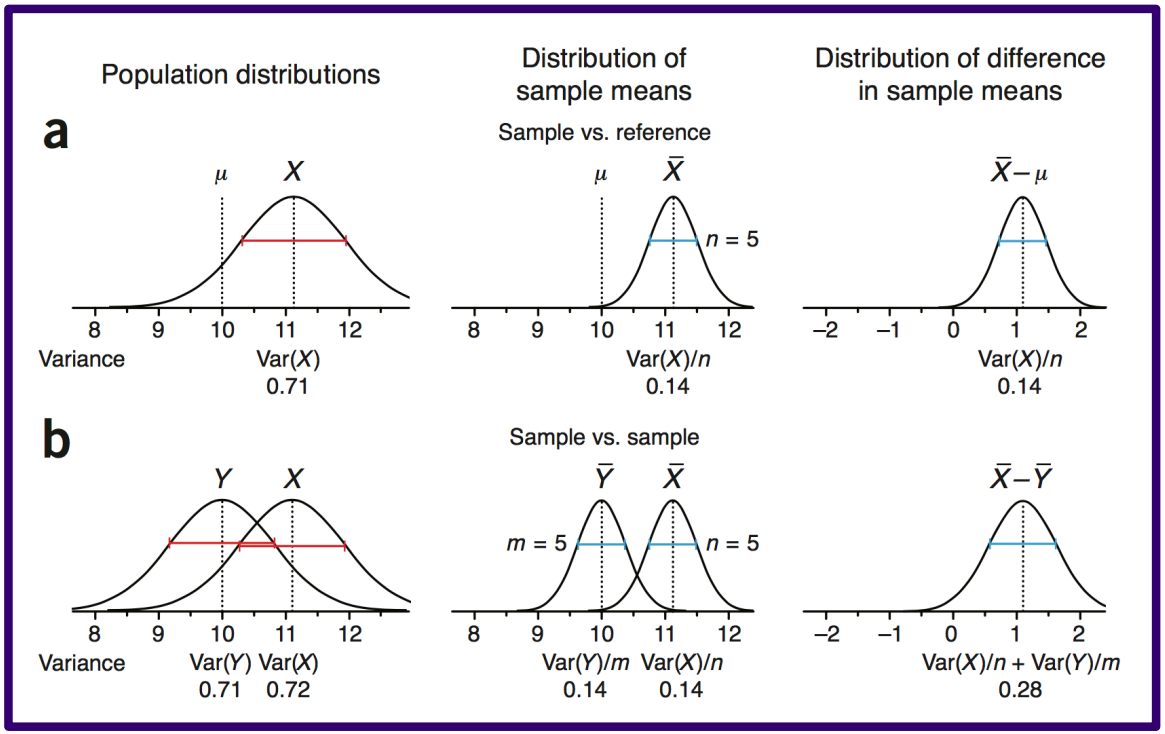
Figure 3 | The distribution of sample means from most distributions will be approximately normally distributed. Shown are sampling distributions of sample means for 10,000 samples for indicated sample sizes drawn from four different distributions. Mean and s.d. are indicated as in **Figure 1**.

Sampling from distributions and the central limit theorem

- Any big picture ideas why this matters?
- Questions / discussion?

Why do distributions even matter?

- A preview of statistical hypothesis testing
- The question we are often asking is “is there an effect of A?” but this manifests itself differently in statistics…



UNIVERSITY *of* WASHINGTON