Estimation and Sampling

Darren Reger

DSI Lecture

galvanize

- Review common *moments*: μ, σ^2
- Motivation: Why estimate distributions?
- Four standard methods:
  - *Method of Moments* (MOM)
  - *Maximum Likelihood Estimation* (MLE)
  - *Maximum a Posteriori* (MAP)
  - *Kernel Density Estimation* (KDE)
- Parametric vs nonparametric methods

Review

galvanize

Expectation

- Discrete: Probability weighted average of all possible values

$$E(X) = x_1 * p_1 + x_2 * p_2 + \dots + x_k * p_k$$

- Continuous: Same idea, except replace Σ with integral, and replace probabilities with probability densities

$$E(X) = \int_{-\infty}^{\infty} x * f(x) dx$$

Review: *Moments* of a random variable *X*

n^{th} Raw Moment:

$$\mu_n = E[(X)^n]$$

What does the E operator do?

n^{th} Central Moment:

$$\mu'_n = E[(X - E[X])^n]$$

Why do we call this the "central" moment?

(hint: look at the equation)

galvanize

galvanize

$$\mu = E[X]$$

For discrete random variables with *k* possible outcomes:

$$E[X] = \sum_{i=1}^{k} x_i P(X = x_i)$$

Fill in the blank:

P is the "Probability ___ Function"

For continuous random variables:

$$E[X] = \int_{-\infty}^{\infty} x f(x)\, dx$$

Fill in the blank:

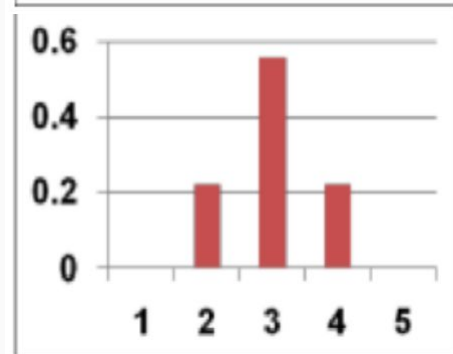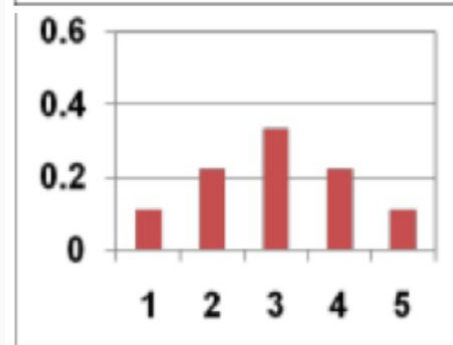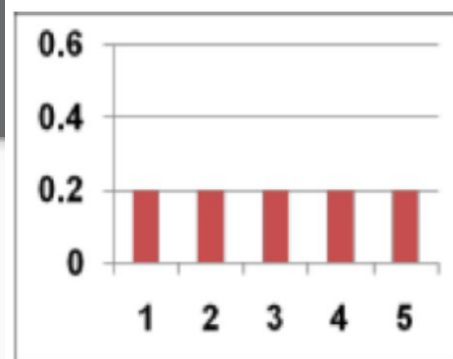*f* is the "Probability ___ Function"

$$\sigma^2 = E[(X - E[X])^2]$$

For discrete random variables with *k* possible outcomes:

$$\sigma^2 = \sum_{i=1}^{k} (x_i - \mu)^2 P(X = x_i)$$

For continuous random variables:

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\, dx$$

Why Estimate Distributions?

Why estimate distributions?

Example 1: You have data on how many people order cakes every day at your bakery, and you want to estimate the probability of selling out.

Example 2: You have data on how often your car breaks down, and you want to know your chances of safely crossing the country in it.

Example 3: You have data on how many people visit your website each day, and you want to know the probability of your servers being overloaded.

Credit to Jancsy for these examples.

![galvanize]

Econometrician's Philosophy

If you lack the information to determine the value directly, estimate the value to the best of your ability using the information you do have.
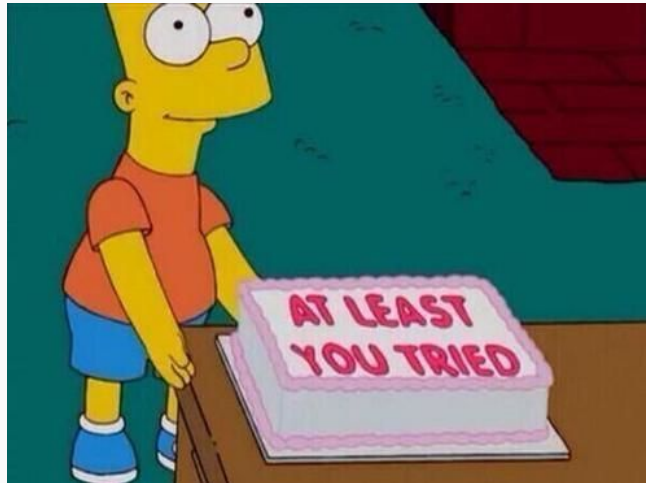
Method of Moments (MOM)

Method of Moments (MOM)

Overview:

1. **Assume an underlying distribution for your domain.**
   E.g. Poisson, Bernoulli, Binomial, Gaussian

2. **Compute the relevant sample moments.**
   E.g. Mean, Variance

3. **Plug the sample moments into the PMF/PDF of the assumed distribution.**

Your website visitor log shows the following number of visits for each of the last seven days: [6, 4, 7, 4, 9, 3, 5]. What's the probability of zero visitors tomorrow?

Which underlying distribution should we assume?

Poisson! Let's look at Wikipedia to remind ourselves what it is. :)

What moment should we estimate?

The mean. Our mean estimate will become the estimate for the only parameter used in the Poisson distribution: λ

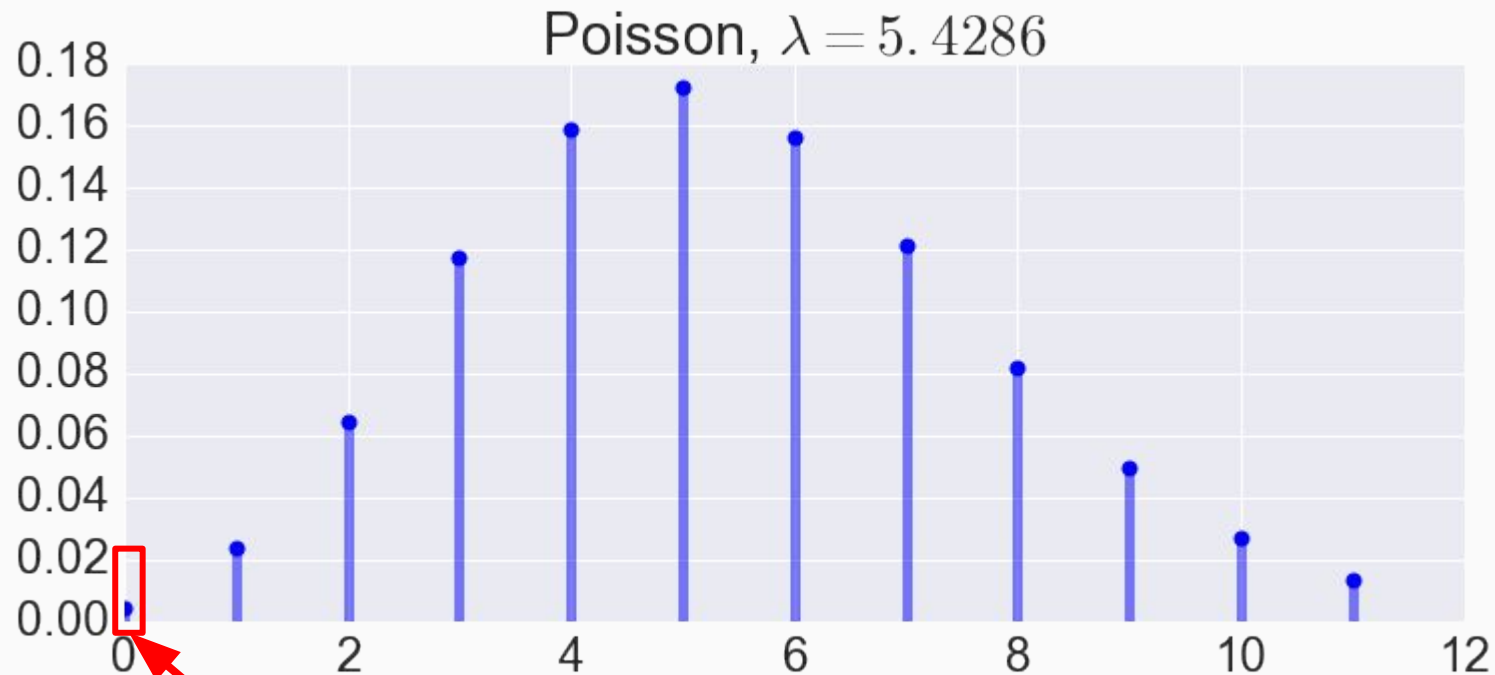Example 1 (con't): Method of Moments

Your website visitor log shows the following number of visits for each of the last seven days: [6, 4, 7, 4, 9, 3, 5]. What's the probability of zero visitors tomorrow?

From our samples, estimate the mean:

$$\bar{x} =$$

Our mean estimate is used to estimate λ:

$$\lambda =$$

Example 1 (con't): Method of Moments

galvanize

Poisson, $\lambda = 5.4286$


```
print scipy.stats.poisson.???
0.00438936184278
```

You flip a coin 100 times. It comes up heads 52 times. What's the MOM estimate that in the next 100 flips the coin will be heads <= 45 times?

Which underlying distribution should we assume?

Binomial… note: We really only have one binomial sample here.

What moment should we estimate?

The mean. We actually only have one sample here where the result is 52. So the mean is 52.

Example 2 (con't): Method of Moments

galvanize

You flip a coin 100 times. It comes up heads 52 times. What's the MOM estimate that in the next 100 flips the coin will be heads <= 45 times?

From our one binomial sample, we know:

The binomial distribution has mean:

$$\bar{x} = 52$$

$$\mu = np$$

What does MOM say to do next?

Use the sample moments to estimate the distribution parameters. In this case, the parameter we need to estimate is p.
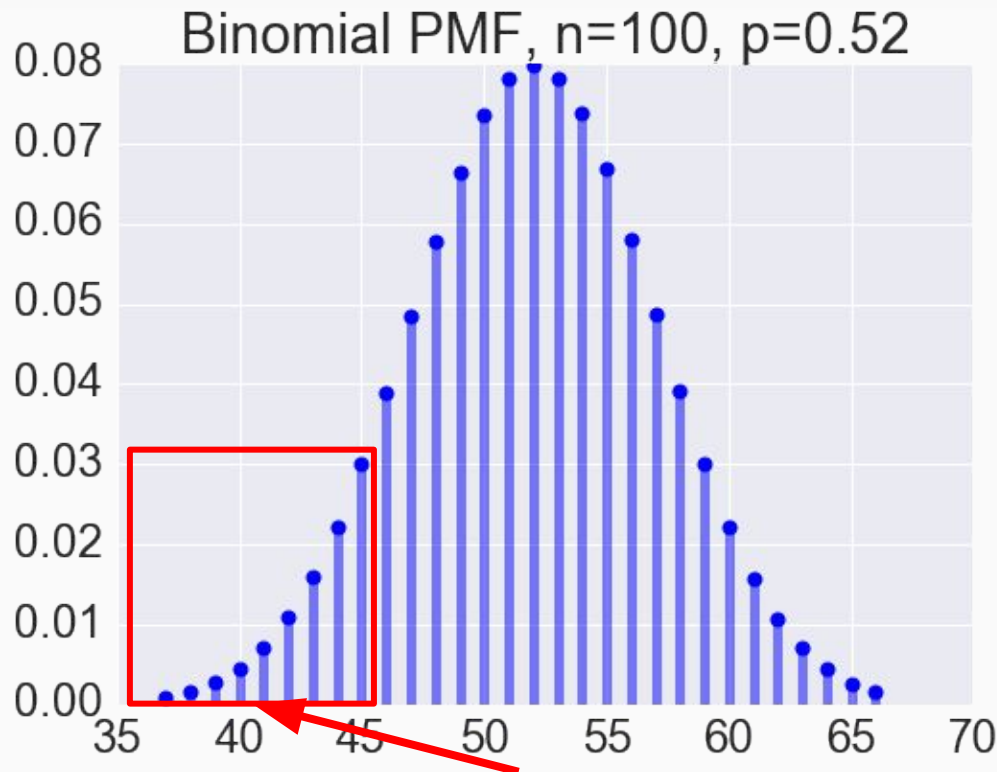
Example 2 (con't): Method of Moments

$$52 = np$$

$$n = 100$$

$$p = 52/100$$

$$p = 0.52$$

Binomial PMF, n=100, p=0.52

Probability in the next 100 flips the coin will be heads <= 45 times:

```
print scipy.stats.binom.???
```

Maximum Likelihood Estimation (MLE)

galvanıze

Maximum Likelihood Estimation (MLE)

Overview:

Law of Likelihood:
 If $P(X|H1) > P(X|H2)$, then the evidence supports *H1* over *H2*.

Question:
 Which hypothesis does the evidence most strongly support?

Answer:
 The hypothesis H that maximizes $P(X|H)$, which is found via *MLE*.

Maximum Likelihood Estimation (MLE)

Overview:

1. **Assume an underlying distribution for your domain.** (just like with MOM)
   E.g. Poisson, Bernoulli, Binomial, Gaussian

2. **Define the likelihood function.**
   We want to know the likelihood of the data we observe under different distribution parameterizations.

3. **Choose the parameter set that maximizes the likelihood function.**

$$f(x_1, x_2, \ldots, x_n | \theta) = f(x_1 | \theta) f(x_2 | \theta) \ldots f(x_n | \theta)$$

True because we assume X is i.i.d.
Recall, what does i.i.d. Mean?
What's the i. part? What's the i.d. part?

$$\mathcal{L}(\theta | x_1, x_2, \ldots, x_n) = f(x_1, x_2, \ldots, x_n | \theta) = \prod_{i=1}^{n} f(x_i | \theta)$$

We will find θ to maximize the log-likelihood function:

$$\hat{\theta}_{\mathrm{mle}} = \arg \max_{\theta \in \Theta} \log \left(\mathcal{L}(\theta | x_1, \ldots, x_n) \right)$$

You flip a coin 100 times. It comes up heads 52 times. What's the MOM estimate that in the next 100 flips the coin will be heads <= 45 times?

Yep, same example...

Which underlying distribution should we assume?

Binomial… (like last time)

… now we need to define our likelihood function...

Example 3 (con't): Maximum Likelihood Estimation

galvanize

$$X_i \overset{iid}{\sim} Bin(n, p) \qquad i = 1, 2, \ldots, n \qquad f(x_i|p) = \binom{n}{x_i} p^{x_i} (1 - p)^{n - x_i}$$

$$\log \mathcal{L}(p) = \sum_{i=1}^{n} \left[\log \binom{n}{x_i} + x_i \log p + (n - x_i) \log(1 - p) \right]$$

$$\frac{\partial \log \mathcal{L}(p)}{\partial p} = \sum_{i=1}^{n} \left[\frac{x_i}{p} - \frac{n - x_i}{1 - p} \right] = 0$$

$$\hat{p}_{MLE} = \boxed{\frac{\bar{x}}{n}}$$

For the Binomial distribution, MOM and MLE give the same answer!

Maximum a Posteriori (MAP)

galvanize

Maximum a Posteriori (MAP)

Recall, MLE finds θ to maximize:

$$f(x_1, x_2, \ldots, x_n | \theta)$$

Whereas, MAP finds θ to maximize:

$$f(\theta | x_1, x_2, \ldots, x_n)$$

Maximum a Posteriori (MAP)

MLE → MAP, just use Bayes' Theorem

$$f(\theta|x) = \frac{f(x|\theta)g(\theta)}{\int_{\theta' \in \Theta} f(x|\theta')g(\theta')\,d\theta'} \propto \boxed{f(x|\theta)}g(\theta)$$

MLE is just this part.

MLE vs MAP

MLE solves:

$$\hat{\theta}_{\text{mle}} = \arg\max_{\theta \in \Theta} f(x_1, x_2, \ldots, x_n | \theta)$$

MAP includes the prior belief.

MAP solves:

$$\hat{\theta}_{\text{map}} = \arg\max_{\theta \in \Theta} f(x_1, x_2, \ldots, x_n | \theta) \boxed{g(\theta)}$$

Nonparametric Techniques

Question: How can we model data that does not follow a known distribution?

Answer: Use a nonparametric technique.

Kernel Density Estimation (KDE)

Kernel Density Estimation Motivation
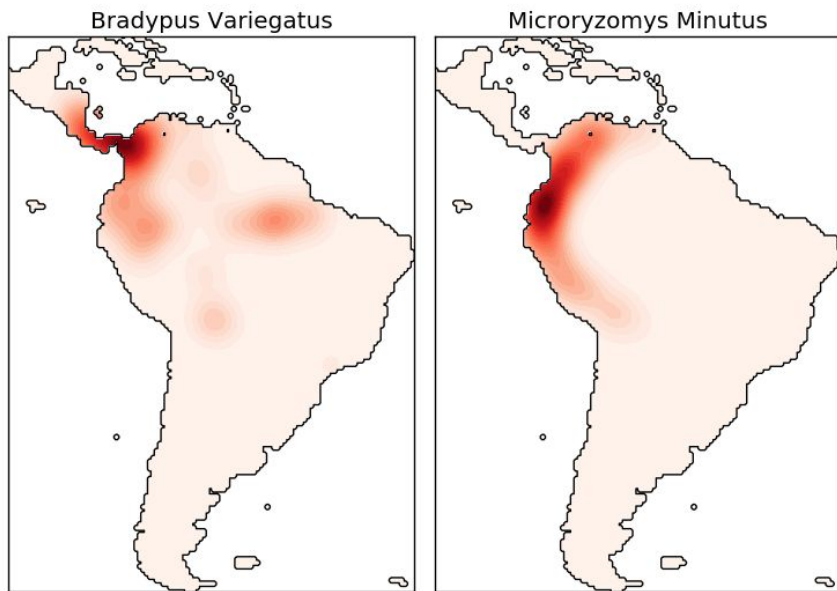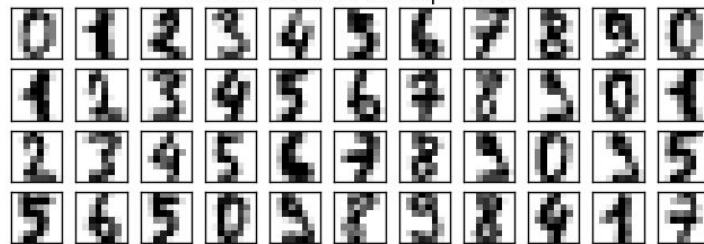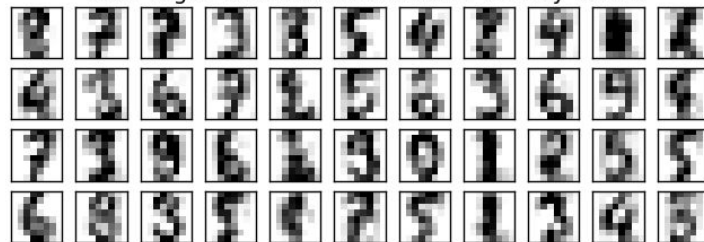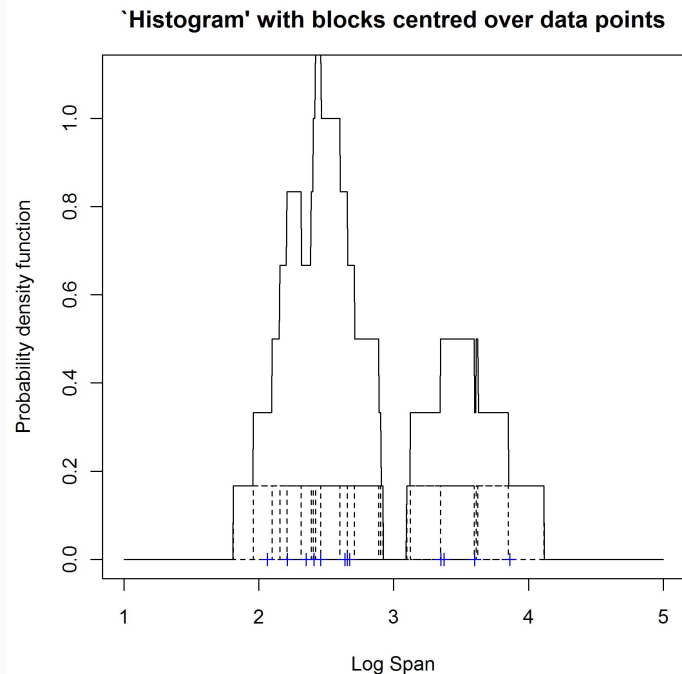
Histogram Troubles
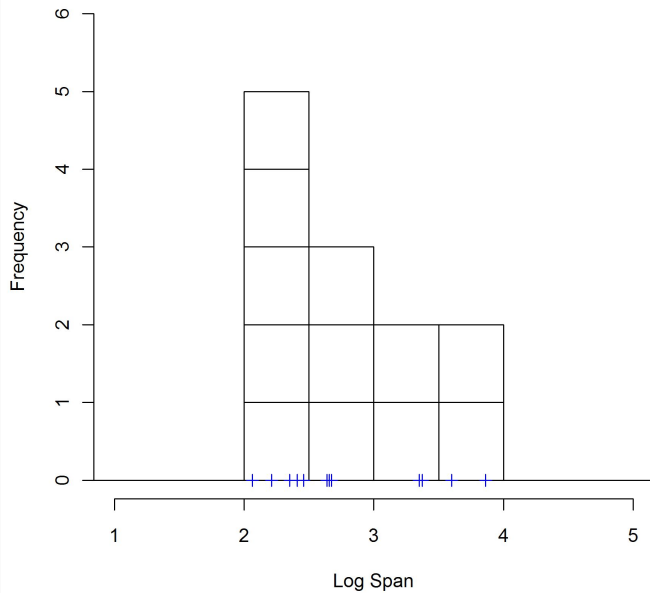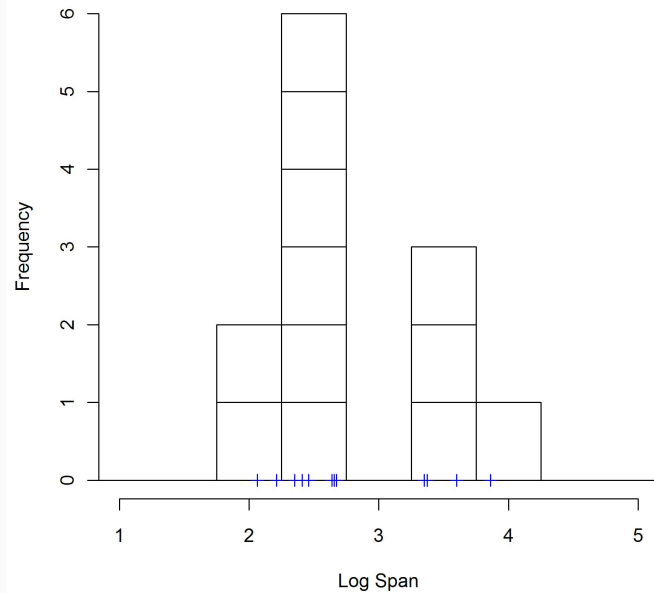
Box Kernel Density Estimate

KDE is a nonparametric way to estimate the PDF of a random variable. KDE makes the histogram by summing "kernel functions."
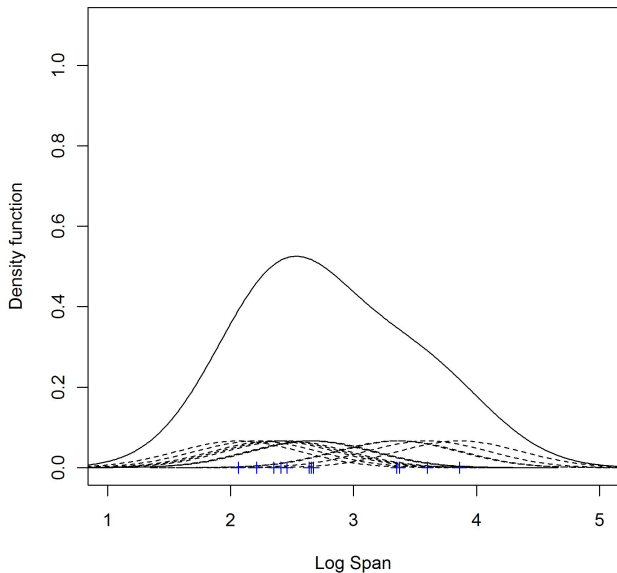
In Simpler Terms:

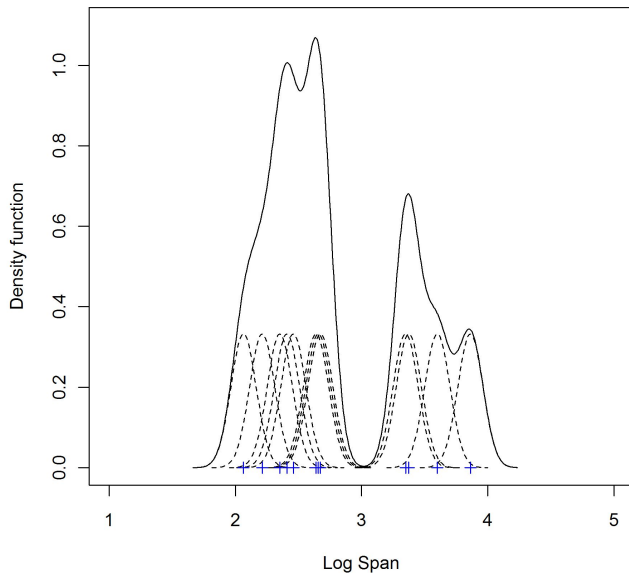1) Center the kernel (box in this case) over each data point
2) Stack them up

`Histogram' with blocks centred over data points

Am I Too Smooth… Or Not Smooth Enough?

Parametric vs Nonparametric Methods

galvanize
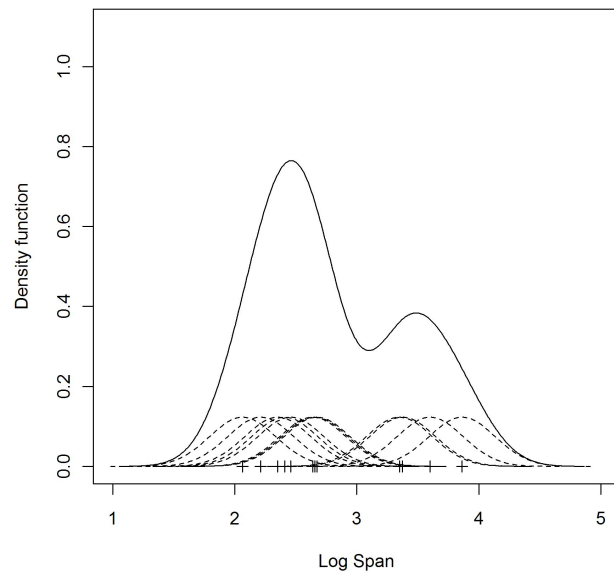
Estimating Distributions
Parametric vs Nonparametric Methods

Parametric: We assume an underlying distribution, then we use our data to estimate the parameters of that underlying distribution. E.g. Using:

- *Method of Moments* (MOM)
- *Maximum Likelihood Estimation* (MLE)
- *Maximum a Posteriori* (MAP)

Nonparametric: We don't assume any *single* underlying distribution, but instead we fit a combination of distributions to the observed data. E.g. Using:

- *Kernel Density Estimation* (KDE)

Parametric methods:

1. Based on assumptions about the distribution of the underlying population and the parameters from which the sample was taken.
2. If the data deviates strongly from the assumptions, could lead to incorrect conclusions.

Nonparametric methods:

1. NOT based on assumptions about the distribution of the underlying population.
2. Generally not as powerful -- less inference can be drawn.
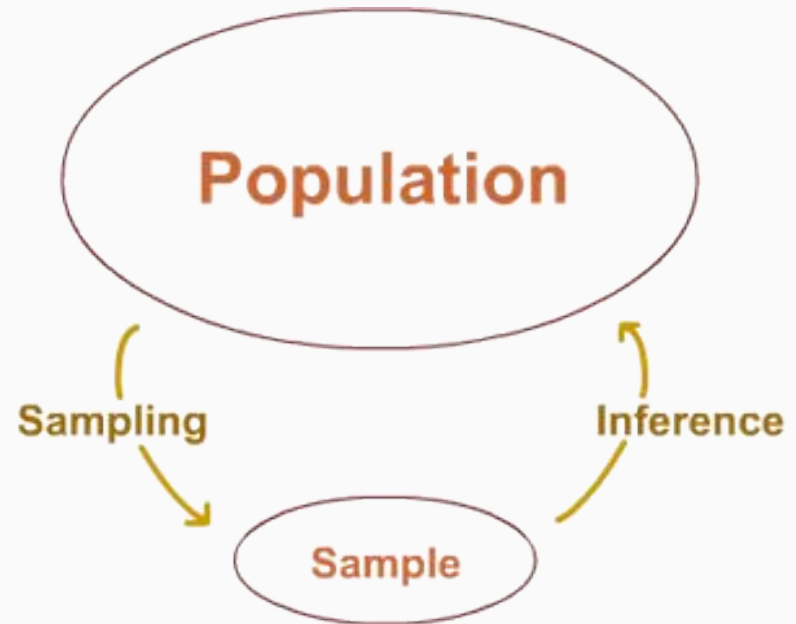3. Interpretation can be difficult… what does the wiggly curve mean?

Sampling

- Population Inference & Sampling
- Central Limit Theorem
- Confidence Intervals
- Bootstrapping

galvanize

Population Inference & Sampling

galvanize

Population Inference

- Start with a question/hypothesis
- Design an experiment
- Collect data
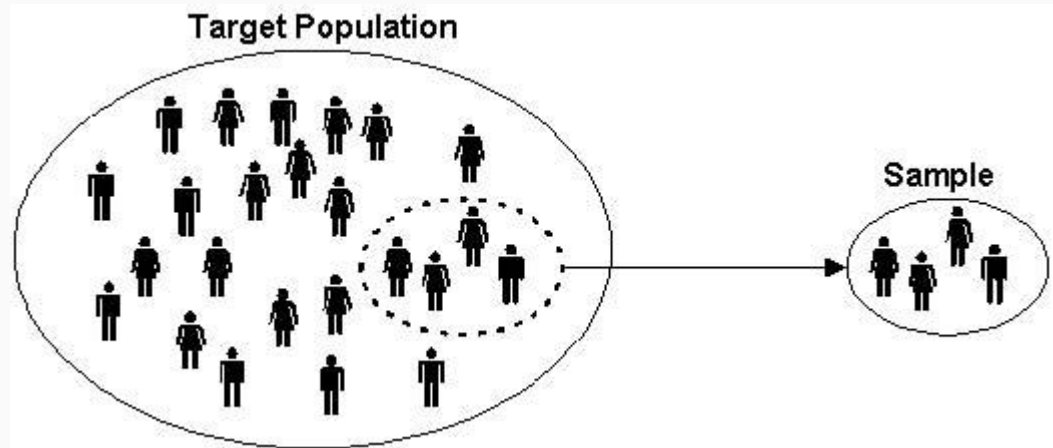- Analyze
- Check the results
- Repeat? Redesign?

Collecting data: Taking a sample

A sample should be representative of the population.

Random sampling is often the best way to achieve this. Ideally: **each subject has an equal chance of being in the sample**.

Types of Sampling

1) Simple Random Sampling
2) Stratified Sampling
3) Cluster Sampling

Discussion of Sampling Galvanize Students

Random sampling is harder than it sounds...

Scenario: You want to estimate the percentage of dog owners in SF.

Method 1: Go to the nearby dog park and ask **random** people if they own dogs until you have *n* responses.

Method 2: Stand on 24th and Mission and ask **random** people if they own dogs until you have *n* responses.

Method 3: Repeat *n* times: Pick a **random** neighborhood in SF (weighted by census data per neighborhood), go to that neighborhood, ask **random** people you see if they own dogs until you get 1 response.

Random sampling... just do the best you can.

Often it's impossible to do *perfect* random sampling.

So...

1. do the best you can,
2. call out possible objections, and
3. make a case for why you think your results are valid.
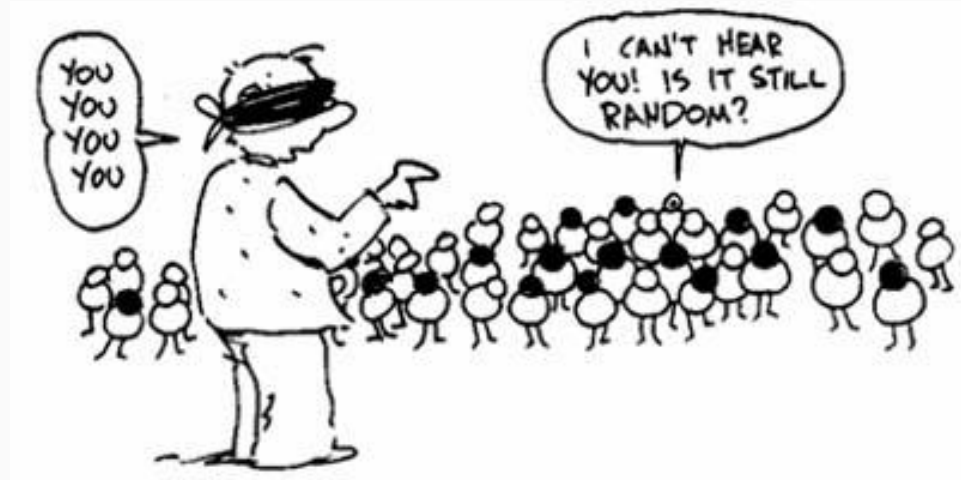
Random sampling in the digital age...

You might think that random sampling in a digital context is easier, and you're right! But there are still gotchas.

Scenario: *Slack* is testing a new feature ("channel polling", a way to survey people in a channel). They'd like to test the feature on only a subset of their users (*n*), then draw inference about their entire userbase.

Method 1: SELECT user_id FROM users LIMIT n;

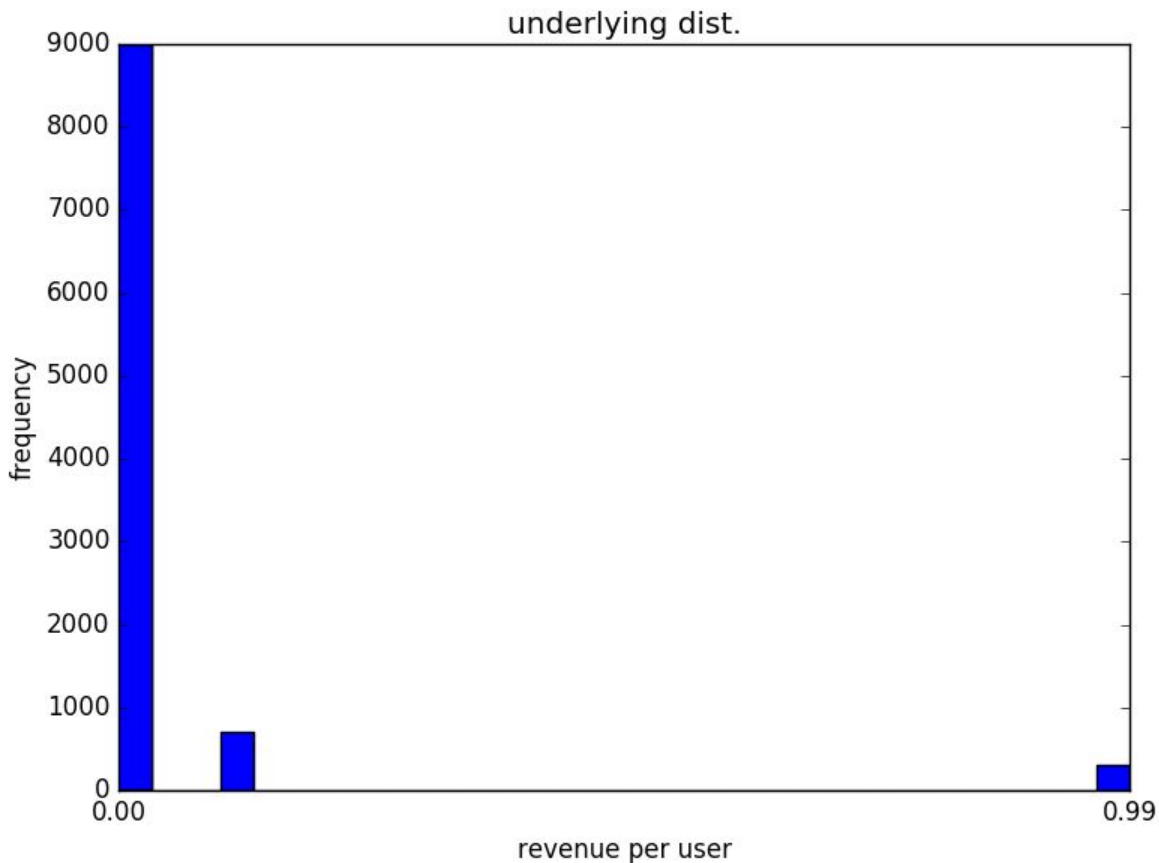**Method 2:** SELECT user_id FROM users ORDER BY RAND() LIMIT n;

Central Limit
Theorem

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

galvanize

galvanize

Underlying Distribution:

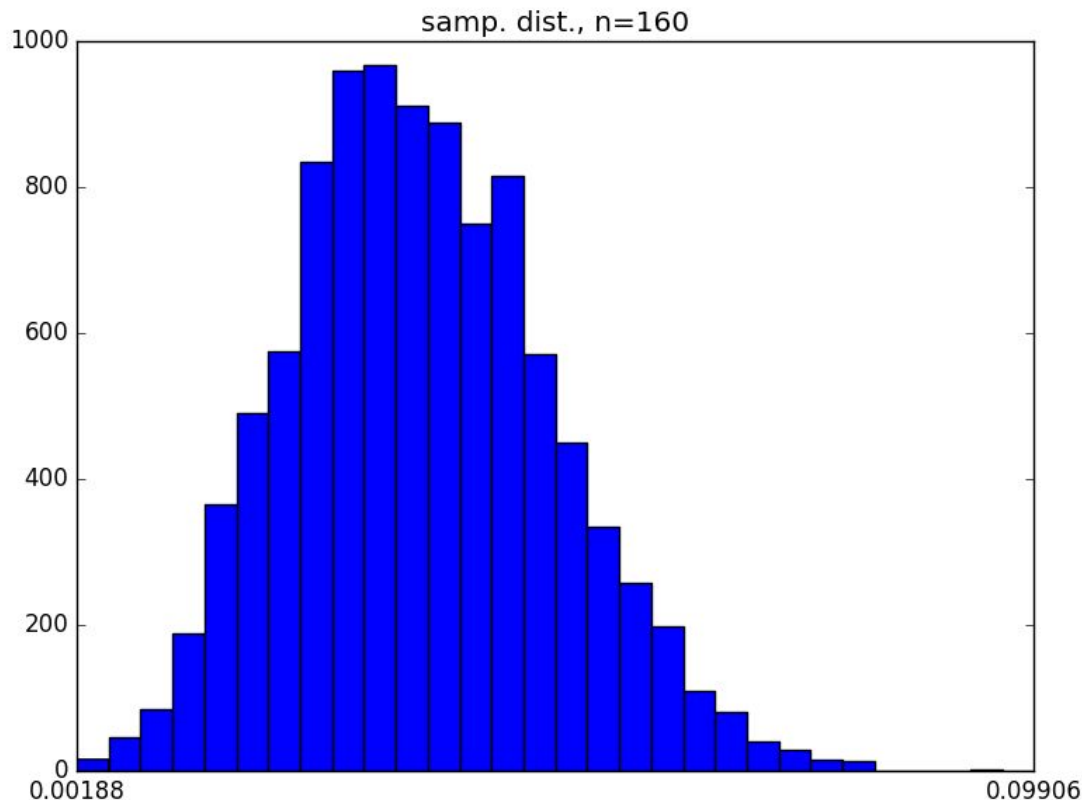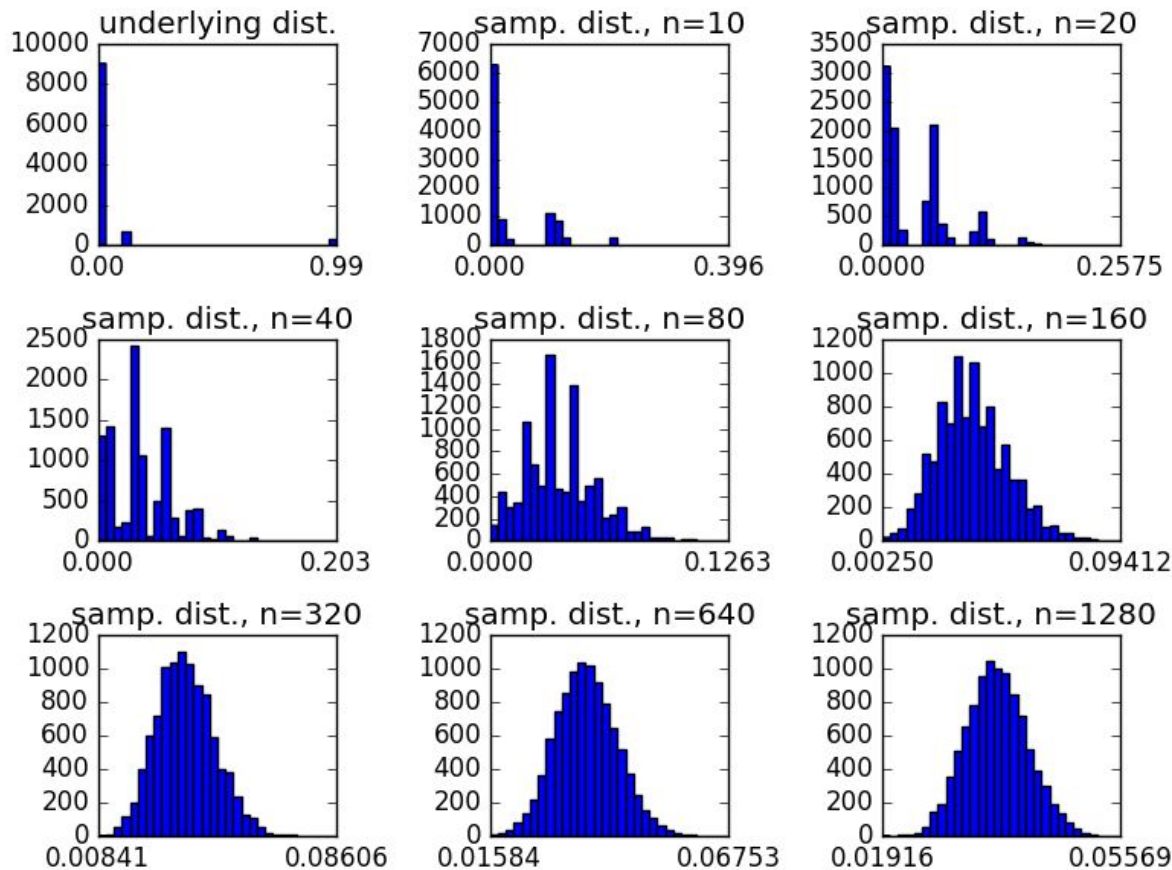| Random variable: X = revenue per visitor	P(X):
X = $0.00 (no revenue)	90%
X = $0.10 (ad-click)	7%
X = $0.99 (app purchase)	3%

Collect *n* samples from the website revenue distribution, calculate the sample mean $\bar{x}$

Repeat 10,000 times, we get:

$$\bar{x}_0, \bar{x}_1, \dots, \bar{x}_{9999}$$
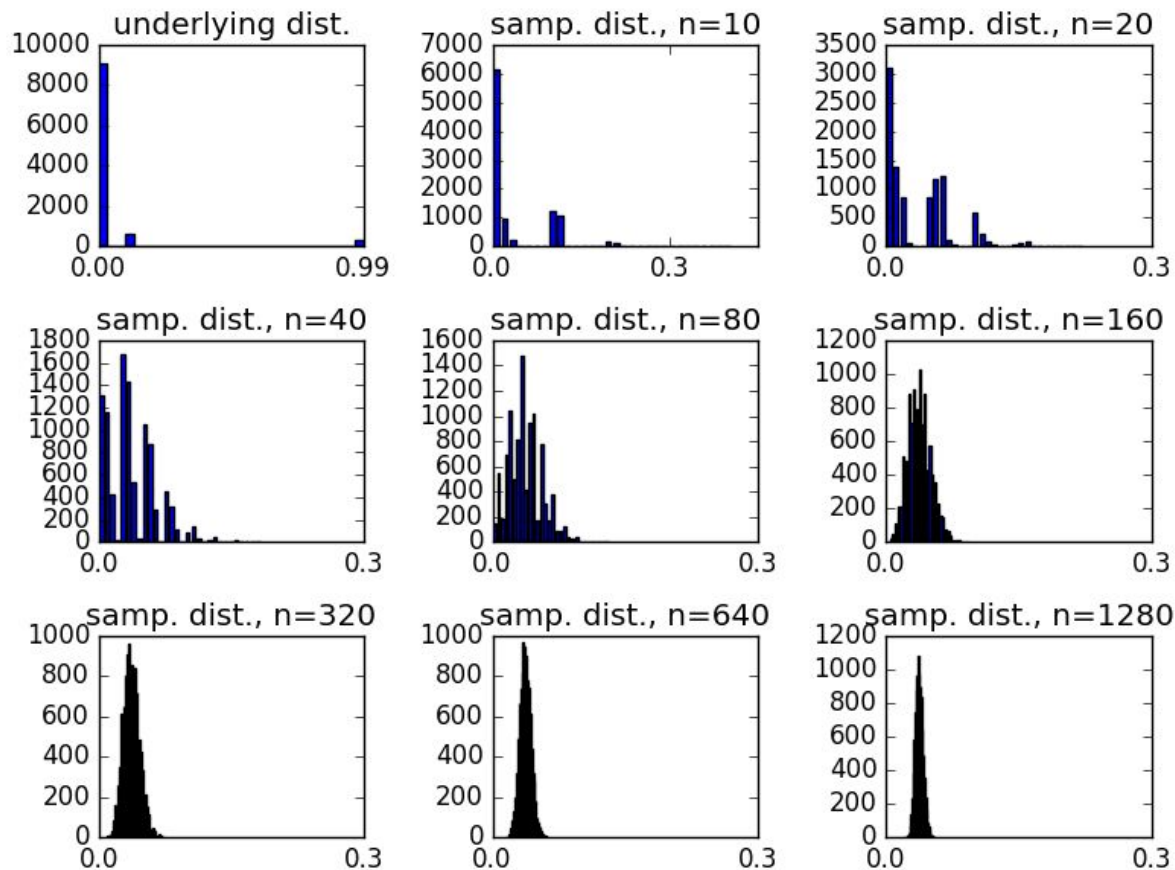
Plot all 10,000 sample means.

samp. dist., n=160

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

The distribution of sample means (aka, the "sampling distribution") is normally distributed. *

* Under certain conditions; e.g. sufficiently large sample sizes, and i.i.d. r.v.

Same charts as the previous slide, but now the scale of each x-axis is the same!

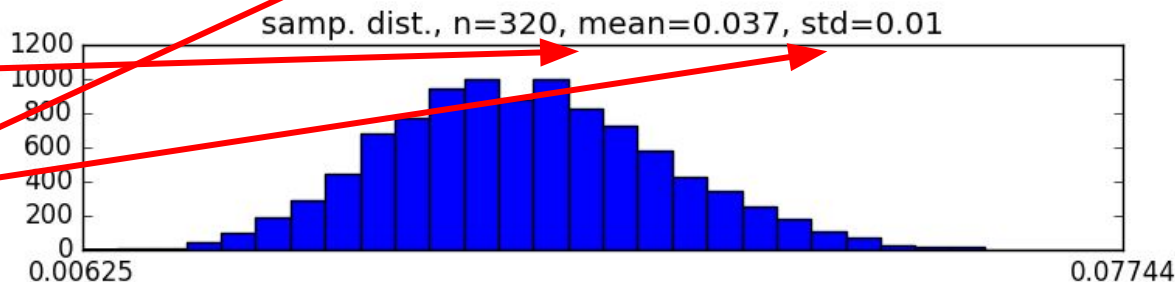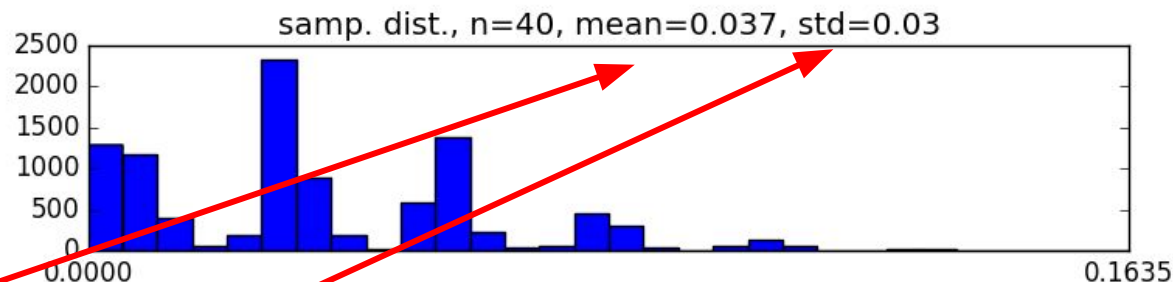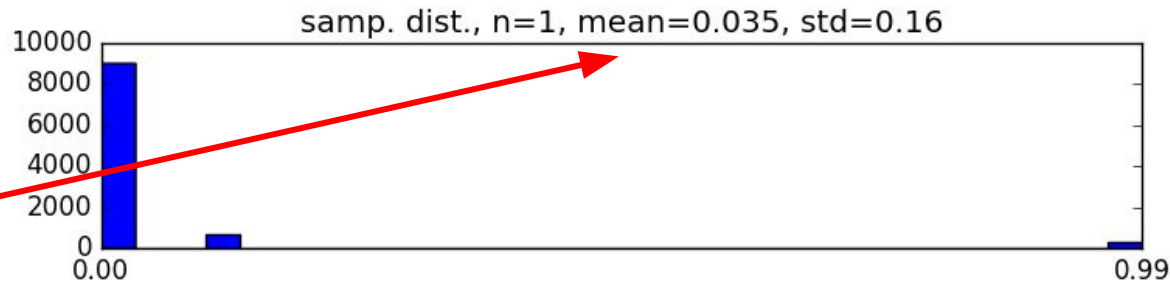Now we can see: What happens when the sample size increase?

galvanize

Let the underlying distribution have mean and std. dev.

$$\mu \text{ and } \sigma$$

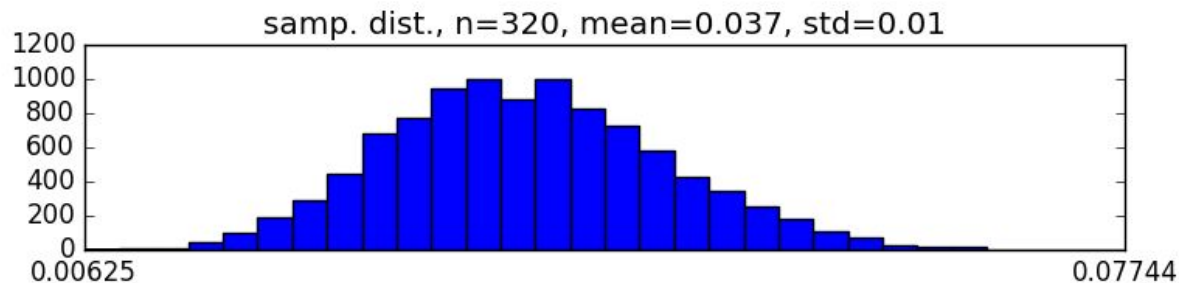The sampling distribution's mean and std. dev. will equal:

$$\mu' = \mu$$
$$\sigma' = \sigma / \sqrt{n}$$

samp. dist., n=1, mean=0.035, std=0.16

samp. dist., n=40, mean=0.037, std=0.03

samp. dist., n=320, mean=0.037, std=0.01

galvanize

Intuitively, why does the mean stay the same in each histogram?

Intuitively, why does the std. dev. decrease as the sample size increases?

Confidence Intervals

galvanıze
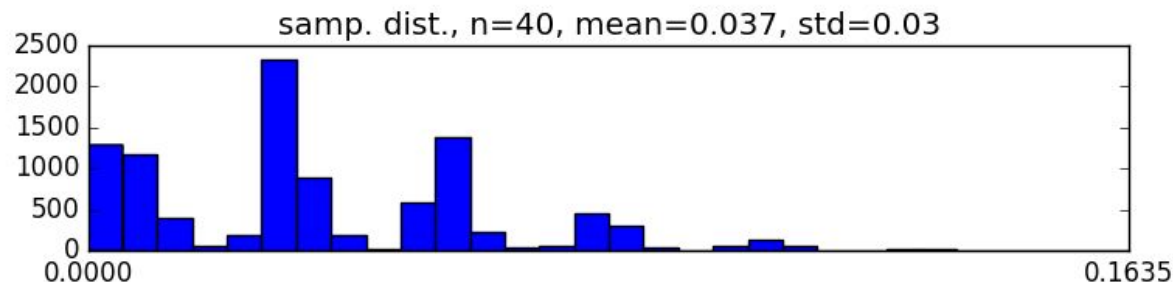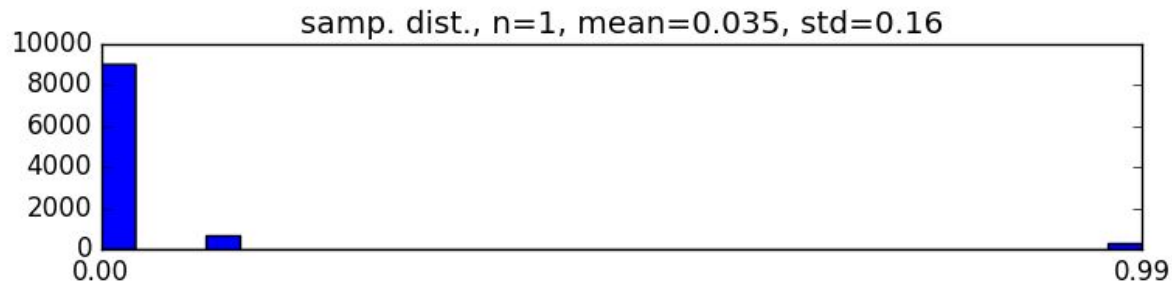
Confidence Interval

A *confidence interval* (CI) is an interval estimate of a population parameter.

The typical level of confidence is 95%, but they can be calculated for any level.

For example, a 95% CI for the population mean is given by:

$$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

Grr. We don't know sigma...

Why assuming the normal distribution here?

Where does the sqrt(n) come from?

```
conf = 0.95
scipy.stats.norm.ppf((1 + conf) / 2.)   # <-- handling two-sided-ness
```

Confidence Interval (con't)

Since we don't know sigma, we can substitute *s* for it:

$$\bar{x} \pm 1.96 \boxed{\frac{s}{\sqrt{n}}}$$
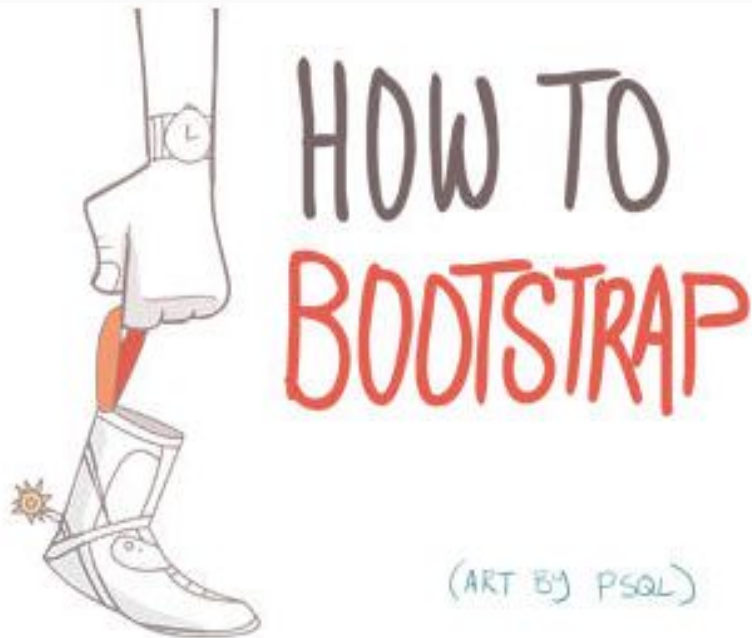
Btw, this value is the "standard error of the mean (SEM)"

When *n* is small (<30), we should use the t-distribution instead of the normal:

$$\bar{x} \pm t_{(\alpha/2, n-1)} \frac{s}{\sqrt{n}}$$

Bootstrapping

Pull Yourself Up By Your Bootstraps

HOW TO BOOTSTRAP

(ART BY PSQL)

Bootstrap Sampling

Estimates the **sampling distribution** of an estimator by sampling with replacement from the original sample.

Advantages:

- Completely automatic
- Available regardless of how complicated the estimator may be

Often used to estimate the standard errors and confidence intervals of an unknown population parameter or something normally not confidence intervalable (I invented that word).

Bootstrap Sampling

Method:

1. Start with your dataset of size n
2. Sample from your dataset with replacement to create 1 bootstrap sample of size n
3. Repeat B times
4. Each bootstrap sample can then be used as a separate dataset for estimation and model fitting

Bootstrap Sampling
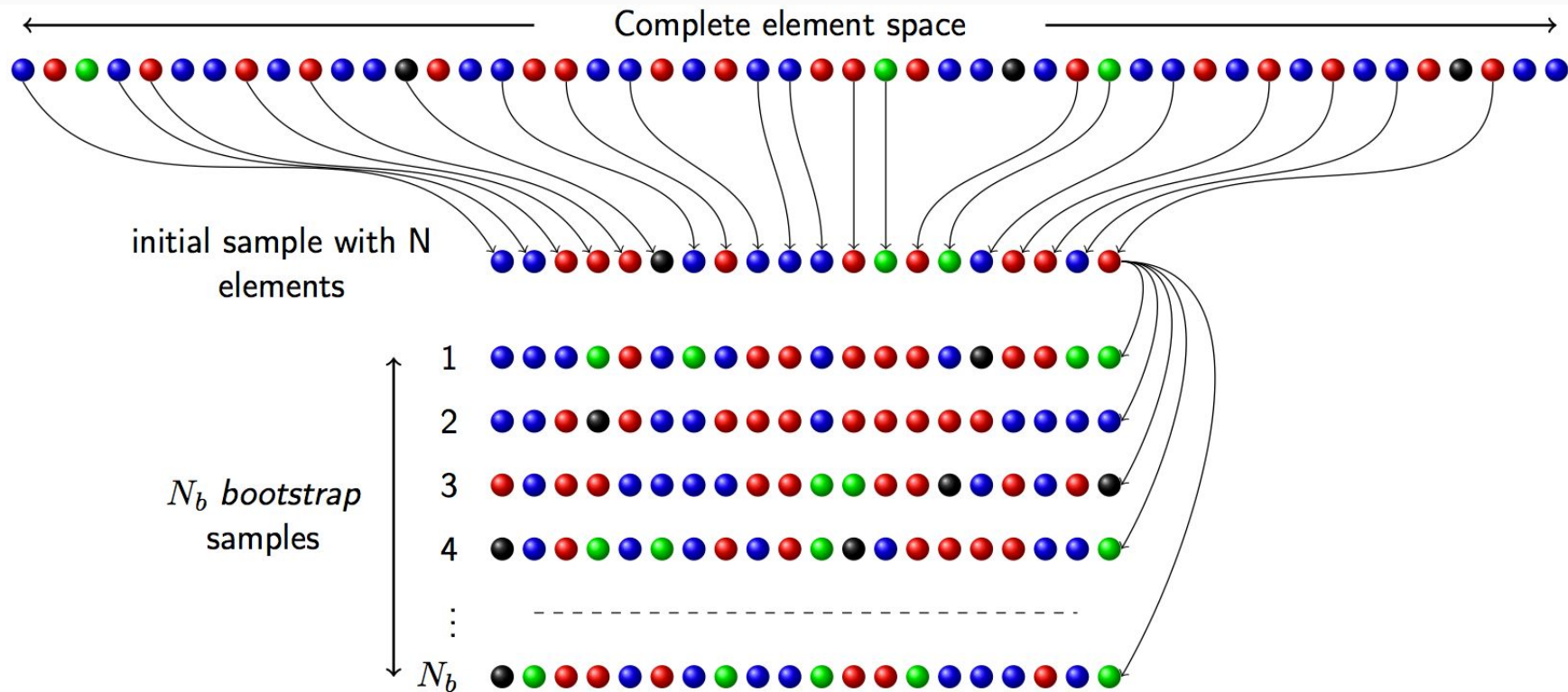
1. Draw a bootstrap sample:

$$X_1^*, X_2^*, \ldots, X_n^*$$

2. Calculate bootstrap estimate of your parameter (the parameter you're interested in):

$$\hat{\theta}^* = t(X_1^*, X_2^*, \ldots, X_n^*)$$

3. Repeat steps 1 and 2, B times to get:

$$\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_B^*$$

4. Calculate the bootstrapped variance (maybe):

$$s_{\text{boot}}^2 = \frac{1}{B} \sum_{b=1}^{B} (\hat{\theta}_b^* - \bar{\theta}^*)^2 \qquad \text{where } \bar{\theta}^* = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_b^*$$

Bootstrap Confidence Intervals

Percentile method:

$$(\hat{\theta}^*_{\alpha/2}, \hat{\theta}^*_{1-\alpha/2})$$

Interval assuming approximately *normal* bootstrap sampling distribution:

$$\bar{\theta}^* \pm 1.96\, s_{\text{boot}}$$

galvanize

When to Bootstrap

When the theoretical distribution of the statistic (parameter) is complicated or unknown.   (E.g. Median or Correlation)

When the sample size is too small for traditional methods.

Favor accuracy over computational cost.

Questions

- MOM vs. MLE
    - What do they solve for?   Parameter Estimation
    - How does each approach tackle the problem?
        - Both assume a specific distribution already.
        - MOM uses moment matching to get at parameters
        - MLE asks what parameter would maximize the the likelihood of the resulting data

- How about MAP?
    - How does it relate to the MLE?   Similar to MLE, but need to account for Prior

- What's bootstrapping? Random sampling w/ replacement technique
    - When might I think of using it? Want sense of accuracy of some sample estimate
    - What are the steps to setting up a bootstrap estimate?
        - Sample w/ replacement, B times → Compute B estimates from B samples → Get Standard Errors, Confidence Intervals, etc.