

Estimating Distributions

Ryan Henning

Customized by: Frank Burkholder

Further Customized by: Brent Lemieux

4/11/2017



- 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 common *moments*



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)

$$\mu = E[X]$$

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:

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

What does
variance tell
us?

For continuous random variables:

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

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

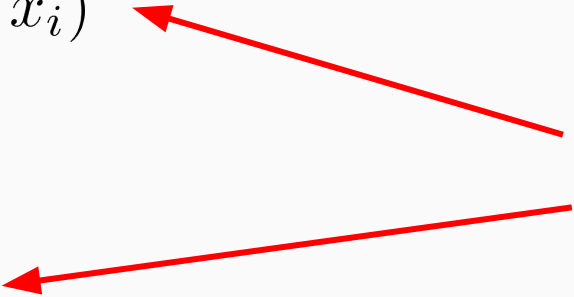
For discrete random variables with k possible outcomes:

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

For continuous random variables:

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

Variance
measures the
“spread” of data.
Can you see
that in the
equations?



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.

Distribution	PDF or PMF	Mean	Variance
$Bernoulli(p)$	$\begin{cases} p, & \text{if } x = 1 \\ 1 - p, & \text{if } x = 0. \end{cases}$	p	$p(1 - p)$
$Binomial(n, p)$	$\binom{n}{k} p^k (1 - p)^{n-k}$ for $0 \leq k \leq n$	np	npq
$Geometric(p)$	$p(1 - p)^{k-1}$ for $k = 1, 2, \dots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
$Poisson(\lambda)$	$e^{-\lambda} \lambda^x / x!$ for $k = 0, 1, 2, \dots$	λ	λ
$Uniform(a, b)$	$\frac{1}{b-a} \quad \forall x \in (a, b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$Gaussian(\mu, \sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2
$Exponential(\lambda)$	$\lambda e^{-\lambda x} \quad x \geq 0, \lambda > 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

- $X \sim Bernoulli(p)$ = Single coin flip turns out to be Heads
- $X \sim Binomial(100, p)$ = # of coin flips out of 100 that turn out to be Heads
- $X \sim Geometric(p)$ = # of Trials until coin flip turns out to be Heads
- $X \sim Poisson(\lambda=10)$ = # of taxis passing a street corner in a given hour (on avg 10/hr)
- $X \sim Exponential(\lambda=10)$ = Time until taxi will pass street corner
- $X \sim Uniform(0,360)$ = Degrees between hour hand and minute hand
- $X \sim Gaussian(100, 10)$ = IQ Score

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)

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?

What moment should we estimate?

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!

What moment should we estimate?

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

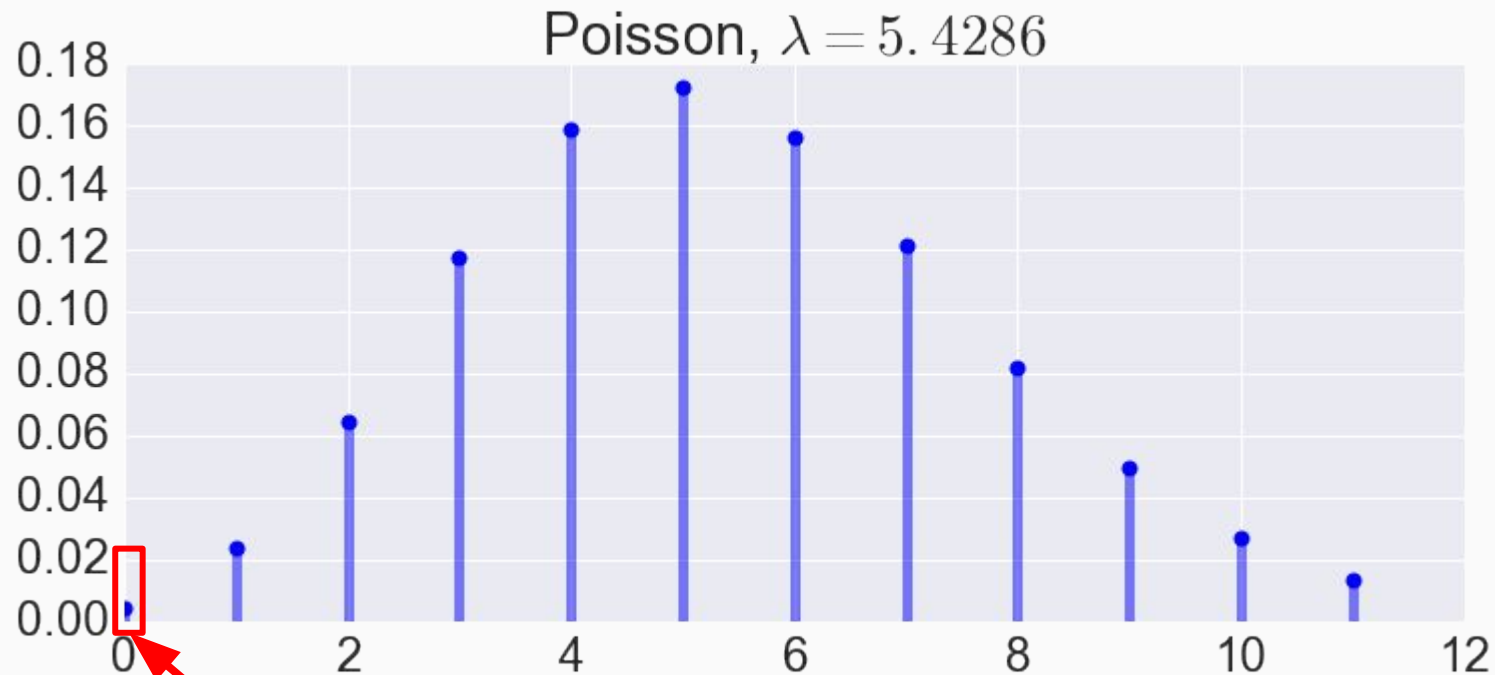
Your website visitor log shows the following number of visits for each of the last seven days: [6, 4, 7, 4, 9, 3, 5]. What's the probability of zero visitors tomorrow?

From our samples, estimate the mean:

$$\bar{x} = (6 + 4 + 7 + 4 + 9 + 3 + 5)/7 = 5.43$$

Our mean estimate is used to estimate λ :

$$\lambda = 5.43$$



```
print scipy.stats.poisson.pmf(0, lmda)
```

```
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?

What moment should we estimate?

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.

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:

$$\bar{x} = 52$$

The binomial distribution has mean:

$$\mu = np$$

What does MOM say to do next?

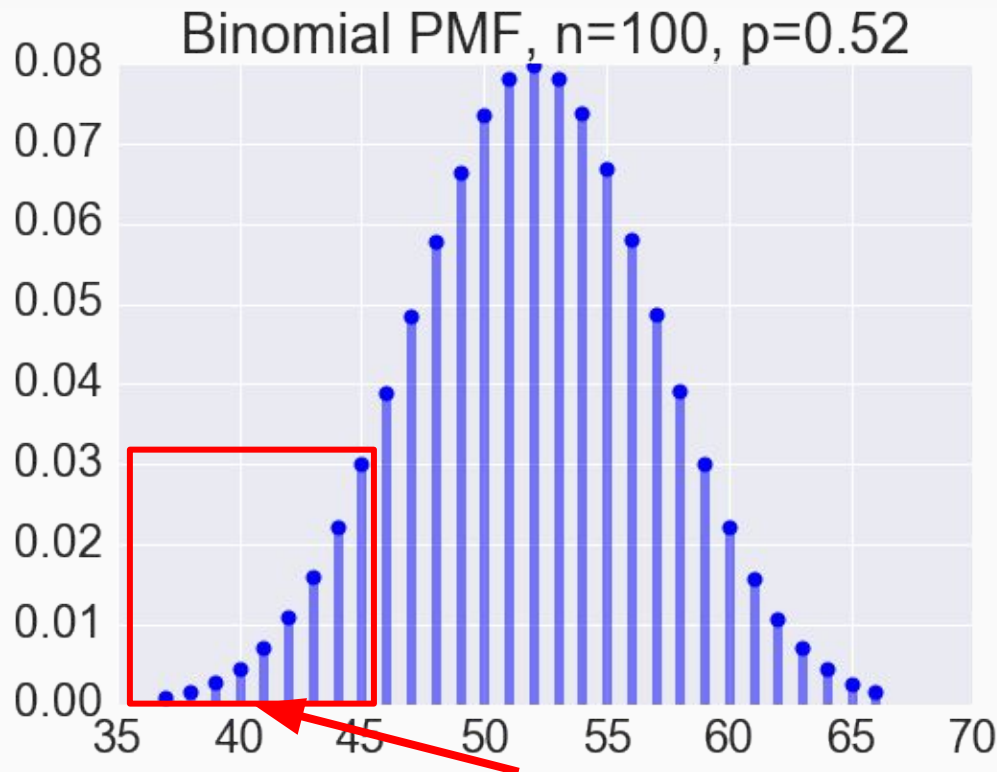
Use the sample moments to estimate the distribution parameters. In this case, the parameter we need to estimate is p .

$$52 = np$$

$$n = 100$$

$$p = 52/100$$

$$p = 0.52$$



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

```
print scipy.stats.binom.cdf(45, n, p) ⇒ 0.096653350327
```

Maximum Likelihood Estimation (MLE)



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.**

Assume our data is drawn independently from some distribution, and we'd like to estimate the parameters for that distribution. The probability distribution function for the data is:

$$f(x|\theta)$$

where:

x Is the data

θ Are the parameters for a given distribution that we are trying to estimate

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

True because we assume X is i.i.d.
Recall, what does i.i.d. Mean?

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

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

$$\hat{\theta}_{\text{mle}} = \arg \max_{\theta \in \Theta} \log (\mathcal{L}(\theta | x_1, \dots, x_n))$$

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?

Use the Poisson distribution and estimate lambda with MLE instead of MOM

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

visits: [6, 4, 7, 4, 9, 3, 5]

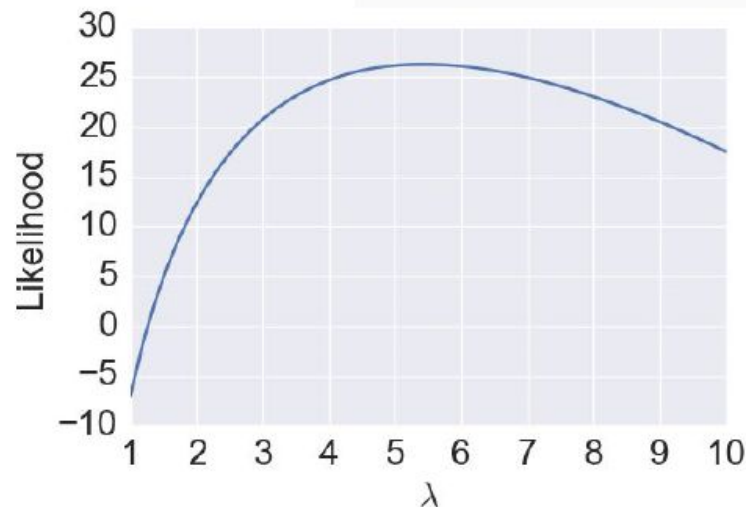
$$L(\lambda) = P(X = 6) \cdot P(X = 4) \cdot P(X = 7) \cdot P(X = 4) \cdot P(X = 9) \cdot P(X = 3) \cdot P(X = 5)$$

$$L(\lambda) = \frac{\lambda^6 e^{-\lambda}}{6!} \cdot \frac{\lambda^4 e^{-\lambda}}{4!} \cdot \frac{\lambda^7 e^{-\lambda}}{7!} \cdot \frac{\lambda^4 e^{-\lambda}}{4!} \cdot \frac{\lambda^9 e^{-\lambda}}{9!} \cdot \frac{\lambda^3 e^{-\lambda}}{3!} \cdot \frac{\lambda^5 e^{-\lambda}}{5!}$$

$$L(\lambda) = \frac{\lambda^{38} e^{-7\lambda}}{6! \cdot 4! \cdot 7! \cdot 4! \cdot 9! \cdot 3! \cdot 5!}$$

$$L(\lambda) \propto \lambda^{38} e^{-7\lambda}$$

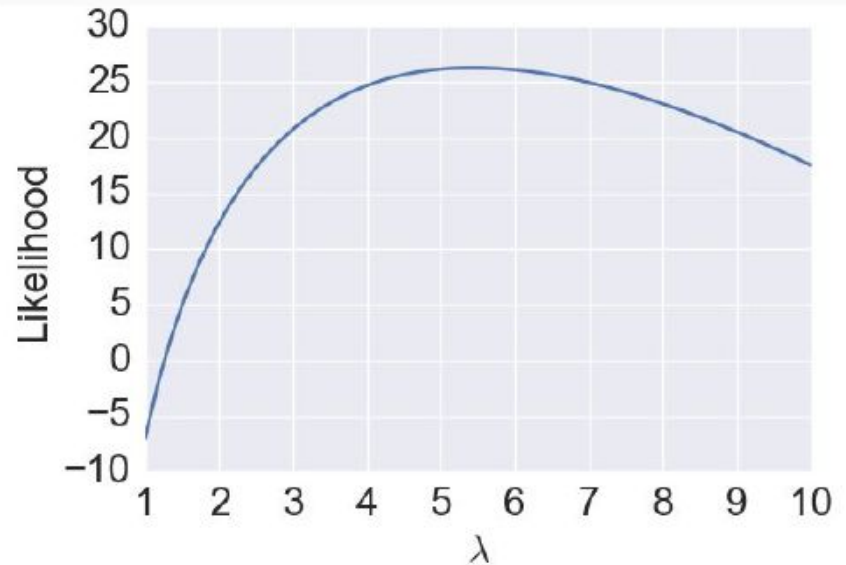
$$\ln L(\lambda) \propto 38 \cdot \ln(\lambda) - 7\lambda$$



$$\frac{d}{d\lambda}(\ln L(\lambda)) = \frac{38}{\lambda} - 7 = 0$$

$$\lambda = \frac{38}{7} = 5.429$$

$$\frac{d^2}{d\lambda^2}(\ln L(\lambda)) = -\frac{38}{\lambda^2}$$



Alternatively...

Maximize (`scs.poisson(lambda).pmf(X)`)

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?

You flip a coin 100 times. It comes up heads 52 times. What's the MLE 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...

$$X_i \stackrel{iid}{\sim} \text{Bin}(n, p) \quad i = 1, 2, \dots, n \quad f(x_i|p) = \binom{n}{x_i} p^{x_i} (1-p)^{n-x_i}$$

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

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

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



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

Maximum a Posteriori (MAP)



Maximum a Posteriori (MAP)

Recall, MLE finds θ to maximize:

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

Whereas, MAP finds θ to maximize:

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

Maximum a Posteriori (MAP)

MLE \rightarrow 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:

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

MAP solves:

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

MAP includes the prior belief.



What if all θ are equally likely?

Kernel Density Estimation (KDE)



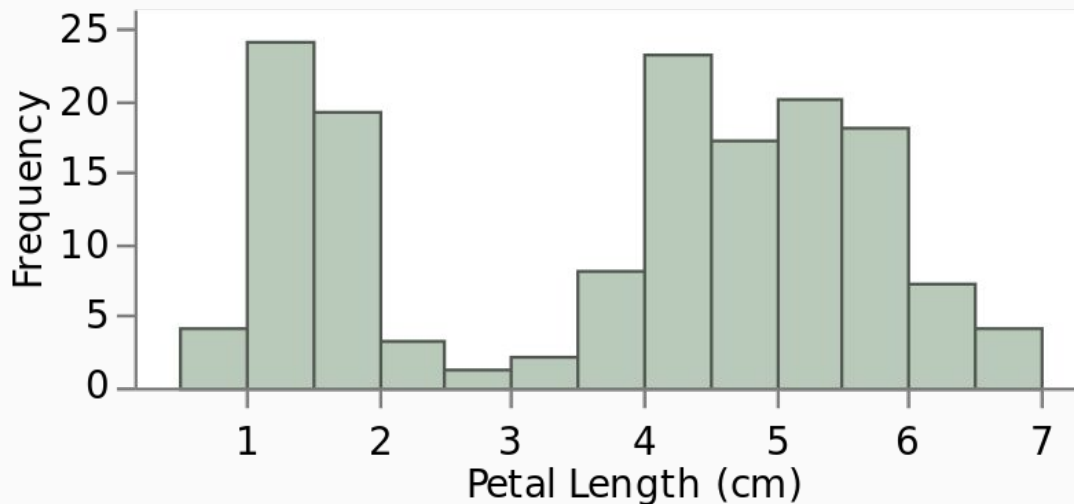
Nonparametric Techniques

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

Answer: Use a nonparametric technique.

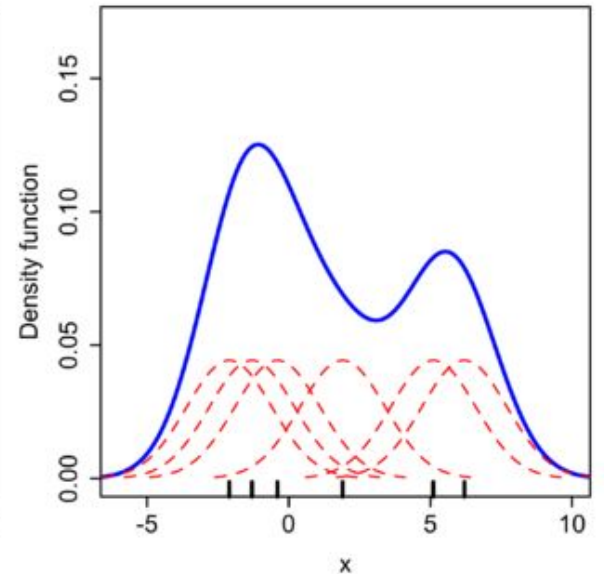
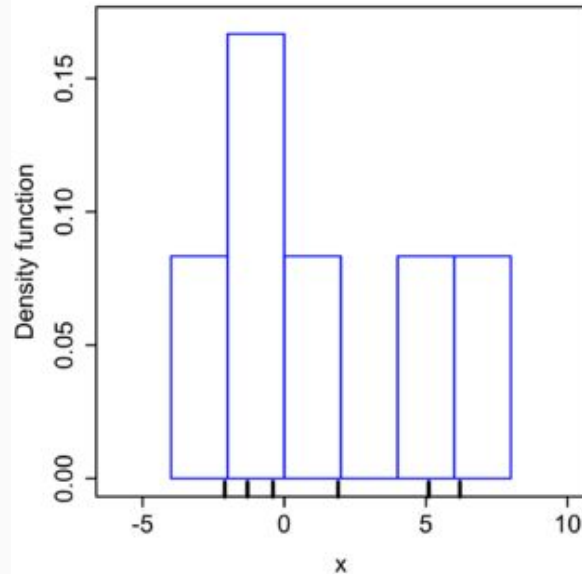
Histograms

A histogram groups continuous data into discrete intervals and displays relative frequencies. But it's not a smooth distribution. :(



Kernel Density Estimation (KDE)

KDE is a nonparametric way to estimate the PDF of a random variable. KDE smooths the histogram by summing “kernel functions” (usually Gaussians) instead of binning into rectangles.

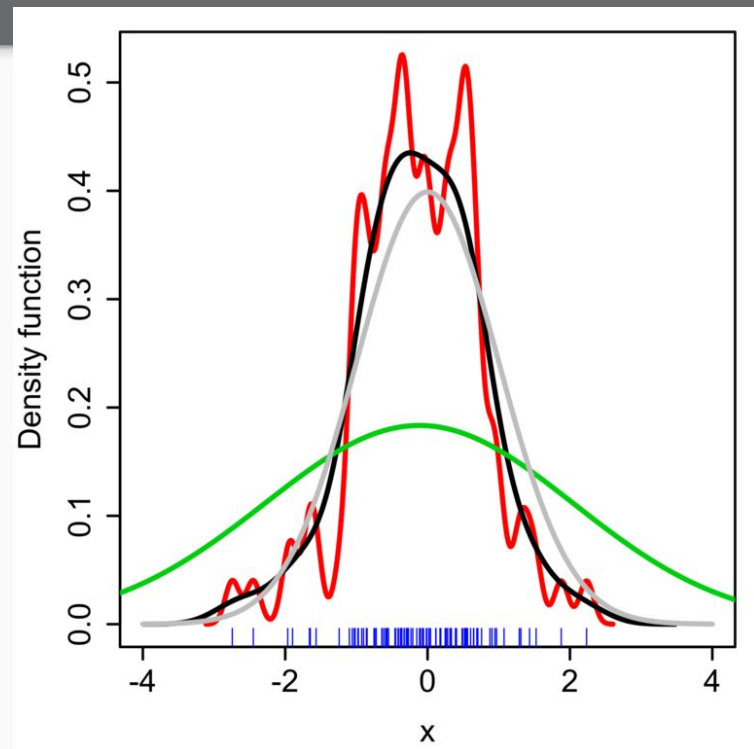


Kernel Density Estimation (KDE)

Kernel functions have a *bandwidth* parameter to control under- and over-fitting.

Each curve on the right shows an estimated PDF with different bandwidths.

See Adam's "sampling-estimation.ipynb" in the repo.



Parametric vs Nonparametric Methods



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

Ryan Henning



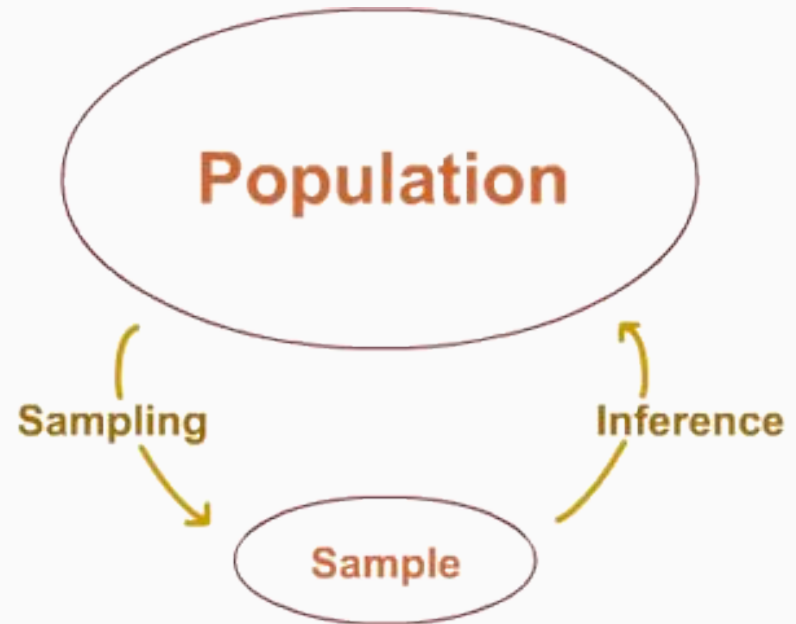
- Population Inference & Sampling
- Central Limit Theorem
- Confidence Intervals
- Bootstrapping

Population Inference & Sampling



Population Inference

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

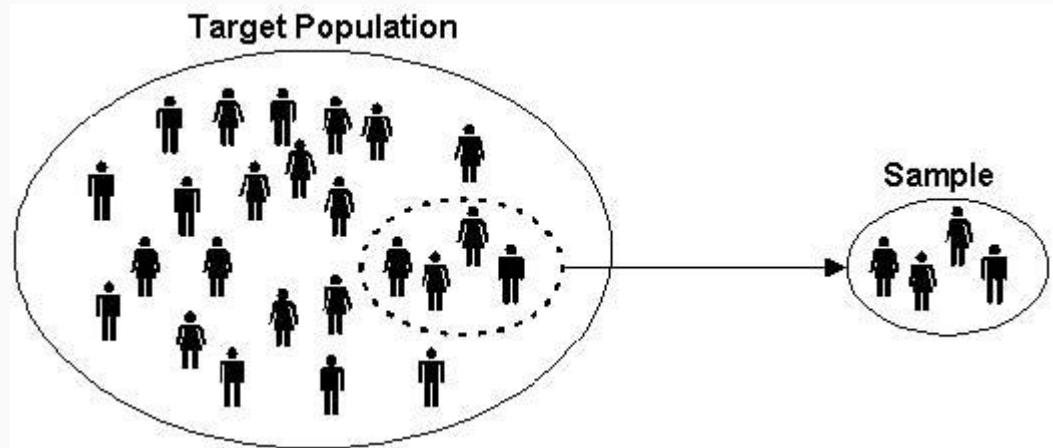


Collecting data: Taking a sample

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.**



Random sampling is surprisingly hard to do...

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

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 15th and Platte and ask **random** people if they own dogs until you have n responses.

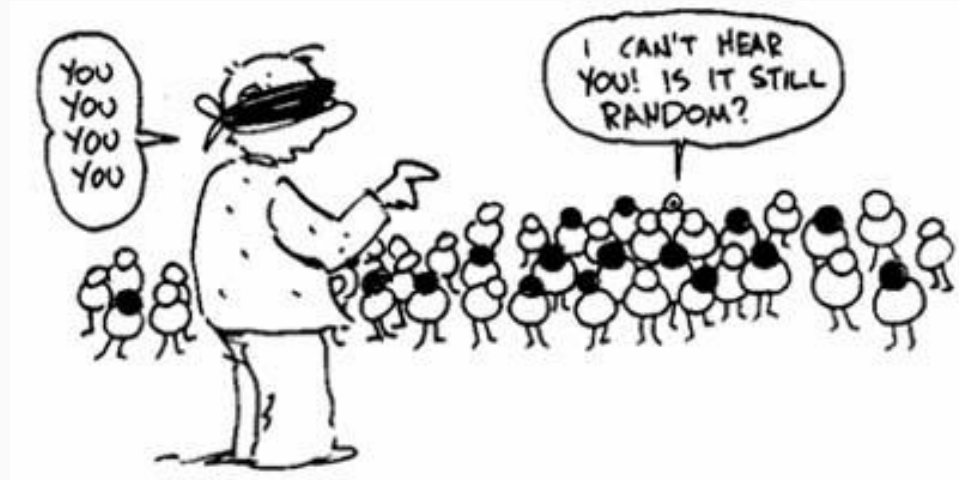
Method 3: Repeat n times: Pick a **random** neighborhood in Denver (weighted by census data per neighborhood), go to that neighborhood, ask **random** people you see if they own dogs until you get 1 response.

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 features ("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;`

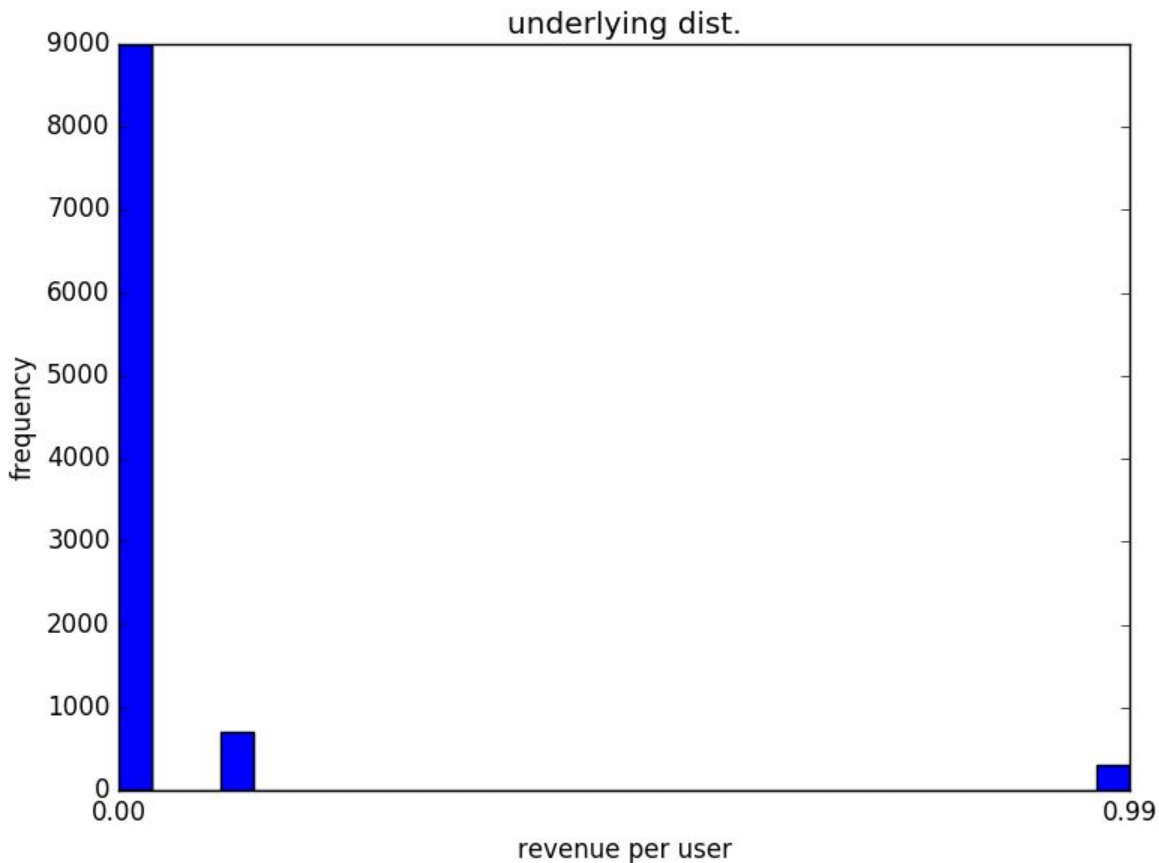
Central Limit Theorem

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

 galvanize

Underlying Distribution:

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

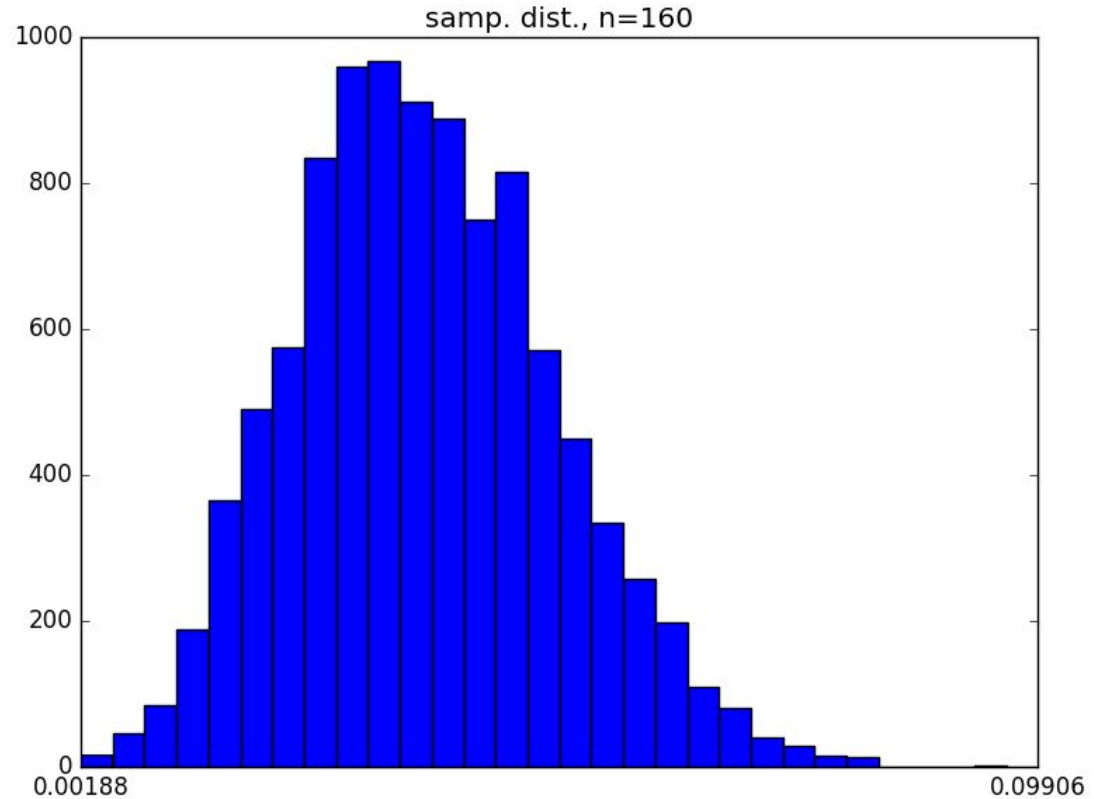


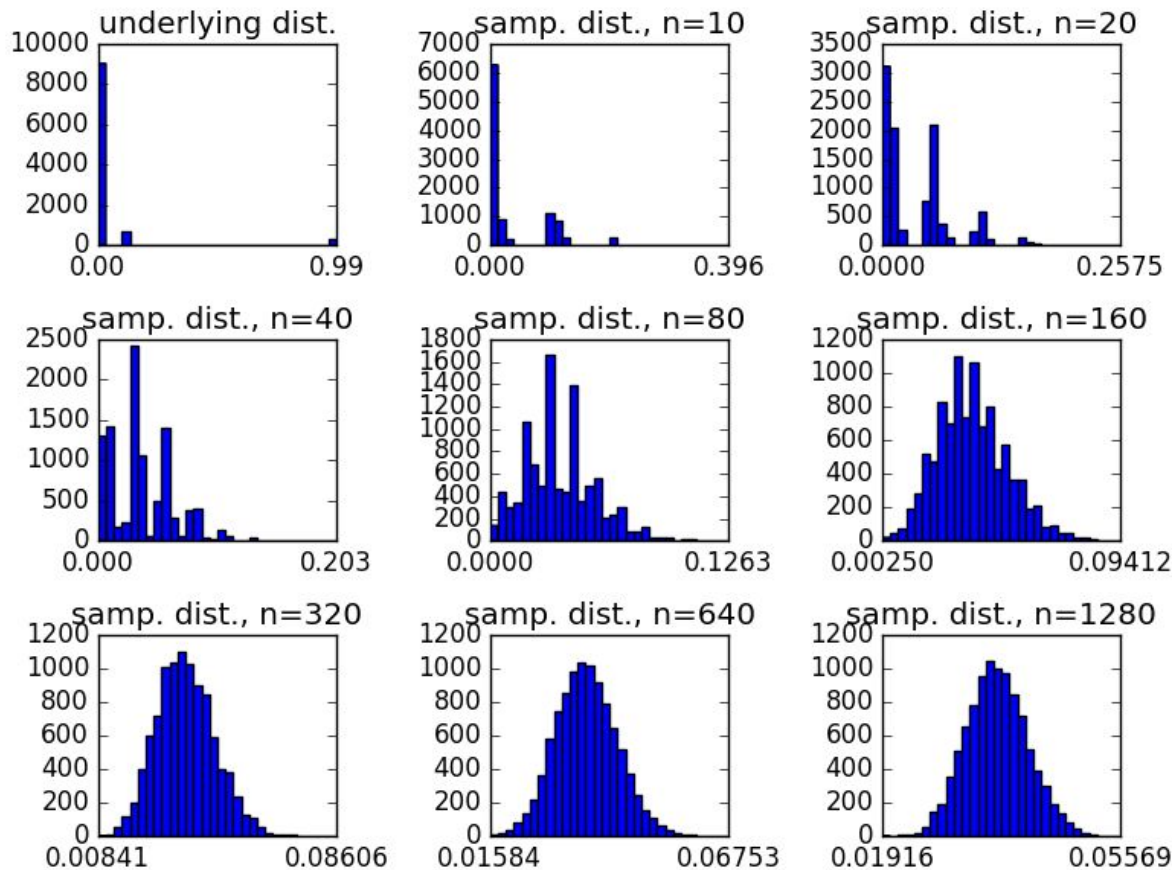
Collect n samples from the website revenue distribution, calculate the sample mean \bar{x}

Repeat 10,000 times, we get:

$\bar{x}_0, \bar{x}_1, \dots, \bar{x}_{9999}$

Plot all 10,000 sample means.



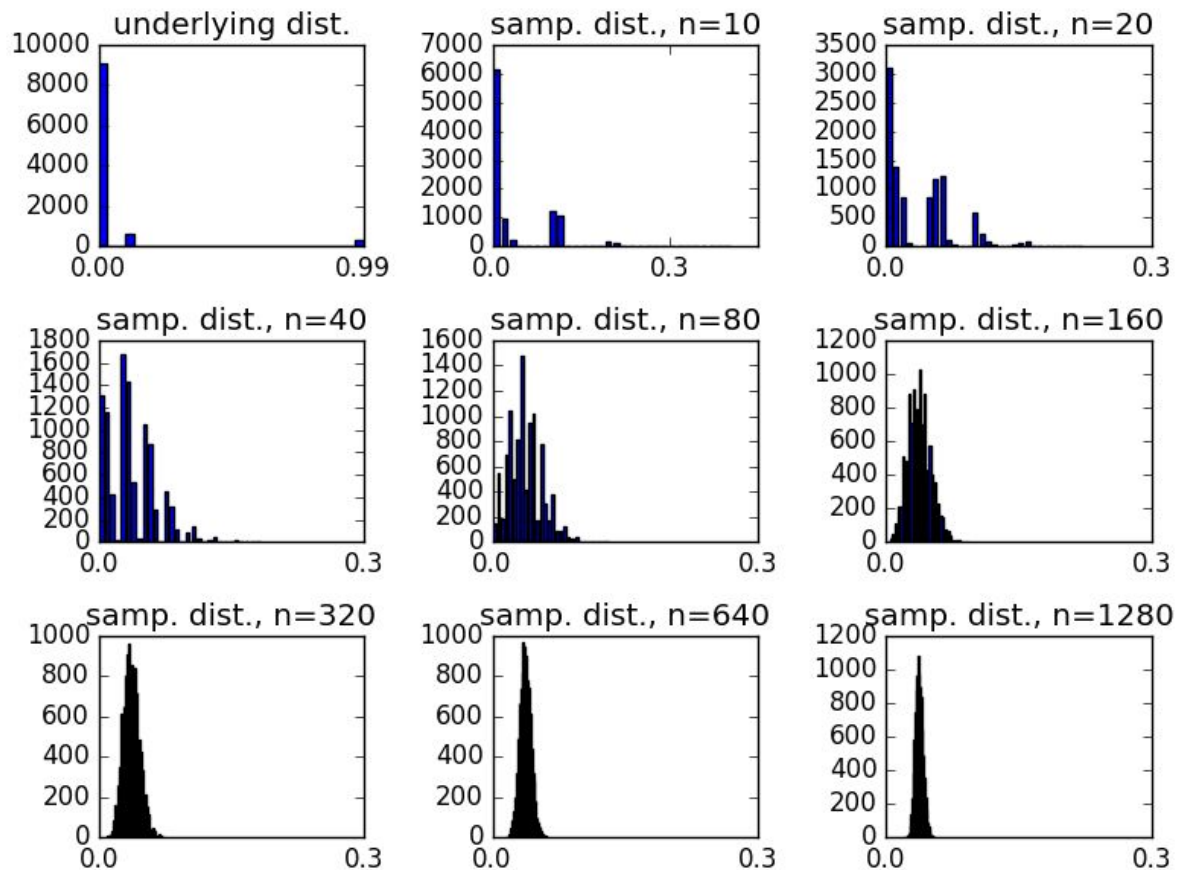


$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

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.

Central Limit Theorem: What happens when the sample size increases?



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?**

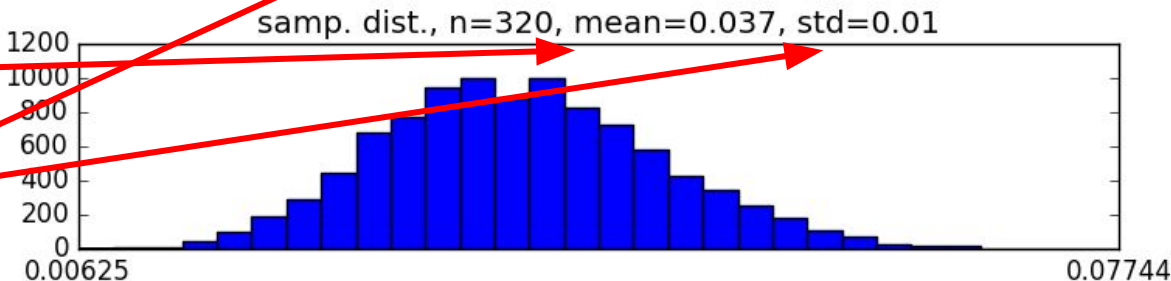
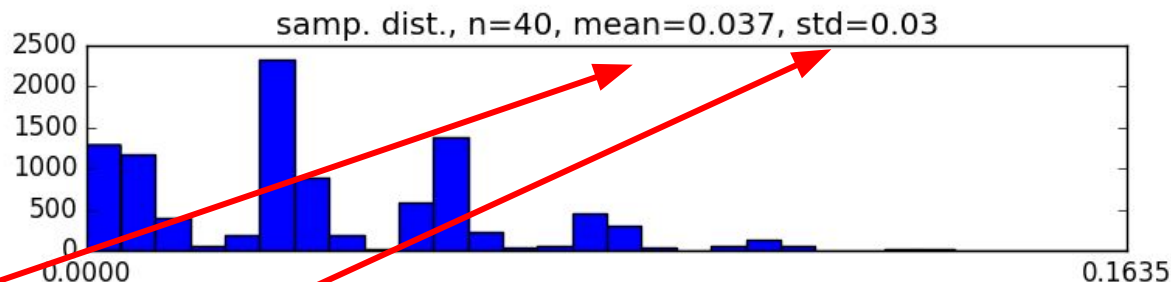
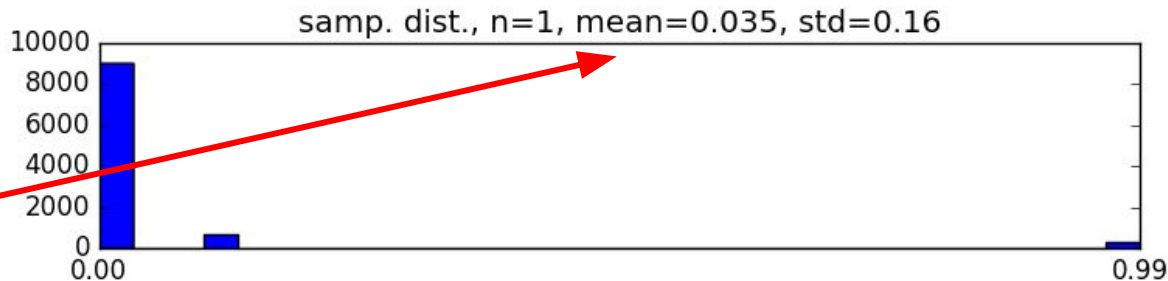
Let the underlying distribution have mean and std. dev.

μ and σ

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

$$\mu' = \mu$$

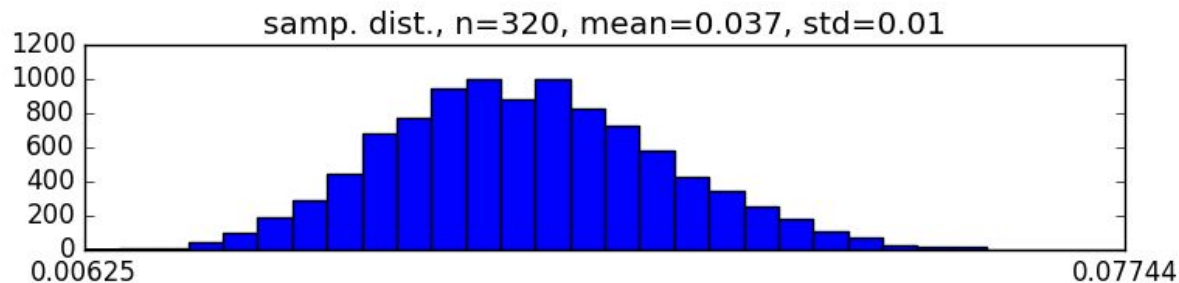
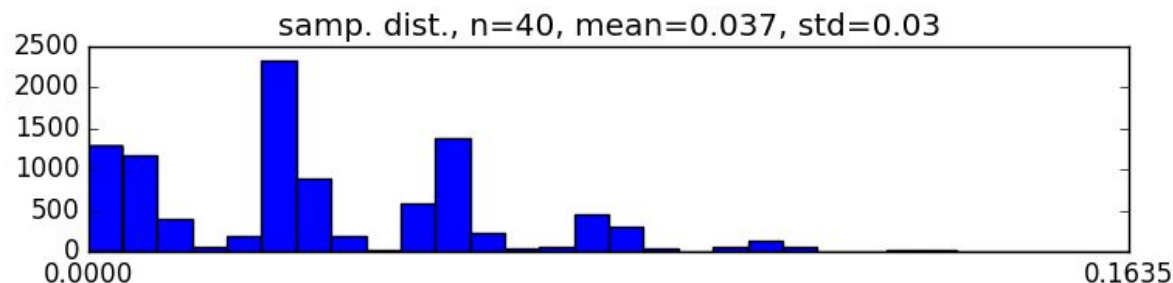
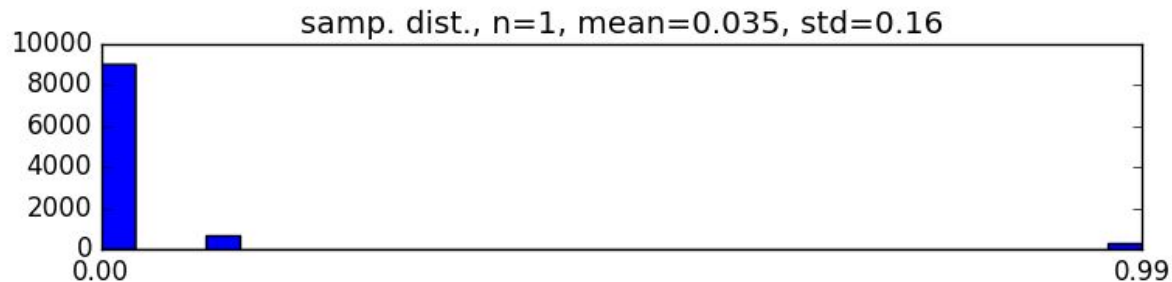
$$\sigma' = \sigma / \sqrt{n}$$



Intuitively, does the Central Limit Theorem make sense?

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



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:

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

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

 galvanize

Bootstrap Sampling

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

Recall, how is the “sampling distribution” distributed when the estimator is the sample average?



Advantages:

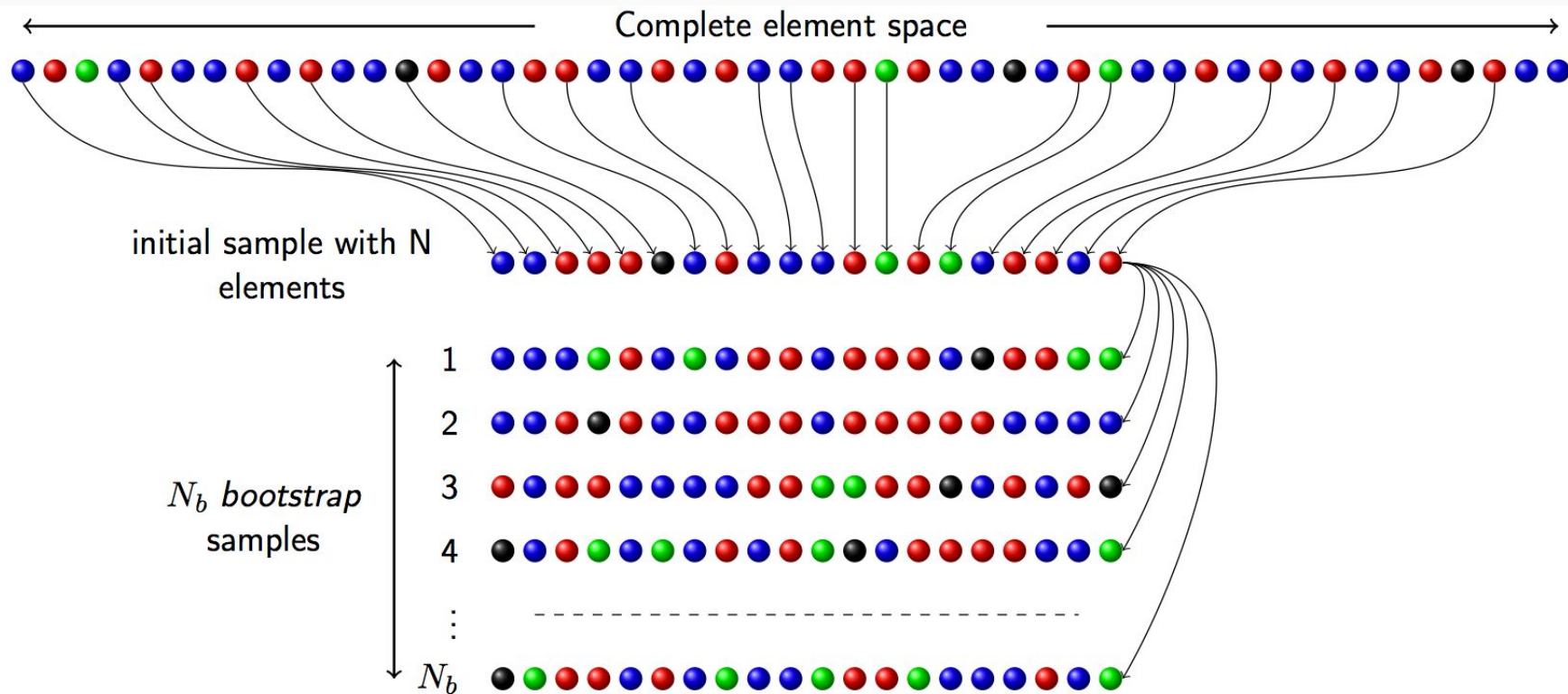
- Completely automatic
- Available regardless of how complicated the estimator may be

Often used to estimate the standard errors and confidence intervals of an unknown population parameter.

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



1. Draw a bootstrap sample:

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

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

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

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

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

4. Calculate the bootstrapped variance:

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

Bootstrap Confidence Intervals

Percentile method:

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

Interval assuming approximately *normal* bootstrap sampling distribution:

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

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.