Estimation
&
Sampling

Overview

- Review
  - Expected Value, Variance
- Statistics
  - Parametric vs. Non-Parametric
- Inference
  - MOM, MLE, MAP
  - KDE

- Sampling
  - CLT
  - Population Inference
  - Confidence Intervals
  - Bootstrapping

Review - Expectation

- Discrete: Probability weighted average of all possible values

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

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

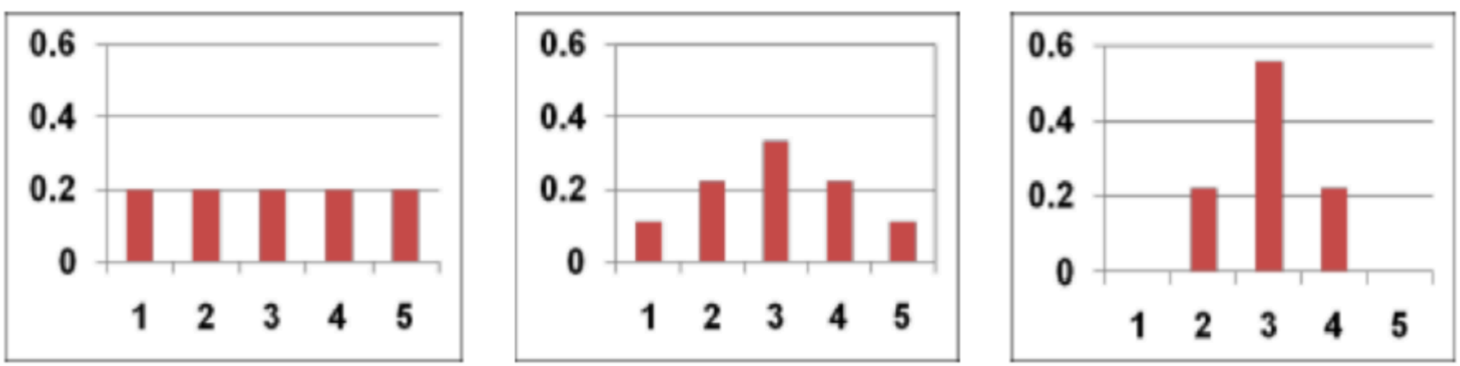
$$E(X) = \int_{-\infty}^{\infty} x * f(x) dx$$

Review - Variance

- Intuition
  - Measures how much "spread" there is in a set of numbers
  - The mean squared distance of random variable X from the mean, μ.

$$\operatorname{Var}(X) = \boxed{\operatorname{E}\left[(X-\mu)^2\right]}.$$

$$\begin{aligned}\operatorname{Var}(X) &= \operatorname{E}\left[(X-\operatorname{E}[X])^2\right] \\ &= \operatorname{E}\left[X^2 - 2X\operatorname{E}[X] + (\operatorname{E}[X])^2\right] \\ &= \operatorname{E}\left[X^2\right] - 2\operatorname{E}[X]\operatorname{E}[X] + (\operatorname{E}[X])^2 \\ &= \boxed{\operatorname{E}\left[X^2\right] - (\operatorname{E}[X])^2}\end{aligned}$$



Review - Variance

- Discrete: Probability weighted average of all possible deviations from mean, squared.

  Suppose discrete r.v. X can take on ***k*** distinct values.

$$E(X) = \mu$$

$$Var(X) = E[(X-\mu)^2] = \sum_{i=1}^{k} p_i \; * \; (x_i - \mu)^2$$

- Continuous: Same idea, except replace Σ with integral, and replace probabilities p_i with probability densities f(x)

$$Var(X) = \sigma^2 = \int (x-\mu)^2 f(x)dx$$

Modeling – Parametric vs. Nonparametric

- Parametric
  - Assumes data comes from a type of probability distribution and makes inferences about the parameters
  - For example, $Normal(\mu, \sigma^2),\ Poisson(\lambda)$
  - May make use of some common sample statistics

$$\boxed{\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i} \qquad \boxed{s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Non-parametric
  - Unlike parametric case, non-parametric statistics make no assumptions about the probability distributions from which the variables arise

Inference – MOM and MLE

Method of Moments (MOM) and Method of Maximum Likelihood (MLE) are two parameter estimation strategies.

- $Normal(\underline{\mu}, \underline{\sigma}^2),\ \ Poisson(\underline{\lambda})$
- May or may not result in the same estimate

For fixed set of data and underlying statistical model…

- MOM – Derive equations related to population moments

  What's a moment? $\underbrace{E(X)}_{\text{first moment}}, \underbrace{E(X^2)}_{\text{second moment}}, \underbrace{E(X^3)}_{\text{third moment}} \ldots$

- MLE – Set values of parameters to maximize the likelihood f(n)

Inference – MOM

MOM – Derive equations related to population moments

What's a moment? $E(X), E(X^2), E(X^3)...$

$$X_i \underset{\text{iid}}{\sim} Binomial(N, \underline{p}), \quad i = 1, 2, ..., n$$

$$\Rightarrow E(X_i) = Np$$

Assume data comes from some distribution

$$\bar{x} = Np$$

Compute first moment from **sample data**.

$$\hat{p} = \frac{\bar{x}}{N}$$

Estimate parameter *p* based on first moment

Inference – MOM

$$X_i \sim Uniform(-\theta, \theta), \quad i = 1, 2, ..., n$$
$$\Rightarrow E(X_i) = 0$$

Again, assume data comes from some distribution

E(X) does not depend on parameters, so first moment doesn't help, but…

$$Var(X_i) = \frac{\theta^2}{3}$$

Compute first and second moments from **sample data**.
Recall $Var(X) = E(X^2) - [E(X)]^2$

$$s^2 = \frac{\hat{\theta}^2}{3}$$

Estimate parameter Θ based on first moment and second moments.

Inference – MLE

MLE – Set values of parameters to maximize the likelihood f(n)

First off…what's a likelihood function?

– Since we assume x_1, x_2, … x_n are i.i.d., we have the joint density function

$$\boxed{f(x_1, x_2, ..., x_n|\theta) = f(x_1|\theta) * f(x_2|\theta) * f(x_3|\theta) * ... * f(x_n|\theta)}$$

– Just call this joint density the "Likelihood", and the log of that joint density the "Log Likelihood" just to make calculus easier)

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

$$log\mathcal{L}(\theta|x_1, ..., x_n) = \sum_{i=1}^{n} log[f(x_i|\theta)]$$ "Log Likelihood"

– Parameter estimate is simply the one that maximizes the likelihood f(n)

$$\hat{\theta}_{mle} = \underset{\theta \in \Theta}{argmax}\ log\mathcal{L}(\theta|x_1, ..., x_n)$$

Inference – MLE

MLE – Set values of parameters to maximize the likelihood f(n).

$$X_i \sim Binomial(N,p),\ \ i = 1, 2, \ldots, n$$

As with MOM, assume data comes from some distribution

$$\Rightarrow \quad f\left(x_i\right) = \binom{N}{x_i} p^{x_i} \left(1-p\right)^{N-x_i}$$

$$\Rightarrow \quad \mathcal{L}(p|x) = \prod_{i=1}^{n} \binom{N}{x_i} p^{x_i} \left(1-p\right)^{N-x_i}$$

Define Likelihood

$$\Rightarrow \quad log\mathcal{L}(p|x) = \sum_{i=1}^{n} \log \binom{N}{x_i} + x_i \log p$$

$$+ \left(N - x_i\right) \log\left(1-p\right)$$

Log Likelihood

$$\Rightarrow \quad \frac{\partial\, log\mathcal{L}(p|x)}{\partial p} = \sum_{i=1}^{n} \left[\frac{x_i}{\widehat{p}} - \frac{N - x_i}{1 - \widehat{p}}\right] = 0$$

$$\Rightarrow \quad \hat{p} = \frac{\bar{x}}{N}$$

Estimate parameter using some calculus!

Inference – MLE

Logistic Regression

- For each data point, have feature vector, x_i, and observed response y_i

$$\mathcal{L}(\beta_0, \beta | x_1, ..., x_n) = \prod_{i=1}^{n} p(x_i)^{y_i} (1 - p(x_i))^{1-y_i}$$

Can think of each observation as a Bernoulli trial where probabilities are estimated by your Logistic Model

Pick coefficients that maximize the joint likelihood!

Inference – MLE

Logistic Regression

- For each data point, have feature vector, x_i, and observed response y_i

$$\mathcal{L}(\beta_0, \beta | x_1, \dots, x_n) = \prod_{i=1}^{n} p(x_i)^{y_i}(1 - p(x_i))^{1-y_i}$$

Can think of each observation as a Bernoulli trial where probabilities are estimated by your Logistic Model

Pick coefficients that maximize the joint likelihood!

Note: no closed form solution, but no matter!
We can use some numerical method, such as Gradient Descent.

MOM vs. MLE

- MOM introduced in 1894. Some say MLE has supplanted MOM.
- Still MOM has some good qualities
  - Fairly simple
  - Useful if MLE computationally intractable
  - Can be useful as stepping stone to solving MLE
    - First approximation to solutions of likelihood equations

Inference – MAP

- Maximum a posteriori (MAP) – mode of the posterior distribution
  - For MLE, we have

$$\hat{\theta}_{mle} = \underset{\theta \in \Theta}{argmax}\ f(x|\theta) = \underset{\theta \in \Theta}{argmax}\ log\mathcal{L}(\theta|x_1, ..., x_n)$$

  - For MAP, we assume a prior *g* over Θ, and go one step further to get the posterior.

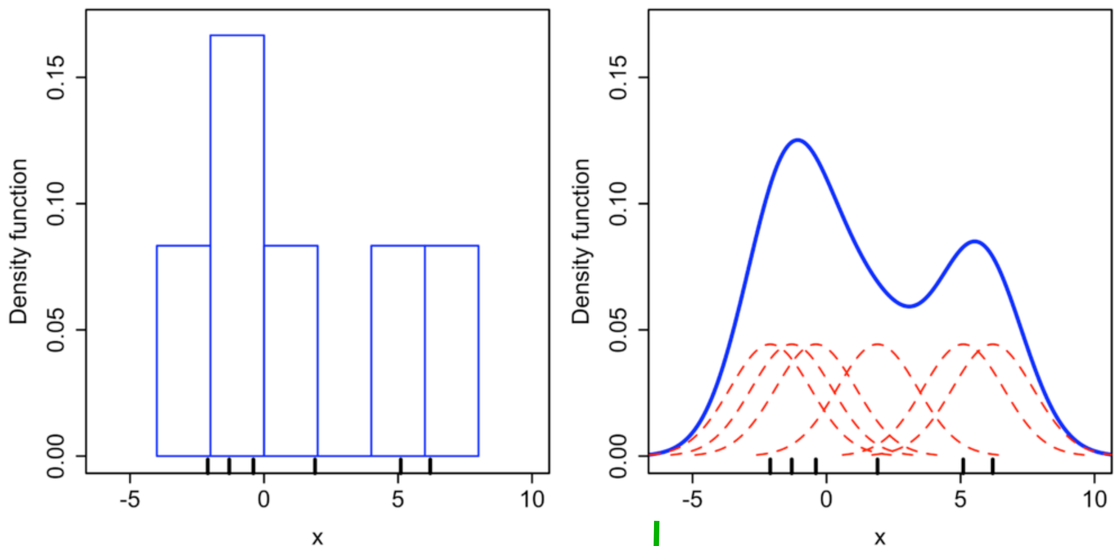
$$\theta \mapsto f(\theta|x) = \frac{f(x|\theta)\, g(\theta)}{\int_{\vartheta \in \Theta} f(x|\vartheta)\, g(\vartheta)\, d\vartheta}$$

Simply get Posterior using Bayes

$$\hat{\theta}_{map} = \underset{\theta \in \Theta}{argmax} \frac{f(x|\theta)\, g(\theta)}{\int_{\vartheta} f(x|\vartheta)\, g(\vartheta)\, d\vartheta} = \underset{\theta \in \Theta}{argmax} f(x|\theta)\, g(\theta).$$

Inference – KDE

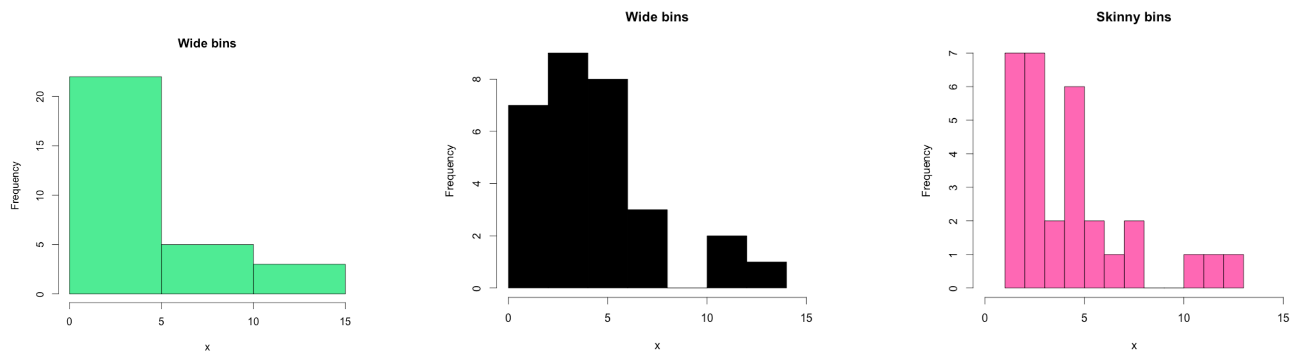
- Kernel Density Estimation (KDE)
  - Non-parametric way to estimate pdf of a random variable
  - Really, a data smoothing problem
    - Very similar to histograms
    - *Data:* $x_1 = -2.1$, $x_2 = -1.3$, $x_3 = -0.4$, $x_4 = 1.9$, $x_5 = 5.1$, $x_6 = 6.2$



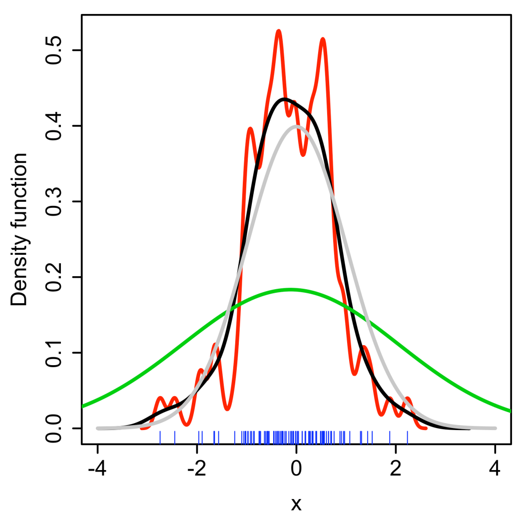
Instead of summing rectangles, can sum gaussian curves

Inference – KDE

- Kernel Density Estimation (KDE)
  - Varying Bandwidth (for histograms)



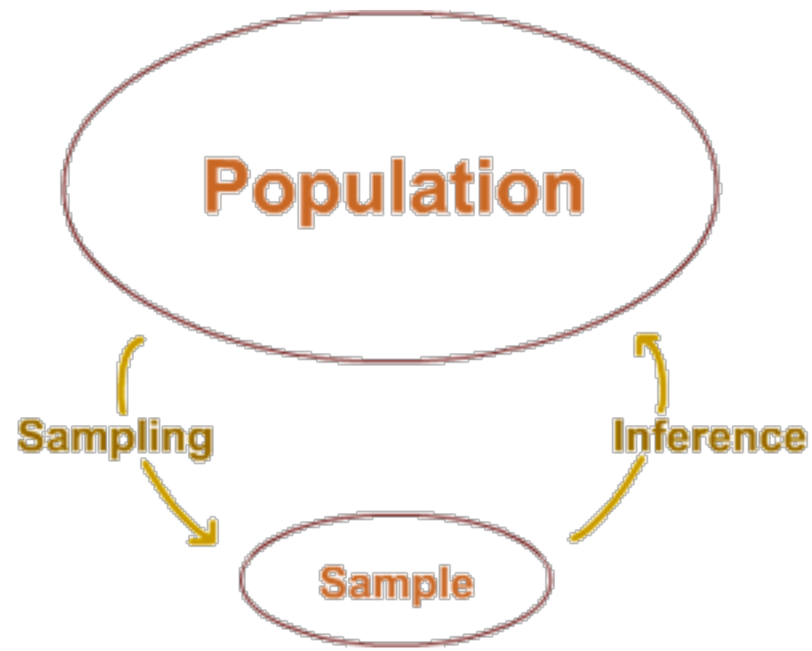
  - Instead, can use Gaussian kernels



Which bandwidth seems to be overfitting? Underfitting?

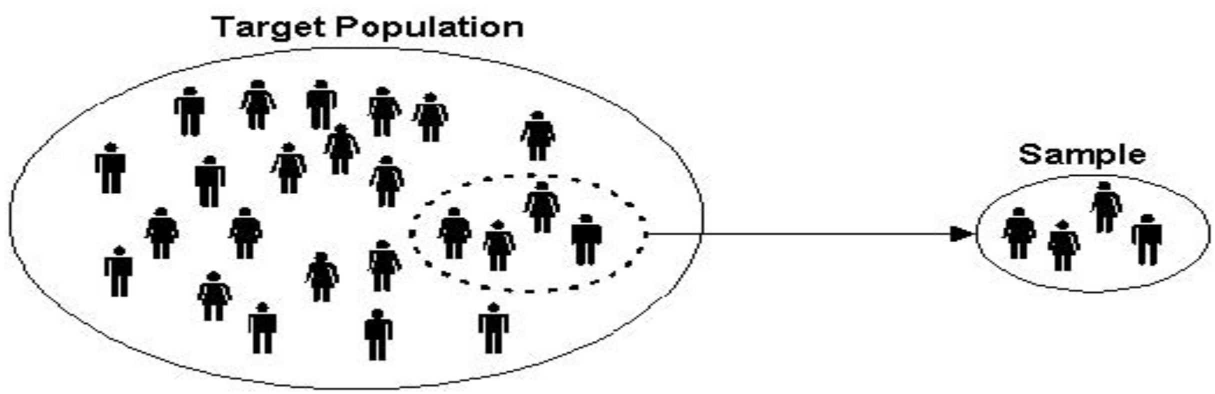
Statistical Data Discovery in General

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



Getting (Good) Data

- A sample should be representative of the population (junk in = junk out)

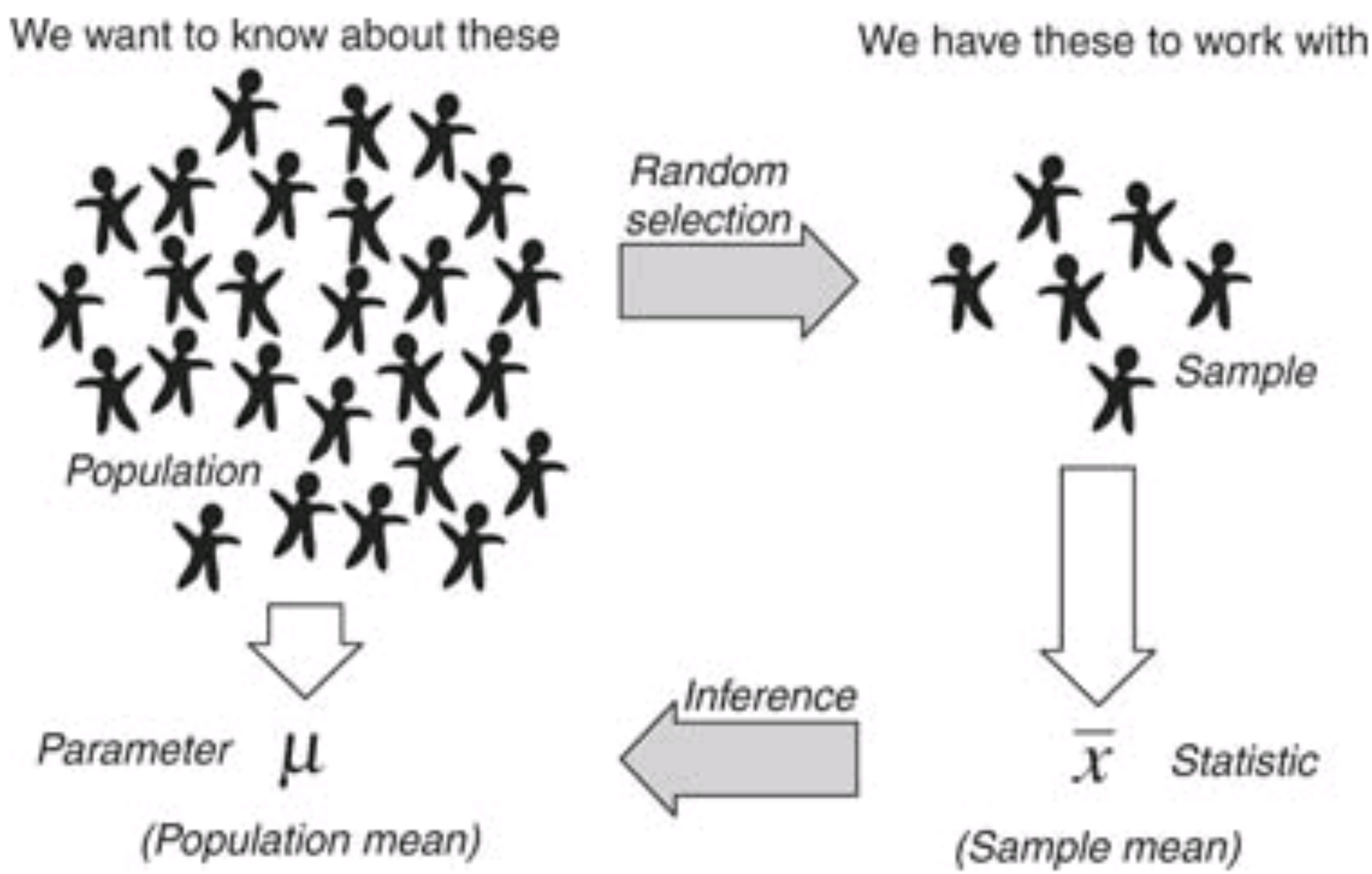


- Random sampling is often the best way to achieve this

Sampling Methods

- Simple random sampling (SRS)
  - The easiest most widespread form of sampling
  - Each subject has an equal chance to being in the sample
- Other common sampling methods:
  - Systematic sampling
  - Stratified sampling
  - Cluster sampling

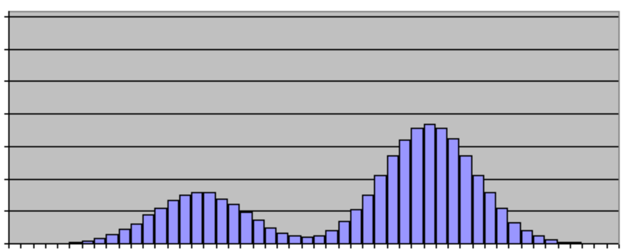
Sampling and Inference



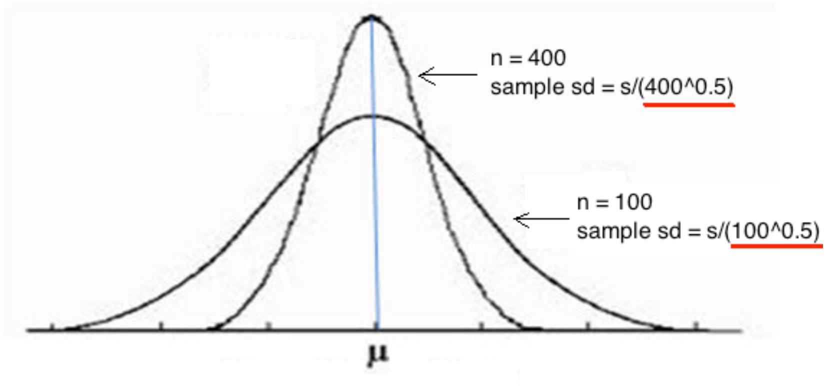
Central Limit Theorem

- Given certain conditions, the **mean** of a sufficiently large number of i.i.d. random variables, will be approximately normal, ***regardless*** of the underlying distribution.

Underlying distribution



draw i.i.d. samples
and average them



Central Limit Theorem

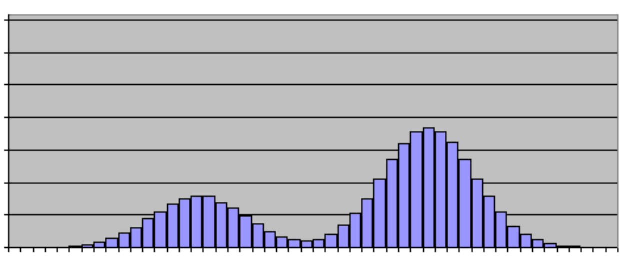
- Not only is the sample mean normally distributed, we have....

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

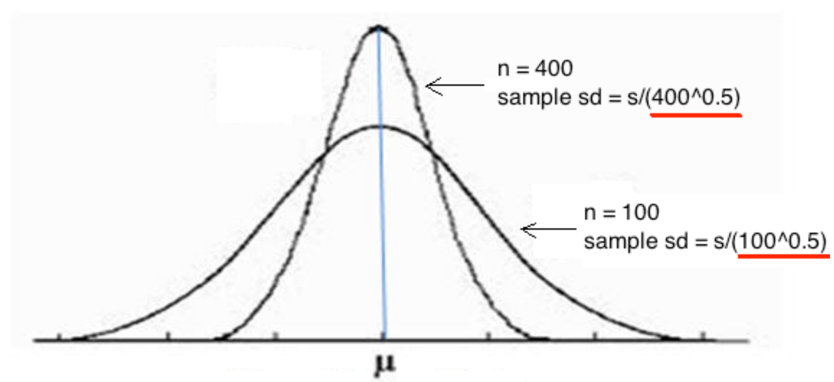
- And as usual, from any normally distributed random variable, we can derive a standard normal variable. In this case...

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

Underlying distribution



draw i.i.d. samples and average them



Confidence Interval

- A confidence interval (CI) is an interval estimate of a population parameter
- They are typically stated at 95% confidence level, but they can be shown at any confidence level, e.g. 50%, 90%, 99%
- The confidence interval for the mean is given by

$$(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}) \quad \text{or} \quad \bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

Confidence Interval - cont

- Since we do not know σ, if N > 30, we can substitute s for it

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

- When N is small, we would use

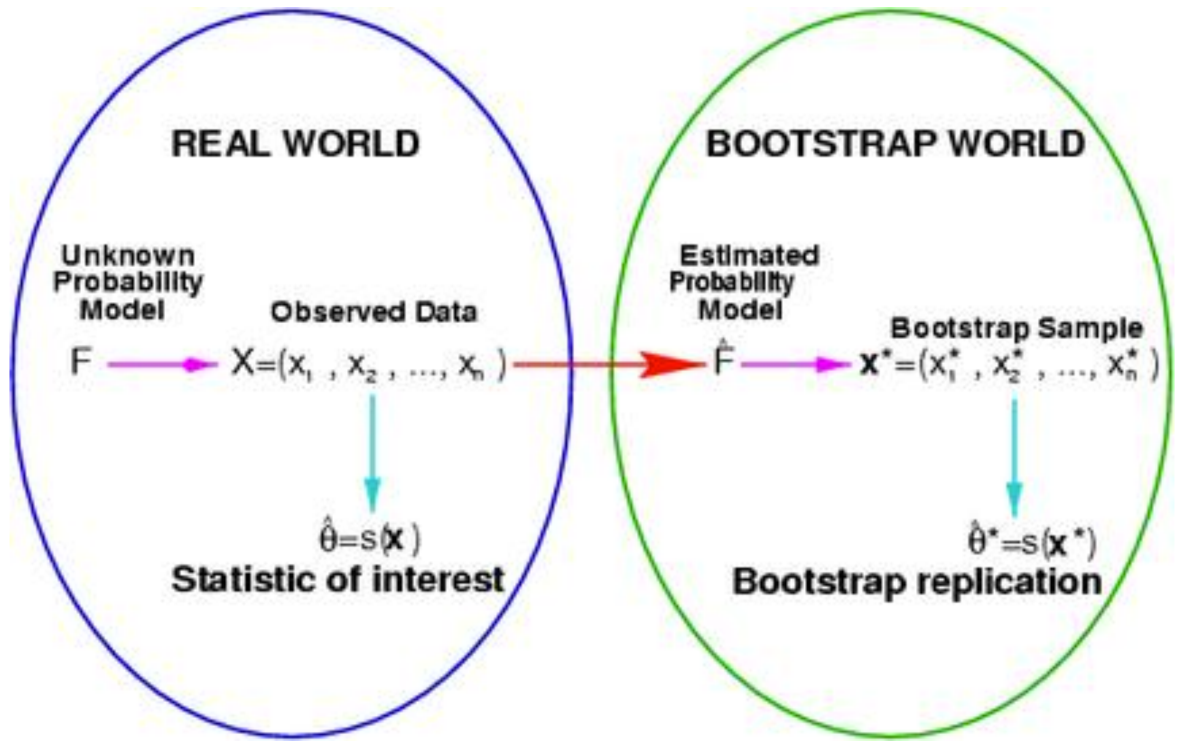
$$\bar{x} \pm t_{(\alpha/2, n-1)} \frac{s}{\sqrt{n}}$$

Resampling

- Resampling: drawing repeated samples from the given data
- Common resampling techniques:
  - Bootstrapping
  - Jackknifing
  - Cross-validation
  - Permutation tests

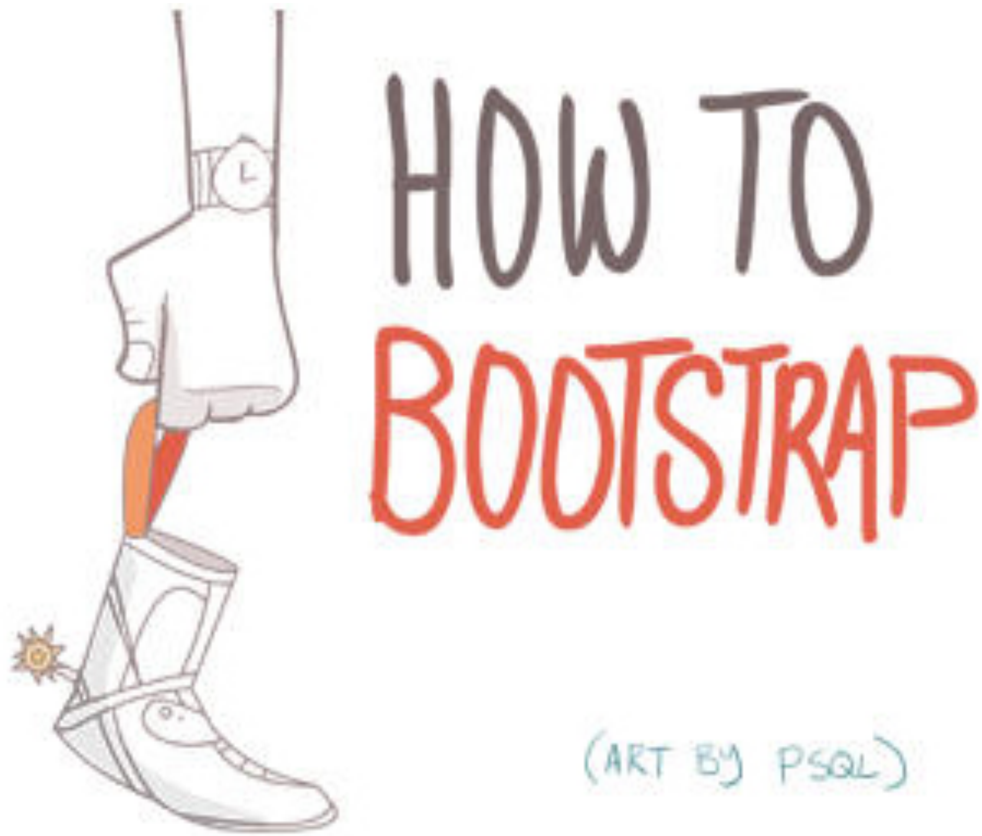
Bootstrapping

- Estimates the sampling distribution of an estimator by sampling with replacement from the original sample
- Often used to estimate the standard errors and confidence intervals of a population parameter

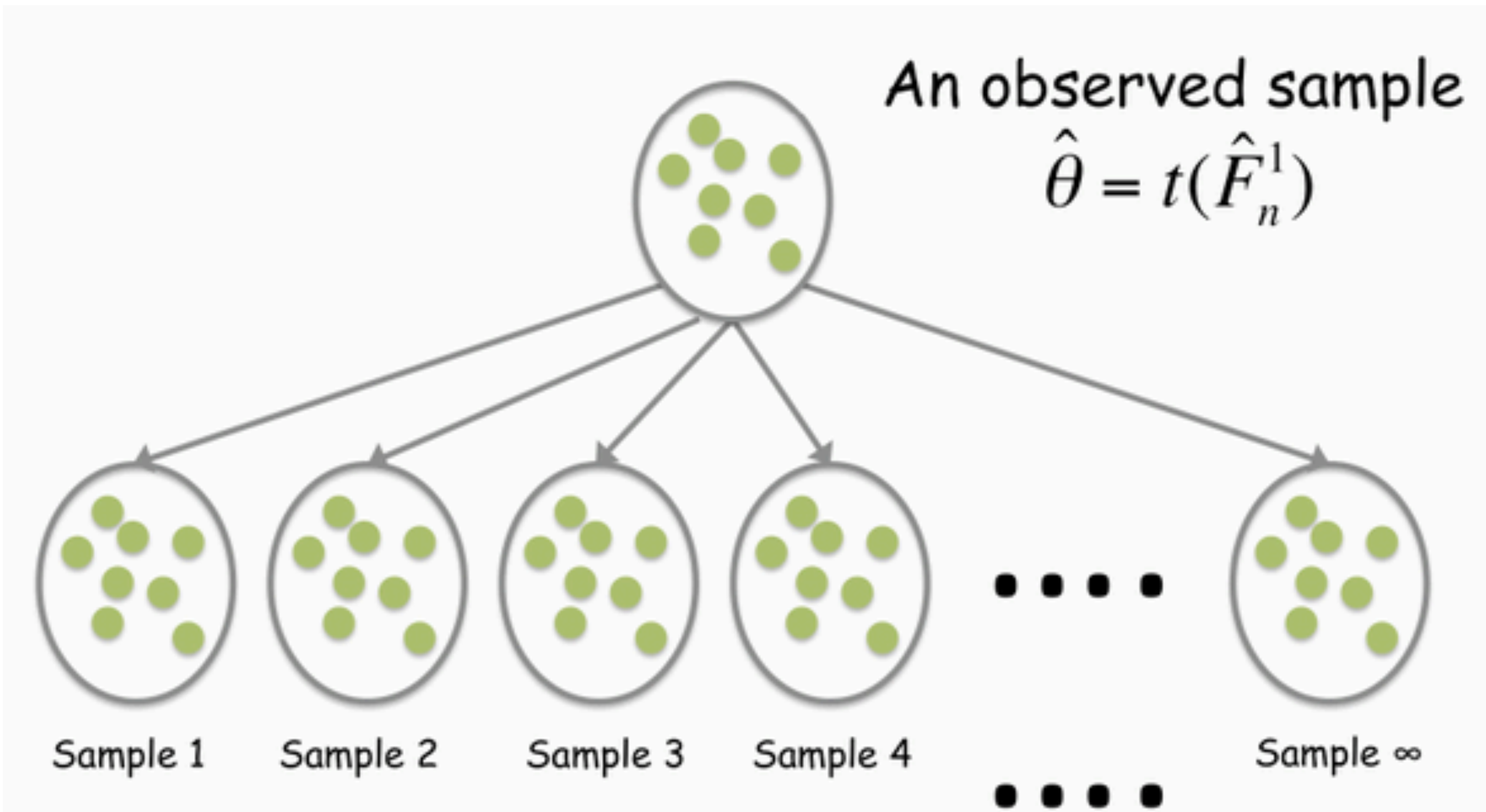


How To Bootstrap

To pull oneself up by one's bootstrap..



Bootstrapping



Bootstrap Variance Estimation

1. Draw $X_1^*, \ldots, X_n^* \sim \hat{F}_n$

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

3. Repeat steps 1 and 2, B times, to get $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$

4. Let $v_{boot} = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}_b^* - \frac{1}{B}\sum_{r=1}^{B}\hat{\theta}_r^*)$

$$(\hat{se}_{boot} = \sqrt{v_{boot}})$$

Bootstrap Confidence Intervals

- Percentile method

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

- The Normal interval

$$\hat{\theta} \pm z_{\alpha/2} \hat{se}_{boot}$$

When Do We Use Bootstrapping?

- When the theoretical distribution of the statistic is complicated or unknown
- When the sample size is too small
- When estimating the variance of a statistic using a small pilot sample for power calculations

Questions

- MOM vs. MLE
  - What do they solve for?
  - How does each approach tackle the problem?

- How about MAP?
  - How does it relate to the MLE?

- What's bootstrapping?
  - When might I think of using it?
  - What are the steps to setting up a bootstrap estimate?

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.