Models, Estimators, Maximum-Likelihood Estimator

Tuesday 22nd November, 2016

Scientific Models

Scientific models are frequently used to

- Better understand a phenomenon
- Enable prediction
- Classify phenomena
- Simulate data
- Visualize data

Scientific Models

Examples

- seasons of the year
- classical mechanics
- supply and demand

Statistical Models

A specific type of statistical model

- A collection of assumptions regarding the data

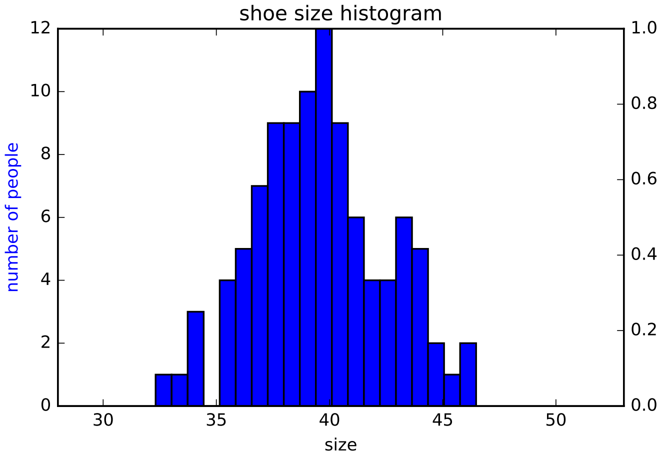
Statistical Models

A shoemaker

- Wants to make shoes that will fit 95% of the population.
- Measures the shoe sizes of 100 random people
- Assumes: shoe-size $\sim \mathcal{N}(\mu, \sigma^2)$

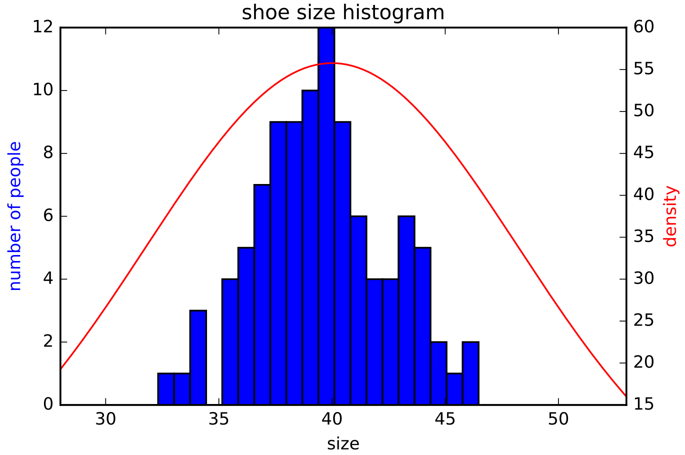
Statistical Models

A shoemaker



Statistical Models

A shoemaker



Models Vs. Reality

- don't have to be true in an objective sense
- a simplification of reality

Models Vs. Reality

- shoemaker example
  - model - sizes are distributed normally
  - they're not
  - shoe-size $\in [a, b]$
  - samples from a normal distribution $\in (-\infty, \infty)$

Parametric Model

A family of distributions, described by a finite number of parameters

Example: shoemaker

- shoe sizes are normally distributed
- a parametric model with two parameters μ, σ

$$\text{shoe-size} \sim M$$

$$M = M(\mu, \sigma)$$

$$M \in \left\{ N\left(\mu, \sigma^2\right) \right\}_{\substack{\mu \in \mathcal{R} \\ \sigma \geq 0}}$$

Parametric Model

A family of distributions, described by a finite number of parameters

Example: rain on a cloudy day

- given today is cloudy it will rain with probability p
- a parametric model (Bernoulli) with one parameter p

$$\text{rain} \sim M$$

$$M = M(p)$$

$$M \in \{M(p)\}_{0 \le p \le 1}$$

I'm not a parameter I'm a free man

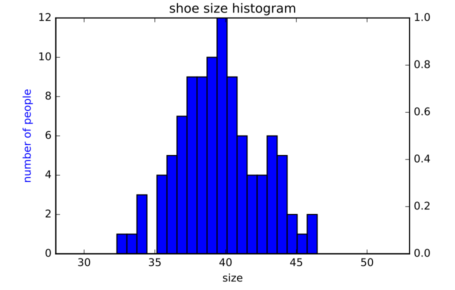
for the sake of completeness

I'm not a parameter I'm a free man

- Not all models are parametric

I'm not a parameter I'm a free man

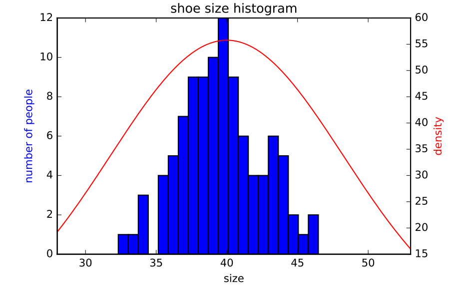
- Not all models are parametric
- Some model are distribution free



- A histogram
- Pros
  fewer assumptions
- Cons
  fewer assumptions

I'm not a parameter I'm a free man

- Not all models are parametric
- Some model are distribution free



- A histogram
- Pros
  fewer assumptions
- Cons
  fewer assumptions

I'm not a parameter I'm a free man

- Not all models are parametric
- Some models do not assume a fixed structure
    - In a nutshell
        - Start with a simple model
        - Increase complexity of the model with the size of the data

I'm not a parameter I'm a free man

IsraelTech challenge

- Not all models are parametric
- Some models do not assume a fixed structure
  - In a nutshell
    - Start with a simple model
    - Increase complexity of the model with the size of the data
  - Shoemaker example
    - Given 100 samples - model is a normal distribution
    - Given 300 samples - one normal distribution for women and one for men

Parameter Estimation

A parametric model

finding the correct model for the data	=	estimating correct parameter values

- Example - rain on a cloudy day
  - model - $p = Pr\left(rain \mid cloudy\right)$
  - estimate p

Intuition

- N cloudy days
- N_{rain} cloudy days with rain
- $N_{no-rain}$ cloudy days with $\boxed{no}$ rain
- How would you estimate p

Likelihood

Likelihood function

How likely (probable) is the data, given the model?

- likelihood

$$\mathcal{L}(\theta) = \prod_{x} Pr(x \mid \theta)$$

- log likelihood

$$l(\theta) = \log \mathcal{L}(\theta)$$

The Maximum Likelihood Principle

The "best" model for the data - makes the data "most likely"

$$\theta_{ml} = \arg\max_{\theta} \; Pr\left(X \mid \theta\right)$$

The Maximum Likelihood Principle

The "best" model for the data - makes the data "most likely"

The Maximum Likelihood Principle

The "best" model for the data - makes the data "most likely"

- Why not the most likely parameter value given the data?

$$\hat{\theta} = \arg\max_{\theta} \; Pr\left(\theta \mid X\right)$$

The Maximum Likelihood Principle

The "best" model for the data - makes the data "most likely"

- Why not the most likely parameter value given the data?
- our model assigns probability to data given the parameters and not vice versa

The Maximum Likelihood Principle

The "best" model for the data - makes the data "most likely"

- Why not the most likely parameter value given the data?
- our model assigns probability to data given the parameters and not vice versa
- what's the probability, a random Fellow knows French?

  is a parameter a random variable?

Maximum Likelihood Estimator

$$
\begin{aligned}
\theta_{mle} &= \arg\max_{\theta} \; \mathcal{L}(\theta) \\
&= \arg\max_{\theta} \; l(\theta)
\end{aligned}
$$

Maximum Likelihood Estimator

$$\begin{aligned}\theta_{mle} &= \arg\max_{\theta}\ \mathcal{L}(\theta) \\ &= \arg\max_{\theta}\ l(\theta)\end{aligned}$$

- When is this easy?

Maximum Likelihood Estimator

Israeltech challenge

$$\begin{aligned}\theta_{mle} &= \arg\max_{\theta} \ \mathcal{L}(\theta) \\ &= \arg\max_{\theta} \ l(\theta)\end{aligned}$$

- When is this easy?
  - when solving $\nabla\mathcal{L}(\theta) = 0$ is easy and the likelihood is $\boxed{concave}$
  - we can compute the likelihood for all values of θ

Maximum Likelihood Estimator

$$\theta_{mle} = \arg\max_{\theta} \ \mathcal{L}(\theta)$$
$$= \arg\max_{\theta} \ l(\theta)$$

- When is this easy?
  - when solving $\nabla\mathcal{L}(\theta) = 0$ is easy and the likelihood is $\boxed{concave}$
  - we can compute the likelihood for all values of θ
- When do you derive the estimator yourself?

Maximum Likelihood Estimators

Bernoulli

- IID binary random variables...

Maximum Likelihood Estimators

Bernoulli

- IID binary random variables...
- coin tosses

Maximum Likelihood Estimators

Bernoulli

- IID binary random variables...
- coin tosses

$$X \sim B\left(p\right)$$

$$P_p\left(X = 1\right) = p$$

Maximum Likelihood Estimators

Bernoulli

- IID binary random variables...
- coin tosses

$$X \sim B\left(p\right)$$

$$P_p\left(X = x\right) = p^x\left(1 - p\right)^{1-x}$$

Maximum Likelihood Estimators

Bernoulli

- IID binary random variables...
- coin tosses

$$X \sim B\left(p\right)$$

$$P_p\left(X_1 = x_1, \ldots, X_n = x_n\right) = \prod_{i=1}^{n} p^{x_i} \left(1-p\right)^{1-x_i}$$

Maximum Likelihood Estimators

$$\begin{aligned}\mathcal{L}(p) &= P_p(X_1 = x_1, \ldots, X_n = x_n) \\ &= \prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i}\end{aligned}$$

Maximum Likelihood Estimators

Israeltech challenge

$$\begin{aligned}\mathcal{L}(p) &= P_p(X_1 = x_1, \ldots, X_n = x_n) \\ &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \\ &= p^{\sum_{i=1} x_i}(1-p)^{\sum_{i=1}^{n} 1-x_i}\end{aligned}$$

Maximum Likelihood Estimators

IsraelTech challenge

$$
\begin{aligned}
\mathcal{L}(p) &= P_p\left(X_1 = x_1, \ldots, X_n = x_n\right) \\
&= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \\
&= p^{\sum_{i=1} x_i}(1-p)^{\sum_{i=1}^{n} 1-x_i} \\
\left[S \triangleq \textstyle\sum_{i=1}^{n} x_i\right] &= p^S (1-p)^{n-S}
\end{aligned}
$$

Maximum Likelihood Estimators

Israeltech challenge

$$\mathcal{L}(p) = p^S (1-p)^{n-S}$$

$$\begin{aligned} l(p) &= \log \mathcal{L}(p) \\ &= S \log(p) + (n-S) \log(1-p) \end{aligned}$$

Maximum Likelihood Estimators

Israeltech challenge

$$l\left(p\right) = S\log\left(p\right) + \left(n - S\right)\log\left(1 - p\right)$$

$$\begin{aligned} l'\left(p\right) &= \frac{S}{p} - \frac{n - S}{1 - p} \\ &\stackrel{!}{=} 0 \\ &\Downarrow \\ p_{mle} &= \frac{S}{n} \end{aligned}$$

Maximum Likelihood Estimators

Normal distribution

$$\mu_{mle} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\sigma^2_{mle} = \frac{1}{n}\sum_{i=1}^{n} \left(x_i - \mu_{mle}\right)^2$$

Estimators

A parameter may be a constant, but an estimator?

Estimators

An estimator is a random variable, a function of the sample

Estimators

An estimator is a random variable, a function of the sample

The shoemaker empire !

- each shoemaker independently samples 2 people and estimates shoe size mean and variance...

Estimators

IsraelTech challenge

An estimator is a random variable, a function of the sample

The shoemaker empire !

- each shoemaker independently samples 2 people and estimates shoe size mean and variance...
- each shoemaker independently samples 2000 people in Eilat and estimates shoe size mean and variance...

Estimators

An estimator is a random variable, a function of the sample

The shoemaker empire !

- each shoemaker independently samples 2 people and estimates shoe size mean and variance...
- each shoemaker independently samples 2000 people in Eilat and estimates shoe size mean and variance...
- each shoemaker independently samples 200 people and estimates shoe size mean and variance...

Estimators - Bias

We use θ_{mle} to estimate θ
how are we doing?

Estimators - Bias

We use θ_{mle} to estimate θ
how are we doing?

Bias - the mean deviation from the true value

$$bias\left(\hat{\theta}\right) = E\left[\hat{\theta}\right] - \theta$$

Estimators - Bias

Bernoulli

$$
\begin{aligned}
bias\left(p_{mle}\right) &= E\left[\frac{S}{n}\right] - p \\
&= E\left[\frac{\sum_{i=1}^{n} x_i}{n}\right] - p \\
&= \frac{1}{n}\sum_{i=1}^{n} E\left[x_i\right] - p \\
&= \frac{1}{n} n \cdot p - p \\
&= 0
\end{aligned}
$$

Estimators - Bias

Bernoulli

$$
\begin{aligned}
bias\left(p_{mle}\right) &= E\left[\frac{S}{n}\right] - p \\
&= E\left[\frac{\sum_{i=1}^{n} x_i}{n}\right] - p \\
&= \frac{1}{n}\sum_{i=1}^{n} E\left[x_i\right] - p \\
&= \frac{1}{n} n \cdot p - p \\
&= 0
\end{aligned}
$$

Unbiased

Estimators - Standard Error

We use θ_{mle} to estimate θ
how are we doing?

Estimators - Standard Error

We use θ_{mle} to estimate θ
how are we doing?

Standard Error - standard deviation of the estimator

$$se\left(\hat{\theta}\right) = \sigma_{\hat{\theta}}$$

abbreviated se

- often unknown
- frequently estimated - $\hat{se}$

Estimators - Standard Error

Bernoulli

$$
\begin{aligned}
se\left(p_{mle}\right) &= \sqrt{V\left[p_{mle}\right]} \\
&= \sqrt{V\left[\frac{\sum_{i=1}^{n} x_i}{n}\right]} \\
&= \frac{1}{n}\sqrt{\sum_{i=1}^{n} V\left[x_i\right]} \\
&= \frac{1}{n}\sqrt{n \cdot \sigma_X^2}
\end{aligned}
$$

Estimators - Standard Error

IsraelTech challenge

Bernoulli

$$se\left(p_{mle}\right) = \sqrt{\frac{\sigma_X^2}{n}}$$

$$= \sqrt{\frac{p\left(1-p\right)}{n}}$$

$$\hat{se}\left(p_{mle}\right) = \sqrt{\frac{p_{mle}\left(1-p_{mle}\right)}{n}}$$

Maximum Likelihood Estimators - properties

- The MLE is consistent, $\theta_{mle} \to \theta$, as $n \to \infty$
- The MLE is asymptotically normal - the distribution of the estimator is approximately $\mathcal{N}\left(\theta, \sigma_\theta^2\right)$

Estimators - Confidence Interval

Let's play a game...

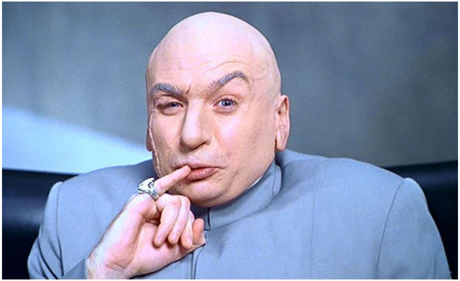
Estimators - Confidence Interval

Let's play a game...

- I know the mean shoe size in Israel
- I'll show you N people and let you measure their shoe size
- I'll ask you to estimate boundaries for the mean $\mu \in (a, b)$
- If you are right and $\mu \in (a, b)$, I'll pay you $\dfrac{50}{b-a}$ ILS
- If you are wrong you'll pay me 8 ILS

Estimators - Confidence Interval

Let's play a game...



Estimators - Confidence Interval

Let's play a game...

- How should you choose a and b?
- Is this game fair?

Estimators - Confidence interval

a $1-\alpha$ confidence interval for parameter θ, is an interval $C=(a,b)$ where $a=a\left(x_1,\ldots,x_n\right)$ and $b=b\left(x_1,\ldots,x_n\right)$ such that $Pr\left(\theta\in C\right)\geq 1-\alpha$ ...

Estimators - Confidence interval

Israeltech challenge

a $1-\alpha$ confidence interval for parameter θ, is an interval $C=(a,b)$ where $a=a\left(x_1,\ldots,x_n\right)$ and $b=b\left(x_1,\ldots,x_n\right)$

- a parameter is a constant
- once a sample is taken and $C=(a,b)$ calculated either θ is in C or isn't
- where's the probability?

Estimators - Confidence interval

a $1-\alpha$ confidence interval for parameter θ, is an interval $C=(a,b)$ where $a=a\left(x_1,\ldots,x_n\right)$ and $b=b\left(x_1,\ldots,x_n\right)$

---

a better question: what is the probability over?

Estimators - Confidence interval

IsraelTech challenge

a $1 - \alpha$ confidence interval for parameter θ, is an interval $C = (a, b)$ where $a = a(x_1, \ldots, x_n)$ and $b = b(x_1, \ldots, x_n)$

---

a better question: what is the probability over?

it's over samples.

Estimators - Confidence interval

a $1-\alpha$ confidence interval for parameter θ, is an interval $C=(a,b)$ where $a=a\left(x_1,\ldots,x_n\right)$ and $b=b\left(x_1,\ldots,x_n\right)$

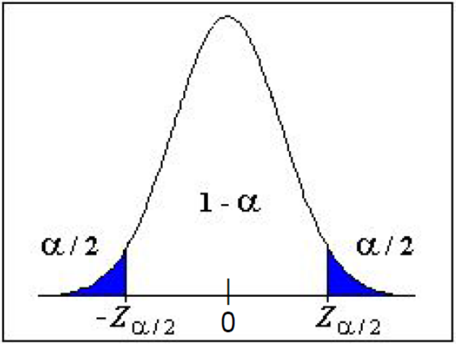
---

possible statements:

- if the experiment is repeated M times
  - $C_1,\ldots,C_M$ calculated
  - $\theta\in C_i$, $1-\alpha$ of the times
- we are about to perform an experiment and calculate C
  - $Pr\left(\theta\in C\right)=1-\alpha$

Confidence Interval - MLE

Let Z be a standard normal random variable, Φ the CDF of Z and $z_{\alpha/2} = \Phi(1 - \alpha/2)$



Confidence Interval - MLE

Let Z be a standard normal random variable, Φ the CDF of Z and $z_{\alpha/2} = \Phi(1 - \alpha/2)$

Theorem If $C = \left(\hat{\theta} - z_{\alpha/2}\hat{se}, \hat{\theta} + z_{\alpha/2}\hat{se}\right)$, then $Pr\left(\theta \in C\right) \to 1 - \alpha$ as $n \to \infty$

Confidence Interval - MLE

- a confidence interval can be computed for every confidence level
- in many cases $\alpha = 0.05$ or the 95% confidence interval is used

Confidence Interval - MLE

- a confidence interval can be computed for every confidence level
- in many cases $\alpha = 0.05$ or the 95% confidence interval is used
- for $\alpha = 0.05$, $z_{\alpha/2} \approx 2$
- which translates to $C = \left(\hat{\theta} - 2\hat{se}, \hat{\theta} + 2\hat{se}\right)$ being an approximate 95% confidence interval

Confidence Interval - MLE

Example: Bernoulli:

- $\alpha = 0.05$
- $\hat{p} = \frac{S}{n}$
- $\hat{se} = \sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}}$
- $C = \left(\frac{S}{n} - \frac{2}{n}\sqrt{S\left(1-\frac{S}{n}\right)}, \frac{S}{n} + \frac{2}{n}\sqrt{S\left(1-\frac{S}{n}\right)}\right)$

Conclusion

- Statistical models
  - parametric models
- Parameter estimation
  - Maximum likelihood
- Estimator properties
  - bias
  - standard error
  - confidence interval

Credits

figures

- confused Mr. Bean - http://mafab.hu/
- Dr. Evil - http://www.quickmeme.com/