Statistics for Astronomers: Lecture 5, 2020.10.13

Prof. Sundar Srinivasan

IRyA/UNAM

Review

Poisson, Uniform, Exponential, and Normal distributions.

Central Limit Theorem. No matter which distribution X is drawn from, $\bar{X}$ is approximately normally distributed about the population mean of X with variance equal to that of X divided by the sample size.

PDFs of functions:

Probability Integral Transform: $Y = F_X(x) \Longrightarrow Y \sim \mathrm{Uniform}(0, 1)$.

CDF, PDF, and convolution methods.

Some review questions.

(Frequentist) Statistical inference

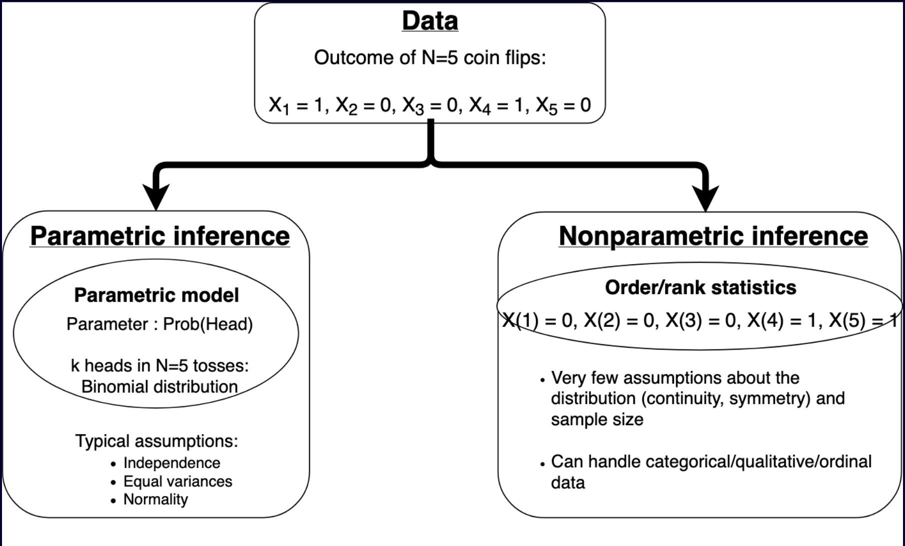
References

Wall & Jenkins

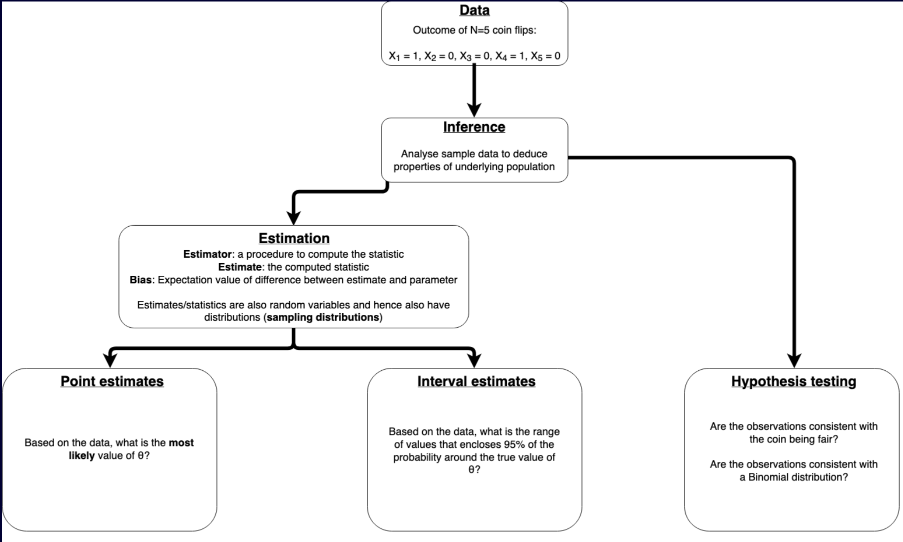
Feigelsen & Babu

R. Andrae, "Error estimation in astronomy: A guide" (ADS)

Inference


Data
Outcome of N=5 coin flips:
$X_1 = 1, X_2 = 0, X_3 = 0, X_4 = 1, X_5 = 0$
Parametric inference
Parametric model
Parameter : Prob(Head)
k heads in N=5 tosses:
Binomial distribution
Typical assumptions:
• Independence
• Equal variances
• Normality
Nonparametric inference
Order/rank statistics
X(1) = 0, X(2) = 0, X(3) = 0, X(4) = 1, X(5) = 1
• Very few assumptions about the distribution (continuity, symmetry) and sample size
• Can handle categorical/qualitative/ordinal data

Inference


Data
Outcome of N=5 coin flips:
$X_1 = 1, X_2 = 0, X_3 = 0, X_4 = 1, X_5 = 0$
Inference
Analyse sample data to deduce properties of underlying population
Estimation
Estimator: a procedure to compute the statistic
Estimate: the computed statistic
Bias: Expectation value of difference between estimate and parameter
Estimates/statistics are also random variables and hence also have distributions (**sampling distributions**)
Point estimates
Based on the data, what is the **most likely** value of θ?
Interval estimates
Based on the data, what is the range of values that encloses 95% of the probability around the true value of θ?
Hypothesis testing
Are the observations consistent with the coin being fair?
Are the observations consistent with a Binomial distribution?

Bayesian vs frequentist inference

Parameters: (frequentist) constants, not random.
Probabilistic aspect contained in likelihood of parameter value given the data.

(Bayesian) random variables modeled by a prior distribution.
Result of analysis is a posterior probability distribution for each parameter.

Bayesian vs frequentist inference

Parameters: (frequentist) constants, not random.
Probabilistic aspect contained in likelihood of parameter value given the data.

(Bayesian) random variables modeled by a prior distribution.
Result of analysis is a posterior probability distribution for each parameter.

Estimators: (frequentist) an estimate for the constant parameter. Random variable.
Uncertainty in estimate based on functional form of likelihood.

(Bayesian) an estimate for the weighted average of the posterior distribution of parameter values. Random variable.
Uncertainty in estimate computed from posterior PDF.

Statistics (Wall & Jenkins Sec. 3.2)

Recall: Populations are summarised by parameters and samples are summarised by statistics.

A statistic is a function $T(X_j; j = 1, \cdots, N)$ of only the data.

Examples: $\max(X_j; j = 1, \cdots, N)$, sample mean, sample median, sample variance, ...

Statistics (Wall & Jenkins Sec. 3.2)

Recall: Populations are summarised by parameters and samples are summarised by statistics.

A statistic is a function $T(X_j; j = 1, \cdots, N)$ of only the data.

Examples: $\max(X_j; j = 1, \cdots, N)$, sample mean, sample median, sample variance, ...

Data X_j are random variables $\Longrightarrow$ a statistic also has a PMF/PDF – the sampling distribution.

Example: according to CLT, the sampling distribution of the sample mean is $\approx$ normal.

Statistics can be compared to parameters via expectation values.

Statistics (Wall & Jenkins Sec. 3.2)

Recall: Populations are summarised by parameters and samples are summarised by statistics.

A statistic is a function $T(X_j; j = 1, \cdots, N)$ of only the data.

Examples: $\max(X_j; j = 1, \cdots, N)$, sample mean, sample median, sample variance, ...

Data X_j are random variables $\Longrightarrow$ a statistic also has a PMF/PDF – the sampling distribution.

Example: according to CLT, the sampling distribution of the sample mean is $\approx$ normal.

Statistics can be compared to parameters via expectation values. Some characteristics of a "good" statistic:

1. Efficiency: reproduce parameter with as few samples as possible.
   Sample mean more efficient than median (variance of mean $\to 0$ as N^{-1}).
2. Robustness: reproduce parameter accurately by being insensitive to outliers in the sample.
   Sample median extremely robust (breakdown point of 50%); breakdown point of sample mean: 0%.
3. Lack of bias: expectation value of the statistic = true parameter value.
   Asymptotically unbiased: bias $\to 0$ as $N \to \infty$. Sample mean always unbiased.
4. Consistency: reproduces true parameter value for very large sample size.
   Sample standard deviation biased, but consistent.

Statistics (Wall & Jenkins Sec. 3.2)

Recall: Populations are summarised by parameters and samples are summarised by statistics.

A statistic is a function $T(X_j; j = 1, \cdots, N)$ of only the data.

Examples: $\max(X_j; j = 1, \cdots, N)$, sample mean, sample median, sample variance, ...

Data X_j are random variables $\Longrightarrow$ a statistic also has a PMF/PDF – the sampling distribution.

Example: according to CLT, the sampling distribution of the sample mean is $\approx$ normal.

Statistics can be compared to parameters via expectation values. Some characteristics of a "good" statistic:

1. Efficiency: reproduce parameter with as few samples as possible.
   Sample mean more efficient than median (variance of mean $\to 0$ as N^{-1}).
2. Robustness: reproduce parameter accurately by being insensitive to outliers in the sample.
   Sample median extremely robust (breakdown point of 50%); breakdown point of sample mean: 0%.
3. Lack of bias: expectation value of the statistic = true parameter value.
   Asymptotically unbiased: bias $\to 0$ as $N \to \infty$. Sample mean always unbiased.
4. Consistency: reproduces true parameter value for very large sample size.
   Sample standard deviation biased, but consistent.

"[S]tatistics of known usefulness are quite rare [most of them having been developed for normally-distributed data]."

In practice useful statistics need to be derived for a specific research problem.
Common approach: Maximum Likelihood Estimation.

Estimators

Parameter values can be guessed from finite samples by computing statistics called estimates.

The "rules" that specify how to compute these estimates are called estimators.

There are estimators for point as well as interval estimates (later).

Estimators

Parameter values can be guessed from finite samples by computing statistics called estimates.
The "rules" that specify how to compute these estimates are called estimators.
There are estimators for point as well as interval estimates (later).

Notation Parameter: θ. Estimator for θ: $\hat{\theta}$.
If X is a random variable, $\hat{\theta}(X)$ is a function of the variable; $\hat{\theta}(x)$ is the value of $\hat{\theta}(X)$ at $X = x$.

Estimators

Parameter values can be guessed from finite samples by computing statistics called estimates.
The "rules" that specify how to compute these estimates are called estimators.
There are estimators for point as well as interval estimates (later).

Notation Parameter: θ. Estimator for θ: $\hat{\theta}$.
If X is a random variable, $\hat{\theta}(X)$ is a function of the variable; $\hat{\theta}(x)$ is the value of $\hat{\theta}(X)$ at $X = x$.

As with a single sample point, we can compare the estimate
– to the parameter being estimated using the (parameter) error:

$e(x) = \hat{\theta}(x) - \theta$. We can estimate $\mathbb{E}[e]$ and $\mathbb{E}[e^2]$:

$\mathbb{E}[e] = \mathbb{E}[\hat{\theta}(x) - \theta] \equiv$ Bias.

$\mathbb{E}[e^2] = \mathbb{E}[(\hat{\theta}(x) - \theta)^2] \equiv$ Mean square error (MSE).

Estimators

Parameter values can be guessed from finite samples by computing statistics called estimates.
The "rules" that specify how to compute these estimates are called estimators.
There are estimators for point as well as interval estimates (later).

Notation Parameter: θ. Estimator for θ: $\hat{\theta}$.
If X is a random variable, $\hat{\theta}(X)$ is a function of the variable; $\hat{\theta}(x)$ is the value of $\hat{\theta}(X)$ at $X = x$.

As with a single sample point, we can compare the estimate

– to the parameter being estimated using the (parameter) error:

$e(x) = \hat{\theta}(x) - \theta$. We can estimate $\mathbb{E}[e]$ and $\mathbb{E}[e^2]$:

$\mathbb{E}[e] = \mathbb{E}[\hat{\theta}(x) - \theta] \equiv$ Bias.

$\mathbb{E}[e^2] = \mathbb{E}[(\hat{\theta}(x) - \theta)^2] \equiv$ Mean square error (MSE).

– or to the expectation value of the estimate using the (sampling) deviation:

$d(x) = \hat{\theta}(x) - \mathbb{E}[\hat{\theta}(x)]$. $\mathbb{E}[d] = 0$, but we can estimate $\mathbb{E}[d^2]$:

$\mathbb{E}[d^2] = \mathbb{E}[(\hat{\theta}(x) - \mathbb{E}[\hat{\theta}(x)])^2] \equiv$ Variance.

Estimators

Parameter values can be guessed from finite samples by computing statistics called estimates.
The "rules" that specify how to compute these estimates are called estimators.
There are estimators for point as well as interval estimates (later).

Notation Parameter: θ. Estimator for θ: $\hat{\theta}$.
If X is a random variable, $\hat{\theta}(X)$ is a function of the variable; $\hat{\theta}(x)$ is the value of $\hat{\theta}(X)$ at $X = x$.

As with a single sample point, we can compare the estimate

– to the parameter being estimated using the (parameter) error:

$e(x) = \hat{\theta}(x) - \theta$. We can estimate $\mathbb{E}[e]$ and $\mathbb{E}[e^2]$:

$\mathbb{E}[e] = \mathbb{E}[\hat{\theta}(x) - \theta] \equiv$ Bias.

$\mathbb{E}[e^2] = \mathbb{E}[(\hat{\theta}(x) - \theta)^2] \equiv$ Mean square error (MSE).

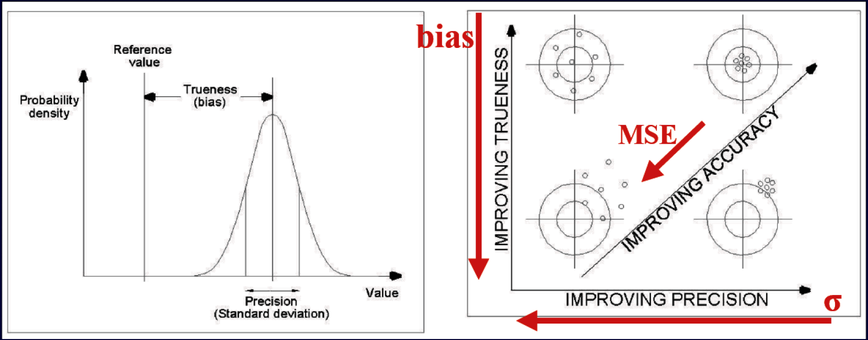
– or to the expectation value of the estimate using the (sampling) deviation:

$d(x) = \hat{\theta}(x) - \mathbb{E}[\hat{\theta}(x)]$. $\mathbb{E}[d] = 0$, but we can estimate $\mathbb{E}[d^2]$:

$\mathbb{E}[d^2] = \mathbb{E}[(\hat{\theta}(x) - \mathbb{E}[\hat{\theta}(x)])^2] \equiv$ Variance.

We can show that $\mathrm{MSE}(\hat{\theta}) = V(\hat{\theta}) + B(\hat{\theta})^2$. Bias-Variance Tradeoff

Bias-Variance Tradeoff



Source: Ivezić+ AstroML book.

Point estimate: intuition

Outcome of $N = 5$ coin tosses:
$X_1 = 1, X_2 = 0, X_3 = 0, X_4 = 1, X_5 = 0.$

Given this data, what is θ, the probability of obtaining a head on a single coin toss?

Sample mean and sample variance

Let $\hat{m}$ be a location estimate based on the data.

Examples: the largest data point, the 10^{th} data point in ascending order, the sample mean, the sample median.

Sample mean and sample variance

Let $\hat{m}$ be a location estimate based on the data.

Examples: the largest data point, the 10^{th} data point in ascending order, the sample mean, the sample median.

Compare the data X_i to the population mean:

Error: deviation of data point from the population mean: $X_i - \mu$ (in general, called bias).
$\mathbb{E}[\bar{X} - \mu] = 0$ (CLT) $\implies$ the sample mean is an unbiased estimator of μ.

Sample mean and sample variance

Let $\hat{m}$ be a location estimate based on the data.

Examples: the largest data point, the 10^{th} data point in ascending order, the sample mean, the sample median.

Compare the data X_i to the population mean:

Error: deviation of data point from the population mean: $X_i - \mu$ (in general, called bias).
$\mathbb{E}[\bar{X} - \mu] = 0$ (CLT) $\Longrightarrow$ the sample mean is an unbiased estimator of μ.

Compare the data X_i to the location estimate:

Residual: deviation of data point from location estimate: $X_i - \hat{m}$.

Estimate population variance using sum of squares of residues, $SSR = \sum_{i=1}^{N} \left(X_i - \hat{m}\right)^2$.

Sample mean and sample variance

Let $\hat{m}$ be a location estimate based on the data.

Examples: the largest data point, the 10^{th} data point in ascending order, the sample mean, the sample median.

Compare the data X_i to the population mean:

Error: deviation of data point from the population mean: $X_i - \mu$ (in general, called bias).
$\mathbb{E}[\bar{X} - \mu] = 0$ (CLT) $\implies$ the sample mean is an unbiased estimator of μ.

Compare the data X_i to the location estimate:

Residual: deviation of data point from location estimate: $X_i - \hat{m}$.

Estimate population variance using sum of squares of residues, $SSR = \sum_{i=1}^{N} \left(X_i - \hat{m}\right)^2$.

The SSR is minimised if we choose $\hat{m} = \bar{X}$, biasing our variance estimate.

Sample mean and sample variance

Let $\hat{m}$ be a location estimate based on the data.

Examples: the largest data point, the 10^{th} data point in ascending order, the sample mean, the sample median.

Compare the data X_i to the population mean:

Error: deviation of data point from the population mean: $X_i - \mu$ (in general, called bias).
$\mathbb{E}[\bar{X} - \mu] = 0$ (CLT) $\Longrightarrow$ the sample mean is an unbiased estimator of μ.

Compare the data X_i to the location estimate:

Residual: deviation of data point from location estimate: $X_i - \hat{m}$.

Estimate population variance using sum of squares of residues, $SSR = \sum_{i=1}^{N} \left(X_i - \hat{m}\right)^2$.

The SSR is minimised if we choose $\hat{m} = \bar{X}$, biasing our variance estimate.

Bessel's Correction compensates for this: choose $\widehat{S^2} = \dfrac{SSR}{N-1}$ instead of $\widehat{S^2} = \dfrac{SSR}{N}$.

Sample mean and sample variance

Let $\hat{m}$ be a location estimate based on the data.

Examples: the largest data point, the 10^{th} data point in ascending order, the sample mean, the sample median.

Compare the data X_i to the population mean:

Error: deviation of data point from the population mean: $X_i - \mu$ (in general, called bias).
$\mathbb{E}[\bar{X} - \mu] = 0$ (CLT) $\implies$ the sample mean is an unbiased estimator of μ.

Compare the data X_i to the location estimate:

Residual: deviation of data point from location estimate: $X_i - \hat{m}$.

Estimate population variance using sum of squares of residues, $SSR = \sum_{i=1}^{N} \left(X_i - \hat{m}\right)^2$.

The SSR is minimised if we choose $\hat{m} = \bar{X}$, biasing our variance estimate.

Bessel's Correction compensates for this: choose $\widehat{S^2} = \frac{SSR}{N-1}$ instead of $\widehat{S^2} = \frac{SSR}{N}$.

While the sample mean has N degrees of freedom (it is based on N measurements), the $\#dof$ for the variance depends on whether the population mean is known ($\#dof = N$) or is estimated from the data ($\#dof = N - 1$).

Bessel's correction for sample variance

Recall: for any random variable $X \sim (\mu, \sigma^2)$, $\mathbb{E}[X^2] = \mu^2 + \sigma^2$.

Since $\bar{X} \sim (\mu, \sigma^2/N)$ (CLT), $\mathbb{E}[\bar{X}^2] = \mu^2 + \sigma^2/N$.

We now compute the expectation value of the sum of squares of residues:

Bessel's correction for sample variance

Recall: for any random variable $X \sim (\mu, \sigma^2)$, $\mathbb{E}[X^2] = \mu^2 + \sigma^2$.

Since $\bar{X} \sim (\mu, \sigma^2/N)$ (CLT), $\mathbb{E}[\bar{X}^2] = \mu^2 + \sigma^2/N$.

We now compute the expectation value of the sum of squares of residues:

$$\mathbb{E}\Big[\sum_{i=1}^{N}(X_j - \bar{X})^2\Big] = \mathbb{E}\Big[\sum_{i=1}^{N} X_j^2 - N\,\bar{X}^2\Big]$$

Bessel's correction for sample variance

Recall: for any random variable $X \sim (\mu, \sigma^2)$, $\mathbb{E}[X^2] = \mu^2 + \sigma^2$.

Since $\bar{X} \sim (\mu, \sigma^2/N)$ (CLT), $\mathbb{E}[\bar{X}^2] = \mu^2 + \sigma^2/N$.

We now compute the expectation value of the sum of squares of residues:

$$\mathbb{E}\Big[\sum_{i=1}^{N}(X_j - \bar{X})^2\Big] = \mathbb{E}\Big[\sum_{i=1}^{N} X_j^2 - N\,\bar{X}^2\Big]$$

$$= N\left(\mathbb{E}[X^2] - \mathbb{E}[\bar{X}^2]\right) = N\left(\left(\cancel{\mu^2} + \sigma^2\right) - \left(\cancel{\mu^2} + \frac{\sigma^2}{N}\right)\right)$$

Bessel's correction for sample variance

Recall: for any random variable $X \sim (\mu, \sigma^2)$, $\mathbb{E}[X^2] = \mu^2 + \sigma^2$.

Since $\bar{X} \sim (\mu, \sigma^2/N)$ (CLT), $\mathbb{E}[\bar{X}^2] = \mu^2 + \sigma^2/N$.

We now compute the expectation value of the sum of squares of residues:

$$\mathbb{E}\Big[\sum_{i=1}^{N}(X_j - \bar{X})^2\Big] = \mathbb{E}\Big[\sum_{i=1}^{N} X_j^2 - N\,\bar{X}^2\Big]$$

$$= N\,\left(\mathbb{E}[X^2] - \mathbb{E}[\bar{X}^2]\right) = N\,\left(\left(\cancel{\mu^2} + \sigma^2\right) - \left(\cancel{\mu^2} + \frac{\sigma^2}{N}\right)\right)$$

$$= (N-1)\,\sigma^2.$$

Bessel's correction for sample variance

Recall: for any random variable $X \sim (\mu, \sigma^2)$, $\mathbb{E}[X^2] = \mu^2 + \sigma^2$.

Since $\bar{X} \sim (\mu, \sigma^2/N)$ (CLT), $\mathbb{E}[\bar{X}^2] = \mu^2 + \sigma^2/N$.

We now compute the expectation value of the sum of squares of residues:

$$\mathbb{E}\Big[\sum_{i=1}^{N}(X_j - \bar{X})^2\Big] = \mathbb{E}\Big[\sum_{i=1}^{N} X_j^2 - N\,\bar{X}^2\Big]$$

$$= N\,\left(\mathbb{E}[X^2] - \mathbb{E}[\bar{X}^2]\right) = N\,\left(\left(\cancel{\mu^2} + \sigma^2\right) - \left(\cancel{\mu^2} + \frac{\sigma^2}{N}\right)\right)$$

$$= (N-1)\,\sigma^2.$$

$$\implies \mathbb{E}\Big[\frac{1}{N-1}\sum_{i=1}^{N}(X_j - \bar{X})^2\Big] \equiv \mathbb{E}[\widehat{S^2}] = \sigma^2.$$

Bessel's correction for sample variance

Recall: for any random variable $X \sim (\mu, \sigma^2)$, $\mathbb{E}[X^2] = \mu^2 + \sigma^2$.

Since $\bar{X} \sim (\mu, \sigma^2/N)$ (CLT), $\mathbb{E}[\bar{X}^2] = \mu^2 + \sigma^2/N$.

We now compute the expectation value of the sum of squares of residues:

$$\mathbb{E}\Big[\sum_{i=1}^{N}(X_j - \bar{X})^2\Big] = \mathbb{E}\Big[\sum_{i=1}^{N} X_j^2 - N\,\bar{X}^2\Big]$$

$$= N\,\Big(\mathbb{E}[X^2] - \mathbb{E}[\bar{X}^2]\Big) = N\,\Big(\big(\cancel{\mu^2} + \sigma^2\big) - \Big(\cancel{\mu^2} + \frac{\sigma^2}{N}\Big)\Big)$$

$$= (N-1)\,\sigma^2.$$

$$\implies \mathbb{E}\Big[\frac{1}{N-1}\sum_{i=1}^{N}(X_j - \bar{X})^2\Big] \equiv \mathbb{E}[\widehat{S^2}] = \sigma^2.$$

Thus, the unbiased estimator for the population variance requires us to reduce $\#dof$ to $N-1$.
Only required if population mean estimated using sample mean, or N small.

Probability vs. likelihood

Experiment: ten coin tosses.
Model: Outcome of each toss $\sim \mathrm{Bernoulli}(\theta)$. Outcome of ten tosses $\sim \mathrm{Binomial}(10, \theta)$.

Probability vs. likelihood

Experiment: ten coin tosses.
Model: Outcome of each toss $\sim \mathrm{Bernoulli}(\theta)$. Outcome of ten tosses $\sim \mathrm{Binomial}(10, \theta)$.

Probability: Predicting future outcomes

Predict what the data will look like given a particular model.

Always normalised.

Probability vs. likelihood

Experiment: ten coin tosses.
Model: Outcome of each toss $\sim \text{Bernoulli}(\theta)$. Outcome of ten tosses $\sim \text{Binomial}(10, \theta)$.

Probability: Predicting future outcomes

Predict what the data will look like given a particular model.

Always normalised.

What is the probability that we obtain eight heads in ten tosses?

$$P(\text{data}|\text{model}) = P(X = 8, N = 10 \mid \theta) = \binom{10}{8} \theta^8 (1 - \theta)^2$$

Probability vs. likelihood

Experiment: ten coin tosses.
Model: Outcome of each toss $\sim \mathrm{Bernoulli}(\theta)$. Outcome of ten tosses $\sim \mathrm{Binomial}(10, \theta)$.

Probability: Predicting future outcomes

Predict what the data will look like given a particular model.

Always normalised.

What is the probability that we obtain eight heads in ten tosses?

$$P(\mathrm{data}|\mathrm{model}) = P(X = 8, N = 10 \mid \theta) = \binom{10}{8} \theta^8 \, (1-\theta)^2$$

Likelihood: Explaining existing observations

Given a sample, gauge the plausibility (believability) that it was drawn from a particular model. Function of model alone.

Probability vs. likelihood

Experiment: ten coin tosses.
Model: Outcome of each toss $\sim \mathrm{Bernoulli}(\theta)$. Outcome of ten tosses $\sim \mathrm{Binomial}(10, \theta)$.

Probability: Predicting future outcomes

Predict what the data will look like given a particular model.

Always normalised.

What is the probability that we obtain eight heads in ten tosses?

$$P(\mathrm{data}|\mathrm{model}) = P(X = 8, N = 10 \mid \theta) = \binom{10}{8} \theta^8 (1-\theta)^2$$

Likelihood: Explaining existing observations

Given a sample, gauge the plausibility (believability) that it was drawn from a particular model. Function of model alone.

When comparing models, only ratio matters.
Not always normalised.

Probability vs. likelihood

Experiment: ten coin tosses.
Model: Outcome of each toss $\sim \mathrm{Bernoulli}(\theta)$. Outcome of ten tosses $\sim \mathrm{Binomial}(10, \theta)$.

Probability: Predicting future outcomes

Predict what the data will look like given a particular model.

Always normalised.

What is the probability that we obtain eight heads in ten tosses?

$$P(\mathrm{data}|\mathrm{model}) = P(X = 8, N = 10 \mid \theta)$$
$$= \binom{10}{8} \theta^8 (1-\theta)^2$$

Likelihood: Explaining existing observations

Given a sample, gauge the plausibility (believability) that it was drawn from a particular model. Function of model alone.

When comparing models, only ratio matters.
Not always normalised.

Given that we obtain eight heads in ten tosses, what is the likelihood of parameter value θ?

$$\mathscr{L}(\mathrm{model}|\mathrm{data}) = P(X = 8, N = 10 \mid \theta)$$
$$= \binom{10}{8} \theta^8 (1-\theta)^2$$

Probability vs. likelihood

Experiment: ten coin tosses.
Model: Outcome of each toss $\sim \mathrm{Bernoulli}(\theta)$. Outcome of ten tosses $\sim \mathrm{Binomial}(10, \theta)$.

Probability: Predicting future outcomes

Predict what the data will look like given a particular model.

Always normalised.

What is the probability that we obtain eight heads in ten tosses?

$$P(\mathrm{data}|\mathrm{model}) = P(X = 8, N = 10 \mid \theta) = \binom{10}{8} \theta^8 (1-\theta)^2$$

Likelihood: Explaining existing observations

Given a sample, gauge the plausibility (believability) that it was drawn from a particular model. Function of model alone.

When comparing models, only ratio matters.
Not always normalised.

Given that we obtain eight heads in ten tosses, what is the likelihood of parameter value θ?

$$\mathscr{L}(\mathrm{model}|\mathrm{data}) = P(X = 8, N = 10 \mid \theta) = \binom{10}{8} \theta^8 (1-\theta)^2$$

> "When two different models, or perhaps two variants of the same model differing only in the value of some adjustable parameter(s), are to be compared as explanations for the same observed outcome, the probability of obtaining this particular outcome can be calculated for each and is then known as the likelihood for the model or parameter value(s) given the data.'"
>
> — A. W. F. Edwards, "Likelihood"

Likelihood in the Bayesian interpretation

Definition (Bayes' Theorem)

$$\underbrace{P(\text{model}|\text{data})}_{\text{"combined/posterior likelihood"}} = P(\text{data}|\text{model})\frac{P(\text{model})}{P(\text{data})}$$

Likelihood in the Bayesian interpretation

Definition (Bayes' Theorem)

$$\underbrace{P(\text{model}|\text{data})}_{\text{“combined/posterior likelihood”}} = P(\text{data}|\text{model})\frac{P(\text{model})}{P(\text{data})} = \overbrace{\mathscr{L}(\text{model})}^{\text{current knowledge}} \frac{\overbrace{P(\text{model})}^{\text{“prior likelihood”}}}{P(\text{data})}$$

Likelihood in the Bayesian interpretation

Definition (Bayes' Theorem)

$$\underbrace{P(\text{model}|\text{data})}_{\text{“combined/posterior likelihood”}} = P(\text{data}|\text{model})\frac{P(\text{model})}{P(\text{data})} = \overbrace{\mathscr{L}(\text{model})}^{\text{current knowledge}} \frac{\overbrace{P(\text{model})}^{\text{“prior likelihood”}}}{P(\text{data})}$$

Bayesian inference: maximise posterior distribution.

If prior distribution is uninformative or diffuse, same as MLE.

Likelihood in the Bayesian interpretation

Definition (Bayes' Theorem)

$$\underbrace{P(\text{model}|\text{data})}_{\text{“combined/posterior likelihood”}} = P(\text{data}|\text{model})\frac{P(\text{model})}{P(\text{data})} = \overbrace{\mathscr{L}(\text{model})}^{\text{current knowledge}} \frac{\overbrace{P(\text{model})}^{\text{“prior likelihood”}}}{P(\text{data})}$$

Bayesian inference: maximise posterior distribution.
If prior distribution is uninformative or diffuse, same as MLE.

When comparing two models given the same data, only the ratio of likelihoods matters.
So $\mathscr{L}(\theta)$ is only known up to a multiplicative constant (not normalised).

Likelihood as a function of parameter values

Observation: $N = 10$ coin tosses result in $X = 8$ heads.
What is the likelihood that the coin is fair?

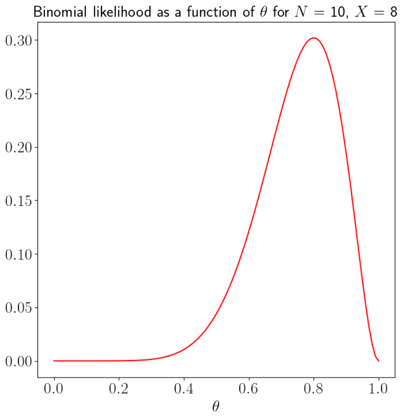
$$\mathscr{L}(\theta) = P(X = 8, N = 10 \mid \theta) = \binom{10}{8} \theta^8 \, (1 - \theta)^2$$

Likelihood as a function of parameter values

Observation: $N = 10$ coin tosses result in $X = 8$ heads.
What is the likelihood that the coin is fair?

$$\mathscr{L}(\theta) = P(X = 8, N = 10 \mid \theta) = \binom{10}{8} \theta^8 (1-\theta)^2$$

$\mathscr{L}(\theta = 0.5)$ is quite small!



Code for plot available here.

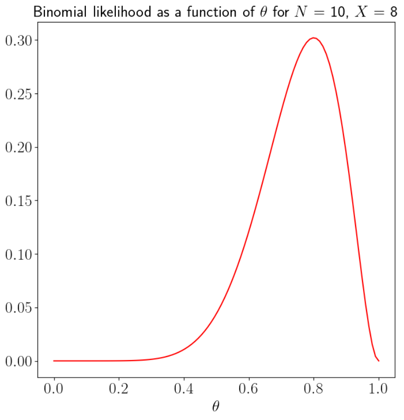
Likelihood as a function of parameter values

Observation: $N = 10$ coin tosses result in $X = 8$ heads.
What is the likelihood that the coin is fair?

$$\mathscr{L}(\theta) = P(X = 8, N = 10 \mid \theta) = \binom{10}{8} \theta^8 (1 - \theta)^2$$

$\mathscr{L}(\theta = 0.5)$ is quite small!

Likelihood = relative preference for various parameter values.



Code for plot available here.

Likelihood as a function of parameter values

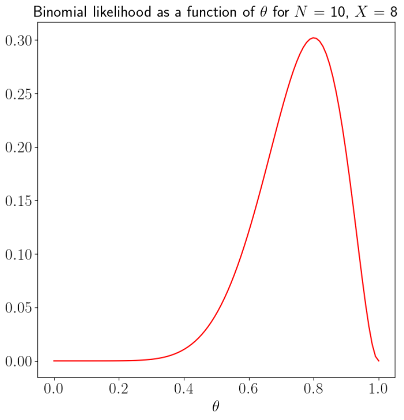
Observation: $N = 10$ coin tosses result in $X = 8$ heads.
What is the likelihood that the coin is fair?

$$\mathscr{L}(\theta) = P(X = 8, N = 10 \mid \theta) = \binom{10}{8} \theta^8 \, (1-\theta)^2$$

$\mathscr{L}(\theta = 0.5)$ is quite small!

Likelihood = relative preference for various parameter values.

While this particular $\mathscr{L}(\theta)$ is a relatively sharply peaked function, others may not be.



Code for plot available here.

Likelihood as a function of parameter values

Observation: $N = 10$ coin tosses result in $X = 8$ heads.
What is the likelihood that the coin is fair?

$$\mathscr{L}(\theta) = P(X = 8, N = 10 \mid \theta) = \binom{10}{8} \theta^8 (1 - \theta)^2$$

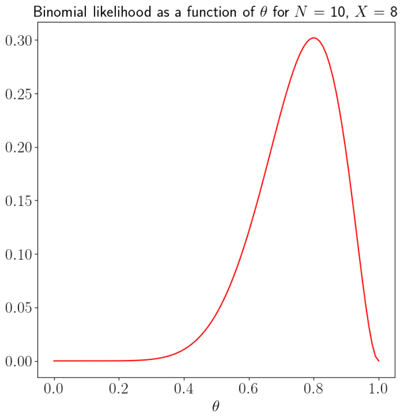
$\mathscr{L}(\theta = 0.5)$ is quite small!

Likelihood = relative preference for various parameter values.

While this particular $\mathscr{L}(\theta)$ is a relatively sharply peaked function, others may not be.

Important to investigate the entire range of values for the function.

Loss of information if we restrict ourselves to location of $max(\mathscr{L})$!



Code for plot available here.