Statistics and Experimental Design: Statistical Estimation

Module 2 - Statistics

academy.
constructor.org

Statistical Inference

Statistical inference allows to draw insights about the population based on specific statistics obtained from a sample.

Key goals:

1. Estimation of unknown parameters of the statistical model
2. Select the best model for your data
3. Do predictions with your model
4. Hypothesis testing

Methodologies:

5. Parametric
   - Inferential approach
   - Bayesian approach
6. Non-parametric

Day 2: Statistical estimation

Part 1 - Basic notions

- Estimation components
- Unbiased estimator

Part 2 - Likelihood function

- Definition
- Likelihood maximization

Part 3 - Analysis of variance: Confidence intervals

- Classic method (parametric approach)
- Bootstrap (non-parametric approach)

Part 4 - Bayesian approach

- Prior and posterior
- Conjugate priors
- EM - algorithm

Part 1: Basic notions: Estimation, Estimator, and Estimate

- Populations can be at least partially described by population parameters like mean, proportion, variance etc.
- Statistics or point estimators are used to estimate population parameters, based on a sample of the the population.
- An estimator or statistic is a function of the sample; it is a rule that tells you how to calculate an estimate of a parameter from a sample.
- An estimate is a numeric value computed by the estimator based on the sample data.

Recap: Estimators and Estimates

- **Estimator:** is a statistic (arbitrary function of a random sample), used to extract information of a parameter from the random sample

➢ expressed as a function of X

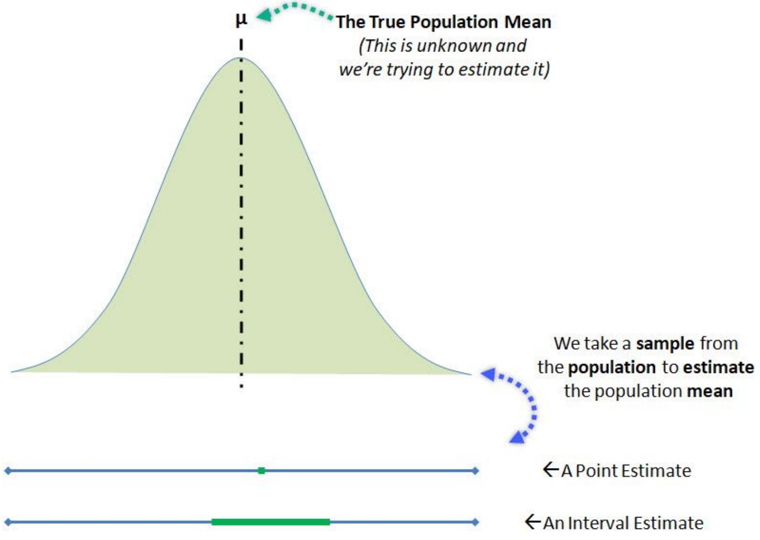
- **Estimate:** value of the evaluated estimator computed based on the data (realizations of the random sample)

➢ computed based on x_1, x_2, …, x_n

- usually denoted by $\hat{\theta}$ for a parameter θ.

Two types of estimators:

- **Point estimator:** single value
- **Interval estimator:** defined by lower and upper limit


μ
The True Population Mean
(This is unknown and we're trying to estimate it)
We take a **sample** from the **population** to **estimate** the population **mean**
←A Point Estimate
←An Interval Estimate

Source

Point Estimate: Example

- Let X represent the random variable ‘height of male freshmen students’, and suppose we would like to know $\mu_x = E(X)$.

- We can’t possibly get the height of all male freshmen students in the world, but we can get the data from a sample of people: let x consist of N i.i.d. samples drawn from X:

$$x = \{X_1, X_2, ..., X_{□}\}$$

Let:

$$\bar{X} = (1/n) \sum(i=1 \text{ to } n) X_i$$

$\bar{X}$ is a random variable!

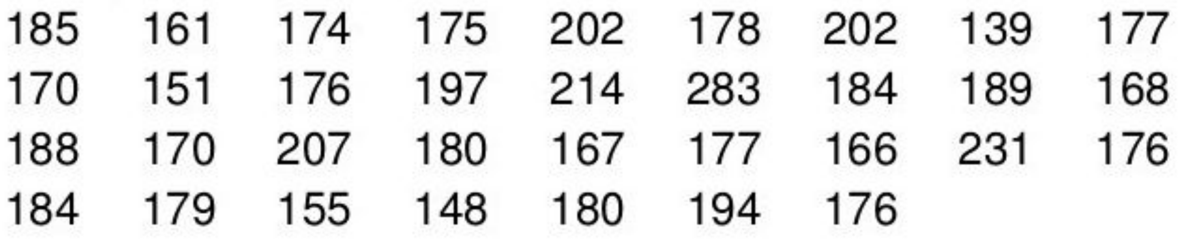
Also: $\bar{X}$ is an estimator for the mean μ_x of X.

Further Reading

Point Estimate: Example

A point estimator of the population mean is the sample mean: $\overline{\boldsymbol{X}} = \frac{\mathbf{1}}{\boldsymbol{n}} \sum \boldsymbol{X_i}$

A sample of height of 34 male freshman students was obtained:



185	161	174	175	202	178	202	139	177
170	151	176	197	214	283	184	189	168
188	170	207	180	167	177	166	231	176
184	179	155	148	180	194	176		

If one wanted to estimate the true mean of all male freshman students, you might use the sample mean as a point estimator for the true mean:

Sample mean = $\overline{x} \ = \mathbf{182.44}$

Comparison of Estimators: What is a “good” estimator?

Main characteristics:

- **Unbiasedness**: the expected value equals the true value.

➢On average, it hits the true parameter value. Independent of sample size**.**

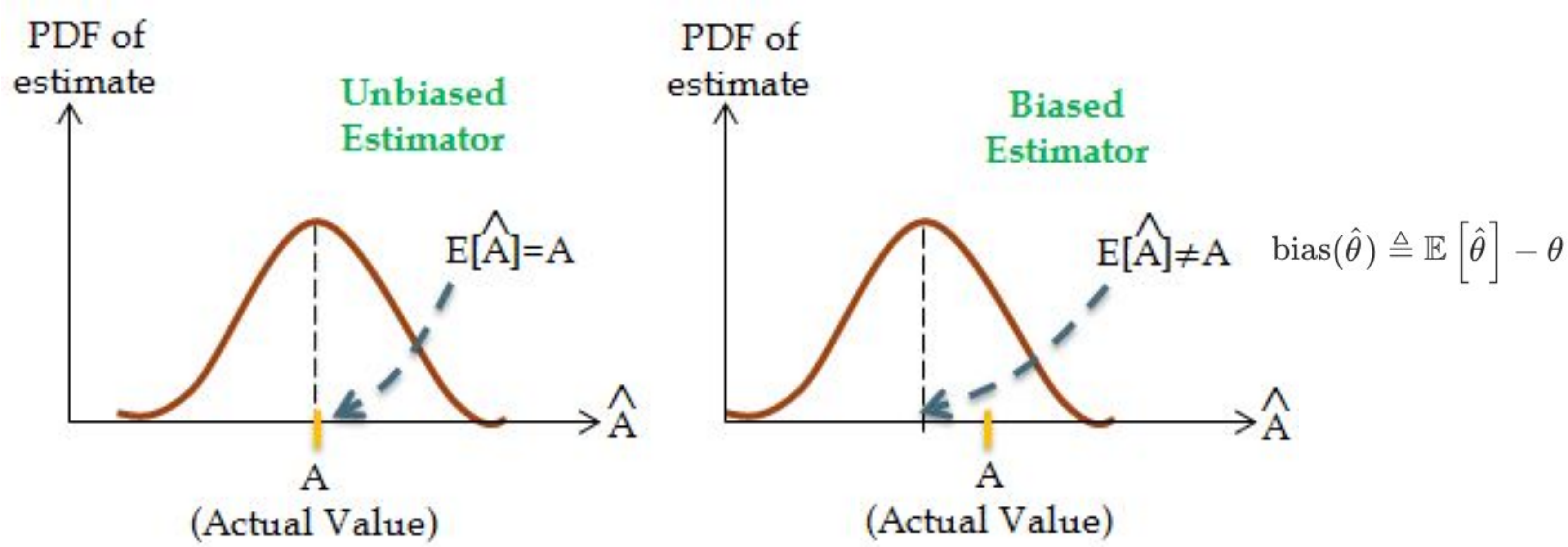
For an unbiased estimator we are looking for the most efficient (estimates the quantity of interest in some “best possible” manner.)

•**Efficiency**: refers to the ability of the estimator to estimate the quantity of interest as accurate as possible.

•**Consistency:** the larger the sample size the more “accurate” the estimate.

➢ The more data you collect, a consistent estimator will be close to the real population parameter you’re trying to measure.

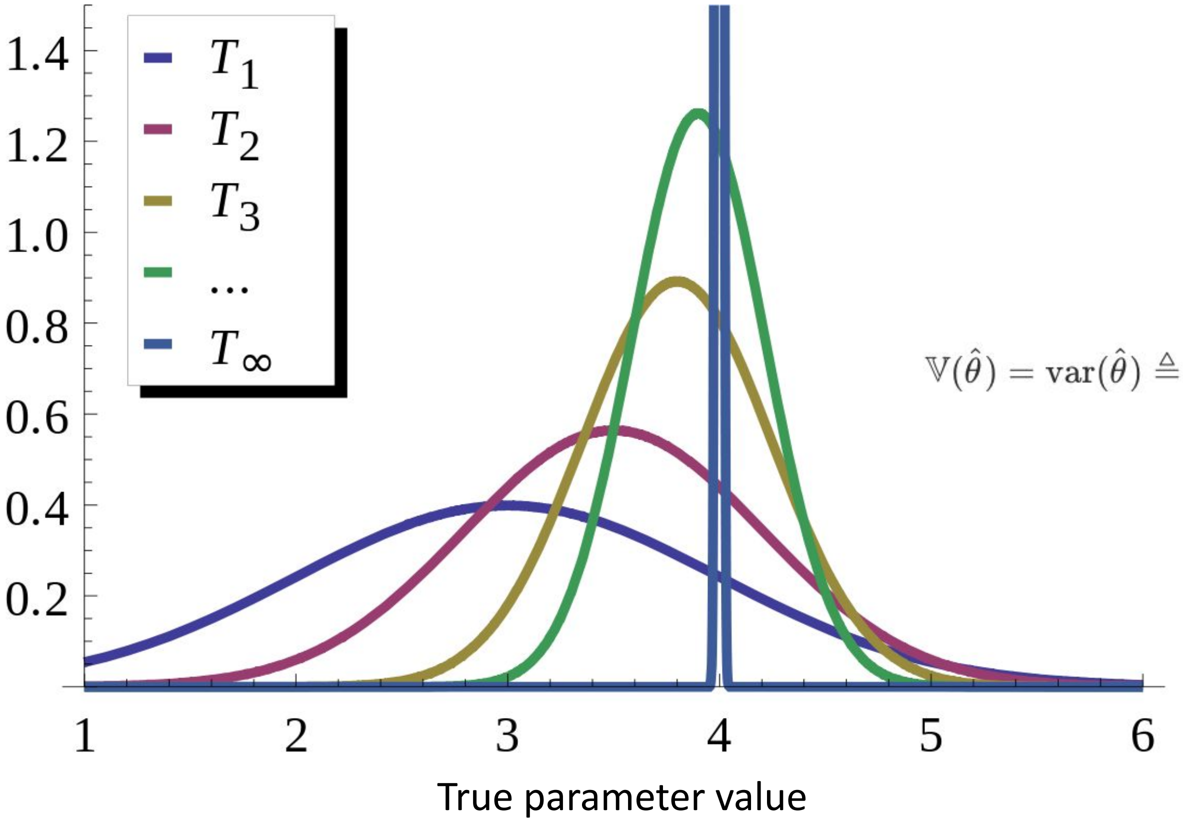
What is a "good" estimator: Bias



$$\text{bias}(\hat{\theta}) \triangleq \mathbb{E}\left[\hat{\theta}\right] - \theta$$

Further Reading

What is a “good” estimator: Consistency



$$\mathbb{V}(\hat{\theta}) = \mathrm{var}(\hat{\theta}) \triangleq \mathbb{E}\left[(\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right])^2\right]$$

Source

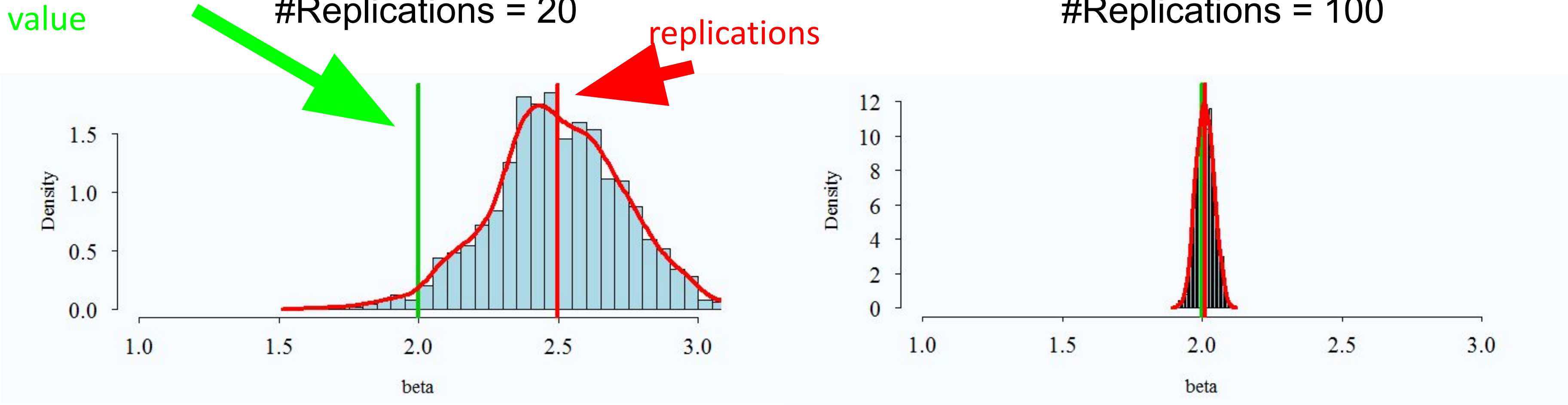
Example: What is a “good” estimator?

True parameter value

#Replications = 20

average of the simulated replications

#Replications = 100



Is the estimate unbiased and/or consistent? Why?

Source

Basic notions: Biased and unbiased estimate: Example

- *Task:* Prove m(x) = $\sum_{i=1}^{N} x_i/N$ is an unbiased estimator of μ of X if E(x) = μ

Part 2: Likelihood function

A likelihood function (often simply called the likelihood) measures how well a statistical model explains observed data by calculating the probability of seeing that data under different parameter values of the model. It is constructed from the joint probability distribution of the random variable that (presumably) generated the observations.

Part 2: Likelihood function

Likelihood function is a function of parameters given data

$$\mathbf{L(\theta) = L(\theta | x_1, x_2, \ldots, x_n) = f(x_1, x_2, \ldots, x_n | \theta)}$$

- Likelihood function is the **probability** of observing the **data** given a set of parameter values

- It measures the support provided by the data for each possible value of the parameter.

- If we observe that $\mathbf{L(\theta_1 | x) > L(\theta_2 | x)}$ at two set of parameter values θ_1 and θ_2 then it can be interpreted that $\boldsymbol{\theta_1}$ is the more likely value for the parameter θ.

Likelihood function - A simple example

Problem:
M&M’s sold in the United States have 50% red candies compared to 30% in those sold in Canada. In an experimental study, a sample of 5 candies were drawn from an unlabelled bag and 2 red candies were observed. Is it more plausible that this bag was from the United States or from Canada?

source

Likelihood function - A simple example

Problem:
M&M’s sold in the United States have 50% red candies compared to 30% in those sold in Canada. In an experimental study, a sample of 5 candies were drawn from an unlabelled bag and 2 red candies were observed. Is it more plausible that this bag was from the United States or from Canada?

The study was repeated twice. The second time out of 5 candies 3 were red. Does it change the conclusion?

source

Likelihood inference

The **likelihood function** is the **probability mass** or **density function** of the *observed data*, viewed as **function of the unknown parameter**.

Maximum likelihood:

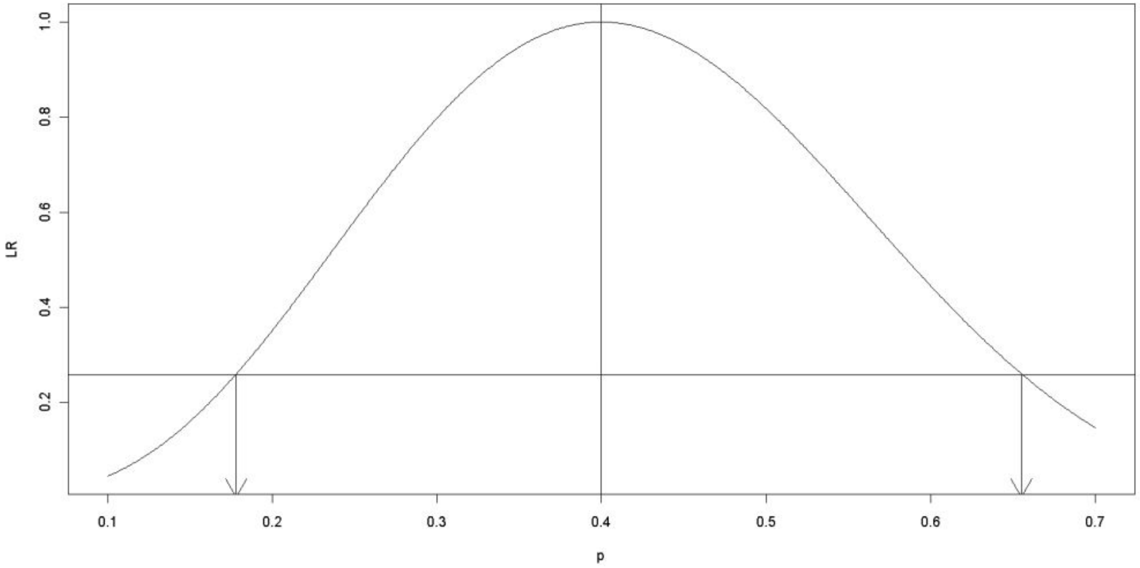
- Maximum of the likelihood function
- Represents the most plausible value of the unknown parameter

Relative likelihood:

- Transforms the arbitrary scale of the likelihood into a standardized scale between 0 and 1

$$\tilde{L}(\theta) = \frac{L(\theta)}{L(\hat{\theta}_{\mathrm{ML}})}$$

Relative Likelihood - Example



- Measures how likely different values of p are relative to p=0.4
- Can be used to define a group of values of p that are supported by the data, i.e. a group of values whose likelihood ratio is above a critical value.

Maximum Likelihood Estimation

The likelihood function is the **probability mass** or **density function** of the observed data, viewed as function of the unknown parameter.

Maximum likelihood:

- Maximum of the likelihood function
- Represents the most plausible value of the unknown parameter

How to solve Likelihood Function?

- Maximum likelihood estimation (MLE):
  - MLE solves for Log-likelihood function
  - Derivative based method
- Brute Force Method (Permutation)

Maximum Likelihood Estimation

How to solve Likelihood Function?

- Maximum likelihood estimation (MLE):

  1. Consider a PMF/PDF of the data given the parameters $f(x|\theta)$
  2. Likelihood under i.i.d assumptions: $L(\theta) = f(x|\theta) = f(x_1, \dots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$
  3. Log-Likelihood: $\ln L(\theta) = \ln \prod_{i=1}^{n} f(x_i|\theta) = \sum_{i=1}^{n} \ln f(x_i|\theta) = LL(\theta)$
  4. Take the derivative of $LL(\theta)$ w.r.t. θ to find the optimum: $\frac{\partial LL(\theta)}{\partial \theta} = 0$
  5. Take the second derivative of $LL(\theta)$ w.rt. θ to ensure that it is negative, which confirms the found optimum is maximum: $\frac{\partial^2 LL(\theta)}{\partial^2 \theta} < 0$

Maximum Likelihood Estimation: Example

Binomial Likelihood $f(k, n, p) = \Pr(k; n, p) = \Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$

MLE: Let’s find p that maximizes P(X=k) given k.

Part 3: Variance analysis: Confidence intervals

Main point: **Draw information from obtained data about a statistic**

➢ Estimation refers to the process by which one makes inferences about a population, based on information obtained from a sample.

Keep in mind: *data is considered a realization of a random sample with a certain distribution.*

Goal of estimation: make statements about a statistic (i.e the mean)

- *parametric approach*: by evaluating the parameters of the underlying distribution based on the distribution
- *non-parametric*: by calculating the statistics directly from the sample

Confidence Intervals: Classic approach

- Since we observe random data, the point estimate of the mean is not enough
- The confidence interval serves to capture the variance of the error relative to the point estimate
- Because of the CLT we assume that the error is normally distributed
- To capture the quantile we use the standard normal distribution corrected to the sample
- Quantiles can be obtained from the standard normal table (i.e Standard Normal Table (sjsu.edu))

Source

Confidence Intervals: Classic approach

Point Estimate – The single value used to approximate the population parameter.

Confidence Interval – The range of values, sometimes called the interval estimate, that is used to estimate the true value of the population parameter.

Confidence Level – The probability that the confidence interval does, in fact, contain the true population parameter, assuming that the estimation process is repeated many times. $(1 - \alpha)$.

Critical Value – A critical value is the z-score that separates sample statistics likely to occur from those unlikely to occur. The number $Z\alpha/_{2}$ is the z-score that separates a region of $\alpha/_{2}$ from the rest of the standard normal curve.

Common Critical Values		
Confidence Level of 90%	$\alpha = .10$	$Z\alpha/_{2} = 1.645$
Confidence Level of 95%	$\alpha = .05$	$Z\alpha/_{2} = 1.96$
Confidence Level of 99%	$\alpha = .01$	$Z\alpha/_{2} = 2.575$

Margin of Error – The maximum likely difference, given a certain confidence level, between the observed sample proportion ($\hat{p}$) and the true value of the population proportion (p); or between the observed sample mean ($\bar{x}$) and the true value of the population mean (μ).

Confidence Intervals: Inferential approach

- interval estimates attached with a confidence level used to express the precision and uncertainty associated with a particular sampling method.

Consisting of three parts:

confidence level:	commonly: 95%
statistic:	point estimate
margin of error:	critical value * standard deviation

$$\hat{\theta} \pm z_{1-\alpha/2}\sqrt{\mathrm{Var}(\hat{\theta})},$$

Example: 95%-CI for the mean: $\bar{X} \mp 1.96 * \frac{\sigma}{\sqrt{N}}$

Further Reading

Confidence Intervals

confidence level

$$\bar{X} \mp \underbrace{z_{1-\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{N}}}_{\text{margin of error}}$$

point estimate

Confidence Intervals: Exercise

confidence level

$$\underset{\text{point estimate}}{\bar{X}} \quad \mp \quad \underbrace{z_{1-\frac{\alpha}{2}} \quad * \quad \frac{\sigma}{\sqrt{N}}}_{\text{margin of error}}$$

- What is the 5%-confidence interval if the mean is 0, n =10000 and the variance is 100?
- What is the 5%-confidence interval if the mean is 0, n =10000 and the variance is 100?
- What is the 10%-confidence interval if the mean is 25, n =100 and the variance is 2?

https://www.sjsu.edu/faculty/gerstman/EpiInfo/z-table.htm

Non-parametric approach: Bootstrapping

- Resampling is impossible or/and the distribution cannot be figured out
- **Data is the population**
- **No distributional assumptions**

- Repeated sampling from a population may be impractical, expensive or impossible (i.e sample items destroyed during sampling).

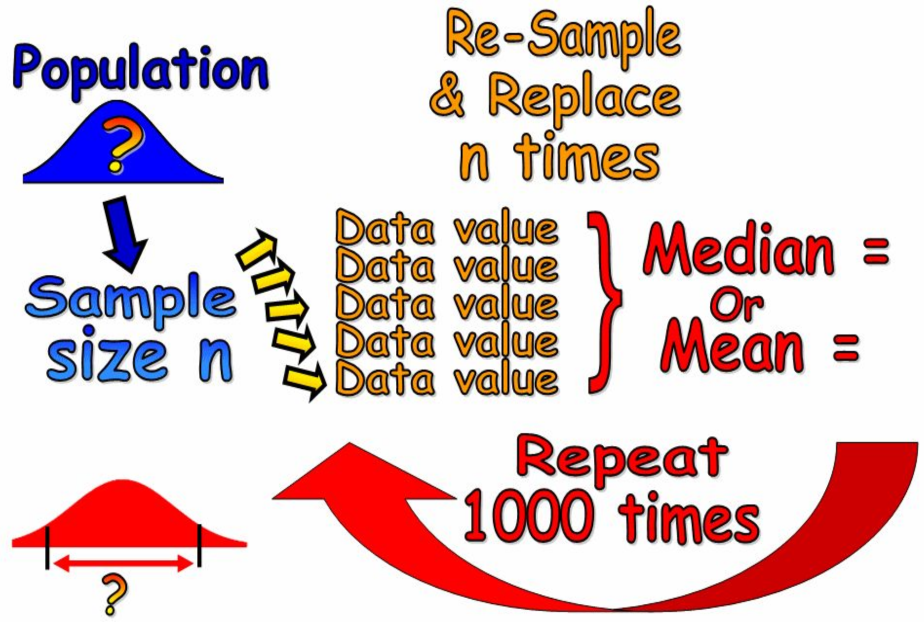
- Impossible to resample from the population (or/and the true sampling distribution is unavailable):
  ⟶ Best approximation of the population is the sample itself
  ⟶ Let’s resample the only sample!

Source

Bootstrapping: Algorithm

1. Use the original sample to represent the population.
2. Take re-samples with replacement from the original sample m times
3. Use each re-sample to calculate the statistic of interest
4. This process produces a distribution of m statistics
5. Order this distribution incrementally

→ the middle (1-n)% of the resampling distribution is **a n%-confidence interval**



Source

Bootstrapping versus Classical approach: Coding exercise

Let’s generate a random sample of n = 10000 from **the normal distribution N(0,100)**.
Now let’s pretend that we don’t know which distribution this data comes from. We are interested to evaluate its expected value.

1. Calculate the point estimate of the mean (by LLN it is converges to the expected value)
2. Calculate the 5-% confidence interval using **the classic/inferential approach**
3. Calculate it using **bootstrap** with m = 100, 1000, 10000
   a. Make a function given m that returns the confidence interval as a list of two elements
   b. What happens to the bootstrap confidence interval as m increases?
4. Repeat the same for n=10000 and **the uniform distribution** U[-50,50]
5. For which distribution the classic method seems to be working better? For which - the bootstrap? Why do you think it is this way?

Go to the helper notebook for performing an exercise

Part 4: Two *philosophical* approaches to probability

Frequentist

- *One and only* probability of something does exist
- The parameters of the distribution are considered **fixed**
- I.e probability describes relative frequency as the limit.

$$\mathbf{p} = \frac{k}{n}$$

- Increasing n makes the parameter estimate to converge

$$p = \lim_{n\to\infty} \frac{k}{n}$$

Bayesian

- One and only probability of something does not exist (since it is not provable)
- The parameters of the distribution are considered **random** and has their distribution that is refined with the new data
- Probability describes our uncertainty about the world
- Probability only reflects our knowledge

Bayes’ Rule for Conditional Probability

Bayes’ Rule

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

- A and B are events
- P(B) is not 0

P(A) and P(B) are the unconditional probabilities of A and B occurring
P(A|B) is the conditional probability of A occurring, given that B occurs
P(B|A) is the conditional probability of B occurring, given that A occurs

$P(B) = P(B|A)*P(A) + P(B|A^{-})*P(A^{-})$
$P(A|B) = 1 - P(A^{-}|B)$

Posterior update based on the prior

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

Treating the data y as fixed,

$$p(\theta|y) \propto L(\theta)p(\theta)$$

Credit: Joe Blitzstein

Bayes formula when data is involved

Prob(Data | Model) - **likelihood term**

Prob(Data) - Sum of probabilities of the data across all possible parameter values

Prob(Model) - **Prior**

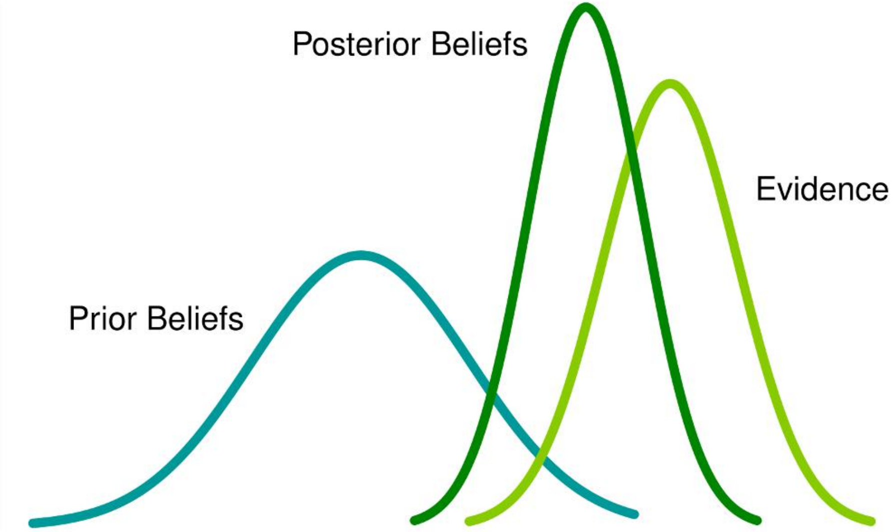
Prob(Model | Data) - **Posterior**

$$Prob(Model \mid Data) \;=\; \frac{Prob(Data \mid Model) \;\cdot\; Prob(Model))}{Prob(Data)}$$

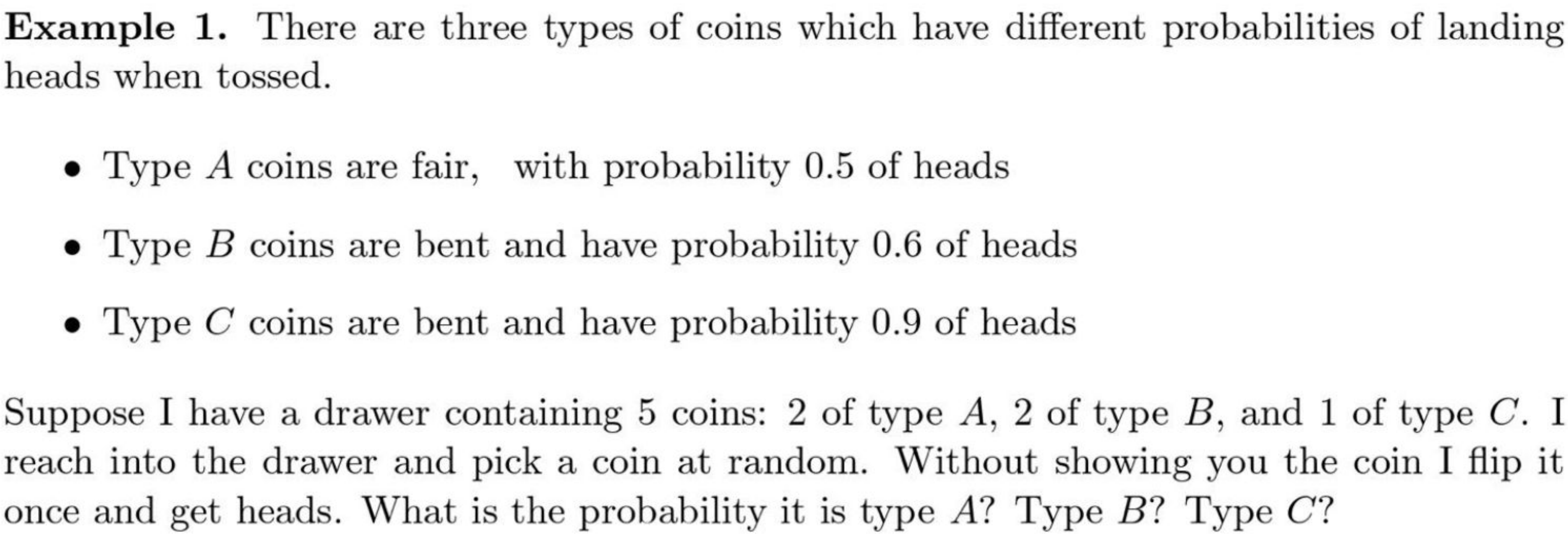
Bayesian statistics

new understanding = prior understanding + evidence

posterior distribution = prior distribution + data



Bayesian statistics



Example 1. There are three types of coins which have different probabilities of landing heads when tossed.

- Type A coins are fair, with probability 0.5 of heads
- Type B coins are bent and have probability 0.6 of heads
- Type C coins are bent and have probability 0.9 of heads

Suppose I have a drawer containing 5 coins: 2 of type A, 2 of type B, and 1 of type C. I reach into the drawer and pick a coin at random. Without showing you the coin I flip it once and get heads. What is the probability it is type A? Type B? Type C?

Before applying Bayes' theorem, let's introduce some terminology.

- Experiment: pick a coin from the drawer at random, flip it, and record the result.

- Data: the result of our experiment. In this case the event D = 'heads'. We think of D as data that provides evidence for or against each hypothesis.

- Hypotheses: we are testing three hypotheses: the coin is type A, B or C.

- Prior probability: the probability of each hypothesis prior to tossing the coin (collecting data). Since the drawer has 2 coins of type A, 2 of type B and 1 or type C we have

$$P(A) = 0.4, \qquad P(B) = 0.4, \qquad P(C) = 0.2.$$

- Likelihood: (This is the same likelihood we used for the MLE.) The likelihood function is P(D|H), i.e., the probability of the data assuming that the hypothesis is true. Most often we will consider the data as fixed and let the hypothesis vary.

For example, P(D|A) = probability of heads if the coin is type A. In our case the likelihoods are

$$P(D|A) = 0.5, \qquad P(D|B) = 0.6, \qquad P(D|C) = 0.9.$$

The name likelihood is so well established in the literature that we have to teach it to you. However in colloquial language likelihood and probability are synonyms. This leads to the likelihood function often being confused with the probability of a hypothesis. Because of this we'd prefer to use the name Bayes' term. However since we are stuck with 'likelihood' we will try to use it very carefully and in a way that minimizes any confusion.

- Posterior probability: the probability (posterior to) of each hypothesis given the data from tossing the coin.

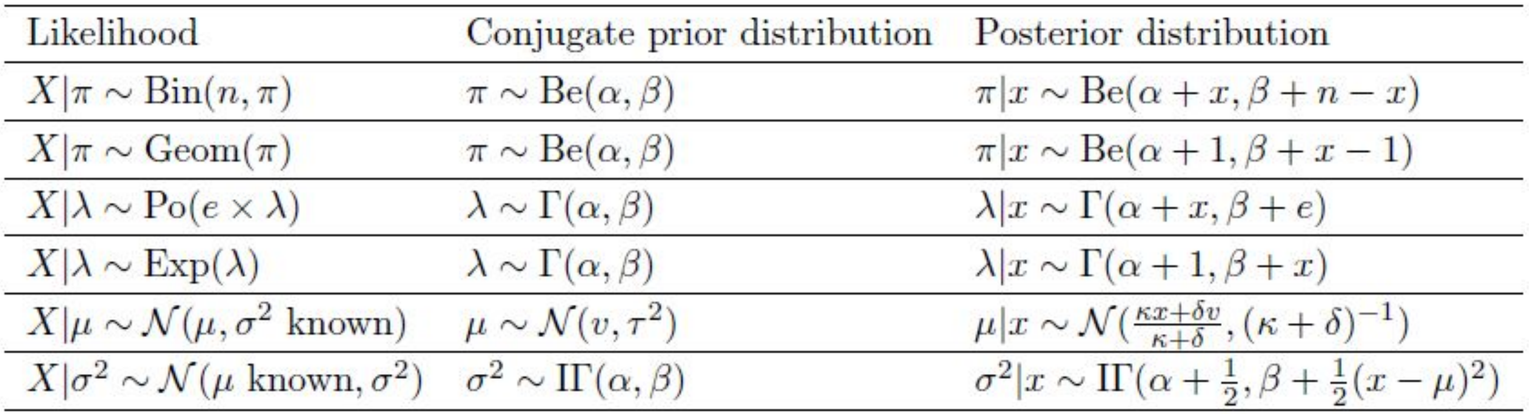
$$P(A|D), \qquad P(B|D), \qquad P(C|D).$$

These posterior probabilities are what the problem asks us to find.

How should a prior be chosen?

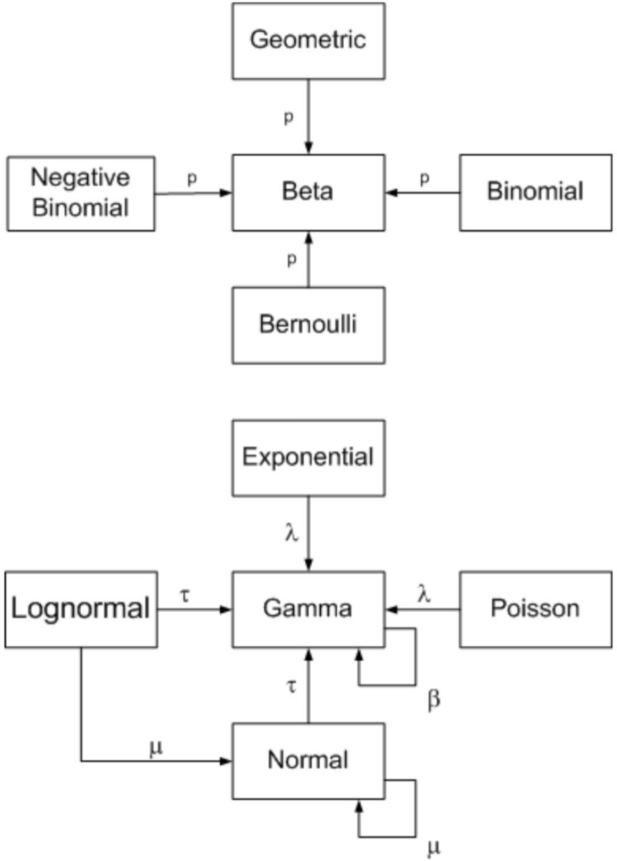
Conjugate Priors

Definition: A family G of probability distributions is conjugate with respect to the likelihood function if the posterior distribution is in the same family G for every f(θ) within G.



Likelihood	Conjugate prior distribution	Posterior distribution
$X\|\pi \sim \mathrm{Bin}(n, \pi)$	$\pi \sim \mathrm{Be}(\alpha, \beta)$	$\pi\|x \sim \mathrm{Be}(\alpha + x, \beta + n - x)$
$X\|\pi \sim \mathrm{Geom}(\pi)$	$\pi \sim \mathrm{Be}(\alpha, \beta)$	$\pi\|x \sim \mathrm{Be}(\alpha + 1, \beta + x - 1)$
$X\|\lambda \sim \mathrm{Po}(e \times \lambda)$	$\lambda \sim \Gamma(\alpha, \beta)$	$\lambda\|x \sim \Gamma(\alpha + x, \beta + e)$
$X\|\lambda \sim \mathrm{Exp}(\lambda)$	$\lambda \sim \Gamma(\alpha, \beta)$	$\lambda\|x \sim \Gamma(\alpha + 1, \beta + x)$
$X\|\mu \sim \mathcal{N}(\mu, \sigma^2 \text{ known})$	$\mu \sim \mathcal{N}(v, \tau^2)$	$\mu\|x \sim \mathcal{N}(\frac{\kappa x + \delta v}{\kappa + \delta}, (\kappa + \delta)^{-1})$
$X\|\sigma^2 \sim \mathcal{N}(\mu \text{ known}, \sigma^2)$	$\sigma^2 \sim \mathrm{I}\Gamma(\alpha, \beta)$	$\sigma^2\|x \sim \mathrm{I}\Gamma(\alpha + \frac{1}{2}, \beta + \frac{1}{2}(x - \mu)^2)$

Conjugate Priors



source

Choosing a prior

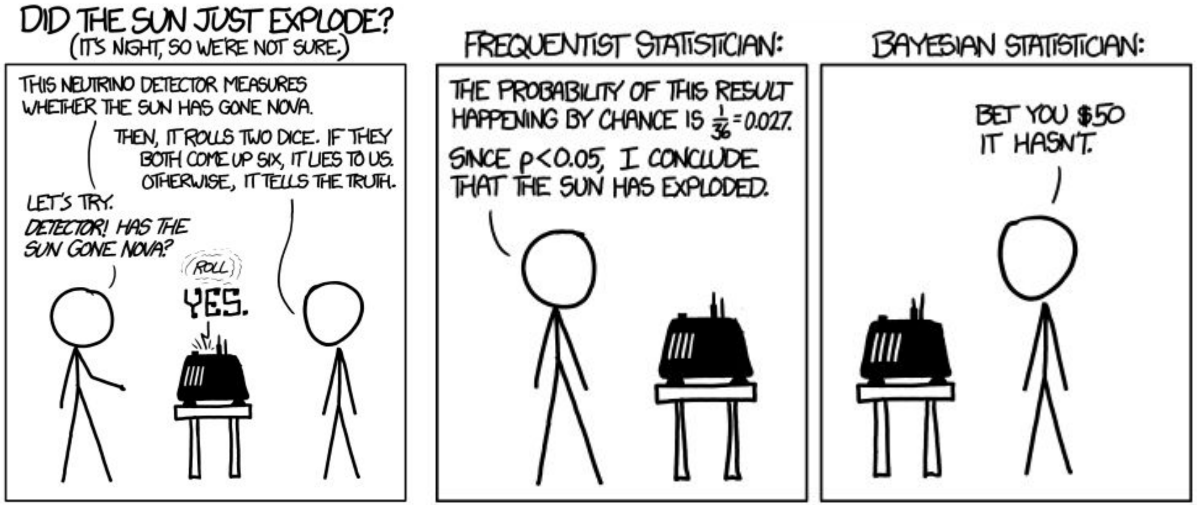
- Should not be too restrictive nor too broad.
- Available prior information may not be precise enough to determine a single prior distribution.
- Several distributions may be consistent with the prior information
- No unique way of choosing.

- Non-informative or dummy prior:
  - No prior information is available.
  - Gives every value of the parameter the same likelihood.
  - Minimizes the influence of the prior.

Why did we learn all this?

- Bayesian statistics is picking up in the data science field (especially medical and finance)
- Bayesian reasoning can sometimes be more intuitive
- **Bayesian Neural Networks** are heavily used in medical image analytics
- Bayesian Hyper-Parameter Optimization (Machine Learning Week)

XKCD: Frequentists vs. Bayesians


DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY. DETECTOR! HAS THE SUN GONE NOVA?
ROLL
YES.
FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS $\frac{1}{36}=0.027$. SINCE $p<0.05$, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.

Source: https://xkcd.com/1132/

Construction of Estimators: Methods

- **Maximum likelihood approach:**
  Differential calculus to determine the maximum of the probability function of a number of sample parameters.

- **Method of Moments:**
  Equates values of sample moments (functions describing the parameter) to population moments.

- **Bayesian Methods:**
  Introducing a frequency function for the parameter being estimated.

Exercise

- Exercise 1: Maximum likelihood estimation of frog survival
- Exercise 2: Point estimates and confidence intervals
- Exercise 3: COVID Rt Estimation