CSC380: Principles of Data Science

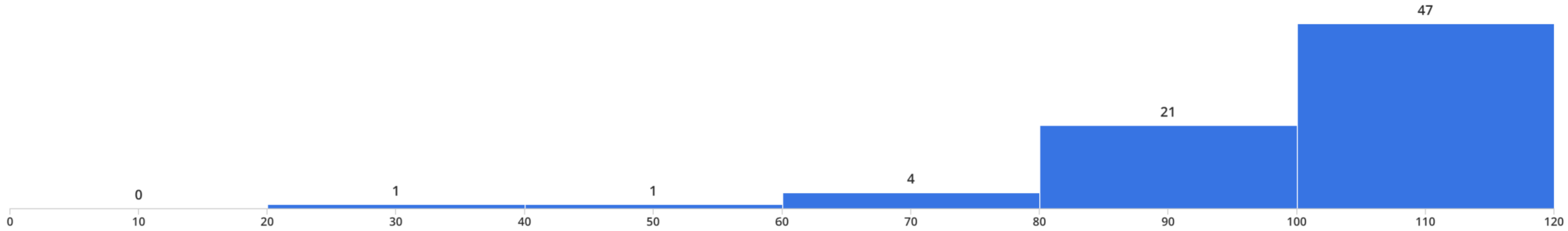
Midterm review

Chicheng Zhang

credit: Jason Pacheco, Kwang-Sung Jun’s slides & Watkins, J. “Intro. to the Science of Statistics”

Summary statistics for HWs

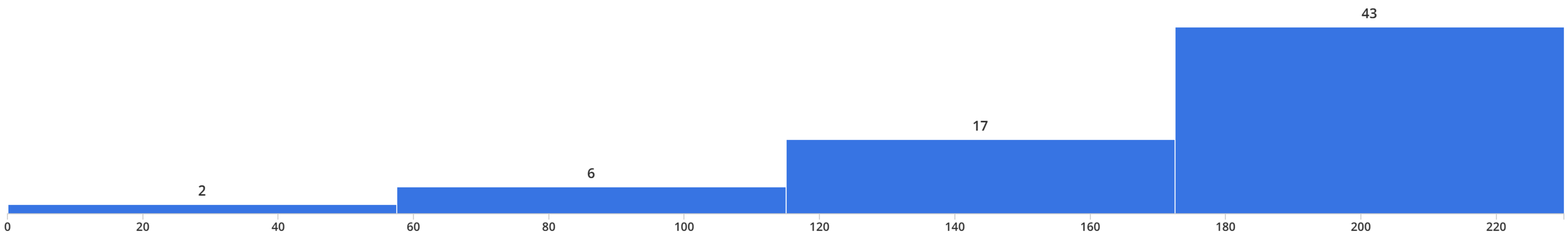
- HW1



Minimum	Median	Maximum	Mean	Std Dev
38.0	**109.5**	**120.0**	**102.78**	**16.62**

Summary statistics for HWs

- HW2



Minimum	Median	Maximum	Mean	Std Dev
40.0	**184.25**	**230.0**	**173.29**	**45.61**

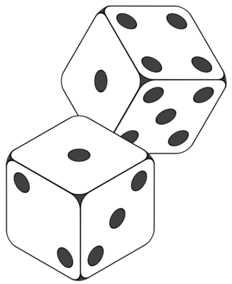
General suggestions on midterm preparation

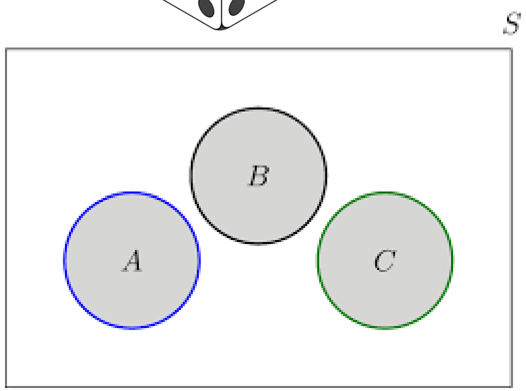
- Prioritize reviewing basic concepts & ideas
  - Remember, you can bring a cheatsheet with necessary technical details
  - Understand the motivations of concepts
  - “Math should be there to aid understanding, not hinder it.” – Hal Daume III

- “Memorization with understanding”

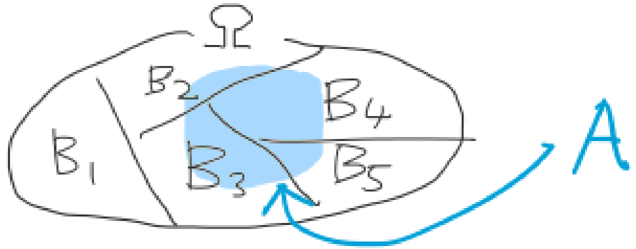
- Try to solve these on your own, then discuss with classmates
  - “Quiz Candidate” questions
  - Sample midterm
  - HW questions (esp. if you did not get them right the first time)

Probability



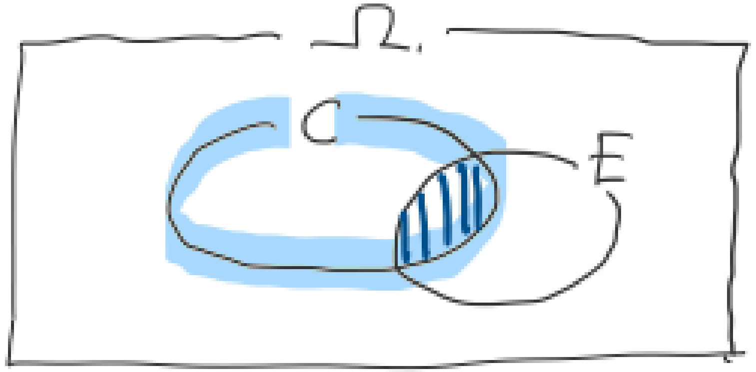


- Basic definitions: outcome space, events
- Probability P: maps events to [0,1] values
  - Three axioms
  - Axiom 3: additivity
- Special case of P: all outcomes are equally likely
  - Happens when we flip a fair coin, roll two fair dice, etc
  - Does *not* apply when we flip unfair coins, etc



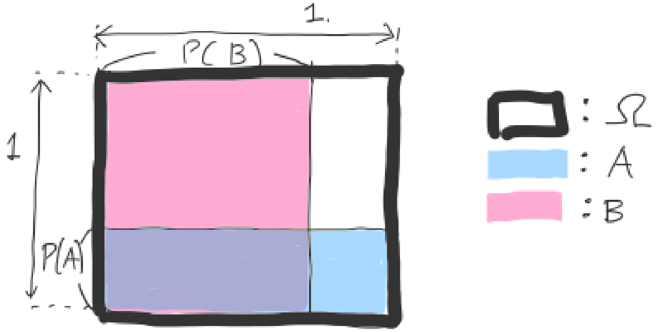
- Calculating probability: Inclusion-exclusion rule; law of total probability

Probability



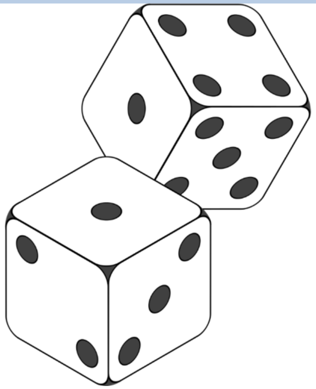
- Conditional probability $P(E|C) := \frac{P(E \cap C)}{P(C)}$
  - Is $P(A|B) \neq P(B|A)$ in general?

- Properties: chain rule, law of total probability, Bayes rule
  - Important application: medical diagnosis
  - Approach: write down the joint probability table



- Independence of events: $P(A, B) = P(A)P(B)$

Probability

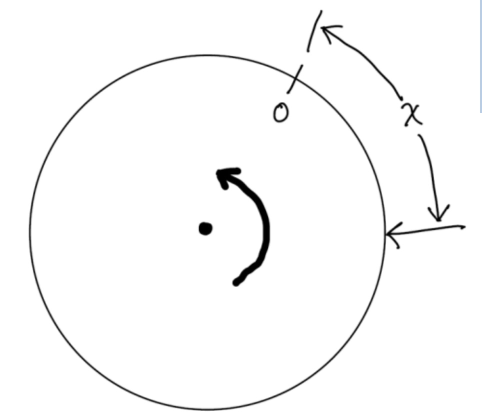


- Discrete random variable (RV) X (e.g. outcome of a die roll)
  $\{X = x\}$ is an event
- Representation of its distn: probability mass function (PMF)
  - Tabular representation of joint distribution of two RV's (X, Y)

value	prob.
1	0.2
2	0.8

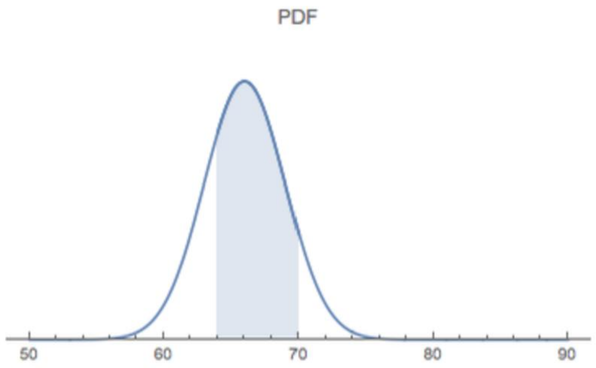
- Law of total probability, conditional probability, chain rule, Bayes rule on RVs
- Independence and conditional independence of RV's
- Useful discrete distns: uniform, Bernoulli, binomial, multinomial

Probability



- Continuous RVs X: $\mathrm{P}(X = x) = 0$ for any x

- Key representation tool for distn: p, probability density function (PDF)

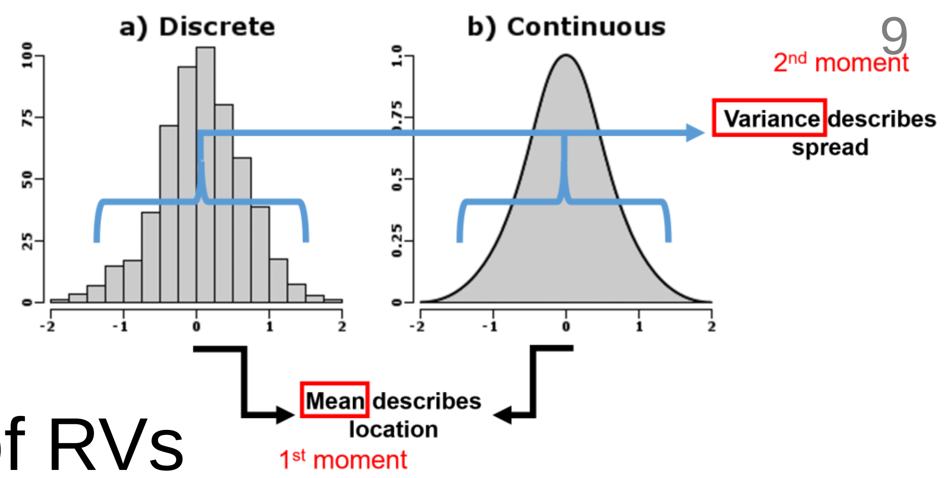


$P(a < X \leq b) = \int_a^b p(x)\mathrm{d}x$: area under the PDF

- Useful continuous distns and their PDFs:
  - Uniform, exponential, Gaussian (important properties)

- Cumulative density functions (CDF): $F(t) = P(X \leq t)$
  - Well-defined for discrete & continuous RVs (its shape in respective settings?)

Probability



- Moments of random variables

- Calculate mean (expectation) and variance of RVs
- (Very useful) Expectation value formula:

$$\mathrm{E}[f(Z)] = \sum_z f(z) P(Z = z)$$

  - Also applies for $Z = (X, Y)$
- Linearity of expectation: $\mathrm{E}[X + cY] = \mathrm{E}[X] + c\mathrm{E}[Y]$ for constant c
  - What about $\mathrm{Var}[X + cY]$?

- Moments of useful distns: Bernoulli, binomial, Gaussian

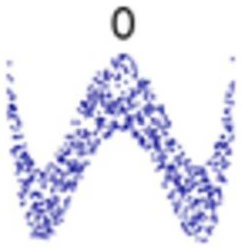
Probability

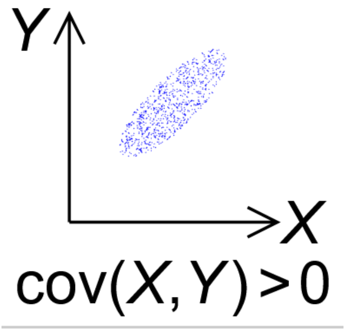
- A “mixed” 2nd moment: covariance

$$\mathrm{Cov}(X,Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$$

- Measures *linear relationship* between X, Y

$$\mathrm{Cov}(X,Y) = 0 \nRightarrow X \perp Y$$





- Pearson correlation: $\rho = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$, where $\sigma_X = \sqrt{\mathrm{Var}(X)}$

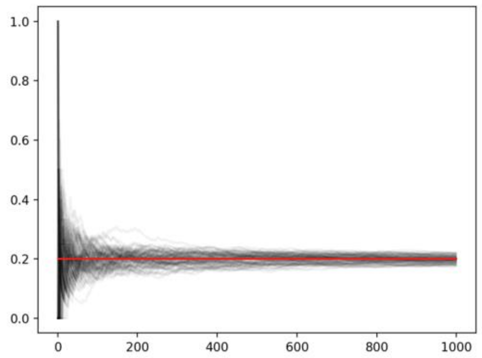
- Important property: $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y)$
  - What if X, Y are independent?

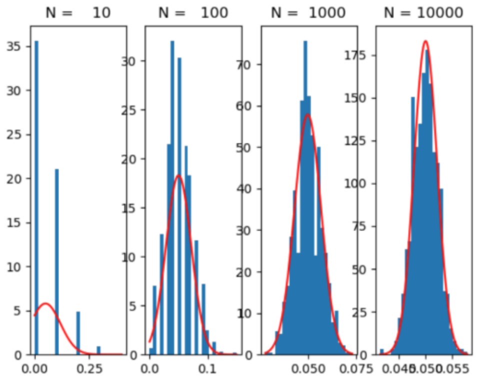
Statistics

- Statistics: make statements about data generation process based on data seen; reverse engineering

- Point estimation
  - Given iid samples $X_1, \ldots, X_n \sim \mathcal{D}_\theta$, estimate θ by constructing *statistics* $\hat{\theta}_n$
  - Basic estimators: sample mean, sample variance
  - Performance measures: unbiasedness, consistency, MSE (efficiency)
  - Bias-variance decomposition:
    - $\mathrm{MSE}(\hat{\theta}) = \mathrm{bias}(\hat{\theta})^2 + \mathrm{var}(\hat{\theta})$

- Useful probability tools:
  - Law of Large Numbers
  - Central Limit Theorem

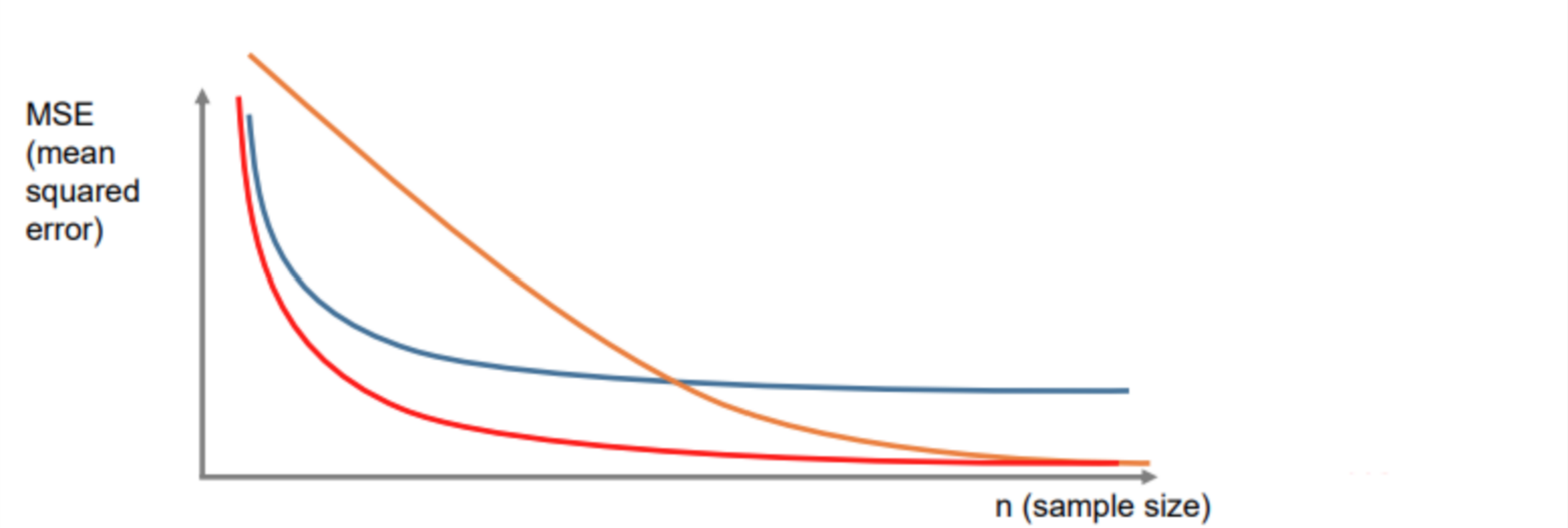




Piazza review questions this week

Resolved Unresolved @112_f1

Nam Nhat Do 2 days ago



Please label how consistent and efficient each line is (or rank them by how consistent/efficient they are). Here are the definitions:

➢ Consistency (asymptotic notion): Given enough data, the estimator converges to the true parameter value

➢ Efficiency (nonasymptotic notion): It should have low error with finite n

Piazza review questions this week

Resolved Unresolved @112_f6 Actions

Sean Eddy 23 hours ago

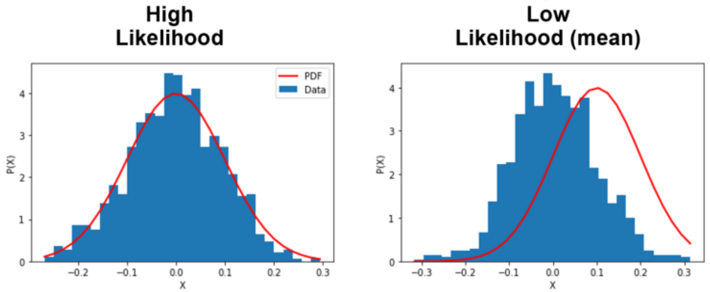
I would appreciate some further clarification on the bias-variance tradeoff. I understand the incentives to limit bias and variance in a model, but why can't both be easily limited simultaneously? I understand the target analogy comparing bias-variance to accuracy and precision, but this analogy doesn't convey to me why there is a tradeoff (how does increasing accuracy hurt precision?).

- Bias-variance tradeoff happens when we fix the dataset and vary estimators
- An elementary example:
  - $X_1, \dots, X_n \sim \mathcal{D}$ with population mean μ
  - $\hat{\mu} = \lambda \cdot \frac{1}{n} \sum_{i=1}^{n} X_i$
    - $\lambda = 0$ => high bias, zero variance
    - $\lambda = 1$ => zero bias, high variance
- Recommended video: [The weirdest paradox in statistics (and machine learning)]
- Revisit Faraz's answer after lecture "predictive modeling"

- Maximum likelihood (MLE): a general approach for point estimation

- Given $X_1, \dots, X_n \sim \mathcal{D}_{\theta^*}$, estimate θ^* by finding the maximizer of the likelihood function

$$\mathcal{L}_n(\theta) = p(x_1, \dots, x_n; \theta) = p(x_1; \theta) \cdot \dots \cdot p(x_n; \theta)$$



- Intuition: $\mathcal{L}_n(\theta)$ measures the "goodness of fit" of $\mathcal{D}_\theta$ to data $x_1, \dots, x_n$

- $\mathcal{D}_\theta$ can be general, e.g. Bernoulli, Gaussian, Poisson (in HW3)

- MLE in action
- E.g. HW3, P3

$$\text{Poisson}(x; \lambda) = \frac{1}{x!} \lambda^x e^{-\lambda}.$$

During my last three office hours I received $X_1 = 5, X_2 = 6, X_3 = 8$ students.

1. Write down the (log)-likelihood function

$$\log L_n(\lambda) = \log \left(\prod_{i=1}^{n} p(x_i) \right)$$

$$= \sum_{i=1}^{n} \log \left(\frac{1}{x_i!} \lambda^{x_i} e^{-\lambda} \right) = -\sum_{i=1}^{n} \log(x_i!) + \log(\lambda) \sum_{i=1}^{n} x_i - n\lambda$$

2. Find the parameter that maximizes the likelihood

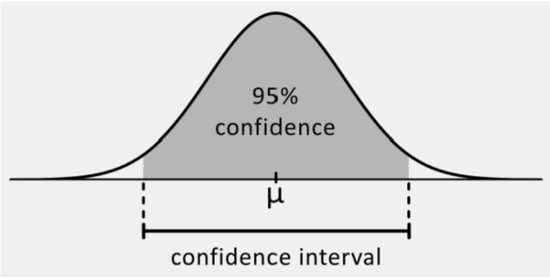
$$\Rightarrow \lambda^{MLE} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Statistics

- Confidence interval (interval estimation)

- Definition of confidence intervals:
  - Given data $X_1, \dots, X_n \sim \mathcal{D}_\theta$ with unknown θ (say, $\mathcal{D}_\theta = \mathcal{N}(\theta, 1)$)
  - Construct a_n, b_n (that depends on $X_1, \dots, X_n$), such that

$$P(\theta \in [a_n, b_n]) \geq 1 - \alpha$$



- Confidence intervals for population mean: Gaussian(naive), Gaussian(corrected)
  - We expect you to understand how they are computed

- Confidence intervals for general population properties: bootstrap

Piazza review questions this week

Resolved Unresolved @112_f2

Actions

Nam Nhat Do 2 days ago

The question above was from Week 5. Sorry! Here is one from Week 6:

Given that v, the distribution of $\hat{\theta}_n - \theta$, is unknown, what method should we use to estimate the confidence interval for θ. What are the steps to do this method?

Piazza review questions this week

Resolved Unresolved @112_f4

Actions

Thomas Everitte 1 day ago

Why does the confidence interval not mean the probability a variable falling within the interval for an arbitrary distribution?

helpful! 0

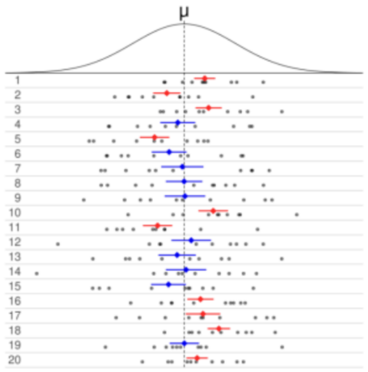
I am having difficulty understanding the justification for confidence intervals. What would a 95% confidence interval mean exactly? Is it that we can be 95% sure that it represents the entire set of data? Perhaps an example would help me understand better.

Piazza review questions this week

Ali Elbekov 1 day ago

Recommended point of view:

- Assume: Heights of UA students follow a normal distribution $\mathcal{N}(\mu, 1)$ with unknown μ
- Fork m **parallel universes**. For each universe $u \in \{1,2,\ldots,m\}$,
  - Subsample n UA students randomly, take the sample mean $\hat{\mu}^{(u)}$.
  - Compute the confidence bound $\left[\hat{\mu}^{(u)} - \frac{1.96\sigma}{\sqrt{n}}, \hat{\mu}^{(u)} + \frac{1.96\sigma}{\sqrt{n}}\right]$



- The fraction of parallel universes where the random interval includes μ is *approximately* at least 0.95 if m is large enough.
- As m goes to infinity, the fraction will become arbitrarily close to a value that is at least 0.95.

https://en.wikipedia.org/wiki/Confidence_interval

What does "fork m parallel universes" mean here?

Sayyed Faraz Mohseni 21 hours ago

It's another way of saying that you take m samples of n random students from the UA with replacements.

Piazza review questions this week

Resolved Unresolved @112_f3

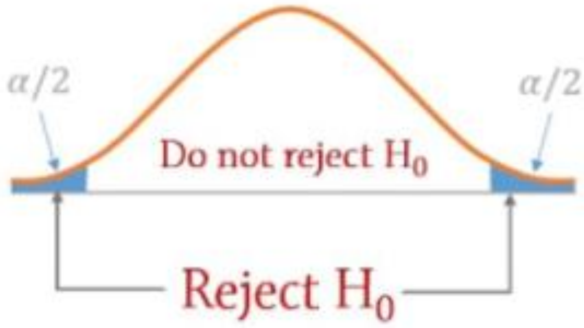
Actions

Henry Einhaus 1 day ago

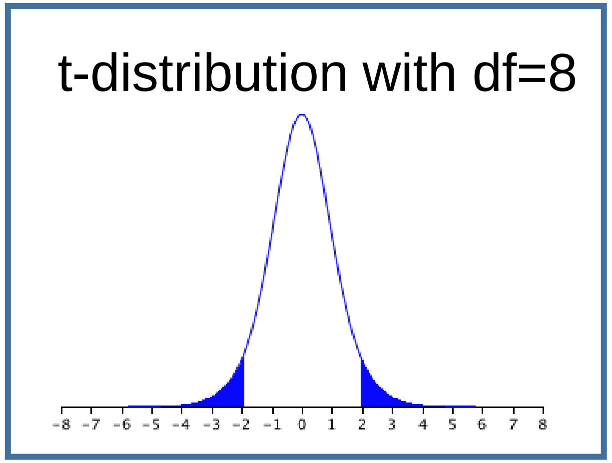
My question is related to Gaussians and arbitrary distribution testing. I know that in class we covered a couple different ways to go about this, (Gaussian(Naïve), Gaussian(Corrected), and bootstrap) but I am not completing sure when we should be using these methods, and how to know the steps to take to figure out what method is best.

- First, bootstrap applies to *general* interval estimation beyond population mean
  - Drawback: computational cost
- For population mean:
  - Gaussian(Corrected) is always preferred than Gaussian (Naive)
  - But for large sample size N, Gaussian(Naive) is almost equivalent to Gaussian(Corrected) (can you see why?)
  - https://math.stackexchange.com/questions/1350635/when-do-i-use-a-z-score-vs-a-t-score-for-confidence-intervals

- Hypothesis testing
  - Given dataset S
  - Goal: decide whether the data distribution satisfies:
    - H_0: null hypothesis
    - H_1: alternative hypothesis



- Paired t-test
  - Applications?
  - What satisfies the t-distribution, under what setting?
  - What is the test?



Data analysis

- Types of data:
  - Qualitative: nominal and ordinal
  - Quantitative: discrete and continuous

- Summary statistics
  - Median