Computer Science

CSC380: Principles of Data Science

Statistics 5 & Midterm review

Xinchen Yu

Announcements

- HW5 has been out.
  - Due Friday, Mar 15

- Practice problems will be out by end of this weekend.
  - Solutions will be out by Mar 10.

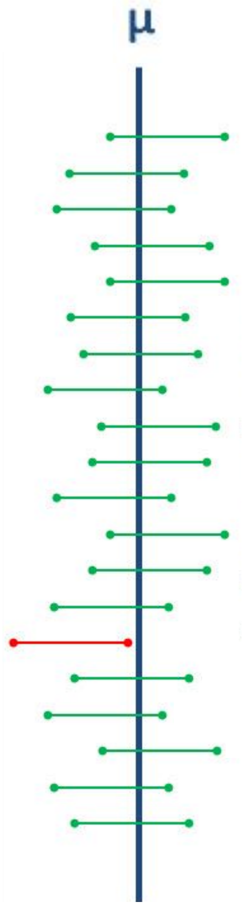
- Lecture on Tuesday Mar 12:
  - Another review session
  - revisit solutions of some questions in HW1 - 4
  - Q & A

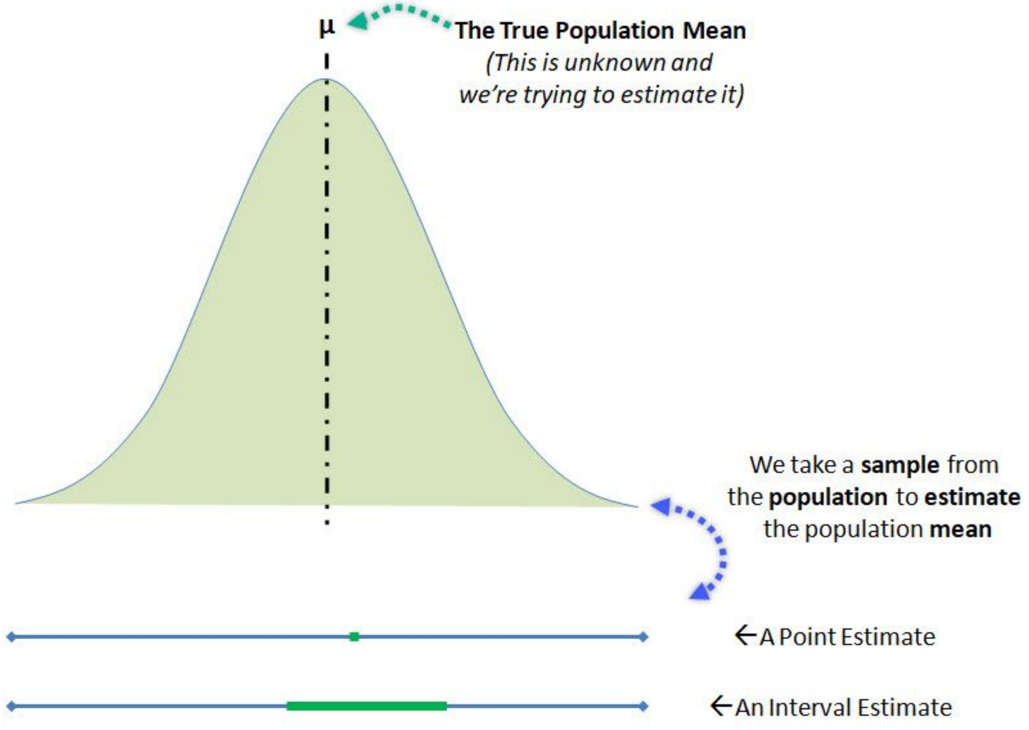
Announcements

- Midterm
  - What you can bring?
    - Cheat sheet: letter size, double-sided
    - Scientific calculator
  - Time: Mar 14, Thursday, 3:30-4:45 pm
  - Location: same as lecture room

$\frac{\sqrt{2}}{3}$ is ok

Review: Interval estimate





Review: Gaussian (Corrected)

Suppose $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ with unknown μ & known σ^2.

<u>(Fact 1)</u> $\hat{\mu} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ $\quad \sqrt{n}\frac{\hat{\mu}-\mu}{\sigma} \sim N(0,1)$

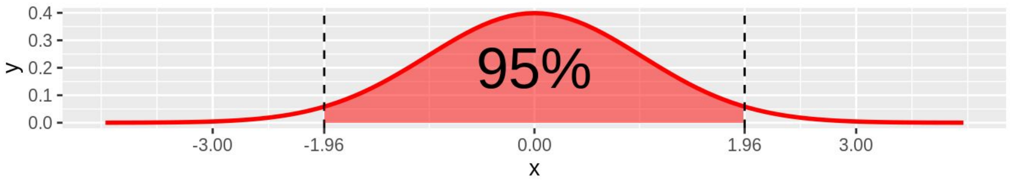
<u>(Fact 2)</u> If $Z \sim \mathcal{N}(0,1)$,

$$P(Z \in [-z, z]) = 1 - 2(1 - \Phi(z))$$

where $\Phi(z) := \mathrm{P}(Z \leq z)$ is the CDF of Z.

z = 1.96: RHS ≈ .95, 95% confident

z = 2.58: RHS ≈ .99,



$$\phi(1.96) = P(Z \in [-\infty, 1.96]) = 0.975$$
$$P(Z \in [-1.96, 1.96]) = 1 - 2 \cdot (1 - 0.975) = 0.95$$

Let: $Z \longrightarrow \sqrt{n}\,\frac{\hat{\mu} - \mu}{\sigma}$

$$P\left(\hat{\mu} \in \left[\mu - \frac{1.96\sigma}{\sqrt{n}}, \mu + \frac{1.96\sigma}{\sqrt{n}}\right]\right) \geq 0.95$$

$$P\left(\hat{\mu} \in \left[\mu - \frac{2.58\sigma}{\sqrt{n}}, \mu + \frac{2.58\sigma}{\sqrt{n}}\right]\right) \geq 0.99$$

=> Compute $\left[\hat{\mu} - \frac{1.96\sigma}{\sqrt{n}}, \hat{\mu} + \frac{1.96\sigma}{\sqrt{n}}\right]$. Done!

Review: Gaussian (Corrected)

Suppose $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ with unknown μ & known σ^2.

<u>(Fact 1)</u> $\hat{\mu} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ $\sqrt{n}\frac{\hat{\mu}-\mu}{\boxed{\sigma}} \sim \boxed{N(0,1)}$ → T-dist

$\hat{\sigma}$

<u>(Fact 2)</u> If $Z \sim \mathcal{N}(0,1)$,

$$P(Z \in [-z, z]) = 1 - 2(1 - \Phi(z))$$

where $\Phi(z) := \mathrm{P}(\mathrm{Z} \le z)$ is the CDF of Z.

z = 1.96: RHS ≈ .95, 95% confident

z = 2.58: RHS ≈ .99,

Let: $Z \longrightarrow \sqrt{n}\frac{\hat{\mu}-\mu}{\sigma}$

$$P\left(\hat{\mu} \in \left[\mu - \frac{1.96\sigma}{\sqrt{n}}, \mu + \frac{1.96\sigma}{\sqrt{n}}\right]\right) \ge 0.95$$

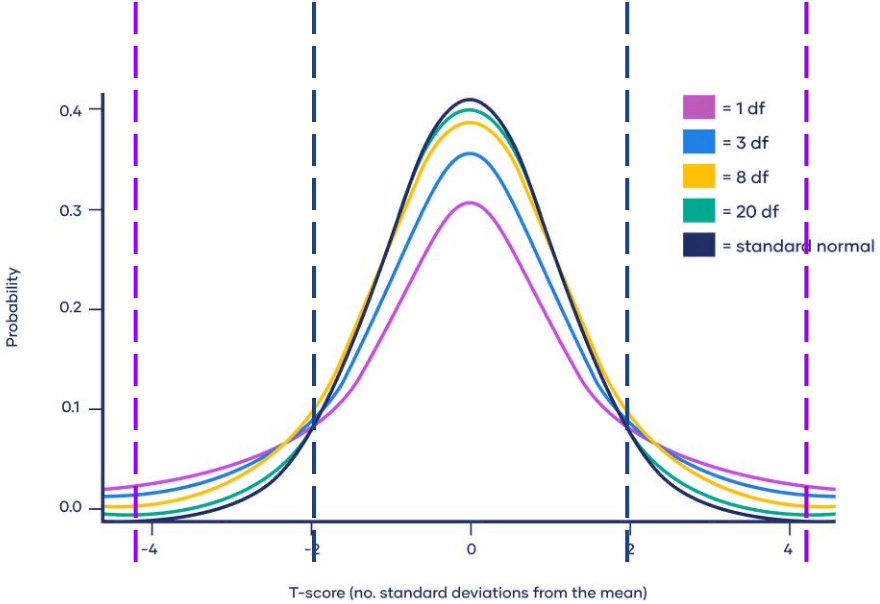
$$P\left(\hat{\mu} \in \left[\mu - \frac{2.58\sigma}{\sqrt{n}}, \mu + \frac{2.58\sigma}{\sqrt{n}}\right]\right) \ge 0.99$$

=> Compute $\left[\hat{\mu} - \frac{1.96\sigma}{\sqrt{n}}, \hat{\mu} + \frac{1.96\sigma}{\sqrt{n}}\right]$. Done!

Q: what if X from an arbitrary distribution (e.g. uniform)?

Q: what if σ^2 is unknown and sample size is small (< 30)?

Z score versus T score



When alpha is 0.05:

- For standard normal distribution:

$$P(X \in [-1.96,\ 1.96]) = 0.95$$

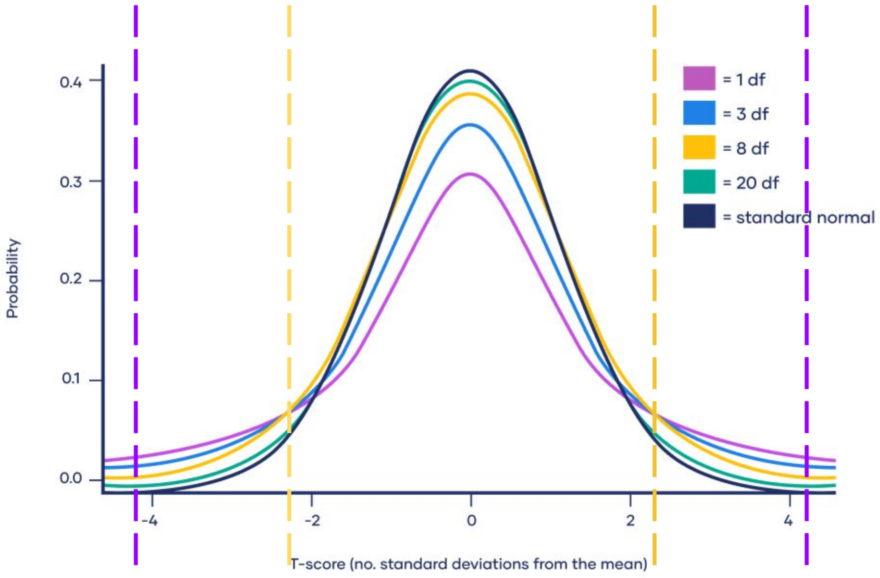
- For T distribution when n = 2:

$$P(X \in [-4.30,\ 4.30]) = 0.95$$

Review: T scores for different df

Let's compare t scores when we only have 3 and 6 observations in the sample:

$$\left[\hat{\mu} - t_{\alpha/2,n-1}\frac{\hat{\sigma}}{\sqrt{n}}, \hat{\mu} + t_{\alpha/2,n-1}\frac{\hat{\sigma}}{\sqrt{n}}\right]$$



(recall: 1.96 for gaussian)

```
import scipy.stats as st
alpha = 0.05
st.t.ppf(1-alpha/2,df=2)
=> 4.302652729911275

st.t.ppf(1-alpha/2,df=5)
=> 2.5705818366147395

st.t.ppf(1-alpha/2,df=10)
=> 2.2281388519649385

st.t.ppf(1-alpha/2,df=30)
=> 2.0422724563012373

st.t.ppf(1-alpha/2,df=100)
=> 1.9839715184496334
```

Method 2: Bootstrap

Suppose $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ with unknown μ & known σ^2.

(Fact 1) $\hat{\mu} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ $\sqrt{n}\frac{\boxed{\hat{\mu}-\mu}}{\sigma} \sim \mathcal{N}(0,1)$

Directly approximate distributions of $\hat{\mu} - \mu$

(Fact 2) If $Z \sim \mathcal{N}(0,1)$,

$$P(Z \in [-z, z]) = 1 - 2(1 - \Phi(z))$$

where $\Phi(z) \coloneqq \mathrm{P}(\mathrm{Z} \leq z)$ is the CDF of Z.

z = 1.96: RHS ≈ .95, 95% confident

z = 2.58: RHS ≈ .99,

Let: $Z \longrightarrow \sqrt{n}\frac{\hat{\mu}-\mu}{\sigma}$

$$P\left(\hat{\mu} \in \left[\mu - \frac{1.96\sigma}{\sqrt{n}}, \mu + \frac{1.96\sigma}{\sqrt{n}}\right]\right) \geq 0.95$$

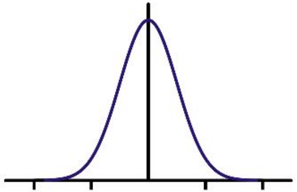
$$P\left(\hat{\mu} \in \left[\mu - \frac{2.58\sigma}{\sqrt{n}}, \mu + \frac{2.58\sigma}{\sqrt{n}}\right]\right) \geq 0.99$$

=> Compute $\left[\hat{\mu} - \frac{1.96\sigma}{\sqrt{n}}, \hat{\mu} + \frac{1.96\sigma}{\sqrt{n}}\right]$. Done!

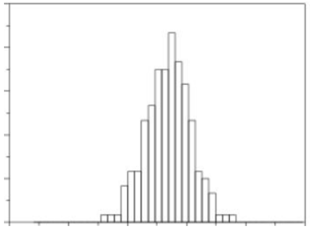
Method 2: Bootstrap

- Key idea: approximate ν, the distribution of $\hat{\theta}_n - \theta$
- Insight:

θ

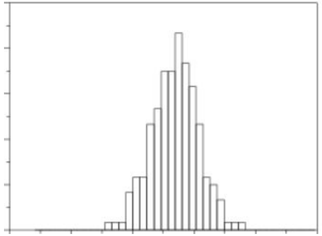


$$D_\theta \xrightarrow{n \text{ iid samples}} S$$

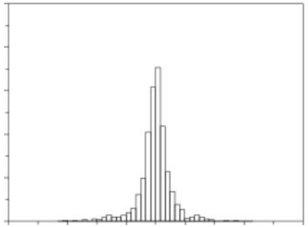


$\hat{\theta}_n$

$\hat{\theta}_n$



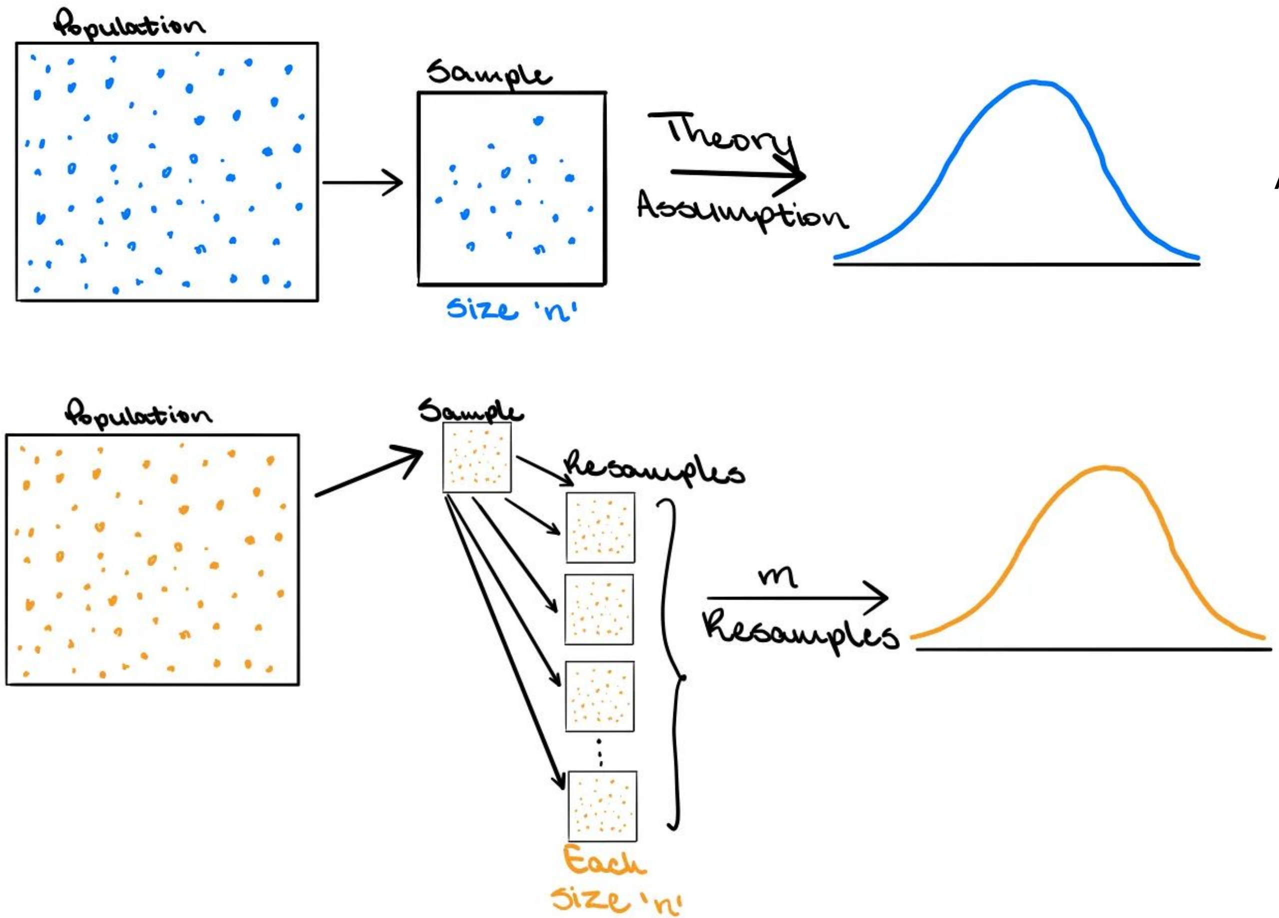
$$\text{Uniform}(S) \xrightarrow{n \text{ iid samples}} S_{\text{b}}$$



$\hat{\theta}_{n,b}$

- Use empirical distribution of $\hat{\theta}_{n,b} - \hat{\theta}_{\text{n}}$'s to approximate ν, obtaining approximations of $v_{\alpha/2}$ and $v_{1-\alpha/2}$
- This empirical distribution can be obtained by drawing multiple S_{b}'s (bootstrap subsample)

Method 2: Bootstrap



Approximate distribution

$\hat{\theta} - \theta$

$\{\hat{\theta}_{n,b} - \hat{\theta}\}_{b=1}^{B}$

Method 2: Bootstrap example

Sample data: 30, 37, 36, 43, 42, 43, 43, 46, 41, 42

Sample mean: $\overline{x} = 40.3$

We want to know the distribution of: $\delta = \overline{x} - \mu$

Can approximate the distribution: $\delta^* = \overline{x}^* - \overline{x}$

Let's resample data with same size and generate 20 bootstrap samples:

43	36	46	30	43	43	43	37	42	42	43	37	36	42	43	43	42	43	42	43
43	41	37	37	43	43	46	36	41	43	43	42	41	43	46	36	43	43	43	42
42	43	37	43	46	37	36	41	36	43	41	36	37	30	46	46	42	36	36	43
37	42	43	41	41	42	36	42	42	43	42	43	41	43	36	43	43	41	42	46
42	36	43	43	42	37	42	42	42	46	30	43	36	43	43	42	37	36	42	30
36	36	42	42	36	36	43	41	30	42	37	43	41	41	43	43	42	46	43	37
43	37	41	43	41	42	43	46	46	36	43	42	43	30	41	46	43	46	30	43
41	42	30	42	37	43	43	42	43	43	46	43	30	42	30	42	30	43	43	42
46	42	42	43	41	42	30	37	30	42	43	42	43	37	37	37	42	43	43	46
42	43	43	41	42	36	43	30	37	43	42	43	41	36	37	41	43	42	43	43

Method 2: Bootstrap example

43	36	46	30	43	43	43	37	42	42	43	37	36	42	43	43	42	43	42	43
43	41	37	37	43	43	46	36	41	43	43	42	41	43	46	36	43	43	43	42
42	43	37	43	46	37	36	41	36	43	41	36	37	30	46	46	42	36	36	43
37	42	43	41	41	42	36	42	42	43	42	43	41	43	36	43	43	41	42	46
42	36	43	43	42	37	42	42	42	46	30	43	36	43	43	42	37	36	42	30
36	36	42	42	36	36	43	41	30	42	37	43	41	41	43	43	42	46	43	37
43	37	41	43	41	42	43	46	46	36	43	42	43	30	41	46	43	46	30	43
41	42	30	42	37	43	43	42	43	43	46	43	30	42	30	42	30	43	43	42
46	42	42	43	41	42	30	37	30	42	43	42	43	37	37	37	42	43	43	46
42	43	43	41	42	36	43	30	37	43	42	43	41	36	37	41	43	42	43	43

Calculate sample mean for each column (bootstrap sample), compute: $\delta^* = \overline{x}^* - \overline{x}$

Sort the 20 differences:

$$-1.6, -1.4, -1.4, -0.9, -0.5, -0.2, -0.1, 0.1, 0.2, 0.2, 0.4, 0.4, 0.7, 0.9, 1.1, 1.2, 1.2, 1.6, 1.6, 2.0$$

If confidence level is 80%, find out top 10% and bottom 10%:

$$-1.6, \boxed{-1.4} -1.4, -0.9, -0.5, -0.2, -0.1, 0.1, 0.2, 0.2, 0.4, 0.4, 0.7, 0.9, 1.1, 1.2, 1.2, 1.6, \boxed{1.6,} 2.0$$

The bootstrap confidence interval is:

$$[\overline{x} - \delta^*_{.1},\ \overline{x} - \delta^*_{.9}] \ = \ [40.3 - 1.6,\ 40.3 + 1.4] \ = \ [38.7,\ 41.7]$$

Method 2: Bootstrap

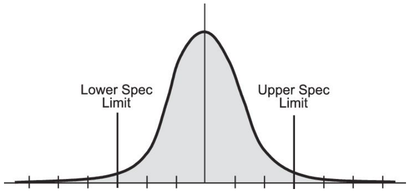
Suppose we observe data $X_1, X_2, \ldots, X_n \sim P(X;\theta)$:

1. Sample new "dataset" $X_1^*, \ldots, X_n^*$ uniformly from $X_1, \ldots, X_n$ **with replacement**

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

3. Repeat B times to get the estimators $\hat{\theta}_{n,1}, \ldots, \hat{\theta}_{n,B}$

4. Consider the **empirical distribution** of $\left\{\hat{\theta}_{n,b} - \frac{1}{n}\sum_{i=1}^{n} X_i\right\}_{b=1}^{B}$ and find its top $\frac{\alpha}{2}$ quantile and bottom $\frac{\alpha}{2}$ quantile (denoted by Q_U and Q_L respectively).



5. (1-α) Confidence Interval: $\left[\frac{1}{n}\sum_{i=1}^{n} X_i - |Q_U|, \frac{1}{n}\sum_{i=1}^{n} X_i + |Q_L|\right]$

counterintuitively, upper quantile for lower width, lower quantile for upper width. Why?

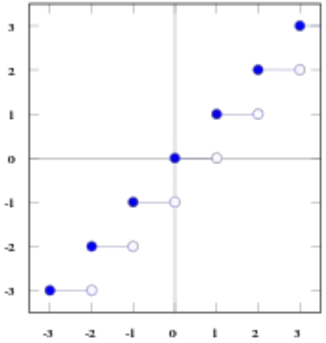
$$P\left(v_{\frac{\alpha}{2}} \leq \hat{\theta}_n - \theta \leq v_{1-\frac{\alpha}{2}}\right) \geq 1 - \alpha$$

Method 2: Bootstrap

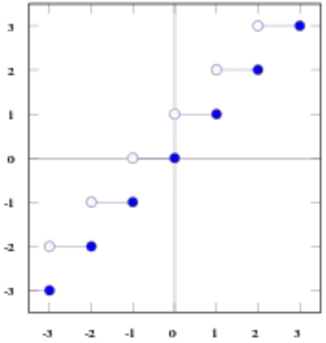
Pseudocode

Input: $X_1, \dots, X_n, B, \alpha$

- Compute $\bar{X}_n$
- Bootstrapping B times to obtain $\{\hat{\theta}_{n,b} - \bar{X}_n\}_{b=1}^{B}$; call this array S
- Sorted S in increasing order.
- $Q_U :=$ the top $\frac{\alpha}{2}$ quantile; i.e., `S[int(np.ceil( (1-alpha/2)*(B-1) ))]`
- $Q_L :=$ the bottom $\frac{\alpha}{2}$ quantile; i.e., `S[int(np.floor((  alpha/2)*(B-1) ))]`
- Return $[\bar{X}_n - |Q_U|, \bar{X}_n + |Q_L|]$



floor



ceil

Midterm Review

General tips on midterm preparation

- Prioritize reviewing basic concepts & ideas
- Understand the motivations and links between concepts
- “Memorization with understanding”

- Try to solve these on your own, then discuss with classmates
    - examples in the slides
    - HW questions (esp. if you did not get them right the first time)
    - practice problems

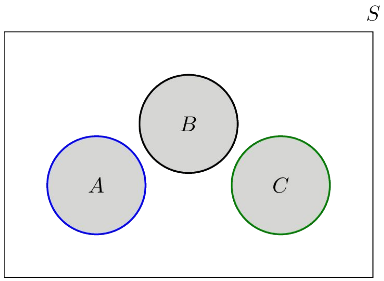
General tips on midterm preparation

- What will not included in the midterm?
  - Code related questions
  - Pure proof questions
    - But may need you to provide justifications

Probability

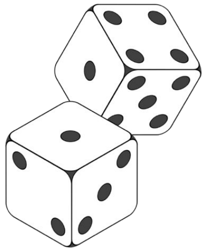
Probability

- Basic definitions: outcome space, events



- Probability P: maps events to [0, 1] values
    - Three axioms
    - Axiom 3: additivity

- Special case of P: each outcomes is equally likely

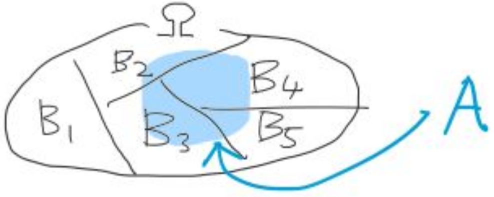


$$P(E) = \frac{|E|}{|\Omega|}$$

Number of elements in event set

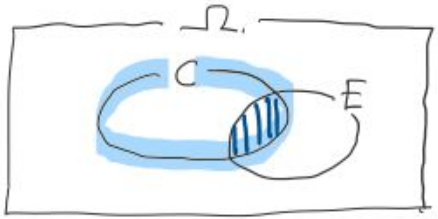
Number of possible outcomes (36)

- distributive law, inclusion-exclusion rule; law of total probability



Probability

- $P(E \cap C) = P(E|C)P(C) = P(C|E)P(E)$



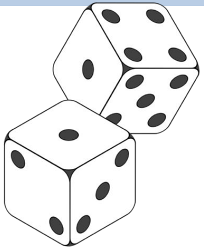
- Conditional probability
  - Chain rule, chain rule + law of total probability, bayes rule
  - Important application: medical diagnosis
  - Approach: write down the joint probability table

- Independence of events:

$$P(A, B) = P(A)P(B)$$

- Conditional / joint / marginal probability

Probability



- Discrete random variable X (e.g., sum of two dice)

- Representation of its distribution: probability mass function (PMF)
    - Tabular representation of joint distribution of 2 RVs (X,Y)

- RVs: law of total probability, conditional probability, chain rule, bayes rule, independence, conditional independence

- Useful discrete distributions
    - Uniform
    - Bernoulli
    - Binominal

Probability (this part will not be in midterm)

- Continuous random variable X: $P(X = x) = 0$ for any x

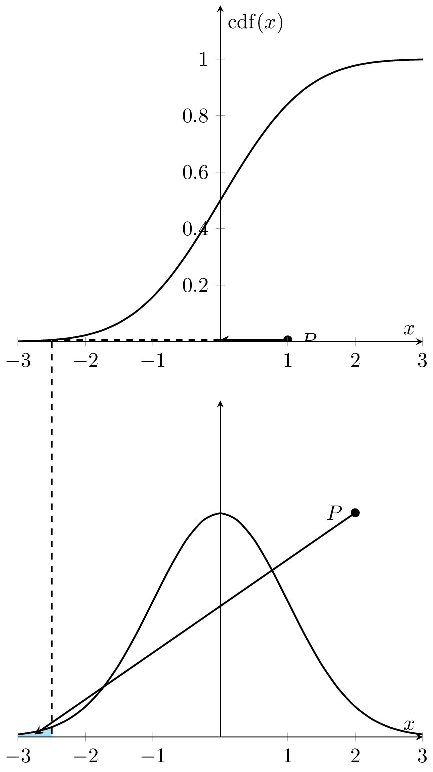
- Probability density function (PDF)

$$P(a < X \leq b) = \int_a^b p(x)\mathrm{d}x \quad p(x) = \frac{dF(x)}{dx}$$

- Cumulative distribution function (CDF)

$$P(a < X \leq b) = F(b) - F(a)$$

- Useful continuous distributions
  - Uniform
  - Gaussian (important properties)



Probability

- Moments of random variables: expectation, variance, covariance

- Calculate mean (expectation) and variance of RVs
  - Linearity of expectation: $E[X + cY] = E[X] + cE[Y]$ for constant c
  - $E[X^2]$
  - $E[XY]$
    - If independent: $E[X]E[Y]$
    - If not independent: $E[XY] = \sum_{(x,y)} \sum xy \cdot p(x, y)$
  - $E[X \mid Y = y]$
  - $\mathrm{Var}[c] = 0$
  - $\mathrm{Var}[cX]$
  - $\mathrm{Var}[X + c] = \mathrm{Var}[X]$
  - $\mathrm{Var}[X+Y]$ when independent

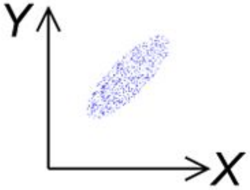
$$\boxed{\mathbf{Var}[X] = \mathbf{E}[(X - \mathbf{E}[X])^2]}$$

- Expectation and variance of useful distributions (esp. Bernoulli, Gaussian)

- Measures *linear relationship* between X, Y

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

0



cov(X,Y)>0

- 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?

Probability: example

Let random variable X and Y independent from each other.

The PMF for X is: $P(X = 0) = 0.5, \ P(X = 1) = 0.5$

The PMF for Y is: $P(Y = 0) = 0.25, \ P(Y = 1) = 0.5, \ P(Y = 2) = 0.25$

$E\left[XY^2\right]$?

$$P\left(XY^2 = 1\right) = P(X = 1, \ Y = 1) = 0.5 \cdot 0.5 = 0.25$$

$$P\left(XY^2 = 4\right) = P(X = 1, \ Y = 2) = 0.5 \cdot 0.25 = 0.125$$

$$P\left(XY^2 = 0\right) = 1 - 0.25 - 0.125 = 0.625$$

$$\begin{aligned} P\left(XY^2 = 0\right) &= P(X = 0, \ Y = 0) + P(X = 0, \ Y = 1) + P(X = 0, \ Y = 2) + P(X = 1, \ Y = 0) \\ &= 0.125 + 0.25 + 0.125 + 0.125 = 0.625 \end{aligned}$$

$$E\left[XY^2\right] = 1 \cdot 0.25 + 4 \cdot 0.125 + 0 \cdot 0.625 = 0.75$$

Probability: example

Let random variable X and Y independent from each other.

The PMF for X is: $P(X = 0) = 0.5,\ P(X = 1) = 0.5$

The PMF for Y is: $P(Y = 0) = 0.25,\ P(Y = 1) = 0.5,\ P(Y = 2) = 0.25$

$E\left[XY^2\right]$?

$$E\left[XY^2\right] = E[X] \cdot E\left[Y^2\right]$$

$$E[X] = 0.5$$

$$E\left[Y^2\right] = 0.25 \cdot 0^2 + 0.5 \cdot 1^2 + 0.25 \cdot 2^2 = 1.5$$

$$E\left[XY^2\right] = E[X] \cdot E\left[Y^2\right] = 0.5 \cdot 1.5 = 0.75$$

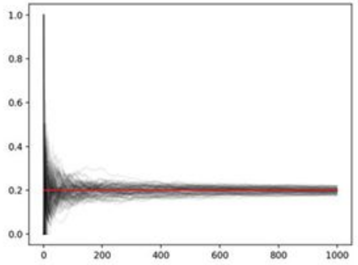
Statistics

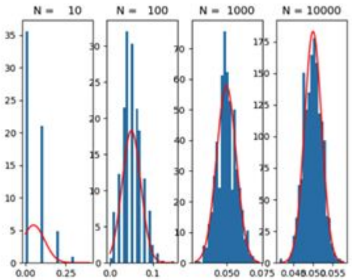
Statistics

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

- Point estimation
  - Given iid samples $X_1, \dots, X_n \sim \mathcal{D}_\theta$, 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



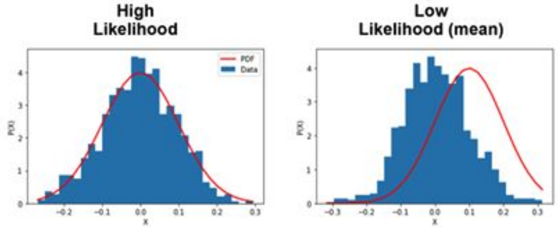


Statistics

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

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

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



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

Statistics

- Sample mean
  - Expectation (unbiased)
  - Variance

- Sample variance
  - biased version
  - unbiased version
  - Compare MSE of two versions

- How to determine an estimator is biased or unbiased?
  - statistics1, page 25; statistics3, page 9

Statistics

- MSE, Bias, Variance
  - how to calculate expectation and variance if there are more than 1 random variable -- use what we learned in probability lecture 5 & 6
  - Calculate bias and variance

$$
\begin{aligned}
\mathrm{MSE}(\hat{\theta}_n) &= \mathbf{E}[(\hat{\theta}_n - \theta)^2] \\
&= \left(\mathbf{E}[\hat{\theta}] - \theta\right)^2 + \mathbf{E}[(\hat{\theta} - \mathbf{E}[\hat{\theta}])^2] \\
&= \mathrm{bias}^2(\hat{\theta}) + \mathrm{Var}(\hat{\theta})
\end{aligned}
$$

Statistics

Important properties of Gaussian

- Closed under additivity:

$$X \sim \mathcal{N}(\mu_x, \sigma_x^2) \qquad Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$$

$$X + Y \sim \mathcal{N}(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$$

- Closed under affine transformation (a and b constant):

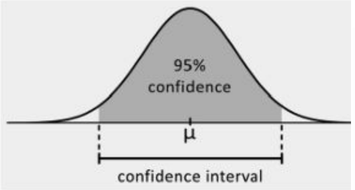
$$aX + b \sim \mathcal{N}(a\mu_x + b, a^2\sigma_x^2)$$

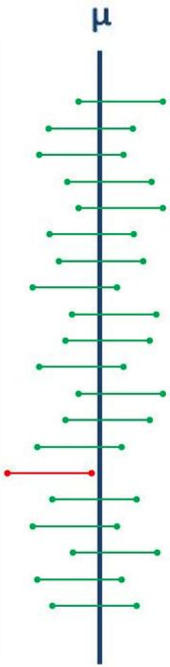
Statistics

- Confidence interval (interval estimation)

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

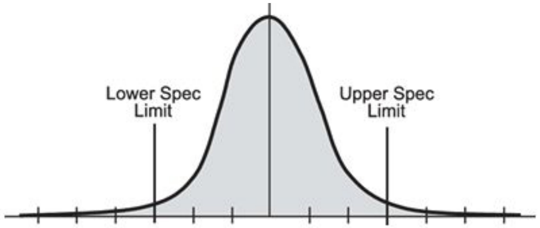
$$P(\theta \in [a_n, b_n]) \geq 1 - \alpha$$





- Interpretation: unless we are extremely unlucky (in that we encounter an unrepresentative dataset, which happens with prob. $\leq \alpha$), our confidence interval always contains the underlying parameter

Statistics



- Confidence intervals for population mean:
  - Gaussian(naive):
    $\left[\hat{\mu} - \frac{z_{1-\alpha/2}\hat{\sigma}}{\sqrt{n}}, \hat{\mu} + \frac{z_{1-\alpha/2}\hat{\sigma}}{\sqrt{n}}\right]$, $z_{1-\alpha/2} = 1 - \alpha/2$-quantile of $\mathcal{N}(0,1)$
  - Gaussian(corrected):
    $\left[\hat{\mu} - \frac{t_{1-\alpha/2}\hat{\sigma}}{\sqrt{n}}, \hat{\mu} + \frac{t_{1-\alpha/2}\hat{\sigma}}{\sqrt{n}}\right]$, $t_{1-\alpha/2} = 1 - \alpha/2$-quantile of t distribution (degree of freedom=?)

  - We expect you to be able to compute them on a small dataset

- Confidence intervals for general population parameters: bootstrap

HW4: Problem 2

I would like to build a simple model to predict how many students are likely to come to my office hours this semester. Because this is an arrival process, I will model the number of arrivals during office hours as Poisson distributed. Recall that the Poisson is a discrete distribution over the number of arrivals (or events) in a fixed time-frame. The Poisson distribution has a probability mass function (PMF) of the form,

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

Likelihood function: $$L_n(\lambda) = p(x_1, x_2, x_3, .., x_n; \lambda) = \prod_{i=1}^{n} p(x_i; \lambda)$$

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

HW4: Problem 2

Take the log:

$$
\begin{aligned}
f(\lambda) = \log L_n(\lambda) &= \log\left(\prod_{i=1}^{n} p(x_i)\right) \\
&= \sum_{i=1}^{n} \log\left(\frac{1}{x_i!}\lambda^{x_i} e^{-\lambda}\right) \\
&= \sum_{i=1}^{n} \Big(\log(1) - \log(x_i!) + x_i \log\lambda + (-\lambda)\Big) \\
&= -\sum_{i=1}^{n} \log(x_i!) + \log(\lambda)\sum_{i=1}^{n} x_i - n\lambda
\end{aligned}
$$

HW4: Problem 2

Take the log:
$$f(\lambda) = \log L_n(\lambda) = \log\left(\prod_{i=1}^{n} p(x_i)\right)$$
$$= -\sum_{i=1}^{n} \log(x_i!) + \log(\lambda)\sum_{i=1}^{n} x_i - n\lambda$$

Take the derivative:
$$\frac{df}{d\lambda} = \frac{\sum_{i=1}^{n} x_i}{\lambda} - n = 0$$
$$\Rightarrow \frac{\sum_{i=1}^{n} x_i}{\lambda} = n$$
$$\Rightarrow \lambda^{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

HW2 Problem 4 a)

I have decided to get myself tested for COVID-19 antibodies. However, being comfortable with statistics, I am curious about what the test means for my actual status. Let's investigate these questions, showing all your work.

a) The antibody test I take has a *sensitivity* (a.k.a. true positive rate) of 97.5% and a *specificity* (a.k.a. true negative rate) of 99.1%. If you are not familiar with sensitivity vs specificity, please see Wikipedia. Assume that 4% of the population actually have COVID-19 antibodies. Write down the joint probability distribution $P(S, R)$ with events for antibody state $S \in \{\text{true}, \text{false}\}$ and test result $R \in \{\text{true}, \text{false}\}$.

Law of total probability + Conditional probability: $\cancel{P(A)} = \sum_i P(A \cap B_i) = \sum_i P(B_i)P(A|B_i) = \sum_i \cancel{P(A)}P(B_i|A)$

P(R=True | S=True) = 0.975

P(R=False | S=False) = 0.991

$P(R\|S)$	$S = True$	$S = False$
$R = True$	0.975	0.009
$R = False$	0.025	0.991

$P(S = true) = 0.04$

$P(S = false) = 0.96$

P(R and S)	$S = True$	$S = False$
$R = True$	0.039	0.00864
$R = False$	0.001	0.95136

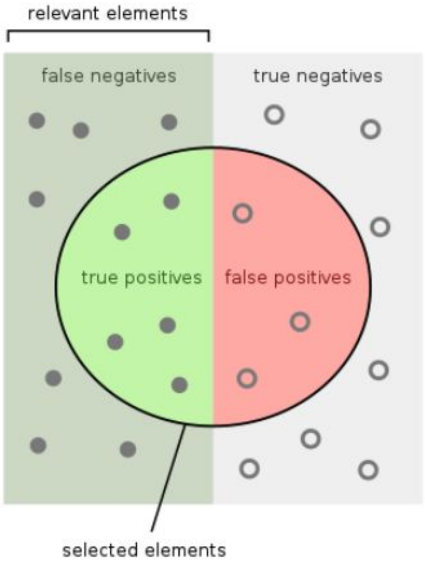
HW2 Problem 4 a)

I have decided to get myself tested for COVID-19 antibodies. However, being comfortable with statistics, I am curious about what the test means for my actual status. Let's investigate these questions, showing all your work.

a) The antibody test I take has a *sensitivity* (a.k.a. true positive rate) of 97.5% and a *specificity* (a.k.a. true negative rate) of 99.1%. If you are not familiar with sensitivity vs specificity, please see Wikipedia. Assume that 4% of the population actually have COVID-19 antibodies. Write down the joint probability distribution $P(S, R)$ with events for antibody state $S \in \{\text{true}, \text{false}\}$ and test result $R \in \{\text{true}, \text{false}\}$.

P(R=True | S=True) = 0.975

P(R=False | S=False) = 0.991



False positive: test says antibody T when antibody is not T
False negative: test says antibody F when antibody is not F

HW2 Problem 4 d)

d) Assume I take the test twice, and receive a positive result in the first test and a negative result in the second test. Assume that the two test results are conditionally independent given the existence of the antibody. What is the probability that I have COVID-19 antibodies according to Bayes' rule?

$$\boxed{P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}}$$

$$P(S = T|R_1 = T, R_2 = F) = \frac{P(R_1 = T, R_2 = F|S = T)P(S = T)}{P(R_1 = T, R_2 = F)}$$

Law of total probability

$$\begin{aligned} &P(R_1 = T, R_2 = F) \\ &= \mathrm{P}(R_1 = T, R_2 = F, S = T) + P(R_1 = T, R_2 = F, S = F) \\ &= P(R_1 = T, R_2 = F|S = T)P(S = T) + P(R_1 = T, R_2 = F|S = F)P(S = F) \\ &= P(R_1 = T|S = T)P(R_2 = F|S = T)P(S = T) + P(R_1 = T|S = F)P(R_2 = F|S = F)\mathrm{P}(S = \mathrm{F}) \end{aligned}$$

HW2 Problem 4 d)

d) Assume I take the test twice, and receive a positive result in the first test and a negative result in the second test. Assume that the two test results are conditionally independent given the existence of the antibody. What is the probability that I have COVID-19 antibodies according to Bayes' rule?

$P(R\|S)$	$S = True$	$S = False$
$R = True$	0.975	0.009
$R = False$	0.025	0.991

$P(S = true) = 0.04$

$P(S = false) = 0.96$

Let T=true and F=false.

$$P(S = T \mid R_1 = T, R_2 = F)$$
$$= \frac{P(R_1 = T, R_2 = F \mid S = T)P(S = T)}{P(R_1 = T, R_2 = F \mid S = T)P(S = T) + P(R_1 = T, R_2 = F \mid S = F)P(S = F)}$$
$$= \frac{P(R_1 = T \mid S = T)P(R_1 = F \mid S = T)P(S = T)}{P(R_1 = T \mid S = T)P(R_2 = F \mid S = T)P(S = T) + P(R_1 = T \mid S = F)P(R_2 = F \mid S = F)P(S = F)}$$
$$= \frac{0.975 \cdot 0.025 \cdot 0.04}{0.975 \cdot 0.025 \cdot 0.04 + 0.009 \cdot 0.991 \cdot 0.96}$$
$$\approx 0.1022$$