Computer Science

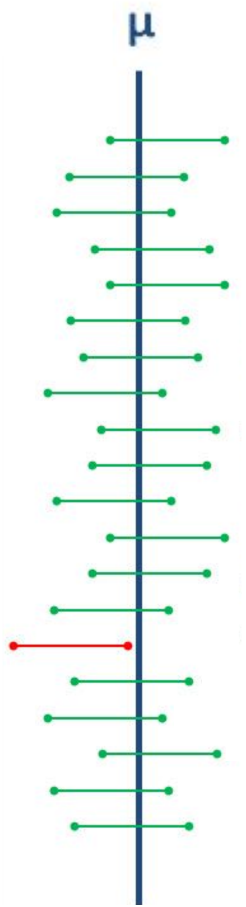
CSC380: Principles of Data Science

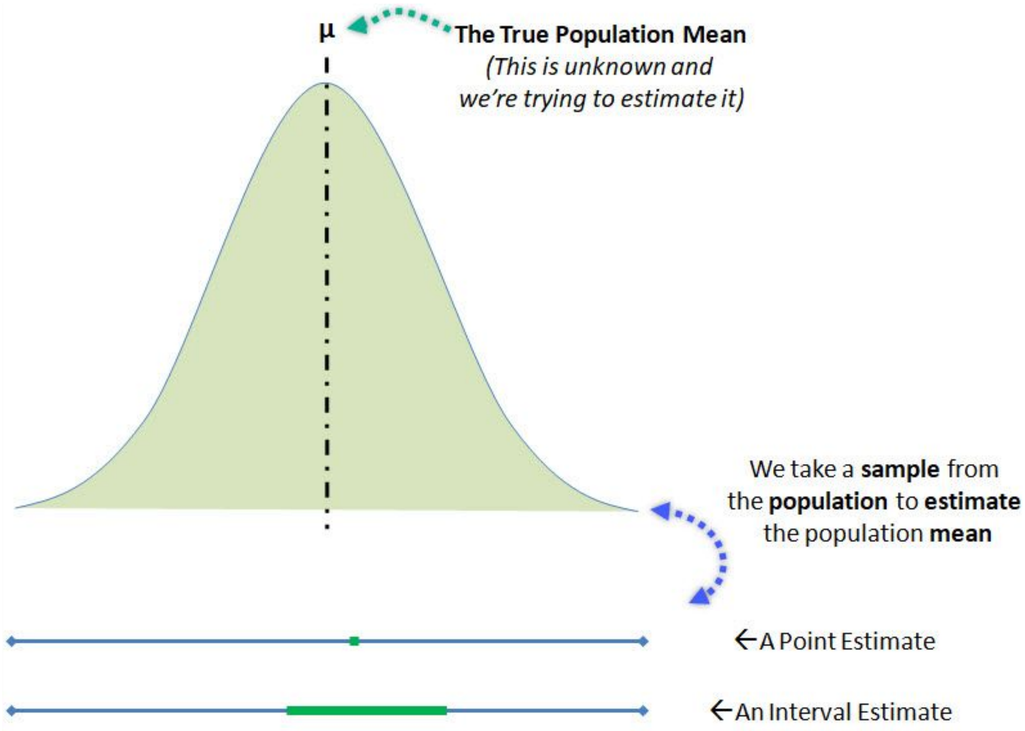
Statistics 5

Xinchen Yu

Acknowledgement: Built on Jason Pacheco, Kwang-Sung Jun, Chicheng Zhang’s slides

Review: Interval estimate





Review: Gaussian (Corrected)

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

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

$\hat{\sigma}$

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

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

where $\Phi(z) \coloneqq \mathrm{P}(\mathrm{Z} \le z)$ 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?

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

Review: Gaussian (Corrected)

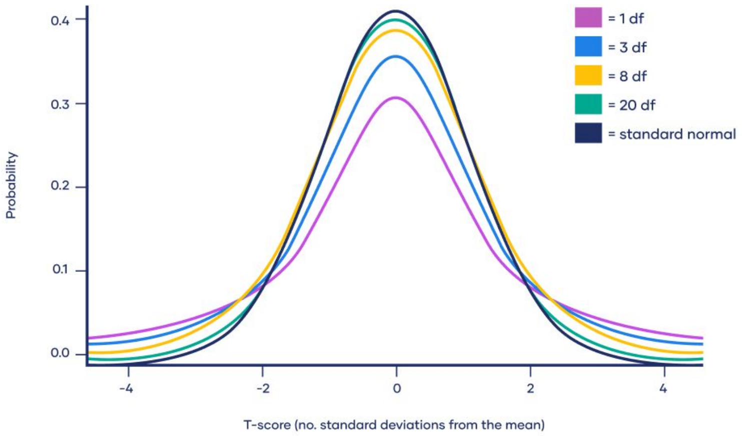
Recall: Gaussian confidence interval with $\sqrt{n}\frac{\hat{\mu}_n-\mu}{\sigma} \sim \mathcal{N}(0,1)$.

What if we use $\hat{\sigma}$ instead of σ?

(Theorem) $X_1, \ldots, X_n$ with unknown μ, σ^2.

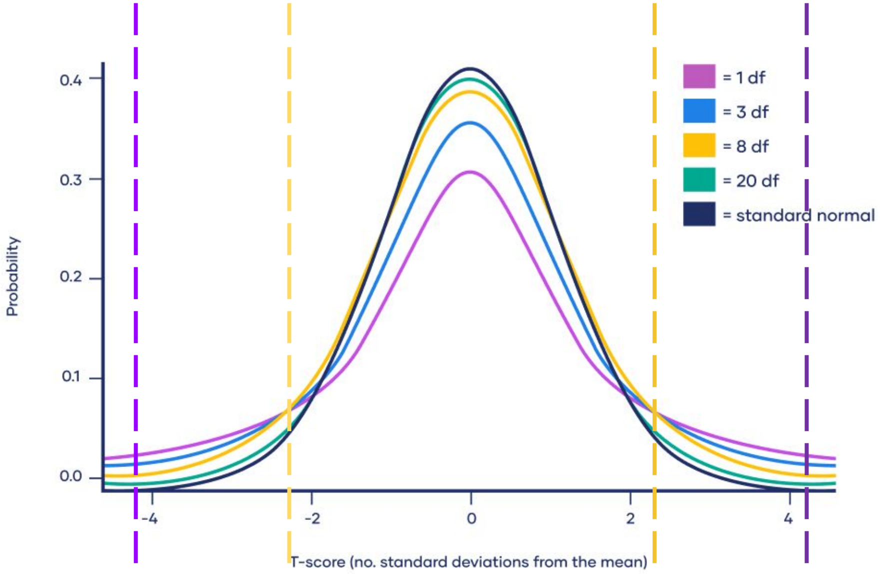
Let $\widehat{UVar}_n := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu}_n)^2$ (unbiased version of sample variance). Then,

$$\sqrt{n}\frac{\hat{\mu}_n-\mu}{\sqrt{\widehat{UVar}_n}} \sim \text{student-t(mean 0, scale 1, degrees of freedom} = n-1)$$



As df approaches infinity, T distribution becomes gaussian

Review: T scores for different df



much larger number compensates for the inaccuracy of $\hat{\sigma}^2$

(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
```

Review: Gaussian (Corrected)

With a similar derivation we have done before,
With at least 95% confidence:

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

Where $t_{\alpha/2,n-1}$ can be computed numerically.

Key take away: more conservative!
=> more likely to be correct.

Common practice*: Apply this method even if we do not know whether true distribution is Gaussian.*

(recall: 1.96 for gaussian)

much larger number compensates for the inaccuracy of $\hat{\sigma}^2$

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

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

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

z = 1.96: RHS ≈ .95, 95% confident

z = 2.58: RHS ≈ .99,

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

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

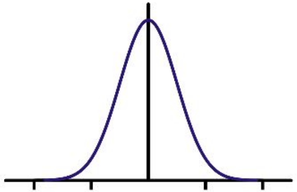
$$P\left(\hat{\mu} \in \left[\mu - \frac{2.58\sigma}{\sqrt{n}}, \mu + \frac{2.58\sigma}{\sqrt{n}}\right]\right) \geq 0.99$$

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

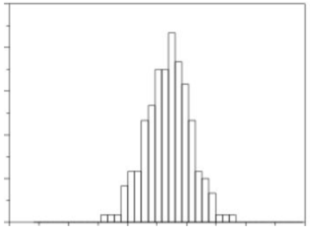
Method 2: Bootstrap

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

θ

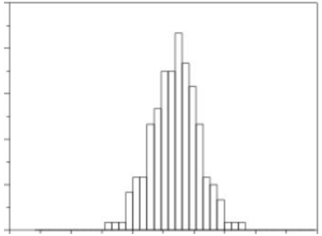


$$D_\theta \xrightarrow{n \text{ iid samples}} S$$

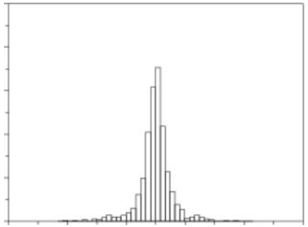


$\hat{\theta}_n$

$\hat{\theta}_n$



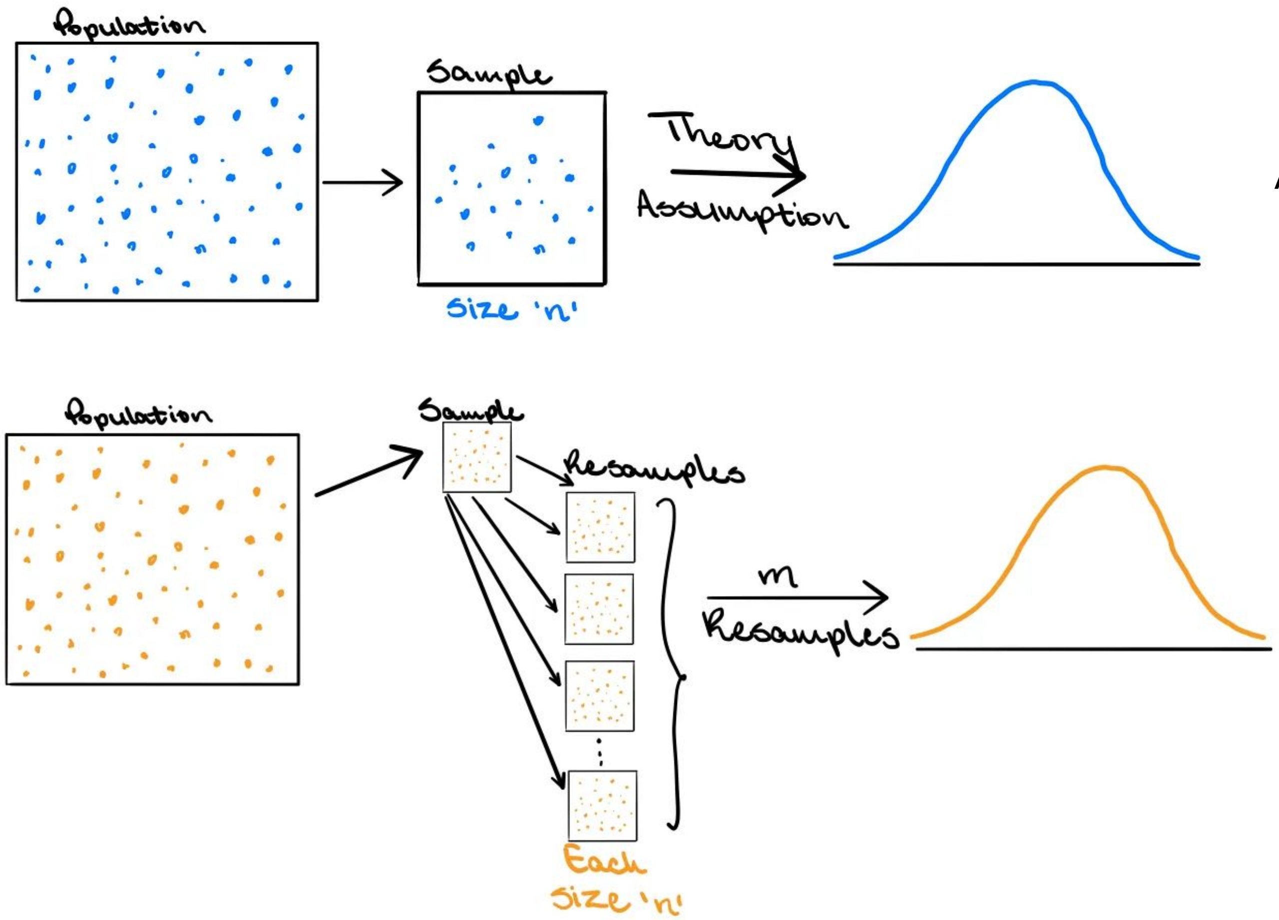
$$\text{Uniform}(S) \xrightarrow{n \text{ iid samples}} S_{\text{b}}$$



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

- 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

$\widehat{\theta} - \theta$

$\left\{\hat{\theta}_{n,b} - \frac{1}{n}\sum_{i=1}^{n} X_i\right\}_{b=1}^{B}$

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

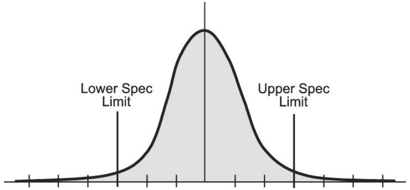
Suppose we observe data $X_1, X_2, \ldots, X_n \sim P(X;\theta)$:

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 <u>**empirical distribution**</u> 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?

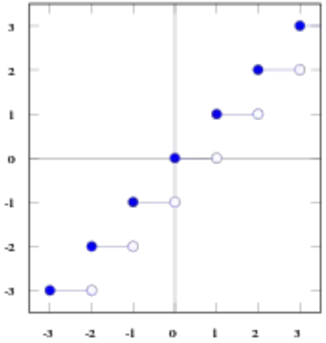
$$P\left(v_{\frac{\alpha}{2}} \leq \hat{\theta}_n - \theta \leq v_{1-\frac{\alpha}{2}}\right) \geq 1 - \alpha$$

Method 2: Bootstrap

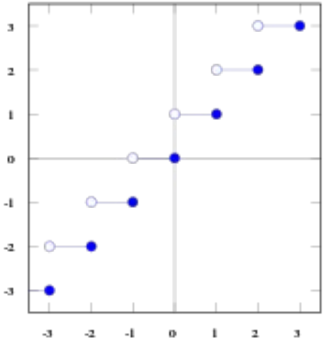
Pseudocode

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

Confidence Intervals Comparison

good = correct
bad = incorrect

	Gaussian (corrected)	Bootstrap
small n	Bad	Bad
moderate n	Okay / Bad	Okay
large n	Good	Very Good
Computational complexity	Low	High, depends on B

bad if the estimator takes a long time to compute

Q: When could it be bad?

When the distribution is far from Gaussian

Classical Statistics Review

- **Statistical Estimation** infers unknown parameters θ of a distribution $p(X;\theta)$ from observed data $X_1, \ldots, X_n$
- An estimator is a function of the data $\hat{\theta}(X_1, \ldots, X_n)$, it is a **random variable**, so it has a distribution
- **Confidence Intervals** measure uncertainty of an estimator, e.g.

$$P(\theta \in (a(X), b(X))) \geq 0.95$$

- **Bootstrap** A simple method for estimating confidence intervals

↑ Q: when is this good?

Caution

- Confidence intervals are often misinterpreted!
- Confidence intervals in practice may not be true for small n

Classical Statistics Review

- **Estimator bias** describes systematic error of an estimator

- **Mean squared error (MSE)** measures estimator quality / efficiency,

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

- **Law of Large Numbers (LLN)** guarantees that sample mean approaches (piles up near) true mean in the limit of infinite data

- **Central Limit Theorem (CLT)** says sample mean approaches a Normal distribution with enough data. Also means $\frac{1}{\sqrt{n}}$ convergence.

- **LLN** and **CLT** are *asymptotic statements* and do not hold for small/medium data in general

- Probability
- Statistics

- Data Visualization

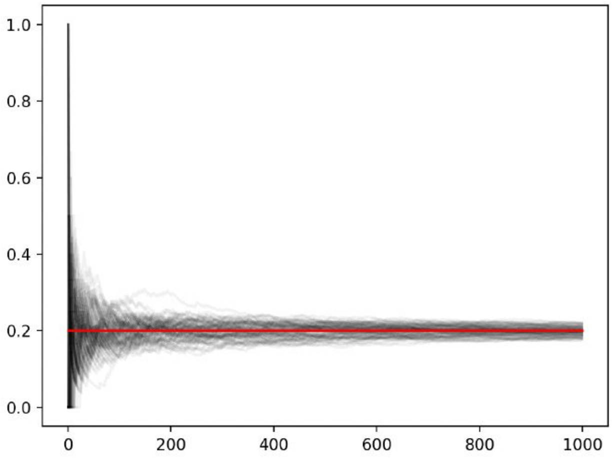
- Predictive modeling
- Linear models
- Nonlinear models
- Clustering

HW3: Problem 1 a)

a) Let us numerically verify the law of large numbers. We will simulate $m = 100$ sample mean trajectories of $X_1, \ldots, X_N \sim \text{Bernoulli}(\mu = 0.2)$ and plot them altogether in one plot. Here, a sample mean trajectory means a sequence of $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_N$ where $\bar{X}_i$ is the sample mean using samples $X_1, \ldots, X_i$. We will plot $\bar{X}_n$ as a function of n, but do this multiple times. Take n from 1 to $N = 1000$. An ideal plot would look like the following:

N = 1, 2, 3, 4, 5, 6, …… 1000

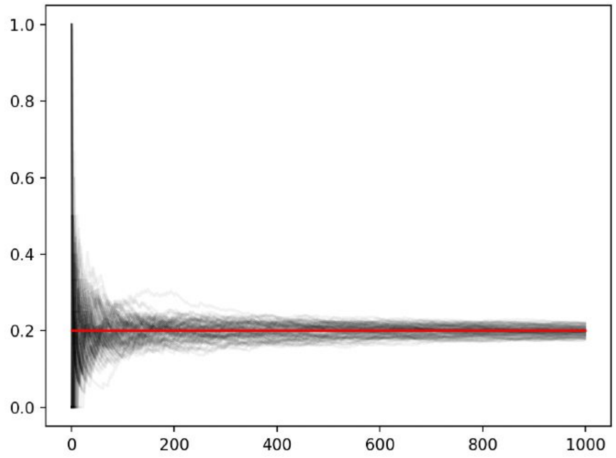
0, 1, 1, 0, 0, 0, …… 0, 0, 1



HW3: Problem 1 a)

a) Let us numerically verify the law of large numbers. We will simulate $m = 100$ sample mean trajectories of $X_1, \ldots, X_N \sim \text{Bernoulli}(\mu = 0.2)$ and plot them altogether in one plot. Here, a sample mean trajectory means a sequence of $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_N$ where $\bar{X}_i$ is the sample mean using samples $X_1, \ldots, X_i$. We will plot $\bar{X}_n$ as a function of n, but do this multiple times. Take n from 1 to $N = 1000$. An ideal plot would look like the following:

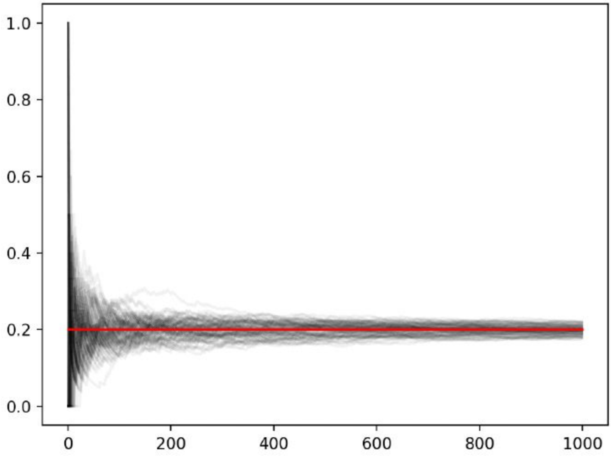
$$
\begin{array}{cc}
 & \text{N = } \; 1,\; 2,\; 3,\; 4,\; 5,\; 6,\quad \ldots\ldots \quad 1000 \\
\text{m = 100} & \left(\begin{array}{c}
[\,0,\; 1,\; 1,\; 0,\; 0,\; 0,\quad \ldots\ldots \quad 0,\; 0,\; 1\,] \\
[\,1,\; 0,\; 0,\; 0,\; 0,\; 0,\quad \ldots\ldots \quad 0,\; 1,\; 0\,] \\
[\,0,\; 1,\; 0,\; 0,\; 0,\; 0,\quad \ldots\ldots \quad 1,\; 0,\; 1\,] \\
\ldots\ldots \\
[\,0,\; 0,\; 0,\; 0,\; 0,\; 1,\quad \ldots\ldots \quad 0,\; 0,\; 0\,]
\end{array}\right)
\end{array}
$$



HW3: Problem 1 a)

a) Let us numerically verify the law of large numbers. We will simulate $m = 100$ sample mean trajectories of $X_1, \ldots, X_N \sim \text{Bernoulli}(\mu = 0.2)$ and plot them altogether in one plot. Here, a sample mean trajectory means a sequence of $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_N$ where $\bar{X}_i$ is the sample mean using samples $X_1, \ldots, X_i$. We will plot $\bar{X}_n$ as a function of n, but do this multiple times. Take n from 1 to $N = 1000$. An ideal plot would look like the following:

$$
\begin{array}{cc}
N = & 1,\ 2,\ 3,\ 4,\ 5,\ 6,\ \ldots\ldots\ 1000 \\
m = 100 & \left(\begin{array}{l}
\boxed{0,}\ 1,\ 1,\ 0,\ 0,\ 0,\ \ldots\ldots\ 0,\ 0,\ 1 \\
\boxed{1,}\ 0,\ 0,\ 0,\ 0,\ 0,\ \ldots\ldots\ 0,\ 1,\ 0 \\
\boxed{0,}\ 1,\ 0,\ 0,\ 0,\ 0,\ \ldots\ldots\ 1,\ 0,\ 1 \\
\qquad\qquad\qquad \ldots\ldots \\
\boxed{0,}\ 0,\ 0,\ 0,\ 0,\ 1,\ \ldots\ldots\ 0,\ 0,\ 0
\end{array}\right)
\end{array}
$$



HW3: Problem 1 a)

a) Let us numerically verify the law of large numbers. We will simulate $m = 100$ sample mean trajectories of $X_1, \dots, X_N \sim \text{Bernoulli}(\mu = 0.2)$ and plot them altogether in one plot. Here, a sample mean trajectory means a sequence of $\bar{X}_1, \bar{X}_2, \dots, \bar{X}_N$ where $\bar{X}_i$ is the sample mean using samples $X_1, \dots, X_i$. We will plot $\bar{X}_n$ as a function of n, but do this multiple times. Take n from 1 to $N = 1000$. An ideal plot would look like the following:

N = 1, 2, 3, 4, 5, 6, …… 1000

m = 100

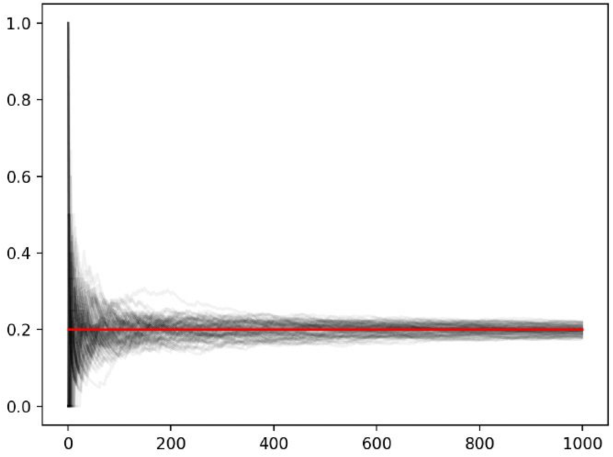
0, 1, 1, 0, 0, 0, …… 0, 0, 1

1, 0, 0, 0, 0, 0, …… 0, 1, 0

0, 1, 0, 0, 0, 0, …… 1, 0, 1

……

0, 0, 0, 0, 0, 1, …… 0, 0, 0



HW3: Problem 1 a)

a) Let us numerically verify the law of large numbers. We will simulate $m = 100$ sample mean trajectories of $X_1, \ldots, X_N \sim \text{Bernoulli}(\mu = 0.2)$ and plot them altogether in one plot. Here, a sample mean trajectory means a sequence of $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_N$ where $\bar{X}_i$ is the sample mean using samples $X_1, \ldots, X_i$. We will plot $\bar{X}_n$ as a function of n, but do this multiple times. Take n from 1 to $N = 1000$. An ideal plot would look like the following:

N = 1, 2, 3, 4, 5, 6, …… **1000**

m = 100

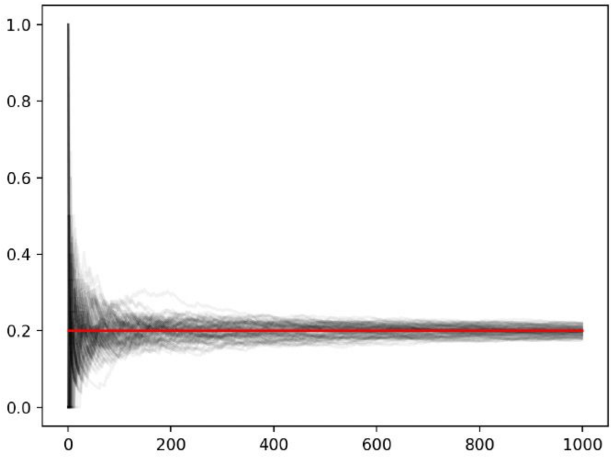
0, 1, 1, 0, 0, 0, …… 0, 0, 1

1, 0, 0, 0, 0, 0, …… 0, 1, 0

0, 1, 0, 0, 0, 0, …… 1, 0, 1

……

0, 0, 0, 0, 0, 1, …… 0, 0, 0



HW3: 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: $$f(\lambda) = \log L_n(\lambda) = \log \left(\prod_{i=1}^{n} p(x_i) \right)$$

HW3: Problem 2

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

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

$$= \sum_{i=1}^{n}\Big(\log(1) - \log(x_i!) + x_i\log\lambda + (-\lambda)\Big)$$

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

HW3: 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: $$\frac{df}{d\lambda} = \frac{\sum_{i=1}^{n} x_i}{\lambda} - n = 0$$

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

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

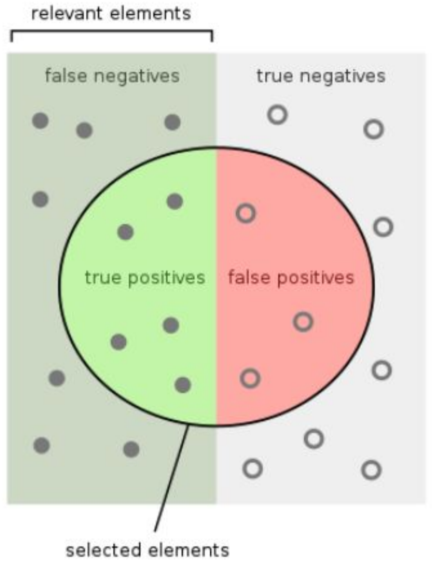
HW2 Problem 4 d)

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)

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 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?

Let T=true and F=false.

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

Midterm

- Cheat sheet: US Letter size (1 page, 2-sided)
- Practice questions will release next Tuesday Oct 3rd
  - Solutions will release on Oct 8th
- Midterm review next Thursday
- How to prepare?
  - Slides
  - Examples
  - Homework questions