Lecture 6: Interval estimation

Statistical Methods for Data Science

Yinan Yu

Department of Computer Science and Engineering

November 17 and 21, 2022

Today

1. Central limit theorem
   - Terminology
   - Standardization
   - Central limit theorem
2. Interval estimation
   - Confidence interval
   - Credible interval
3. Summary

Learning outcome

- Be able to explain the following terminology:
  - Sample statistic, sampling distribution, sample mean, sample variance, standardization, z-table, t-table
  - Point estimation, interval estimation
  - Confidence interval, credible interval
- Be able to explain the central limit theorem (CLT)
- Be able to construct the following interval estimates:
  - Confidence interval for
    - sample mean of i.i.d. sample with unknown σ
    - unknown sampling distribution using bootstrap
  - Credible interval for a given posterior function

Today

1. Central limit theorem
   - Terminology
   - Standardization
   - Central limit theorem
2. Interval estimation
3. Summary

Terminology

Terminology

- (Statistical) population: all items of interest (e.g., all ducks in the world)
- **Sample**: a random data set $\{x_1, x_2, \cdots, x_N\}$; the corresponding random variables are denoted as $X_1, X_2 \cdots, X_N$; a subset of the population (e.g., the 20 ducks you have weighed)
- **i.i.d. sample**: $X_1, X_2 \cdots, X_N$ are i.i.d. random variables
- **Sample statistic**: a statistic computed from a sample
  For example,
  - **Sample mean**:

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

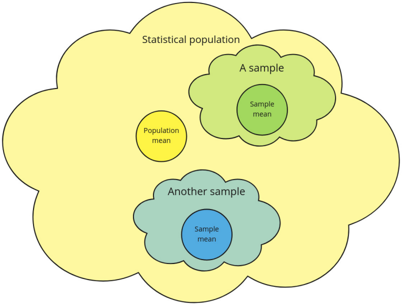
  - **Sample variance**:

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

Note: capital letters and small letters are used to denote random variables and the values, respectively.

Terminology (cont.)

- **Sampling distribution**: the probability distribution of a sample statistic that is computed from a random sample (of size N)



- Asymptotic: in this context, asymptotic means $N \to \infty$

What's the difference between the mean of a Gaussian distribution is random (Bayesianist) vs the sample mean is random?

Awesome properties of Gaussian random variables

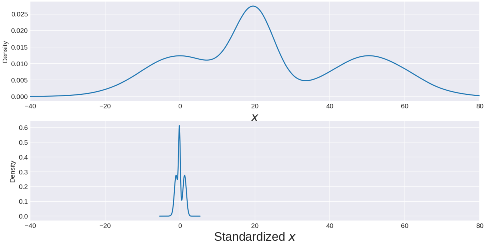
- Let $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ be a Gaussian random variable, then the following random variables are also Gaussian (location scale family)
  - Scaling (scale): $tX \sim \mathcal{N}(t\mu_X, t^2\sigma_X^2)$, $t \neq 0$ is a constant
  - Translation (location): $X + c \sim \mathcal{N}(\mu_X + c, \sigma_X^2)$, c is a constant
  - $tX + c \sim \mathcal{N}(t\mu_X + c, t^2\sigma_X^2)$
- Let $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ be two **independent** Gaussian random variables, then the following random variables are also Gaussian
  - $X + Y \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$
  - $X - Y \sim \mathcal{N}(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2)$

Standardization

Standardization

- Why standardization? We want to translate and scale data into a **standard shape** so that we can use standard tools to compare and analyze it
- Let X be a random variable that follows **any probability distribution** with mean μ and standard deviation σ. The standardization of X is

$$Y = \frac{X - \mu}{\sigma}$$

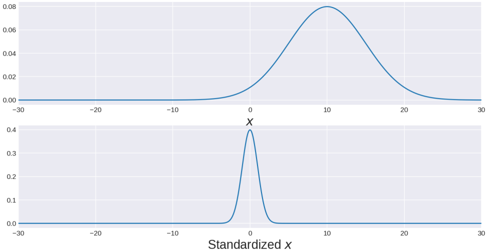


Question: what is the mean and standard deviation of Y? Random variable Y has **mean 0 and standard deviation 1**

Standardization

- Let X be a random variable following a **Gaussian distribution** with mean μ and standard deviation σ, i.e. $X \sim \mathcal{N}(\mu, \sigma^2)$; the standardization of X is

$$Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1) \tag{1}$$



The distribution $\mathcal{N}(0, 1)$ is called a **standard Gaussian (normal) distribution**

Standard Gaussian distribution

- Remember how much we love Gaussian distributions? **We love the standard Gaussian distribution even more!** We love it so much that we gave its CDF a special name: $\Phi(z)$
- There is a table describing the quantiles of the standard Gaussian (the **z-table**)

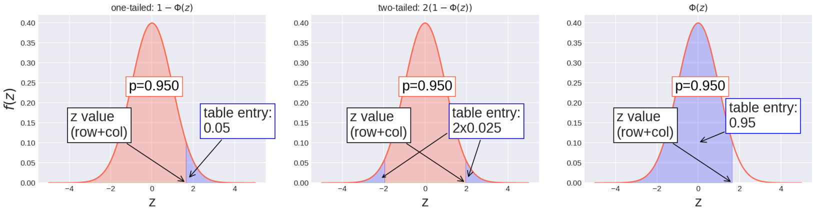
z	+ 0.00	+ 0.01	+ 0.02	+ 0.03	+ 0.04	+ 0.05	+ 0.06	+ 0.07	+ 0.08	+ 0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56360	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91308	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670

  - Each row represents the integer and the first decimal of z
  - Each column represents the second decimal of z
  - Each cell is the

$$
\begin{aligned}
P(Z \leq \text{row} + \text{column}) &= \Phi(\text{row} + \text{column}) \\
&= \textbf{stats.norm.cdf(x=row + column, loc=0, scale=1)}
\end{aligned}
$$

Standard Gaussian distribution (cont.)

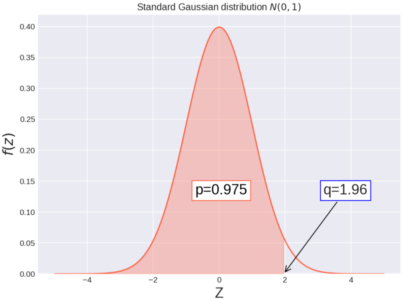
- There are different representations of the z-table; the difference is what is inside each cell, e.g. $\Phi(\text{row} + \text{column})$, $2(1 - \Phi(\text{row} + \text{column}))$, $1 - \Phi(\text{row} + \text{column})$ or $\frac{1}{2}(1 - \Phi(\text{row} + \text{column}))$; but the principle is the same; for now we use the version with $\Phi(\text{row} + \text{column})$



- Due to symmetry, there are only positive values for z in the z-table

Standard Gaussian distribution (cont.)

Exercise:



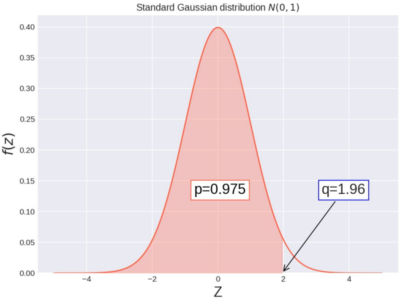
z-table

z	+ 0.00	+ 0.01	+ 0.02	+ 0.03	+ 0.04	+ 0.05	+ 0.06	+ 0.07	+ 0.08	+ 0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56360	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91308	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670

Try to find the corresponding pair $(p, q) = (0.975, 1.96)$ in the z-table (60 secs).

Standard Gaussian distribution (cont.)

Answer:



z	+ 0.00	+ 0.01	+ 0.02	+ 0.03	+ 0.04	+ 0.05	+ 0.06	+ 0.07	+ 0.08	+ 0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56360	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91308	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670

q=1.9+0.06=1.96

p=0.9750

Note: the table itself is not important (we use a computer these days); the point is to reflect on the meaning of z values (quantiles) and the related probabilities (CDFs)

Central limit theorem

Motivation and use cases

So far, we have been looking at **distributions** (centrality and spread); the central limit theorem is about the **mean** (centrality only); do we care about the mean that much?

- Yes, we do!
- Example: we want to test the effectiveness of a drug; a patient can be either **cured** by this drug or **not cured**, i.e., we can model the data using a (2 secs) *Bernoulli distribution* with parameter (2 second) p (**cure rate**) and the maximum likelihood estimation of p is the (4 secs) **sample mean**
- In general, we are often interested in how things work “on average”

Distribution of the sample mean

- You have 1000 ducks
- Now, you take 30 of them and measure the sample mean of their weights x_i:

$$\hat{\mu}_1 = \frac{1}{30}\sum_{i=1}^{30} x_i$$

- Then you take another 30 ducks to measure the sample mean of their weights y_i:

$$\hat{\mu}_2 = \frac{1}{30}\sum_{i=1}^{30} y_i$$

- You do this experiment 100 times and plot the histogram of these 100 sample means $\hat{\mu}_j$ for $j = 1, \cdots, 100$
- Then you realize **these sample means $\hat{\mu}_j$ seem to follow a Gaussian distribution** 🤔

Distribution of the sample mean (cont.)

- The colors of your 1000 ducks can be either red $t_i = 0$ or blue $t_i = 1$
- Now, you take 30 of them and measure the sample mean of their color t_i:

$$\hat{n}_1 = \frac{1}{30}\sum_{i=1}^{30} t_i = \frac{1}{30}(1 + 1 + 0 + 1 + \cdots \ldots 1 + 1)$$

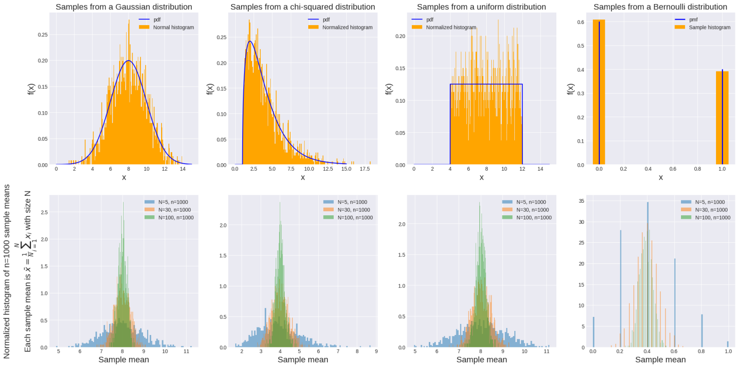
  Note: here $t_i \in \{0, 1\}$ is discrete
- You take another 30 ducks and measure the sample mean of their color t_i:

$$\hat{n}_2 = \frac{1}{30}\sum_{i=1}^{30} t_i = \frac{1}{30}(0 + 0 + 0 + 1 + \cdots \ldots 1 + 0)$$

- You do this experiment 100 times and plot the histogram of these 100 sample means $\hat{n}_j$
- Then you realize **these sample means $\hat{n}_j$ also seem to follow a Gaussian distribution** 🤯 !!???

Distribution of the sample mean (cont.)

- In fact, this is true for i.i.d. samples drawn from ANY probability distribution



- The larger the sample size N (in the previous example $N = 30$), the "more Gaussian" it becomes
- A rule of thumb: $N \geq 30$
- If the data distribution is Gaussian-like (bell-shaped, symmetric), only a small sample size is needed for the sample mean to be Gaussian

Central limit theorem

- One of the most important results in probability theory and statistics
- Given an **i.i.d. sample** $X_1, X_2, \cdots, X_N$ from **ANY probability distribution** with *finite mean* μ *and variance* σ^2 (most distributions satisfy this!), when the sample size N is sufficiently large, the **sample mean** approximately follows a Gaussian distribution with mean μ and variance $\frac{\sigma^2}{N}$, i.e.,

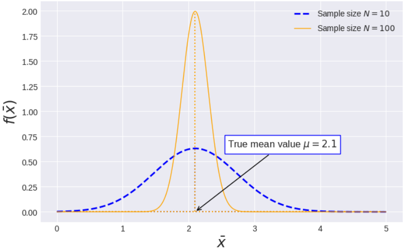
$$\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{N}) \tag{2}$$

where $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$ is the sample mean

Central limit theorem (cont.)

How to interpret this?

$$\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{N})$$



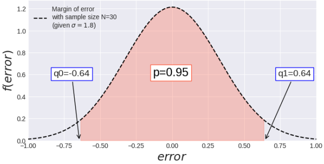
- The sample mean $\bar{X}$ is around the true mean value μ
- The "deviation" of $\bar{X}$ from μ is $\frac{\sigma^2}{N}$; the larger N, the smaller the deviation

Estimation error $\bar{X} - \mu$

We are interested in estimating the mean value μ
We use the sample mean $\bar{X}$ to estimate the mean value μ
We are interested in how good this estimation is

Analysis of the estimation error $\bar{X} - \mu$

- In many use cases, we want to estimate **the population mean μ using the sample mean** from **one sample** $\{x_1, \cdots, x_N\}$ and we are interested in the **statistics of the estimation error**
- **Random variable $X_1, \cdots, X_N$** from **any distribution**
- **Assumption**: i.i.d. with **known** standard deviation σ and **unknown** mean μ
- **Random variable of interest**: $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$
  - From CLT (page 21), we know that for a large N, the sample mean approximately follows a Gaussian distribution $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{N})$: $\bar{X}$ is around the true mean μ
  - Let $\mathcal{E} = \bar{X} - \mu$ be the estimation **error**; what distribution does $\mathcal{E}$ follow (awesome properties of Gaussian - 30 secs)? $\mathcal{E} \sim \mathcal{N}(0, \frac{\sigma^2}{N})$; can we plot the PDF of $\mathcal{E}$? (5 secs) Yes! σ and N are both **known**!



  - Given $q0$ and $q1$, how to interpret this plot (5 secs)? **95% of the time, the error $\bar{X} - \mu$ is within $q0 = -0.64$ and $q1 = 0.64$**
  - Now it's pretty cool because not only can we estimate the mean (using the sample mean), but we also get the margin of error!
  - This **95%** is called the **confidence level**; for a given confidence level, we can find a corresponding **interval** $(q0, q1)$

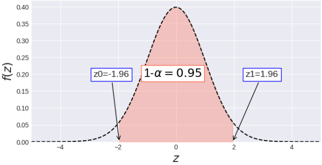
Calculate the margin of error

- For a given confidence level, denoted as $1-\alpha$, how do we find this interval for the error term in Python? We can use the function **ppf** from **scipy.stats**

```
std = 1.8 # known standard deviation
N = 30
alpha = 0.05
confidence_level = 1 - alpha # 95% confidence level
q0 = stats.norm.ppf(alpha/2,
                    0, std/math.sqrt(N))
q1 = stats.norm.ppf(confidence_level+alpha/2,
                    0, std/math.sqrt(N))
>> (-0.6441098917381766, 0.6441098917381766)
```

Find a standardized expression for the margin of error

- Standardize (cf. page 11) $\mathcal{E}$ by $\frac{\mathcal{E}}{\sigma/\sqrt{N}} = \frac{\bar{X}-\mu}{\sigma/\sqrt{N}} \sim \mathcal{N}(0,1)$
- We just learned that there is a special name for the standard Gaussian distributed random variable - $Z \sim \mathcal{N}(0,1)$ - let $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{N}}$
- Now we have an expression for the error term in terms of Z: $\mathcal{E} = \bar{X} - \mu = Z\frac{\sigma}{\sqrt{N}}$
- The only random variable here is $Z \sim \mathcal{N}(0,1)$



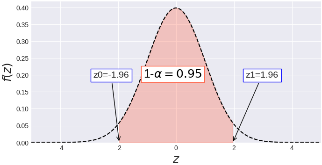
- We can use a two-tailed z-table (cf. page 13) to find the values for $z0$ and $z1$
- In order to find an interval for $\mathcal{E}$, we just need to look at

$$(z0\frac{\sigma}{\sqrt{N}}, z1\frac{\sigma}{\sqrt{N}})$$

Find a standardized expression for the margin of error (cont.)

- For example, with $1-\alpha = 95\%$ confidence level, the error is within

$$\left(-1.96\frac{\sigma}{\sqrt{N}},\ 1.96\frac{\sigma}{\sqrt{N}}\right)$$



- Generally speaking, the value $z1$ (denoted by $z_{\alpha/2}$) is the quantile at $1-\alpha/2$; the value of $z_{\alpha/2}$ is called the **(right) critical value**; $\frac{\sigma}{\sqrt{N}}$ is called the **standard error**; in this example, we have $z_{\alpha/2} = z1 = -z0 = 1.96$
- Why **two-tailed** z-table: there are two tails $z \leq -z_{\alpha/2}$ and $z \geq z_{\alpha/2}$

Find a standardized expression for the margin of error (cont.)

- In Python

```
std = 1.8
N = 30
alpha = 0.05
confidence_level = 1 - alpha # 95% confidence level
z0 = stats.norm.ppf(alpha/2, 0, 1)
z1 = stats.norm.ppf(confidence_level+alpha/2, 0, 1)
print(z0*std/math.sqrt(N), z1*std/math.sqrt(N))
>> (-0.6441098917381766, 0.6441098917381766)
```

Find a standardized expression for the margin of error (cont.)

- For a given sample with an estimate $\bar{x}$ (note: here the small letter $\bar{x}$ denotes the value of the estimate itself instead of a random variable), it's more convenient to have this margin of error around $\bar{x}$ instead - so that we can say: the estimated mean is $\bar{x}$ with this uncertainty:

$$\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{N}},\ \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{N}}\right)$$

- This is called the **confidence interval**
- The confidence interval for the sample mean is *exact* when the data distribution is Gaussian, otherwise it is an approximation under the central limit theorem
- This calculation is called **interval estimation**, because it gives an interval estimate $\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{N}},\ \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{N}}\right)$ instead of a single value estimate as in MAP or MLE

Today

1. Central limit theorem
2. Interval estimation
   - Confidence interval
   - Credible interval
3. Summary

One quick question

- What does it mean by something being **random**?
  - We can use a **random variable** to model the data
  - The random variable has an underlying **probability distribution**
  - We can use **distribution functions (e.g. CDF, PDF/PMF)** to describe the probability distribution
- Actually another quick question... What does it mean by **unknown probability distribution**?
  - The **family** of the probability distribution is unknown, e.g. Gaussian? Uniform?
  - Or given a family of probability distributions, some **parameters** of the probability distribution are unknown, e.g. a Gaussian distribution with unknown mean or unknown variance

More questions

- How to estimate an **unknown probability distribution**?
  - Q-Q plot to estimate the family of probability distributions, e.g. data distribution vs Gaussian distribution
  - Parameter estimation techniques (e.g. MLE, MAP) to estimate the unknown parameters
- Given a probability distribution with unknown mean value μ, is μ **random**?
  - In Bayesian approaches (e.g. MAP), it is assumed random
  - In Frequentist approaches (e.g. MLE), it is not assumed random; it is an **unknown constant**
- Is the **sample mean** random?
  - Yes, it is random. The sample mean is a **sample statistic**; a sample statistic is computed from a sample; a sample is random and hence the sample statistic is random.

I can't promise but one last question...

- Is the **sample mean** always the MLE for the mean?
    - It is the MLE for the mean value of Gaussian distributions, but it is not always the MLE for the mean value of any distribution

Now we are done with questions
UNLESS...
Nah we are done

Interval estimation

- MLE and MAP are **point estimation** techniques since they only return one single value, i.e., a point, for the parameter estimation
- However, we are often interested in the **uncertainty** associated with the point estimate; a point estimate + uncertainty is called an **interval estimate** since they return an interval instead a single value

Confidence interval

Confidence interval (CI)

- **Data**: $x_1, \cdots, x_N$
- **Random variable**: $X_1, \cdots, X_N$ with i.i.d. assumption
- **Parameter of interest**: θ, e.g. the mean μ
- **Estimate**: $\hat{\theta}$, e.g. the sample mean $\bar{x}$
- **Confidence interval** for a given confidence level $1 - \alpha$ (e.g. 95%)
  - Definition:

    confidence interval = ($\hat{\theta}$ - **margin of error**, $\hat{\theta}$ + **margin of error**)

    where

    margin of error = critical value $\times$ standard error of $\hat{\theta}$
  - Calculation:

Distribution of X_i	Scenario	θ	$\hat{\theta}$ (sampling distribution)	Critical value	Standard error	Confidence interval	Note
i.i.d. Gaussian	✓ σ known	mean	sample mean $\bar{x}$	$z_{\alpha/2}$	$\frac{\sigma}{\sqrt{N}}$	$\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{N}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{N}}\right)$	exact
	? σ unknown		(Gaussian distribution)	$t_{\alpha/2}$	$\frac{s}{\sqrt{N}}$	$\left(\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{N}}, \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{N}}\right)$	
i.i.d.	✓ σ known		sample mean $\bar{x}$	$z_{\alpha/2}$	$\frac{\sigma}{\sqrt{N}}$	$\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{N}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{N}}\right)$	approximate
	? σ unknown		(approximately Gaussian under CLT)	$t_{\alpha/2}$	$\frac{s}{\sqrt{N}}$	$\left(\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{N}}, \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{N}}\right)$	for large N
i.i.d.	-	any	MLE (asymptotically Gaussian)	$z_{\alpha/2}$	$\frac{1}{\sqrt{NI_N(\hat{\theta})}}$	$\left(\hat{\theta} - z_{\alpha/2}\frac{1}{\sqrt{NI_N(\hat{\theta})}}, \hat{\theta} + z_{\alpha/2}\frac{1}{\sqrt{NI_N(\hat{\theta})}}\right)$	asymptotic
i.i.d.	? -	any	any statistic (any distribution)	bootstrap the error quantile		$\left(\hat{\theta} - \epsilon_{1-\alpha/2}, \hat{\theta} - \epsilon_{\alpha/2}\right)$	approximate

where σ is the standard deviation of the X_i and s the sample standard deviation

Calculation of the confidence interval

Data: $x_1, \cdots, x_N$
Random variable: $X_1, \cdots, X_N$ i.i.d. with standard deviation σ

- CI for Gaussian sampling distribution (exact, approximate, asymptotic):
  - **Parameter of interest:** mean value
    Estimation method: sample mean $\bar{x}$
    - ✅ σ known: $\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{N}},\ \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{N}}\right)$ (cf. page 29)
    - ? σ unknown: $\left(\bar{x} - t_{\alpha/2}\frac{\sigma}{\sqrt{N}},\ \bar{x} + t_{\alpha/2}\frac{\sigma}{\sqrt{N}}\right)$
  - **Parameter of interest:** any statistic
    Estimation method: MLE (cf. lecture 3 properties of MLE)
    - 🙈 [not required] $\left(\hat{\theta} - z_{\alpha/2}\frac{1}{\sqrt{NI_N(\hat{\theta})}}, \hat{\theta} + z_{\alpha/2}\frac{1}{\sqrt{NI_N(\hat{\theta})}}\right)$
- CI for unknown sampling distribution
  - **Parameter of interest:** any parameter, e.g. median
    Estimation method: any method
    - ? Bootstrap $\left(\hat{\theta} - \epsilon_{1-\alpha/2}, \hat{\theta} - \epsilon_{\alpha/2}\right)$

CI for unknown σ

- When the standard deviation σ is **known**, we have shown the standardization of the error term $\frac{\mathcal{E}}{\sigma/\sqrt{N}} = \frac{\bar{X}-\mu}{\sigma/\sqrt{N}} \sim \mathcal{N}(0,1)$ (cf. page. 26).
- When σ is **unknown**, which is the most common case, we replace σ by its estimate $\hat{\sigma}$ - the **sample standard deviation** s - random variable S

$$\hat{\sigma} = S = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

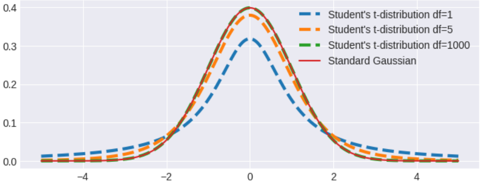
- Now the standardization of the error term becomes (**random**, **constant**):

$$\frac{\mathcal{E}}{\sigma/\sqrt{N}} \rightarrow \frac{\mathcal{E}}{S/\sqrt{N}} = \frac{\bar{X}-\mu}{S/\sqrt{N}} \sim t(N-1)$$

- Compared to the case with known σ, $\frac{\bar{X}-\mu}{\sigma/\sqrt{N}} \sim \mathcal{N}(0,1)$, the distribution of $\frac{\bar{X}-\mu}{S/\sqrt{N}}$ is no longer the standard Gaussian ($\frac{\mu}{S/\sqrt{N}}$ is no longer a constant because S is a random variable). Instead, it follows a **Student's t-distribution** t. The Student's t-distribution has one parameter $df = N-1$ (**degrees of freedom**).

CI for unknown σ (cont.)

- The Student's t-distribution is a function of the sample size: $df = N - 1$
- Think of it as a standard Gaussian compensated for the small sample size. For a large N, they become very similar.



CI for unknown σ (cont.)

- **t-table**: similar to the z-table for the standard Gaussian distribution, there is a t-table for the Student's t-distribution (image from `http://www.ttable.org/`).
- **each cell = stats.t.ppf(q=cum.prob, df=N-1, loc=0, scale=1)**
- α = **two-tails** and **confidence level** $= 1 - \alpha$

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

Summary

Data: $x_1, \cdots, x_N$
Random variable: $X_1, \cdots, X_N$ i.i.d. with standard deviation σ
CI for unknown σ with Gaussian sampling distribution

$$\left(\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{N}}, \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{N}}\right)$$

CI for unknown sampling distribution

- Sample mean approximately follows a Gaussian distribution under the central limit theorem, but most other statistics do not have such luxury
- When the sampling distribution is unknown, we cannot use the t-table or z-table to find the critical values
- Recall the definition of CI:
  confidence interval = ($\hat{\theta}$ - **margin of error**, $\hat{\theta}$ + **margin of error**)
- One way to find this CI is to approximate the **margin of error** using **bootstrap**

Bootstrap

- **Data**: $x_1, \cdots, x_N$
- **Random variables**: $X_1, \cdots, X_N$ i.i.d. **from any distribution**
- **Parameter of interest**: any θ
- **Estimation method**: any method
- **Confidence interval**: $\left(\hat{\theta} - \epsilon_{1-\alpha/2}, \hat{\theta} - \epsilon_{\alpha/2}\right)$, where ϵ_p denotes the quantile of the error term at p
- Where does it come from?
  - Confidence interval: $100 * (1-\alpha)\%$ of the time, the error term $\hat{\theta} - \theta$ is within interval $(\epsilon_{\alpha/2}, \epsilon_{1-\alpha/2})$
  - $P(\epsilon_{\alpha/2} \leq \hat{\theta} - \theta \leq \epsilon_{1-\alpha/2}) = 1 - \alpha$
  - $P(\epsilon_{\alpha/2} \leq \hat{\theta} - \theta \leq \epsilon_{1-\alpha/2}) \iff P(\hat{\theta} - \epsilon_{1-\alpha/2} \leq \theta \leq \hat{\theta} - \epsilon_{\alpha/2})$
- But the error term ϵ_p is unknown - we had CLT for μ, but not for any θ

The idea of bootstrap is to approximate the error ϵ_p directly from data

Bootstrap example

Given a data set $\mathcal{X} = \{1, 2, 3, 4, 5\}$ with size $N = 5$ and $\hat{\theta} = median(\mathcal{X}) = 3$ estimated from this data set, construct CI with 95% confidence level

Intuition:

- $\hat{\theta} = 3$ is approximating the true median θ
- Construct statistic m_i to approximate $\hat{\theta}$
- Then we can use the "mock up" error $m_i - \hat{\theta}$ to approximate the actual error $\hat{\theta} - \theta$

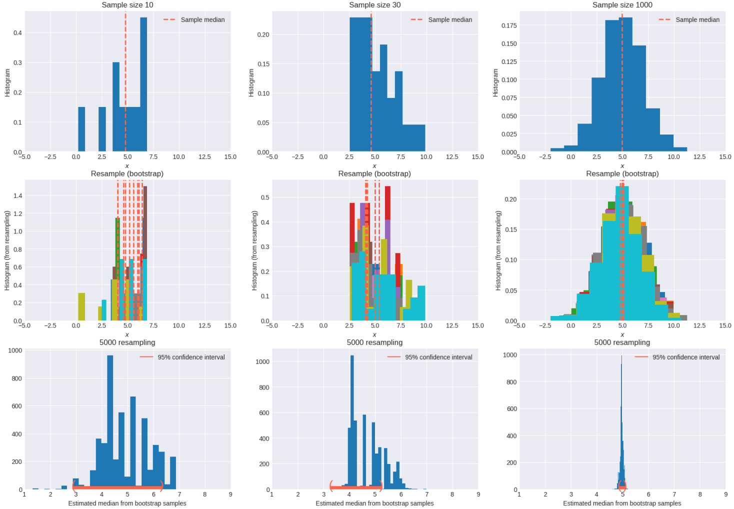
Steps:

- **Sample with replacement**

  Step 1.1: Randomly choose 5 elements from $\mathcal{X}$: $\mathcal{X}_1^* = \{1, 2, 1, 1, 4\}$
  Step 1.2: Compute the median from $\mathcal{X}_1^*$: $m_1 = 1.0$
  Step 2.1: Randomly choose 5 elements from $\mathcal{X}$: $\mathcal{X}_2^* = \{2, 5, 2, 4, 4\}$
  Step 2.2: Compute the median from $\mathcal{X}_2^*$: $m_2 = 4.0$
  ...

- Repeat this 100 times and get the set $\{m_1, \cdots, m_{100}\}$
- Compute $\epsilon^i = m_i - 3$ for $i = 1, \cdots, 100$
- Compute 0.025-quantile $\epsilon_{0.025}$ and 0.975-quantile $\epsilon_{0.975}$ from the set $\{\epsilon^1, \cdots, \epsilon^{100}\}$
- The 95% CI is constructed as $(3 - \epsilon_{0.975}, 3 - \epsilon_{0.025})$

Bootstrap example (cont.)

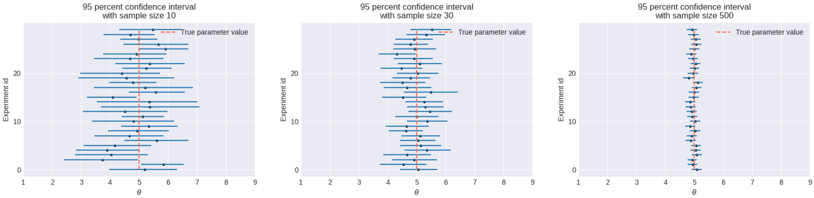


CI for unknown sampling distribution using bootstrap

- **Intuition**:
  - $\hat{\theta}$ is approximating θ
  - $\hat{\theta}_i$ is approximating $\hat{\theta}$
  - We can use $\hat{\theta}_i - \hat{\theta}$ to approximate $\hat{\theta} - \theta$
- **Steps** Given a data set $\mathcal{X}$ with size N and a statistic $\hat{\theta}$ computed from this data set, construct CI with $1 - \alpha$ confidence level:
  - Choose a large n
  - For $i = 1, \cdots, n$, repeat
    - Sample N elements from $\mathcal{X}$ with replacement: $\mathcal{X}_i^*$
    - Estimate the parameter of interest from $\mathcal{X}_i^*$: $\hat{\theta}_i$
    - Compute $\epsilon^i = \hat{\theta}_i - \hat{\theta}$
  - Compute $\alpha/2$-quantile $\epsilon_{\alpha/2}$ and $1 - \alpha/2$-quantile $\epsilon_{1-\alpha/2}$ from the set $\{\epsilon^1, \cdots, \epsilon^n\}$
  - The $1 - \alpha$ CI is constructed as $\left(\hat{\theta} - \epsilon_{1-\alpha/2}, \hat{\theta} - \epsilon_{\alpha/2}\right)$
- **Note**: there are many alternative methods for bootstrap; the exact method needs to be described when you talk about bootstrap

Confidence interval interpretation

- Confidence interval is **random** (data is random; statistic is random); the true parameter value θ is **not random** (illustrated in the image)
- A 95% confidence interval means that 95% of the time, the interval covers the true value θ



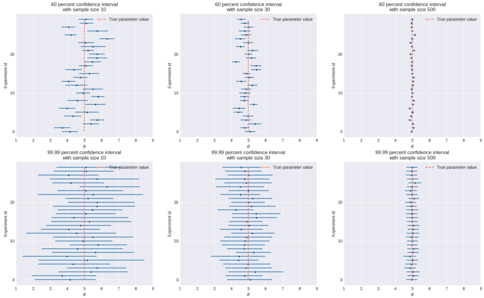
- Question 1: with the same problem setup, the larger the confidence level,
  - A. the wider the confidence interval
  - B. the narrower the confidence interval

  Answer: A
- Question 2: for a given confidence level, a good estimate has
  - A. a wide confidence interval
  - B. a narrow confidence interval

  Answer: B

Confidence level interpretation

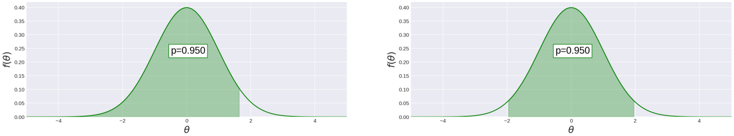
- If we compare 60% CI with 99.99% CI, the 60% CI does not always cover the true value $\theta = 5$ (it only covers it 60% of the time). On the other hand, the 99.99% CI covers the true value pretty much all the time. From this perspective, 99.99% CI is more meaningful to use as a quality measure.
- However, 99.99% CI can be very wide - of course - since it promises to cover the true value 99.99% of the time. A wide interval might not be meaningful sometimes, e.g. if you claim that you have estimated $\hat{\theta} = 4.3$ and you are 100% sure that the interval $(4.3 - \infty, 4.3 + \infty)$ contains the true value, your client might get mad.



Credible interval

Credible interval for Bayesian approach

- In maximum a posteriori estimation, the parameter of interest θ is modeled as **a random variable** - θ is generated from an underlying probability distribution described by $f(\theta)$
- Technically, any interval (a, b) with $P(a \leq \Theta \leq b) = 0.95$ is a 95% credible interval, but not all of them make sense, e.g.



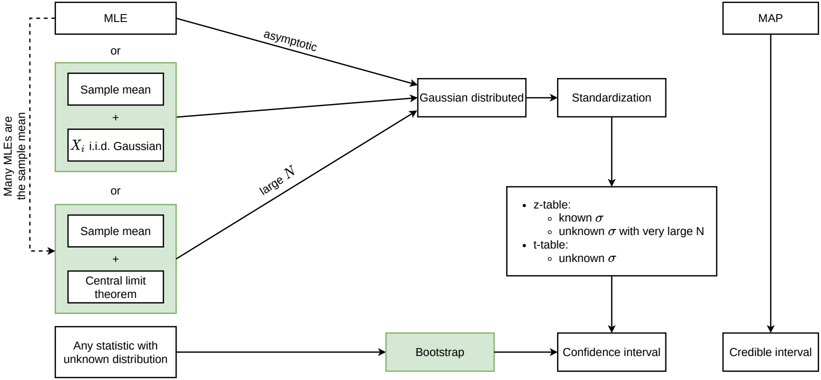
- There are different techniques for choosing this interval

Credible interval for Bayesian approach (cont.)

- In Python, for a given posterior (e.g. a standard Gaussian distribution $\mathcal{N}(0,1)$), the .interval method computes the interval with equal areas around the median:

```
posterior = stats.norm(loc=0, scale=1)
credible_interval = posterior.interval(0.95)
```

Recap


MLE
or
Sample mean
+
X_i i.i.d. Gaussian
or
Sample mean
+
Central limit theorem
Many MLEs are the sample mean
Any statistic with unknown distribution
asymptotic
large N
Gaussian distributed
Standardization
z-table:
known σ
unknown σ with very large N
t-table:
unknown σ
Bootstrap
Confidence interval
MAP
Credible interval

Today

1. Central limit theorem
2. Interval estimation
3. Summary

Summary

So far:

- Data types and data containers
- Descriptive data analysis: descriptive statistics, visualization
- Probability distributions, events, random variables, PMF, PDF, parameters
- CDF, Q-Q plot, how to compare two distributions (data vs theoretical, data vs data)
- Modeling
- Parameter estimation: maximum likelihood estimation (MLE) and maximum a posteriori estimation (MAP)
- Classification, multinomial naive Bayes classifier, Gaussian naive Bayes classifier
- Central limit theorem, interval estimation

Next:

- Clustering

Before next lecture:

- Bayes' rule, MLE

See you next week!