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
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**:

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

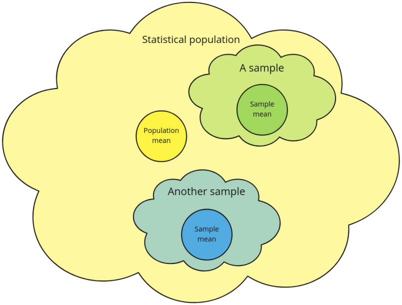
  - **Sample variance**:

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

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

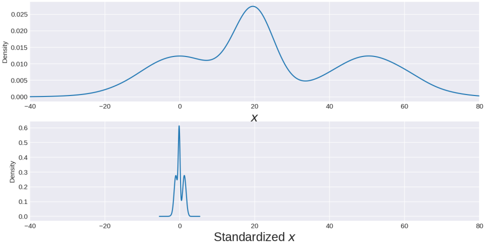
- Let $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ be a Gaussian random variable, then the following random variables are also Gaussian (location scale family)
  - Scaling (scale): $tX \sim \mathcal{N}(t\mu_X, t^2\sigma_X^2)$, $t \neq 0$ is a constant
  - Translation (location): $X + c \sim \mathcal{N}(\mu_X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

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

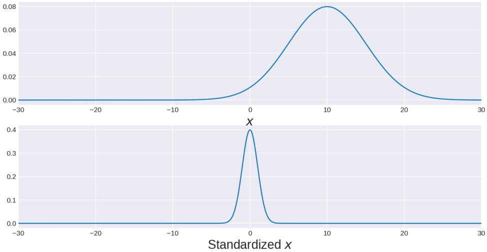


Question: what is the mean and standard deviation of Y? Random variable Y has **mean 0 and standard deviation 1**

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**)

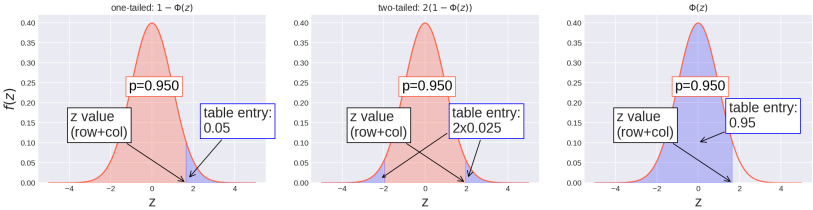
z	+ 0.00	+ 0.01	+ 0.02	+ 0.03	+ 0.04	+ 0.05	+ 0.06	+ 0.07	+ 0.08	+ 0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56360	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	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.)

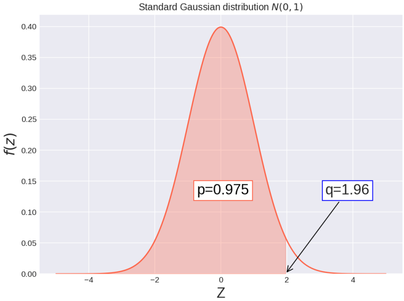
- There are different representations of the z-table; the difference is what is inside each cell, e.g. $\Phi(\text{row} + \text{column})$, $2(1 - \Phi(\text{row} + \text{column}))$, $1 - \Phi(\text{row} + \text{column})$ or $\frac{1}{2}(1 - \Phi(\text{row} + \text{column}))$; but the principle is the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



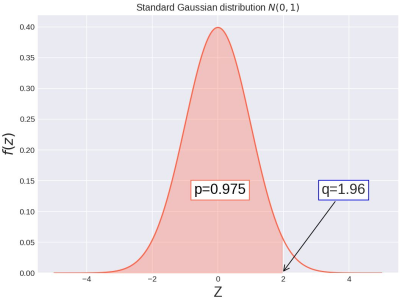
z-table

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

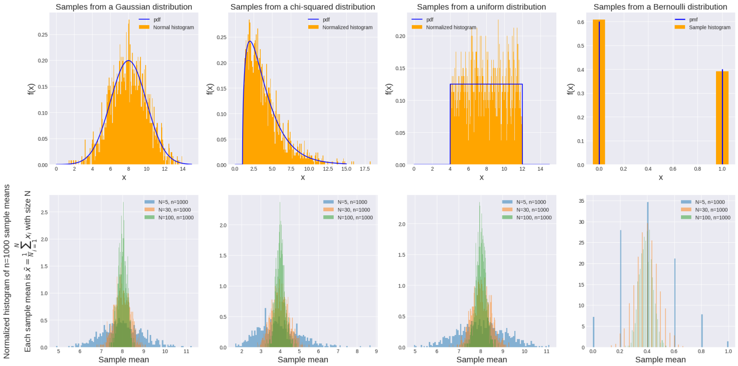
- You take another 30 ducks and measure the sample mean of their color t_i:

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

- 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.,

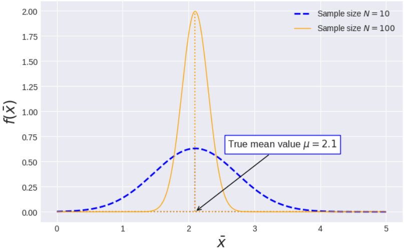
$$\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{N}) \tag{2}$$

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

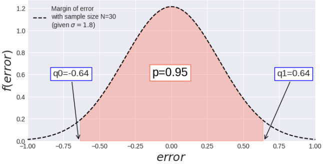
Estimation error $\bar{X} - \mu$

We are interested in 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 $\bar{X} - \mu$

Random variable: $X_1, \cdots, X_N$
Assumption: i.i.d. with **known** standard deviation σ and **unknown** mean μ

- In many use cases, we want to **estimate μ using the sample mean $\hat{\mu} = \bar{X}$** from **one sample** and we are interested in the **statistics of the estimation error**
- From CLT (cf. Eq. (22)), 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**!



- Interpretation of the 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 can also give a margin of error!
- This **95%** is called the **confidence level**; for a given confidence level, we can find a corresponding **interval** $(q0, q1)$

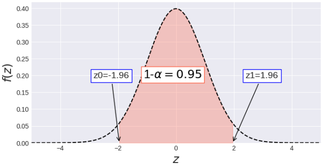
Calculate the margin of error

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

```
std = 1.8 # standard deviation of data
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 $Z \sim \mathcal{N}(0,1)$



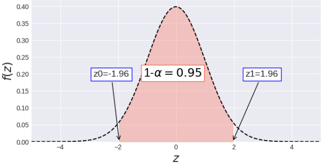
- We can use a two-tailed z-table (cf. page 13) to find the values for $z0$ and $z1$
- In order to find an interval for $\mathcal{E}$, we just need to look at

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

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)
```

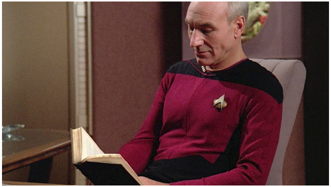
Find a standardized expression for the margin of error (cont.)

- 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

To Be Continued...



Today

Today