Data Analysis: Statistical Inference

Technological University Dublin Tallaght Campus

Department of Computing

Confidence intervals

What is a confidence interval?

- A confidence interval is a measure of how well a statistic, calculated on a data sample, represents a population parameter.
- The parameter for which a confidence interval is most commonly stated is the **mean**.
- A confidence interval is expressed in terms of a percentage-based confidence level (e.g. 95%) and a range within which the *actual parameter* is expected to be found with that level of confidence. For example, the confidence interval

  $$\mathbf{455.5 \pm 5.4}, \text{ with a confidence level of } \mathbf{99\%}$$

  states that we can be 99% confident that the parameter at hand is between 450.1 and 460.9. $\mathbf{455.5}$ is the value of the statistic (calculated on a sample).

The sampling distribution

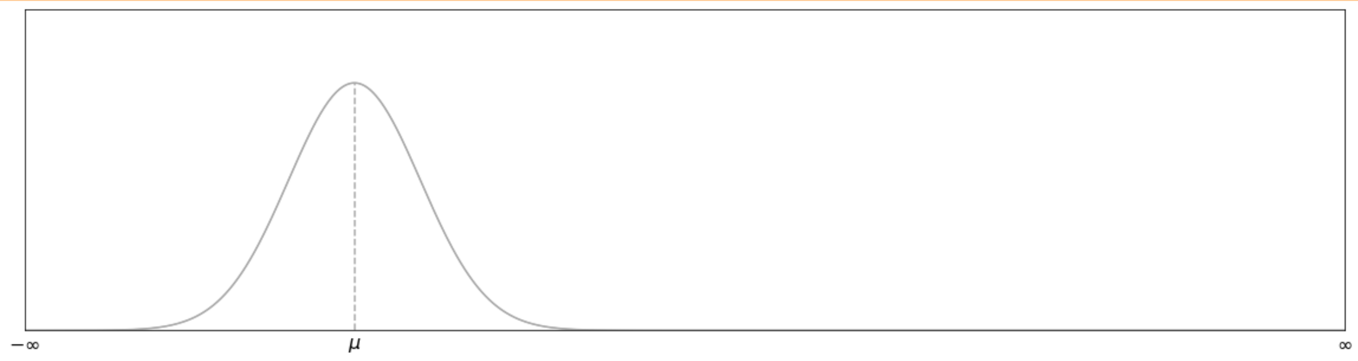
- Any statistic, being calculated from a sample, will vary between samples and consequently will have a *distribution* - **this is what allows us to define confidence intervals**.
- Take the mean: if it is calculated repeatedly for different samples drawn from a population, these values of the mean will vary and will be distributed in some way.
- The distribution of a statistic
  - is called a **sampling distribution**
  - has a **standard error** (corresponding to the standard deviation of a value distribution)
  - has an **expected value** (corresponding to the mean of a value distribution)

The sampling distribution of the mean

Now we focus on the **mean** - the statistic that is most commonly associated with confidence intervals. For the mean of a numeric variable:

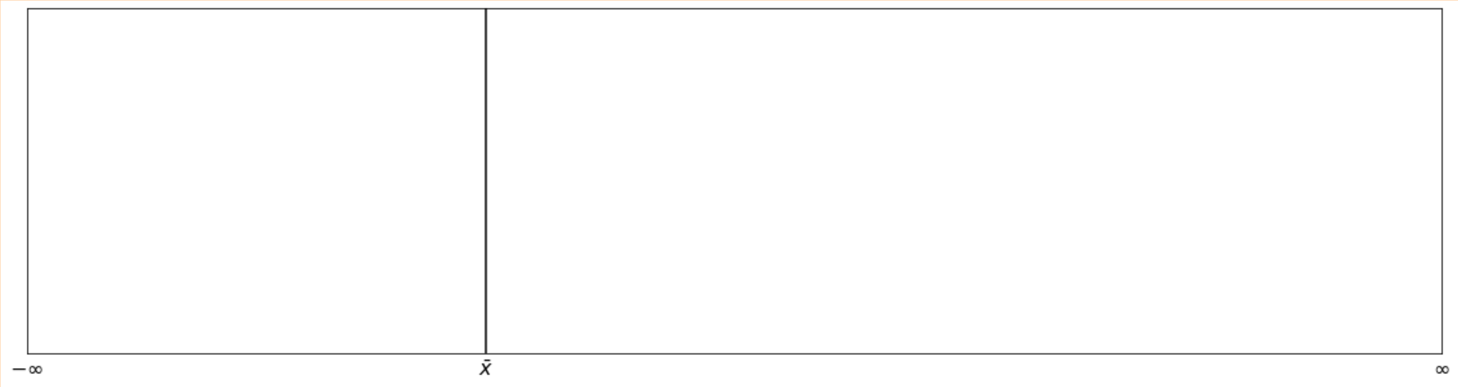
- the **sampling distribution** is *normal* in many cases:
  - when the value distribution is normal
  - if the value distribution is not normal but the sample size is greater than 30 (central limit theorem)

- the **standard error** is: $\boxed{\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}}$
  where $\sigma_{\bar{x}}$ is the standard error, σ is the standard deviation of the variable x and n is the sample size
- the **expected value** is equal to the population mean: $\boxed{E(\bar{x}) = \mu}$
  where $E(\bar{x})$ is the expected value for the mean and μ is the population mean.

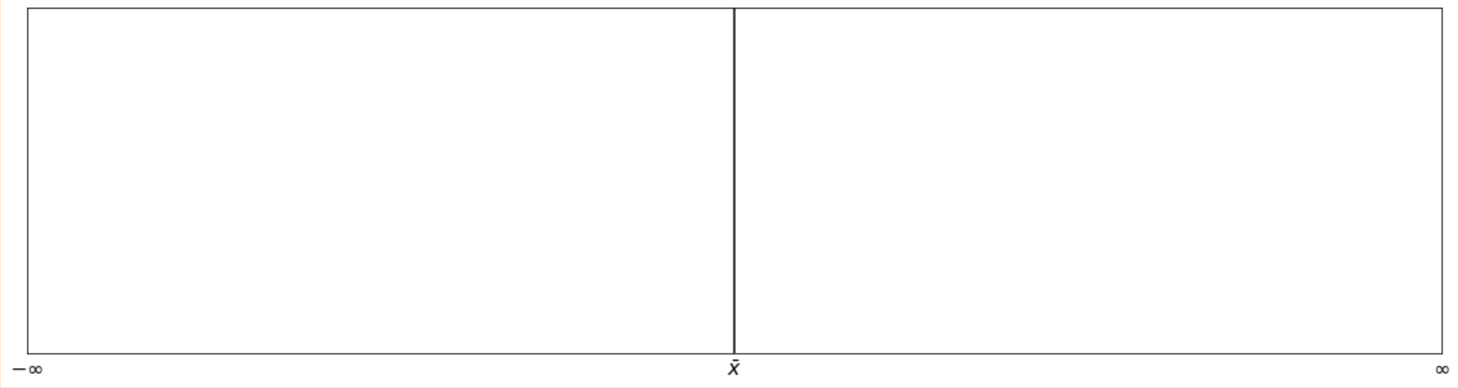


The concepts behind confidence intervals

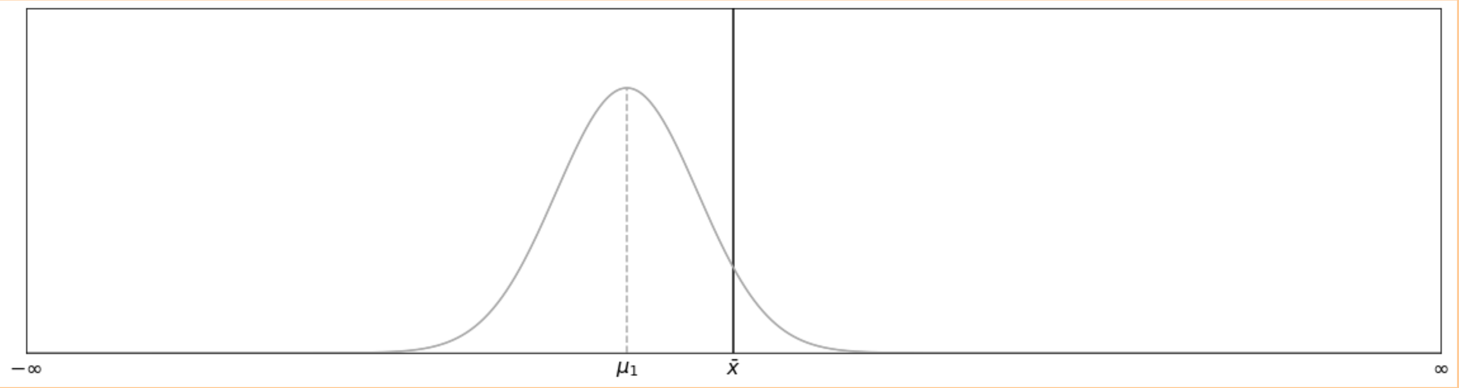
- Let's look at the situation where we know the standard deviation of our variable but do not know the mean. We have taken a sample and calculated the sample mean, $\bar{x}$:



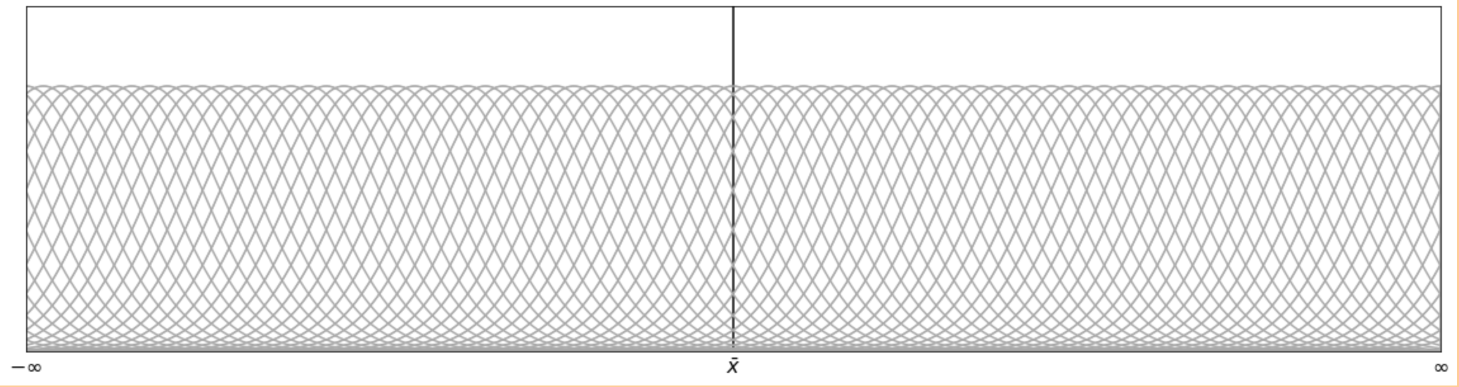
- We will place this sample mean in the middle of the picture, without loss of generality:



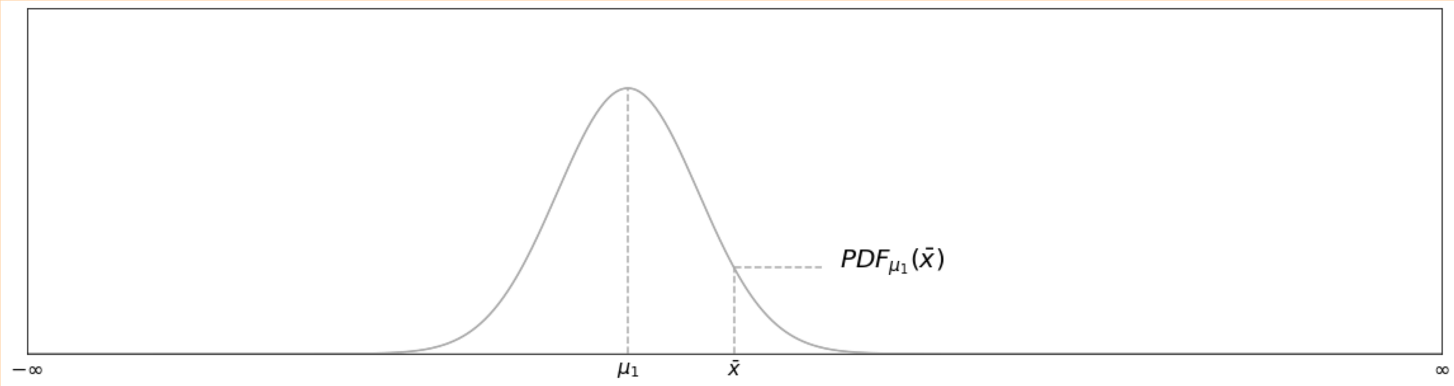
- This sample mean may have come from a distribution with expected value μ_1:



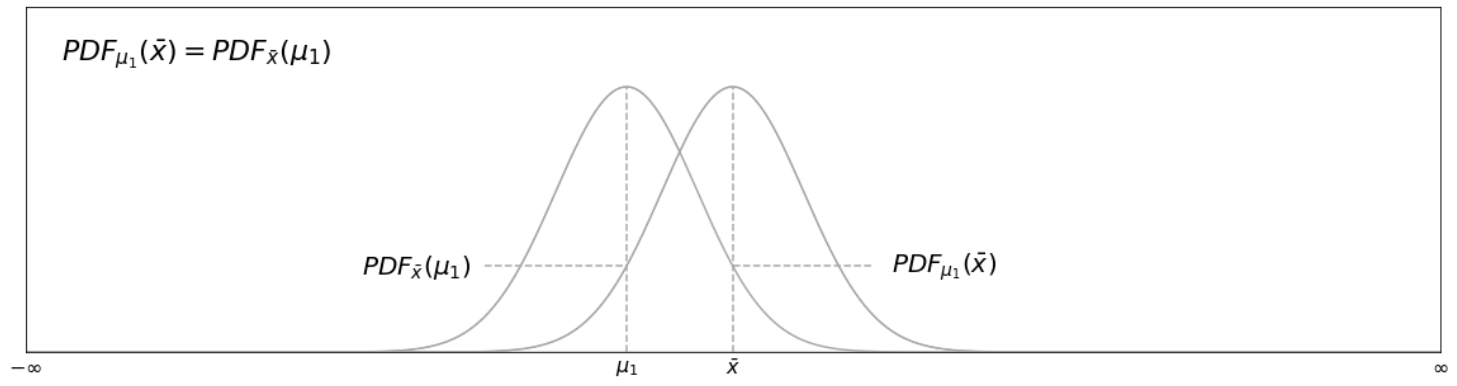
- But, it also may have come from any of an infinite set of equally probable distributions:



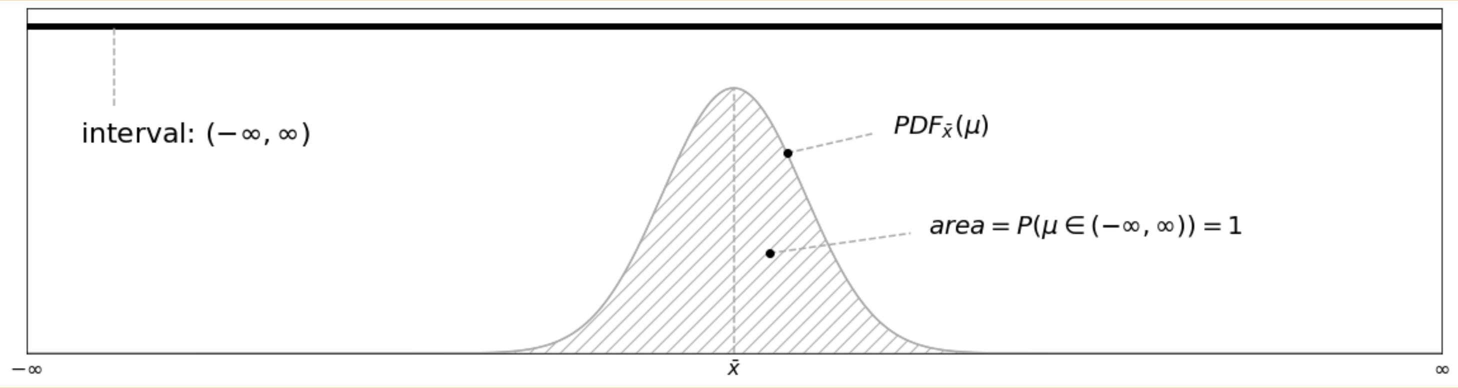
- Let us suppose that the real mean is some value μ_1. Its sampling distribution probability density function (PDF) is shown in the picture. Also shown is our sample mean $\bar{x}$ and the PDF value for $\bar{x}$ in that distribution.

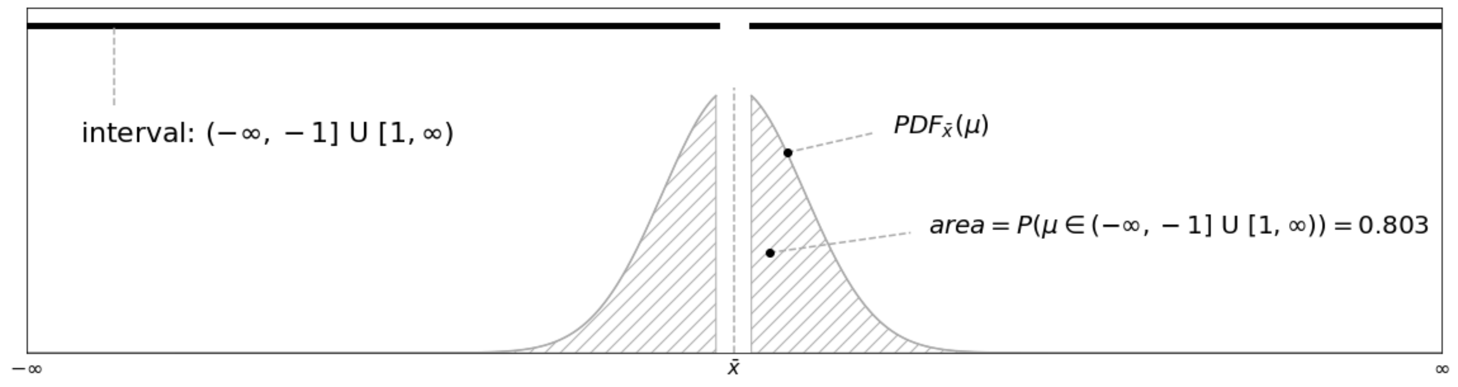


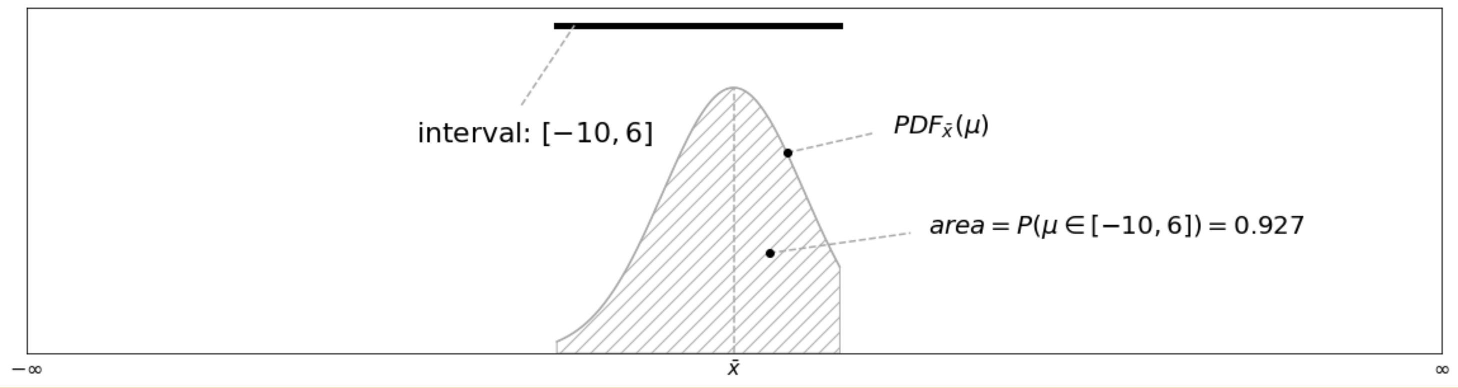
- Because of the way probability works, the probability density at $\bar{x}$ for a real mean of μ_1 is the same as the probability density at μ_1 for the sample mean of $\bar{x}$. This applies for all values of μ, meaning that we can construct a whole PDF around the calculated sample mean of $\bar{x}$ to express the distribution of μ value probabilities. When we are looking at probabilities for population parameter values, based on already calculated and hence fixed sample statistics, we call these probabilities **likelihoods**.

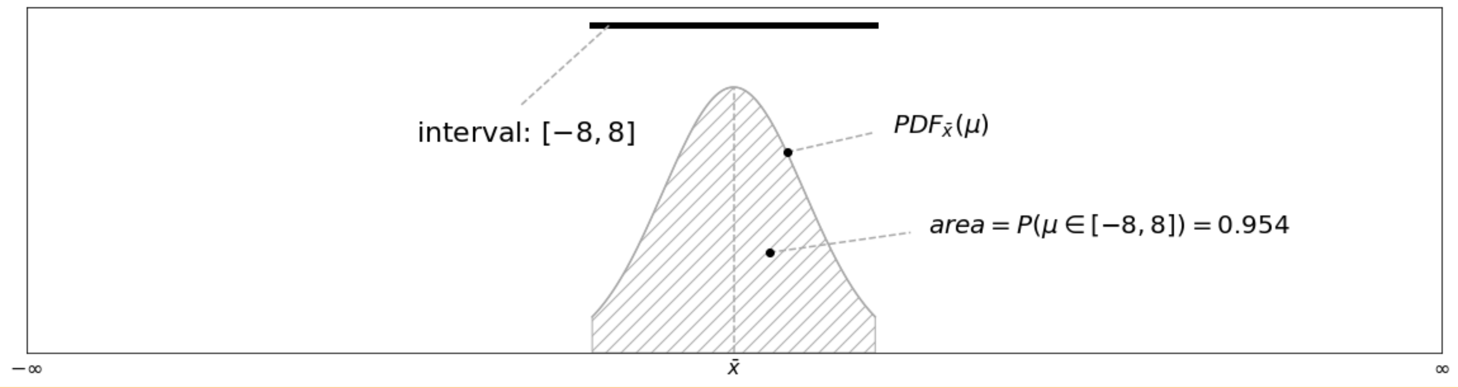


- The following pictures show some ranges of possible values of μ. Each picture shows:
  - the range as a thick line running at the top of the picture, with sections that are outside of the range cut out
  - the μ value likelihood PDF, hatched in the areas included in the range
  - the likelihood of μ belonging to the range, $P(\mu \in RANGE)$, which is calculated as the area under the PDF curve and which is 1 when all values are included (range $(-\infty, \infty)$)




interval: $(-\infty, -1] \cup [1, \infty)$
$PDF_{\bar{x}}(\mu)$
$area = P(\mu \in (-\infty, -1] \cup [1, \infty)) = 0.803$
$-\infty$
$\bar{x}$
∞


interval: $[-10, 6]$
$PDF_{\bar{x}}(\mu)$
$area = P(\mu \in [-10, 6]) = 0.927$
$-\infty$
$\bar{x}$
∞


interval: $[-8, 8]$
$PDF_{\bar{x}}(\mu)$
$area = P(\mu \in [-8, 8]) = 0.954$
$-\infty$
$\bar{x}$
∞

The meaning of a confidence interval

- A confidence interval states, with a certain level of confidence, that the real mean is in a particular range. The four pictures of ranges and likelihoods lead to the following statements:
  1. The mean is somewhere between $-\infty$ and ∞, with a confidence level of 100%.
  2. The mean is somewhere between $-\infty$ and $\bar{x} - 1$ or between $\bar{x} + 1$ and ∞, with a confidence level of 80%.
  3. The mean is somewhere between $\bar{x} - 10$ and $\bar{x} + 6$, with a confidence level of 93%.
  4. The mean is somewhere between $\bar{x} - 8$ and $\bar{x} + 8$, with a confidence level of 95%.

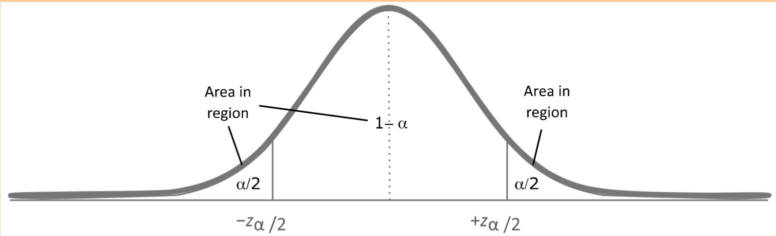
Confidence interval as used in statistics

- Most of the intervals listed above are not very useful ($-\infty$ to ∞?!)
- In statistics the interval used has the following properties:
  - is symmetrical around the point of highest likelihood (the sample mean, $\bar{x}$)
  - maximises the likelihood among all intervals of the same size
- It can be stated as:

  $\bar{x}\pm$ $<$half-interval for LoC$>$, with level of confidence $<$LoC$>$

Deriving a confidence interval in practice

- In practice, the distribution that is used is the z-distribution (a normal distribution with standard deviation 1 and mean 0):



Source: [MSD]

- If we define $\alpha = 1 - \frac{LoC}{100}$, where LoC is the percentual value of the level of confidence (e.g. if the required level of confidence is 95%, $LoC = 95$), then we can find two values on the x-axis, $-z_{\alpha/2}$ and $z_{\alpha/2}$, that 'fence off' an area under the distribution curve of $\frac{\alpha}{2}$ to the left and to the right, respectively.
- The above step is performed by looking up a table that maps $\frac{\alpha}{2}$ values to values for $z_{\alpha/2}$.
- The interval between $-z_{\alpha/2}$ and $z_{\alpha/2}$ is the normalised *confidence interval* for level of confidence $1 - \alpha$.
- The normalised confidence interval can be de-normalised i.e. brought back to be applicable to the distribution of the variable at hand.

HOWTO

Deriving a confidence interval

Example scenario: We know the population standard deviation ($\sigma = 10$) and the mean ($\bar{x} = 251$) and size ($n = 100$) of a sample.

---

1. Calculate $\alpha/2$: $\alpha/2 = \dfrac{1 - \dfrac{LoC}{100}}{2}$

   Example: For a level of confidence of 95% this is $\alpha = \dfrac{1 - \dfrac{95}{100}}{2} = 0.025$

2. Lookup cut-off value $z_{\alpha/2}$

   If the table (as the one on the next page) contains upper-tail values, we look for $\alpha/2$ in the table. If the table is for two-tailed values, we look up α.

   Example: The z-value corresponding to $\alpha/2$ of 0.025 is 1.96

3. De-normalise using the following formula: $CI_h = z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$, where CI_h is the half-interval.

   Example: $CI_h = 1.96 \times \dfrac{10}{\sqrt{100}} = 1.96$

4. State the confidence interval: $\bar{x} \pm z_{\alpha/2}(\dfrac{\sigma}{\sqrt{n}})$, with a confidence of LoC%

   Example: 251 ± 1.96, with 95% confidence

Upper-tail percentage points of the standard normal distribution

The table gives the values of z for which $P(Z > z) = p$, where the distribution of Z is $N(0, 1)$.

p	z	p	z	p	z	p	z	p	z
.50	0.000	.15	1.036	.025	1.960	.010	2.326	$.0^{3}4$	3.353
.45	0.126	.14	1.080	.024	1.977	.009	2.366	$.0^{3}3$	3.432
.40	0.253	.13	1.126	.023	1.995	.008	2.409	$.0^{3}2$	3.540
.35	0.385	.12	1.175	.022	2.014	.007	2.457	$.0^{3}1$	3.719
.30	0.524	.11	1.227	.021	2.034	.006	2.512	$.0^{4}5$	3.891
.25	0.674	.10	1.282	.020	2.054	.005	2.576	$.0^{4}1$	4.265
.24	0.706	.09	1.341	.019	2.075	.004	2.652	$.0^{5}5$	4.417
.23	0.739	.08	1.405	.018	2.097	.003	2.748	$.0^{5}1$	4.753
.22	0.772	.07	1.476	.017	2.120	.002	2.878	$.0^{6}5$	4.892
.21	0.806	.06	1.555	.016	2.144	.001	3.090	$.0^{6}1$	5.199
.20	0.842	.050	1.645	.015	2.170	$.0^{3}9$	3.121	$.0^{7}5$	5.327
.19	0.878	.045	1.695	.014	2.197	$.0^{3}8$	3.156	$.0^{7}1$	5.612
.18	0.915	.040	1.751	.013	2.226	$.0^{3}7$	3.195	$.0^{8}5$	5.731
.17	0.954	.035	1.812	.012	2.257	$.0^{3}6$	3.239	$.0^{8}1$	5.998
.16	0.994	.030	1.881	.011	2.290	$.0^{3}5$	3.291	$.0^{9}5$	6.109

Confidence interval for different distributions and sample sizes

While the way we define the confidence interval is in principle always the same, there are three different variants of the method, with applicability depending on:

- whether the standard deviation of the data is known or not
- the distribution of the data
- the size of the sample

The tables show the applicable methods for different combinations of the relevant factors. An **X** indicates a case where a confidence interval cannot be defined.

Standard deviation known (σ)

Sample size → **Distribution ↓**	large	small
Normal	1	1
Any other	1	X

Standard deviation unknown

Sample size → **Distribution ↓**	large	small
Normal	2	3
Any other	2	X

Variant 1 is what has been described already. A look at variants 2 and 3 follows on the next page.

Confidence interval for unknown standard deviation and large sample size (variant 2)

When the sample size is large, the sampling distribution tends towards being normal, regardless of the distribution of the data itself. This means that we can use the standard deviation estimate based on a sample (S) instead of the actual standard deviation (σ) and then apply the same method to derive the confidence interval as with a known σ.

HOWTO

Deriving a confidence interval (variant 2)

Example scenario: We do not know the population standard deviation but we have a sample with standard deviation $(S = 10)$, mean $(\bar{x} = 251)$ and size of at least 30 $(n = 100)$.

Proceed as in variant 1 but using S instead of σ.

Confidence interval for unknown standard deviation and small sample size (variant 3)

When the sample size is small, we can still define a confidence interval but only for the case that *the data distribution is known to be normal*. In this case we must use a t-distribution instead of a z-distribution. The t-distribution:

- depends on the sample size
- has heavier tails owing to greater uncertainty (because of smaller samples)

See a graph of some t-distribution PDFs on Wikipedia.

HOWTO

Deriving a confidence interval (variant 3)

Example scenario: We do not know the population standard deviation but we have a sample with standard deviation ($S = 10$), mean ($\bar{x} = 251$) and size smaller than 30 ($n = 16$). Because the sample is small we need to know that the data is normally distributed, otherwise a confidence interval cannot be specified.

---

1. Calculate $\alpha/2$: $\alpha/2 = \dfrac{1 - 0.01LoC}{2}$

   Example: For a level of confidence of 95% this is $\alpha = \dfrac{1 - 0.95}{2} = 0.025$
2. Lookup cut-off value in the t-table, for α and degrees of freedom $df = n - 1$. Look for $\alpha/2$ among upper-tail values or for α among two-tailed values. The result will be the same.

   Example: The cut-off value corresponding to $df = 15$ and α of 0.5 two-tailed is $CO_t = 2.13$
3. De-normalise using the following formula: $CI_h = CO_t \dfrac{S}{\sqrt{n}}$, where CI_h is the half-interval.

   Example: $CI_h = 2.13 \times \dfrac{10}{\sqrt{16}} = 5.33$
4. State the confidence interval: $\bar{x} \pm CO_t(\dfrac{S}{\sqrt{n}})$, with a confidence of LoC%

   Example: 251 ± 5.33, with 95% confidence

T Distribution Table

α (1 tail)	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
α (2 tail)	0.1	0.05	0.02	0.01	0.005	0.002	0.001
df							
1	6.3138	12.7065	31.8193	63.6551	127.3447	318.4930	636.0450
2	2.9200	4.3026	6.9646	9.9247	14.0887	22.3276	31.5989
3	2.3534	3.1824	4.5407	5.8408	7.4534	10.2145	12.9242
4	2.1319	2.7764	3.7470	4.6041	5.5976	7.1732	8.6103
5	2.0150	2.5706	3.3650	4.0322	4.7734	5.8934	6.8688
6	1.9432	2.4469	3.1426	3.7074	4.3168	5.2076	5.9589
7	1.8946	2.3646	2.9980	3.4995	4.0294	4.7852	5.4079
8	1.8595	2.3060	2.8965	3.3554	3.8325	4.5008	5.0414
9	1.8331	2.2621	2.8214	3.2498	3.6896	4.2969	4.7809
10	1.8124	2.2282	2.7638	3.1693	3.5814	4.1437	4.5869
11	1.7959	2.2010	2.7181	3.1058	3.4966	4.0247	4.4369
12	1.7823	2.1788	2.6810	3.0545	3.4284	3.9296	4.3178
13	1.7709	2.1604	2.6503	3.0123	3.3725	3.8520	4.2208
14	1.7613	2.1448	2.6245	2.9768	3.3257	3.7874	4.1404
15	1.7530	2.1314	2.6025	2.9467	3.2860	3.7328	4.0728
16	1.7459	2.1199	2.5835	2.9208	3.2520	3.6861	4.0150
17	1.7396	2.1098	2.5669	2.8983	3.2224	3.6458	3.9651
18	1.7341	2.1009	2.5524	2.8784	3.1966	3.6105	3.9216
19	1.7291	2.0930	2.5395	2.8609	3.1737	3.5794	3.8834
20	1.7247	2.0860	2.5280	2.8454	3.1534	3.5518	3.8495

Hypothesis tests

What is a hypothesis test?

- A hypothesis test is used to check if a sample of data supports a particular hypothesis made about the population from which the sample was drawn.
- Specifically, what is effectively tested is whether the hypothesised parameter value falls within the appropriate confidence interval, defined using a sample statistic.
- What we are really interested in is whether the sample statistic is far enough from the hypothesised parameter value to signal its low likelihood. It is such an extreme value (rather than one within the confidence interval) that is considered interesting i.e. a 'positive' or a detection.
- The most commonly tested parameter (and the only one covered here) is the population mean.

Hypothesis testing terminology

- The **threshold of probability** or **level of significance** is the probability corresponding to α in the description of the confidence interval (see previous slides) and common values used are 5% and 1%.
- A hypothesis test is stated through a null hypothesis and an alternative hypothesis, for example, if the hypothesis is that the mean is equal to 100:

$H_0 : \mu = 100$ **null hypothesis**

$H_a : \mu \neq 100$ **alternative hypothesis**

- The possible outcomes of a hypothesis test are:
  - the null hypothesis is **accepted**
  - the null hypothesis is **rejected**

Performing a hypothesis test

Hypothesis testing steps:

- state the null and alternative hypotheses
- calculate the test statistic, which in the case of the mean is: $T = \frac{\bar{x} - \mu_0}{SE}$, where $\bar{x}$ is the sample mean, μ_0 is the mean specified in the null hypothesis and SE is the *standard deviation*
- compare the absolute value of the test statistic $(|T|)$ with the critical value (CV) read from the table for the *sampling distribution* for the required level of significance, then:
  - **accept the null hypothesis** if $|T| \leq CV$
  - **reject the null hypothesis** if $|T| > CV$

Hypothesis test variants

As with confidence intervals, there are 3 variants of the hypothesis testing method, differing by:

- how the standard error (SE) is calculated, which is either
  - from the population standard deviation (the parameter, σ) **or**
  - from the sample standard deviation (the statistic, S)
- which sampling distribution needs to be used:
  - z-distribution
  - t-distribution

The variants:

1. sampling distribution: **z**, standard error: $SE = \sigma$

   Applicable in any of the following cases:

   - the standard deviation (σ) is known and the data distribution is normal
   - the standard deviation (σ) is known and the sample is large ($n \geq 30$), regardless of distribution

2. sampling distribution: **z**, standard error: $SE = S$

   Applicable when the standard deviation (σ) is not known and the sample is large ($n \geq 30$), regardless of distribution.

3. sampling distribution: **t**, standard error: $SE = S$

   Applicable when the standard deviation (σ) is not known, the sample is small($n < 30$) and the data distribution is normal.

When the data distribution is not known and the sample is small a hypothesis test cannot be performed reliably.

HOWTO

Hypothesis test

Example scenario: The standard deviation is not known but we have a sample of size $n = 100$, with calculated sample standard deviation of $S = 5.55$ and a mean of $\bar{x} = 64.32$.

---

1. Decide which variant of the hypothesis test is appropriate.

   Example: Variant 2 (σ not known, large sample as $100 > 30$) $\Rightarrow SE = \dfrac{S}{\sqrt{n}}$, sampling distribution is **z**

2. State the null and alternative hypotheses:

   Example: Assuming that the hypothesised mean is $\mu_0 = 65$

   $H_0 : \mu = 65$

   $H_a : \mu \neq 65$

3. Calculate the statistic $T = \dfrac{\bar{x} - \mu_0}{SE}$

   Example:

   $$T = \frac{64.32 - 65}{\frac{5.55}{\sqrt{100}}} = -1.225$$

4. Lookup critical value and compare

   Example: Assuming a level of significance of 5%, the critical value for the z-distribution is 1.96

   $|-1.225| = 1.225 < 1.96 \Rightarrow$ null hypothesis **accepted**

Hypothesis test errors

- Type I error - when the null hypothesis is rejected even though it is true

  The probability of a Type I error is exactly equal to the level of significance.

- Type II error - when the null hypothesis is accepted even though it is not true

  The probability of a Type II error depends on the difference between the real and hypothesised value, relative to the standard error, and on the level of significance.

- Changing the probability of errors:
  - The probability of a Type II error can be reduced by increasing the level of significance, but this simultaneously increases the probability of a Type I error.
  - The probability of a type II error can be reduced by increasing the sample size.

References The pictures in this presentation were taken from the following books. The source for each picture is cited beside it.

[DSB] *Data Science for Business: What you need to know about data mining and data-analytic thinking*, by Foster Provost and Tom Fawcett, O'Reilly Media, 2013.

[MSD] *Making Sense of Data I: A Practical Guide to Exploratory Data Analysis and Data Mining*, by Glenn J. Myatt and Wayne P. Johnson, John Wiley & Sons, 2014.

[US] Understanding Statistics, by Graham Upton and Ian Cook, Oxford University Press, 1996.