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

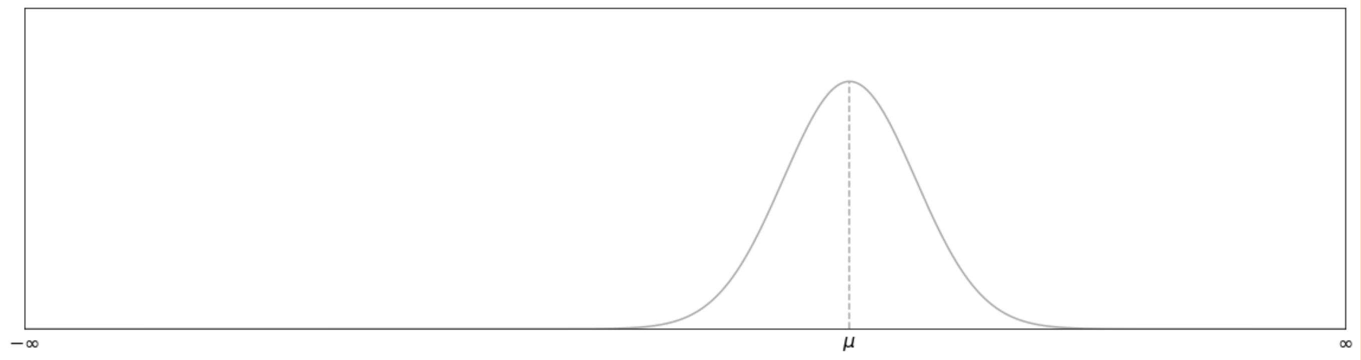
Now we focus on the **mean** - the statistic that is most commonly associated with confidence intervals. For the mean of a numeric variable:

- 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: $\boxed{\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}}$

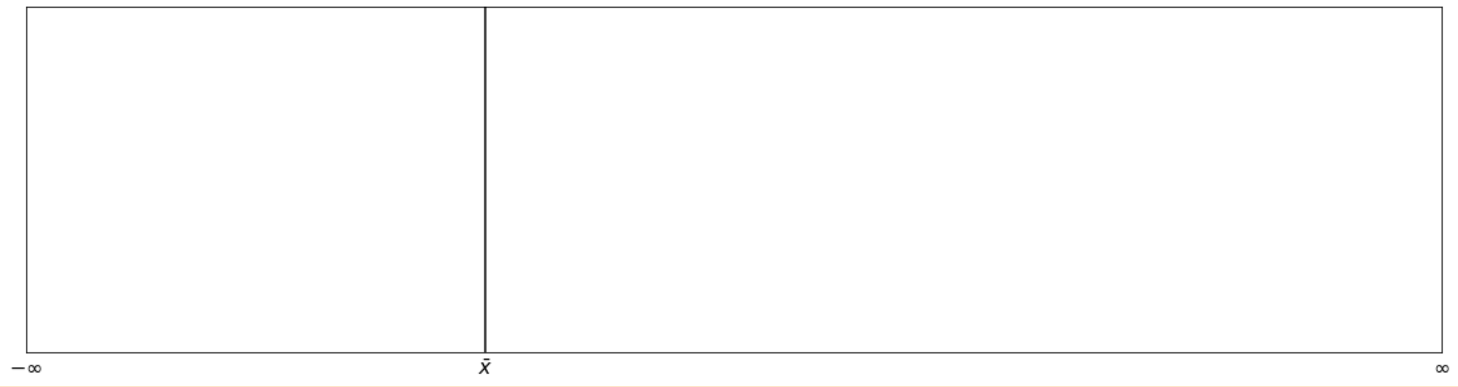
  where $\sigma_{\bar{x}}$ is the standard error, σ is the standard deviation of the variable x and n is the sample size
- the **expected value** is equal to the population mean: $\boxed{E(\bar{x}) = \mu}$

  where $E(\bar{x})$ is the expected value for the mean and μ is the population mean.

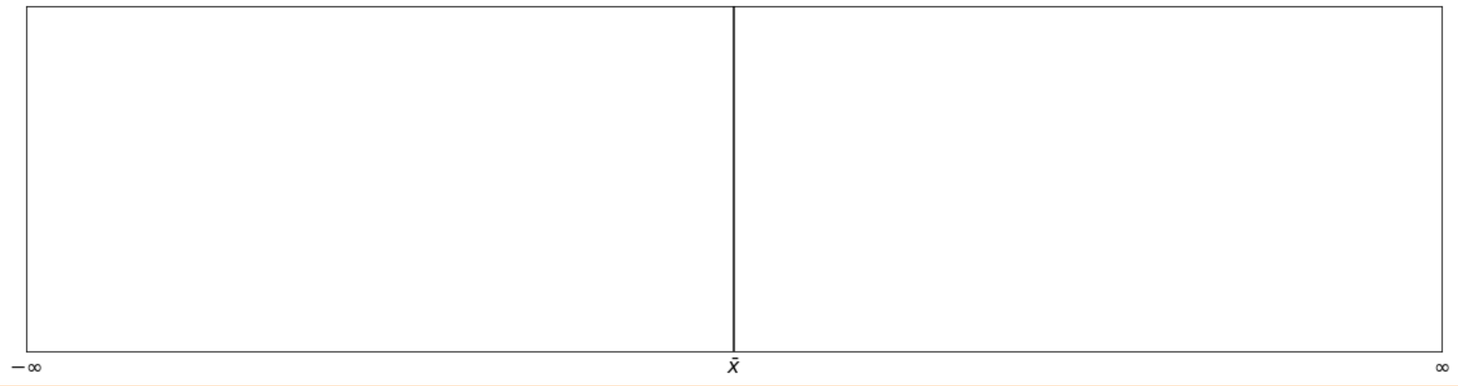


The principle behind confidence intervals

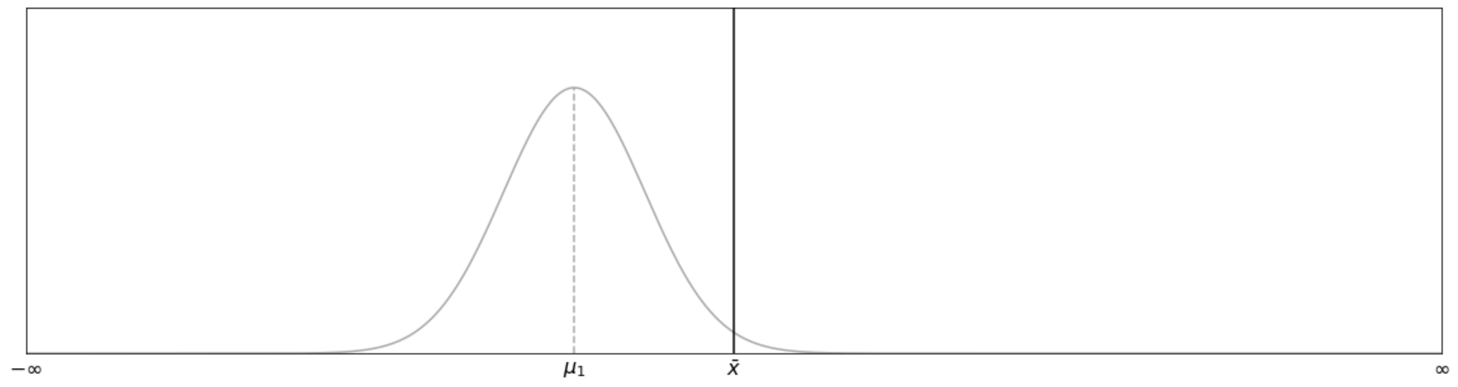
Let's look at the situation where we know the standard deviation of our variable but do not know the mean. We have taken a sample and calculated the sample mean, $\bar{x}$:



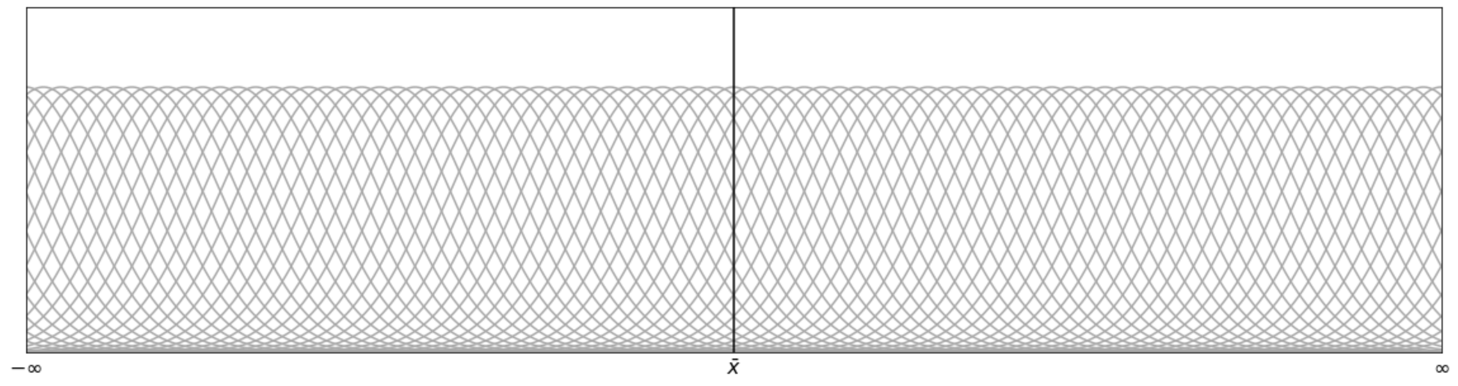
We will place this sample mean in the middle of the picture, without loss of generality:



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

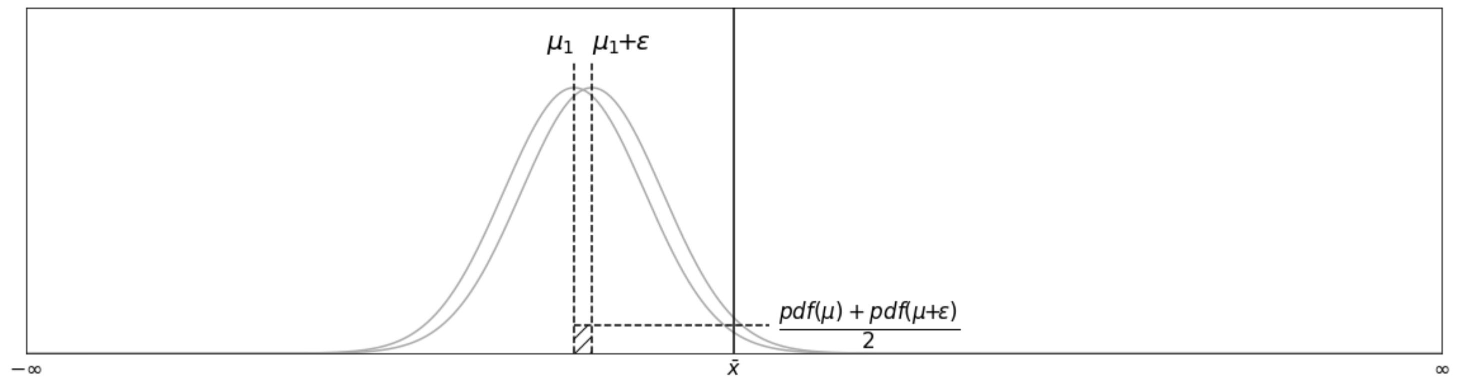


But, it also may have come from any of an infinite set of equally probable distributions:

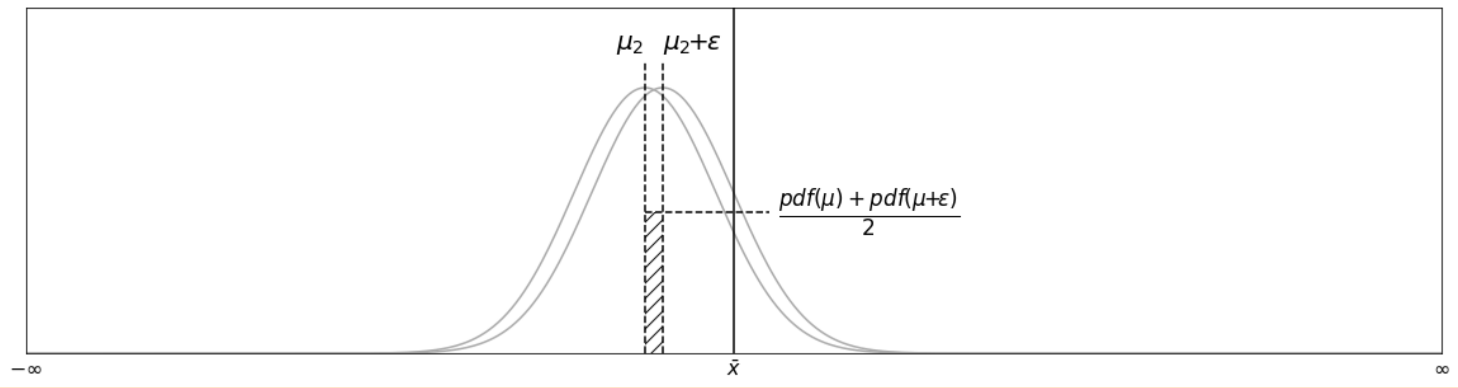


We can think of the full range of possible values for μ as consisting of a very large number (tending to infinity) of very narrow (tending to zero) ranges. We will label the width of the narrow range ϵ and the number of ranges N_ϵ.

Now, let's look at one such range, extending to the right from μ_1. In the case that the real mean (μ) were in this range, the probability of the sample mean having the calculated value ($\bar{x}$) would be approximately equal to the area of the shaded rectangle (and get closer to exactly equal as ϵ gets smaller).



The probability of the sample mean having the obtained value, shown as the area of a shaded rectangle for another value of μ, $\mu 2$.

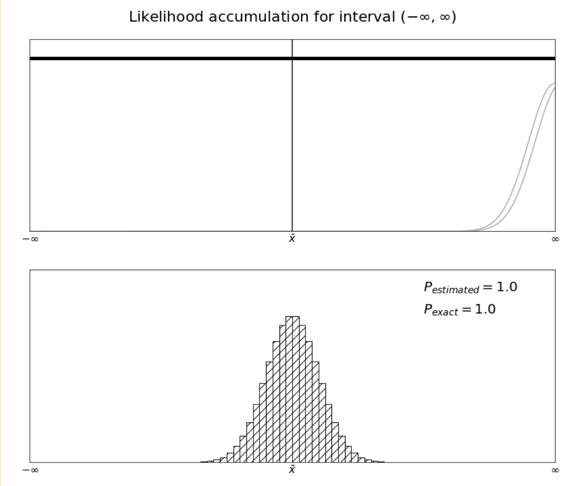


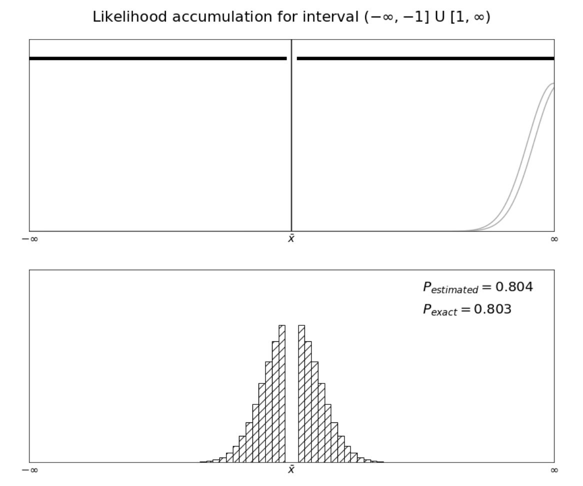
When we draw a sample and calculate the sample mean, that mean has become a certain outcome i.e. it's probability after the fact is 1. The 'narrow range probabilities' (represented by shaded rectangle areas) from the previous two pictures represent the **likelihoods** of the real value of μ being in the respective ranges ($[\mu_1, \mu_1 + \epsilon]$, $[\mu_2, \mu_2 + \epsilon]$). These can be viewed as contributions by the various narrow ranges of μ towards the outcome (obtained value for $\bar{x}$) probability of 1 (it has already happened).

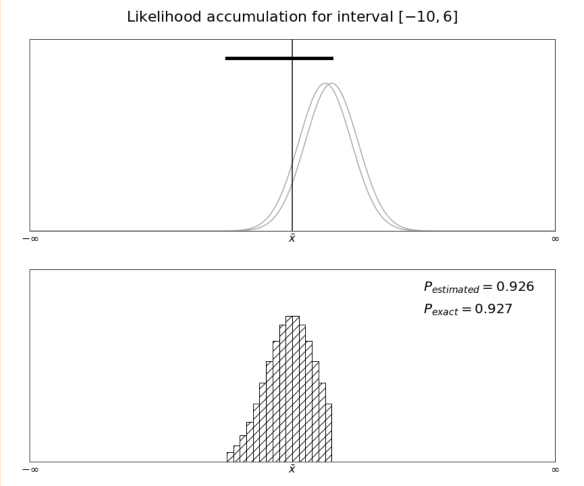
Click on the pictures on the following page to see a videos of how these likelihoods accumulate across the full range and some subranges of values for μ.

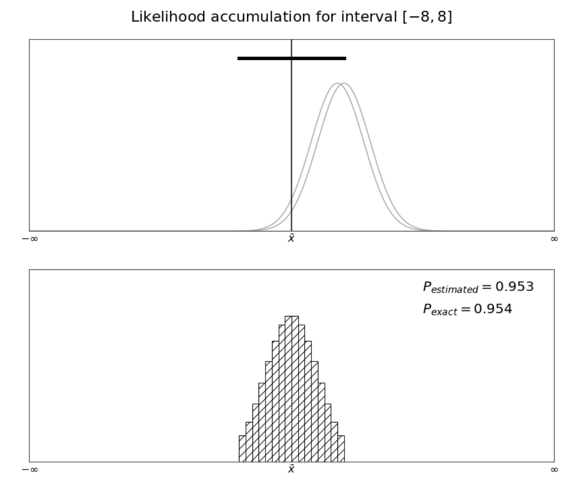
The cumulative likelihood tells us how confident we can be that the real value of μ is in the subrange.

Likelihood accumulations


Likelihood accumulation for interval $(-\infty, \infty)$


Likelihood accumulation for interval $(-\infty, -1]$ U $[1, \infty)$


Likelihood accumulation for interval $[-10, 6]$


Likelihood accumulation for interval $[-8, 8]$

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. Our videos 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 93%.

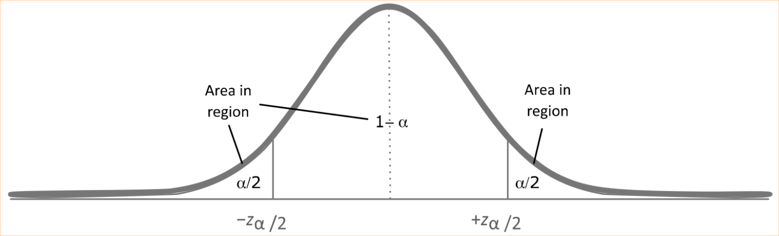
Confidence interval as used in statistics

- 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
- Hence, 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

Assumptions:

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

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

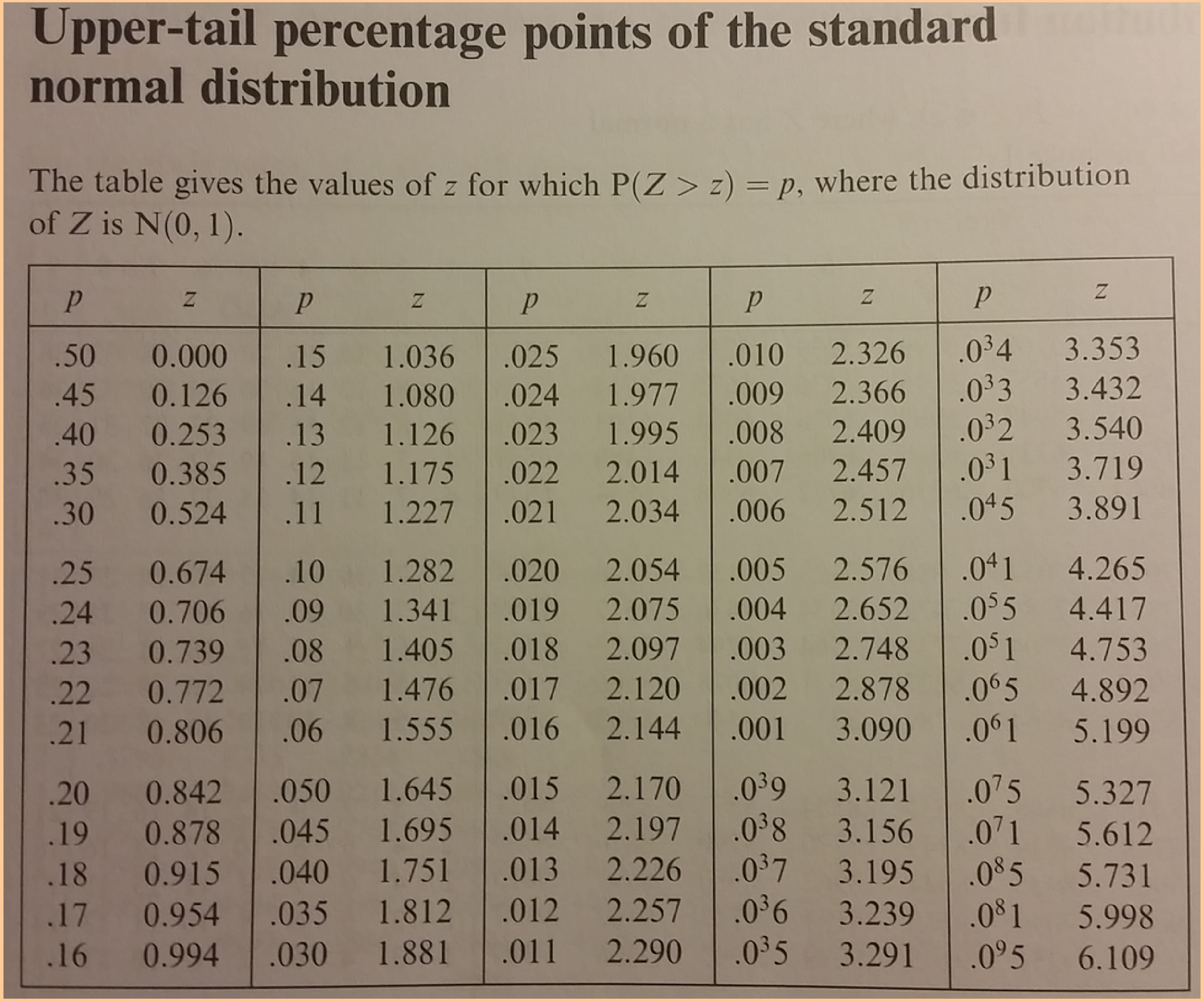
   Example: For a level of confidence of 95% this is $\alpha = \frac{1 - \frac{95}{100}}{2} = 0.025$
2. Lookup cut-off value $z_{\alpha/2}$

   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: $CI_h = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, where CI_h is the half-interval, σ is the standard deviation of our variable and n is the size of the sample.

   Example: $CI_h = 1.96 \times \frac{10}{\sqrt{10}} = 1.96$
4. State the confidence interval: $\bar{x} \pm z_{\alpha/2}(\frac{\sigma}{\sqrt{n}})$, 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

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.

Temporary page!

LATEX was unable to guess the total number of pages correctly. As there was some unprocessed data that should have been added to the final page this extra page has been added to receive it.

If you rerun the document (without altering it) this surplus page will go away, because LATEX now knows how many pages to expect for this document.