Introduction to inference

Population vs. Sample

- Population is set of entities with some common feature.
- Sample is a representative subset of the population. It must be chosen according to:
  - Unbiasedness: same probability
  - Representativeness: same proportion
  - Dimension
- Random sample: All elements of the population have the same probability of being chosen.

Population Probability Distribution

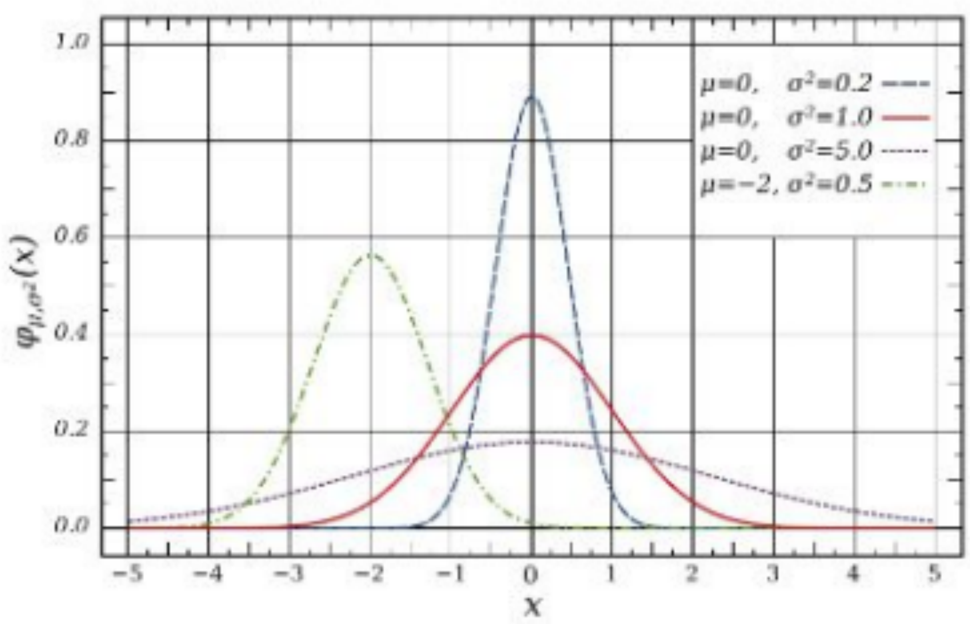
- Population probability distribution is the list of values and corresponding probabilities that a population can have.

x	**abs. freq.**	**Pr(x)**
5	1	1/5
7	2	2/5
8	1	1/5
12	1	1/5
	N=5	$\sum = 1$

- From the population probability distribution we can compute parameters, such as mean μ and standard deviation σ.

Normal Distribution

- The normal distribution is a continuous probability distribution with two parameters: μ and σ.



- The standard normal population can be used to compute the probability of a given interval for any μ and σ.

Standard Normal Distribution

- The standard normal distribution is a normal distribution with $\mu = 0$ and $\sigma = 1$.

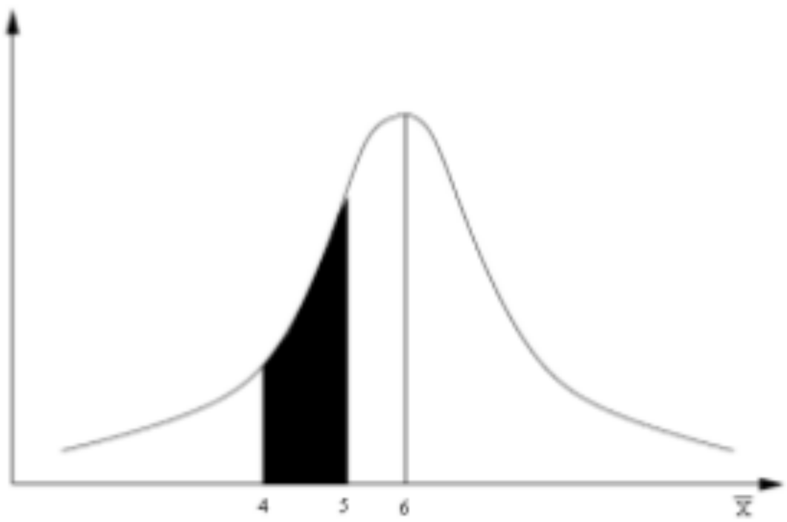
- Normal distributions can be **transformed** to the standard normal distribution by the formula:

$$Z = \frac{x - \mu}{\sigma}$$

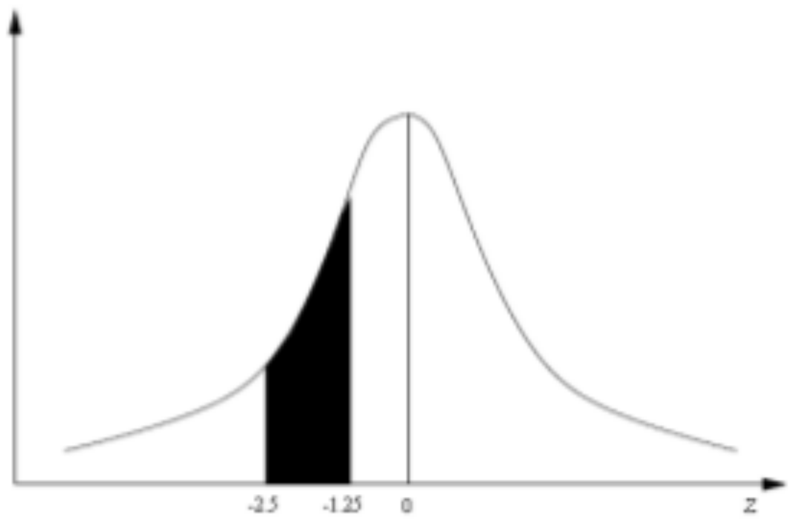
where x is the score from the original distribution

Standard Normal Distribution

- The probabilities associated to the standard normal distribution are tabulated. (more about this later on)



$\Pr(4 \leq X \leq 5) = ?$



$\Pr(-2.5 \leq Z \leq -1.25) = 9.94\%$

Sampling distributions

- A sample $x_1, \ldots, x_n$ is a representative subset of the population.
- Each element x_i is a random variable. Thus, each x_i has the same probability distribution of the population.
- The sample mean $\bar{x}$ changes according to the sample.
- Then, $\bar{x}$ is also a random variable and it has a probability distribution.

Sampling distribution of means

- Consider all samples of 3 elements (there are $\binom{5}{3}$= 10 possible samples) and compute the sample mean for each one.

$\bar{x}$	**abs. freq.**	**Pr($\bar{x}$)**
6.(3)	1	1/10
6.(6)	2	2/10
7.(3)	1	1/10
8	2	2/10
8.(3)	1	1/10
8.(6)	1	1/10
9	2	2/10
	N=10	$\sum = 1$

Sampling distribution of means

- Mean of the sampling distribution of means ($\mu_{\bar{x}}$) is equal to μ:

$$\mu_{\bar{x}} = \mu$$

- Standard deviation of the sampling distribution of means ($\sigma_{\bar{x}}$) is equal to σ, divided by the root square of sample size (n):

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

Note: there is a correction of $\sigma_{\bar{x}}$ for large samples

Sampling distribution of means

- If a given population follows **a normal distribution** with mean μ and standard deviation σ, then the sampling distribution of means also follows a normal distribution with the following parameters:

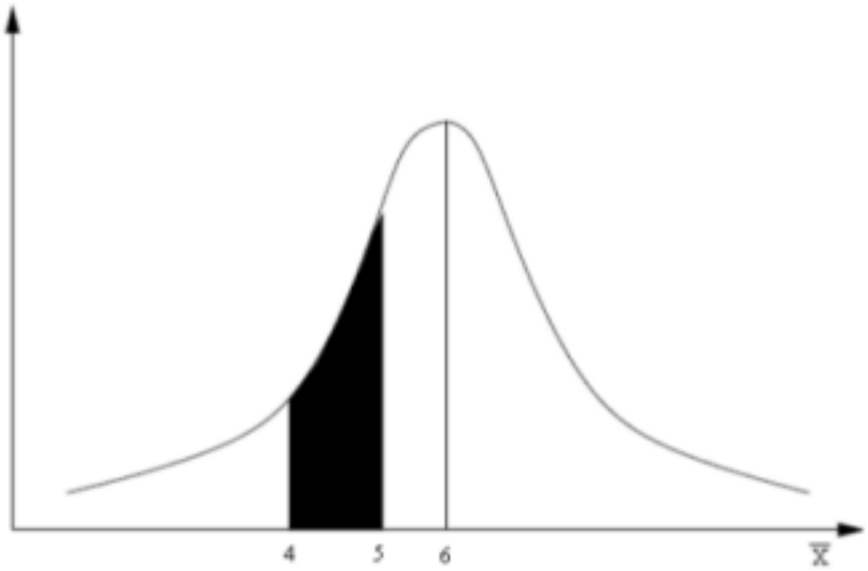
$$\mu_{\bar{x}} = \mu \qquad \text{and} \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Sampling distribution of means

Example: The time of user connection to my blog follows a normal distribution with a mean of 6 minutes and a standard deviation of 4 minutes. In a random sample of 25 user connections, which is the probability that they take between 4 and 5 minutes, in average?

$$\mu_{\bar{x}} = \mu = 6$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4}{5}$$



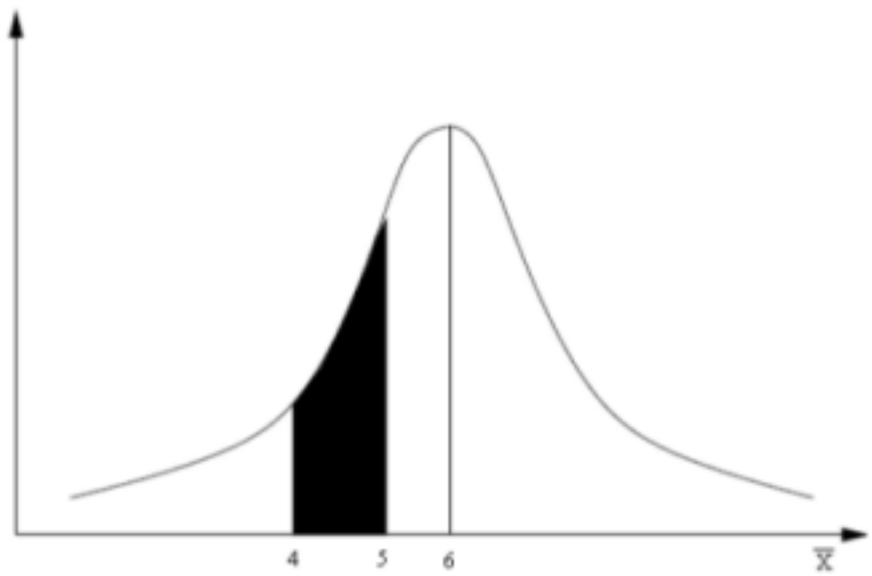
$$\Pr\,(4 \leq \bar{x} \leq 5) = ?$$

Sampling distribution of means

Example: The time of user connection to my blog follows a normal distribution with a mean of 6 minutes and a standard deviation of 4 minutes. In a random sample of 25 user connections, which is the probability that they take between 4 and 5 minutes, in average?

$$\mu_{\bar{x}} = \mu = 6$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4}{5}$$



$$\Pr(-2.5 \le Z \le -1.25) = 9.94\%$$

Sampling distribution of means

- If a given population with **unknown distribution** with mean μ and standard deviation σ, then the sampling distribution of means, for **increasing *n***, also follows a normal distribution with the following parameters:

$$\mu_{\bar{x}} = \mu \qquad \text{and} \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Also known as the Central Limit Theorem

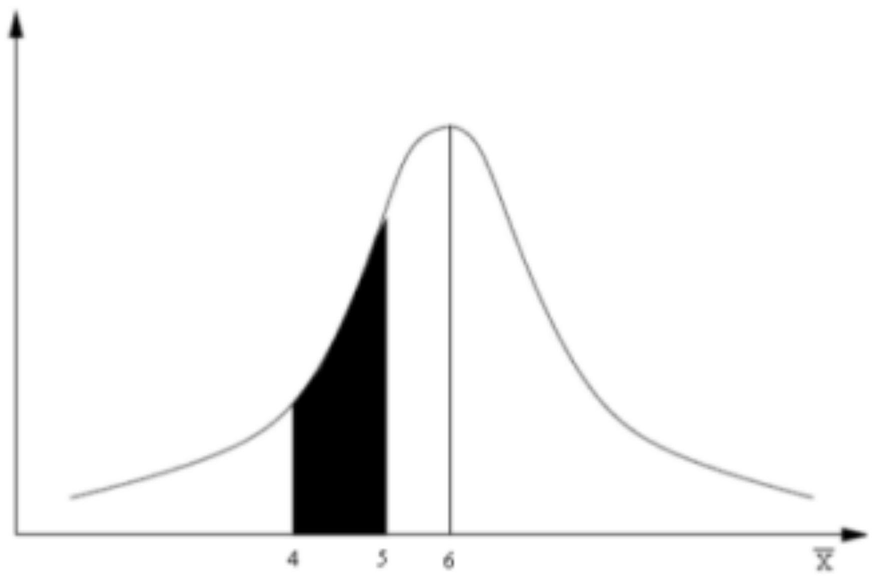
Sampling distribution of means

Example: The time of user connection to my blog follows an unknown distribution with a mean of 6 minutes and a standard deviation of 4 minutes. In a random sample of **36** user connections, which is the probability that they take between 4 and 5 minutes, in average?

$$\mu_{\bar{x}} = \mu = 6$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4}{6}$$

$$n > 30$$



$$\Pr(4 \leq \bar{x} \leq 5) = 6.55\%$$

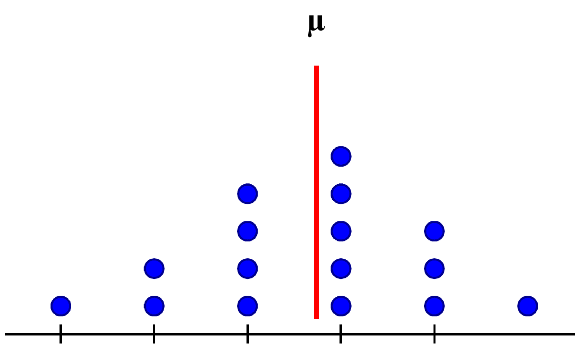
Calculating Confidence Intervals

Confidence intervals (basics)

- When we perform multiple measurements of the same thing, we can calculate confidence intervals
- Assume measurements are samples from a (normal) distribution (real value + random error)
- Characterize the distribution dispersion
- Find the range that includes the desired mass of the probability density (e.g. 90%)

Confidence intervals

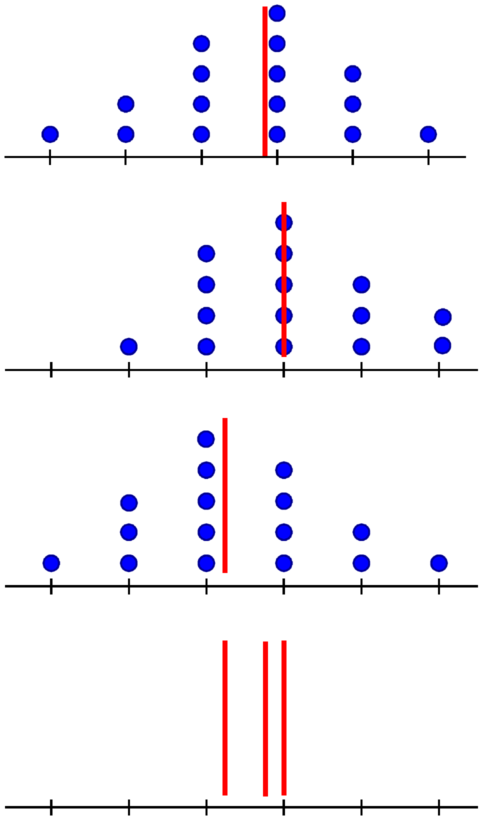
- Assume that the mean of a sample of measurements follows a normal distribution
- This set has a mean $\bar{x}$, which is an estimate of the real mean μ
- If we repeat this with different samples, we will get a slightly different average



(adapted from Dror Feitelson, HU, slides)

Confidence intervals (cont.)

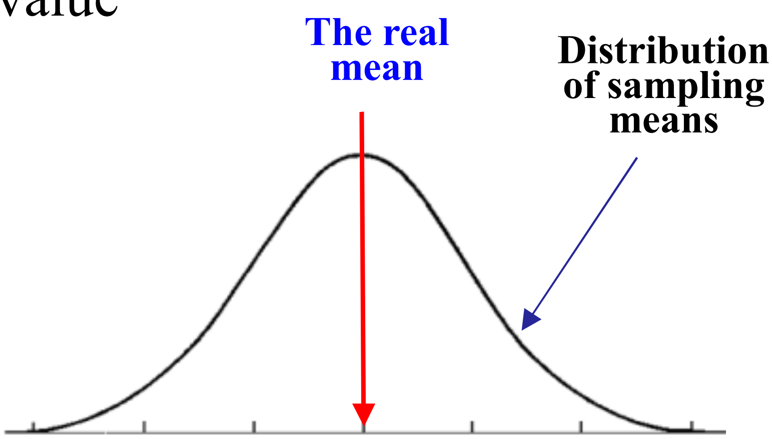
- Multiple sets of samples induce multiple samples from the (sampling) distribution of means
- The sampling distribution of means is narrower than the base distribution
- So it gives a tighter estimate of the real mean μ



(adapted from Dror Feitelson, HU, slides)

Confidence intervals (cont.)

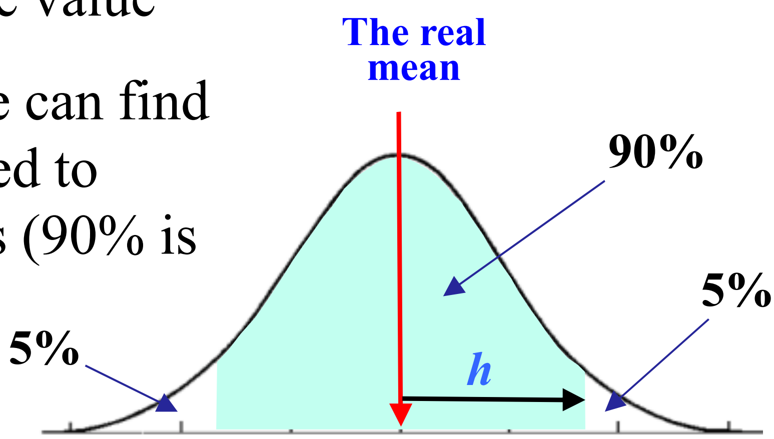
- **Assumption**: the sampling means reflect the true μ plus some random error/noise
- Thus, the sampling means are distributed around the true value



(adapted from Dror Feitelson, HU, slides)

Confidence intervals (cont.)

- **Assumption**: the sampling means reflect the true μ plus some random error/noise
- Thus, the sampling means are distributed around the true value
- Given the distribution, we can find the range ***h*** that is expected to contain 90% of the means (90% is just an example)



(adapted from Dror Feitelson, HU, slides)

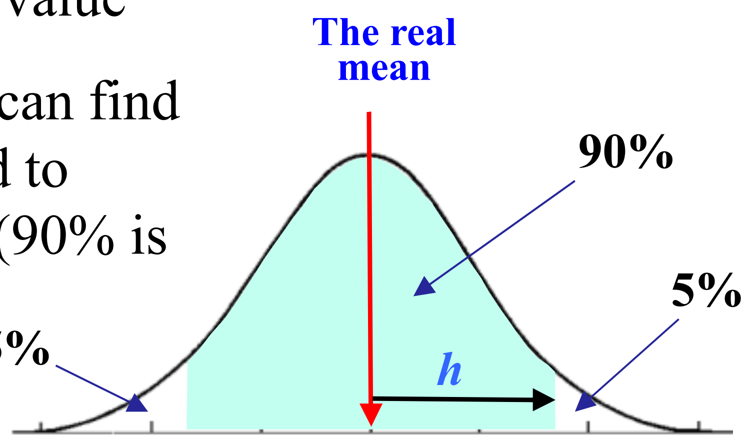
Confidence intervals (cont.)

- **Assumption**: the sampling means
some random
eans are
true value
, we can find
ected to
eans (90% is
just an example)

For 90% of the sampling means, the true mean is within *h*

or

the range sampling mean ± *h* has probability 0.9 to include the real mean



(adapted from Dror Feitelson, HU, slides)

Calculate confidence intervals

- Let μ denote the real mean of the base distribution
- Let $\bar{x}$ denote the mean of n samples
- For large n, then the sampling means have a normal distribution
- Let α denote the acceptable uncertainty (imply that the level of confidence is 1 – α) and define the half-width as

$h = z_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}}$ Then, $\Pr(|\bar{x} - \mu| < h) = 1 - \alpha$

Calculate confidence intervals

- Let μ denote the real mean of the base distribution
- Le
- Fo
dis
- Le evel
of

- $z_{1-\alpha/2}$ comes from tables
- $\frac{\sigma}{\sqrt{n}}$ is the standard deviation of the sampling means.

  Assuming the base samples are independent, this can be ~~calculated~~ estimated as $\frac{s}{\sqrt{n}}$, where s is the standard deviation of the samples

$h = z_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}}$ Then, $\Pr(|\bar{x} - \mu| < h) = 1 - \alpha$

Calculate confidence intervals

- Let μ d

- Let $\bar{x}$ d

- For lar
distrib

With a certainty of 1-α, the distance between a sample of the average $\bar{x}$ and the true mean μ is less than h

If we repeat this many times, and each time we draw a segment of $\pm h$ around $\bar{x}$, then in 1-α of the cases this segment will include μ

- Let α denote the acceptable uncertain that the level of confidence is 1 – α) and define the width as

$$h = z_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$ Then, $\Pr(|\bar{x} - \mu| < h) = 1 - \alpha$

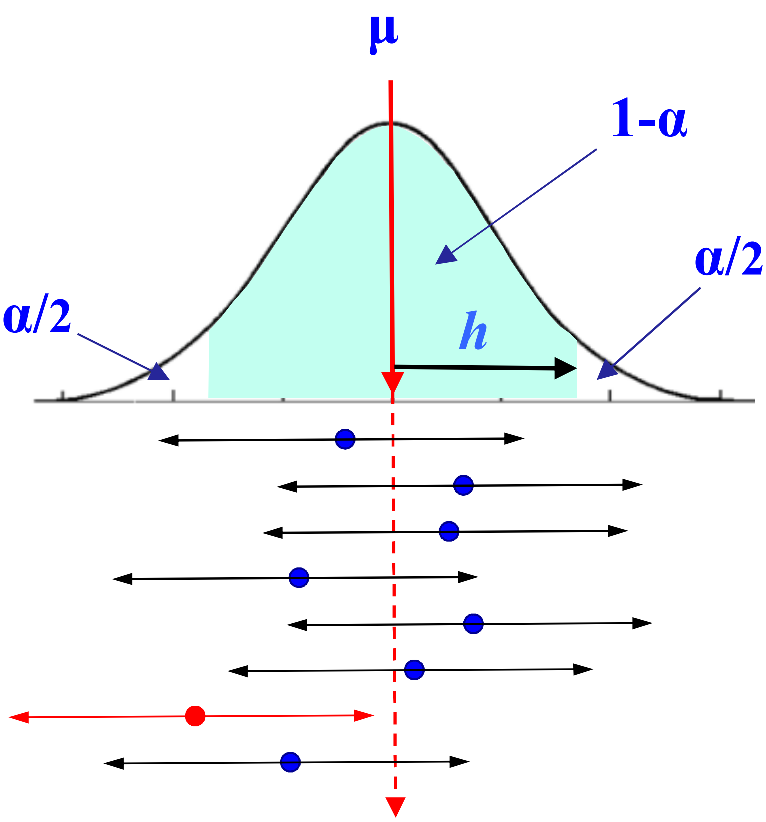
(adapted from Dror Feitelson, HU, slides)

Calculate confidence intervals (cont.)

With a certainty of *1-α* the distance between a <u>sample</u> of the mean $\bar{x}$ and the true mean μ is less than *h*

or

If we repeat a measurement many times, and each time we draw a segment of $\pm h$ around $\bar{x}$, then in *1-α* of the cases this segment will include μ



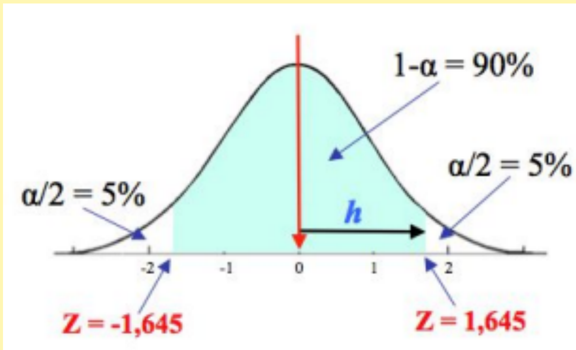
(adapted from Dror Feitelson, HU, slides)

Calculate confidence intervals (cont.)

- L
- L
- If d
- L o

In practice, assuming the base samples are independent, the formula is:

$$\bar{x} \pm z_{1-\alpha/2} \times \frac{s}{\sqrt{n}}$$



Where:

- s is the standard deviation of the n samples
- For α = 0.1 the value z = 1,645. It represents the point in the axis where the area under the standard normal curve is 1 – α (i.e., 90% for α = 0.1)

(adapted from Dror Feitelson, HU, slides)

Calculate confidence intervals

- Let μ denote the real mean of the base distribution
- Let $\bar{x}$ denote the mean of n samples
- For small n, then the standardised sample means follow a Student's t distribution
(note that the sample standard deviation is itself a random variable)
- Let α denote the acceptable uncertainty (imply that the level of confidence is 1 – α) and define the half-width as

$$h = t_{n-1,1-\alpha/2} \times \frac{s}{\sqrt{n}}$$ Then, $\Pr(|\bar{x} - \mu| < h) = 1 - \alpha$

Calculate confidence intervals

- $t_{n-1,1-\alpha/2}$ comes from tables
  - n is the sample size
  - n – 1 degrees of freedom
- $\frac{\sigma}{\sqrt{n}}$ is the true standard deviation of the sampling means. Assuming the base samples are independent, this can be ~~calculated~~ estimated as $\frac{s}{\sqrt{n}}$, where *s* is the standard deviation of the samples

- ... evel of confid...) and define the half-width as

$$h = t_{n-1,1-\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

Then, $\Pr(|\bar{x} - \mu| < h) = 1 - \alpha$

Calculate confidence intervals (cont.)

Assumptions:

- The base samples come from a normal distribution

  If not, but have a finite variance, the sampling means will still be normal, but this will require a larger *n*

- Base samples are independent

  If not, maybe using larger batches will reduce the correlation between them

Calculate confidence intervals (cont.)

Assumptions:

- Th

  If t
  thi

- Ba

  If t
  the

In practice, before computing confidence intervals:

- Clean up the data first
- Remove outliers that indicate interference or spurious measurements. For example:
  - remove top and bottom measurements;
  - look at the data and decide outliers to be removed
- Remove warm-up and history effects

How to find the value Z?

Example: what is the confidence coefficient Z for α = 5%? (two-tailed test)

1. Subtract α from 1
   1 – 0.05 = 0.95

2. Divide result by 2 (because it is two-tailed)
   0.95/2 = 0.475

3. Look at the z-table and locate the results from Step 2 (0.475) in the table.
   The closest value for the coefficient Z is at the intersection of row 1.9 and the column of 0.06. Adding up these two values comes that Z = 1,96 for α = 5%

How to find the value Z?

Example: what is the
(two-tailed test)

1. Subtract α from 1
   1 – 0.05 = 0.95

2. Divide result by 2 (b
   0.95/2 = 0.475

3. Look at the z-table a
   the table.
   The closest value for t
   and the column of 0.0
   1,96 for α = 5%

The entries in this table give the areas under the standard normal curve from 0 to z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	[illegible]	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	[illegible]	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	[illegible]	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	[illegible]	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	[illegible]	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	[illegible]	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	[illegible]	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	[illegible]	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	[illegible]	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	[illegible]	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	[illegible]	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	[illegible]	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	[illegible]	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	[illegible]	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	[illegible]	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	[illegible]	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	[illegible]	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	[illegible]	.4616	.4625	.4633
[illegible]	.4641	.4649	.4656	.4664	.4671	.4678	[illegible]	.4693	.4699	.4706
1.9	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	.4750	.4756	.4761	.4767
[illegible]	.4772	.4778	.4783	.4788	.4793	.4798	[illegible]	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

Common confidence levels and values of Z

Confidence Level	Z
0.70	1.04
0.75	1.15
0.80	1.28
0.85	1.44
0.90	1.645
0.91	1.70
0.92	1.75
0.93	1.81
0.94	1.88
0.95	1.96
0.96	2.05
0.97	2.17
0.98	2.33
0.99	2.575

Example of confidence interval computation

Assume you are measuring the execution time of a given program. You repeat the program execution with different loads and in different moments, in the same computer.

$$x \pm z_{1-\alpha/2} \times \frac{s}{\sqrt{n}}$$

Exec. Time (msec)	
2711	2634
2673	3275
3533	2580
2867	3353
3392	2950
2864	3452
3274	3449
3322	2542
2884	2419
3569	3538
3484	3290
3198	3290
2879	3290
3281	3290
3347	3290
2960	3290

	90%	99%
n of samples	32	32
Z	1.65	2.575
S (std dev)	330.51	330.51
average	3130.31	3130.31
Confidence interval	96.11	150.45
Exec. time minimum	3034.20	2979.86
Exec. time maximum	3226.42	3280.76

Execution time (90%) = 3130.31 ± 96.11

Execution time (99%) = 3130.31 ± 150.45

Notes

- A larger confidence level means a wider and less precise interval
- A smaller confidence level means a more precise interval but will increase the probability of missing μ
- A more precise interval can be obtained by increasing sample size.

$$\bar{x} \pm z_{1-\alpha/2} \times \frac{s}{\sqrt{n}}$$

Confidence interval for the difference between means

- Confidence interval for the difference between means is used to estimate the difference in two population means.
- Independent samples: Two samples from the two populations are independent if the selection of the first sample does not change the selection of the second sample.
- Paired samples: Two samples from the two populations are paired if for each observation in a sample there exists another corresponding observation in the other sample

Confidence interval for the difference between means

- Independent (unpaired) samples (both groups are large)

$$(\bar{x}_1 - \bar{x}_2) \pm z_{1-\alpha/2}\, s_x$$

where s_x is the (estimated) standard deviation of the difference of the means

$$s_x = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Confidence interval for the difference between means

- Independent (unpaired) samples (at least one group is small)

$$(\bar{x}_1 - \bar{x}_2) \pm t_{n_{df}, 1-\alpha/2} s_x$$

where s_x is the (estimated) standard deviation of the difference of the means, as before, and

$$n_{df} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

- n_{df} should be rounded to the nearest integer

Luís Paquete, DEI-FCTUC, 2015

Confidence interval for the difference between means

- Paired samples (large samples)

$$\bar{x}_D \pm z_{1-\alpha/2} \times \frac{s_D}{\sqrt{n}}$$

where $\bar{x}_D$ and S_D are the sample mean and standard deviation of the paired differences

- Paired samples (small samples)

$$\bar{x}_D \pm t_{n-1,1-\alpha/2} \times \frac{s_D}{\sqrt{n}}$$

Population and sample proportion

- Population proportion (p) is the number of elements with a common feature in the size of the population
- Sample proportion ($\bar{p}$) is the number of elements with a common feature in the size of the sample

Sample distribution of proportions

- Mean of the sampling distribution of proportions is equal to the population proportion p:

$$\mu_{\bar{p}} = p$$

- Standard deviation of the sampling distribution of proportions is given by

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Note: There is a correction for large n.

Sample distribution of proportions

- If $np > 5$ and $n(1\text{-}p) > 5$, the sampling distribution of proportions approximates a normal with the following parameters:

$$\mu_{\bar{p}} = p \qquad \sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Known as the Central Limit Theorem for Proportions.

Confidence interval for the proportion

- Let p denote the population proportion and $\bar{p}$ denote the sampling proportion. If $\min(n\bar{p},\ n(1-\bar{p})) > 5$

$$\bar{p} \pm z_{1-\alpha/2} \times \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

Confidence interval for the difference between proportions

- Let p denote the population proportion and $\bar{p}_1$ *and* $\bar{p}_2$ denote the sampling proportions of the two groups.
- If $\min(n_1\bar{p}_1,\ n_1(1-\bar{p}_1),\ n_2\bar{p}_2,\ n_2(1-\bar{p}_2)) > 5$

$$(\bar{p}_1 - \bar{p}_2) \pm z_{1-\alpha/2} \times \sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}$$