Module 3: Probability and Statistics

Week 11 Tutorial

Confidence interval in the mean

Key goals for the class

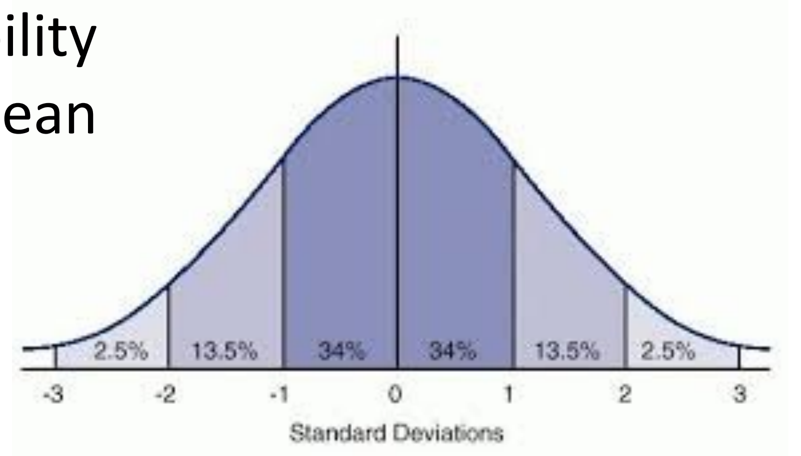
1. How do we predict the probabilities of outcomes for a **normal distribution**, and why is this distribution so important?

2. How do we determine the **confidence interval of an estimate of the population mean**?

Normal or Gaussian distribution

- The **Gaussian** (or "**normal**") probability distribution for a variable x, with mean μ and standard deviation σ is:

$$P_{\text{Gaussian}}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

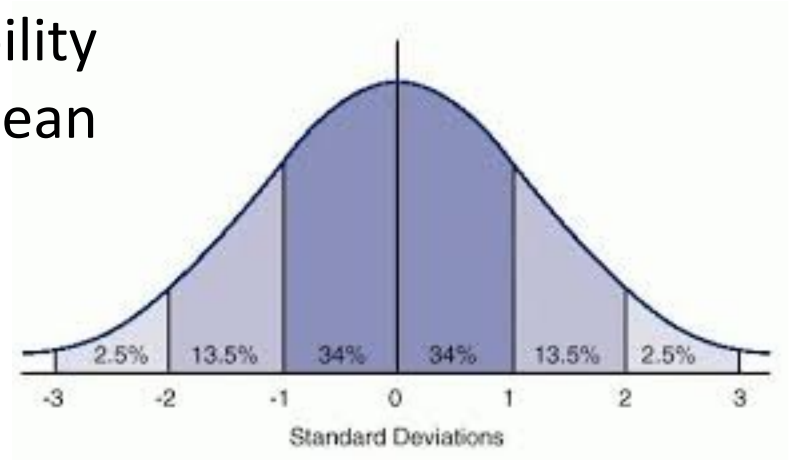


Why is this such an important probability distribution?

Normal or Gaussian distribution

- The **Gaussian** (or "**normal**") probability distribution for a variable x, with mean μ and standard deviation σ is:

$$P_{\text{Gaussian}}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

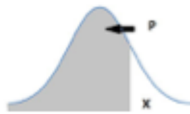


- *Why is this such an ubiquitous and important probability distribution?*
- It is the **high-N limit** for the Binomial and Poisson distributions
- The **central limit theorem** says that if we average together variables drawn many times from any probability distribution, the resulting average will follow a Gaussian!

Reading the normal distribution table

Normal Distribution

Distribution function



The table gives probability $P = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt$.
For $x < 0$ values of $\Phi(x)$ can be obtained from $\Phi(-x) = 1 - \Phi(x)$.

x	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

It's just a single string of values wrapped into a table!

This row from $x = 0$ to $x = 0.09$

This row from $x = 0.1$ to $x = 0.19$

The table gives the one-sided probability P integrating a unit normal distribution from $-\infty$ to x

To get the two-sided probability C (with tails on both sides), we need to find $P = (1 + C)/2$, e.g. $C = 0.9$ maps to $P = 0.95$

Example: for 95% confidence we are looking for $P = 0.975$, hence $x = 1.96$

Tutorial question

Try Q1 on the tutorial sheet (**normal distribution**).

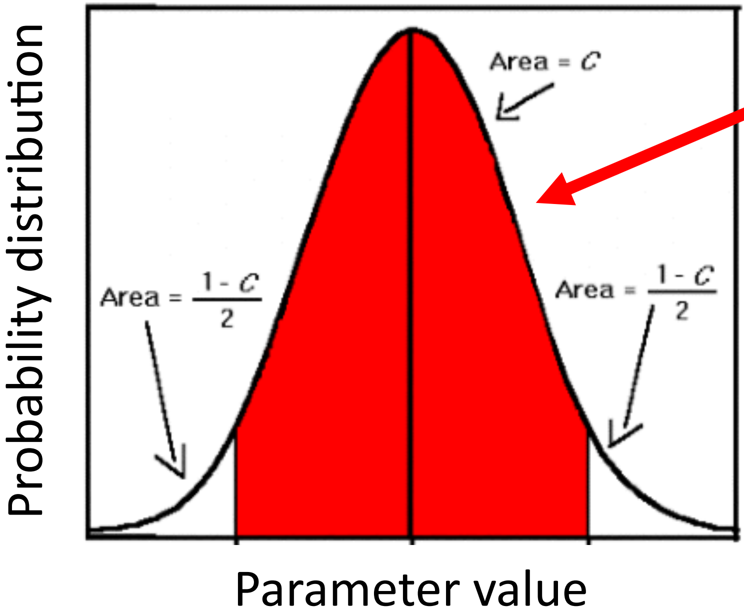
- 1. The random variable X is normally distributed according to $X \sim N(110, 3^2)$.
  - (a) Find b such that $P(X \leq b) = 0.8$
  - (b) Find a such that $P(X < a) = 0.2$
  - (c) Determine $P(X \in [109, 111])$.
  - (d) Determine all quartiles of the X distribution.

Note: the normal distribution is written $N(\mu, \sigma^2)$ so $\mu = 110$ and $\sigma = 3$ for this case

Confidence regions for inference

In **statistical inference**, we estimate the properties of the underlying population from a sample.

We present our results as a **confidence region**, which gives a probability the value lies in a range



Range of values containing probability C (for example, $C = 0.95$ for 95%)

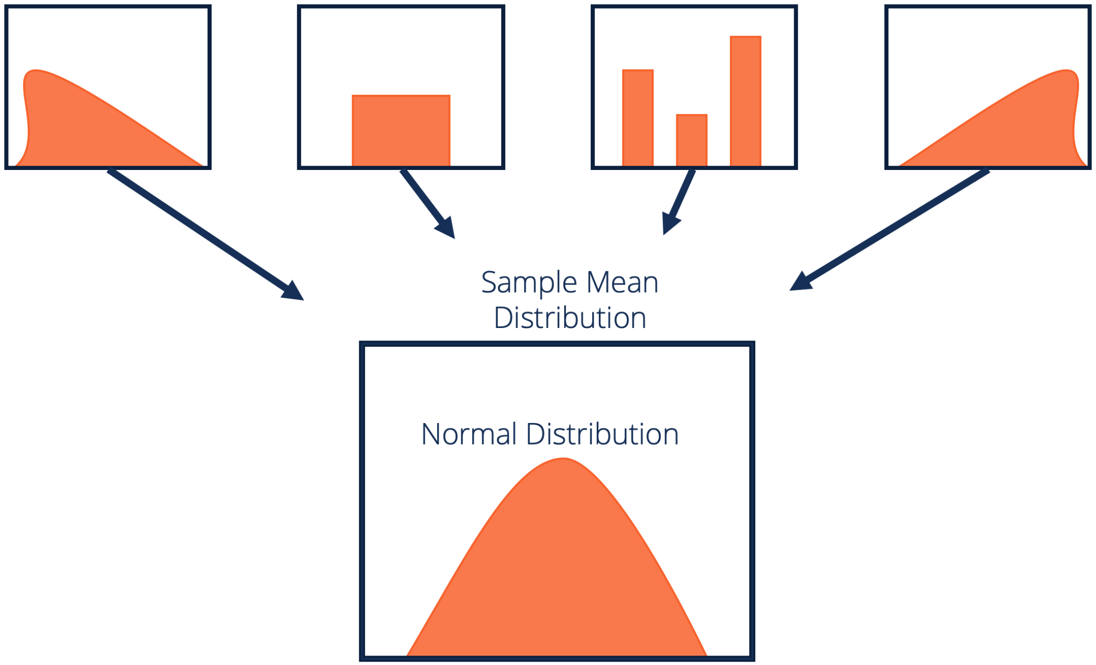
We'll focus on the question: what confidence region can we place on the **population mean**, based on a sample?

Confidence regions for inference

Central limit theorem: if we draw many samples of size N from a population of mean μ and std.dev. σ, the sample mean follows a normal distribution with mean μ and std.dev. $\sigma/\sqrt{N}$

(Exact if the population is normal-distributed, true for large N if not.)

For large N, the sample means are normally-distributed for any type of population distribution – amazing!



Confidence regions for inference

Recipe for the confidence interval in the mean:

- Measure the sample mean, $\bar{x}$
- If the standard deviation σ is **known**, the confidence range is:

Confidence C	Confidence range
90%	$\bar{x} \pm 1.65\ \sigma/\sqrt{N}$
95%	$\bar{x} \pm 1.96\ \sigma/\sqrt{N}$
99%	$\bar{x} \pm 2.58\ \sigma/\sqrt{N}$

These values come from the normal distribution function table by finding the x value which maps to $P = (1 + C)/2$

- If the s.d. s is **estimated from the data**, the confidence range is:

$$\bar{x} \pm t_{N-1,C} \frac{s}{\sqrt{N}}$$

This is the t-distribution critical value with degrees of freedom $N - 1$ and confidence level C

Reading the t-distribution table

TABLE D

t distribution critical values

df	Upper-tail probability p											
	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	[illegible]	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	[illegible]	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
z^*	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											

Degrees of freedom ($= N - 1$ for estimate of mean)

Example: for 98% confidence for a sample of $N = 15$

2-tailed confidence level

Tutorial question

Try Q2 on the tutorial sheet (**confidence interval for the mean when the standard deviation is known**).

- 2. A soft-drink machine is regulated so that the amount of drink dispensed is approximately normally distributed with standard deviation equal to 0.15 deciliters.

  (a) Find a 90% confidence interval for the mean of all drinks dispensed by this machine if a random sample of 36 drinks has an average content of 2.25 deciliters.

  (b) How many drinks should be randomly sampled if we want the 90% confidence interval for the mean to be within 0.09 deciliters?

Tutorial question

Try Q3 on the tutorial sheet (**confidence interval for the mean when the standard deviation is unknown**).

- 3. Watching paint dry.
  The following measurements were recorded for the drying time, in hours, of a certain brand of paint:

  3.4; 2.5; 4.8; 2.9; 3.6; 2.8; 3.3; 5.6; 3.7; 2.8; 4.4; 4.0; 5.2; 3.0; 4.8

  Given that paint drying times are normally distributed with some unknown standard deviation, find a 98% confidence interval for the average time taken for the paint to dry. Hint: use your calculator to obtain sample statistics for the given data.

Tutorial question

Try Q4 on the tutorial sheet (**if time**).

- 4. (*) In case when X does NOT follow normal distribution, σ is unknown and sample size $n > 30$, we use t-score $t = \frac{\bar{X}-\mu}{s/\sqrt{n}}$ (i.e. t-Table) to determine the confidence interval for the mean $\mu = E(X)$, based on the sample statistics $\bar{X} = 1/n \sum_{i=1}^{n} X_i$ and $s^2 = 1/(n-1) \sum_{i=1}^{n} (X_i - \bar{X})^2$. However, z-score $z = \frac{\bar{X}-\mu}{s/\sqrt{n}}$ (i.e. z-Table) can also be used as an approximation. Confirm this statement by inspecting the values of z^* from the t-Table. Note that z^* gives the value of the quantile of the normal distribution for the given confidence level. Compare z^* for each value of the confidence level with the corresponding t-quantiles for $n > 30$.

That’s all for today!