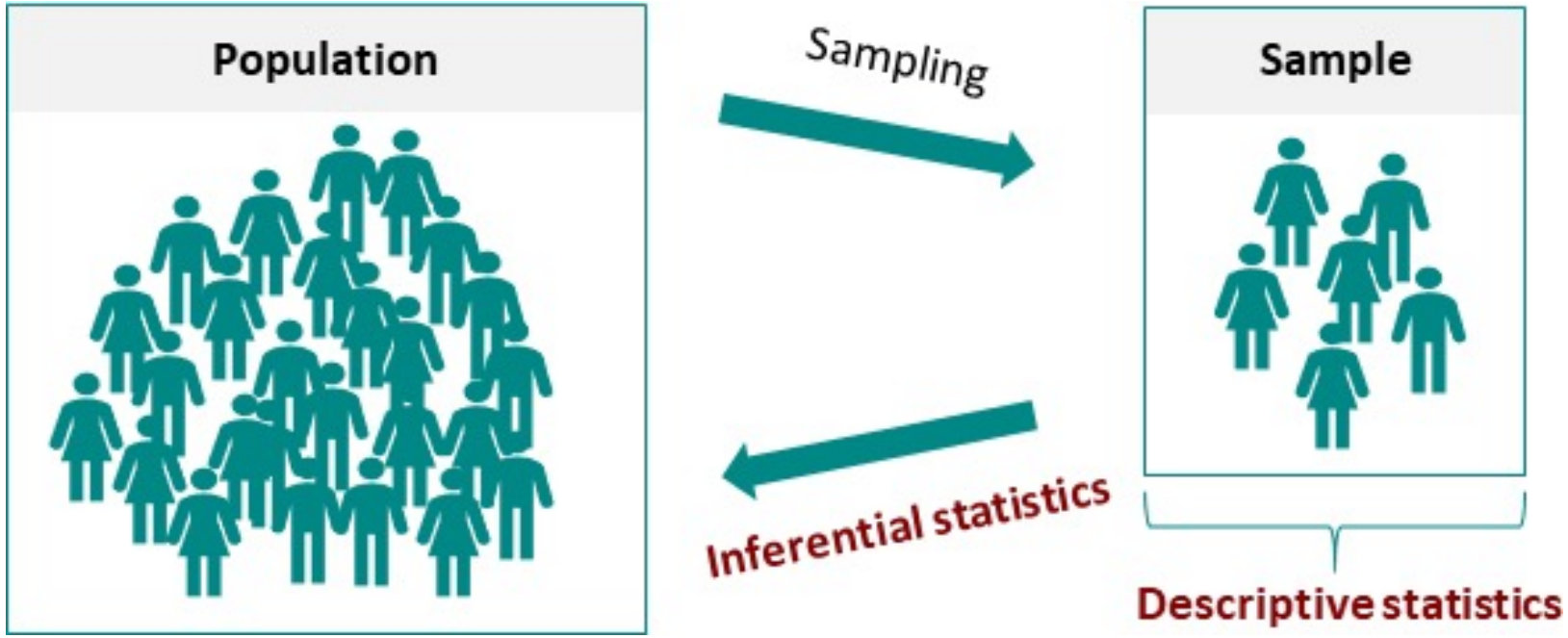
Point Estimation and Confidence Intervals

ANA 500 – Foundations of Data Analytics

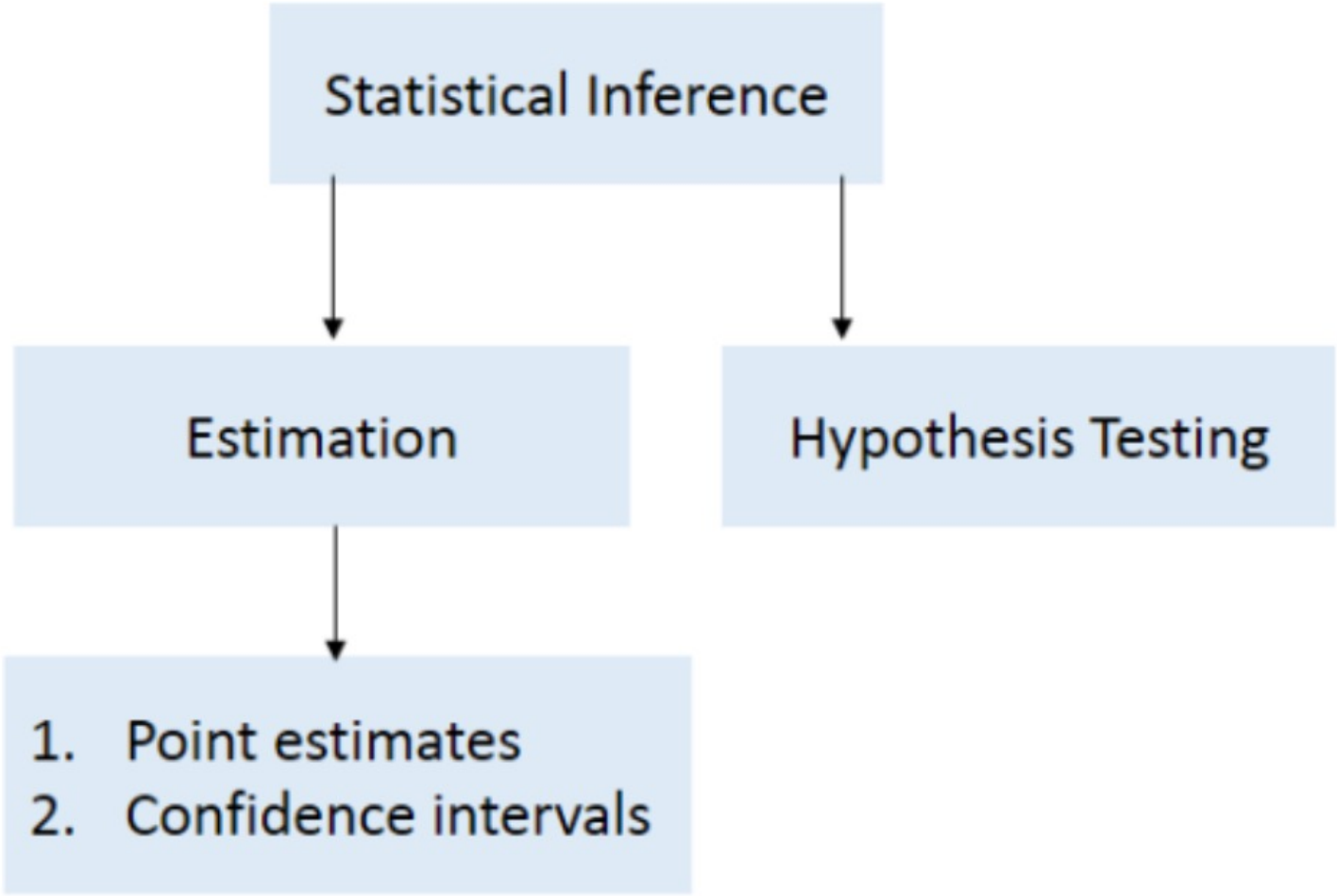
Module 2 - week 3A

Inferential statistics

- Inferential statistics is **a statistical method that deduces from a small but representative sample the characteristics of a bigger population**. In other words, it allows the researcher to make assumptions about a wider group, using a smaller portion of that group as a guideline.



Inferential statistics



Estimation

- We are given information of a sample and using this information , we **estimate** any quantity of population.

- **Point Estimation**: single numerical value used to estimate an unknown population parameter.

- **Estimator**: An estimator is a sample statistics used to estimate a population parameter.

Point Estimation

- What is the mean price of a 4-star hotel room in Washington DC?
  - Gather data (e.g., internet search of prices)
  - Calculate average price
    - This is a point estimate of a population mean
    - $\bar{x}$

- How often are flights delayed?
  - Gather data (e.g., record the flights that are delayed)
  - divide by the total number of flights.
    - This is a point estimate of a population proportion
    - $\hat{p}$

Confidence Intervals

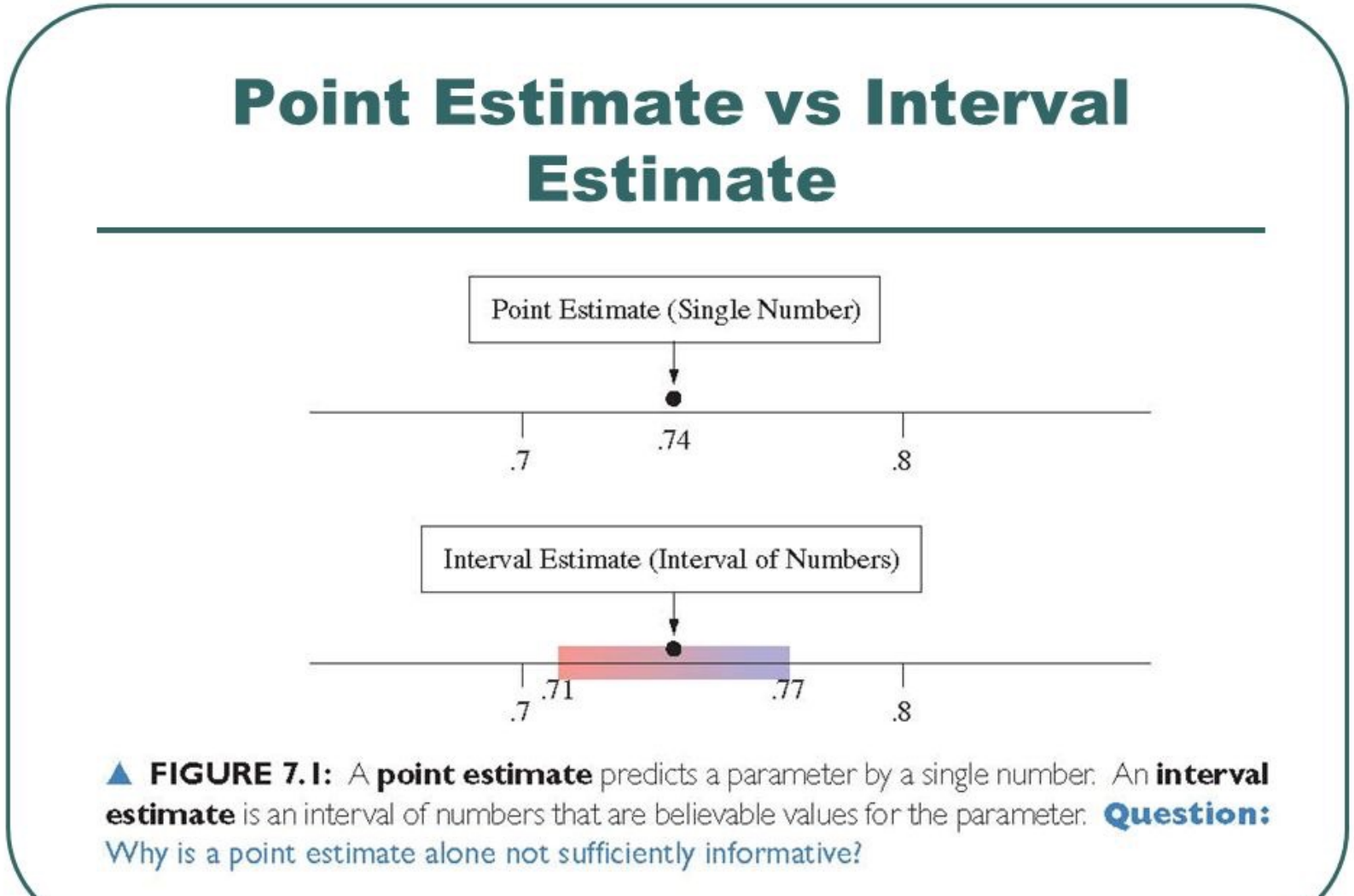
- **Point Estimation**: **single** numerical value used to estimate an unknown population parameter.
  - Sample mean is a point estimate of a population mean.
  - Sample proportion is a point estimate of a population proportion.

- **Interval Estimate**: **Range** of values used to estimate an unknown population parameter.
  - Interval estimates of population parameters are called confidence intervals.

Point Estimation vs Interval Estimate

- The point estimate is unlikely to be exactly equal to the true value of the parameter of interest, we hope it is close though.

- A confidence interval can help us quantify this imprecision.



▲ **FIGURE 7.1:** A **point estimate** predicts a parameter by a single number. An **interval estimate** is an interval of numbers that are believable values for the parameter. **Question:** Why is a point estimate alone not sufficiently informative?

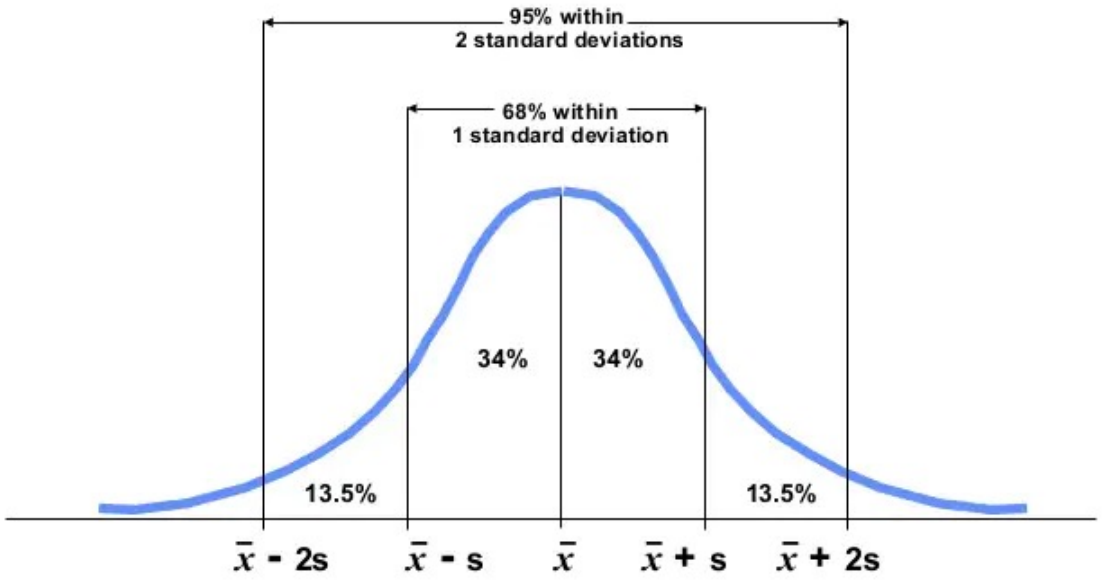
Confidence Intervals

- A confidence interval provides a range of values it is reasonable for the population mean (parameter) to fall in.
  - There is no guarantee that the interval contains the true value of the parameter.
  - We can make probability statements about how likely it is.

Confidence Intervals

- Recall that $\bar{x} \sim N(\mu, \frac{\sigma^2}{n})$
- About 95% of all observations fall within 2 standard deviations of the mean.

FIGURE 2-15 **The Empirical Rule**
(applies to bell-shaped distributions)

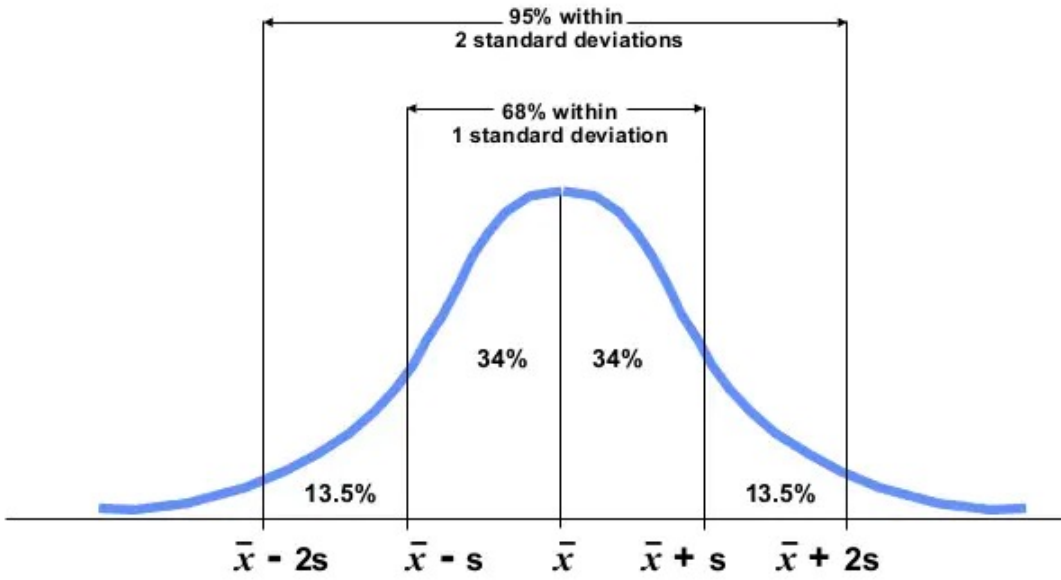


Confidence Intervals

What is the mean price of a 4-star hotel room in Washington DC?

- Suppose $\bar{x}$ = 290 and σ = 200 and n = 100
- Standard deviation for $\bar{x}$: $\frac{\sigma}{\sqrt{n}} = \frac{200}{\sqrt{100}} = 20$
- 2 standard deviations = 40
- 95% confidence interval:

$$290 \pm 40 = (250, 330)$$



- We are 95% confident the population mean falls in this interval

Confidence Intervals

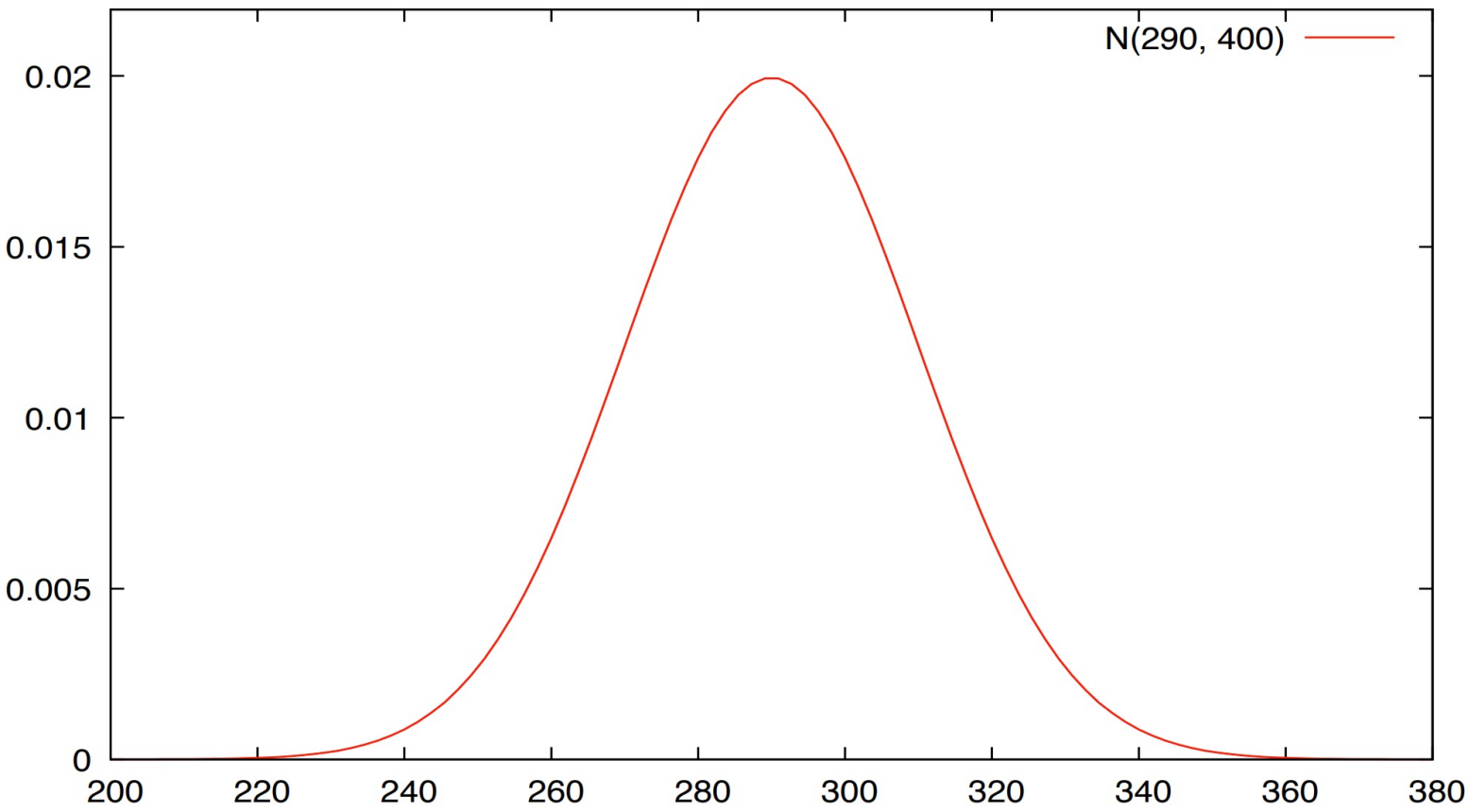
- In general, confidence intervals have the following form:

  Point estimate $\pm$ margin of error

  - Margin of error depends on level of confidence and the standard deviation of the estimate (i.e., standard error)

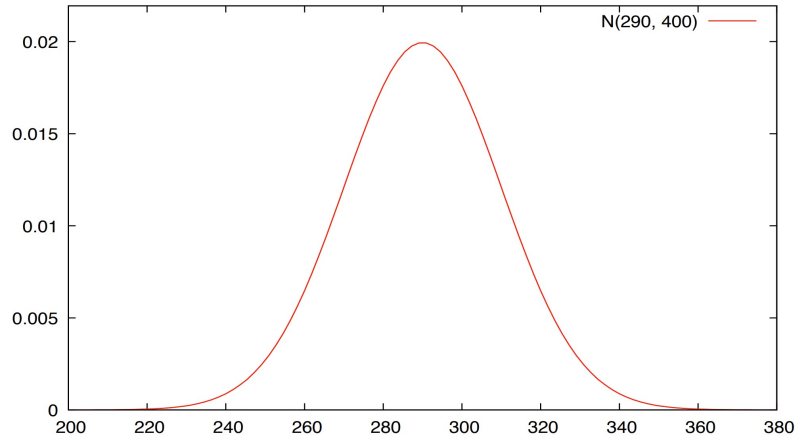
- The confidence level is the percentage of times the interval contains the true parameter value in repeated random samples.
- (1 – level of confidence) = level of significance = α

Confidence Intervals



Confidence Intervals

- If We draw another random sample, we will most likely get a different sample mean different from 290.
  - The confidence interval itself would change every random sample we draw.



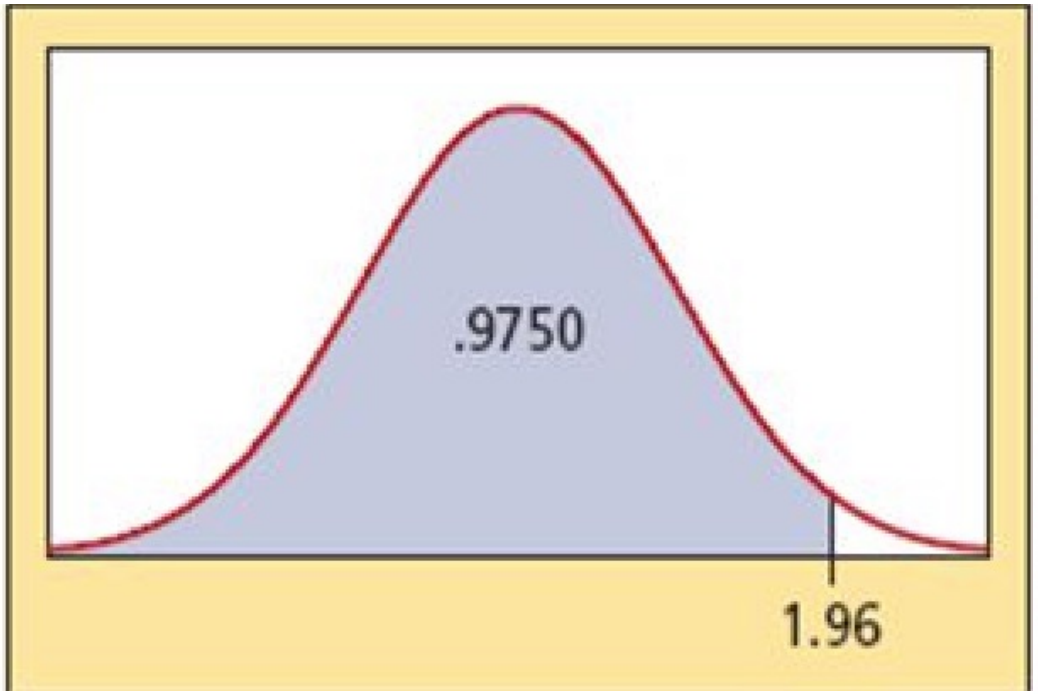
- Remember the true parameter of interest Mu is either in the interval or not. We never know for sure.

Confidence Interval Estimation Sigma Known

- When σ , the population standard deviation, is known, the standard normal distribution is used to calculate the confidence interval.
    - We need to find the value of z that puts the confidence level in the middle of the distribution and the level of significance in the tails.

Confidence Interval Estimation Sigma Known

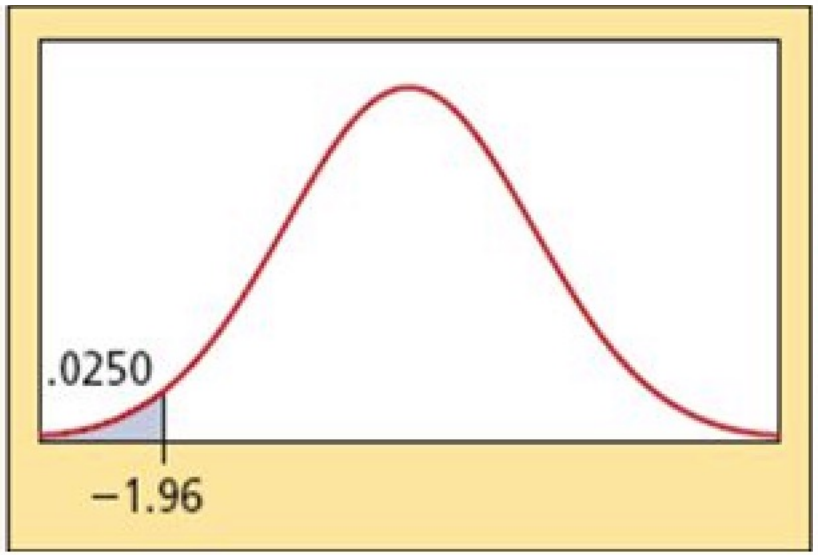
For a 95% confidence interval z = ± 1.96



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884

Confidence Interval Estimation Sigma Known

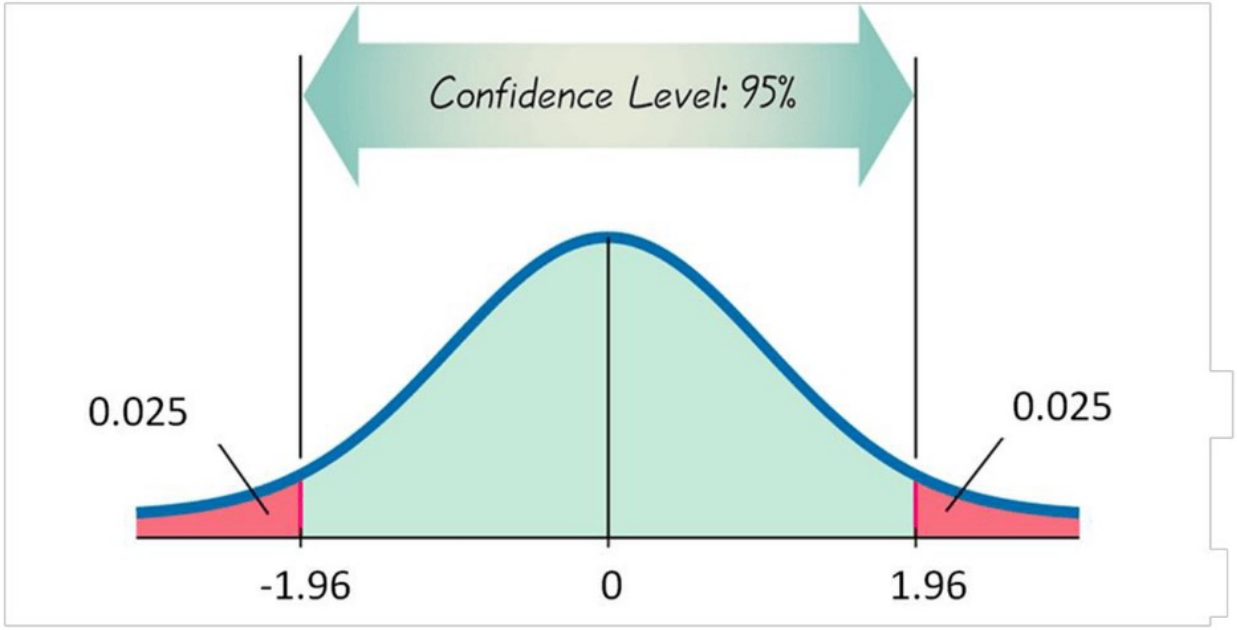
For a 95% confidence interval z = ± 1.96



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0
-1.9	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.0250	0.0244	0.0
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0

Confidence Interval Estimation Sigma Known

For a 95% confidence interval z = $\pm$ 1.96

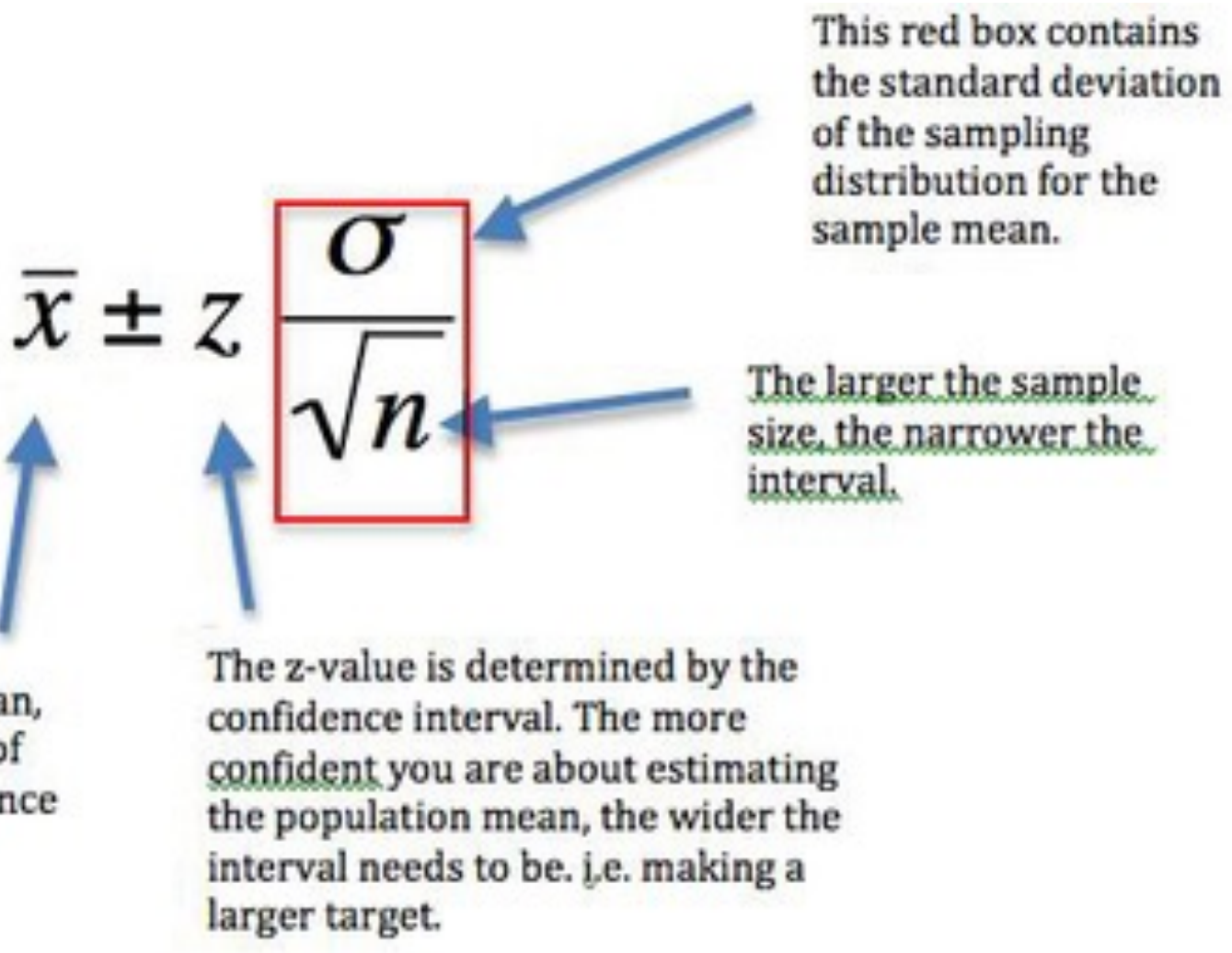


Confidence Interval Estimation Sigma Known

- Example: what is the average amount of time spent commuting to work? Suppose in a random sample of 36 commuters the average commute time is 27 minutes and σ **= 12**.

- Q1: What is the 95% confidence interval?
- Q2: What is the 90% confidence interval?

Confidence Interval Estimation Sigma Known



Confidence Interval Estimation Sigma Known

- Example: what is the average amount of time spent commuting to work? Suppose in a random sample of 36 commuters the average commute time is 27 minutes and σ = 12.

- Q1: What is the 95% confidence interval?
- 95% confidence interval = $27 \pm 1.64\left(\frac{12}{\sqrt{36}}\right) = 27 \pm 1.64\ (2) =$ (23.72, 30.28)

Confidence Interval Estimation Sigma Known

- Example: what is the average amount of time spent commuting to work? Suppose in a random sample of 36 commuters the average commute time is 27 minutes and σ = 12.

- Q1: What is the 90% confidence interval?
- 90% confidence interval = $27 \pm 1.64\left(\frac{12}{\sqrt{36}}\right) = 27 \pm 1.64\ (2) =$ (23.72, 30.28)

Confidence Interval Estimation Sigma Unknown

- When σ is unknown it must be estimated with s and the Student's t-distribution is used to calculate the confidence interval.
  - σ unknown cases: use s as an estimator
  - σ unknown cases: use t distribution instead of Z

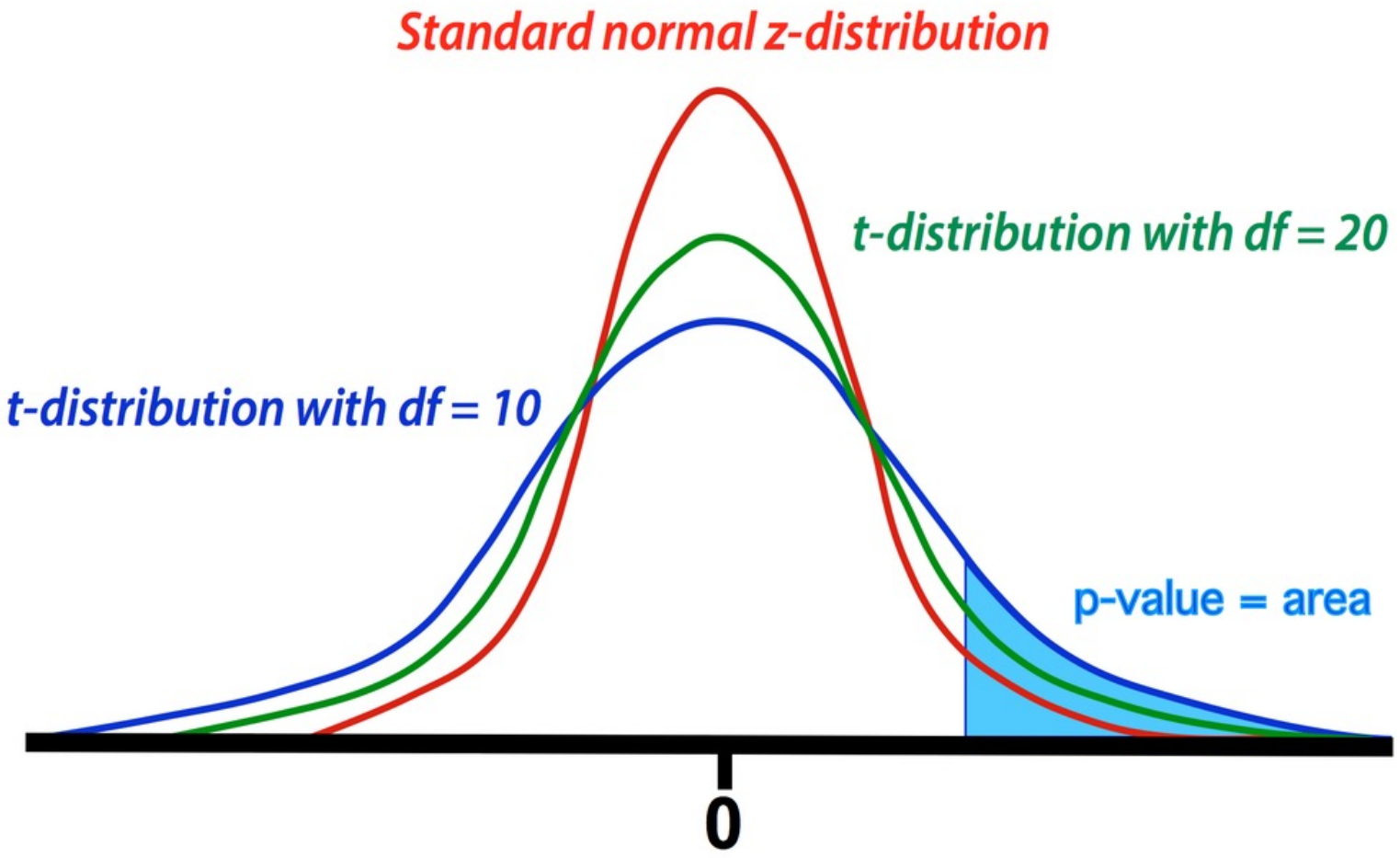
- The approach today is to always use the student's T distribution whenever the standard deviation sigma is estimated with the sample statistic s.

Confidence Interval Estimation Sigma Unknown

- *To read t distribution table we need:*

  - Degree of Freedom: $n - 1$
  - $\alpha/2$ α = level of significance =1-level of confidence
    - For 95% confidence interval: 1-.95=0.05. $\alpha/2$ =0.025

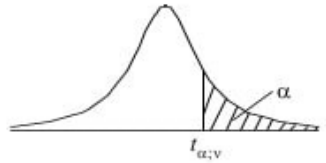
Confidence Interval Estimation Sigma Unknown



Confidence Interval Estimation Sigma Unknown

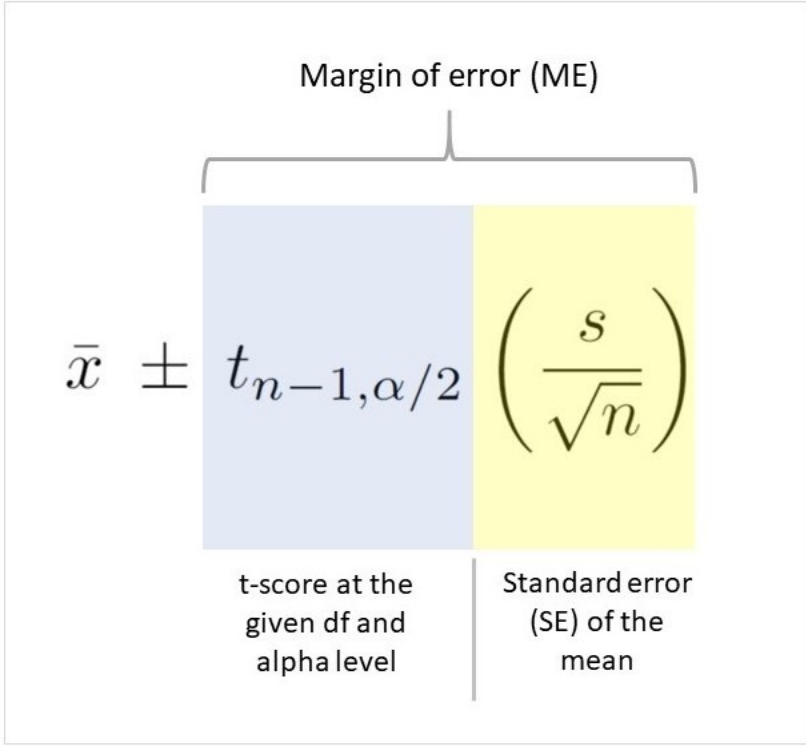
Table of the Student's *t*-distribution

The table gives the values of $t_{\alpha;\nu}$ where
$\Pr(T_\nu > t_{\alpha;\nu}) = \alpha$, with ν degrees of freedom



ν \ α	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
1	3.078	6.314	12.076	31.821	63.657	318.310	636.620
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.767
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Confidence Interval Estimation Sigma Unknown



Carsten Grube. Statistical Data Analysis. dataZ4s.com

Confidence Interval Estimation Sigma Unknown

- **Example**: we want to draw a 90% confidence interval on the true average grade on a population of students taking a statistics exam. Assuming we don't know Sigma, the population standard deviation. We take a sample of n equal 20 (n=20) and we get a sample average of 75 ($\bar{x}$ =75) and a sample standard deviation of 15 (s = 15).

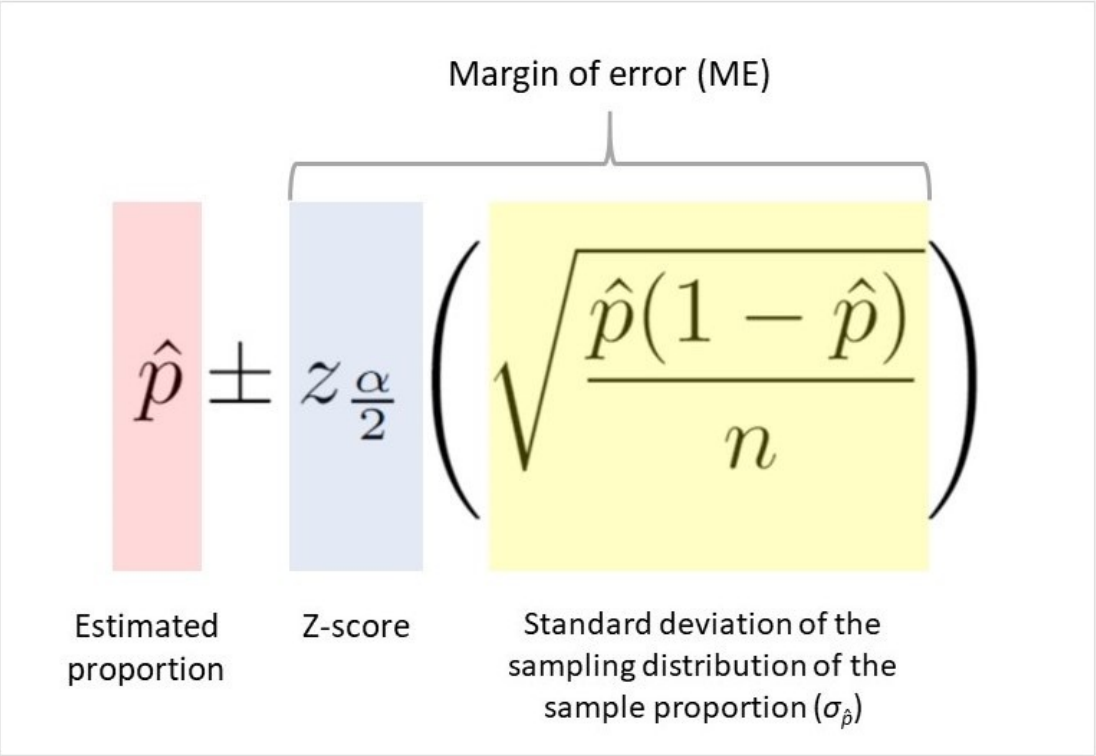
- Level of Significance = α = 1-0.9 = 0.1; $\alpha/2$=0.05
- df = n -1 = 20-1 =19 ; $t_{\alpha/2}$ = 1.729

- $\bar{x} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) = 75 \pm 1.729\left(\frac{15}{\sqrt{20}}\right) = 75 \pm 5.7992 = (69.20, 80.80)$
- 90% confident that the true mean is located somewhere within this interval.
- there is a 10% chance that the true mean is outside.

Confidence Intervals for Proportions

- General form: $\hat{p} \pm Margin\ of\ Error$
- The sampling distribution of $\hat{p}$ can be approximately by the normal distribution when
  - $np \geq 5\ and\ n(1-p) \geq 5$
- The standard error of $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

Confidence Intervals for Proportions



Carsten Grube. Statistical Data Analysis. dataZ4s.com

Confidence Intervals for Proportions

- **Example:** 130 fans out of 200 sampled said they order hot dogs at stadium. What is the 99% confidence interval?
- $\hat{p}$ =130/200=0.65; n=200
- Level of Significance = α = 1-0.99 = 0.01; α/2=0.005
- Z = $\pm$ 2.575

- $\hat{p} \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} = 0.65 \pm 2.575\sqrt{\frac{0.65(1-0.65)}{200}}$

  =0.65 $\pm$2.575(0.0337) = (0.563, 0.737)
- We are 99% confident that the true proportion of fans ordered hotdogs at the stadium is within this interval.