Technische Universität München

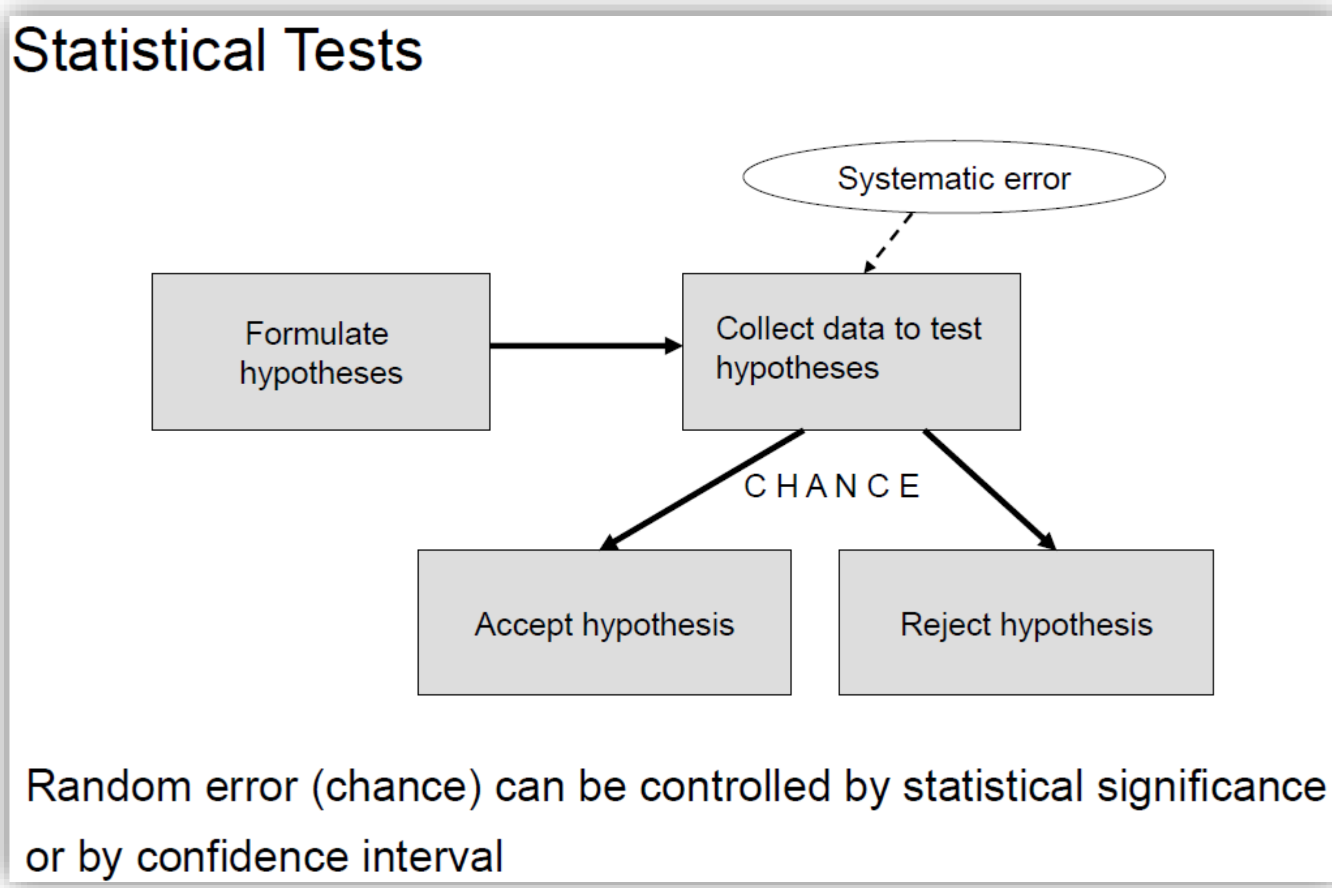
Tutorial Business Analytics

Tutorial 2: Statistics

Decision Sciences & Systems (DSS)
Department of Informatics
TU München

Tutorial 2 Business Analytics: Statistics

What we will focus on in this tutorial:



Tutorial 2 Business Analytics: Statistics

Agenda

1. Theory: How does **Hypothesis testing** work?
2. Calculation **Example**
3. Practice: **Exercises in Live Tutorial Session**

Recommendations

- Use paper and a scientific calculator for the exercises (except R exercises)
- Pay attention to the theory and the example part
- Do all exercises and homework

Tutorial 2 Business Analytics: Statistics

Statistical Testing

- We are trying to validate a claim about a statistic of a population, only based upon (a) sample(s)
- This **statistical hypothesis** is tested by observing random variables
- Common cases are
  - Sample statistic is compared against a synthetic (population) statistic
  - Two samples are compared
- A hypothesis is proposed for the **statistical relationship** between the two statistics; this is compared to a **null hypothesis**
- The comparison is denoted as **statistically significant** if the relationship between the statistics (i.e., drawing respective sample(s)) would be unlikely under the null hypothesis according to a threshold probability

Tutorial 2 Business Analytics: Statistics

"Test Manual" – Overview

1. i) 1 sample or 2 samples

   ii) If 1 sample: σ_x known or unknown
   If 2 samples: dependent or independent

2. State H_0 and H_1 (given)

3. Select and calculate the test statistic

4. Select α (given)

5. Find the critical value in the table

6. Result

Tutorial 2 Business Analytics: Statistics

"Test Manual" – 2nd Step

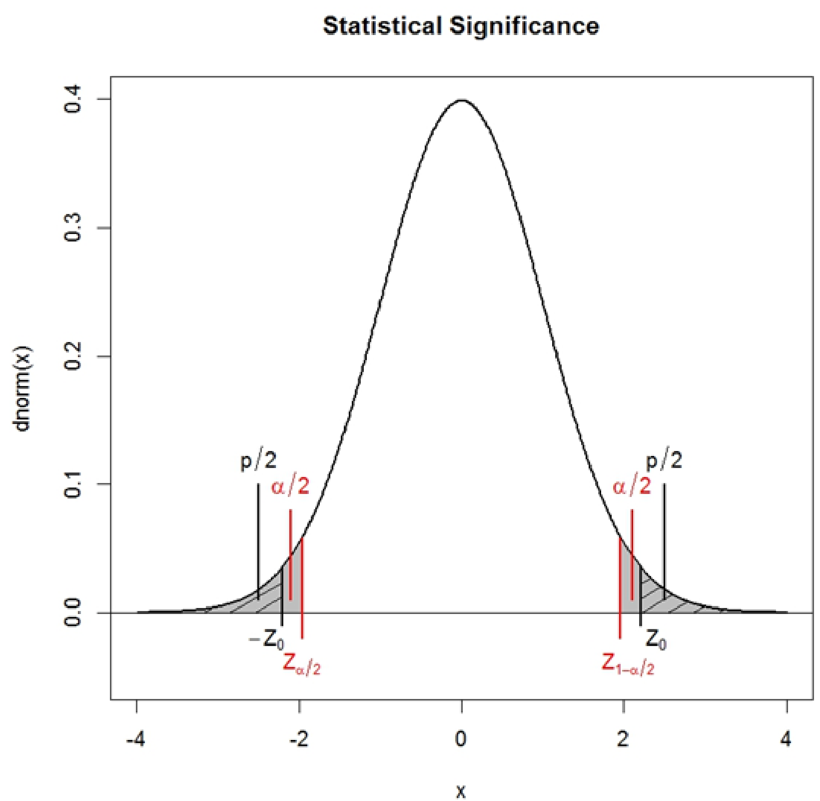
There exist three possible alternative hypotheses H_1:

Hypothesis	H_0	H_1
Two-sided	$\mu_x = \mu_0$	$\mu_x \neq \mu_0$
One-sided	$\mu_x \leq \mu_0$	$\mu_x > \mu_0$
One-sided	$\mu_x \geq \mu_0$	$\mu_x < \mu_0$

Tutorial 2 Business Analytics: Statistics

"Test Manual" – 2nd Step: Two-Sided Hypothesis Test

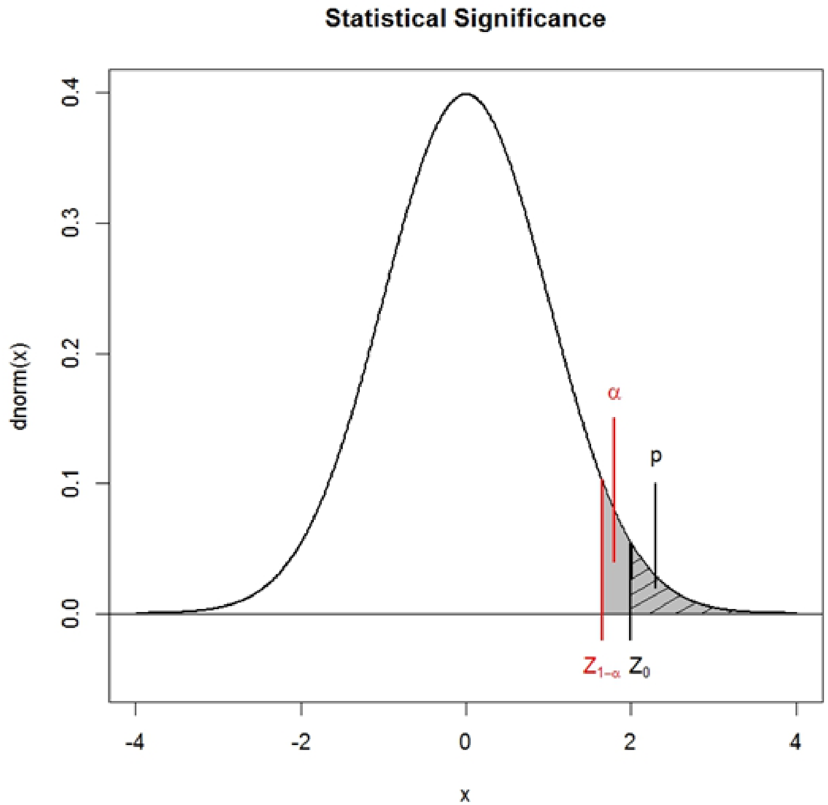
$$H_0\colon \mu_x = \mu_0 \qquad H_1\colon \mu_x \neq \mu_0$$



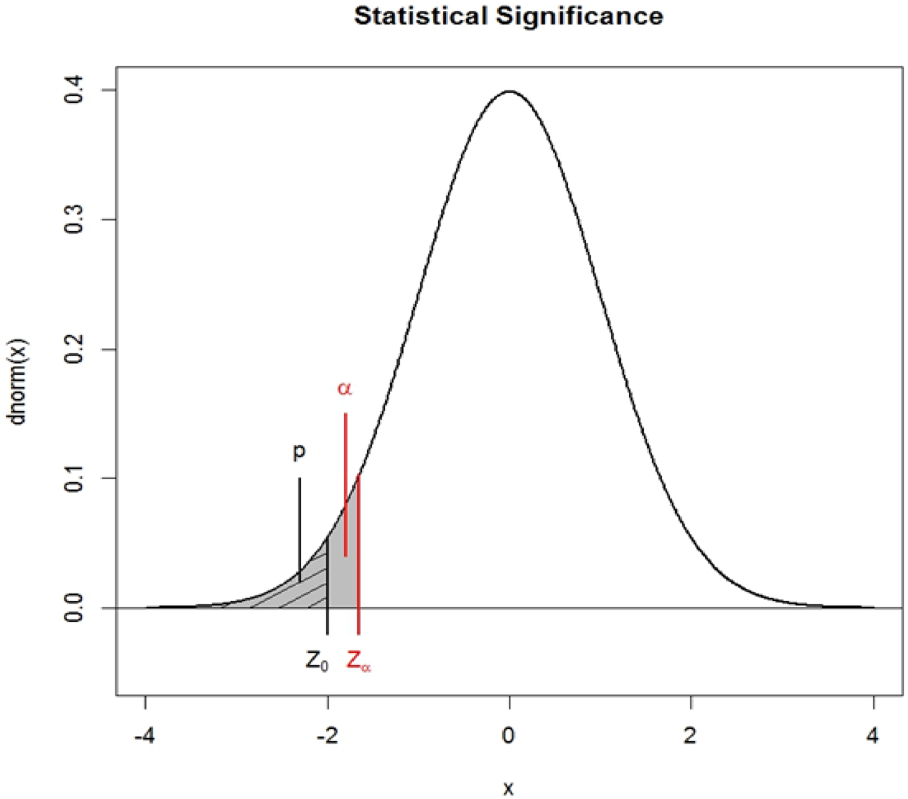
Tutorial 2 Business Analytics: Statistics

"Test Manual" – 2nd Step: One-Sided Hypothesis Test

$H_0: \mu_x \leq \mu_0 \qquad H_1: \mu_x > \mu_0$



$H_0: \mu_x \geq \mu_0 \qquad H_1: \mu_x < \mu_0$



Tutorial 2 Business Analytics: Statistics

“Test Manual” – 3rd Step

When to use which test? We want to make a statement about the mean of a population, μ_x, based on a sample with size n_x and mean $\bar{x}$

1 Sample

- σ_x known → Gauss/z-test $z_0 = \frac{\bar{x}-\mu_0}{\sigma_X}\sqrt{n} \sim N(0,1)$
- σ_x unknown → t-test $t_0 = \frac{\bar{x}-\mu_0}{s_X}\sqrt{n} \sim t_{n-1}$ with $s_x^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n}(x_i - \bar{x})^2$

2 Samples

- independent → Welch-test $t_0 = \frac{\bar{x}-\bar{w}-\mu_0}{s_{\bar{x}-\bar{w}}} \sim_{\text{approx}} t_{\text{df}}$ with $s_{\bar{x}-\bar{w}}^2 = \frac{s_x^2}{n_x} + \frac{s_w^2}{n_w}$ and

  $s_x^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n}(x_i - \bar{x})^2$ $\left(\text{df} = \frac{\left(s_{\bar{x}-\bar{w}}^2\right)^2}{\frac{s_x^4}{n_x^2(n_x-1)} + \frac{s_w^4}{n_w^2(n_w-1)}}\right.$ rounded to nearest integer number)
- dependent → Paired t-test $t_0 = \frac{\bar{d}-\mu_0}{s_d}\sqrt{n} \sim t_{n-1}$ with $s_d^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n}(d_i - \bar{d})^2$ and

  $\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i = \bar{x} - \bar{w}$, $d_i = x_i - w_i$, $\mu_D = \mu_X - \mu_W$

Tutorial 2 Business Analytics: Statistics

“Test Manual” – 5th Step

How to find the critical value in the table? For

- Gauss/z-Test → use normal distribution
- t-Test, Welch-Test and Paired t-Test → use t-distribution

H_1	t^c range	t^c value
$\mu_x \neq \mu_0$	can be any, $\mathbb{R}$	$\left\lvert t^c_{1-\frac{\alpha}{2};\,\mathrm{df}} \right\rvert = \left\lvert t^c_{\frac{\alpha}{2};\,\mathrm{df}} \right\rvert$
$\mu_x > \mu_0$	must be positive, $\mathbb{R}_{>0}$	$t^c_{1-\alpha;\,\mathrm{df}}$
$\mu_x < \mu_0$	must be negative, $\mathbb{R}_{<0}$	$t^c_{\alpha;\,\mathrm{df}}$

Tutorial 2 Business Analytics

Normal Distribution (z-table)

- If X is a normally distribution random variable with mean μ and standard deviation σ,
$$Z = \frac{X - \mu}{\sigma}$$
is **standard normally distributed**
- The table contains the *probabilities* that a statistic is less than z, i.e., between negative infinity and z
- The values are calculated using the cumulative distribution function Φ
- Examples:
  - $\boxed{\Phi(0.72) = 0.76424}$
  - $\boxed{\Phi(-1.48) = 1 - \Phi(1.48) = 0.06944}$
  - If quantile $z_{0.9}$ is needed:
$$\Phi(z_{0.9}) = 0.9 \quad \Rightarrow \quad \boxed{z_{0.9} \approx 1.28}$$

z	+0.00	+0.01	+0.02	+0.03	+0.04	+0.05	+0.06	+0.07	+0.08	+0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55966	0.56360	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91308	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520

Tutorial 2 Business Analytics

t-Distribution (t-table)

- A random variable with t-distribution arises, e.g., when estimating the mean of a normally distributed population in situations with a small sample size and unknown population standard deviation
- The numbers in the body of the table, $t^{c}_{1-\alpha;\,\mathrm{df}}$, are the critical values needed for the t-test
  - df: degrees of freedom
  - α: significance level

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	**0.50**	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
two-tails	**1.00**	**0.50**	**0.40**	**0.30**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.002**	**0.001**
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

Tutorial 2 Business Analytics: Statistics

"Test Manual" – 6th Step

Reject H_0:

H_1	p-value criterion	test statistic criterion
$\mu_x \neq \mu_0$	$p < \alpha$	$\lvert t_0 \rvert > \left\lvert t^c_{1-\frac{\alpha}{2};\,\mathrm{df}} \right\rvert$
$\mu_x > \mu_0$	$p < \alpha$	$t_0 > t^c_{1-\alpha;\,\mathrm{df}}$
$\mu_x < \mu_0$	$p < \alpha$	$t_0 < t^c_{\alpha;\,\mathrm{df}}$

Tutorial 2 Business Analytics: Statistics

Example: Learning Method Comparison

In order to compare two learning methods, results have been measured for a group of students. Test if the students got better (higher) results using method 2. Assume the difference follows a normal distribution, (significance level of 5%, i.e., $\alpha = 0.05$).

student	1	2	3	4	5
method 1 (x)	8	6	8	8	4
method 2 (w)	10	9	7	12	7

1.) i) 2 samples ii) dependent

2.) $H_0\colon \mu_D = \mu_X - \mu_W \geq \mu_0 = 0$ $H_1\colon \mu_D = \mu_X - \mu_W < \mu_0 = 0$

3.) → Paired t-Test: $t_0 = \frac{\bar{d} - \mu_0}{s_d}\sqrt{n} \sim t_{n-1}$ with unbiased sample variance $s_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2$

sample means: $\bar{x} = 6.8$, $\bar{w} = 9.0$, difference $\bar{d} = -2.2$,

$s_d^2 = 3.7, \quad s_d = 1.9235 \quad \Rightarrow \quad t_0 = -2.5574$

4.) $\alpha = 0.05$

5.) $\rightarrow \; t_{\alpha;n-1}^c = -t_{1-\alpha;n-1}^c \; (\text{sym.}) \;\; \Rightarrow \;\; t_{0.05;4}^c = -t_{0.95;4}^c \overset{\text{table}}{=} -2.132$

6.) $t_0 = -2.557 < -2.132 = t_{0.05;4}^c \;\; \Rightarrow$ Reject H_0: Learning method 2 is significantly better.

Tutorial 2 Business Analytics: Statistics

Example: Learning Method Comparison – step 3 details

In order to compare two learning methods, results have been measured for a group of students. Test if the students got better (higher) results using method 2. Assume the difference follows a normal distribution, (significance level of 5%, i.e., $\alpha = 0.05$).

student	**1**	**2**	**3**	**4**	**5**
method 1 (x)	8	6	8	8	4
method 2 (w)	10	9	7	12	7

3.)

sample means: $\bar{x} = \frac{1}{5}(8+6+8+8+4) = 6.8$, $\bar{w} = \frac{1}{5}(10+9+8+12+7) = 9.0$

difference: $\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i = \bar{x} - \bar{w} = -2.2$

sample variance: $s_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2$, $d_i = x_i - w_i$,

$$s_d^2 = \frac{1}{4}\left((8-10+2.2)^2+(6-9+2.2)^2+(8-7+2.2)^2+(8-12+2.2)^2+(4-7+2.2)^2\right) = 3.7$$

$$s_d = 1.9235$$

Tutorial 2 Business Analytics: Statistics

Confidence Intervals

Find confidence intervals for μ_x, which—under H_0—contain the true value μ_x with a probability of at least $1-\alpha$ (confidence level). We differentiate two cases:

- σ_x known:

confidence interval: $$\boxed{[I_u(x),\ I_o(x)] = \left[\bar{x} - z^c_{1-\frac{\alpha}{2}} \frac{\sigma_x}{\sqrt{n}},\ \bar{x} + z^c_{1-\alpha/2} \frac{\sigma_x}{\sqrt{n}}\right]}$$

- σ_x unknown, use s_x as estimate instead:

confidence interval: $$\boxed{[I_u(x),\ I_o(x)] = \left[\bar{x} - t^c_{1-\frac{\alpha}{2};\, n-1} \frac{s_x}{\sqrt{n}},\ \bar{x} + t^c_{1-\frac{\alpha}{2};\, n-1} \frac{s_x}{\sqrt{n}}\right]}$$

- Values of μ_0 within the confidence interval cannot be rejected regarding a significance level of α

  $\rightarrow$ Reject H_0 if μ_0 is not in the confidence interval

Tutorial 2 Business Analytics: Statistics

Exercise 2.1

The consumption per person is measured in index values, where a high index value represents a high consumption. The following table embodies index values for 10 individuals before and after a tax increase.

Individual number, i	Index value		Difference, $d = a - b$
	previous to tax increase, a	after tax increase, b	
1	27	40	-13
2	31	36	-5
3	23	43	-20
4	35	34	1
5	26	25	1
6	27	41	-14
7	26	32	-6
8	18	29	-11
9	22	21	1
10	21	36	-15

a) Determine if there is a significant difference in consumption prior to the tax increase and after, utilizing a hypothesis test (significance level $\alpha = 0.05$). The difference is assumed to be normally distributed.

b) Check your result by applying `t.test()` in R.

Tutorial 2 Business Analytics: Statistics

Exercise 2.2

According to the information supplied by the manufacturer of a certain type of car, its gas consumption in city traffic is approximately normally distributed with expected value $\mu = 9.5\ell/100\text{km}$. The standard deviation $\sigma = 2.5\ell/100\text{km}$ is commonly known (to the general public and the manufacturer). In order to review the manufacturer‘s prediction, a consumer organization has performed a test on 25 cars which yielded the following result:

Average gas consumption: $\bar{x} = 10.5\ell/100\text{km}$

Check the manufacturer‘s statement with a suitable test at significance level of $\alpha = 0.05$ and a second time with $\alpha = 0.01$.

Tutorial 2 Business Analytics: Statistics

Exercise 2.3

During a recent study project, a friend of yours asked 8 men and 10 women how many hours per day they wear a mask during the ongoing COVID-19 pandemic. The following table shows their answers. Afterwards he/she set the hypothesis to "On average, women wear their mask longer per day".

a) Test the hypothesis “by hand” with significance level and 16 degrees of freedom.
b) Try to find out how to solve this exercise using R.

Individual no. i	Hours per day	Gender
1	4	**female**
2	2	**female**
3	3	**female**
4	5	**female**
5	7	**female**
6	2	**female**
7	7	**female**
8	3	**female**
9	5	**female**
10	2	**female**
11	2	**male**
12	1	**male**
13	5	**male**
14	3	**male**
15	1	**male**
16	3	**male**
17	2	**male**
18	3	**male**