GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Univariate Statistics

Prof. Dr. Olaf Rank

Statistical Analysis

- Univariate Statistics
  - Hypothesis testing one variable at a time
  - Test of statistical significance
- Bivariate Statistics: Tests of hypotheses involving two variables.
- Multivariate Statistics: Statistical analysis involving three or more variables or sets of variables.

Hypothesis Testing

Hypothesis:

- Unproven proposition
- Supposition that tentatively explains certain facts or phenomena
- Assumption about nature of the world

Null hypothesis vs. Alternative hypothesis

- Alternative hypothesis
  - Statement that indicates the opposite of the null hypothesis
- Null hypothesis
  - Statement about the status quo
  - No difference

Hypothesis Testing and Statistical Significance

- Statistical inference: statements about population based on a sample
- Differences due to random sampling error vs. statistical significant results
- Results are statistically significant if it is unlikely that they have occurred by chance

Significance Levels and p-Values

Significance Level (Alpha):

- Critical probability in choosing between the null hypothesis and the alternative hypothesis
- Significance level selected is typically .05 or .01

p-value:

- Probability value, or the observed or computed significance level
- p-values are compared to significance levels to test hypotheses

Example

Customer satisfaction in a restaurant: owner wants to show, that actual satisfaction differs form 3 (= neither satisfied nor dissatisfied)

The null hypothesis that the mean is equal to 3.0:

H_0: $\mu = 3.0$

The alternative hypothesis that the mean is not equal to 3.0:

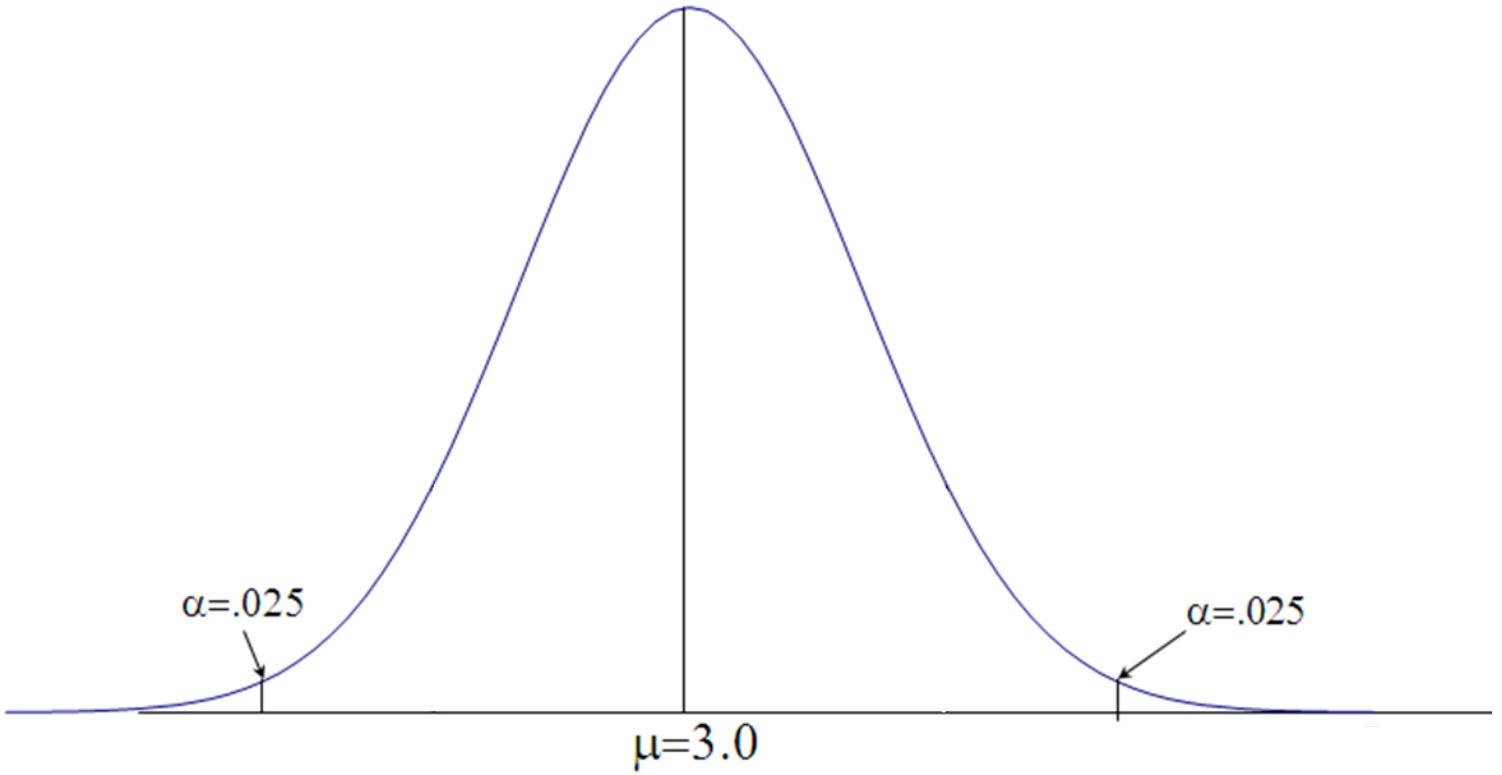
H_1: $\mu \neq 3.0$

N = 225

$\overline{X} = 3.78$

S = 1.5

A Sampling Distribution



Critical values of μ

Critical value - lower limit $= \mu - ZS_{\bar{X}}$ or $\mu - Z\frac{S}{\sqrt{n}}$

$$= 3.0 - 1.96\left(\frac{1.5}{\sqrt{225}}\right) = 2.804$$

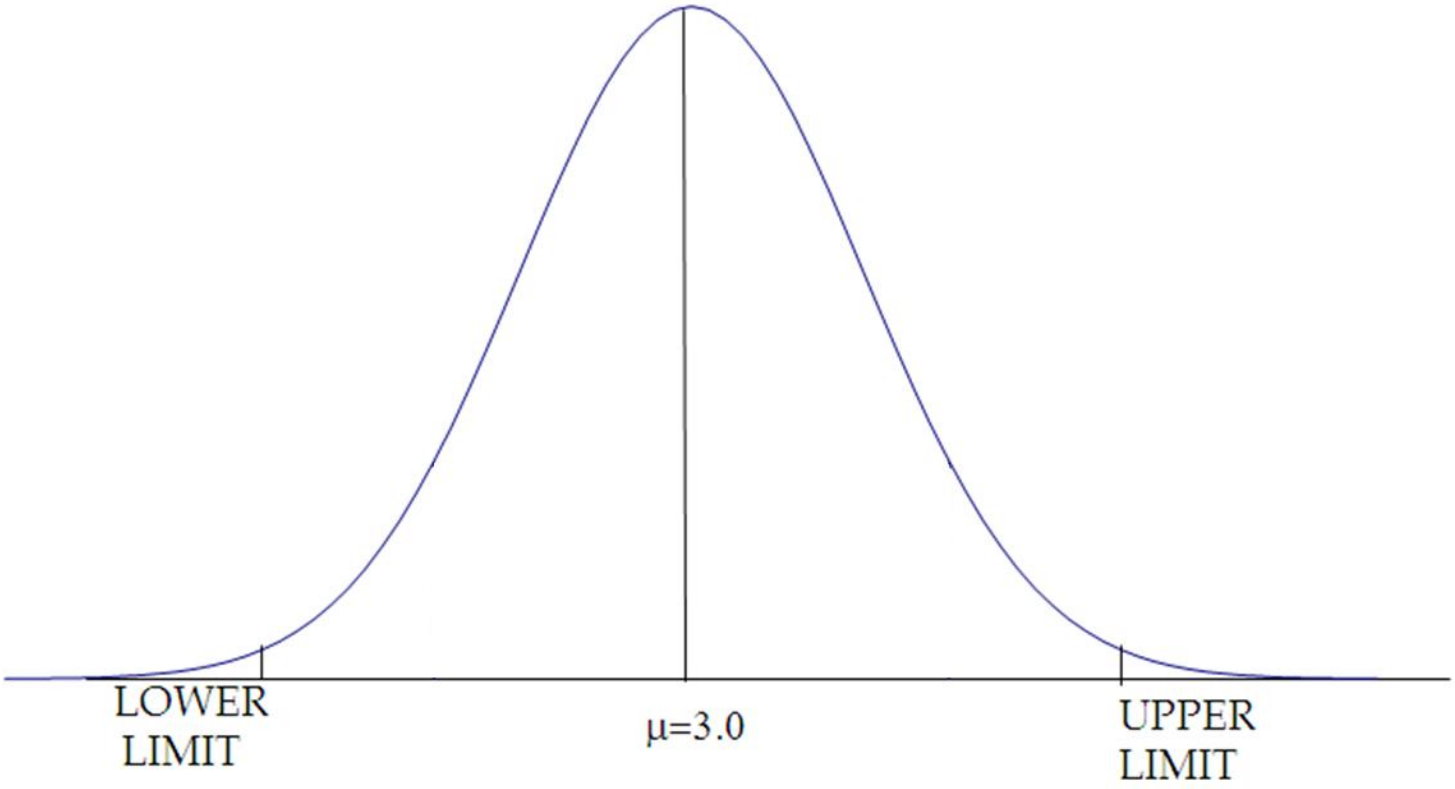
Critical value - upper limit $= \mu + ZS_{\bar{X}}$ or $\mu + Z\frac{S}{\sqrt{n}}$

$$= 3.0 + 1.96\left(\frac{1.5}{\sqrt{225}}\right) = 3.196$$

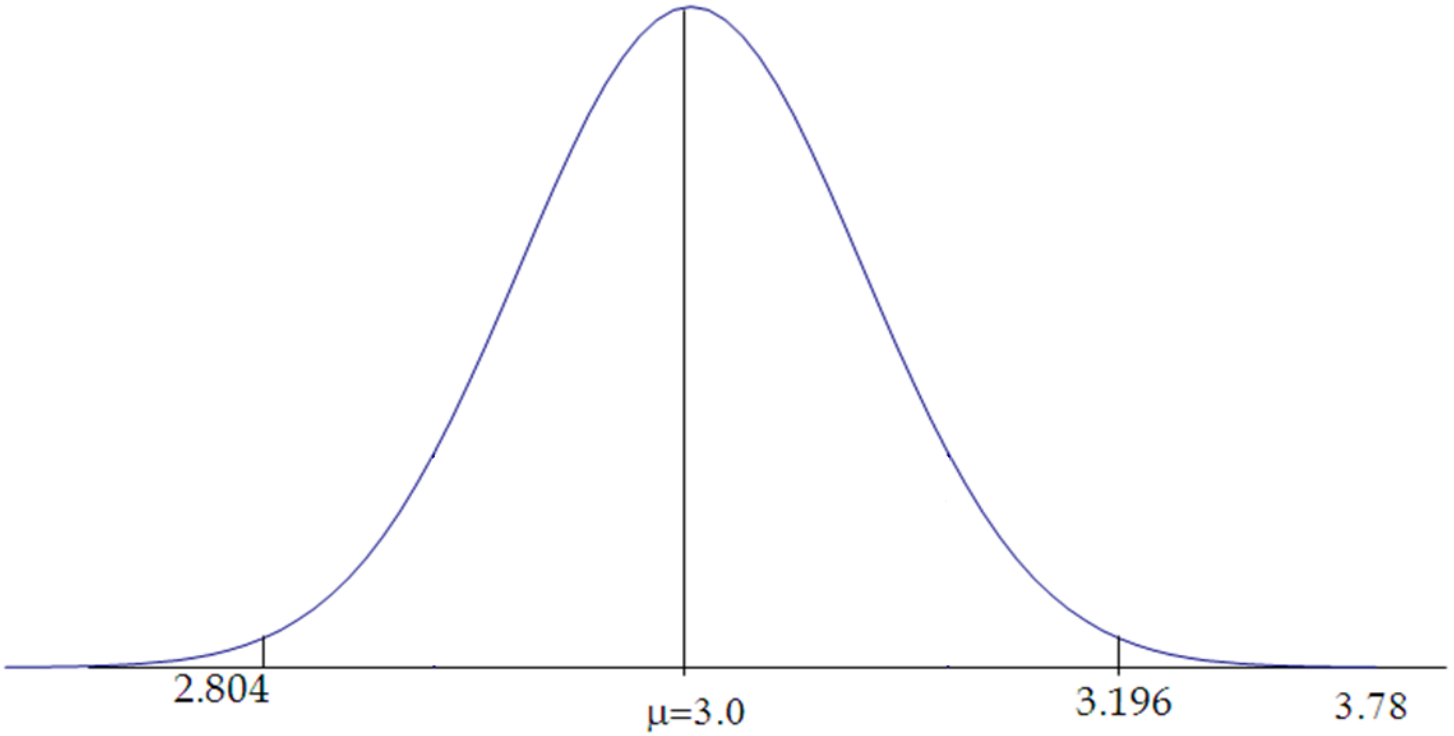
t-/z-Values

	Level of significance for one-tailed test					
	0.10	0.05	0.025	0.1	.005	.0005
	Level of significance for two-tailed test					
d.f.	.20	.10	.05	.02	.01	.001
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.598
3	1.638	2.353	3.182	4.541	5.841	12.941
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.859
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.405
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	2.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.767
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.045	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646
40	1.303	1.684	2.021	2.423	2.704	3.551
60	1.296	1.671	2.000	2.390	2.660	3.460
120	1.289	1.658	1.980	2.358	2.617	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.291

Region of Rejection



Hypothesis Test $\mu = 3.0$



The mean does not equal to 3.0 → rejection of the null hypothesis

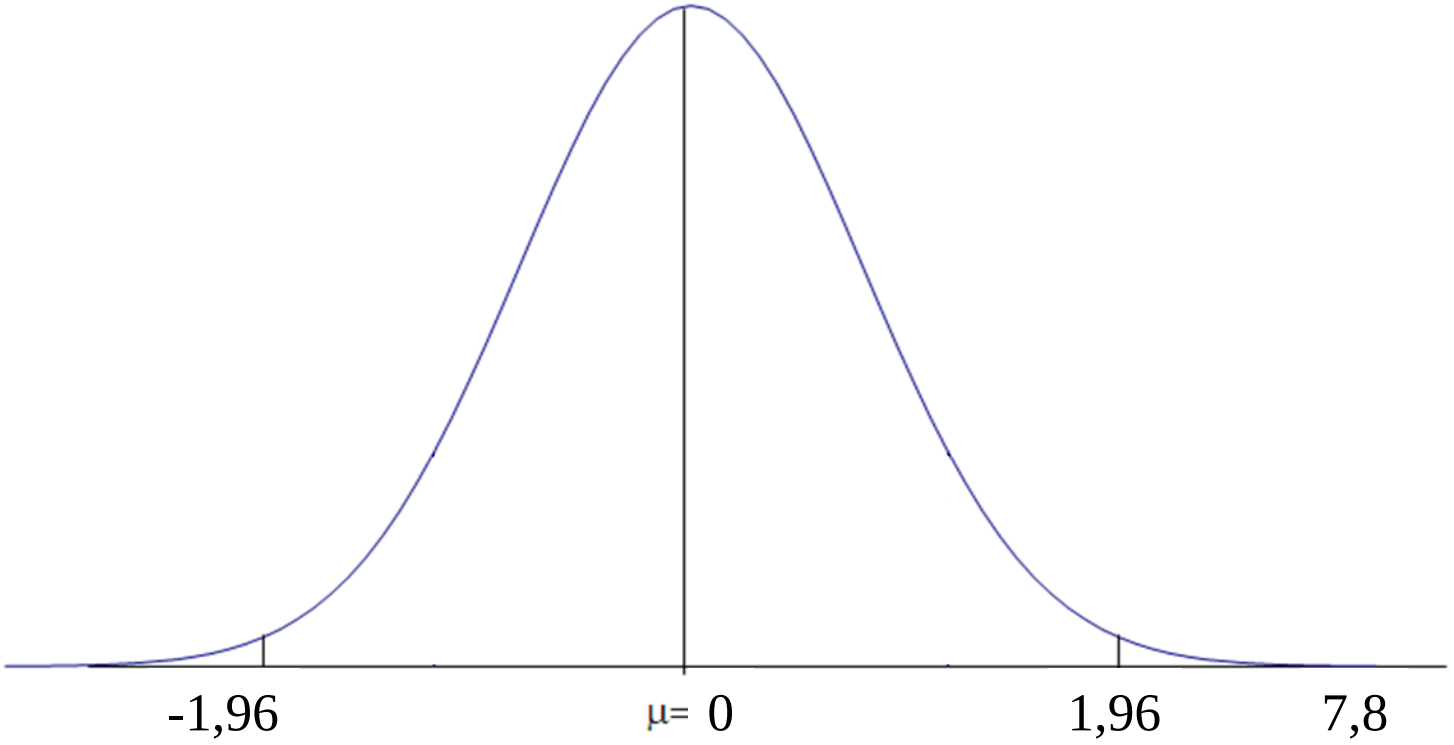
Alternate Way of Testing the Hypothesis

$$Z_{obs} = \frac{\bar{X} - \mu}{S_{\bar{X}}}$$

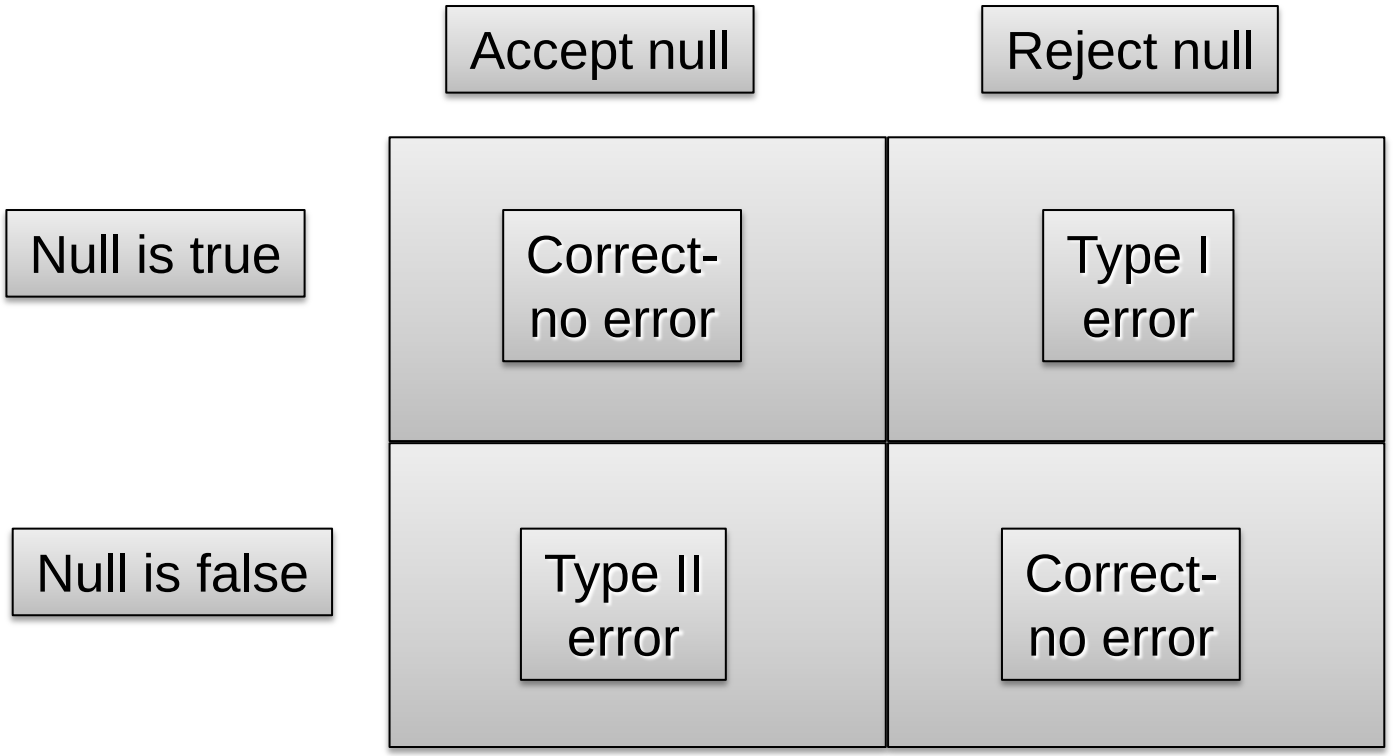
$$Z_{obs} = \frac{3.78 - \mu}{S_{\bar{X}}} = \frac{3.78 - 3.0}{.1}$$

$$= \frac{0.78}{.1} = 7.8$$

Hypothesis Test: Z-Values



Type I and Type II Errors



Choosing Statistical Techniques

Choice of appropriate statistical techniques depends on:

1. Type of question to be answered
2. Number of variables
   - Univariate
   - Bivariate
   - Multivariate
3. Scale of measurement

- Parametric vs. nonparametric statistics

t-Distribution and Degrees of Freedom

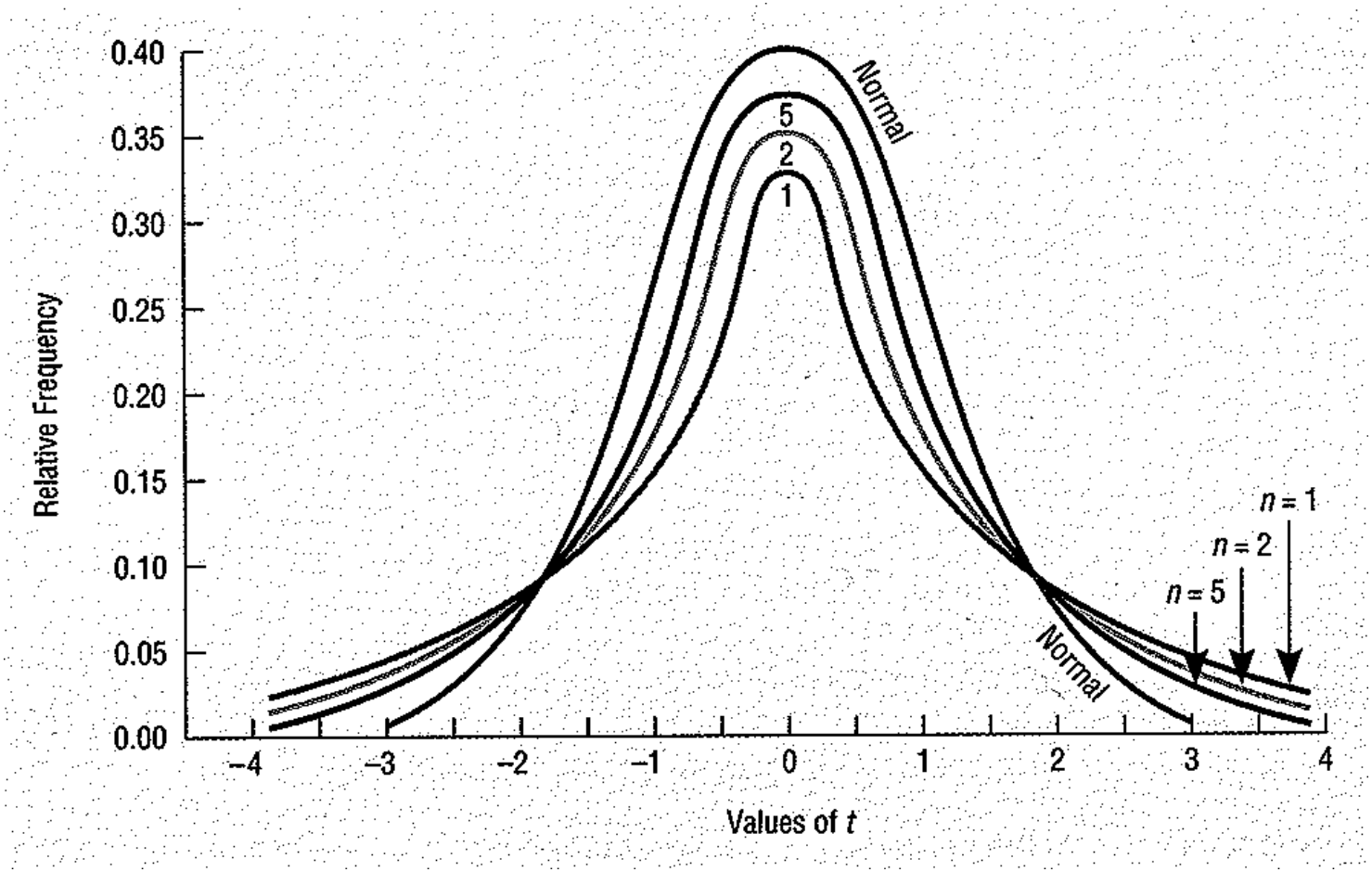
t-Distribution:

- Symmetrical, bell-shaped distribution
- Mean of zero and a unit standard deviation
- Shape influenced by degrees of freedom

Degrees of Freedom (d.f.): the number of observations minus the number of constraints/assumptions

t-Test: A hypothesis test that uses the t-distribution. A univariate t-test is appropriate when the variable analyzed is interval or ratio.

t-Distribution



Confidence Interval Estimate Using the t-distribution

$$\mu = \overline{X} \pm t_{c.l.} S_{\overline{X}}$$

or

$$\text{Upper limit} = \overline{X} + t_{c.l.} \frac{S}{\sqrt{n}}$$

$$\text{Lower limit} = \overline{X} - t_{c.l.} \frac{S}{\sqrt{n}}$$

μ = population mean

$\overline{X}$ = sample mean

$t_{c.l.}$ = critical value of t at a specified confidence level

$S_{\overline{X}}$ = standard error of the mean

S = sample standard deviation

n = sample size

Calculating a Confidence Interval Estimate using the t-Distribution

- Example: How long do MBA students remain on their first job?

$$\mu = \overline{X} \pm t_{cl} s_{\bar{x}}$$

$$\overline{X} = 3.7$$

$$S = 2.66$$

$$n = 17$$

$$\text{upper limit} = 3.7 + 2.12(2.66/\sqrt{17}) = 5.07$$

$$\text{lower limit} = 3.7 - 2.12(2.66/\sqrt{17}) = 2.33$$

- It can be concluded with 95 percent confidence that the population mean for the number of years spent on the first job by MBAs is between 5.07 and 2.33.

t-/z-Values

	Level of significance for one-tailed test					
	0.10	0.05	0.025	0.1	.005	.0005
	Level of significance for two-tailed test					
d.f.	.20	.10	.05	.02	.01	.001
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.598
3	1.638	2.353	3.182	4.541	5.841	12.941
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.859
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.405
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	2.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.767
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.045	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646
40	1.303	1.684	2.021	2.423	2.704	3.551
60	1.296	1.671	2.000	2.390	2.660	3.460
120	1.289	1.658	1.980	2.358	2.617	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.291

Univariate Hypothesis Test Utilizing the t-Distribution

Production manager wants to show, that the average number of defective assemblies each day differs from last years average number of 20

$$H_0 : \mu = 20$$

$$H_1 : \mu \neq 20$$

n = 25 days

$\overline{X}$ = 22

S = 5

Univariate Hypothesis Test Utilizing the t-Distribution

The researcher desired a 95 % confidence, the significance level becomes .05.

The researcher must then find the upper and lower limits of the confidence interval to determine the region of rejection. Thus, the value of *t* is needed. For 24 degrees of freedom (*n*-1, 25-1), the *t*-value is 2.064.

Lower limit:

$$\mu - t_{c.l.} S_{\overline{X}} = 20 - 2.064\left(\frac{5}{\sqrt{25}}\right) = 20 - 2.064 = 17.936$$

Upper limit:

$$\mu + t_{c.l.} S_{\overline{X}} = 20 + 2.064\left(\frac{5}{\sqrt{25}}\right) = 20 + 2.064 = 22.064$$

t-/z-Values

	Level of significance for one-tailed test					
	0.10	0.05	0.025	0.1	.005	.0005
	Level of significance for two-tailed test					
d.f.	.20	.10	.05	.02	.01	.001
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.598
3	1.638	2.353	3.182	4.541	5.841	12.941
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.859
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.405
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	2.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.767
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.045	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646
40	1.303	1.684	2.021	2.423	2.704	3.551
60	1.296	1.671	2.000	2.390	2.660	3.460
120	1.289	1.658	1.980	2.358	2.617	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.291

Univariate Hypothesis Test Utilizing the t-Distribution

Because

$$22 < 22.064$$

- Sample mean of is not included in the region of rejection.
- Consequence: Null Hypothesis cannot be rejected, manager's assumption seems not to be correct.

→ average number of defective assemblies each day does not differ significantly from last years average number of 20

Confidence Interval Estimate (t-distribution)

$$t_{obs} = \frac{\overline{X} - \mu}{S_{\overline{X}}} = \frac{22 - 20}{1} = \frac{2}{1} = 2$$

$$t_{crit} = 2.064 \text{ with } df = 25 - 1 = 24$$

Testing a Hypothesis about a Distribution

- Chi-Square test: Test for significance in the analysis of frequency distributions
- Compare observed frequencies with expected frequencies
- “Goodness of Fit”

Chi-Square Test

$$x^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

$$\chi^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}$$

x^2 = chi-square statistics
O_i = observed frequency in the i^{th} cell
E_i = expected frequency on the i^{th} cell

Example: Chi-Square Test

- Awareness of a brand of automobile tire

Awareness of Tire Manufacturer‘s Brand	Frequency
Aware	60
Unaware	40
	100

- H_0: number of consumers aware of tire brand equals the number unaware of the brand; expected probability (aware or unaware) = .5
- H_1: expected probability (aware or unaware) ≠ .5

Univariate Hypothesis Test: Chi-square Example

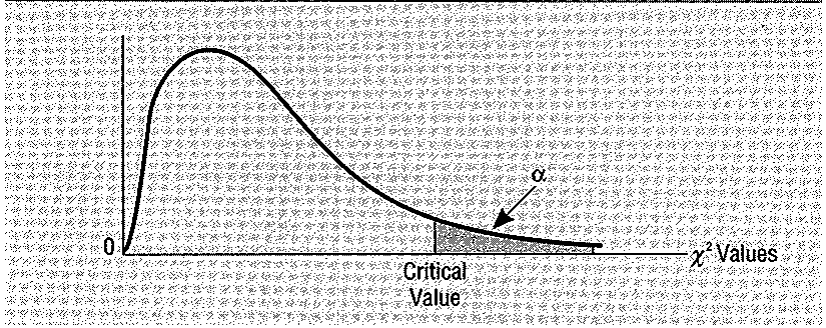
$$\chi^2 = \frac{(60-50)^2}{50} + \frac{(40-50)^2}{50} = 4$$

d.f. = k – 1, with k = number of cells associated with column or row data

d.f. = 2 – 1 = 1

Critical chi^2-value is 3.84 < 4

TABLE 4 CHI-SQUARE DISTRIBUTION



Degrees of Freedom (d.f.)	AREA IN SHADED RIGHT TAIL (α) .10	.05	.01
1	2.706	3.841	6.635
2	4.605	5.991	9.210
3	6.251	7.815	11.345
4	7.779	9.488	13.277
5	9.236	11.070	15.086
6	10.645	12.592	16.812
7	12.017	14.067	18.475
8	13.362	15.507	20.090
9	14.684	16.919	21.666
10	15.987	18.307	23.209
11	17.275	19.675	24.725
12	18.549	21.026	26.217
13	19.812	22.362	27.688
14	21.064	23.685	29.141
15	22.307	24.996	30.578
16	23.542	26.296	32.000
17	24.769	27.587	33.409
18	25.989	28.869	34.805
19	27.204	30.144	36.191
20	28.412	31.410	37.566
21	29.615	32.671	38.932
22	30.813	33.924	40.289
23	32.007	35.172	41.638
24	33.196	36.415	42.980
25	34.382	37.652	44.314
26	35.563	38.885	45.642
27	36.741	40.113	46.963
28	37.916	41.337	48.278
29	39.087	42.557	49.588
30	40.256	43.773	50.892