Data and Tests

PROF. DR. AKIN PALA

Test assumptions

- Distribution is normal
- Variances are homogenous

- Why care about assumptions?
- You make an amazing discovery of differences between two groups that is accepted for publication in a prestigious journal.
- Later you have to write a retraction because your analysis was shown to be invalid by a competitor
  - You'll be very embarrassed.

Test assumptions

- If homogeneity of variance assumption is violated, then the unequal variance *t-test should be used*.
- *If a* parametric statistic is inappropriate, consider the use of non-parametrics or using a transformation method.
- Most problematic assumption violation is the normal distribution, so you can start off by analyzing your data for normality.

Data Transformation

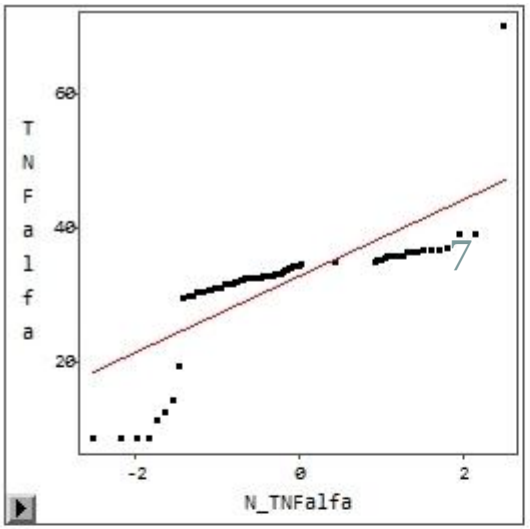
- Statistical procedures assume that the data distribution is normal.
- If not,
  - Use nonparametric methods
  - Transform the data
- Conception rate and weaning weight.
- Use bootstrapping
- Multiply each animal's weaning weight with the breed's conception rate.

Data Transformation

- Angus:
  1. 240kg*0.90
  2. 250*0.90
  3. 260*0.90
- Brangus:
  1. 250*0.80
  2. 260*0.80
  3. 270*0.80
- Angus average ww/cow exposed=250*0.9=225kg
- Brangus average ww/cow exposed=260*0.8=208kg
  - Now you included 0 and 1 conception rate

Data Transformation

- Evaluate the data set and decide which transformations are appropriate.
  - Graph the data, look at its shape, or make a QQ plot
  - Use methods such as Shapiro-Wilk, Anderson-Darling and the Kolmogorov-Smirnov for normality tests



Normal QQ Ref Lines

Line	Intercept	Slope
	32.8088	5.7060

Moments

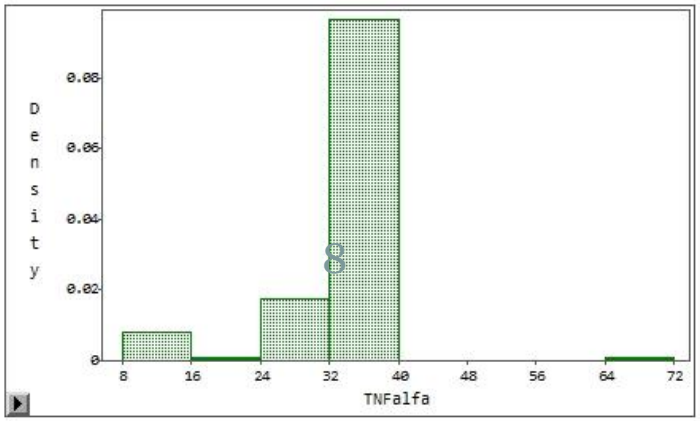
N	107.0000	Sum Wgts	107.0000
Mean	32.7854	Sum	3508.0400
Std Dev	7.2903	Variance	53.1491
Skewness	-0.6481	Kurtosis	10.6891
USS	120646.368	CSS	5633.8011
CV	22.2365	Std Mean	0.7048

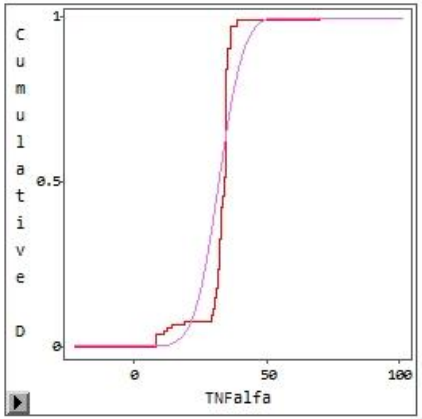
Quantiles

100% Max	70.6000	99.0%	39.4200
75% Q3	35.0000	97.5%	39.3600
50% Med	34.6500	95.0%	36.9800
25% Q1	32.5500	90.0%	35.9500
0% Min	8.6600	10.0%	30.1500
Range	61.9400	5.0%	12.7500
Q3-Q1	2.4500	2.5%	8.7200
Mode	35.0000	1.0%	8.7000

Tests for Normality

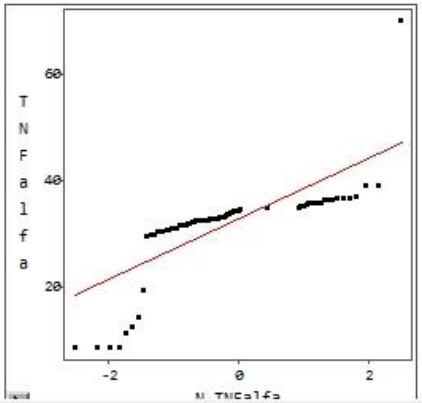
Test Statistic	Value	p-value
Shapiro-Wilk	0.590183	0.0000
Kolmogorov-Smirnov	0.276760	<.0100
Cramer-von Mises	3.103084	<.0050
Anderson-Darling	16.36451	<.0050



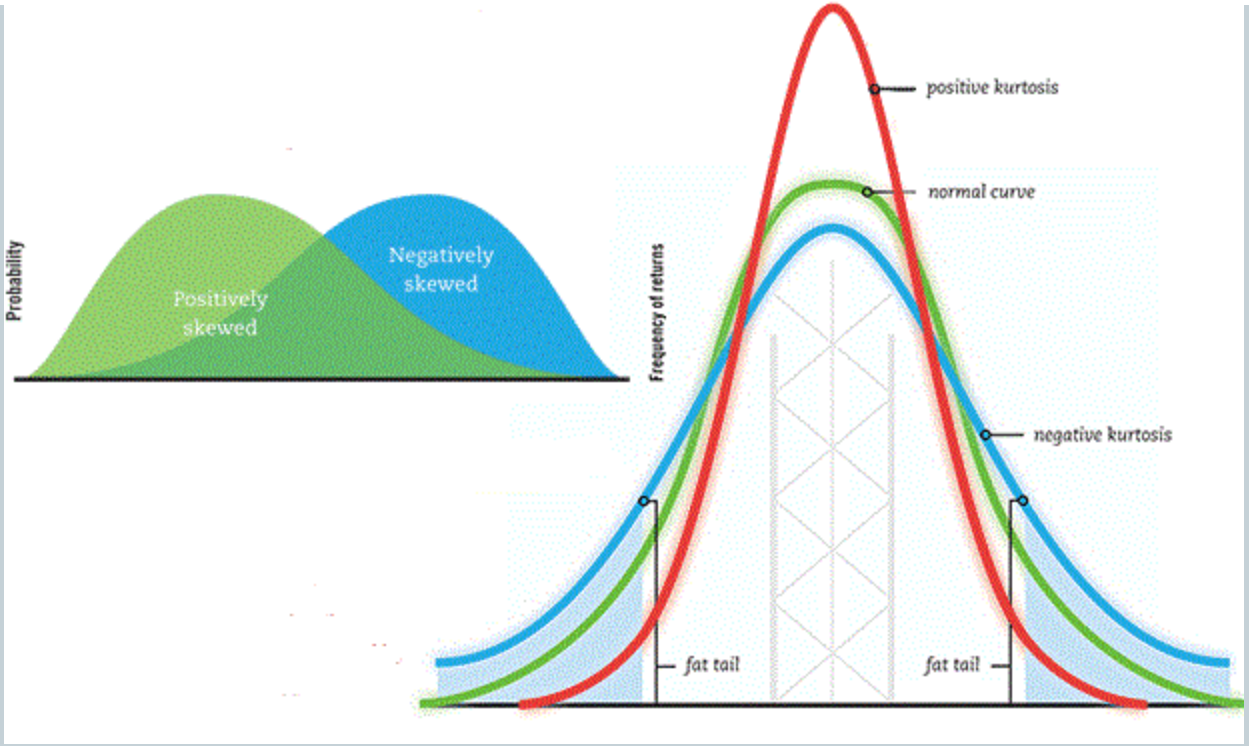


Tests for Distribution

Curve	Distribution	Mean/Theta	Sigma	Kolmogorov D	Pr > D
	Normal	32.7854	7.2903	0.2768	<.01



Skewness and Kurtosis



Kurtosis is how peaked the curve is. Stepper: positive curtosis, flatter: negative kurtosis.
Skewness is the asymmetry of a distribution. Positive skewness: A tail pulled in the positive (right) direction. Negative skewness: A tail pulled in the negative (left) direction.

Right skewed data

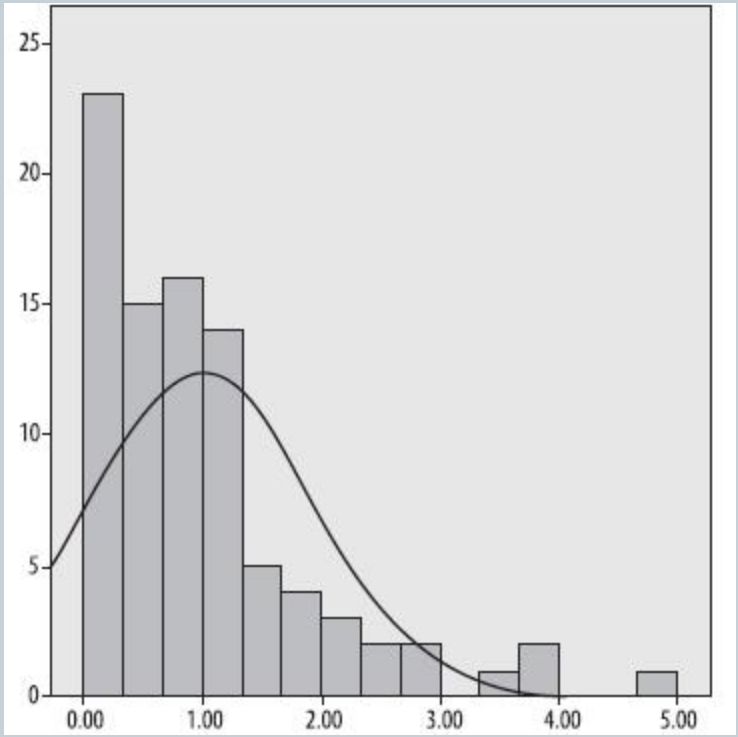
- Data that is right-skewed may be made more normal by either taking squareroot or log transformations.
- The square-root transformation computes the square root of each value.
  - If the data value is 9, the transformed value is 3 ($\sqrt{9} = 3$).
- The log transformation computes the natural log of each value
  - If the data value is 4, the transformed value is 1.386 because $ln(4) = 1.386$.

SAS for Square root and log

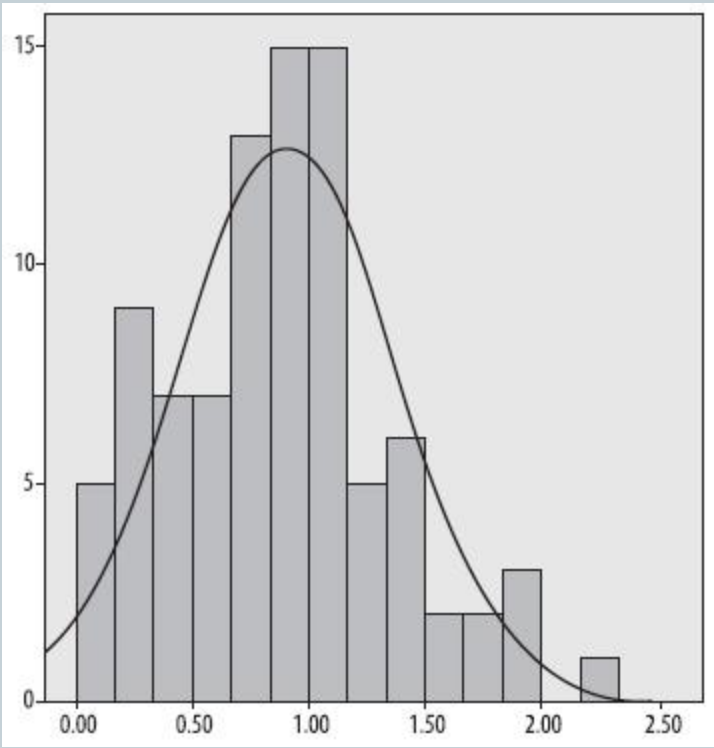
- data log;
- xlog=log(4);
- xsquareroot=sqrt(9);
- proc print;
- run;

Obs	xlog	xsquareroot
1	1.38629	3

Before and After Picture



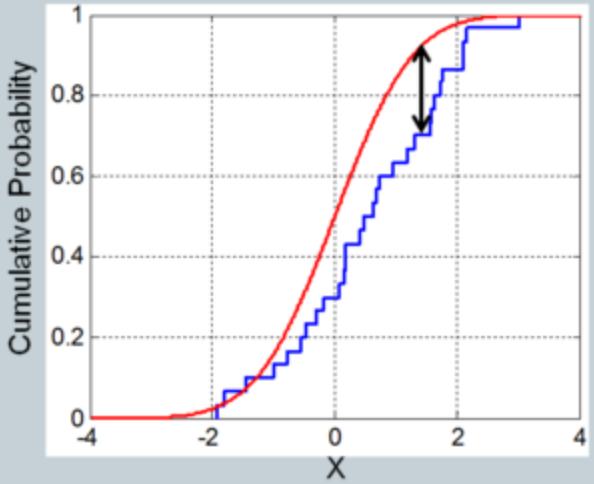
Right skewed raw data



Data after square-root transformation

Kolmogorov-Smirnov test for normality for a data set and two transformations

	Raw data	Square root transformation	Natural log transformation
Kolmogorov-Smirnov Z	1.46	0.66	1.41
P	0.029	0.78	0.04



The Kolmogorov–Smirnov statistic quantifies a distance between the empirical distribution function of the sample (blue line) and the cumulative distribution function of the reference distribution (red line), or between the empirical distribution functions of two samples.

Transformation Changes the Data

- The transformed data should also be evaluated for normality.
- Transformation changes the unit of the data
  - If you apply the log transformation to a population of Somatic Cell Count, your unit is then the log of SCC and not the original scores.
- The effects of other transformation must be kept in mind when interpreting the statistical results.

T Test

- T-*test: makes inferences about single means*, about two means or variances, where sample sizes are small.
- Not a lot of people use it
  - Analysis of Variance (ANOVA) is mathematically equivalent to the *t-test*
  - Most people try and avoid type II error, so they collect a reasonable number of samples
- *Central limit theorem: the sampling distribution of a statistic (like a sample mean) will follow a normal distribution, as long as the sample size is large.*
- *Understanding t test* will make it much easier for you to follow ANOVA & other sophisticated analytical techniques

Student's t Test

- Industrial statistician William Gosset in the 20th century worked at the Guinness Brewery in Dublin for quality assurance for beer brewing.
- After studying statistics with Karl Pearson, Gosset published a paper under the pseudonym "Student,"
  - Guinness did not want their competitors to know that they were employing statistics to improve quality control.
- Gosset's key observation was the dependence on sample size for determining the probability that the mean of the population lies within a given distance of the mean of the sample, if a normal distribution is assumed.

Student's t Test

- Using numerical simulations, Gosset illustrated:
- When you have a normal distribution, and if the number of samples is small,
- The distribution (for the variable *x)* is both flatter, and has more observations appearing in the tails, than a normal distribution, when sample sizes are less than 30, and where $\frac{s}{\sqrt{n}}$ is std err.

$$t = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}$$

x: sample mean
μ: population mean
s: standard deviation of the sample
n: sample size.

Student's t Test

$$t = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}$$

x: sample mean
μ: population mean
s: standard deviation of the sample
n: sample size.

- The formula looks at the differences between the sample and population mean.
- This is called one sample t test
  - There is one sample, and it is compared to the population mean.
- Two sample t test compares two sample means.

Student's t Test

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{x1x2} * \sqrt{\frac{1}{n}}}$$

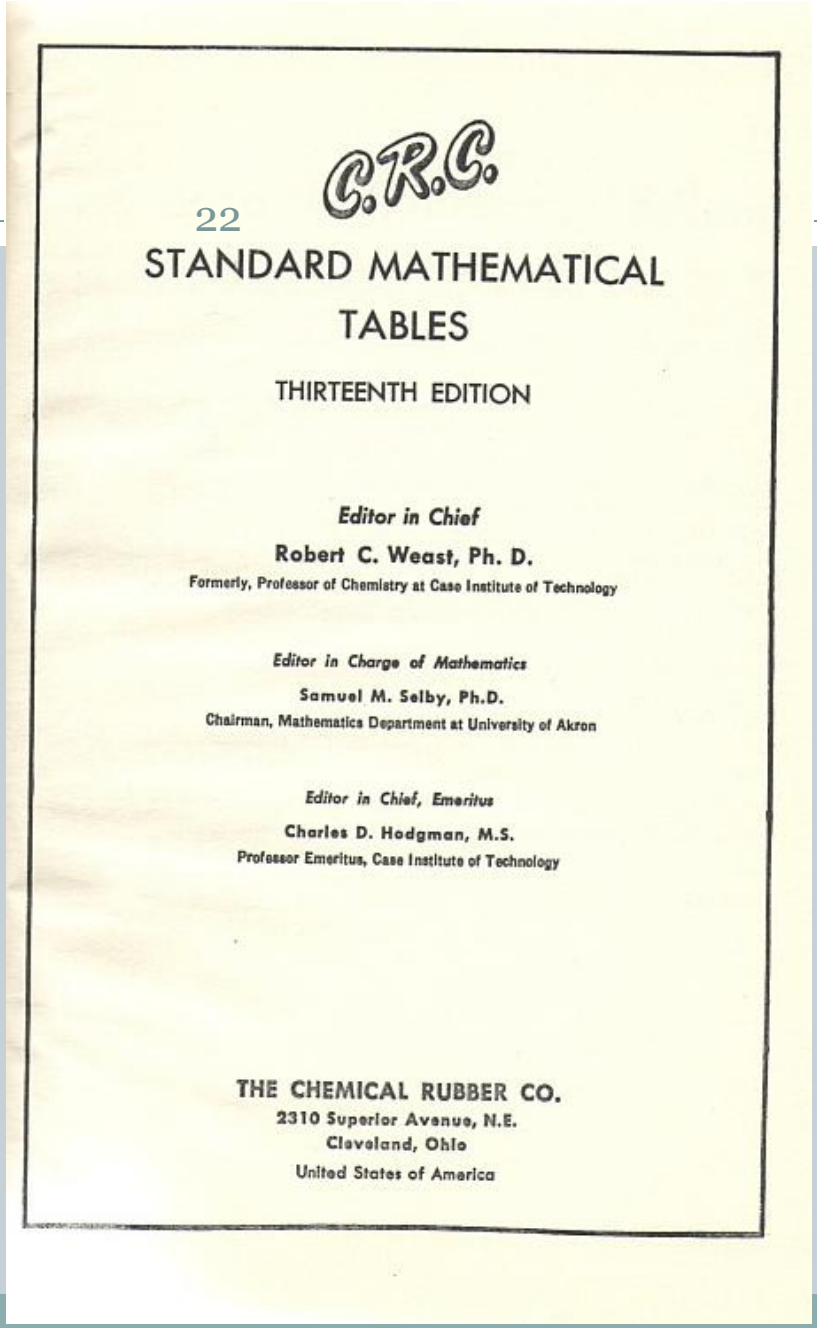
$$s^2_{x1x2} = \sqrt{(s^2_{x1} + s^2_{x2})}$$

- The formula looks at the differences between two different means.
- Denominator is the standard error of the difference between two means.
- s_{x1x2} is the pooled standard deviation

- As the number of samples increases, the distribution becomes normal, given the dependence on *n, and the corresponding effect on df*, since $df = n - 1$.
- *This distribution is known as the t distribution, and* approximates a normal distribution if *n (and by implication df) are large (>30 in* practical terms).

- Books of statistical tables normally provide critical values of *t that can be used at* different degrees of freedom to make inferences about the population, with an associated probability of committing a Type I error (α).
- For example, where *n = 21 and df = 20, then t = 1.725 at the p = 0.05 significance level, and t = 2.528 at the p = 0.01* significance level.
- Expression: $t_{0.05,20} = 1.725$ *and* $t_{0.01,20} = 2.528$

The table in the next slide is from this 1964 book:



C.R.C.

STANDARD MATHEMATICAL TABLES

THIRTEENTH EDITION

Editor in Chief

Robert C. Weast, Ph. D.

Formerly, Professor of Chemistry at Case Institute of Technology

Editor in Charge of Mathematics

Samuel M. Selby, Ph.D.

Chairman, Mathematics Department at University of Akron

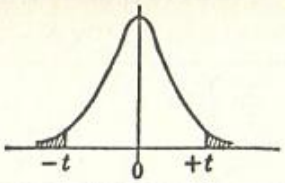
Editor in Chief, Emeritus

Charles D. Hodgman, M.S.

Professor Emeritus, Case Institute of Technology

THE CHEMICAL RUBBER CO.
2310 Superior Avenue, N.E.
Cleveland, Ohio
United States of America

TABLE FOR t TEST OF SIGNIFICANCE BETWEEN TWO SAMPLE MEANS ($\bar{x}_1$ AND $\bar{x}_2$)



Degrees of freedom	*P = 0.9	0.8	0 7	0.6	0.5	0.4	0.3	0.2	0.1	0.05	0.02	0.01
1	0.158	0.325	0.510	0.727	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657
2	0.142	0.289	0.445	0.617	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925
3	0.137	0.277	0.424	0.584	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841
4	0.134	0.271	0.414	0.569	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604
5	0.132	0.267	0.408	0.559	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032
6	0.131	0.265	0.404	0.553	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707
7	0.130	0.263	0.402	0.549	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499
8	0.130	0.262	0.399	0.546	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355
9	0.129	0.261	0.398	0.543	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250
10	0.129	0.260	0.397	0.542	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169
11	0.129	0.260	0.396	0.540	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106
12	0.128	0.259	0.395	0.539	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055
13	0.128	0.259	0.394	0.538	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012
14	0.128	0.258	0.393	0.537	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977
15	0.128	0.258	0.393	0.536	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947
16	0.128	0.258	0.392	0.535	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921
17	0.128	0.257	0.392	0.534	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898
18	0.127	0.257	0.392	0.534	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878
19	0.127	0.257	0.391	0.533	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861
20	0.127	0.257	0.391	0.533	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845
21	0.127	0.257	0.391	0.532	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831
22	0.127	0.256	0.390	0.532	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819
23	0.127	0.256	0.390	0.532	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807
24	0.127	0.256	0.390	0.531	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797
25	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787
26	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779
27	0.127	0.256	0.389	0.531	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771
28	0.127	0.256	0.389	0.530	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763
29	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756
30	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750
∞	0.12566	0.25335	0.38532	0.52440	0.67449	0.84162	1.03643	1.28155	1.64485	1.95996	2.32634	2.57582

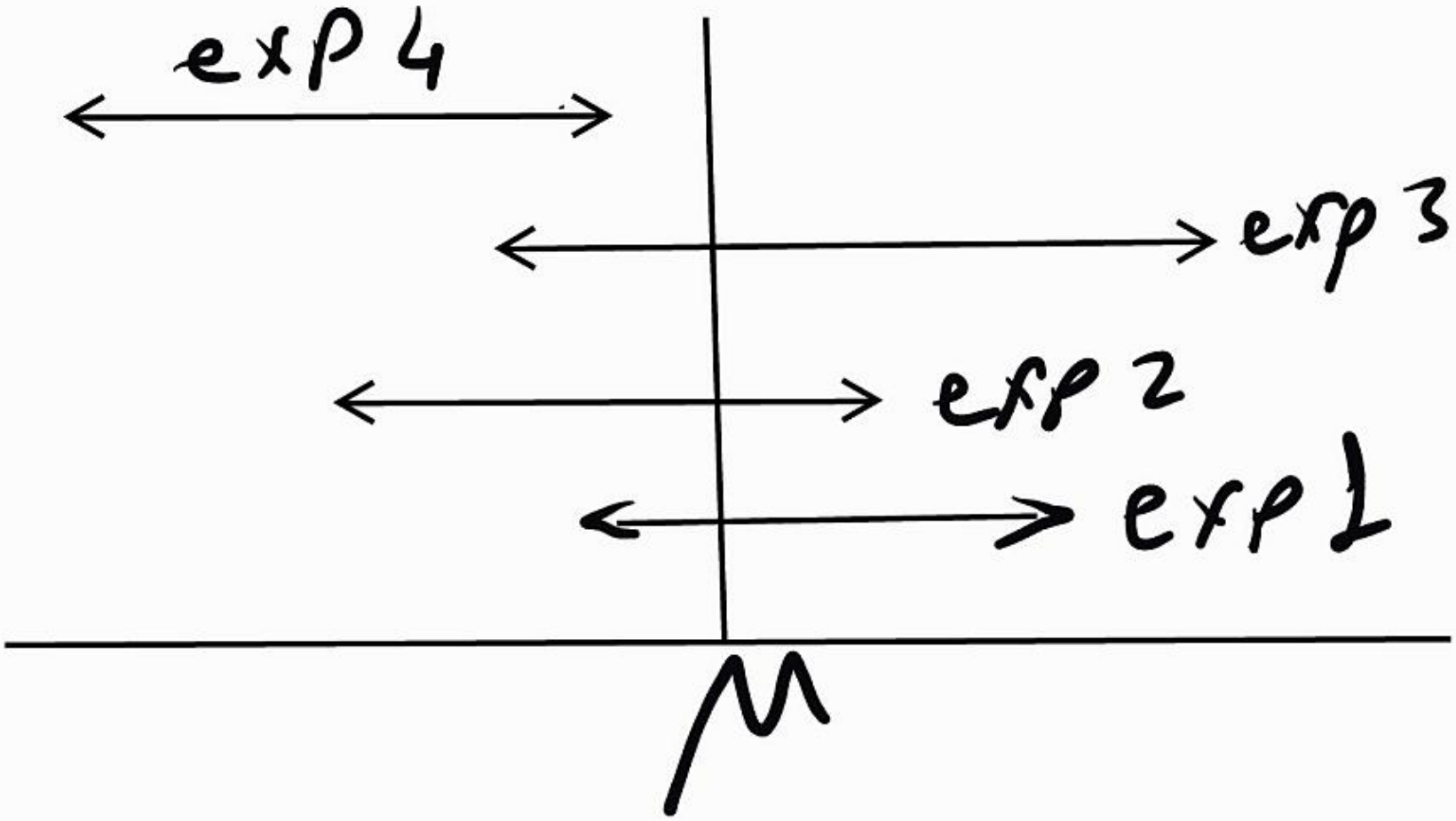
Reproduced from *Statistical Methods for Research Workers*, 6th ed., with the permission of the author, R. A. Fisher, and his publisher, Oliver and Boyd, Edinburgh.

* P is the probability of having t this large or larger in size by chance.

Confidence Interval, level, limit

- **A confidence level (coefficient):** the degree of confidence, or certainty, that the researcher wants to be able to place in the confidence interval.
- **Confidence Interval**: The interval that will include the population parameter a certain percentage (confidence level) of the time in the long run (over repeated sampling)
  - The range of values of a sample statistic that is likely (at a given level of probability: confidence level) to contain a population parameter.
- *Statistics* are used in samples to estimate analogous *parameters* in the population from which the sample was drawn.
  - A sample mean statistic, Y, is often calculated to estimate the population parameter μ.
- **Confidence limit**: the upper and lower values.

95 % of these will cover μ



Factors affecting CI

- The length decreases as n increases.

$$\bar{Y} - t^* \frac{s}{\sqrt{n}} < \mu < \bar{Y} + t^* \frac{s}{\sqrt{n}}$$

- Length increases as s (sample variance) increases
- Length increases as confidence level increases
  - t value increases because alpha decreases

Confidence Interval t value example

- Construct a 95 percent confidence interval for average sugar content of a certain variety of strawberry
- Assume sugar contents form a normal population.
- Take a random sample of 20 sugar contents.
- μ=average sugar for this variety

Confidence Interval T value example

- Average sugar content of this variety for our sample is Y=120 gr/l, s=17gr/l
- t=2.093 from the table 19 df and 0.05 alpha

$$\bar{Y} - t^* \frac{s}{\sqrt{n}} < \mu < \bar{Y} + t^* \frac{s}{\sqrt{n}}$$

- =120-2.093*(17/sqrt20)<μ<120+2.093*(17/sqrt20)

  112.04gr/l<μ<127.96 gr/l
- When we state the sugar content is X<μ<Y, the accuracy is 95 percent.
  - **Not** that the probability is 95 % for being in that interval

Using SAS for formula calculations

- data calculation;
- minus=120-2.093*(17/(sqrt(20)));
- plus=120+2.093*(17/(sqrt(20)));
- proc print;
- run;

Obs	minus	plus
1	112.044	127.956

T test example

- Acme Corporation claims that an average Acme light bulb lasts 300 days.
- A researcher randomly selects 15 bulbs for testing.
  - The sampled bulbs last an average of 290 days
  - Standard deviation of 50 days.
- If the CEO's claim were true, what is the probability that 15 randomly selected bulbs would have an average life of no more than 290 days?

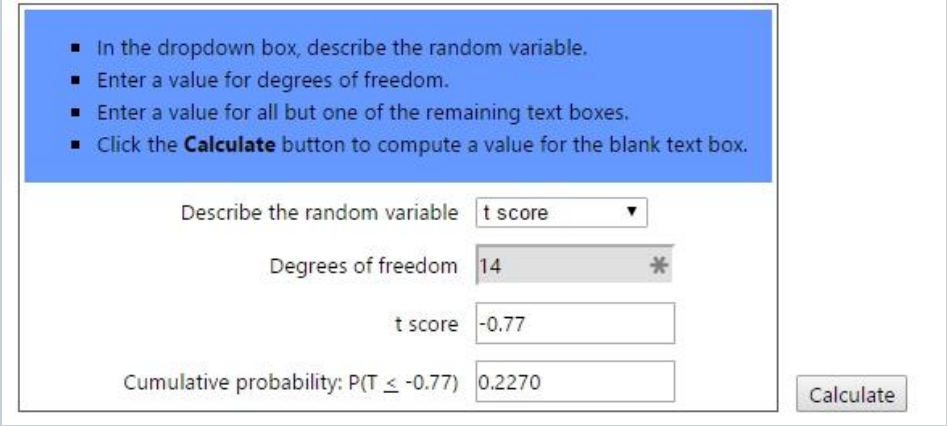
$$t = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}$$

t = (290 - 300) / [50 / sqrt(15)] = -10 / 12.909945 = - 0.7745966

  - where x is the sample mean, μ is the population mean, s is the standard deviation of the sample, and n is the sample size.
- You can compare this to the t value from the table for alpha=0.05, which is 2.145, so P>0.05

Online resources, no code required.

- You can use
  - http://stattrek.com/online-calculator/t-distribution.aspx
- instead of lengthy tables, which does not include negative values.
- The degrees of freedom= 15 - 1 = 14.
- The t score= - 0.7745966.
- The calculator displays the cumulative probability: 0.227 (only one blank)



- If the true bulb life were 300 days, there is a 22.7% chance that the average bulb life for 15 randomly selected bulbs would be less than or equal to 290 days.

T test

- You can use the formula
- To calculate the t value from the data and compare it to the alpha=0.05 t table/online calculator value to see if it differs from
  - The population mean (one sample t test)
  - Any other mean:another group (two sample t test)

T test using SAS


```
data random;
input numbers;
cards;
225.675573406358
230.403723018554
189.516286360279
332.934122981807
329.230863443389
257.930855990253
248.208620360979
327.671778383415
295.673233811555
309.618487511914
270.394292656366
341.765371894855
358.122251787641
336.545748642036
337.948933502283
;
Title 'Random number generators are more successful with more numbers';
title2 'Prof. Dr. Akin Pala';
proc ttest H0=300;
var numbers;
run;
```

T test using SAS

Random number generators are more successful with more numbers
Prof. Dr. Akin Pala

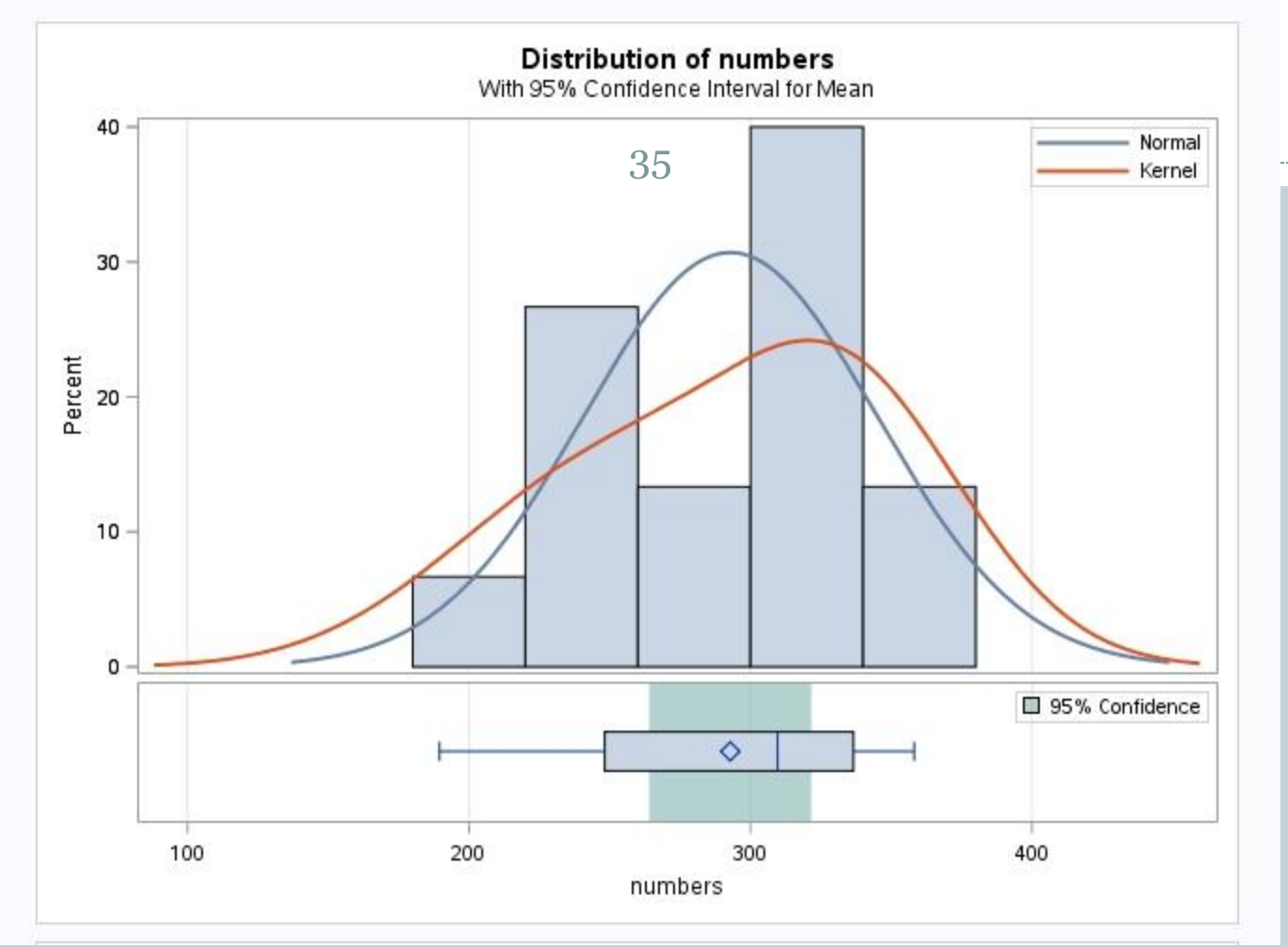
The TTEST Procedure
Variable: numbers

N	Mean	Std Dev	Std Err	Minimum	Maximum
15	292.8	51.9853	13.4225	189.5	358.1

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
292.8	264.0	321.6	51.9853	38.0598	81.9860

DF	t Value	Pr > |t|
14	-0.54	0.5989

the *confidence limits*: The upper and lower values of a confidence interval; the values defining the range of a confidence interval



Our random numbers are closer to 300 than anticipated, and they are more successful, less different than 300. The output graph shows that the data lean towards 300 though the mean is 292.8

T test using SAS

- Using SAS is the easiest
- Pros:
  - Gives you more resolution in terms of output
  - Gives you graphs
  - Gives you CI
  - Once you write one code, you can modify it to apply to other situations
- Cons:
  - You need to be careful, if you don't know the underlying theory, you may be fooled.
  - Computers are not sausage machines

Manual calculation of two groups t test

- An age-old physical performance question is whether male football players are fitter than male ballet dancers, so a sports physiologist organizes a study in partnership with a local hospital research team to answer the question.
- The two groups are independent populations, since no football player is also a ballerina.

Manual calculation of two groups t test

- There are also two lists of ballet dancers and football players located all over the country that are maintained by their respective professional associations, and study members are randomly selected from each group.
- Since ballet dancers and football players are very busy, only 10 study members from each group can be recruited.

Manual calculation of two groups t test

- All will be tested on performance
  - Walking, running, stepping,
- Physiological measures related to fitness
  - Heart-rate variability, pulse-wave velocity
- Combined to form a single fitness score out of 100.
- The participants are all tested
  - in the same facility
  - at the same time of day
- Their responses are assessed and combined using the same clinicians.

Fitness results for football players and ballet dancers

Ballet dancers	Football players
89.2	79.3
78.2	78.3
89.3	85.3
88.3	79.3
87.3	88.9
90.1	91.2
95.2	87.2
94.3	89.2
78.3	93.3
89.3	79.9

Manual calculation of two groups t test

- $\mu_{ballet} = 87.95$, $\mu_{football} = 85.19$
- $s^2_{ballet} = 32.38$, and $s^2_{football} = 31.18$
- Independent two-sample t-test with equal sample sizes, equal variance is only used when:
  - The two sample sizes
    - the number, *n*, of participants of each group are equal;
  - It can be assumed that the two distributions have the same variance
    - If they have similar variances, its OK.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{x1x2} * \sqrt{\frac{1}{n}}}$$

Independent two-sample t-test with equal sample sizes, equal variance

- $\mu_{ballet} = 87.95$, $\mu_{football} = 85.19$
- $s^2_{ballet} = 32.38$, and $s^2_{football} = 31.18$

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{x1x2} * \sqrt{\frac{1}{n}}}$$

$$s_{x1x2} = \sqrt{(s^2_{x1} + s^2_{x2})}$$

$$s_{x1x2} = \sqrt{(32.38 + 31.18)} = 7.97$$

$$t = \frac{87.95 - 85.19}{7.97 * \sqrt{\frac{1}{10}}} = \frac{2.76}{7.97 * 0.32} = 1.09$$

Manual calculation of two groups t test

- Independent two-sample t-test with equal or unequal sample sizes, equal variance is only used when:
  - It can be assumed that the two distributions have the same variance.
- Formula becomes:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{x1x2} * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where $$s_{x1x2} = \sqrt{\frac{(n_1 - 1)s_{x1}^2 + (n_2 - 1)s_{x2}^2}{n_1 + n_2 - 2}}$$

Independent two-sample t-test with equal or unequal sample sizes, equal variance

- $\mu_{ballet} = 87.95$, $\mu_{football} = 85.19$ $s^2_{ballet} = 32.38$, $s^2_{football} = 31.18$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{x1x2} * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$s_{x1x2} = \sqrt{\frac{(n_1 - 1)s^2_{x1} + (n_2 - 1)s^2_{x2}}{n_1 + n_2 - 2}}$$

$$= \sqrt{\frac{(10-1)32.38 + (10-1)31.18}{10+10-2}} = 5.64$$

$$t = \frac{87.95 - 85.19}{5.64 * \sqrt{\frac{1}{10} + \frac{1}{10}}} = \frac{2.76}{5.64 * 0.45} = 1.09$$

Independent two-sample t-test with equal or unequal sample sizes, equal variance

- $\mu_{ballet} = 87.95$, $\mu_{football} = 85.19$ $s^2_{ballet} = 32.38$, $s^2_{football} = 31.18$
- Calculated t=1.09
- The $df = n1 + n2 - 2 = 18$
- $P = 0.05$, $t_{0.95,18} = 1.734$
- Fail to reject the null hypothesis
  - Difference in fitness between the two groups are due to chance for the fitness measure used, and the samples analyzed.
  - They are not large enough to be significant.

Equal or unequal sample sizes, unequal variances

- The formula becomes:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{x}_1 - \bar{x}_2}}$$

- where

$$S_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

- Not pooled variance. S^2 is the unbiased estimator of the variance of the two samples, n_i = number of participants in group i, i=1 or 2.

Dependent t-test (repeated measures)

- The samples are dependent; that is, when there is only one sample that has been tested twice (repeated measures).
- Hearing impaired listeners with: (1) sloping, high-frequency hearing loss are compared with (2) hearing impaired listeners showing a flat hearing loss.
- They're tested on their ability to identify words from a standard speech intelligibility test.
- Independent groups or repeated measures?
- Independent groups. They're different people.

Dependent t-test (repeated measures)

- Children are tested prior to and following a language enrichment program.
- We want to know if the program worked, so language scores prior to the program are compared to those following the program.
- Independent groups or repeated measures?
- Repeated measures - they're the same subjects tested under different conditions.

- , a stimulus is being examined to determine its effect on systolic blood pressure. Twelve men participate in the study. Each man's systolic blood pressure is measured both before and after the stimulus is applied.
- data pressure; input SBPbefore SBPafter @@; datalines; 120 128 124 131 130 131 118 127 140 132 128 125 140 141 135 137 126 118 130 132 126 129 127 135
- ;
- run;

- The variables *SBPbefore* and *SBPafter* denote the systolic blood pressure before and after the stimulus, respectively.
- proc ttest; paired SBPbefore*SBPafter; run;
- The paired statement is used to test whether the mean change in systolic blood pressure is significantly different from zero.

T test Golf example

- Golf scores for males and females in a physical education class are in SAS program for t test.

Data and Tests

- You can use ANOVA, instead of the t test, but if you have more than two groups, you have to use ANOVA or other methods.