Descriptive Statistics

Prof. Dr. Akın Pala

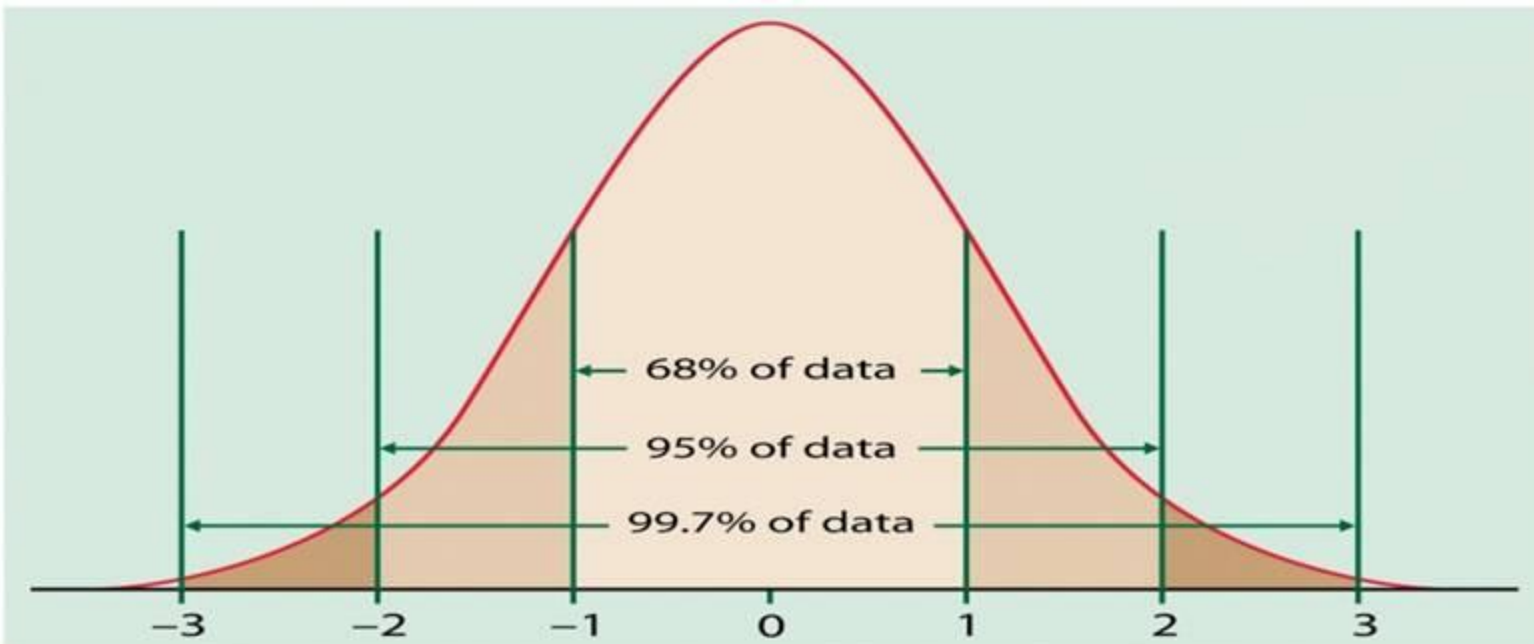
Representation

- Data must be summarized.
- Table form
- Graph form
- Mathematical form

Representation, Table form

- The numerical summary measures
  - For example the center and spread of a distribution of student cycle threshold in real time PCR, identify many important features.
- If we want to know the "typical" CT value
  - given by the center of the distribution
- The spread of the distribution of CT value,
- The relative position of a gene with a particular CT value, the numerical summary measures can provide more detailed information

Distribution



A classic table

Variable Name	Frequency	%
Categories		
Total		

A classic table, example Blood groups

Blood groups	Frequency	%
A B AB O	10 10 20 10	20 20 40 20
Total	50	100

Table example for age

Age (year)	Frequency	%
20<30 30 40 50+	12 18 25 15	24 36 10 30
Total	50	100

Smoking	Lung cancer				Total	
	Cases		Control			
	No.	%	No.	%	No.	%
Smoker	15	75%	8	20%	23	38.33
Non smoker	5	25%	32	80%	37	61.67
Total	20	100	40	100	60	100

Leptin in fat tissue

Least Squares Means							
Effect	group	Time	Estimate	Standard Error	DF	t Value	Pr > |t|
group	1		4.9156	0.4690	45	10.48	<.0001
group	2		5.6428	0.4690	45	12.03	<.0001
group	3		6.1683	0.4690	45	13.15	<.0001
Time		0	4.9789	0.4690	45	10.61	<.0001
Time		12	5.4239	0.4690	45	11.56	<.0001
Time		24	6.3239	0.4690	45	13.48	<.0001

Differences of Least Squares Means

Effect	group	time	_group	_time	Estimate	Std Error	DF	t Value	Pr > |t|
group	1		2		−0.72	0.66	45	−1.10	0.278
group	1		3		−1.25	0.66	45	−1.89	0.065
group	2		3		−0.52	0.66	45	−0.79	0.432
time		0		12	−0.44	0.66	45	−0.67	0.505
time		0		24	−1.34	0.66	45	−2.03	0.048
time		12		24	−0.90	0.66	45	−1.36	0.181

Chicken Descriptive Statistics

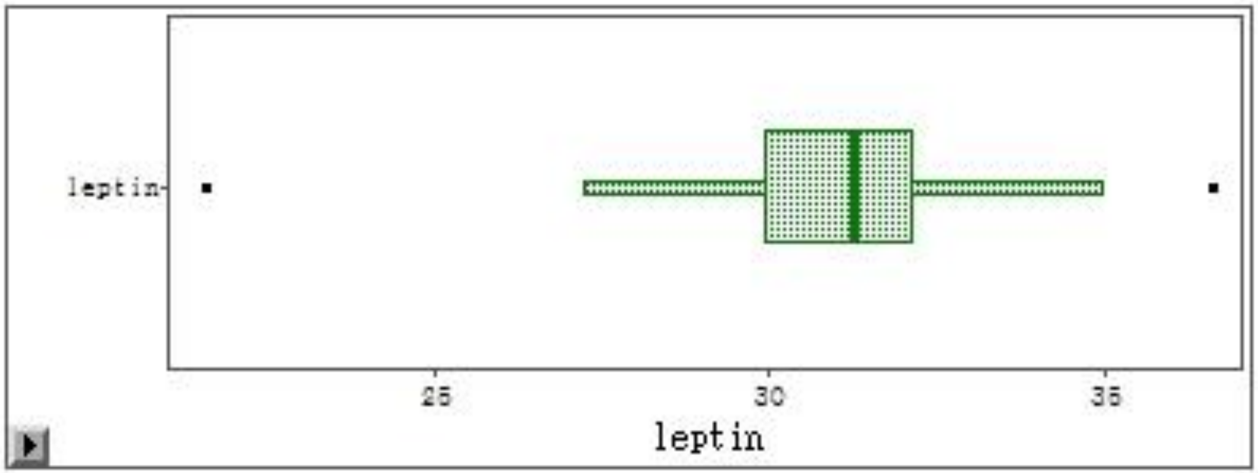
Table 1. Descriptive statistics by lighting programs and sex

23L:1D Lighting program						
Sex	Male	Female	Male		Female	
Traits	$\overline{X} \pm S_{\overline{X}}$	$\overline{X} \pm S_{\overline{X}}$	Min	Max	Min	Max
Live weight (Y) (gr)	7642±352	5245±189	5420	8650	4150	6180
Shank Length (X1) (cm)	16.45±0.49	13.25±0.15	13.50	17.75	12.50	14.25
Breast Length (X2) (cm)	16.25±0.25	14.80±0.11	15.00	17.50	14.00	15.00
Breast Depth (X3) (cm)	6.03±0.19	5.78±0.10	5.25	7.00	5.25	6.25
Breast Circumference (X4) (cm)	72.90±1.51	63.00±0.78	65.00	79.00	60.00	66.00
18L:6D Lighting program						
Live weight (Y) (gr)	7721±386	5225±175	6100	9500	4250	6000
Shank Length (X1) (cm)	15.63±0.35	13.28±0.17	13.25	16.50	12.50	14.00
Breast Length (X2) (cm)	16.15±0.31	14.98±0.19	15.00	17.75	14.00	16.00
Breast Depth (X3) (cm)	5.58±0.10	5.40±0.12	5.00	6.00	4.50	5.75
Breast Circumference (X4) (cm)	72.30±1.05	65.20±1.23	65.00	76.00	60.00	74.00
12L:12D Lighting program						
Live weight (Y) (gr)	6298±228	5443±97.6	5700	7750	5110	5950
Shank Length (X1) (cm)	14.22±0.44	13.08±0.12	13.25	17.00	12.50	13.75
Breast Length (X2) (cm)	15.08±0.16	14.93±0.15	14.25	16.00	14.25	15.75
Breast Depth (X3) (cm)	5.11±0.37	5.40±0.17	4.25	5.75	4.50	6.00
Breast Circumference (X4) (cm)	64.11±0.74	62.60±0.31	62.00	69.00	61.00	64.00

Graph Form

- Pretty easy using excel, power point, SAS

Box Plot (shows distribution)



Multi Dimensional Scaling

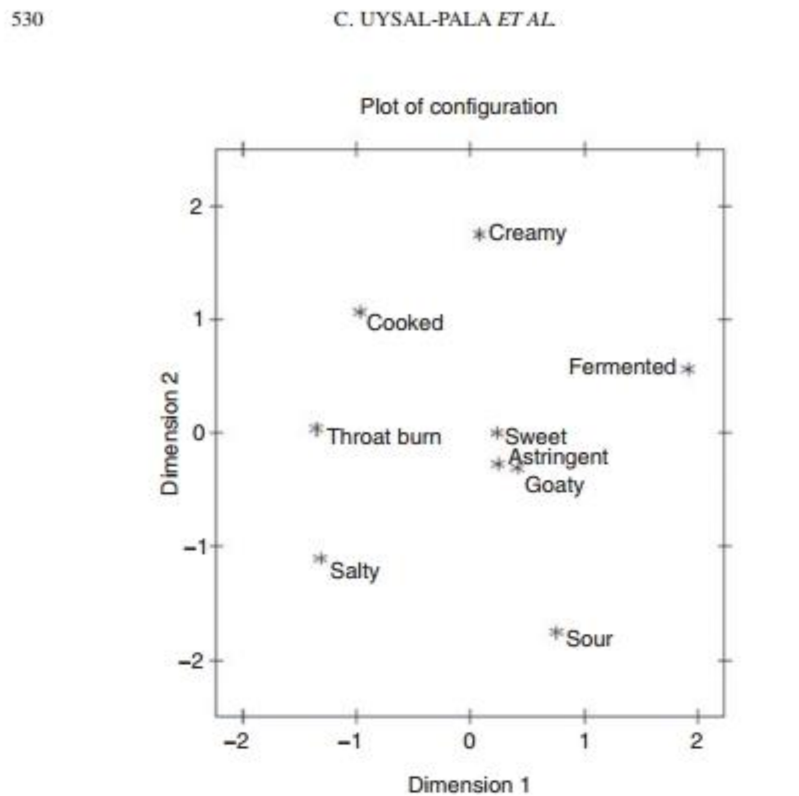
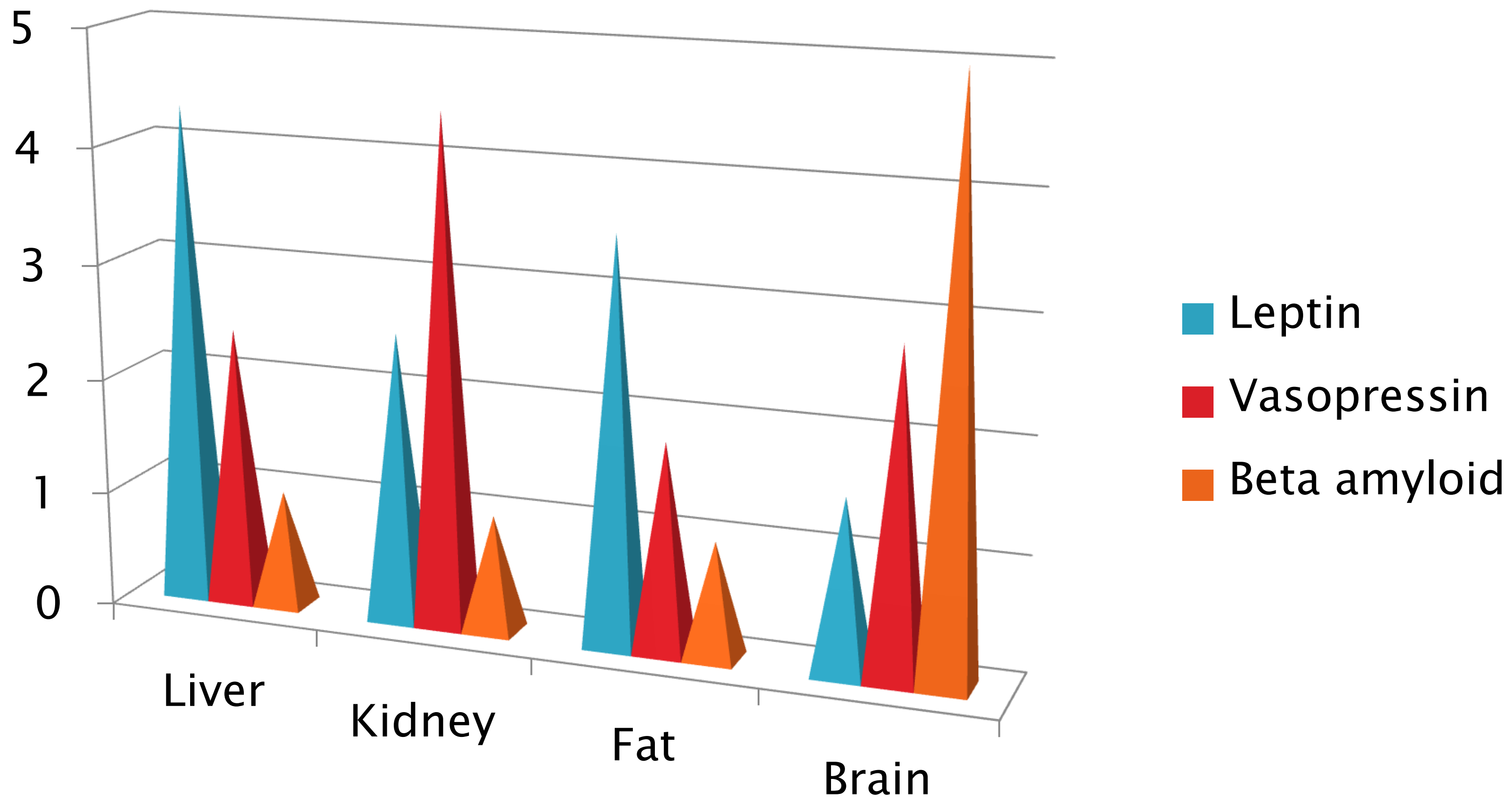


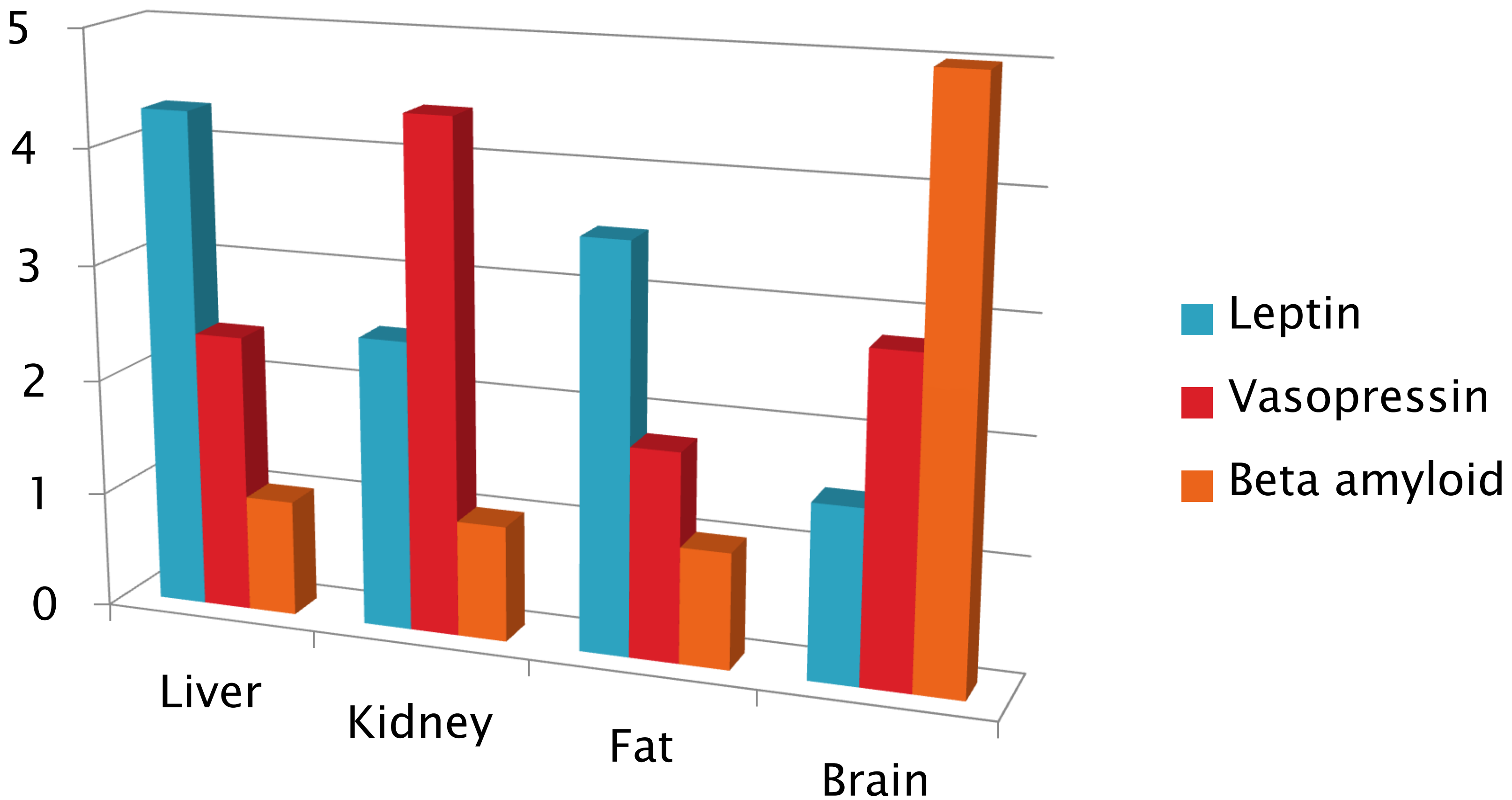
FIG. 1. VISUAL REPRESENTATION OF THE VARIABLES, WITH THE FIRST TWO DIMENSIONS

C. U. Pala, Y. K. Yuceer, A. Pala, and T. Savas. 2006. Sensory Evaluation of Drinkable Yogurt Made from Different Goat Breeds. Journal of Sensory Studies 21: 520-533.

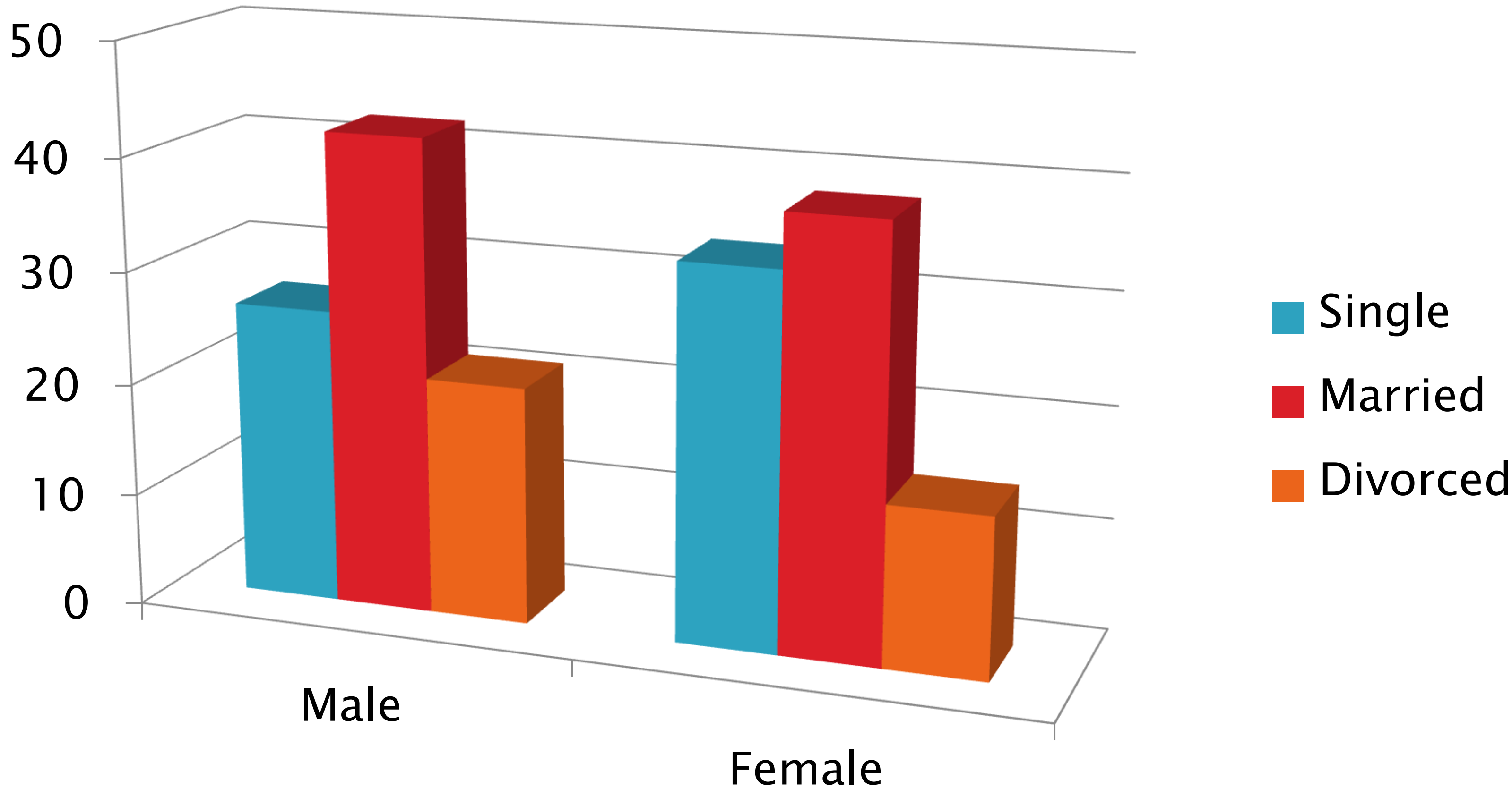
Bar graph, histogram



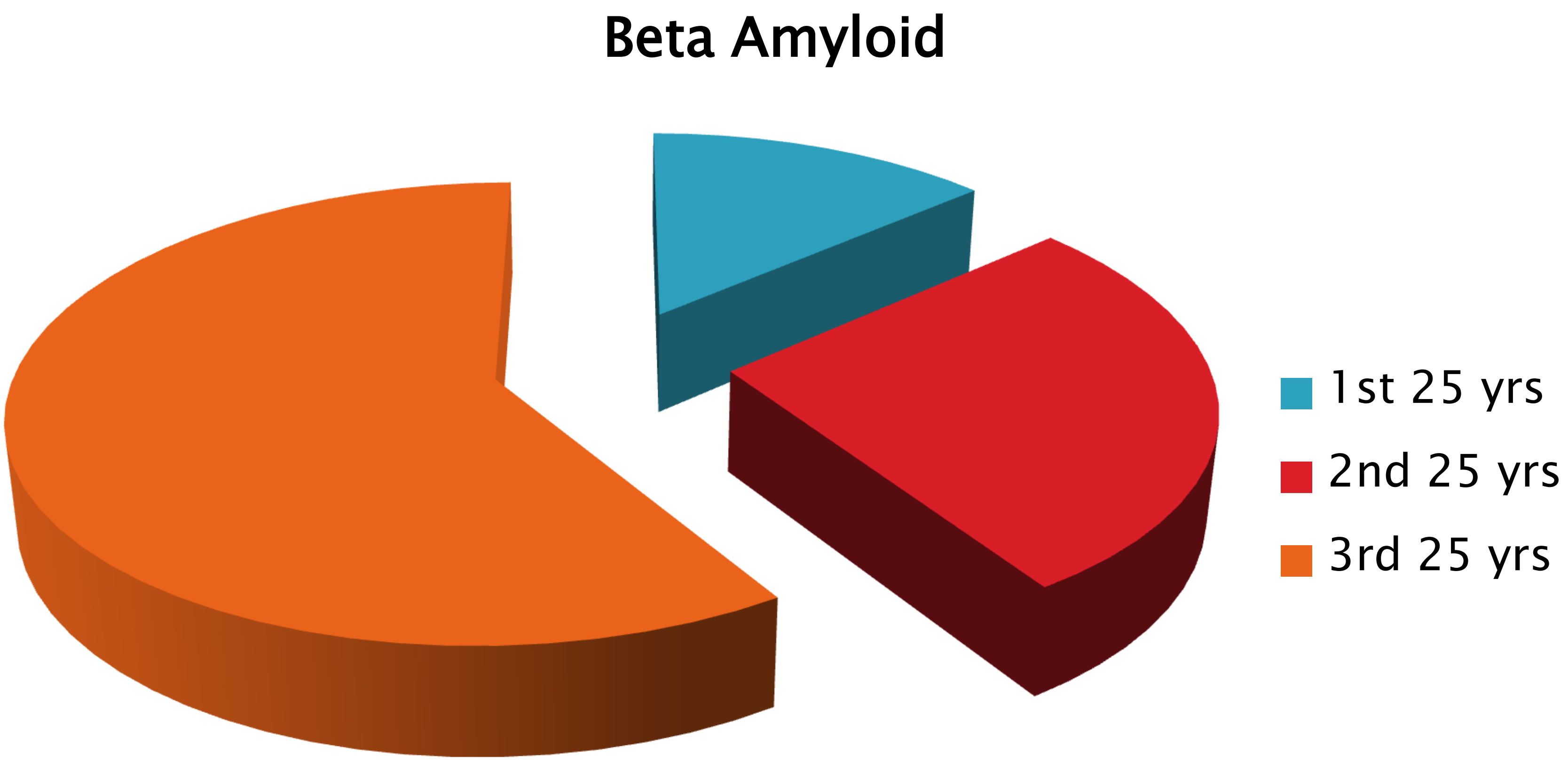
Bar graph, histogram



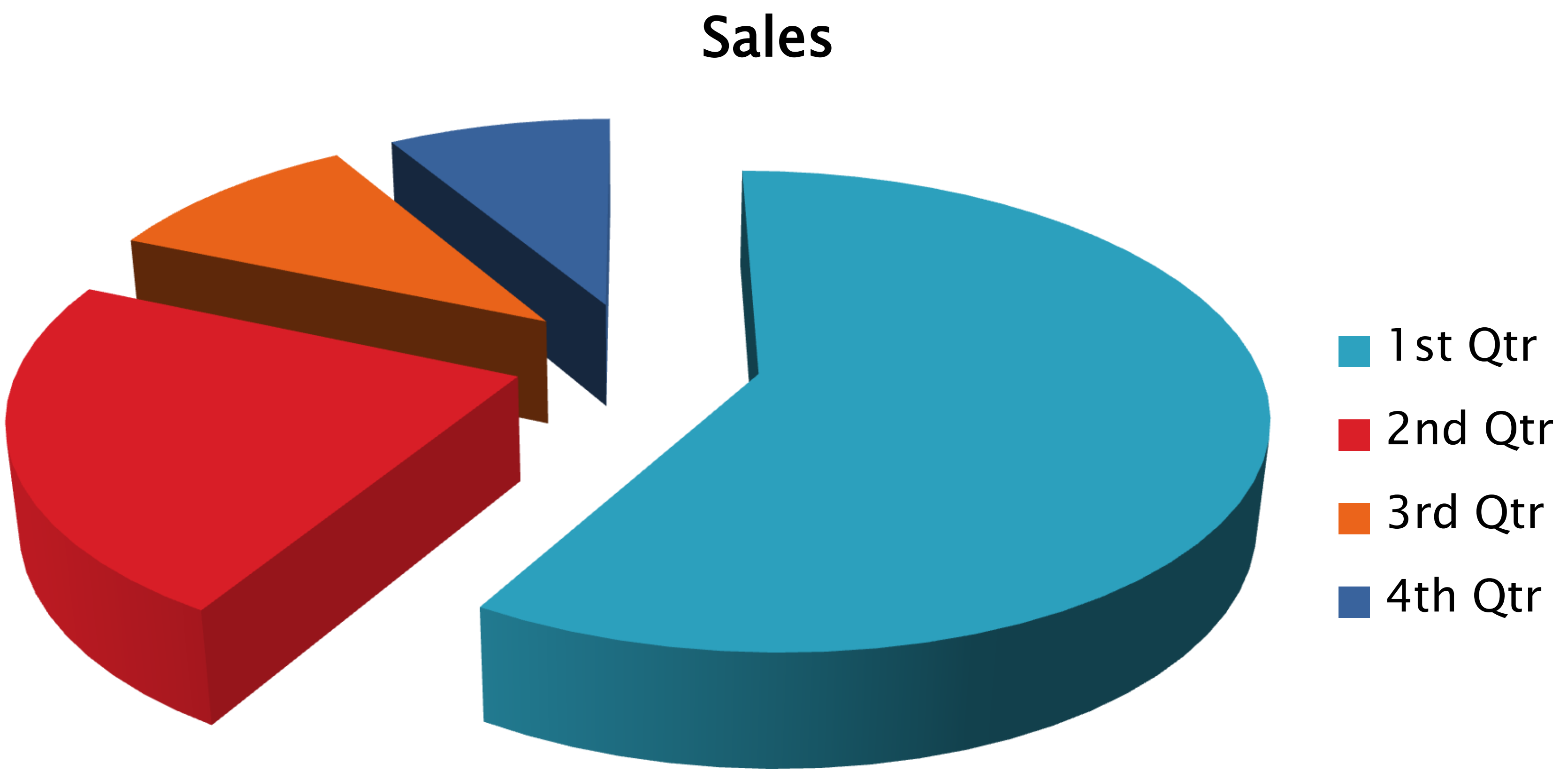
Bar graph, histogram



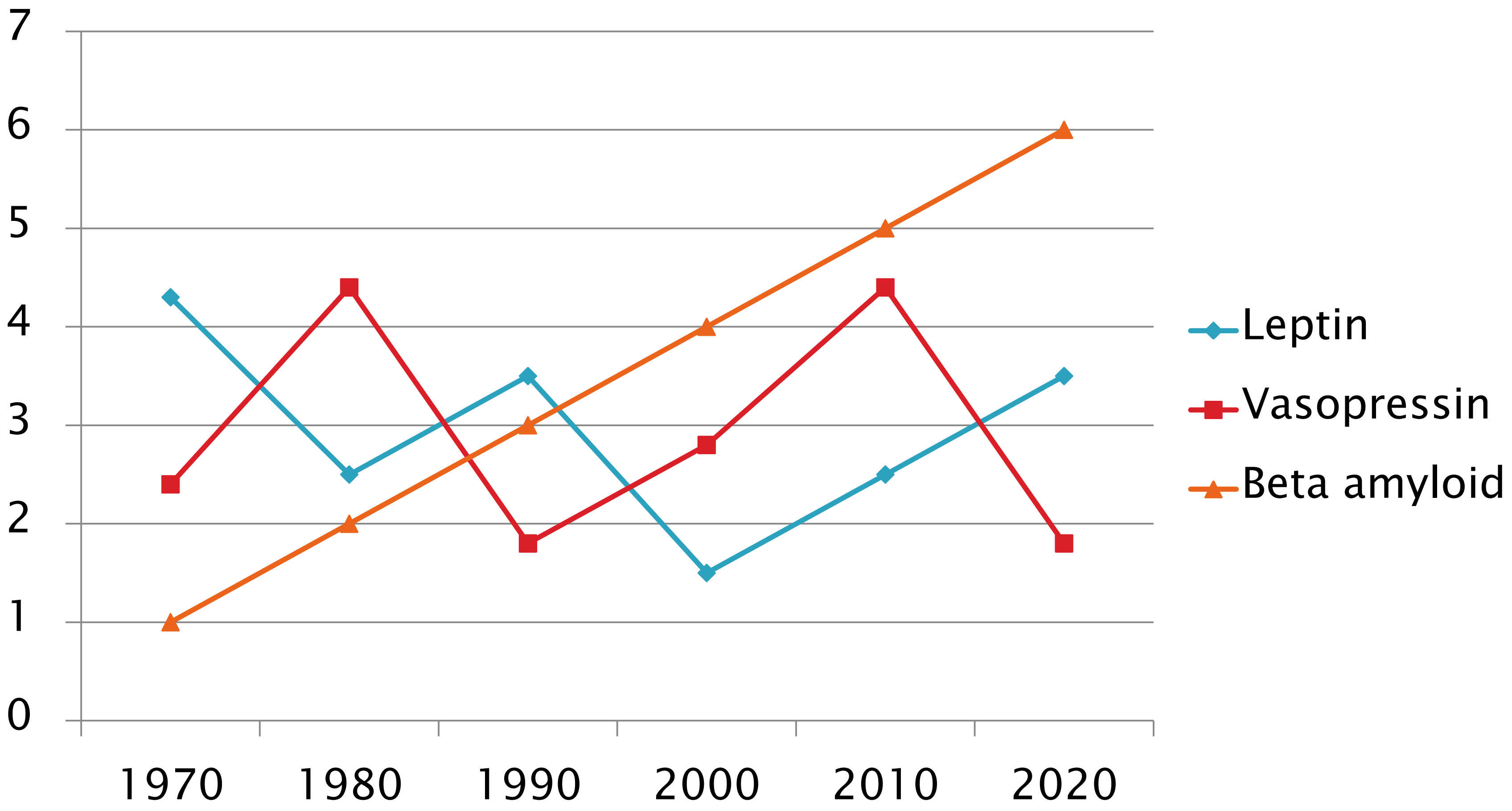
Pie Chart



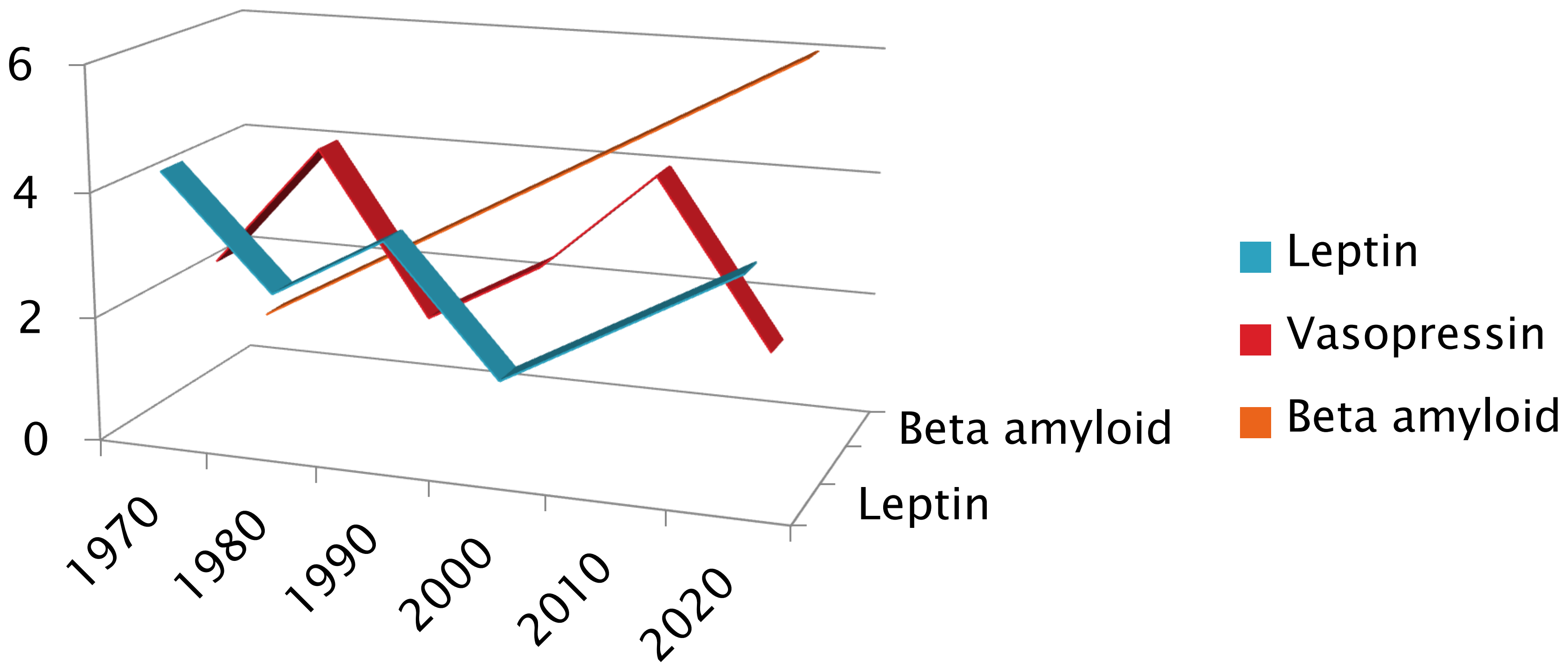
Pie Chart



Line graph


7
6
5
4
3
2
1
0
1970
1980
1990
2000
2010
2020
Leptin
Vasopressin
Beta amyloid

Line graph


6
4
2
0
1970
1980
1990
2000
2010
2020
Beta amyloid
Leptin
Leptin
Vasopressin
Beta amyloid

Line Plot

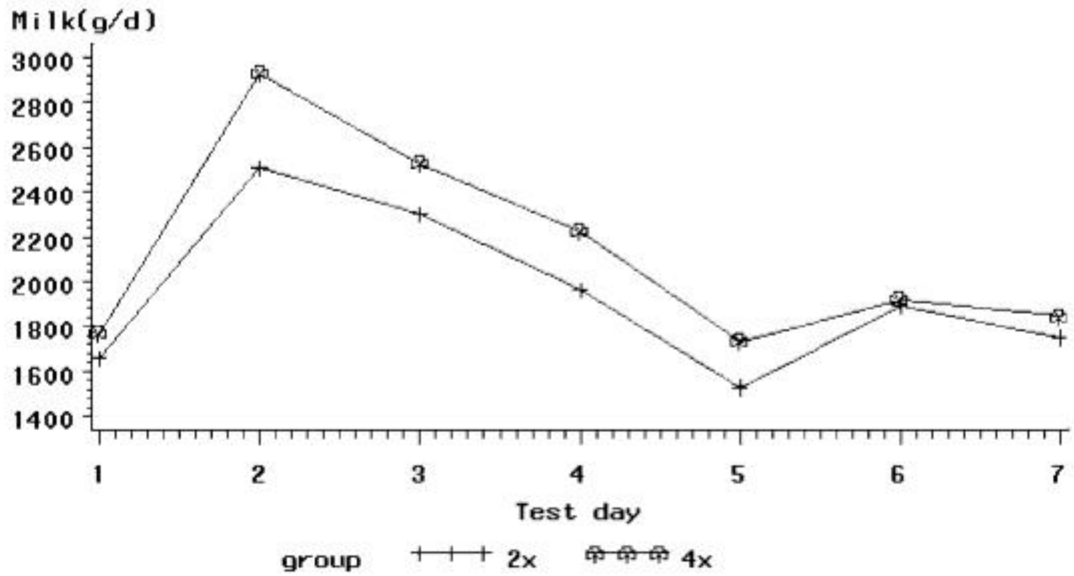


Figure 1 Test day average (least squares mean) milk yield of goats milked four times/d (4×; $n = 14$) for the initial 21 days after weaning the kids or milked twice/day (2×; $n = 14$) for the initial 21 days after weaning the kids. All the goats were milked 2× for the remainder of the lactation. Standard error of the mean ranged from 131 to 134 g/day. The 4× treatment falls into the second interval.

E. Koyuncu and A. Pala. 2008. Effects of Short Period Frequent Milking on Milk Yield and Udder Health in Turkish Saanen Goats. Animal Science Journal 79 (1) 111-115.

Line Plot with standard errors

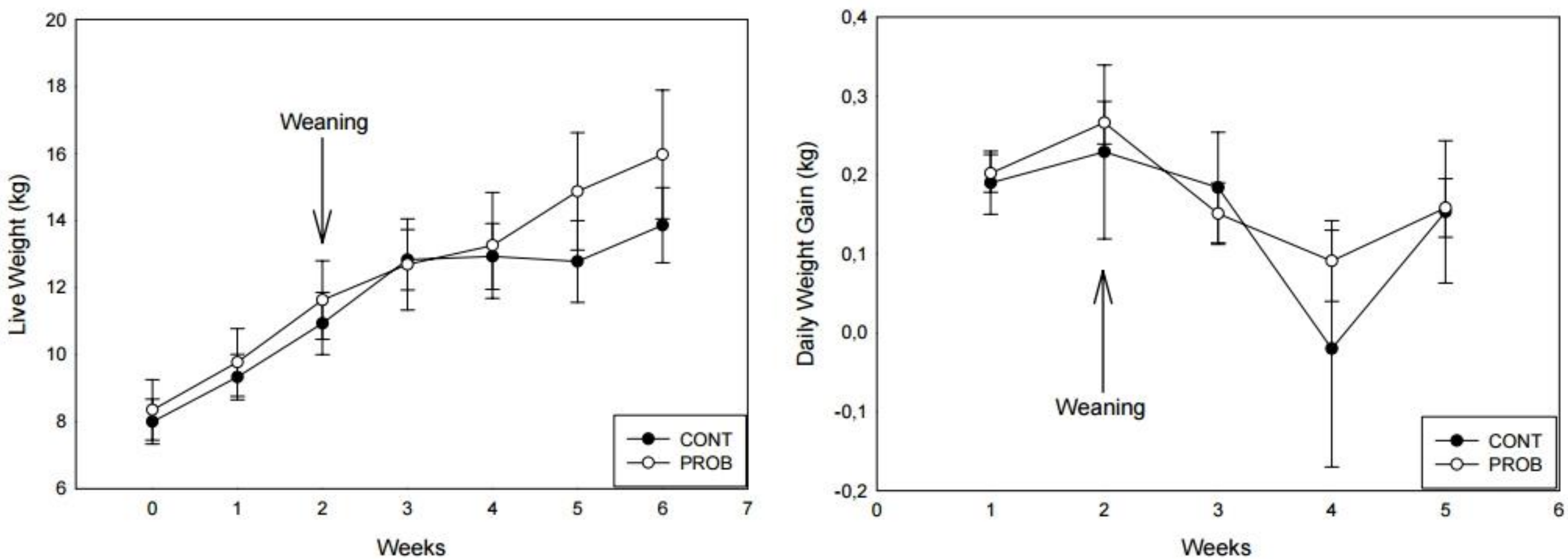


Fig. 1: Average live weights **(a)** and daily weight changes **(b)** with SEDs of kids (Mittelwerte und SEDs für das Lebendgewicht (a) und tägliche die Gewichtszunahme der Lämmer (b))

K. Ayisigi, C. Atasoglu, I. Y. Yurtman, M. Mendes and A. Pala, 2005. Effect of Probiotic Supplementation Shortly Before and After Weaning on Growth of Turkish Saanen Kids. Archiv fur Tierzucht-Archives of Animal Breeding, 48(6) 601-611

Scatter Plot

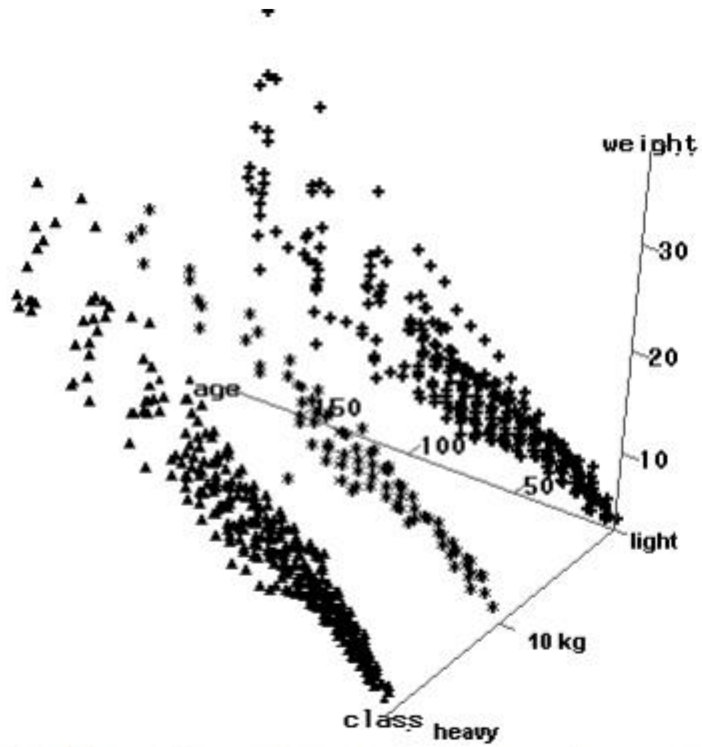
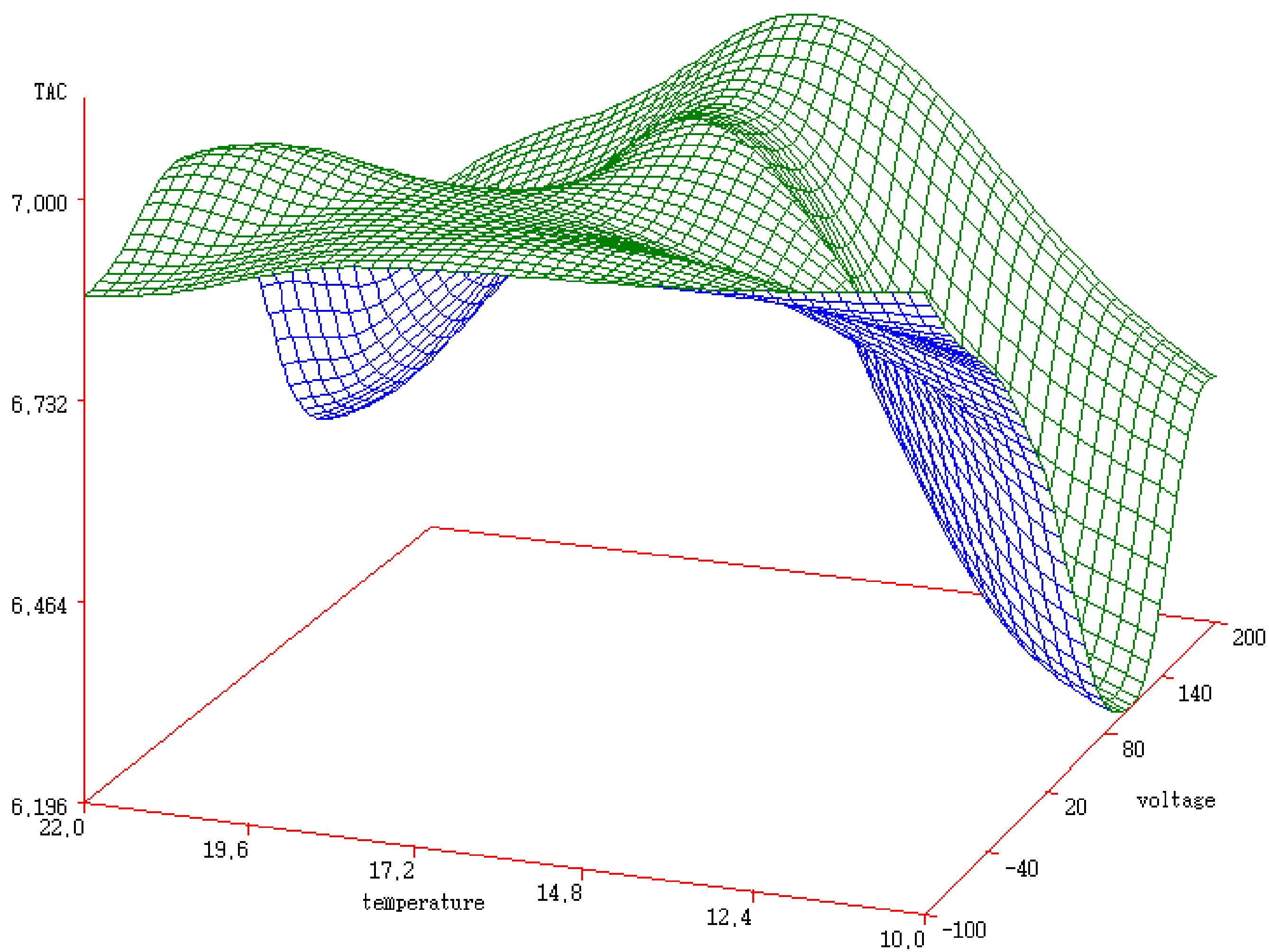


Fig.: Growth curves of animals, heavy, light or 10 kg at weaning (Wachstumskurven der schweren, leichten und der über 10 kg schweren Tiere beim Absetzen)

A. Pala, T. Savas, F. Ugur and G. Das. 2005. Growth Curves of Turkish Saanen Goats' Kids Grouped for Weight and Body Mass Index. Archiv fur Tierzucht-Archives of Animal Breeding 48 (2) 185-193

SAS graph: Silver Nano Particles



Using Tables

Americans' Coffee Consumption

How many cups of coffee, if any, do you drink on an average day?

	Drink one or more cups per day	Average number of cups per day, among drinkers
	%	
Jul 8-12, 2015	64	2.7
Jul 9-12, 2012	64	2.5
Sep 23-26, 1999	63	2.9

GALLUP

http://www.gallup.com/poll/184388/americans-coffee-consumption-steady-few-cut-back.aspx?utm_source=coffee&utm_medium=search&utm_campaign=tiles

Using Tables

Table 1. Total aerobic count (TAC; LSM ± SE) for silver nanoparticle (AgNP) wires created using different accelerating voltages (0, 100, and 200 V)

Milk sample[1]	TAC (log cfu/mL)
Control	$7.08^{A} \pm 0.008$
AgNP0	$7.06^{B} \pm 0.008$
AgNP100	$6.89^{D} \pm 0.008$
AgNP200	$6.95^{C} \pm 0.008$

[A–D]Values with different superscripts differ ($P < 0.05$).

[1]Control = milk sample with no AgNP wire; AgNP0, AgNP100, and AgNP200 = milk samples incubated with AgNP wires coated using 0, 100, and 200 V.

Table 2. Total aerobic count (TAC; LSM ± SE) for different temperatures

Temperature (°C)	TAC (log cfu/mL)
10	$6.86^{C} \pm 0.007$
18	$7.09^{A} \pm 0.007$
22	$7.04^{B} \pm 0.007$

[A–C]Values with different superscripts differ ($P < 0.05$).

N. Yildiz and A. Pala. 2012. Effects Of Small Diameter Silver Nano Particles On Microbial Load In Cow Milk. Journal of Dairy Science. 95(3) 1119-1127

Using Tables

Table 3. Temperature × voltage interactions for total aerobic count (LSM ± SE; log cfu/mL) in control and AgNP100-treated milk samples[1]

Sample[1]	10°C	18°C	22°C
Control	$7.02^{A,c} \pm 0.01$	$7.10^{A,b} \pm 0.01$	$7.14^{A,a} \pm 0.01$
AgNP100	$6.74^{B,b} \pm 0.01$	$7.05^{B,a} \pm 0.01$	$6.86^{B,a} \pm 0.01$

[a–c]Values within a row with different superscripts differ ($P < 0.05$).

[A,B]Values within a column with different superscripts differ ($P < 0.05$).

[1]Control = milk sample with no silver nanoparticle (AgNP) wire; AgNP100 = milk samples incubated with AgNP wires coated using 100 V.

N. Yildiz and *A. Pala. 2012. Effects Of Small Diameter Silver Nano Particles On Microbial Load In Cow Milk. Journal of Dairy Science. 95(3) 1119-1127

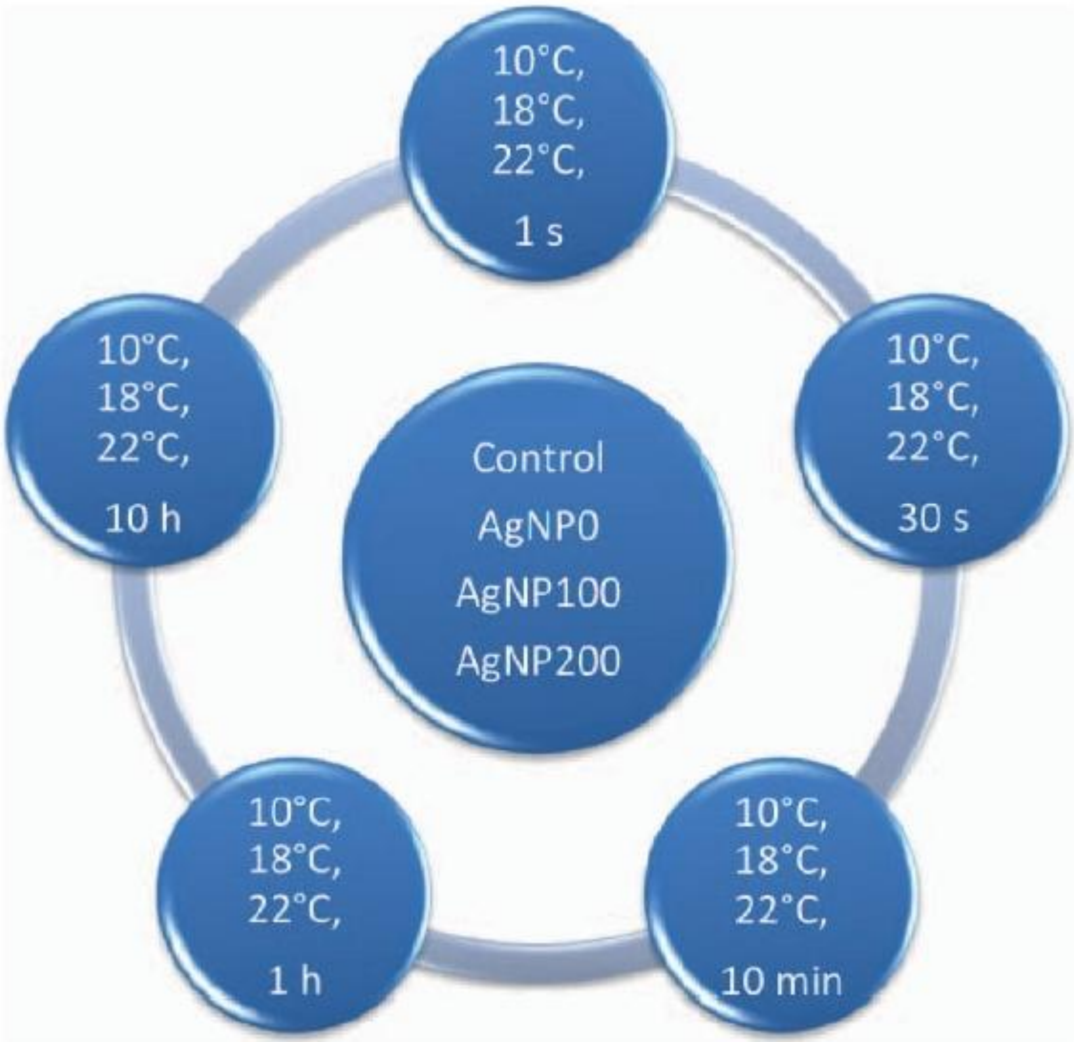


Figure 2. The first trial investigating effects of temperature, time, and accelerating voltage. Control = milk sample with no AgNP wire; AgNP0, AgNP100, and AgNP200 = milk samples incubated with AgNP wires coated using 0, 100, and 200 V. Color version available in the online PDF.

N. Yildiz and *A. Pala. 2012. Effects Of Small Diameter Silver Nano Particles On Microbial Load In Cow Milk. Journal of Dairy Science. 95(3) 1119-1127

Mathematical representation

- Measures of Central Tendency for Ungrouped Data
- Measures of Dispersion for Ungrouped Data
- Mean, Variance, and Standard Deviation for Grouped Data
- Use of Standard Deviation

The Mean

- The mean, is the most frequently used measure of central tendency.
- Mean =average

$$SampleMean = \frac{Sum\_of\_all\_observations}{\#of\_observations} = \overline{X} = \frac{\sum x}{n}$$

$$Population\_Mean = \frac{Sum\_of\_all\_values}{\#of\_values} = \mu = \frac{\sum x}{N}$$

- The number of values in a data set is denoted by n for a sample and by N for a population.

The Mean

- The value of the population mean is constant.
- The value of the sample mean varies from sample to sample.
- Sometimes a data set may contain a few very small or a few very large values.
  - such values are called *outliers or extreme values.*
  - The mean as a measure of central tendency is that it is very sensitive to outliers.

The Median

- The mean is susceptible to the influence of outliers.
- The wages of faculty at a university:

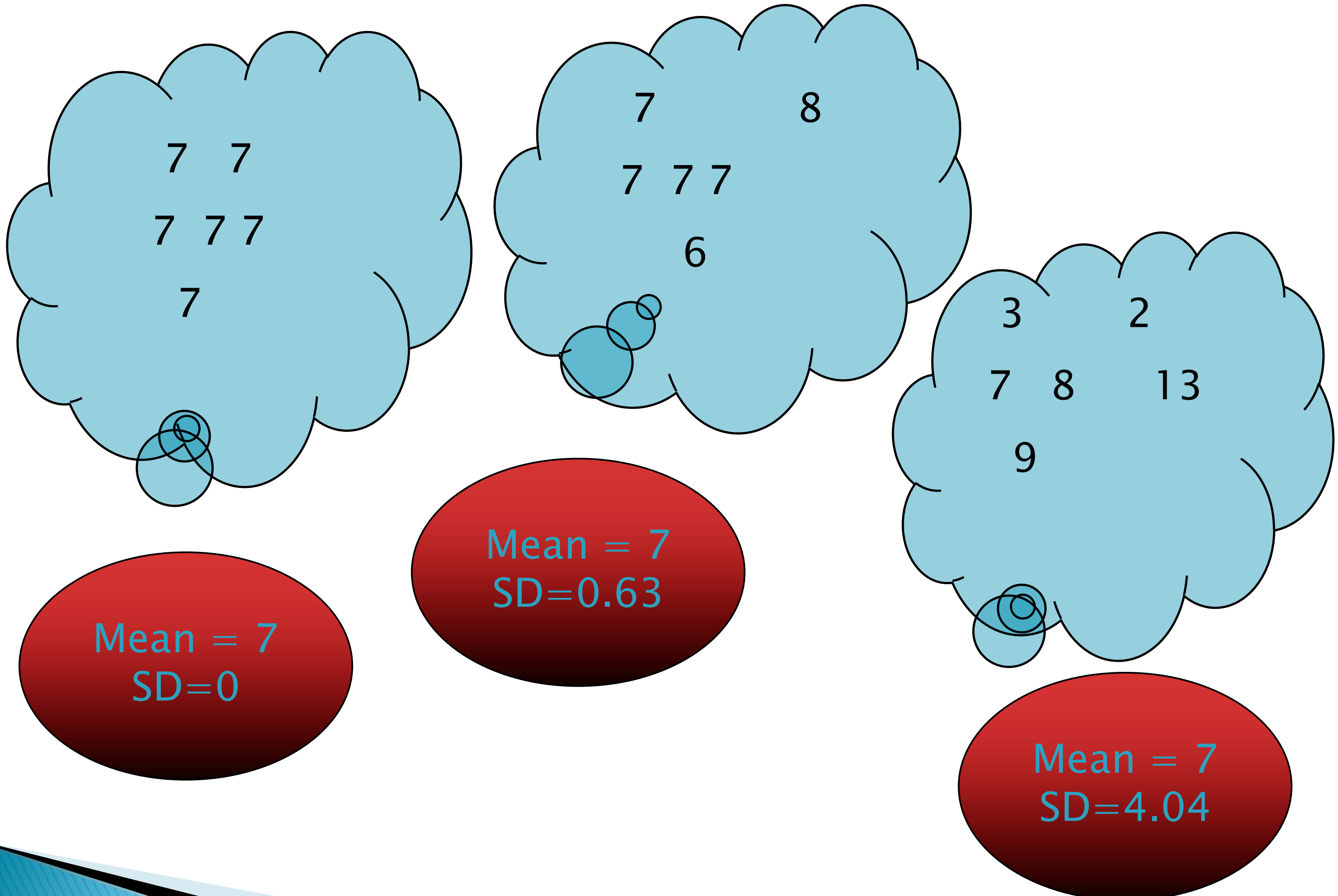
Faculty	1	2	3	4	5	6	7	8	9	10
Salary	30k	32k	35k	40k	45k	55k	60k	65k	190k	195k

- The mean salary for these ten staff is 74.7k.
- This mean value is not the best way to accurately reflect the typical salary of faculty
  - Their salaries are in the 30k to 65k range.
  - The mean is being skewed by the two large salaries.
  - The median would be a better measure of central tendency in this situation.

Summary Statistics, median

- Put the numbers you are given in value order and find the middle number.
- Example: find the Median of
- 30k
- 32k
- 35k
- 40k
- 45k
- 55k
- 60k
- 65k
- 190k
- 195k
- There are two middle numbers, 45 and 55, so you average them: (45+55)/2=50
- X=Median (2, 4, 5, 6, 15)=5

Standard Deviation



Standard Error

- Variance among means of samples selected from certain population

Calculation of Variance and stddev

- Question
- What are the variance and standard deviation of the following data set? Calculate
- this using both the population and sample formulas.
- 1 3 5
- The population formula to calculate variance is:

$$\sigma^2 = \frac{\sum(X_i - \overline{X})^2}{n}$$

- The sample formula is: $\sigma^2 = \frac{\sum(X_i - \overline{X})^2}{n-1}$

Calculation of Variance and stddev

- n = 3, x = 3, and the sum of the squared deviation scores = $(-2)^2 + 0^2 + 2^2 = 8$.
- The population variance= 8/3 = 2.67
  - The population standard deviation is the square root of the variance = 1.63
- The sample variance is 8/2 or 4
  - the sample standard deviation is the square root of the variance = 2.

Descriptive Statistics using SAS

- Data set Blood Pressure.
- The variables Subject (ID value for each subject)
- Drug (with values of Placebo, Drug A, or Drug B),
- SBP (systolic blood pressure)
  - The higher of the two numbers, measures the pressure in the arteries when the heart beats (when the heart muscle contracts).
- DBP (diastolic blood pressure)
  - The lower of the two numbers, measures the pressure in the arteries between heartbeats (when the heart muscle is resting between beats and refilling with blood).

SAS program

```
data palablood;
input subject gender $ drug $ sbp dbp;
cards;
1 male placebo 120 90
2 male placebo 125 93
3 female placebo 118 88
4 female placebo 115 85
5 male drugA 122 97
6 male drugA 123 100
7 female drugA 110 80
8 female drugA 105 86
9 male drugB 100 80
10 male drugB 105 82
11 female drugB 101 77
12 female drugB 102 75
;
title 'Lets see the data';
Footnote 'Prof. Dr. Akin Pala';
proc print;
title 'Normal Proc Means';
proc means;
title 'Proc means for median etc.';
proc means mean median std;
title 'Proc frequency';
proc freq;
*This is a comment line, you can use it and SAS will not read it, your friends might :);
run;
```

SAS program

```
data palablood;
input subject gender $ drug $ sbp dbp;
cards;
1 male placebo 120 90
2 male placebo 125 93
3 female placebo 118 88
4 female placebo 115 85
5 male drugA 122 97
6 male drugA 123 100
7 female drugA 110 80
8 female drugA 105 86
9 male drugB 100 80
10 male drugB 105 82
11 female drugB 101 77
12 female drugB 102 75
;
title 'Making simple graphics';
Footnote 'Prof. Dr. Akin Pala';
proc plot;
plot dbp*sbp ;
run;
```

Other SAS programs

- See https://odamid.oda.sas.com/SASStudio for my course programs:
  - 3d graph smooth silver nano.sas

Computing the Sample Size

- Calculate the number of subjects per group for several scenarios.
- Means
  - 20 versus 30
  - 22 versus 28, are hypothesized (two scenarios).
- Sample sizes for grouped standard deviations of 10 and 15.
- Sample sizes for powers of 80% and 90%.
- A plot of power versus the number of subjects per group (8 different scenarios).

Using SAS for sample size calculations

- title "Sample Size Requirements";
- proc power;
  - twosamplemeans /*to say that this is a t test*/
  - groupmeans = (20 30) (22 28)
  - stddev = 10 15
  - power= .80 .90
  - npergroup = .; *You specify the parameter you want to compute by entering a missing value for that parameter;
  - plot x = power min = .70 max = .90;
- run;

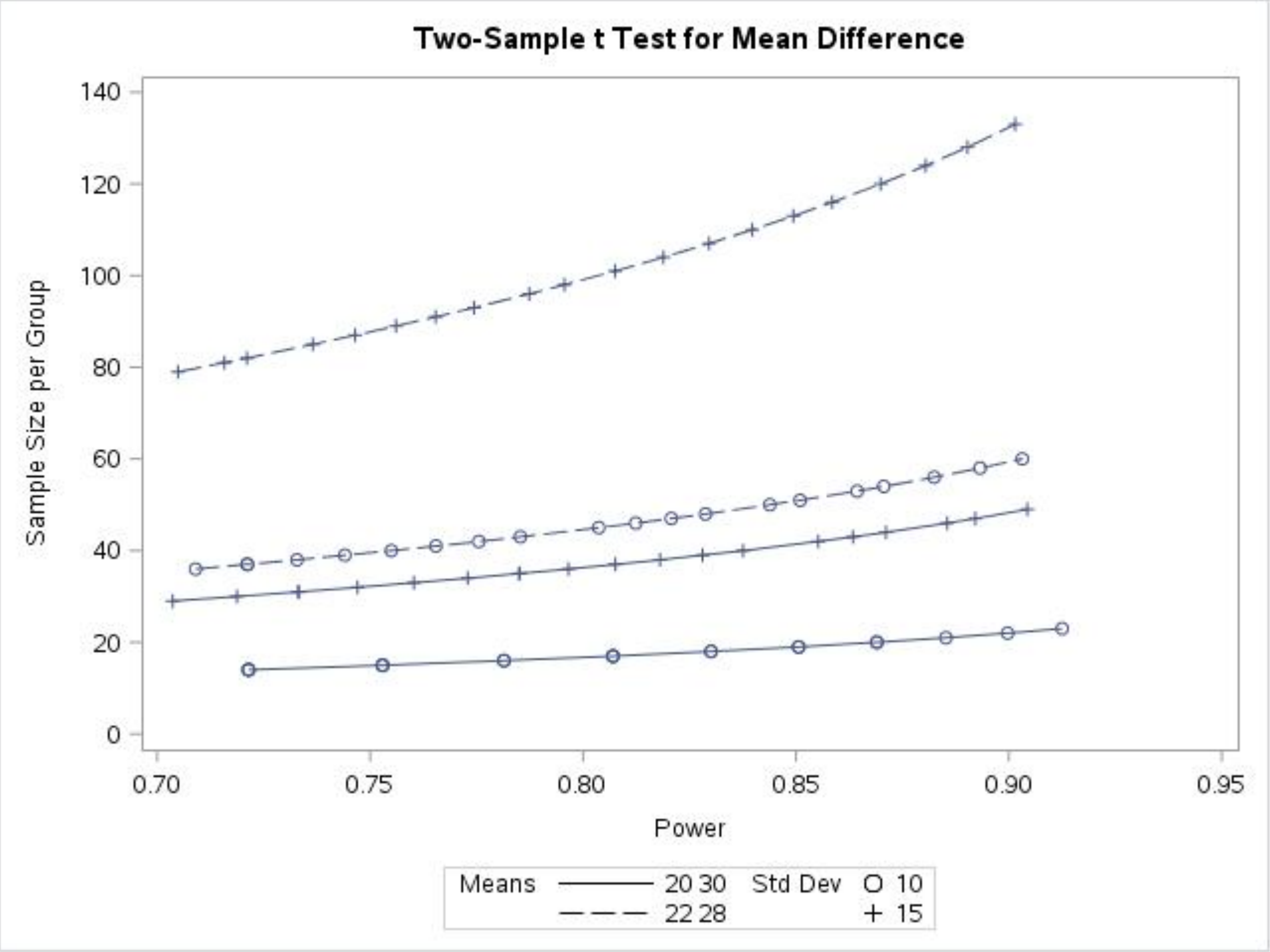
N per Group Output

Computed N per Group						
Index	Mean1	Mean2	Std Dev	Nominal Power	Actual Power	N per Group
1	20	30	10	0.8	0.807	17
2	20	30	10	0.9	0.912	23
3	20	30	15	0.8	0.808	37
4	20	30	15	0.9	0.904	49
5	22	28	10	0.8	0.804	45
6	22	28	10	0.9	0.903	60
7	22	28	15	0.8	0.804	100
8	22	28	15	0.9	0.901	133

Prof. Dr. Akin Pala

Sample Size Requirements for a T-Test

The POWER Procedure
Two-Sample t Test for Mean Difference



Prof. Dr. Akin Pala

Computing the Power

- title "Computing the Power of a T-Test";
- proc power;
  - twosamplemeans
  - groupmeans = 20 | 30 35
  - stddev = 10 15
  - power= .
  - npergroup = 30 35;
  - plot x = n min = 20 max = 50;
- run;
- *The POWER value is set to missing, so the procedure will compute power for the given sample sizes.
- *The vertical bar indicates a crossing—every value on one side of the bar versus every value on the other side of the bar;

Homework

- Use Turnitin and SAS
- Construct a bar graph for the frequency distribution of any data set you can find/construct using appropriate real life data.
- Construct a line graph using means
- Show your work by providing the data set, calculation of the means etc.