Statistical Sampling

Live Session Unit 2

HWs & Labs

HWs

- Get more than 110/130
- 10%

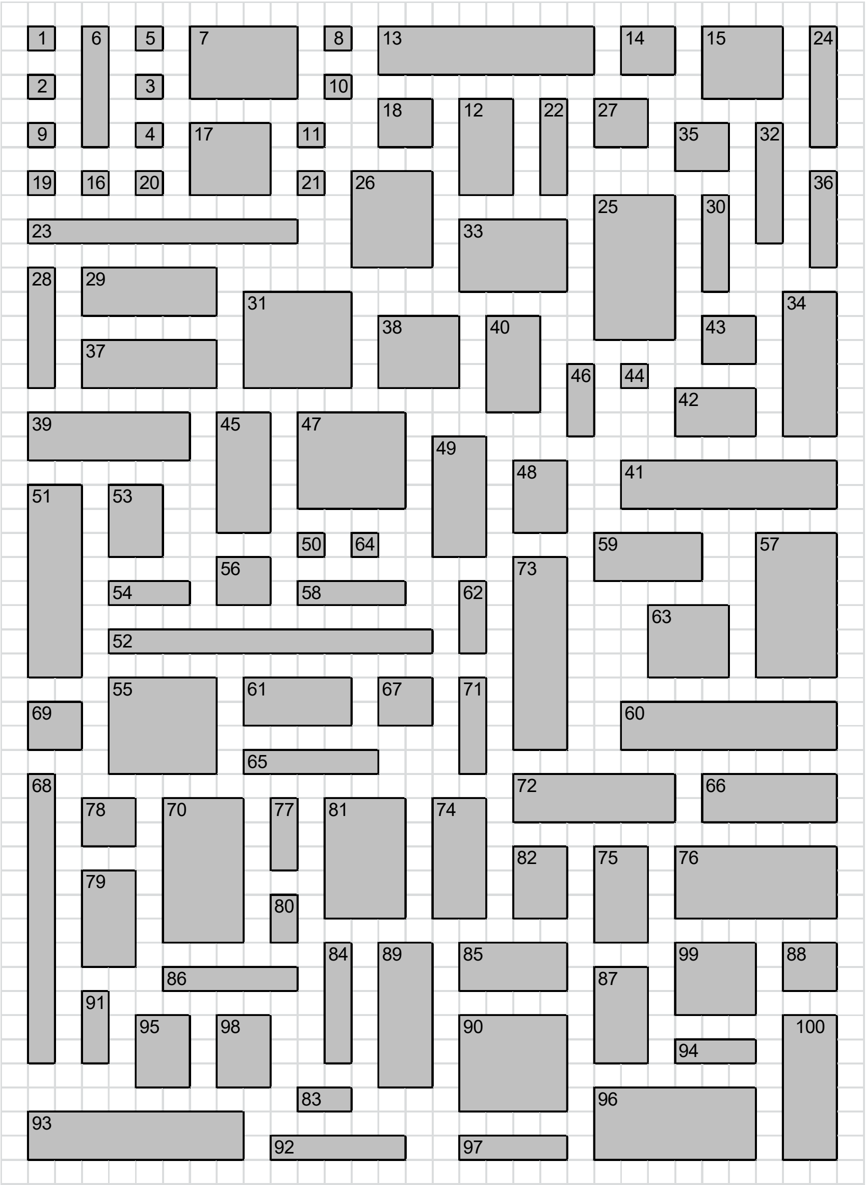
Labs

- Get more than 110/130
- 20%

Future Plan

- May 16 : Live session 02- Checking Randomness
- May 23 : Live session 03- Sampling Distribution
- May 30 : Live session 04- Stratified Design
- Jun 06 : Live session 05- Sample size calculations
- Jun 13 : Live session 06 – Mid term review
- Jun 20 : Mid term

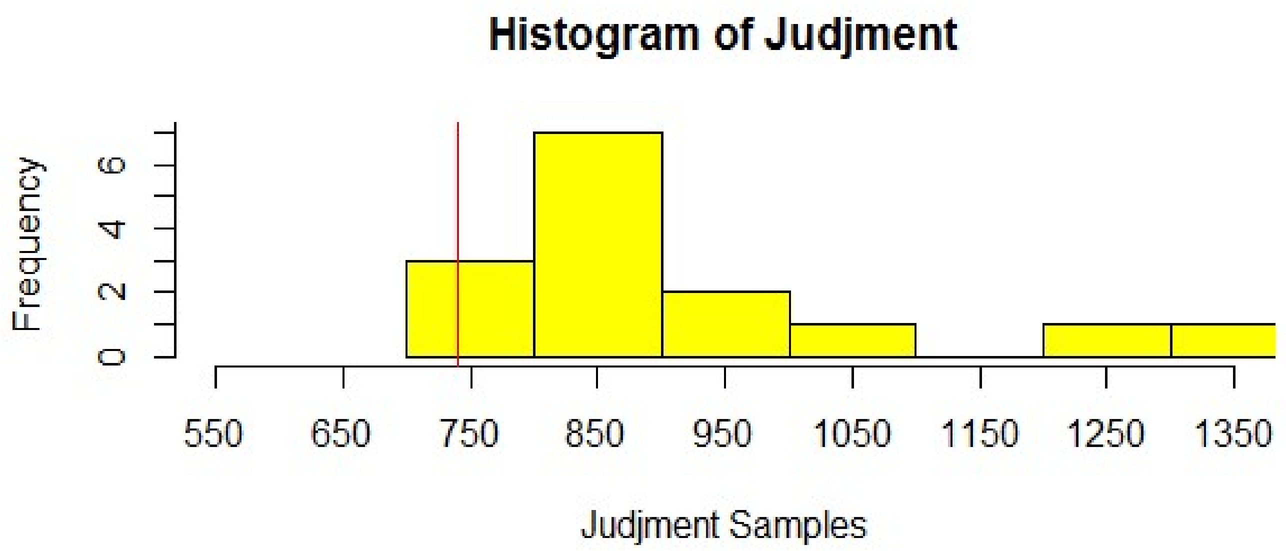
Lab 1 discussion

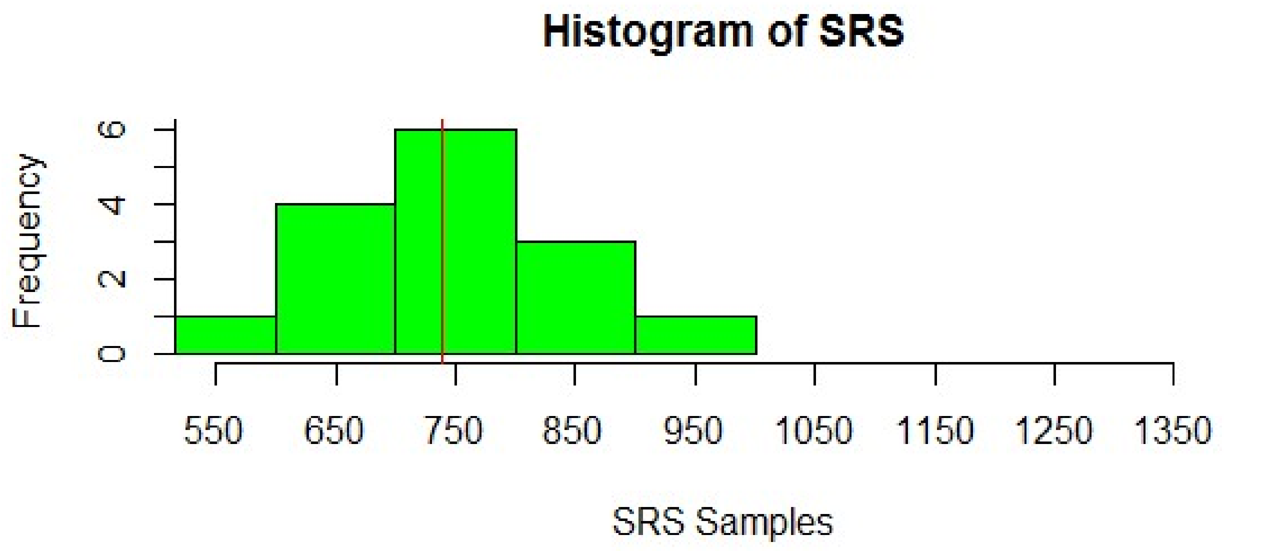

1
6
5
7
8
13
14
15
24
2
3
10
18
12
22
27
9
4
17
11
35
32
19
16
20
21
26
36
25
30
23
33
28
29
31
34
38
40
43
37
46
44
42
39
45
47
49
48
41
51
53
50
64
59
57
56
73
54
58
62
63
52
55
61
67
71
69
60
65
68
72
66
78
70
77
81
74
82
75
76
79
80
84
89
85
99
88
86
87
91
95
98
90
100
94
83
96
93
92
97

Data

	judgment samples	simple random samples
1	920	750
2	940	780
3	850	590
4	790	780
5	890	660
6	780	870
7	900	830
8	890	980
9	1270	720
10	810	650
11	1040	620
12	790	700
13	890	770
14	850	850
15	1320	780

	Judgment samples	SRS
Mean total area	928.7	755.3
Minimum total area	780	590
Maximum total area	1320	980


Histogram of Judjment
Frequency
0
2
4
6
550
650
750
850
950
1050
1150
1250
1350
Judjment Samples


Histogram of SRS
Frequency
0
2
4
6
550
650
750
850
950
1050
1150
1250
1350
SRS Samples

6. The 15 values of estimated total area the class obtained are not the only possible values of the total area that could be obtained from a SRS. What is the minimum possible value of total area that could be obtained from a SRS of size 10 from the rectangle population? ______ . What is the maximum possible value of total area that could be obtained from a SRS of size 10 from the rectangle population? ______

The minimum estimate of total area from a sample of size 10 is 100. Such a sample would have 10 of the 15 rectangles with area 1.

The maximum estimate of total area from a sample of size 10 is 1700. Such a sample would have all 5 rectangles with area 18 and 5 of the 10 rectangles with area 16.

7. Is it possible that estimated total area from a SRS could be worse than your judgment sample estimated total area? Explain.

7. Is it possible that estimated total area from a SRS could be worse than your judgment sample estimated total area? Explain.

- *Technically yes. Because it is random, you have no control over which plots get chosen. It could randomly select several small plots several times, giving you a smaller distribution, which may yield poorer total areas compared to a judgment sample.*

- Yes it could be possible in some cases as the random number generation is by chance, and sometimes all the rectangles selected can be very small or very large, and hence the area can be skewed accordingly. Although, the probability of this happening is less.

8. Is it likely that estimated total area from a SRS could be worse than your judgment sample estimated total area? Explain.

No, it is not likely that an estimated total area form a SRS is further away from the true area than an estimated total from a judgment sample. However, such an event is possible.

8. Is it likely that estimated total area from a SRS could be worse than your judgment sample estimated total area? Explain.

- No, we use probability sampling because it consistently does better than judgement based samples as there is no bias in this form of sampling.
- Unlikely. Human perception does not lend itself to accurate judgment; several selection and bias issues could be present.

An important definition

- A *probability sample* is one in which the probability of selection for every member of the sample is non-zero and known.
  - When you used a random number table to select a sample, this was a probability sample
  - When n = 10 and N = 100, the probability of selection is 10/100 for each "farm"

- A non-probability sample is one in which this condition is NOT true
  - When you used your judgement to select a representative sample, this was a non-probability sample

- There are many ways to select each type of sample

What are samples for?

- Definitions:
  - Population = all the units that you are interested in learning about
  - Sample = a subset of the population
  - Parameter = A numerical characteristic of a population (e.g., mean, total, proportion, difference in means between two parts)
  - Statistic = A numerical characteristic of a sample (e.g., sample mean, sample proportion, or other numerical summary)

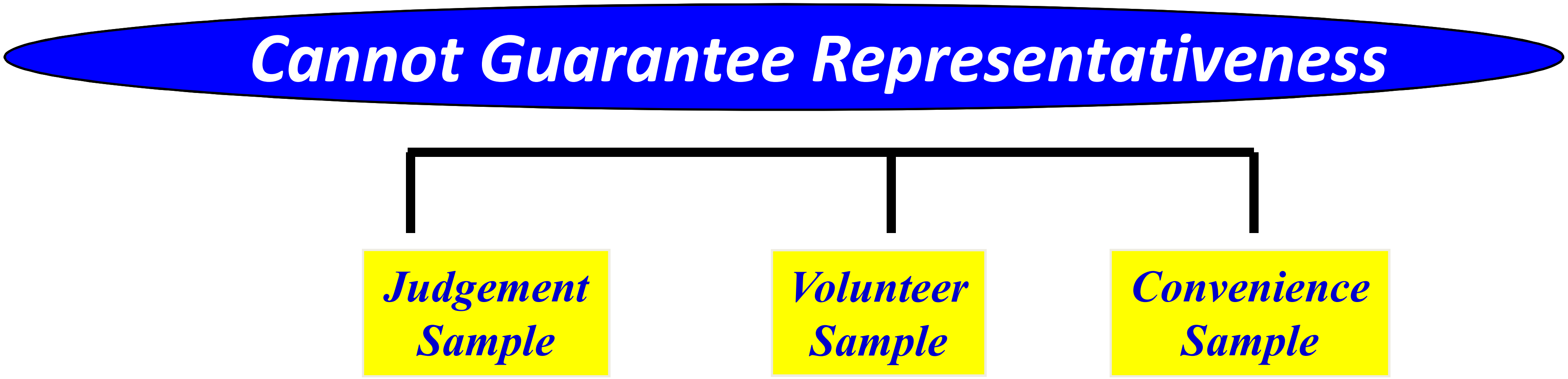
- Goal of a sampling process is to estimate a parameter of a population from a sample statistic
- Thus the sample must be “representative” of the population

How do we assure representativeness?

- In the rectangle exercise, you tried to assure representativeness using human judgement.
  - Quota sampling
- People are not very good at this
- Random selection is a dependable method of assuring representativeness. Its advantage is that…
  - It has a high chance of getting a sample that is close to representative
  - We can compute the probability that it is will NOT be representative.
  - More specifically, we can use the mathematics of probability to compute the probability that the estimate is within a certain distance of the parameter.
- This is NOT true of a nonprobability sample

Non-Probability Samples

- Easy to Obtain, Usually Voluntary Responses
- Self-Selection is a Serious Problem
- Can Contain Useful Information



Cannot Guarantee Representativeness

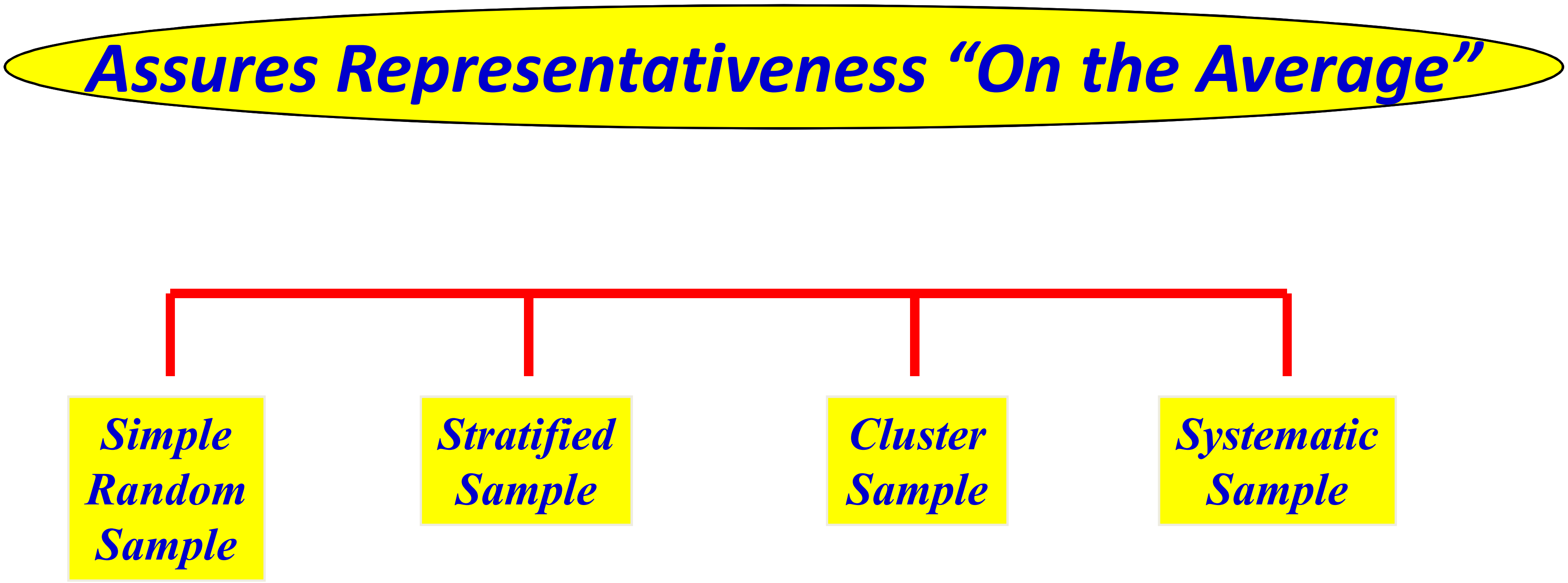
Judgement Sample

Volunteer Sample

Convenience Sample

Probability Samples

Probability Samples



Assures Representativeness "On the Average"

Simple Random Sample

Stratified Sample

Cluster Sample

Systematic Sample

Probability sample designs

- Many ways to select a probability sample
- Simplest way is the method used for sampling farms, called a simple random sample without replacement (SRSWOR or more compactly, SRS)
- The probability of selection for EVERY farm is the same, 10/100 = 1/10.
  - Consider the following: I select one farm at random from those numbered 1 to 20 and one at random from those numbered 21 to 100.
    - Is this a probability sample?
    - YES, because I can compute the probability of selection of every sampled farm
  - Big idea: random does NOT mean that every unit MUST have the same probability of selection!

EPSEM

- Sample designs using an **E**qual **P**robability of **SE**lection **M**ethod are sometimes called EPSEM designs.

  (eg: SRS)

- For EPSEM designs, the usual sample mean is an unbiased estimator of population mean.

HW2

- *A large school district plans to survey parents to measure their opinions on various issues. The survey will be done as follows. A random sample of 20 elementary schools, 10 middle schools and 5 high schools will be drawn. Within selected schools, classes will be randomly drawn: 5 rooms per elementary school, 10 homerooms per middle school and 20 homerooms per high school. A self-administered questionnaire and envelope will be placed in the "home communication" folder of each student in selected classes. Parents will be asked to complete the questionnaire, seal it in the envelope, and send it back to school with their child. Is this an EPSEM sampling procedure? What is the probability of any given student being selected?*

HW2

- $For\ Elementary\ School\ Students\text{:}\ P(Student_i) = \frac{20}{N_{ES}} * \frac{5}{N_{rooms}}$
- $For\ Middle\ School\ Students\text{:}\ P(Student_i) = \frac{10}{N_{MS}} * \frac{10}{N_{rooms}}$
- $For\ High\ School\ Students\text{:}\ P(Student_i) = \frac{5}{N_{HS}} * \frac{20}{N_{rooms}}$

- Given the above equations unless the number of schools of each type times the number of classrooms for the respective school are the same across the board the probability of any one student being selected will be different and the model will not be EPSEM.

Complex Sample designs

- Designs that have unequal probabilities of selection (an example of a *complex design*) must use an estimator of the mean that specifically takes the design into account.
- So before you can analyze data from a sample design, you must have specific types of information about it.
- This is why specialized software is needed for sample data analysis.

How can we use the data from a PPS?

- The Probability Proportion to Size (PPS) is not representative of the population as it is
  - Too many large farms were selected

- We need to downweight units that tend to occur in the sample too frequently (the large ones) and upweight those tend to occur too infrequently (the small ones)

- Thought question: Suppose farms #1 (area of 1) and #6 (area of 5) occurred in your sample. How should you weight them?
  - Each farm of area 5 is 5 times as likely to be chosen on each draw as a farm of area 1. So they will be represented 5 times as often.
  - To make sample representation fair, we need to weight farm 1,

    5 times as much as farm 6 in the estimator of mean

General rule for weighting

- Big idea: If you weight each unit by the reciprocal of its selection probability, and calculate a weighted average, you will always produce an approximately unbiased estimator of the mean.

- The estimator of the mean is:

$$\hat{\mu} = \sum_{i=1}^{n} w_i y_i \Big/ \sum_{i=1}^{n} w_i$$

where $w_i = 1/\pi_i$ and π_i is the probability of selection of the ith sample unit.

This course

- We will spend most of the course on probability samples, but will return to nonprobability samples toward the end of the term.
- They are becoming more important to survey researchers are response rates decrease, causing cost of probability sampling to increase
- They are evaluated by how well they can be made to behave like probability samples

Big idea!

- Two sample designs can be compared by comparing the behavior of the estimators that produce, over many realizations of the sample.
- Two important features of the sampling distribution: its mean and variance.
  - Small (or no!) bias is good.
  - Small variance is good.
- When comparing two designs producing unbiased estimators, the better one is the one with smaller variance
- When comparing two designs producing estimators that are not unbiased, the better one is the one with the smaller mean squared error (MSE), defined as:

$$\text{MSE} = \text{Variance of estimator} + \text{Bias}^2.$$

Important!

- A *probability sample* is one in which the probability of selection for every member of the sample can be calculated.
- A probability sample does not need to have the same probability of selection for every unit; however, if it does not, the analysis method must take into account the varying probabilities.
- A common sample design is one in which units are selected with probabilities proportional to their size.
- Two different sample designs can be compared by comparing the variance or mean squared error of the estimators that they produce. The smaller the variance is, the more efficient is the design.

Randomness

- A pattern is said to be random if it is unpredictable in the short run, but predictable in the long run
- You have learned that random assignment of subjects to treatment groups improves the strength of the inferences you can make from experiments; i.e., allows an inference of *causality*.
- Random selection into a sample is to ensure (at least in the long run) representativeness; note that it does **not** allow causal inference
- The use of randomness in experiments is to increase *internal validity* ("If I applied this treatment to another group like this one, would it cause the same result?")
- The use of randomness in sample selection is to ensure *external validity* ("If I could see the whole population, would its features look like this?")

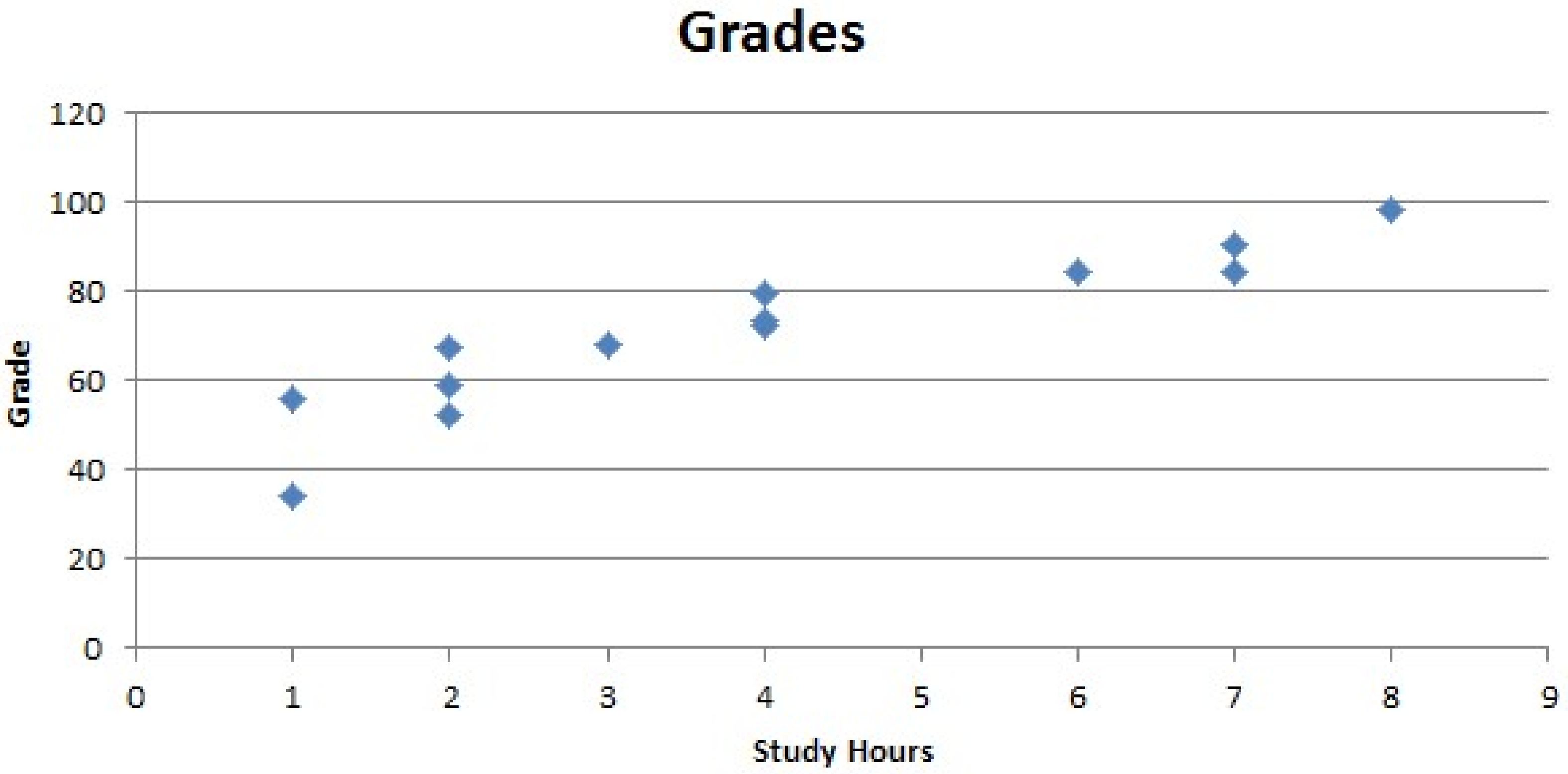
Checking data for randomness

- No formula for checks that guarantees data are random
- A few basic checks for randomness
  - Construct a scatter plot and calculate correlation
  - Examine distributional properties
    - Plot the means within the categories to see if they are approximately equal (assuming they should be).
    - Plot the medians within the categories of interest to see if they are approximately equal (assuming they should be)
      - Keep in mind that a median is robust to unusually large or unusually small observations while mean is not

Key Concept

We would like a numerical measurement of the strength of the relationship between two variables representing quantitative data.

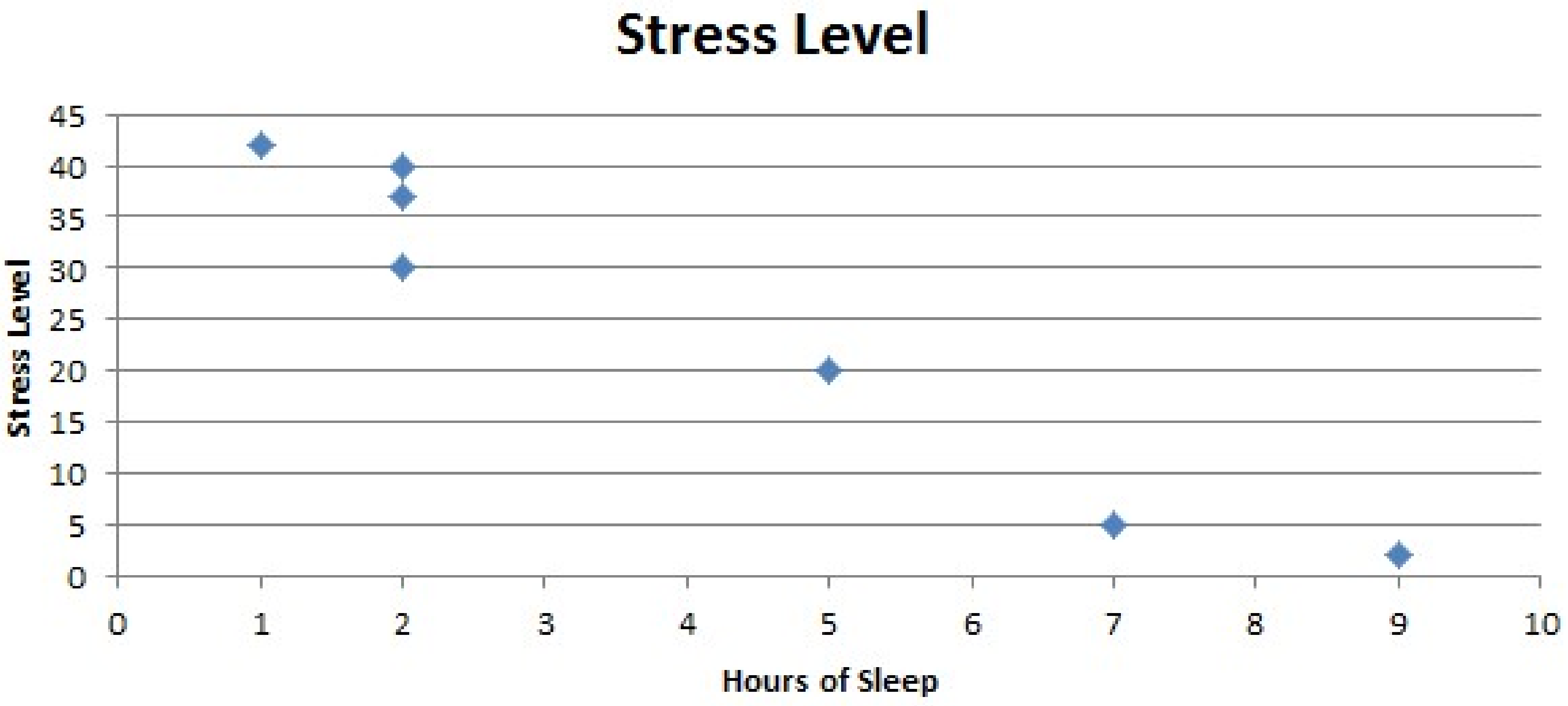
Studying Hours and Grades!



Key Concept

We would like a numerical measurement of the strength of the relationship between two variables representing quantitative data.

Sleep v. Stress Level.



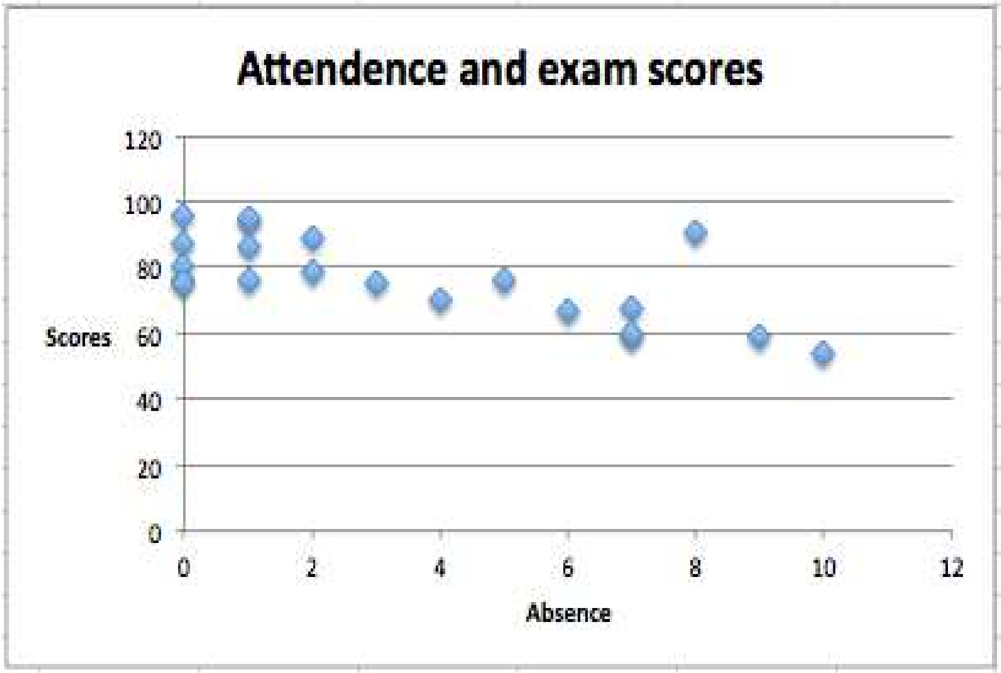
Exploring the Data

We can often see a relationship between two variables by constructing a SCATTERPLOT.

Scatterplot

- A scatterplot displays the relationship between two **quantitative variables**.

- ***Example*** : class attendance and final exam scores.

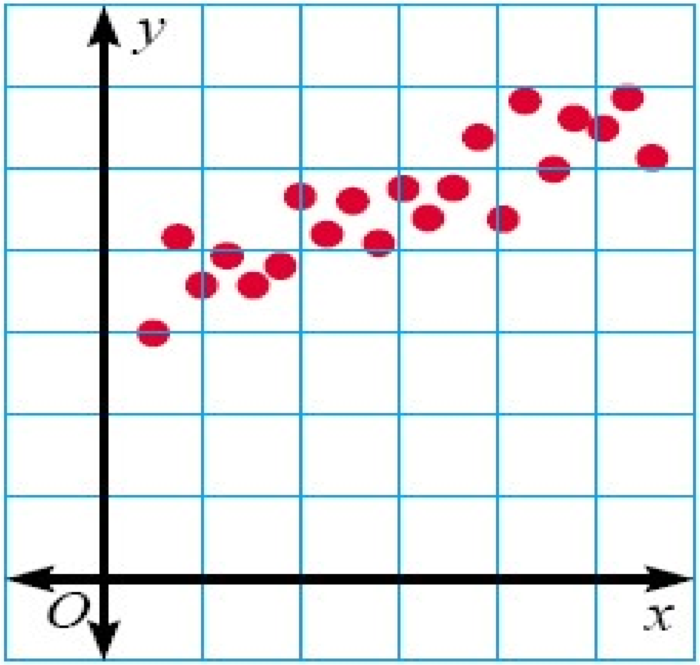


Scatterplots

- Form – overall pattern
- Strength – weak or strong relationship
- Direction – positive, negative or no association
- Outliers – any observations that deviate strongly from the overall pattern

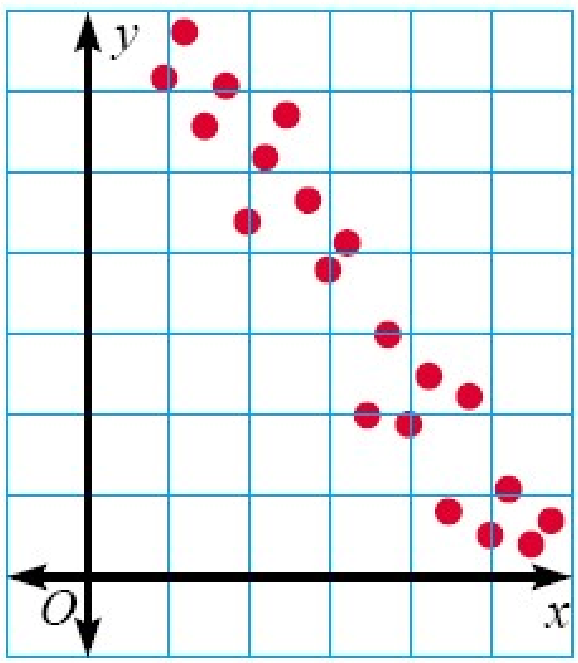
Positive Direction

- If the x-coordinates and the y-coordinates both increase, then it is POSITIVE CORRELATION.
- This means that both are going up, and they are related.



Negative Direction

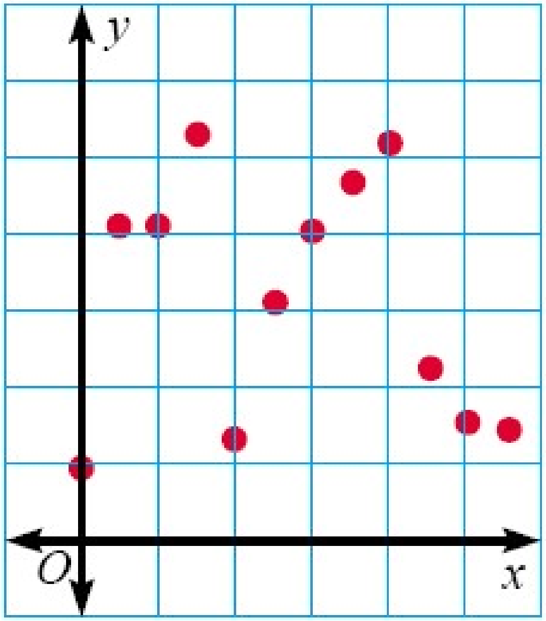
- If the x-coordinates and the y-coordinates have one increasing and one decreasing, then it is NEGATIVE CORRELATION.
- This means that one is going up and one is going down, making a downhill graph. This means the two are related as opposites.



No Linear Correlation

- If there seems to be no pattern, and the points looked scattered, then it is no **linear** correlation.

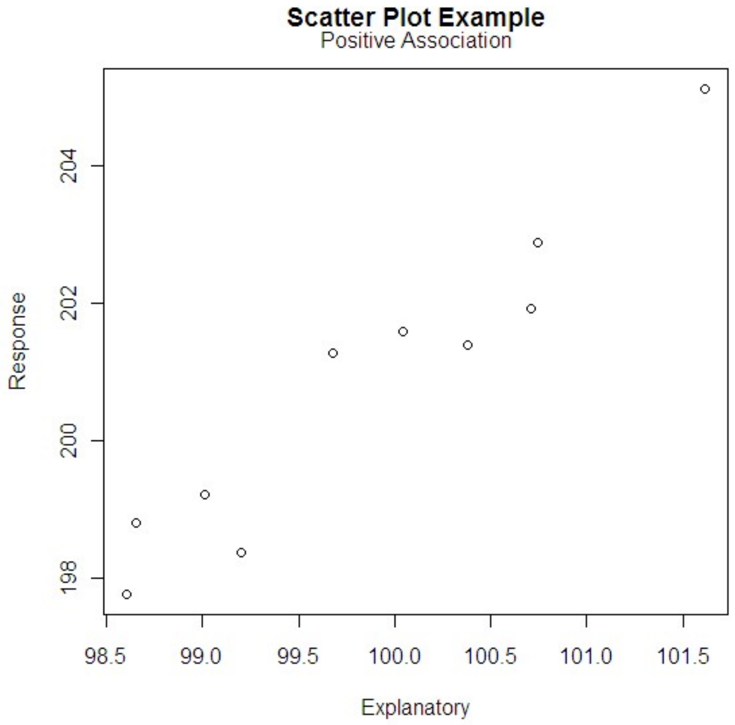
- This means the two are not linearly related.

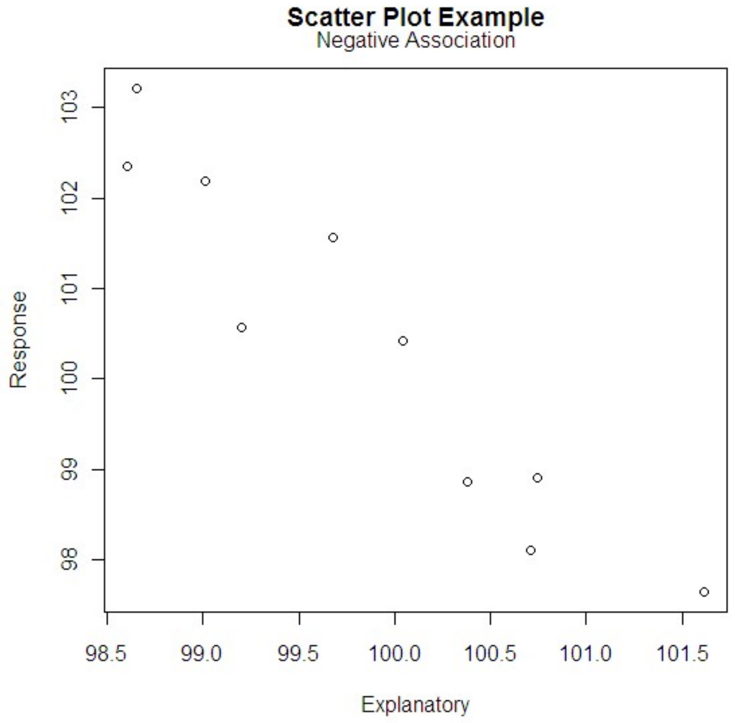


Scatterplot

The best way to graphically examine the relationship between two variables is a scatter plot.

Examples





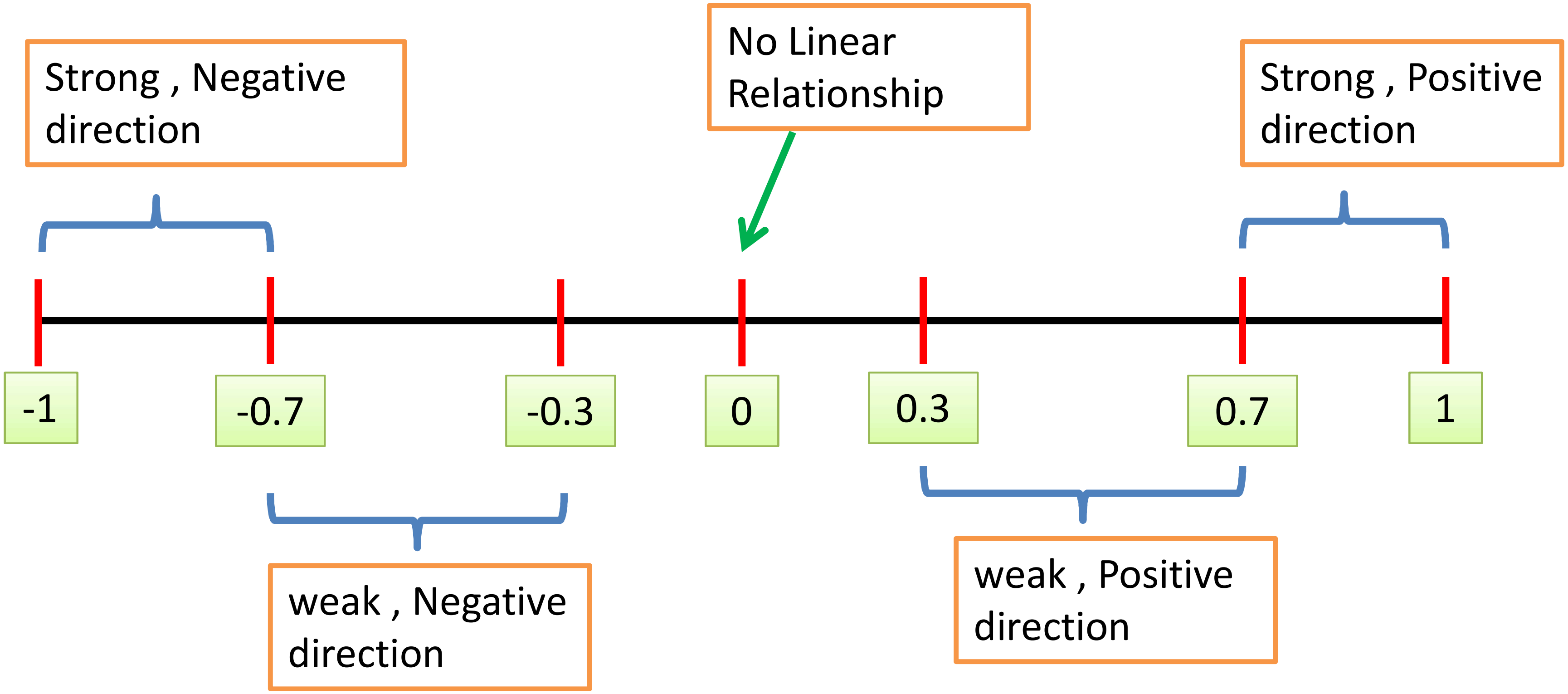
Strength of Linear Relationship

Correlation – measures the direction and strength of the ***linear*** relationship between two ***quantitative*** variables. Usually denoted "r". Has no units.

- $-1 \leq r \leq 1$
- r = 1 is a perfect positive linear relationship
- r = -1 is a perfect negative linear relationship
- r = 0 no linear relationship

Strength of Linear Relationship

- $-1 \leq r \leq 1$



Strong , Negative direction

No Linear Relationship

Strong , Positive direction

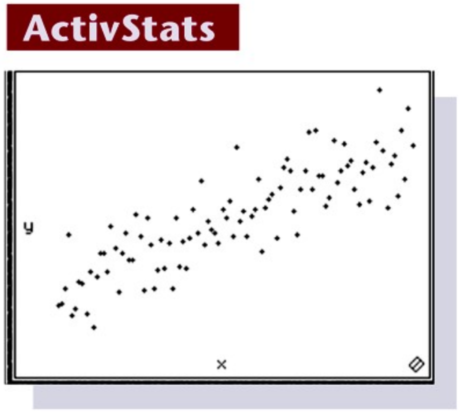
weak , Negative direction

weak , Positive direction

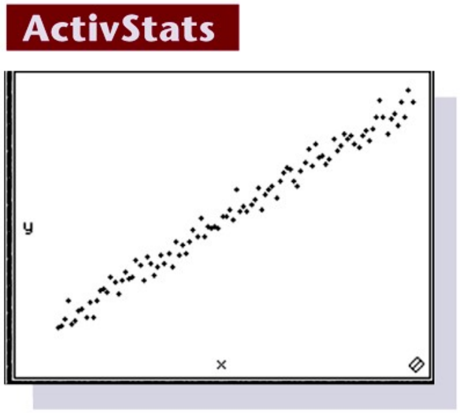
How we can calculate r?

- SAS – PROC CORR
- Excel - correl
- R - cor

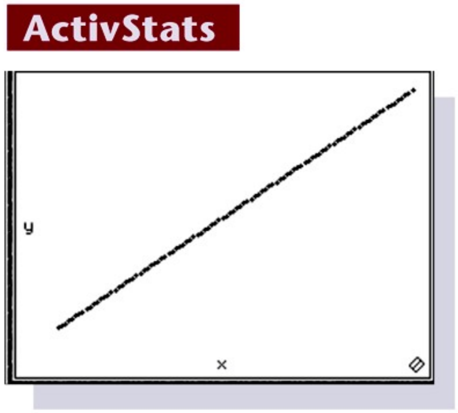
Scatterplots of Paired Data



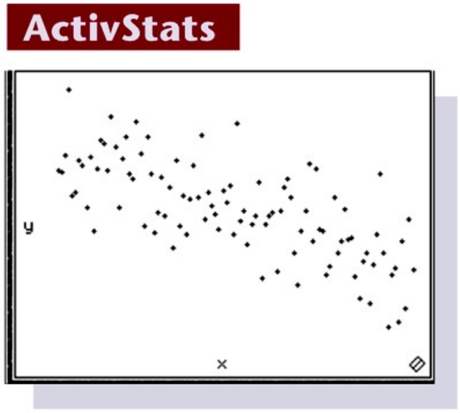
(a) Positive correlation: ***r*** **= 0.851**



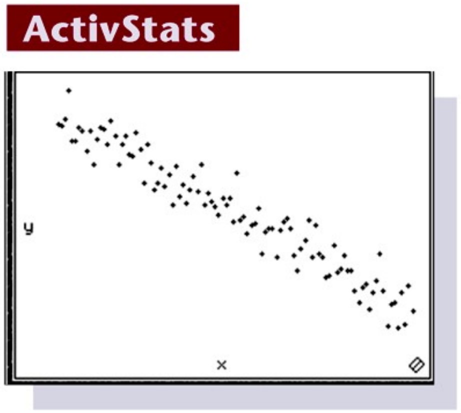
(b) Positive correlation: ***r*** **= 0.991**



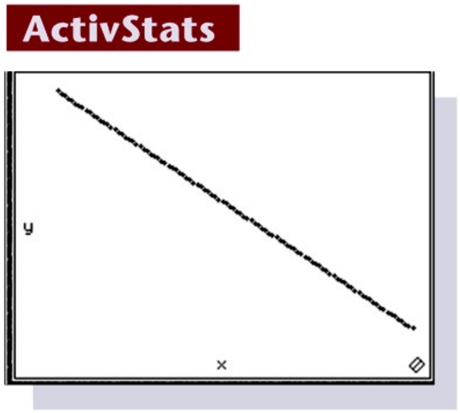
(c) Perfect positive correlation: ***r*** **= 1**



(d) Negative correlation: ***r*** **= −0.702**



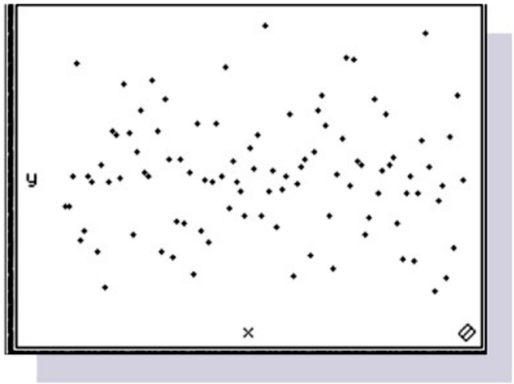
(e) Negative correlation: ***r*** **= −0.965**



(f) Perfect negative correlation: ***r*** **= −1**

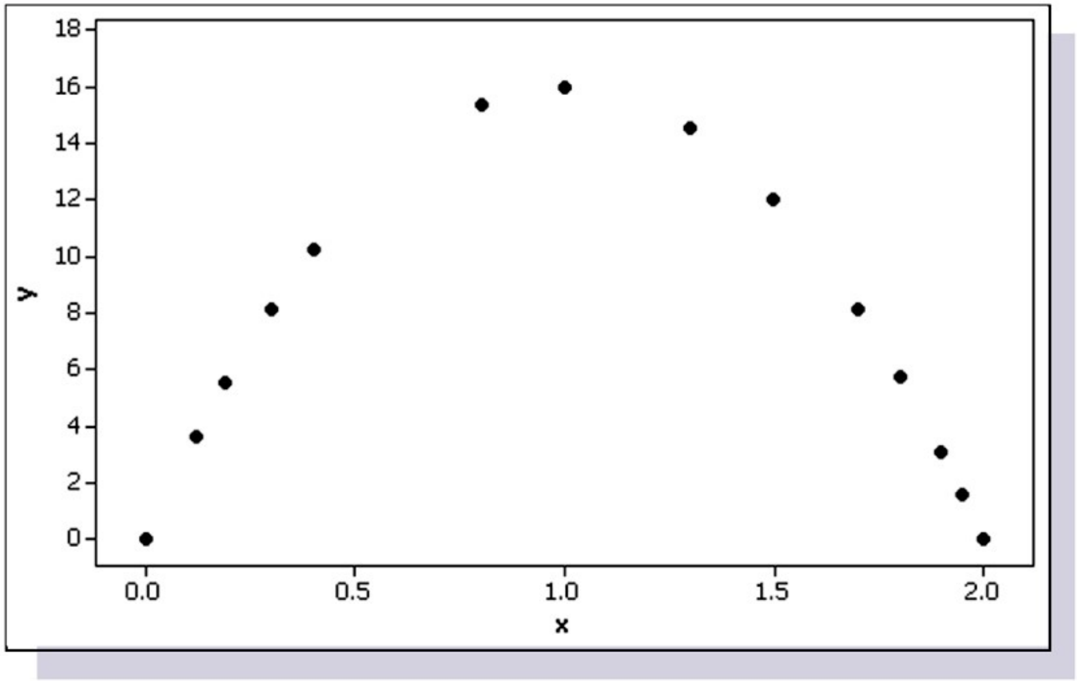
Scatterplots of Paired Data

ActivStats



No Linear Relationship r=0

Minitab



Nonlinear relationship: *r* = −0.087

Example

Which correlation coefficient best describes the scatter plot?

A. r = .32

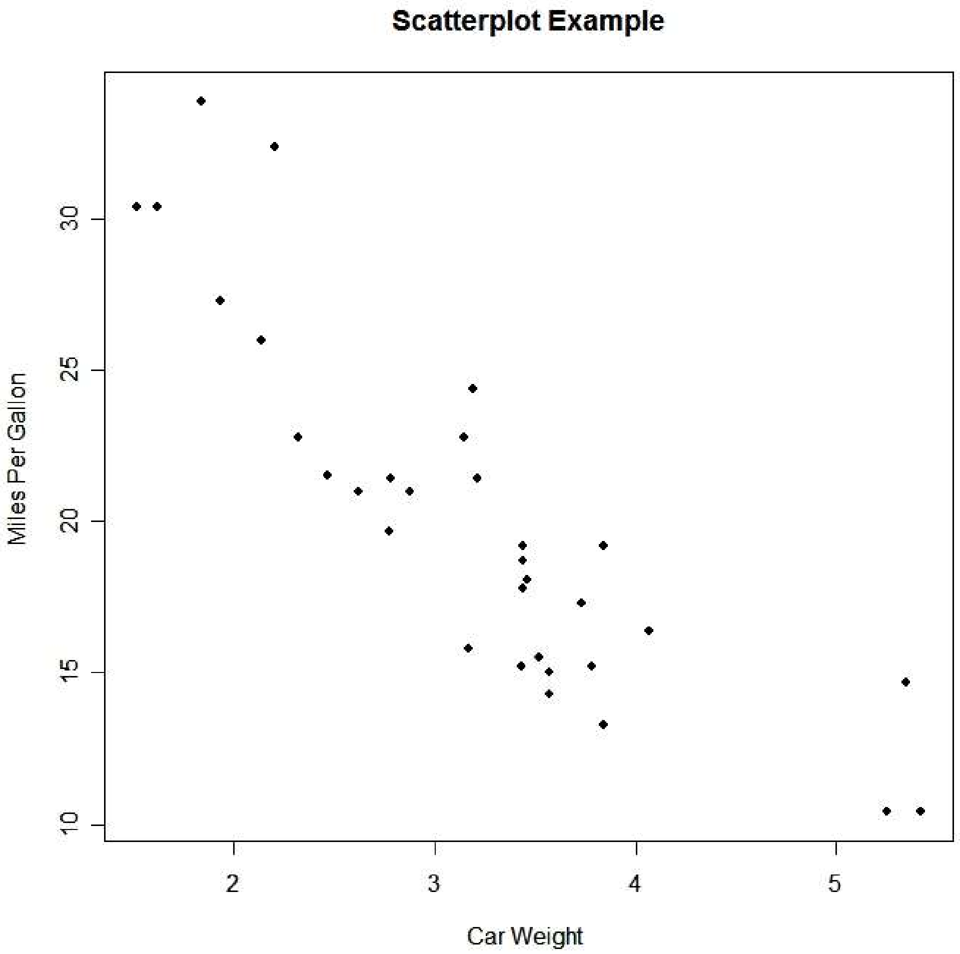
B. r = 0

C. r = -1

[✓ D. r = -.78]

E. r = -.1

F. r = .80



Example

Which correlation coefficient best describes the scatter plot?

A. r = .32

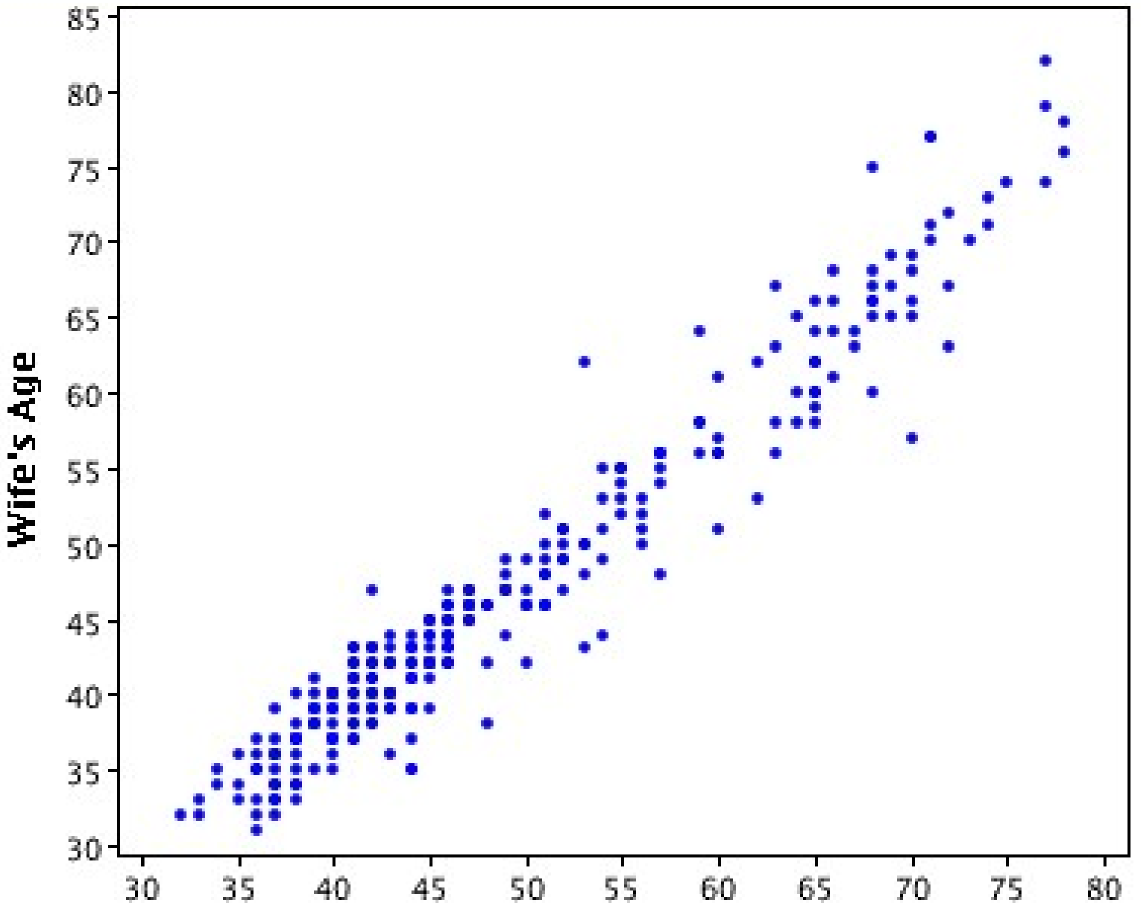
B. r = -.89

C. r = 1

D. r = -.345

[✓ E. r = .95]

F. -.5



Facts About Correlation

- Does not depend on the unit of measurement.

- Example: Say that I have measured the weight and height in pounds and inches respectively. r will be the same if I used kilograms and centimeters instead.

Checking data for randomness

- No formula for checks that guarantees data are random
- A few basic checks for randomness
  - Construct a scatter plot and calculate correlation
  - Examine distributional properties (Lab 02)
    - Plot the means within the categories to see if they are approximately equal (assuming they should be).
    - Plot the medians within the categories of interest to see if they are approximately equal (assuming they should be)
      - Keep in mind that a median is robust to unusually large or unusually small observations while mean is not