Evanston township High School

Data Analysis and Statistics

Introduction to Statistics

Jiangtao Gou
February 13, 2013

Lecture 1: Introduction

Three Reasons to Learn Statistics

- Being a capable “Information Consumer”
  - Extract Information form charts and graphs
  - Follow numerical arguments
  - Know the basics of how data should be gathered, summarized and analyzed to draw statistical conclusions

Three Reasons to Learn Statistics

- Understanding and Making Decisions
  - Decide if existing information is adequate
  - Collect more information in an appropriate way
  - Summarize the available data effectively
  - Analyze the available data
  - Draw conclusions, make decisions, and assess the risks of an incorrect decision

Three Reasons to Learn Statistics

- Evaluate Decisions that Affect Your Life
  - Help understand the validity and appropriateness of processes and decisions that effect your life

Data analysis

THE GATHERING, DISPLAY, AND SUMMARY OF DATA;

Probability

THE LAWS OF CHANCE, IN AND OUT OF THE CASINO;

Statistical inference

THE SCIENCE OF DRAWING STATISTICAL CONCLUSIONS FROM SPECIFIC DATA, USING A KNOWLEDGE OF PROBABILITY.

What is Statistics

- Statistics is the science of
  - Collecting data
  - Analyzing data
  - Drawing conclusions from data

What is your question/hypothesis?

How do you get your data?

How to analyze your data?

How to draw conclusions?

How to draw conclusions?

- Did you know that the great majority of people have more than the average number of legs?

How to draw conclusions?

- [It's obvious really; amongst the 57 million people in Britain there are probably 5,000 people who have only got one leg. Therefore the average number of legs is

$$\frac{(5000 * 1) + (56{,}995{,}000 * 2)}{57{,}000{,}000} = 1.9999123.....$$

- Since most people have 2 legs.....]

Descriptive Statistics

- The methods of organizing and summarizing data
- Graphical techniques
- Numerical techniques

Inferential Statistics

- Involves making generalizations from a sample to a population
- Estimation
- Decision making

Population

- The entire collection of individuals or objects about which information is desired

Sample

- A subset of the population, selected for study in some prescribed manner

Variable

- Any characteristic whose value may change from one individual to another

Data

- Observations on a single variable or simultaneously on two or more variables

Evanston Township High School

Data Analysis and Statistics

Introduction to Statistics

Jiangtao Gou
February 15, 2013

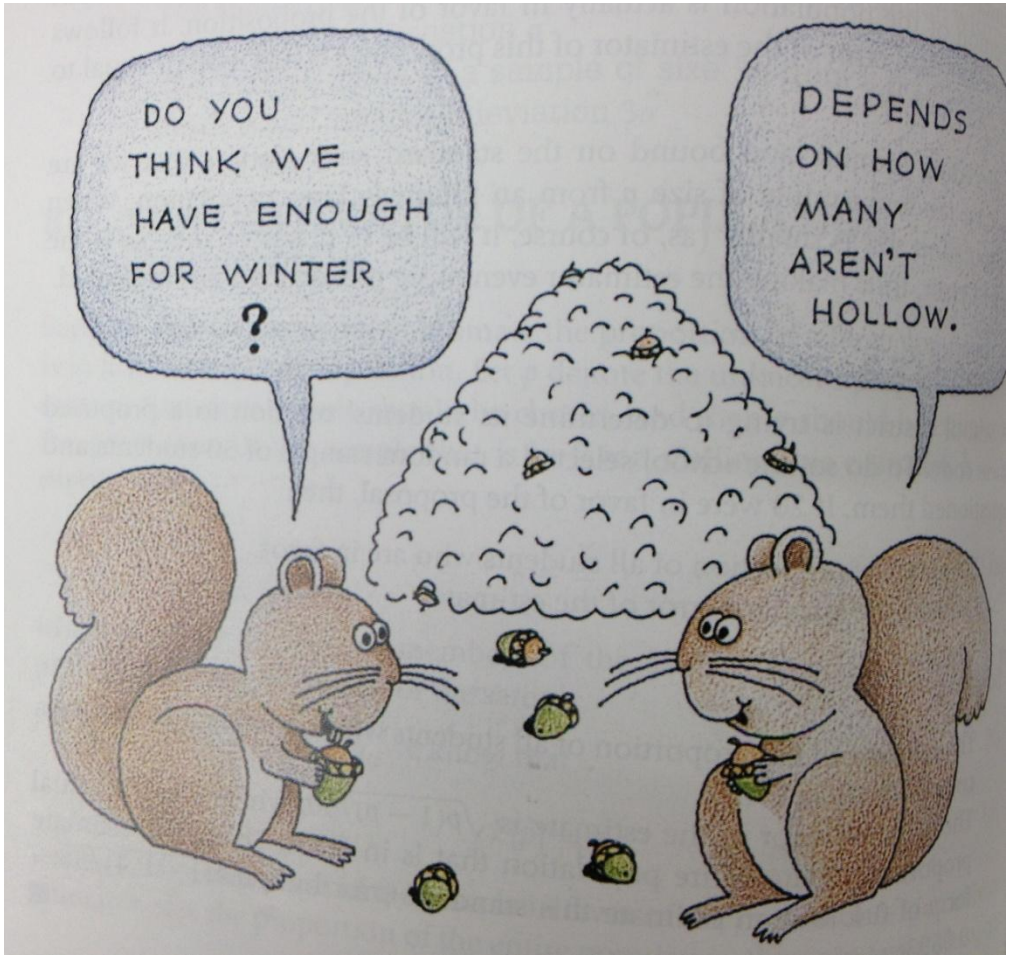
Lecture 2: Data Exploration

What we have learned

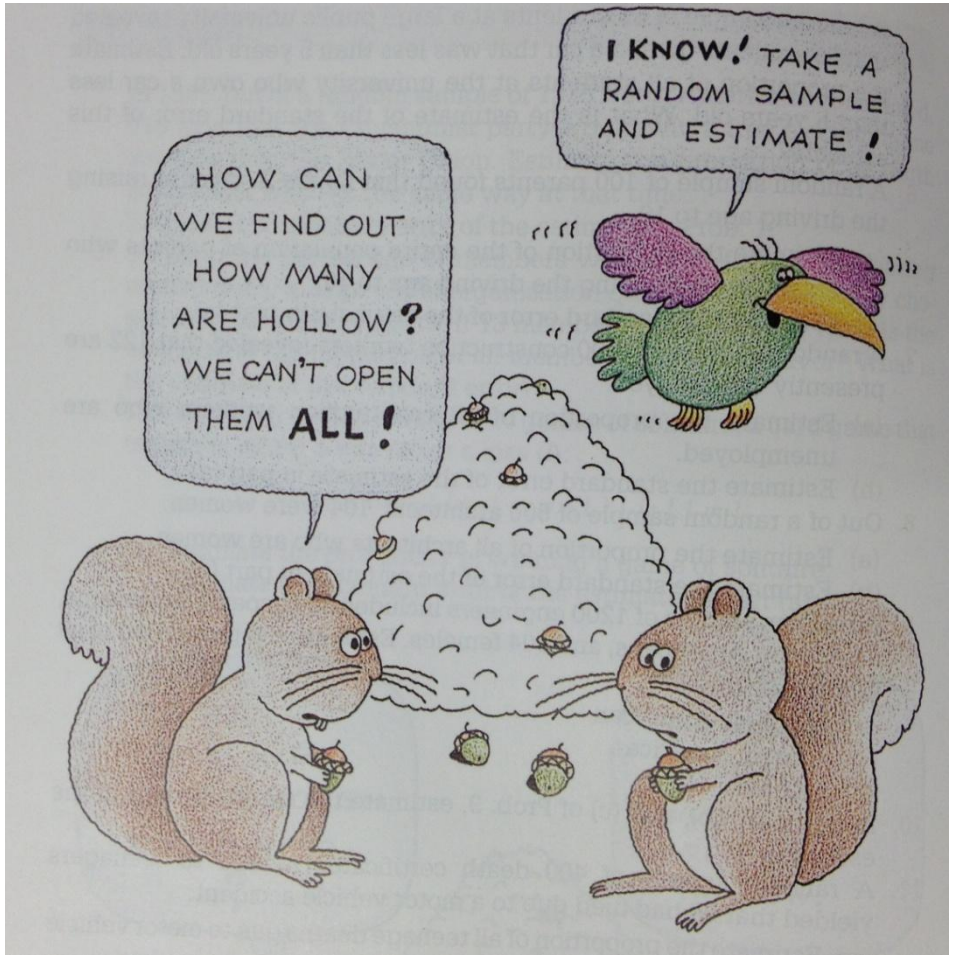
- Population
  - The entire collection of individuals or objects about which information is desired
- Sample
  - A subset of the population, selected for study in some prescribed manner
- A sample of cases are selected from some larger population that we would like to understand

Squirrels and Acorns

Population and Sample



Population and Sample



What we have learned

- Variable
  - Any characteristic whose value may change from one individual to another
- Data
  - Observations on a single variable or simultaneously on two or more variables
- Example
  - Variable: Hourly Temperature
  - Data: 28 F, 16 F, 32 F, 46 F, 19F, …

Outline

- How to picture one-variable data
  - Histogram
  - Stemplot
- How to picture two-variable data
  - Scatter plot

What should we do with Data?

There are **three** things you should always do first with data.

The First Thing

- Make a picture
  - A display of your data will reveal things you are not likely to see in a table of numbers and will help you to **think** clearly about the patterns and relationships that may be hiding in your data.

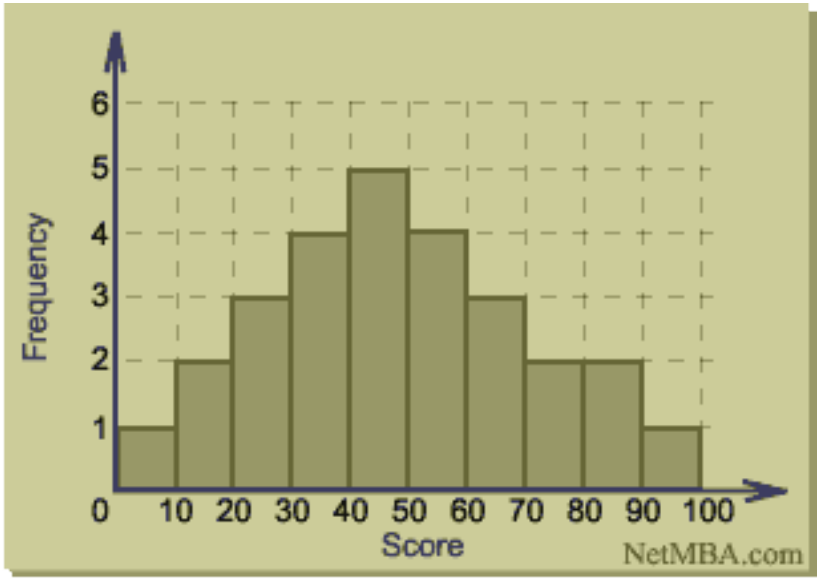
The Second Thing

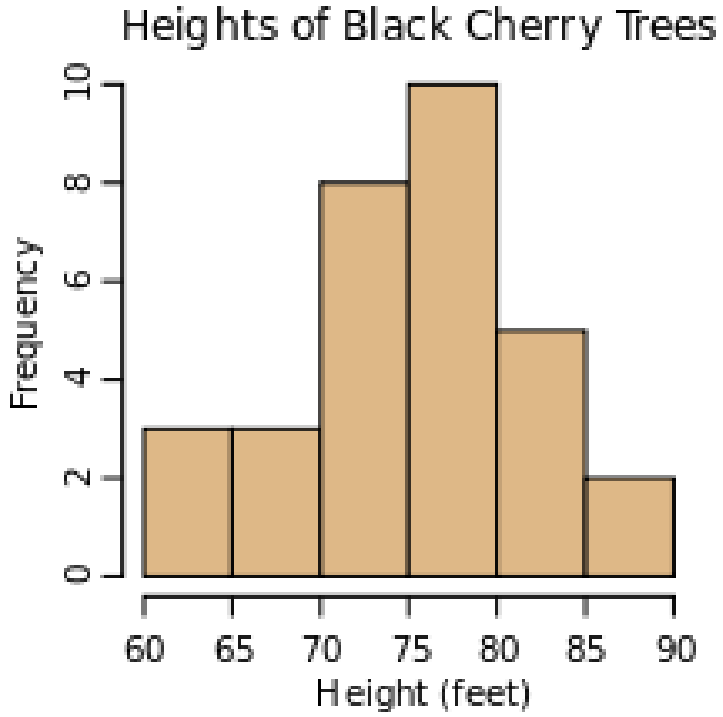
- Make a picture
  - A will-designed display will **show** the important features and patterns in your data. A picture will also show you the things you did not expect to see: the extraordinary (possibly wrong) data values or unexpected patterns.

The Third thing

- Make a picure
  - The best way to **tell** others about your data is with a well-chosen picture.

Histogram


Frequency
0
1
2
3
4
5
6
10
20
30
40
50
60
70
80
90
100
Score
NetMBA.com


Heights of Black Cherry Trees
Frequency
0
2
4
6
8
10
60
65
70
75
80
85
90
Height (feet)

Histogram

- Step 1. Choose the classes
  - Divide the range of the data into classes of equal width
- Step 2. Count the individuals
  - Count the number of individuals in each class
- Step 3. Draw the histogram
  - Mark the scale for the variable whose distribution you are displaying on the horizontal axis. The vertical axis contains the scale of counts.

Fig. 1. Histogram of a sample of college females arranged by height. For example, the two girls at the left are both 4′11″, the next five are all 5 feet tall, and so on. Notice the "bell-shaped" curve formed by the fact that many girls have the same height near the average, while fewer and fewer girls have any given height as one moves farther from the average.

Stemplot

- Stem-and-leaf Display (Dr John Tukey)

Why would I want to use Stemplot?

- It displays a lot of numbers in a neater, cleaner, and organized way.
- It makes it easier to see how spread out the numbers are
- It is easier to identify outliers

Things you need to know…

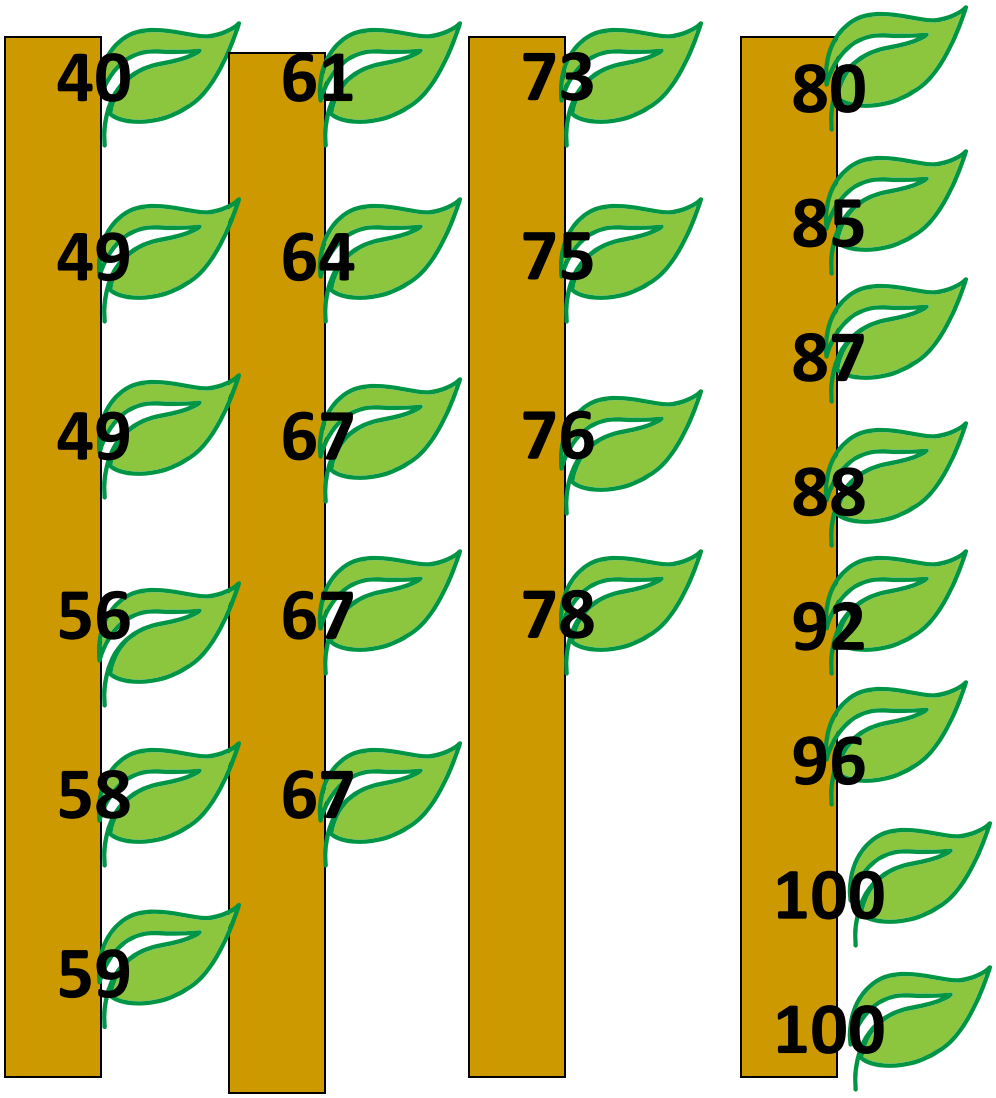
- **Stems**
  - The stems include the digits in the tens place and higher (sometimes the units place too, if data includes decimals)
  - The top of the stem includes all but the last digits of the smallest number
  - The bottom of the stem includes all but the last digit of the largest number
  - Your stems need to include ALL consecutive numbers in between—even if there is no data for it.

Things you need to know…

- **leaves**
  - The leaves are always from the LAST digit from your data (usually ones place)
  - The leaves need to be organized from least to greatest
  - You can have multiples of the same digit per line (interval)

This stem-and-leaf plot shows points that students received on a science quiz.

Stem	Leaves
4	0 9 9
5	6 8 9
6	1 4 7 7 7
7	3 5 6 8
8	0 5 7 8
9	2 6
10	0 0



Practice

- Stemplot and histogram of your pulse rates
  - Please take your own pulse rates and write them down
  - How to measure pulse rate
    - Hold the fingers of one hand on the artery in the neck near the thorax
    - Count your pulse for 30 seconds, then times 2 to get your pulse rate

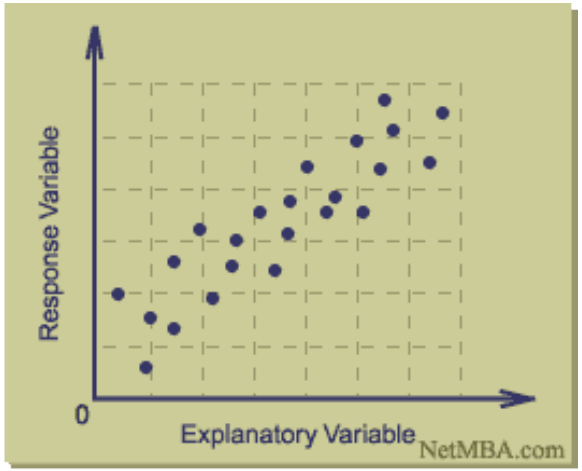
Practice

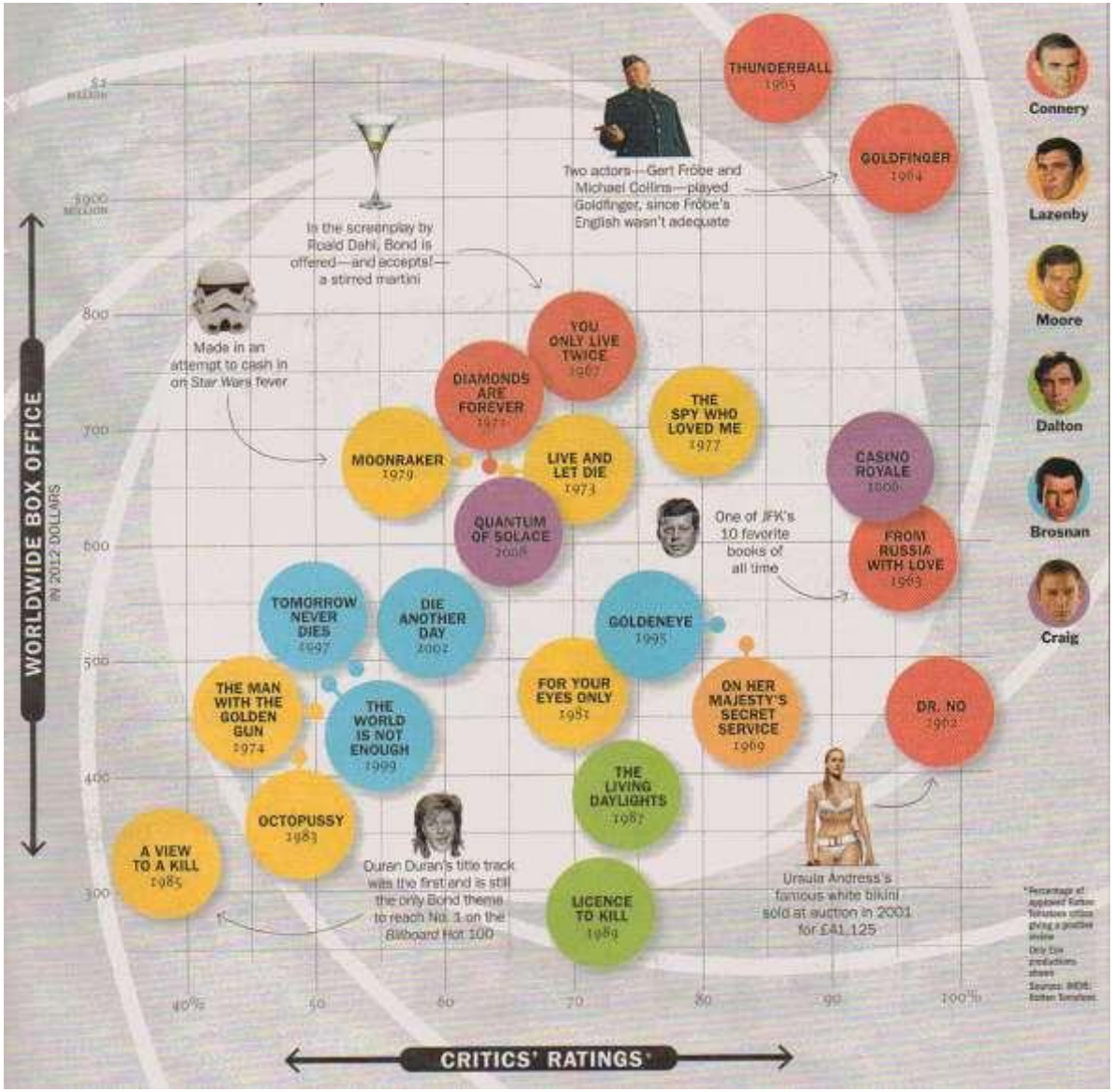

When feeling for the carotid pulse under the angle of the jaw, use very light pressure
A.D.A.M.

Practice

- Now I will collect the data
- and let us plot the stemplot (histogram) together~~

Scatterplot





Scatterplot

- A scatterplot shows the relations between two quantitative variables measured on the same individuals. The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis. Each individual in the data appears as the point in the plot fixed by the values of both variables for that individual.

Fig. 4. Scatter plot of college females arranged according to their height and weight. For example, the girl at the lower left is 4′11″ tall and weights between 90 and 95 pounds. Note the correlation between height and weight depicted by this plot. The average weight of taller girls is appreciably more than the average weight of shorter ones.

Practice

- Scatter plot of your height-weight relation
  - Please write down your height and weight

- I will collect the data and let us plot together

Homework 1

- Due February 20th

Evanston Township High School

Data Analysis and Statistics

Introduction to Statistics

Jiangtao Gou
February 20, 2013

Lecture 3: Sample Surveys and Design of Experiments

What we have learned

- How to picture one-variable data
  - Histogram
  - Stemplot
- How to picture two-variable data
  - Scatter plot

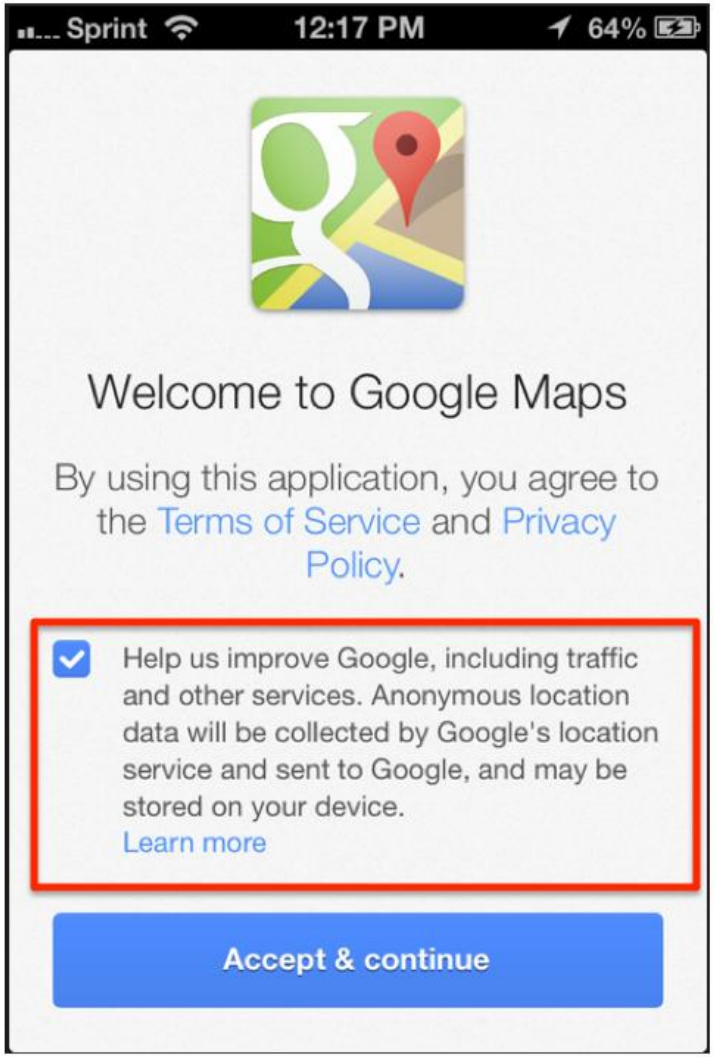
stemplot

34	084
33	045, 565, 739
32	267
31	543, 838, 843
30	449, 704
29	143, 874

Outline

- Data Ethics
- Gathering Data: Sampling
- Gathering Data: Experiments

Data Ethics


Sprint
12:17 PM
64%
Welcome to Google Maps
By using this application, you agree to the Terms of Service and Privacy Policy.
Help us improve Google, including traffic and other services. Anonymous location data will be collected by Google's location service and sent to Google, and may be stored on your device.
Learn more
Accept & continue

Basic Data Ethics

- All planned studies must be reviewed in advance by an **institutional review board** charged with protecting the safety and well-being of the subjects.
- All individuals who are subjects in a study must give their **informed consent** before data are collected.
- All individual data must be kept **confidential**. Only statistical summaries for groups of subjects may by made public.

Institutional Review boards

"To protect the rights and welfare of human subjects (including patients) recruited to participate in research activities"

IRB Institutional Review Board

search

NEW STUDIES

REVISIONS

- Preparing a Revision in eIRB
- Personnel Changes
- Consent Form Changes
- Funding Changes
- Investigator's Brochure Changes

MORE»

CONTINUING REVIEW / TERMINATION

WHEN THINGS GO WRONG

eIRB login with VPN

eIRB Overview and Instructions

NEWS & ANNOUNCEMENTS

Update to the Human Subjects Research Determination Form

The IRB has added a new question to the form. The revised form is dated 12/10/12. Read More »

Update by the FDA on Fungal Meningitis, dated 10-22-12

FDA provides NECC customer list Read More »

Modifications to the Authorized Research Personnel section in eIRB

Definition of research personnel has been revised; NU personnel without Net Ids may be placed in section 2.0 Read More »

EVENTS

FEB 20

Chicago Drop In Hours
11:30am - 1:00pm
Rehabilitation Institute of Chicago (RIC), Room 1301
345 E. Superior

FEB 21

Evanston Drop In Hours
2:00pm - 4:00pm
Chambers Hall, Rm 234
600 Foster

FEB 25

Evanston Drop In Hours
2:00pm - 4:00pm
Chambers Hall, Rm 234
600 Foster

INFORMATION FOR

- IRB Members
- SBS Researchers
- VA Researchers
- Research Participants

Informed Consent

Subjects must be informed in advance about the nature of a study and any risk of harm it may bring.

Informed Consent: A Case

- Golden rice
  - Golden rice is a rice produced through **genetic engineering**. The research was conducted with the goal of producing food to be grown and consumed in areas with a shortage of dietary vitamin A, which is estimated to kill 670,000 children under 5 each year.

Golden Rice

- Golden Rice is a **genetically modified** (**GM**) product.
- There is now a string of evidence that exposure of many species of animals to a variety of genetically modified crops, and food and feed derived from them, can cause illnesses and death, raising the distinct possibility that genetic modification is inherently dangerous. This is reinforced in results obtained in the most recent studies.

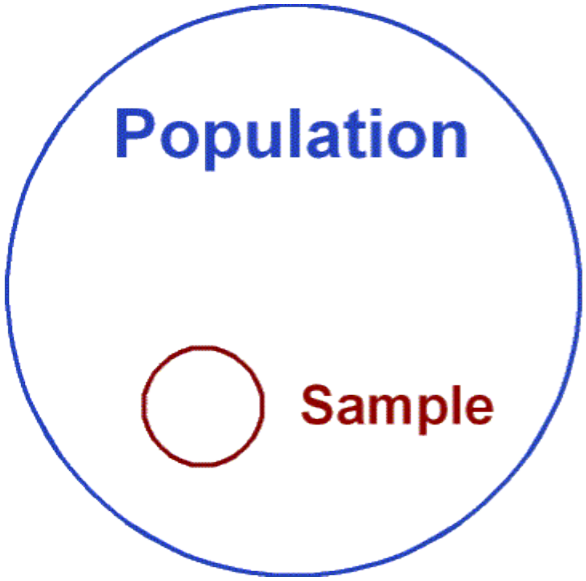
The Golden Rice Scandal

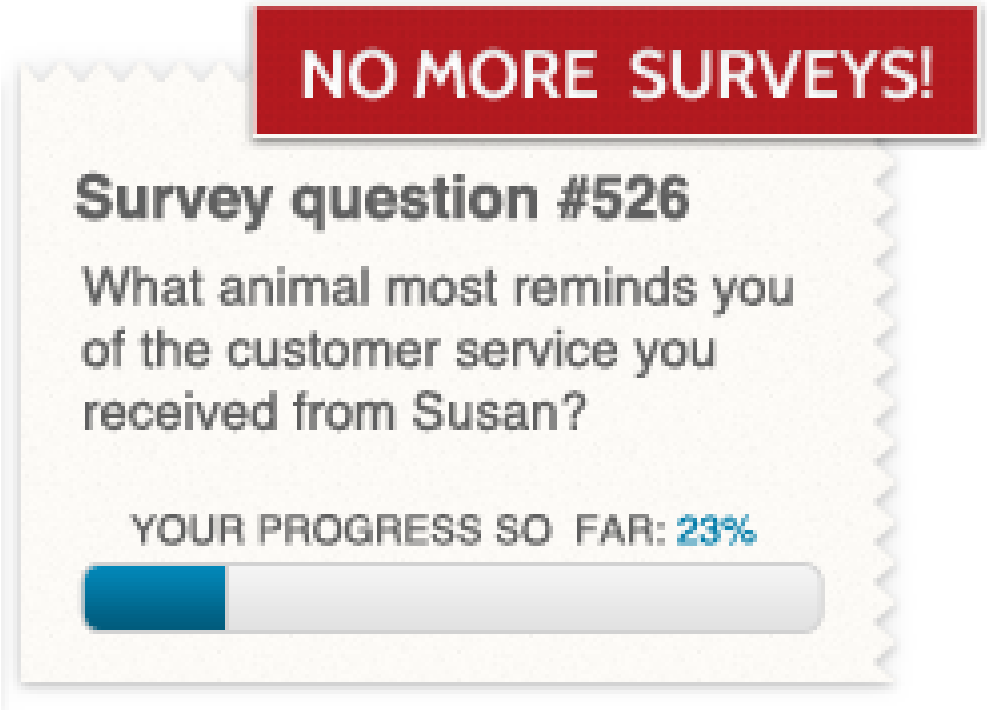
- In August 2012, Tufts University and others published new research on Golden Rice in the *American Journal of Clinical Nutrition* showing that the beta carotene produced by Golden Rice is as good as beta carotene in oil at providing vitamin A to children.
- The study states that "recruitment processes and protocol were approved", but questions have been raised about the use of Chinese children to test the effects of Golden Rice.

Confidentiality

- Protect the subjects' privacy by keeping all data about individuals confidential.

Sample Surveys





Look at the numbers **quickly**, and pick a number **at random**

1 2 3 4

Did you pick **3**?

- If you pick 3, you have got company. Almost 75% of all people pick the number 3.
- About 20% pick either 2 or 4.
- If you pick 1, well, consider yourself a little different. Only about 5% choose 1.

Sampling

- Examine a part of the whole
- Randomize
- Sample Size

- Examine a part of the whole
  - We'd like to know about an entire population of individuals, but examining all of them is usually impractical, if not possible.

- Randomize
  - Randomizing protects us from the influences of all the features of our population by making sure that, on average, the sample looks like the rest of the population.

- Sample size
  - The fraction of the population that you have sampled does not matter. It is the sample size itself that is important.

Bias

- Sampling methods that, by their nature, tend to over- or underemphasize some characteristics of the population are said to be biased.

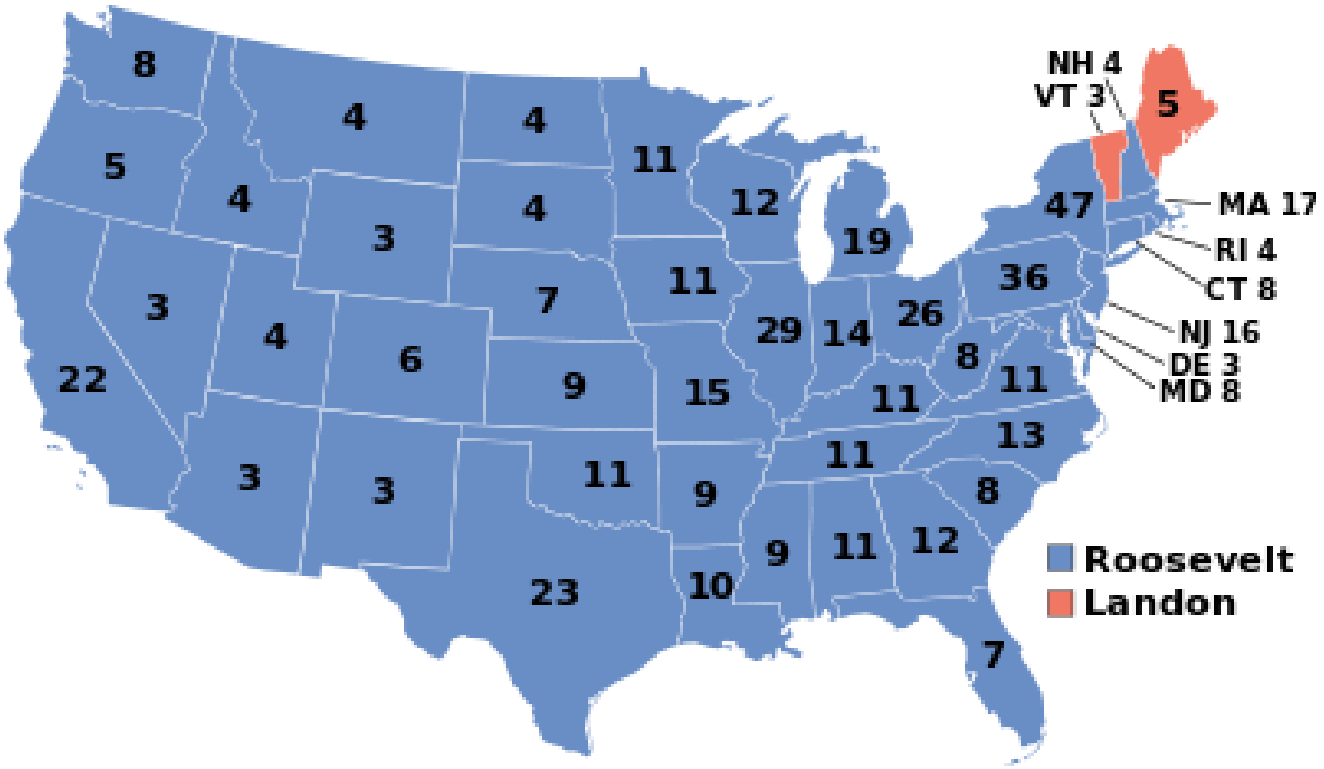
Bias: Example

- United States presidential election, 1936

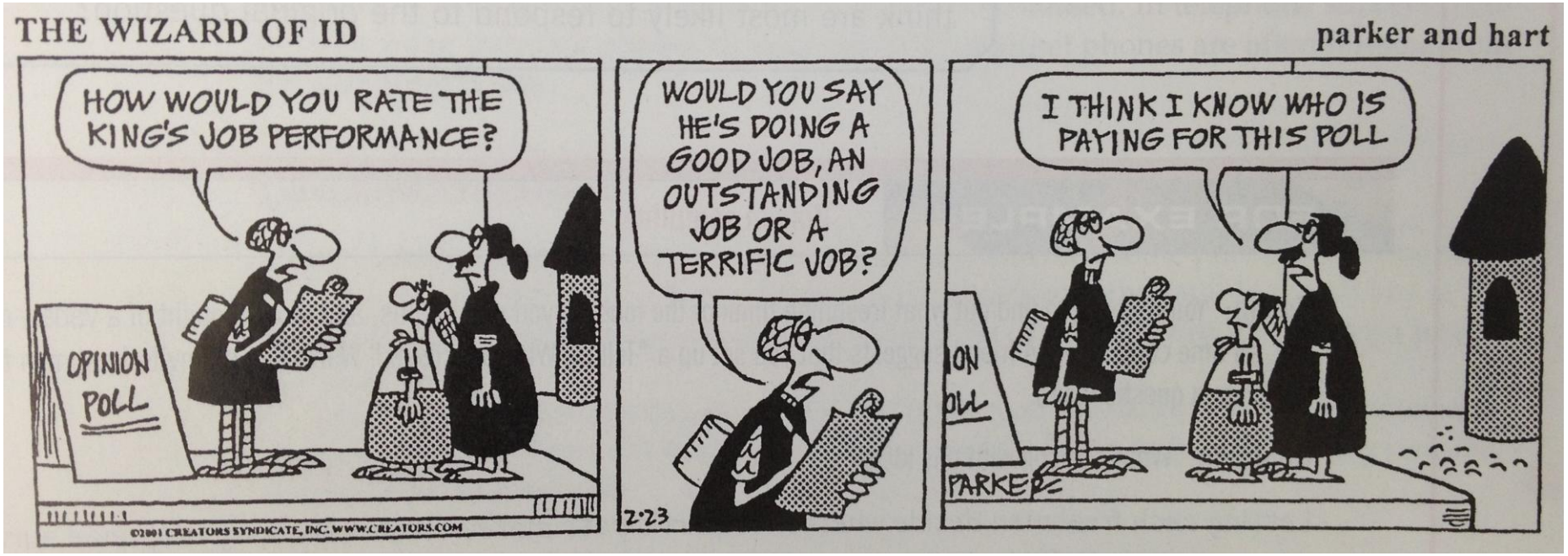
1932 ← | November 3, 1936 | → 1940

Nominee	**Franklin D. Roosevelt**	Alf Landon
Party	Democratic	Republican
Home state	New York	Kansas
Running mate	**John N. Garner**	Frank Knox
Electoral vote	**523**	8
States carried	**46**	2
Popular vote	**27,752,648**	16,681,862
Percentage	**60.8%**	36.5%

Prediction: 57% vs 43%
Results: 37% vs 62%



Bias: Survey is not valid



Sampling

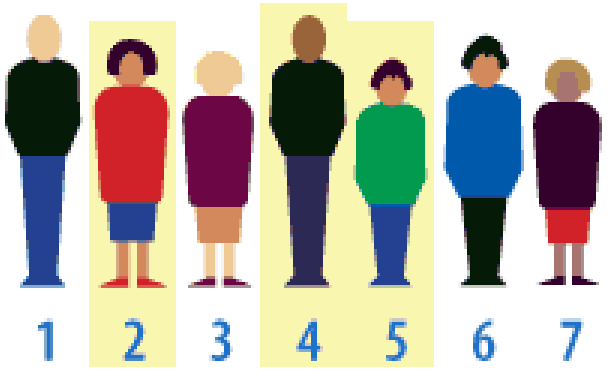
- Simple Random Samples (SRS)
- Stratified Sampling
- Cluster Sampling

Simple Random Samples

- A simple random sample is a subset of individuals (a sample) chosen from a larger set (a population). Each individual is chosen randomly and entirely by chance.

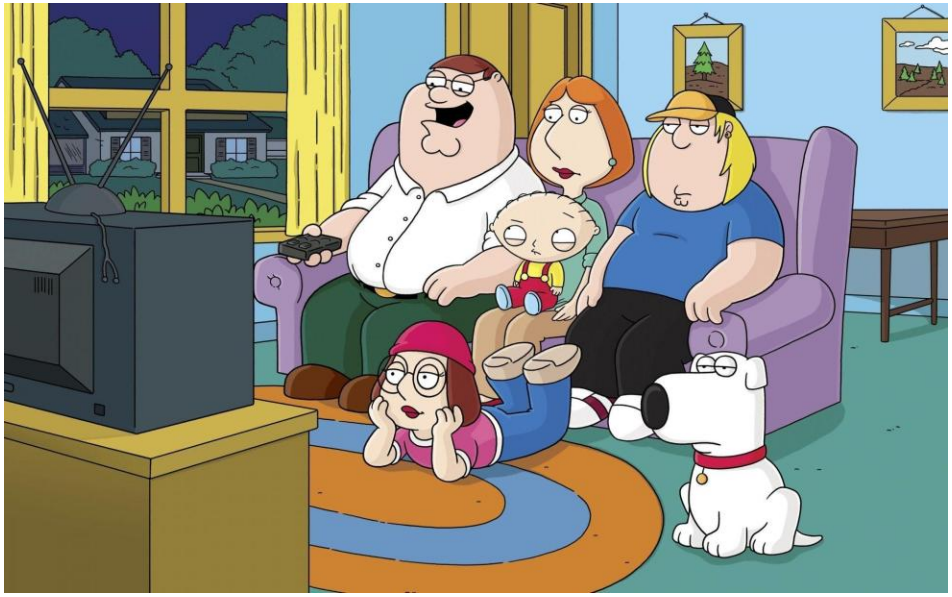
Activity: SRS

- Here are the steps:
  - Assign each member of your population a numerical label.
  - Use statistical software or a random digit table to select numerical labels at random.



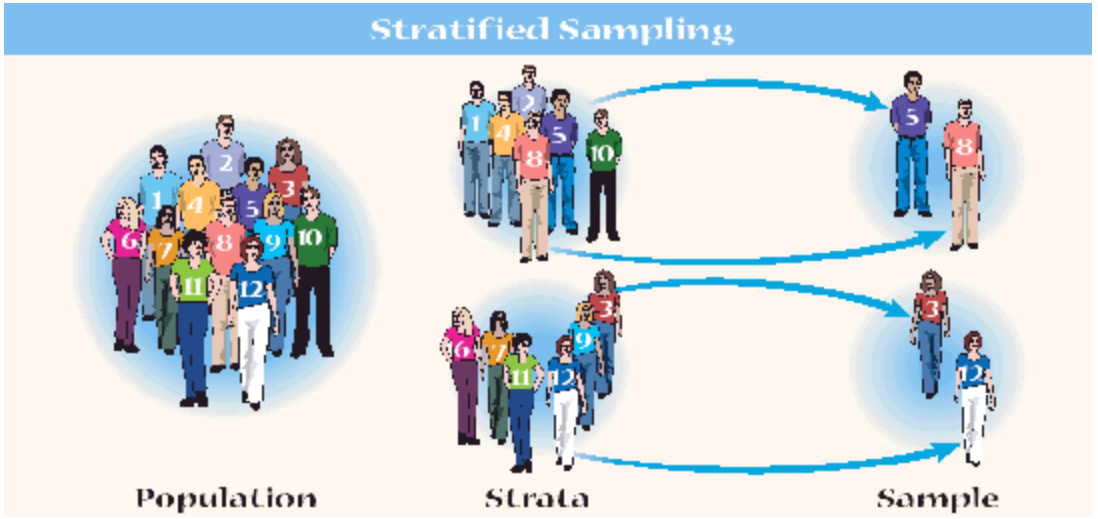
Mini Questionnaire

- Would you like to answer a question?
  - I would like to
  - I am not willing to
- How many hours of TV did you watch last weekend?



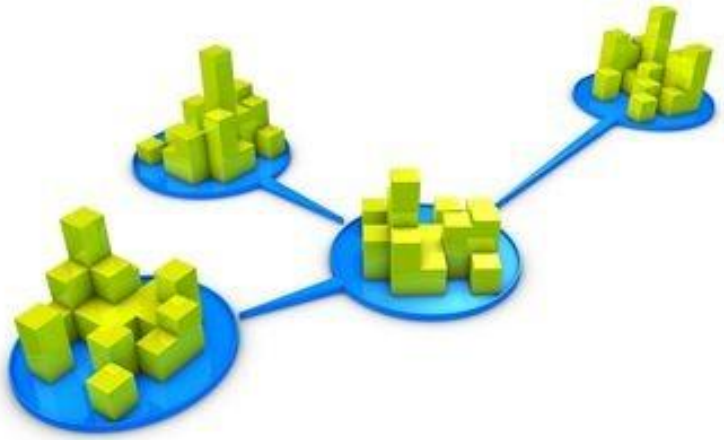
Stratified Sampling

- The population is first sliced into homogeneous groups, called strata, before the sample is selected. Then simple random sampling is used within each stratum before the results are combined.

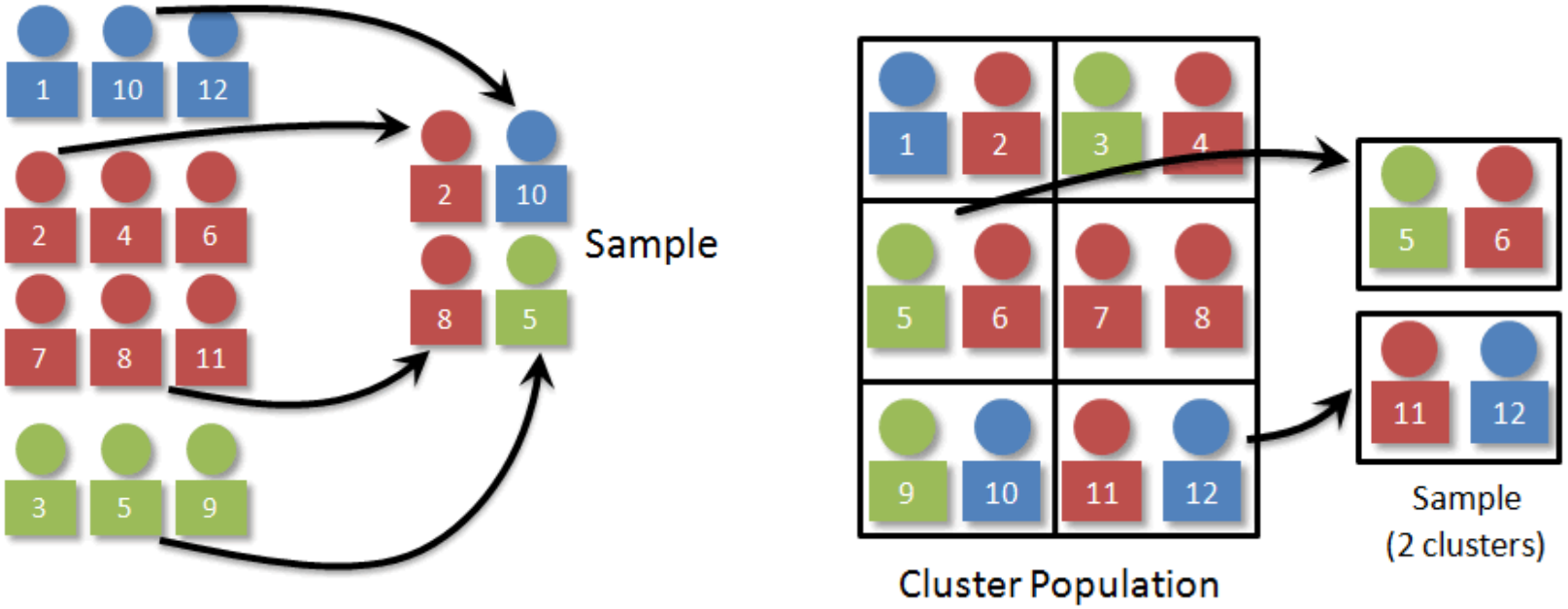


Cluster Sampling

- Splitting the population into representative clusters can make sampling more practical. Then we could simply select one or few clusters at random and perform a census within each of them.



Strata vs. Cluster



Design of Experiments

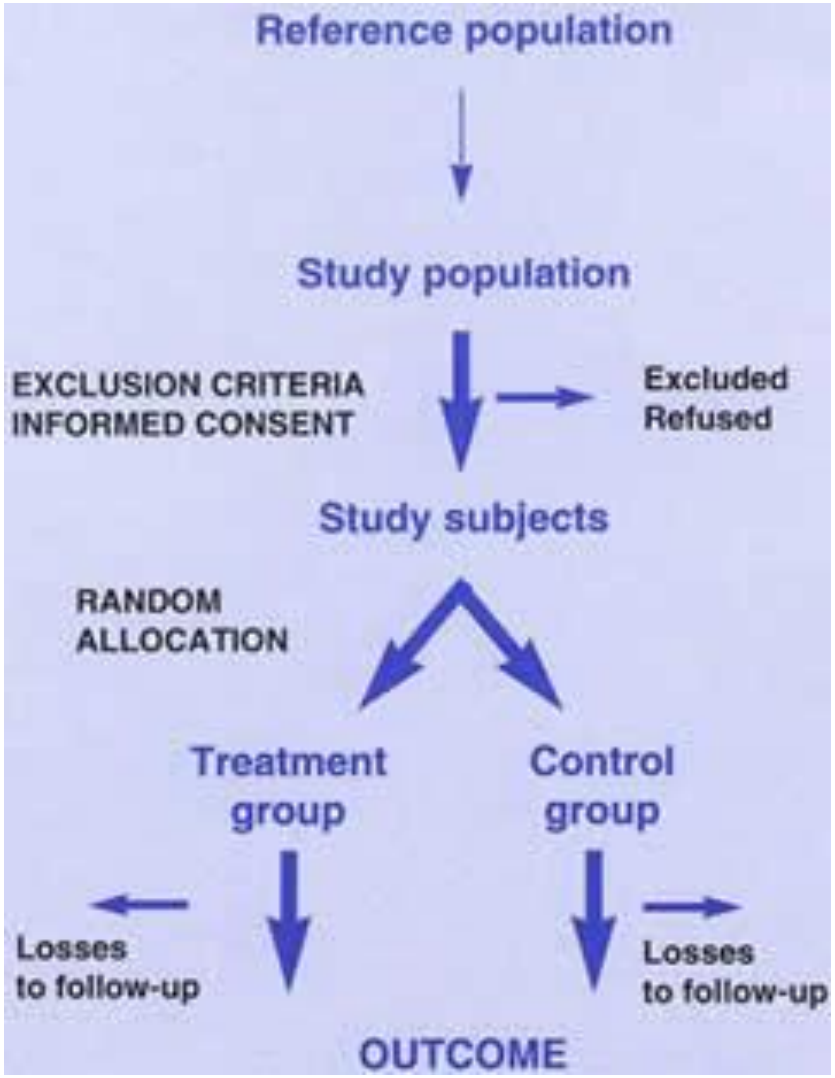
M	A	G	I	C
G	I	C	M	A
C	M	A	G	I
A	G	I	C	M
I	C	M	A	G

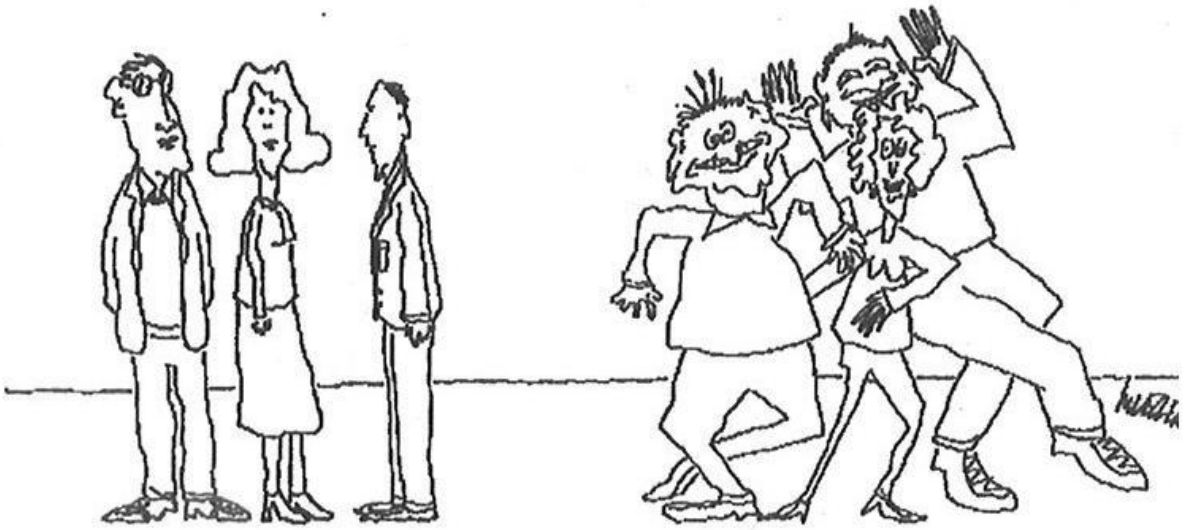
Design of Experiments

- An experiment requires a random assignment of subjects to treatments.

Design of Experiments

- Control
  - We control sources of variation other than the factors we are testing by making conditions as similar as possible for all treatment
- Randomize
  - Randomization allows us to equalize the effects of unknown or uncontrollable sources of variation
- Replicate
  - The outcome of an experiment on a single subect is an anecdote, not data.


Reference population
Study population
EXCLUSION CRITERIA
INFORMED CONSENT
Excluded
Refused
Study subjects
RANDOM
ALLOCATION
Treatment
group
Control
group
Losses
to follow-up
Losses
to follow-up
OUTCOME


CONTROL GROUP
OUT OF CONTROL GROUP.


I'D LIKE TO GET THIS PRESCRIPTION FILLED
PHARMACY
©2007 CREATORS SYNDICATE, INC., WWW.CREATORS.COM
PARKER.
THIS WILL RUN ABOUT $100
HOW MUCH FOR THE PLACEBO?
1.2

Evanston Township High School

Data Analysis and Statistics

Lecture 4: Using Statistics to Summarize Data Sets

2012-02-22

Mean and Deviation

Sample mean is defined to equal the arithmetic average of the data values.

Median and Percentiles

The sample median is the middle value in the ordered list.

Variance and Standard Deviation (Population and Sample)

Normal data set, z-scores, percentiles.

Evanston Township High School

Data Analysis and Statistics

Introduction to Statistics

Jiangtao Gou
February 27, 2013

Lecture 5: Z-Score and Correlation

What we have learned

- Mean and Deviation
  - Sample mean is defined to equal the arithmetic average of the data values.
- Median and Percentiles
  - The sample median is the middle value in the ordered list.
- Variance and Standard Deviation

Review: Standard Deviation

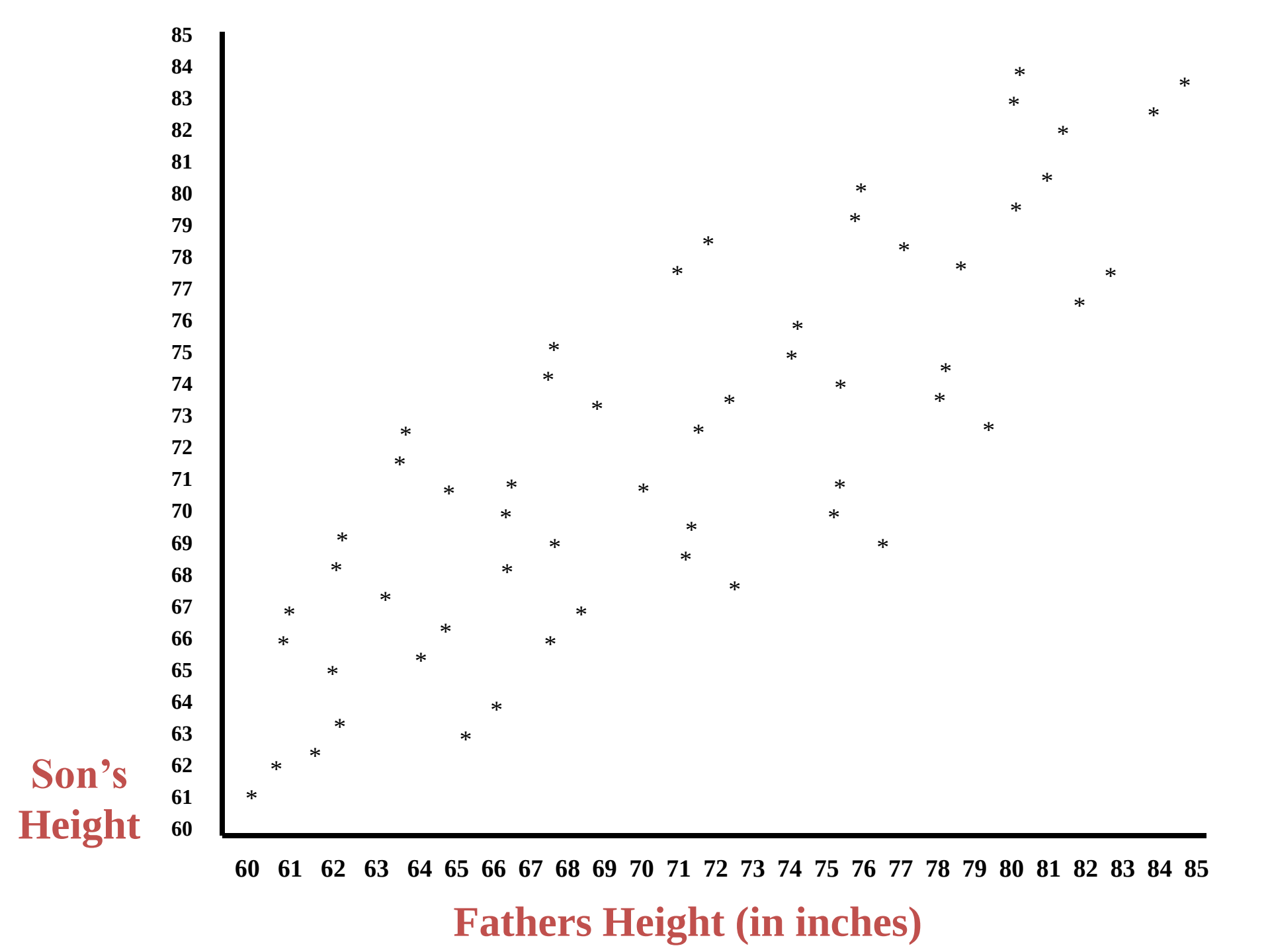
- Statistics is about **variation**, so spread is an important fundamental concept in statistics.
- Standard deviation (SD) takes into account how far each value is from the mean.
- We square each deviation (squaring always gives a positive value), add up these squared deviations and find their average. We call the result the variance.
- To get back to the original units, we take the square root of variance to get SD.

Outline

- Correlation
- Z-Score
- Normal Distribution

Correlation --- Basic Concepts

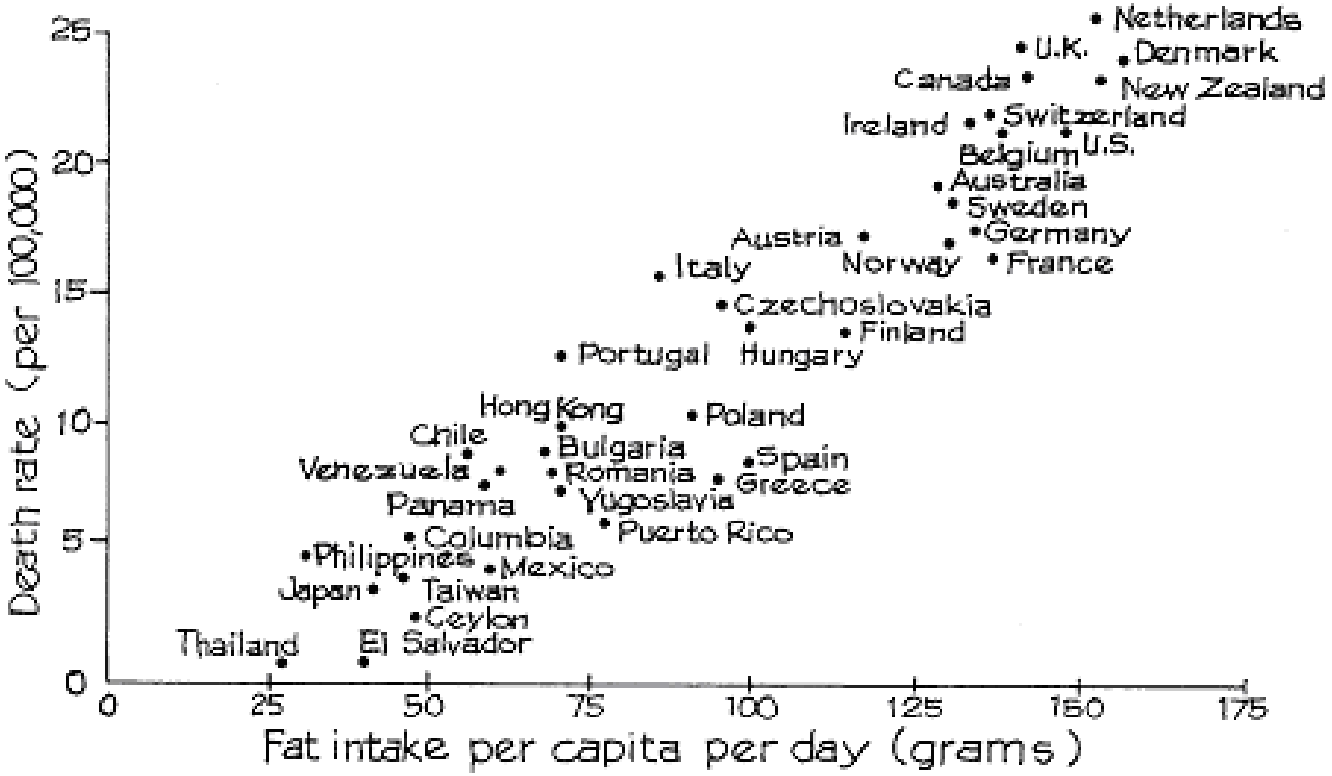
- Review: Scatter Plot
- Consists of a set of ***ordered*** pairs
- Indicates both the ***magnitude*** and ***direction*** of the relationship between variables
- Range is from -1.0 to +1.0


Son's Height
85
84
83
82
81
80
79
78
77
76
75
74
73
72
71
70
69
68
67
66
65
64
63
62
61
60
60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85
Fathers Height (in inches)

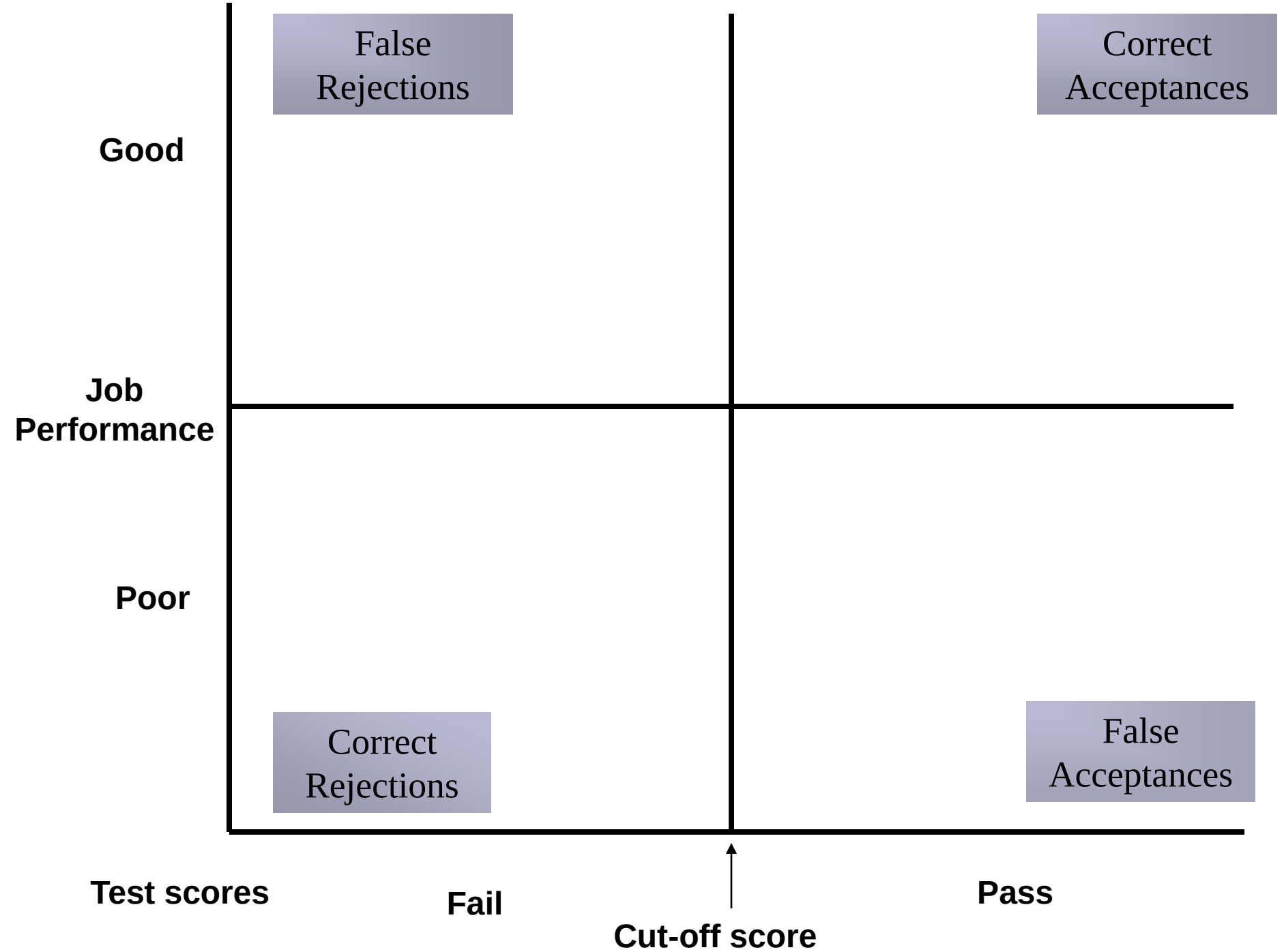
Consider these --- A correlation exists between:

- The total amount of losses in a fire and the number of firemen that were putting out the fire
- Cigarette smokers and lower GPAs
- The number of churches in a city and amount of alcohol consumed
- The amount of fat in diets and cancer rates

Figure 8. Cancer rates plotted against fat in the diet, for a sample of countries.



Source: K. Carroll, "Experimental evidence of dietary factors and hormone-dependent cancers," *Cancer Research* vol. 35 (1975) p. 3379. Copyright by *Cancer Research*. Reproduced by permission.


False Rejections
Correct Acceptances
Good
Job Performance
Poor
Correct Rejections
False Acceptances
Test scores
Fail
Pass
Cut-off score

Computation of Standard Deviation & Variance

Test Scores ↓ X	Deviation scores (scores minus the mean) ↓ x	Squared deviation scores ↓ x^2
10	-20	400
20	-10	100
30	0	0
40	10	100
50	20	400

EX = 150

(EX/N) = 30 (Mean)

EX^2 = 1000 (Sum of the squared deviation scores)

EX^2/N = 200 (the variance or s^2) → **Mean of the sum of the squared deviation scores**

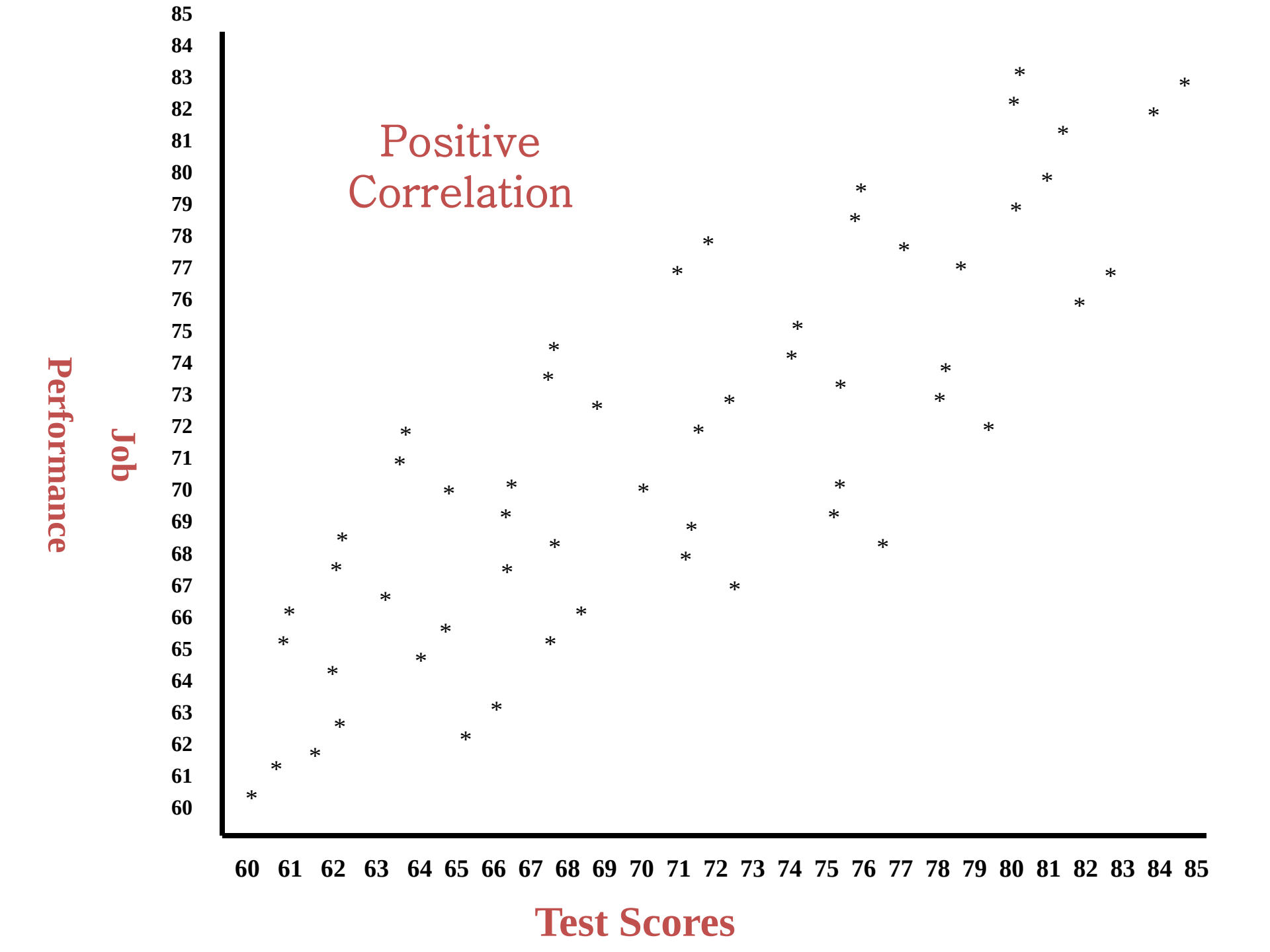
$\sqrt{s^2}$ **= standard deviation or s** → $\sqrt{200}$ **= 14.14 (standard deviation)**

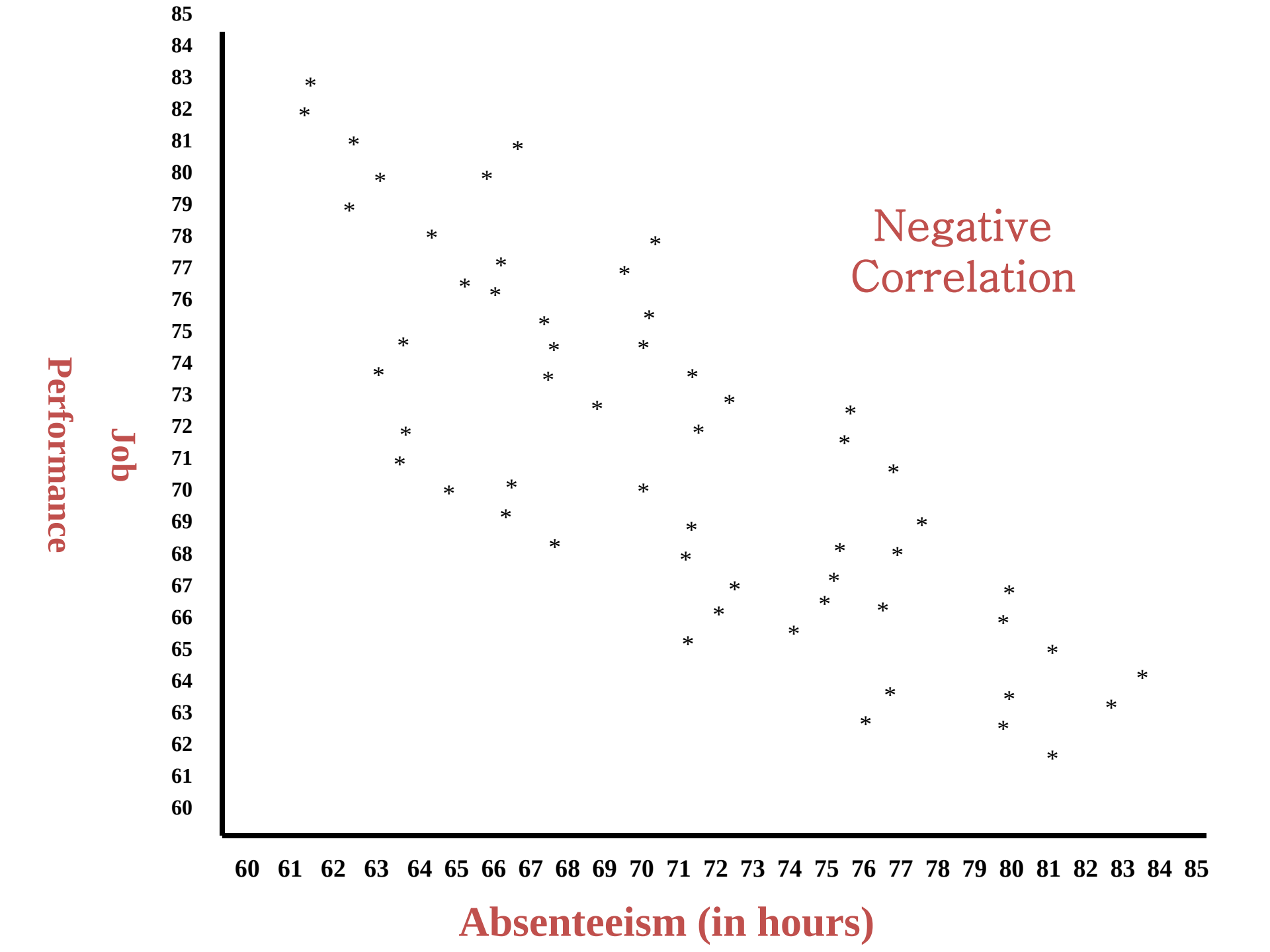
Computational Formula for *r*

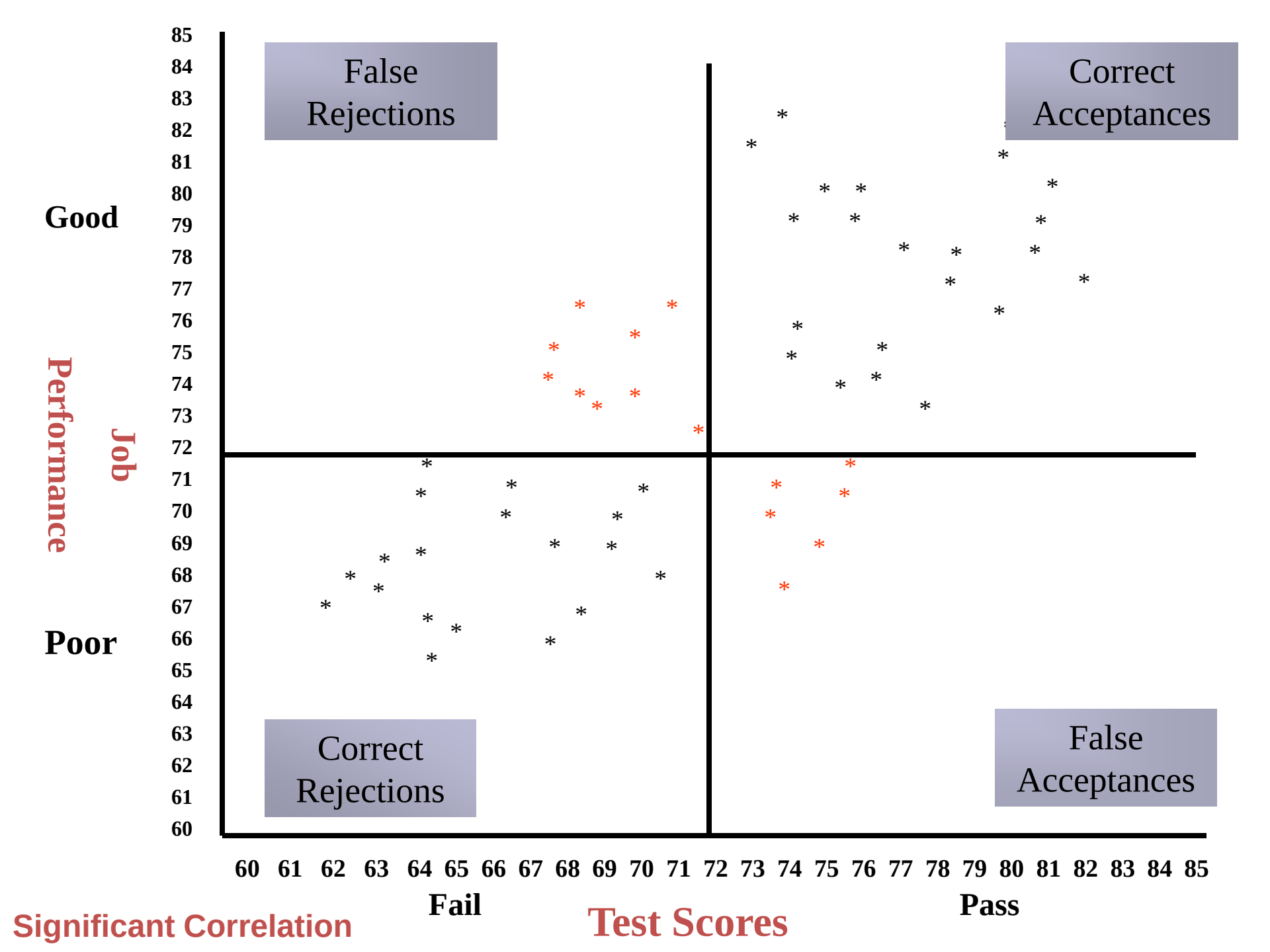
X	Y	XY	X^2	Y^2
1	4	4	1	16
2	3	6	4	9
3	5	15	9	25
4	7	28	16	49
5	6	30	25	36
$\Sigma X = 15$	$\Sigma Y = 25$	$\Sigma XY = 83$	$\Sigma X = 55$	$\Sigma Y = 135$

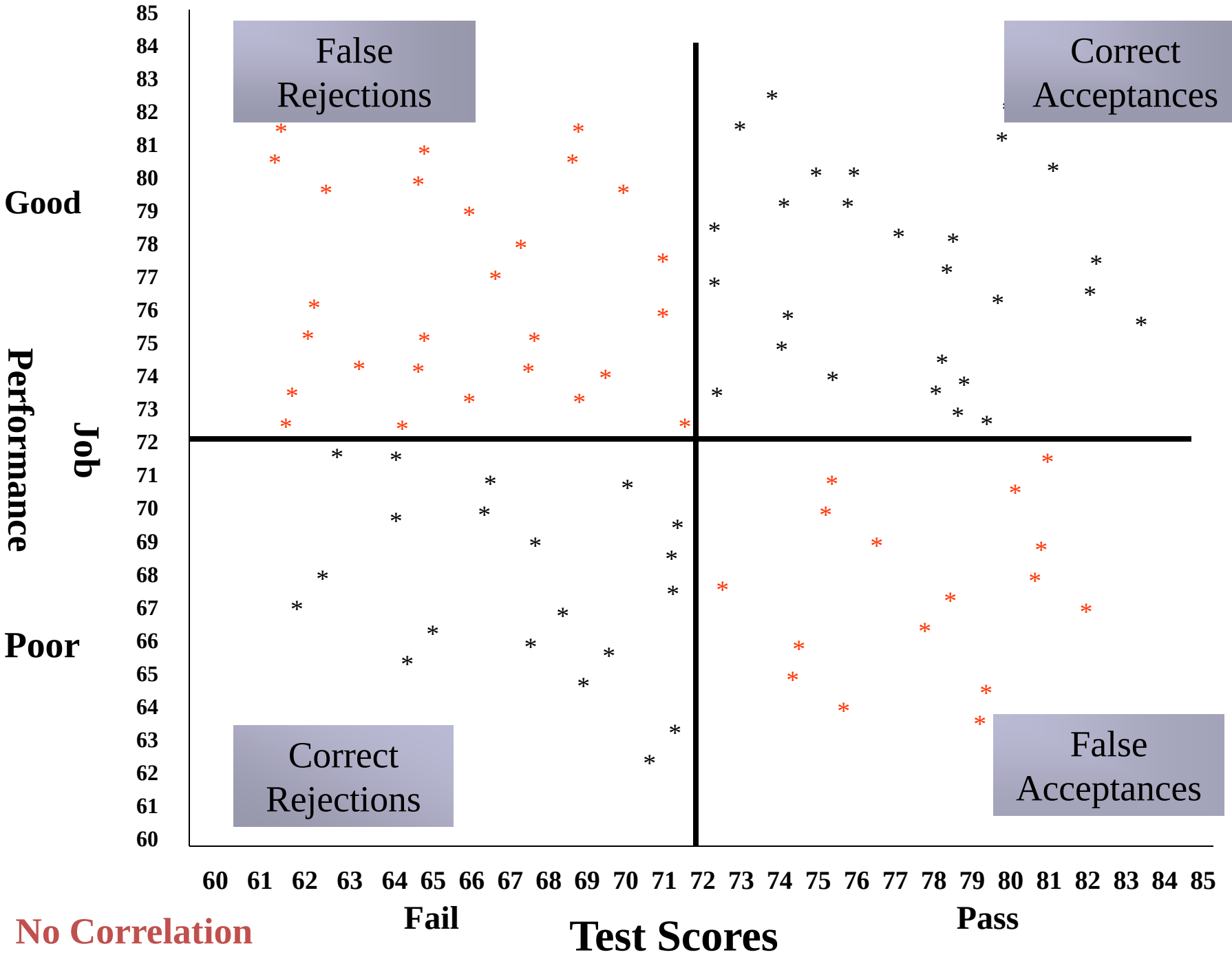
$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}} = \frac{5(83) - (15)(25)}{\sqrt{[5(55) - (15)^2][5(135) - (25)^2}}$$

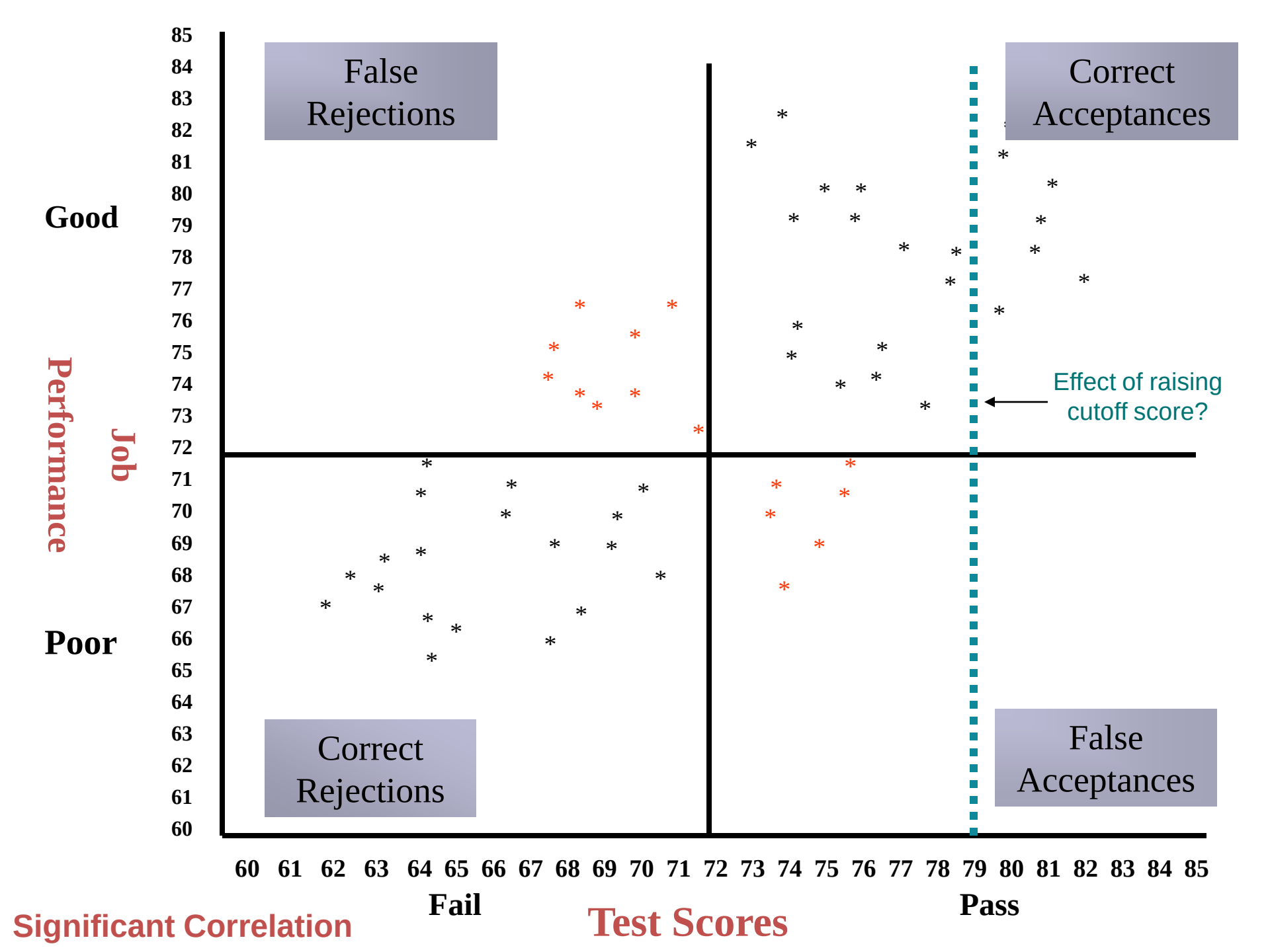
$$= \frac{415 - 375}{\sqrt{(275 - 225)(675 - 625)}} = \frac{40}{\sqrt{(50)(50)}} = \frac{40}{50} = .80$$

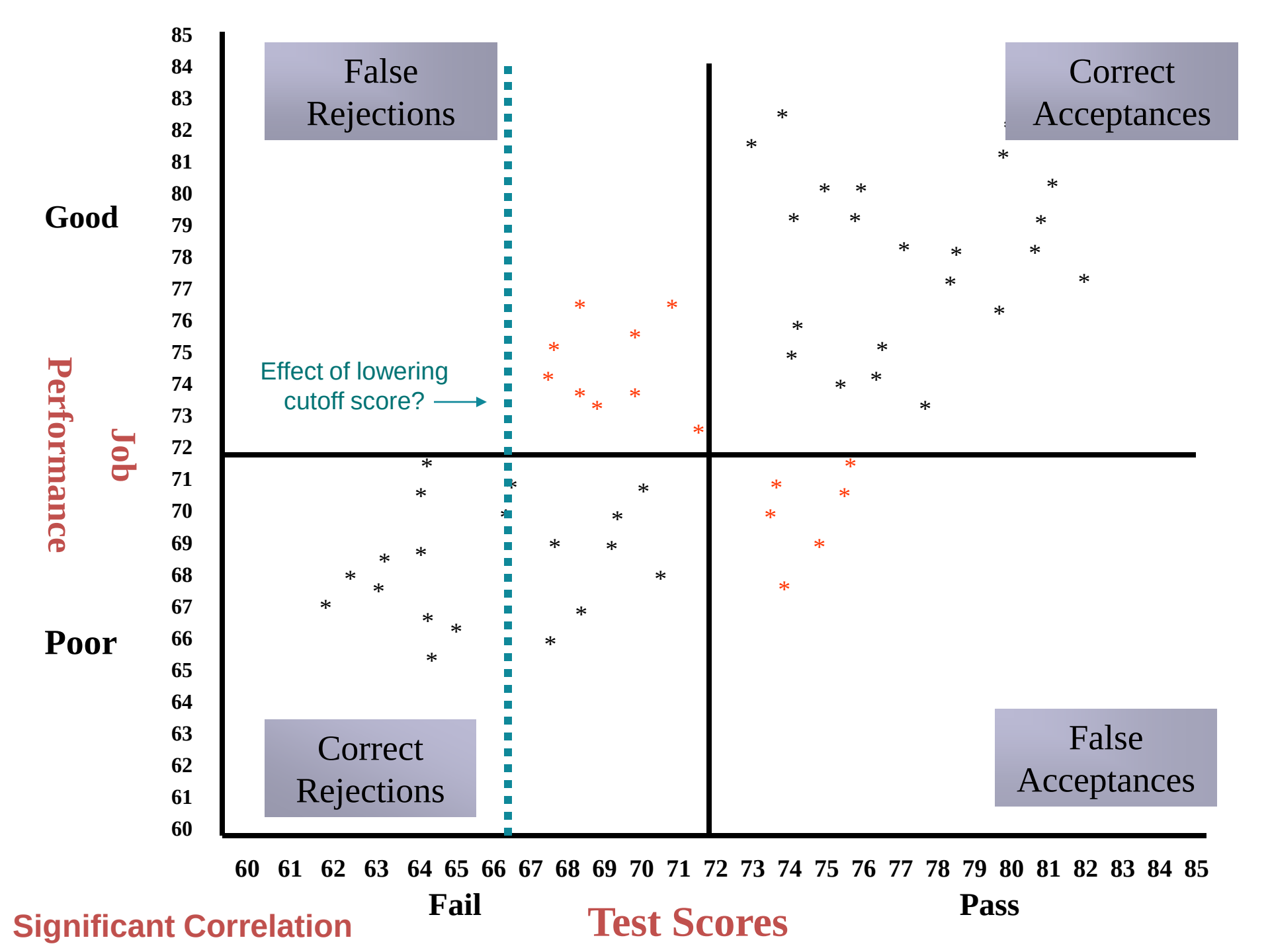

Positive Correlation
Job Performance
85
84
83
82
81
80
79
78
77
76
75
74
73
72
71
70
69
68
67
66
65
64
63
62
61
60
60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85
Test Scores


Negative Correlation
Job Performance
85 84 83 82 81 80 79 78 77 76 75 74 73 72 71 70 69 68 67 66 65 64 63 62 61 60
60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85
Absenteeism (in hours)


False Rejections
Correct Acceptances
Correct Rejections
False Acceptances
Good
Poor
Job Performance
85 84 83 82 81 80 79 78 77 76 75 74 73 72 71 70 69 68 67 66 65 64 63 62 61 60
60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85
Fail
Pass
Test Scores
Significant Correlation


False Rejections
Correct Acceptances
Correct Rejections
False Acceptances
Good
Job Performance
Poor
85 84 83 82 81 80 79 78 77 76 75 74 73 72 71 70 69 68 67 66 65 64 63 62 61 60
60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85
Fail
Test Scores
Pass
No Correlation


False Rejections
Correct Acceptances
Correct Rejections
False Acceptances
Effect of raising cutoff score?
Good
Poor
Job Performance
85 84 83 82 81 80 79 78 77 76 75 74 73 72 71 70 69 68 67 66 65 64 63 62 61 60
60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85
Fail
Pass
Test Scores
Significant Correlation


False Rejections
Correct Acceptances
Effect of lowering cutoff score?
Correct Rejections
False Acceptances
Good
Poor
Job Performance
85 84 83 82 81 80 79 78 77 76 75 74 73 72 71 70 69 68 67 66 65 64 63 62 61 60
60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85
Fail
Pass
Test Scores
Significant Correlation

The Normal distributions

Objectives

The Normal distributions

- Density curves
- Normal distributions
- The 68-95-99.7 rule
- The standard Normal distribution
- Using the calculator to find Normal proportions
- Finding a value given a proportion

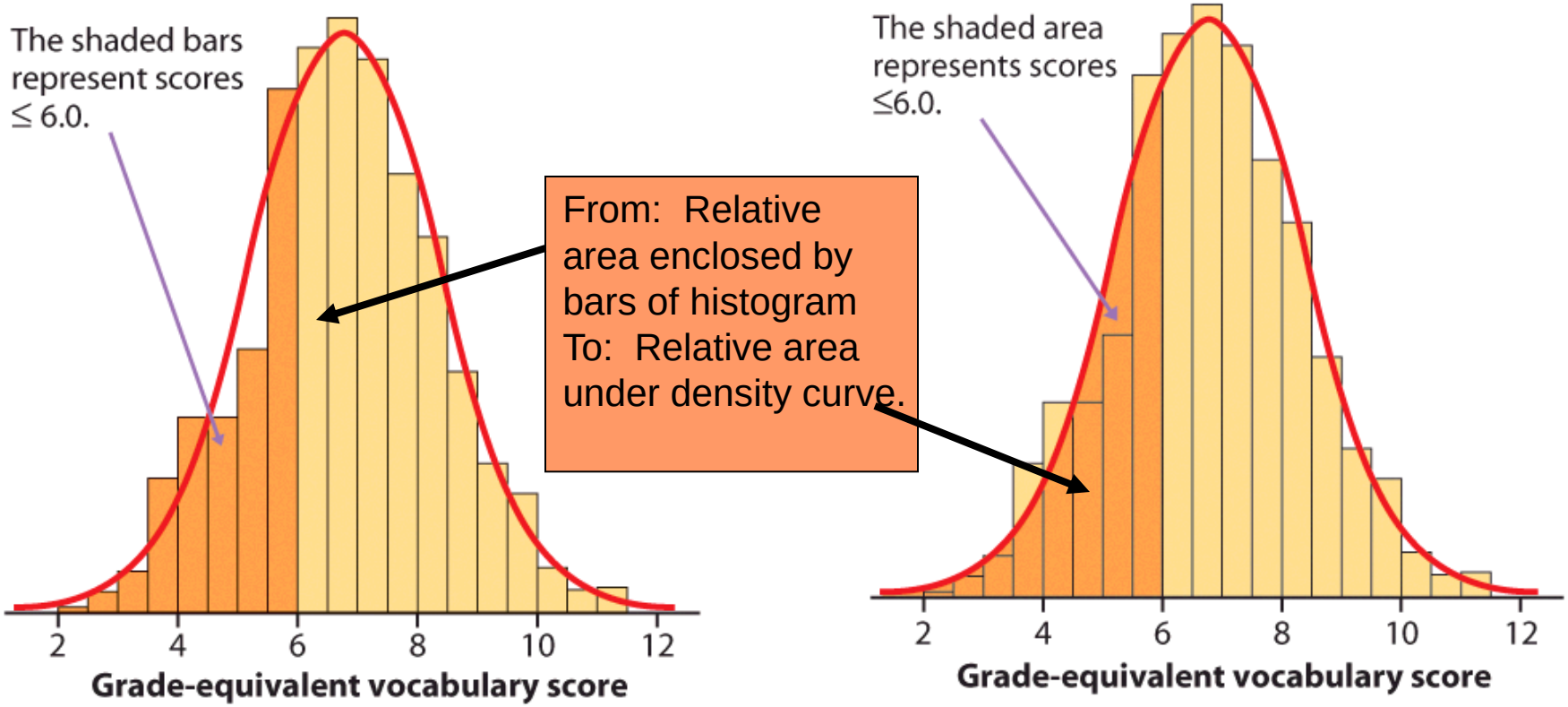
Density curves

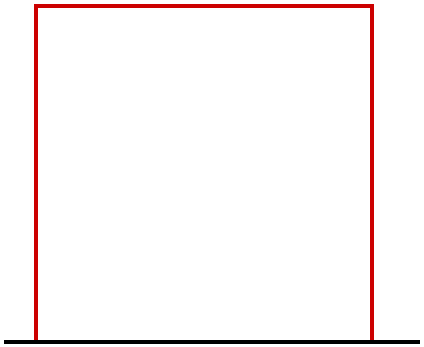
A **density curve** is a **mathematical model** of a distribution.

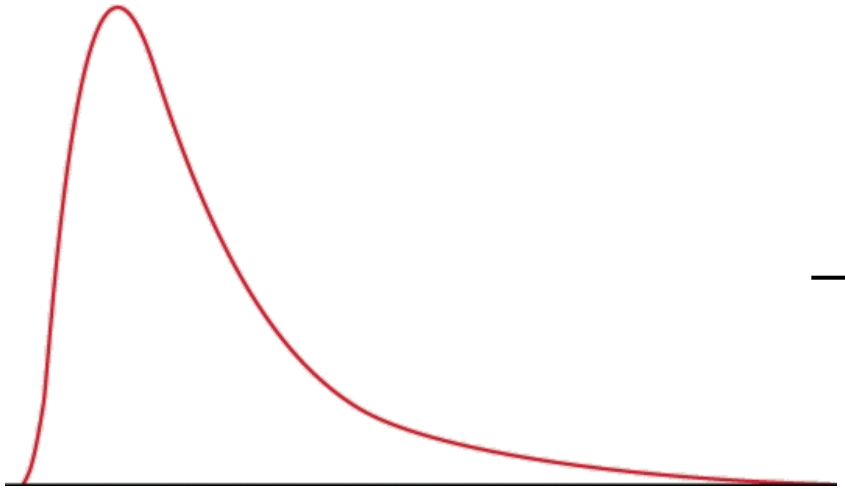
It is always on or above the horizontal axis.

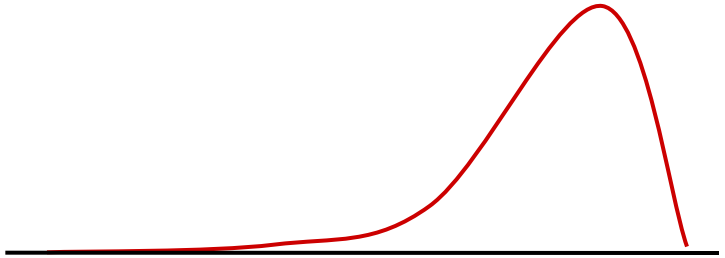
The total area under the curve, by definition, is equal to 1, or 100%.

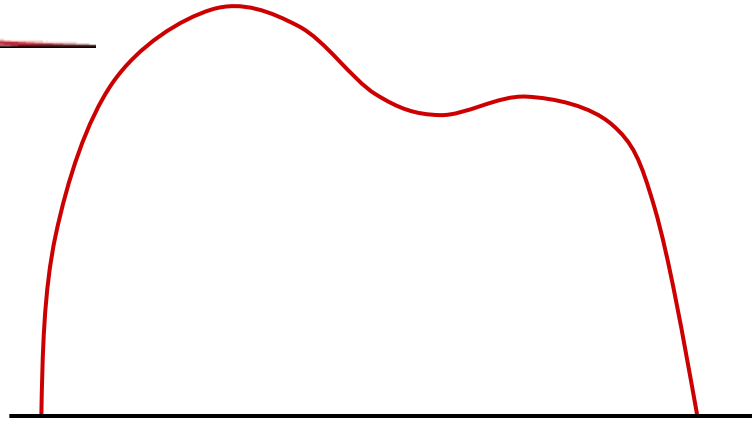
The area under the curve for a range of values is the proportion of all observations for that range.





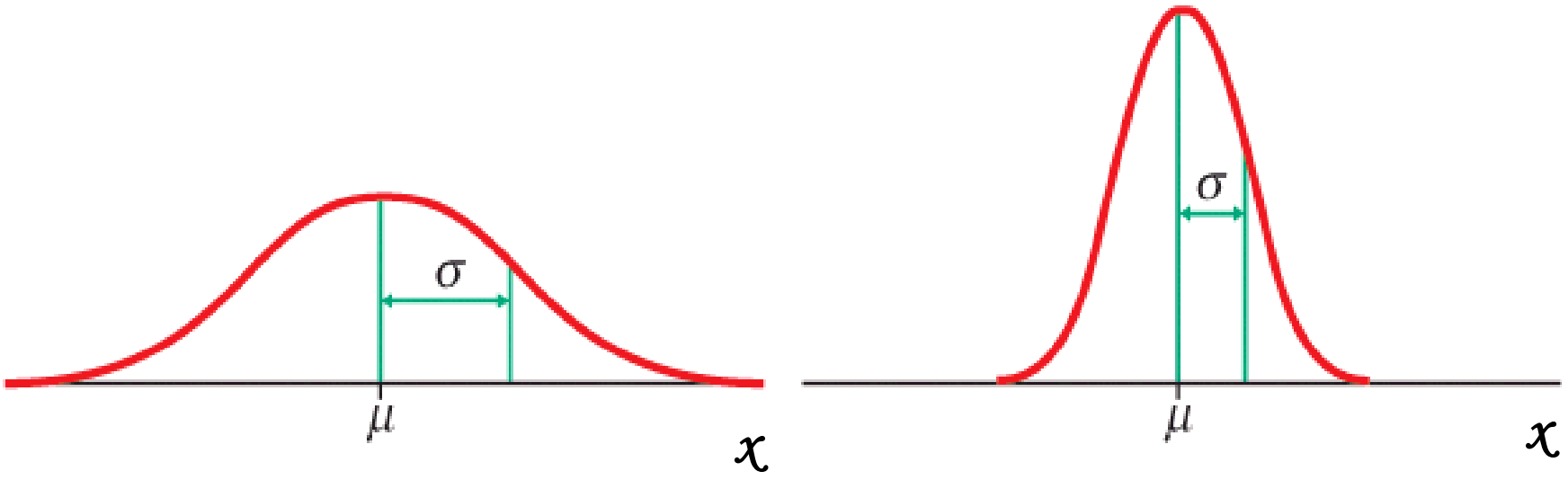






Normal distributions

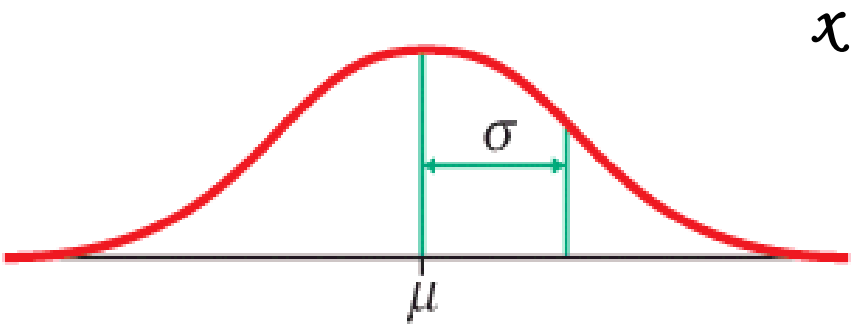
Normal—or Gaussian—distributions are a **family** of symmetrical, bell- shaped density curves defined by a mean μ (*mu*) and a standard deviation σ (*sigma*): $N(\mu, \sigma)$.

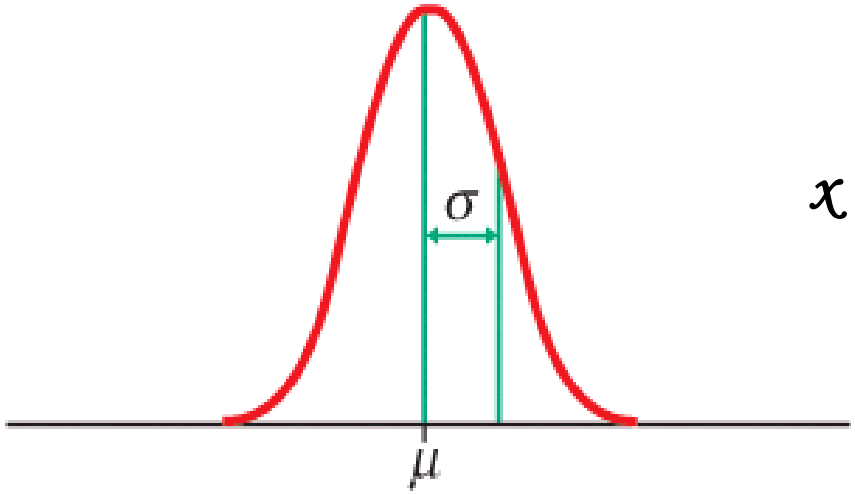


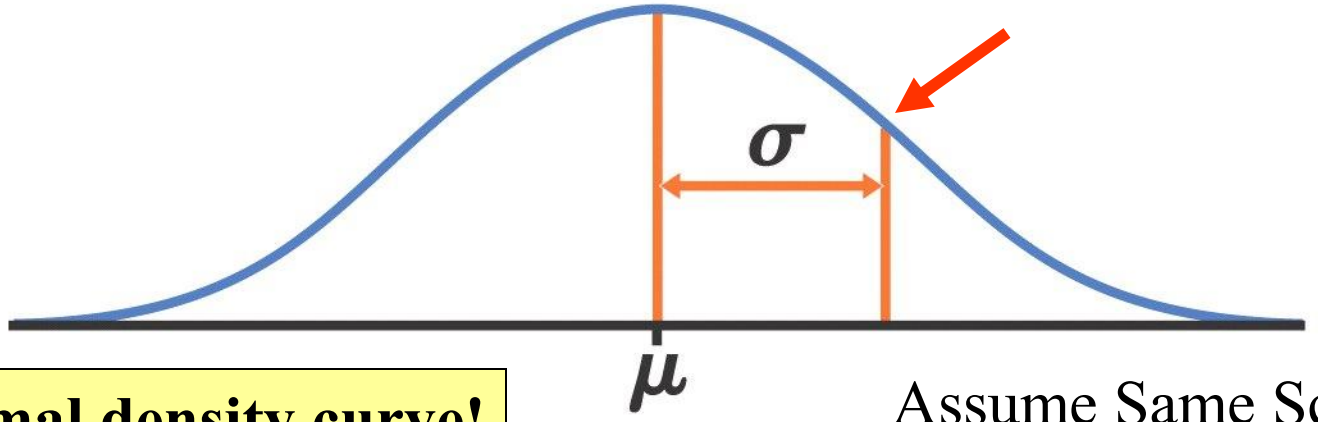
Normal distributions

Normal—or Gaussian—distributions are a **family** of symmetrical, bell- shaped density curves defined by a mean μ (*mu*) and a standard deviation σ (*sigma*): *N* (μ, σ).

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



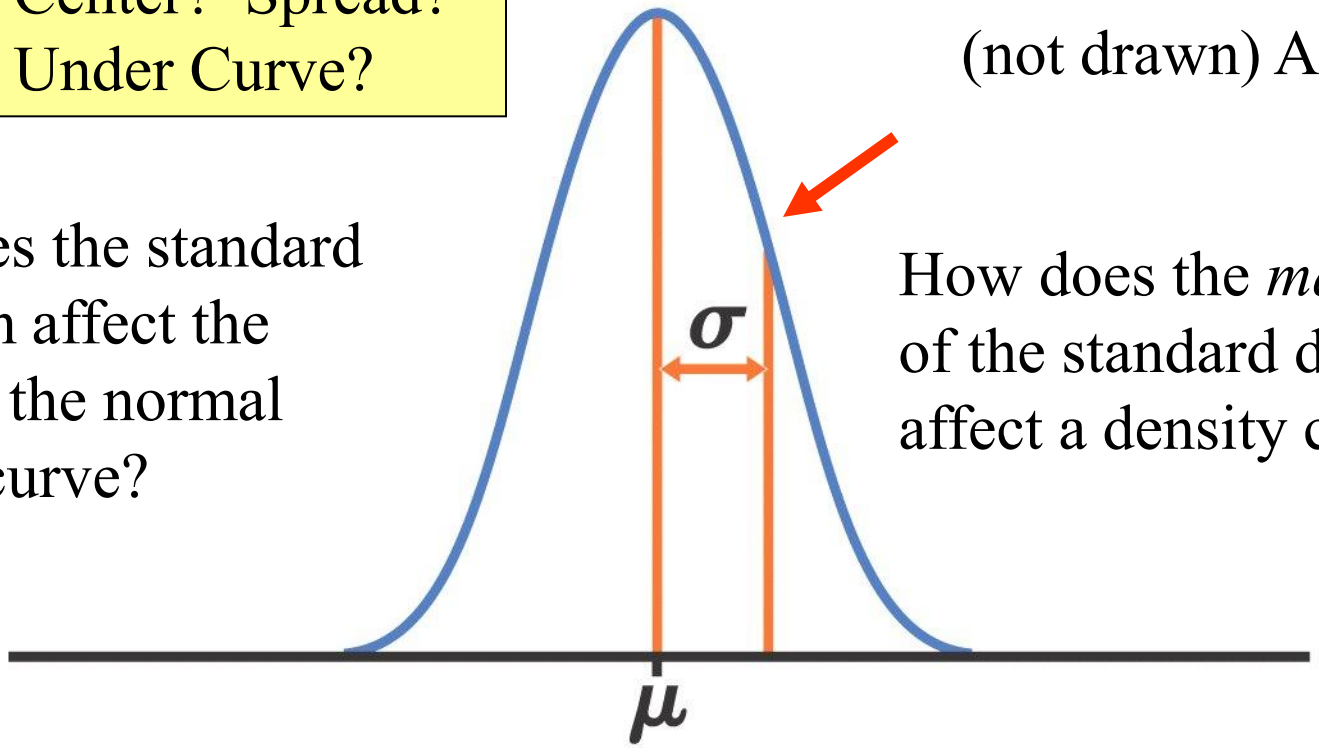




The normal density curve!
Shape? Center? Spread?
Area Under Curve?

Assume Same Scale on Horizontal and Vertical (not drawn) Axes.

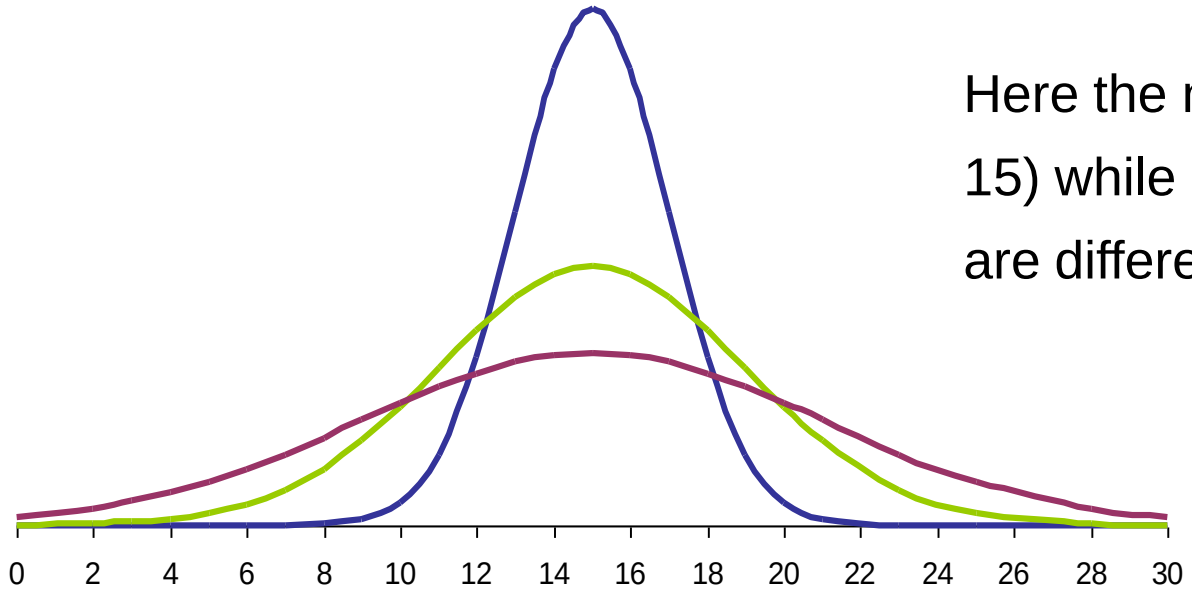
How does the standard deviation affect the shape of the normal density curve?



How does the *magnitude* of the standard deviation affect a density curve?

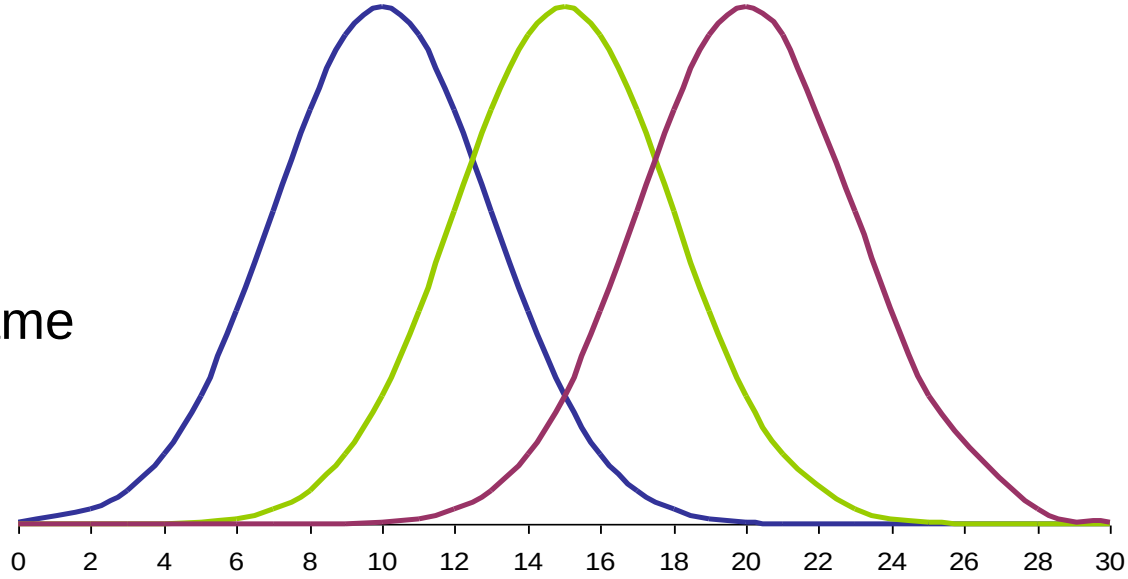
Figure 1-26
Introduction to the Practice of Statistics, Fifth Edition

A family of density curves



Here the means are the same (μ = 15) while the standard deviations are different (σ = 2, 4, and 6).

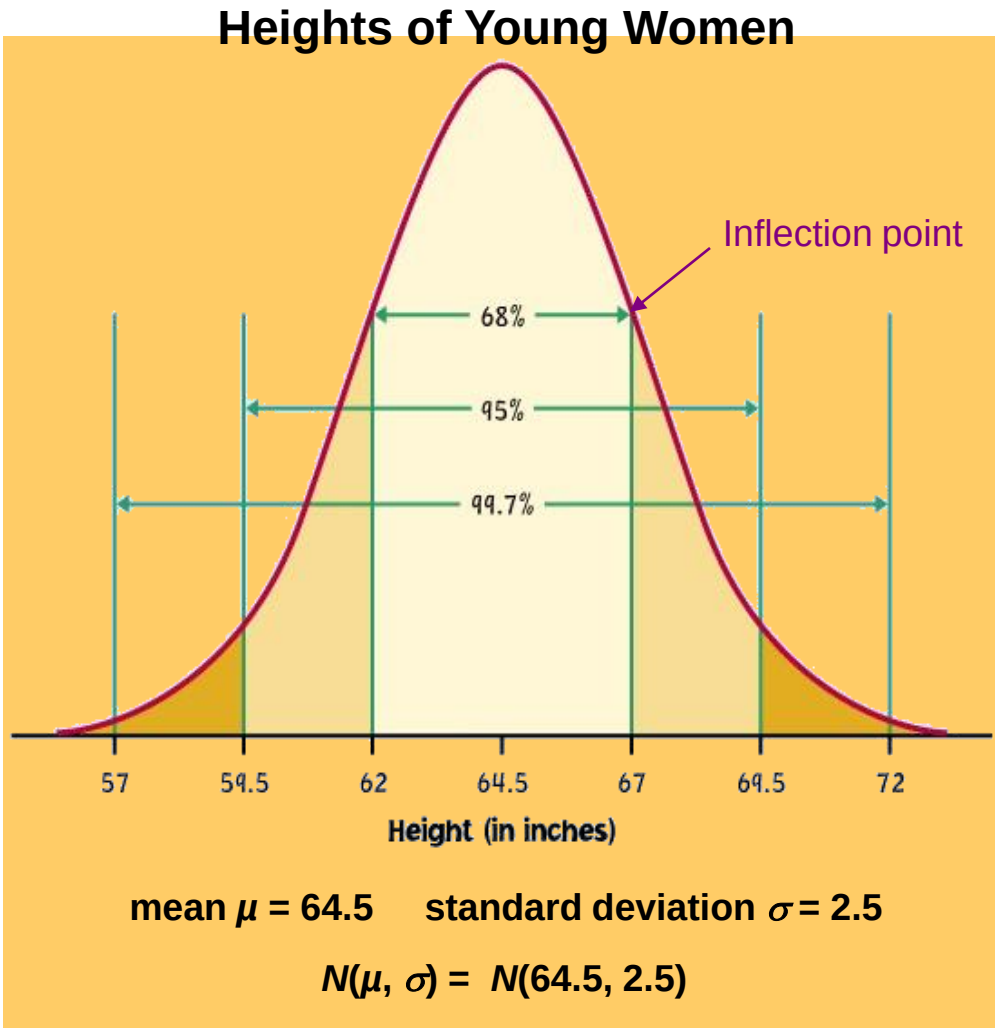
Here the means are different (μ = 10, 15, and 20) while the standard deviations are the same (σ = 3).



All Normal curves $N(\mu, \sigma)$ share the same properties

- About 68% of all observations are within 1 standard deviation (σ) of the mean (μ).
- About 95% of all observations are within 2 σ of the mean μ.
- Almost all (99.7%) observations are within 3 σ of the mean.

- This is called the **68-95-99.7 Rule (aka Empirical Rule)**



$\bar{x}$

How to quickly sketch a normal distribution

- The empirical rule says that 99.7% of the population lies between $\mu - 3\sigma$ and $\mu + 3\sigma$.
- This means that there is very little data outside this range.
  - Thus the normal density curve is almost identical with the x-axis outside the interval from $\mu - 3\sigma$ to $\mu + 3\sigma$.
- So the "visible part" of the bell-shaped density curve is between $\mu - 3\sigma$ and $\mu + 3\sigma$.
- Example: sketch the $N(\mu = 100, \sigma = 15)$ distribution.

My shorthand notation for normal distributions

- I abbreviate the statement "X is normal with mean $\mu = 64.5$ and standard deviation $\sigma = 2.5$" as follows:

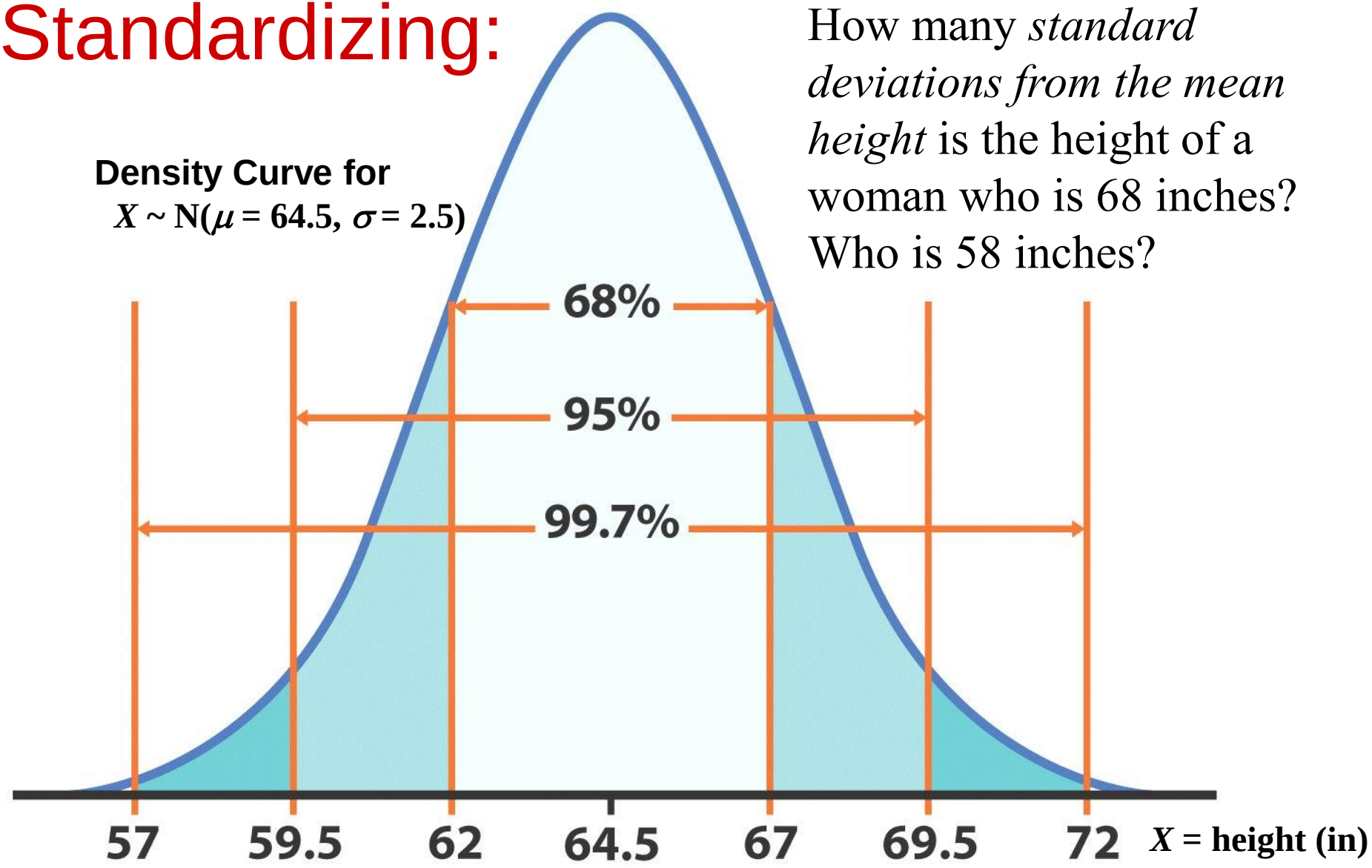
$$X \sim N(\mu = 64.5,\ \sigma = 2.5)$$

- …sometimes I'll leave off the μ and σ:

$$X \sim N(64.5,\ 2.5)$$

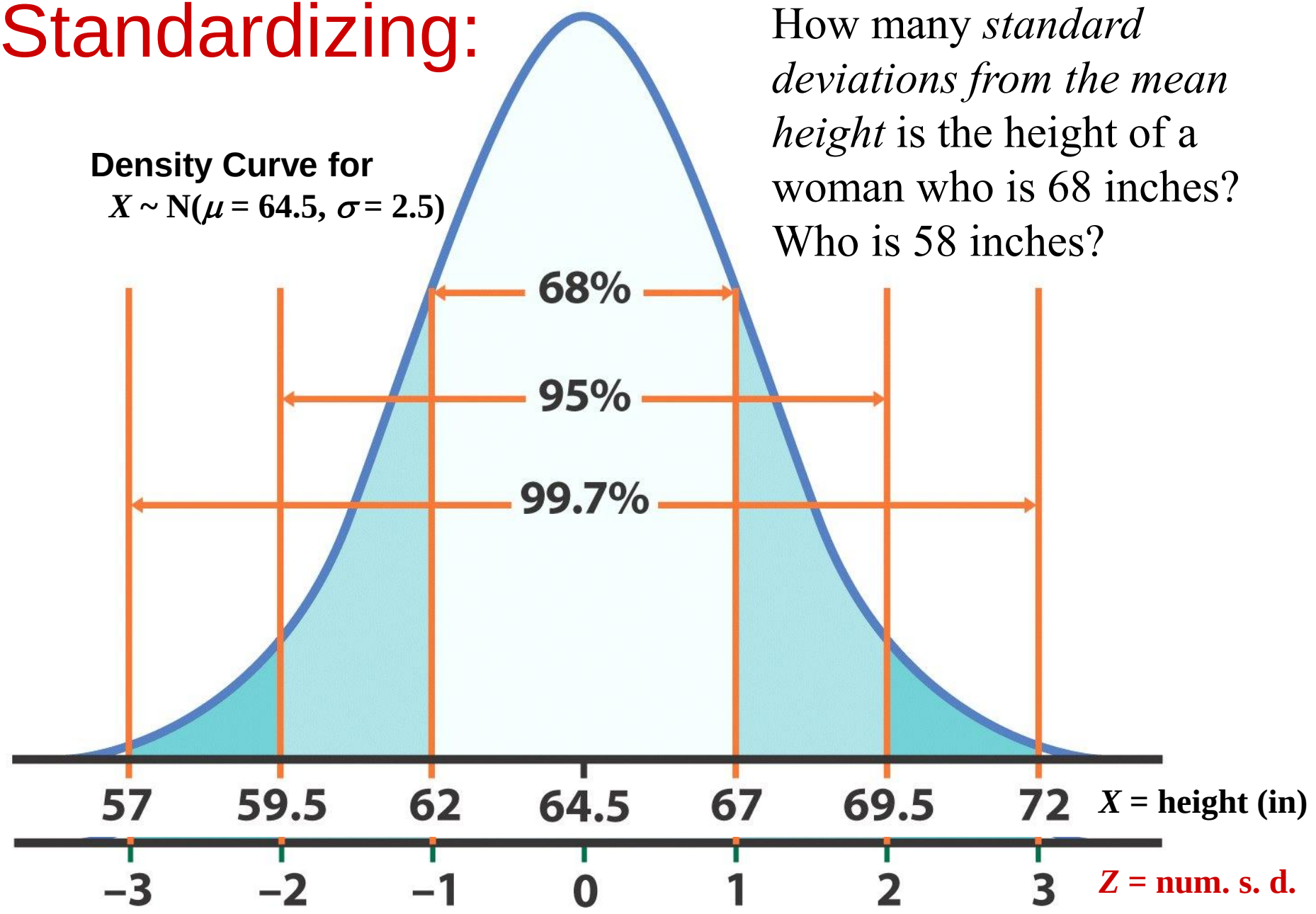
Standardizing:

How many *standard deviations from the mean height* is the height of a woman who is 68 inches? Who is 58 inches?



Standardizing:

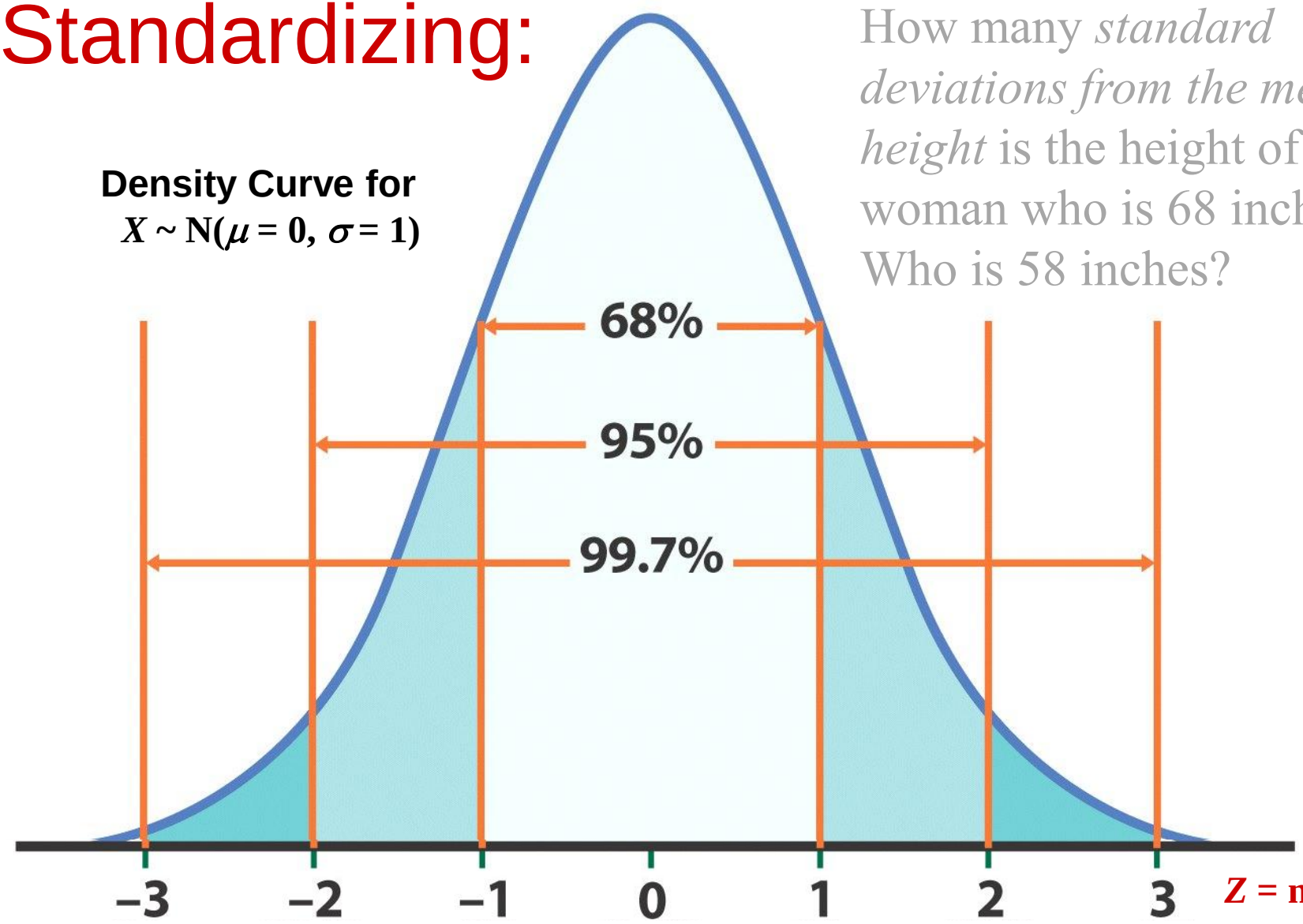
How many *standard deviations from the mean height* is the height of a woman who is 68 inches? Who is 58 inches?



Standardizing:

How many *standard deviations from the mean height* is the height of a woman who is 68 inches? Who is 58 inches?

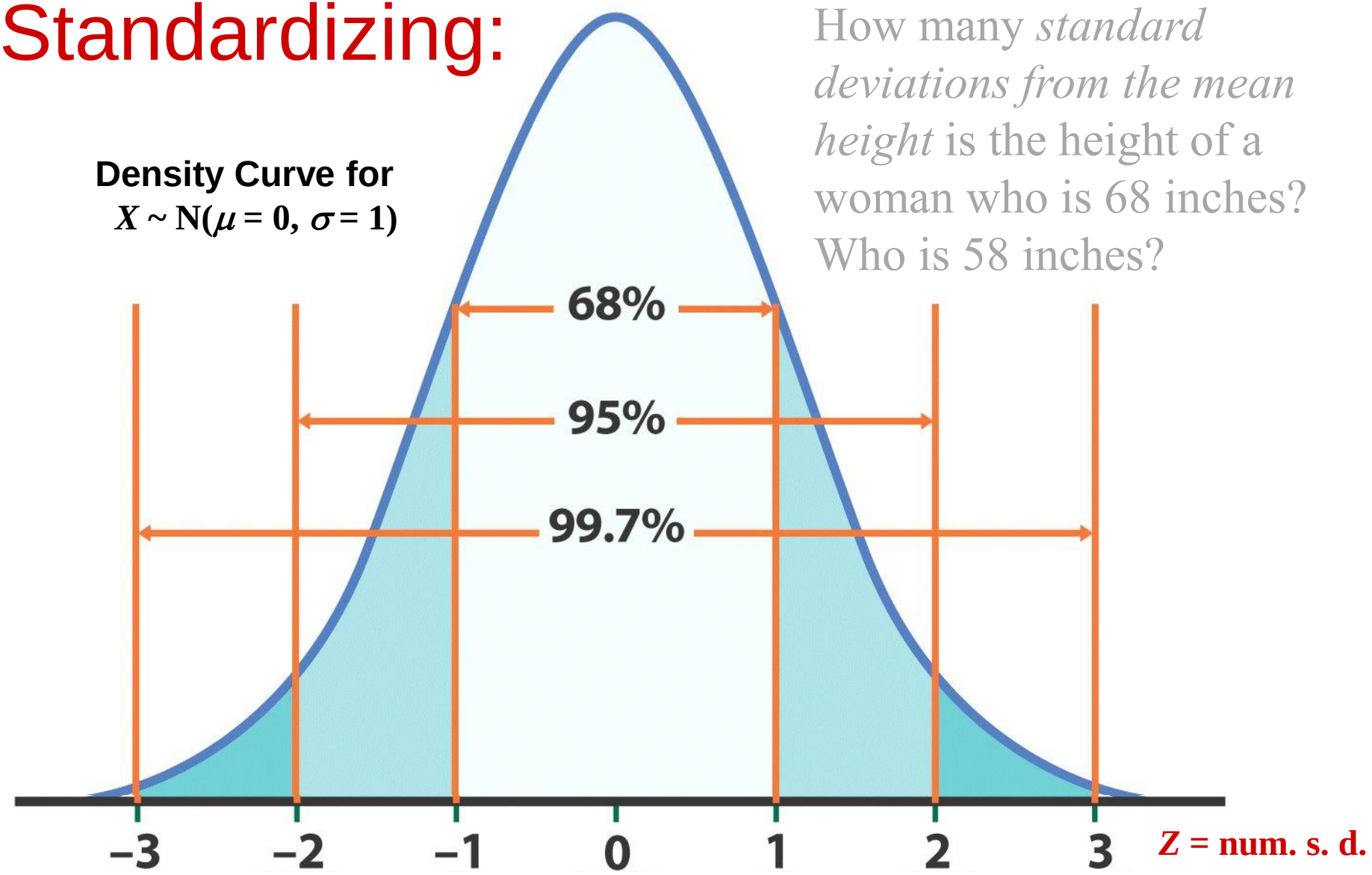
Density Curve for
$X \sim N(\mu = 0, \sigma = 1)$



Z = num. s. d.

Standardizing:

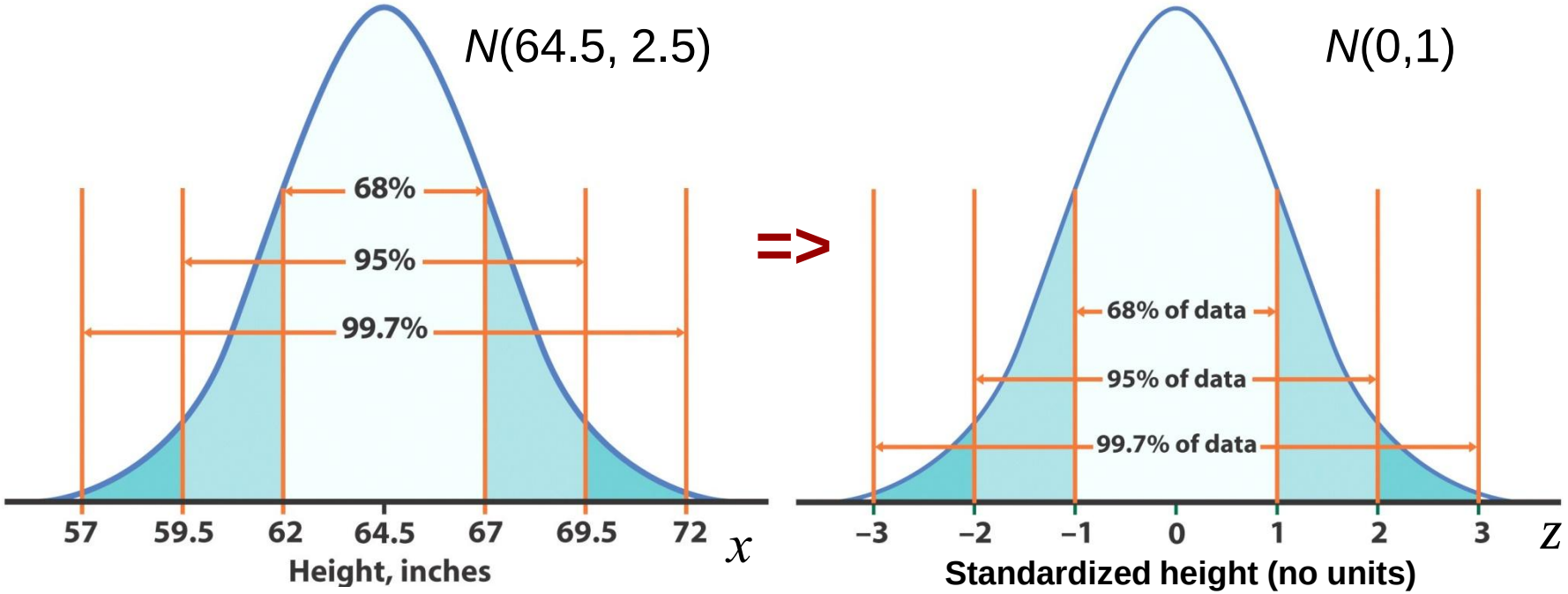
How many *standard deviations from the mean height* is the height of a woman who is 68 inches? Who is 58 inches?



The Standard Normal Distribution ($\mu = 0$, $\sigma = 1$)

Standardizing: Standard Normal Distribution

Because all Normal distributions share the same properties, we can **standardize** our data to transform any Normal curve $N(\mu, \sigma)$ into the standard Normal curve $N(0,1)$.



For each x we calculate a new value, z (called a z-score). The z-score is the number of standard deviations that a data value x is from the mean.

Standardizing: calculating *z*-scores

A ***z*-score** measures the number of standard deviations that a data value *x* is from the mean μ.

$$z = \frac{(x - \mu)}{\sigma}$$

Using the previous slides,
find the z-score of a woman
whose height is 58 inches;
whose height is 73 inches.

- *When x is larger than the mean, z is positive.*
- *When x is smaller than the mean, z is negative.*
- *If x has a normal distribution with mean μ and standard deviation σ, then z will have a standard normal distribution.*

Standardizing: converting *z*-scores back to the original units

Now let's go backwards. Given a ***z*-score** find the corresponding data value *x*. How?

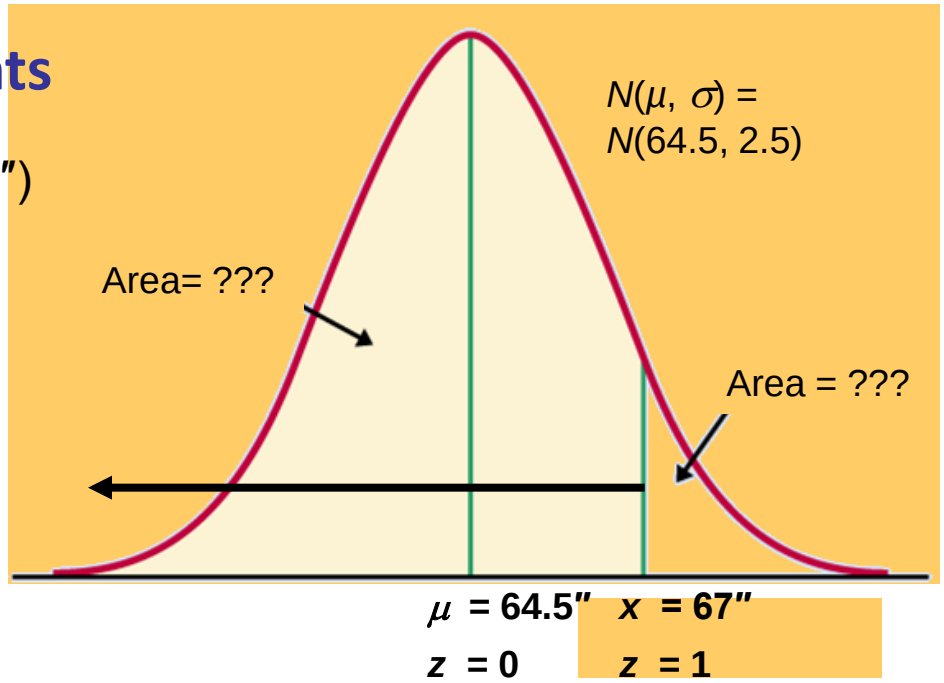
Solve the equation $z = \dfrac{(x - \mu)}{\sigma}$ for *x*

We get $x = \mu + z\sigma$

For the height data on the previous slides, how tall is a woman whose *z*-score is $z = 1.7$?

Example: *X* = Women's heights

Women's heights follow the *N*(64.5″,2.5″) distribution. Using the empirical rule, determine what percent of women are shorter than 67 inches tall (that's 5′7″)?



mean μ = 64.5"
standard deviation σ = 2.5"
x (height) = 67"

We calculate *z*, the standardized value of *x*:

$$z = \frac{(x-\mu)}{\sigma}, \quad z = \frac{(67-64.5)}{2.5} = \frac{2.5}{2.5} = 1 => 1\,\text{stand. dev. from mean}$$

Because of the 68-95-99.7 rule, we can conclude that the percent of women shorter than 67″ should be, approximately, .68 + half of (1 − .68) = .84, or 84%.

Using Our Calculators to Find the Percent of Women Shorter than 67”

- **TI-83 Calculator Command: Distr|normalcdf**
- **Syntax: normalcdf(left, right, mu, sigma) = area under normal curve (with mean mu and standard deviation sigma) from left to right**
- **mu defaults to 0, sigma defaults to 1**
- **Infinity is 1E99 (use the EE key), Minus Infinity is -1E99**

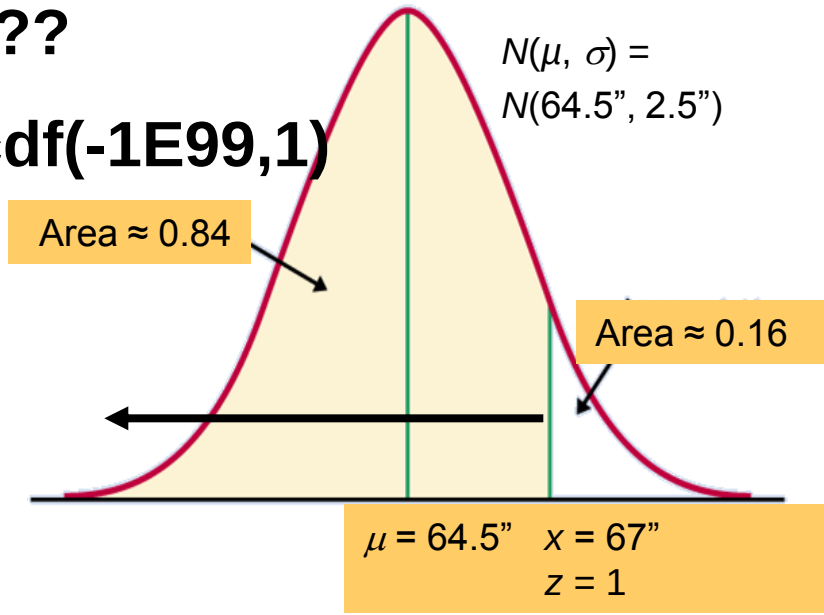
Normalcdf(-1E99, 67, 64.5, 2.5)

Can we compute using z-scores??

Normalcdf(-1E99,1,0,1)=Normalcdf(-1E99,1)

Question:

What percent of women are <u>taller</u> than 67“??



My “P” notation for proportions

- I abbreviate the statement “the proportion of the population with X between 65 and 70” as follows:

$$P(65 < X < 70)$$

- …for a normal distribution, there is no difference between this and

$$P(65 <= X <= 70)$$

(Why?)

The National Collegiate Athletic Association (NCAA) requires Division I athletes to score at least 820 on the combined math and verbal SAT exam to compete in their first college year. The SAT scores of 2003 were approximately normal with mean 1026 and standard deviation 209.

What proportion of all students would be NCAA qualifiers? **P(X ≥ 820)**

$x = 820$

$\mu = 1026$

$\sigma = 209$

$$z = \frac{(x-\mu)}{\sigma}$$

$$z = \frac{(820-1026)}{209}$$

$$z = \frac{-206}{209} \approx -0.99$$

Draw Picture

Compute

Normalcdf(820,1E99,1026,209)

Normalcdf(-.99,1E99)

State Answer

Area right of 820 = Total area − Area left of 820

= 1 − 0.1611

≈ 84%

Answer: Approximately 84% of students who took the SAT in 2003 scored at least 820.

Note: The actual data may contain students who scored exactly 820 on the SAT. However, the proportion of scores exactly equal to 820 being 0 for a normal distribution is a consequence of the idealized smoothing of density curves.

The NCAA defines a “partial qualifier” eligible to practice and receive an athletic scholarship, but not to compete, as a combined SAT score of at least 720.

What proportion of all students who take the SAT would be partial qualifiers?

P(720 <= X <= 820)

$x = 720$

$\mu = 1026$

$\sigma = 209$

$$z = \frac{(x - \mu)}{\sigma}$$

$$z = \frac{(720 - 1026)}{209}$$

$$z = \frac{-306}{209} \approx -1.46$$

Draw Picture

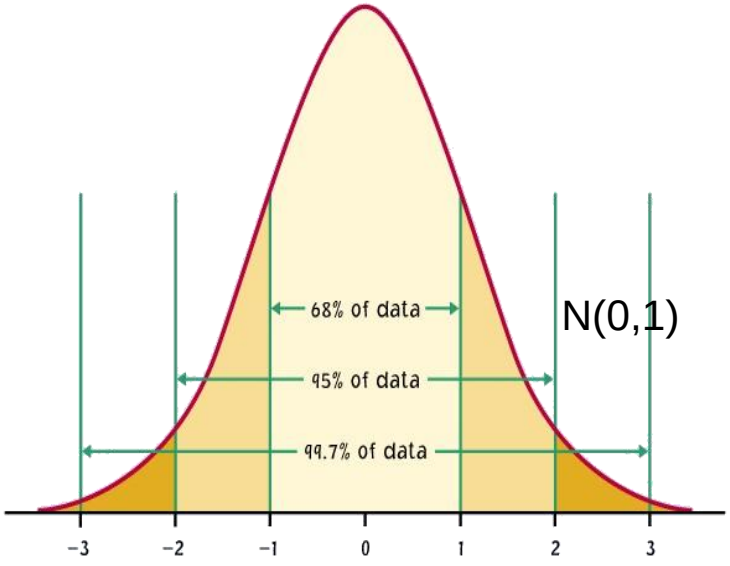
Compute

Normalcdf(720,820,209)
Normalcdf(-1.46,-0.99)

State Answer

About 9% of all students who take the SAT have scores between 720 and 820.

One cool thing about working with normally distributed data is that we can manipulate it and then find answers to questions that involve comparing seemingly non-comparable distributions.



We do this by “standardizing” the data. All this involves is changing the scale so that the mean now equals 0 and the standard deviation equals 1. If you do this to different distributions, it makes them comparable.

$$z = \frac{(x - \mu)}{\sigma}$$

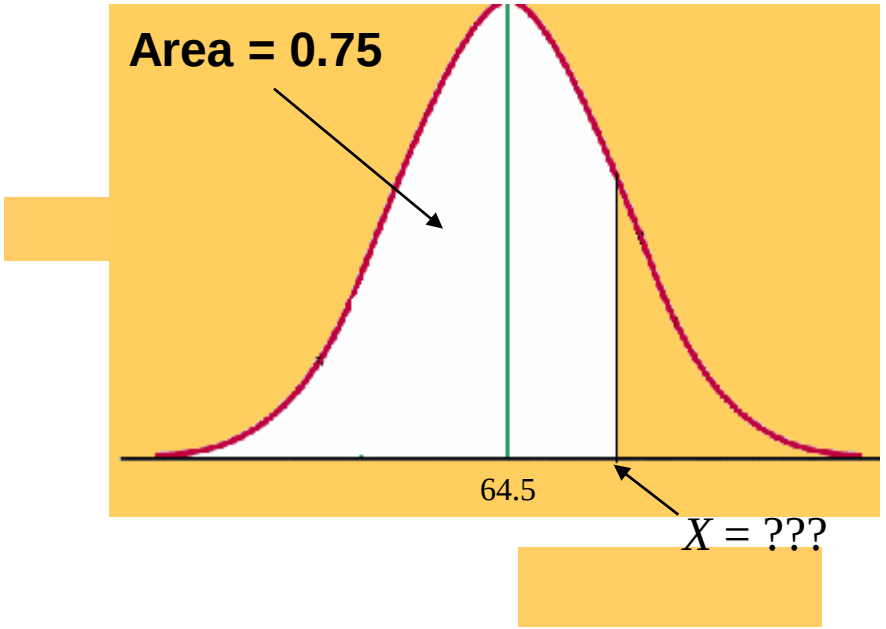
Example: comparing test scores

- Andrew and Brian are in two different math classes.
- Andrew scored 114 out of 120 on his exam. The class average was 96, with a standard deviation of 12.
- Brian scored 162 out of 180 on his exam. His class average was 126, with a standard deviation of 22.
- The distributions of exam scores were normal for both classes.
- Who did better, Andrew or Brian?
  - In terms of raw scores…
  - In terms of z-scores…
  - In terms of cumulative proportions…

Finding a value given a proportion

When you know the proportion, but you don't know the *x*-value that represents the cut-off, you have the inverse problem.

Example: *X* = Women's heights



Women's heights follow the *N*(64.5″,2.5″) distribution. What is the 75th percentile for women's heights?

Draw a picture!

Compute:

We will use our calculator to get X

mean μ = 64.5"

standard deviation σ = 2.5"

proportion = area under curve=0.75

Calculator command:

invNorm(0.75, 64.5, 2.5)

State Answer:

The 75th percentile for women's heights is X = 66.19", or 5' 6.19".

Evanston Township High School

Data Analysis and Statistics

Introduction to Statistics

Jiangtao Gou
February 20, 2013

Lecture 6: Normal Proportions and Confidence Intervals

What we have learned

- Variance and standard deviation
- Correlation
- Normal Distribution

Outline

- Z-Scores
- Normal Proportions
- Confidence Intervals

Z-Scores

- Question
  - Tom scored 680 on the mathematics part of the SAT.
  - Sue took the ACT assessment mathematics test and scored 27.
  - Who actually had the higher score?

Z-Scores

- Further Information
  - The distribution of SAT math scores was Normal with mean 515.
  - The distribution of ACT math scores was Normal with mean 21.0.
  - Can you answer this question now?

Z-Scores

- Further Information
  - The distribution of SAT math scores was Normal with SD 114.
  - The distribution of ACT math scores was Normal with SD 5.1.
  - How about now? Can you answer this question?

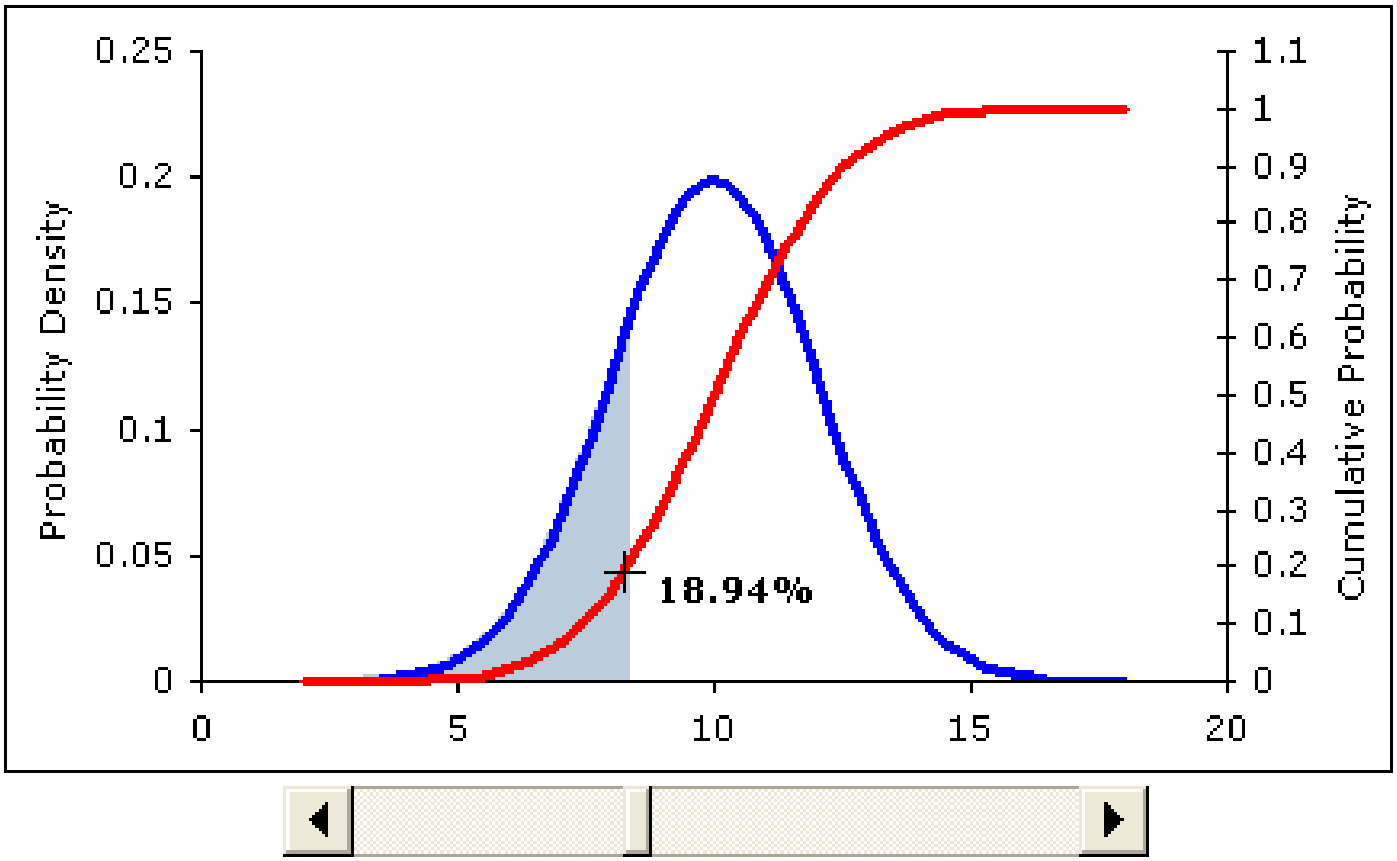
Z-Scores: Exercises

- The heights of women aged 20 to 29 are approximately Normal with mean 64 inches and SD 2.7 inches. Men the same age have mean height 69.3 inches with SD 2.8 inches. What are the Z-scores for a woman 6 feet tall and a man 6 feet tall? Try to say what information the Z-scores give that the actual heights do not.

Normal Proportions

- One-to-one correspondence between
  - Z-score
  - Cumulative proportion
    - The cumulative proportion for a value x in a distribution is the proportion of observations in the distribution that are less than or equal to x.

Cumulative Proportions

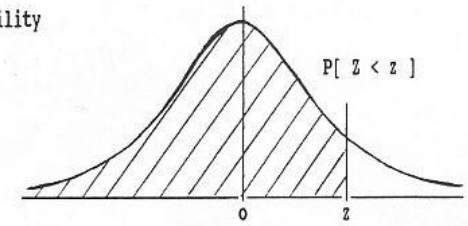


STANDARD STATISTICAL TABLES

1. Areas under the Normal Distribution

The table gives the cumulative probability up to the standardised normal value z

i.e. $$P[\, Z < z \,] = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}Z^2)\, dZ$$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5159	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7854
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8804	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9773	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9865	0.9868	0.9871	0.9874	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9924	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9980	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
z	3.00	3.10	3.20	3.30	3.40	3.50	3.60	3.70	3.80	3.90
P	0.9986	0.9990	0.9993	0.9995	0.9997	0.9998	0.9998	0.9999	0.9999	1.0000

Normal Proportions

- SAT and ACT scores
  - What's Tom's normal proportion?
  - What's Sue's normal proportion?

Exercise

- 6-feet tall woman and 6-feet tall man

Confidence interval

- A confidence interval (CI) gives an estimated range of values which is likely to include an unknown population parameter, the estimated range being calculated from a given set of sample data.

Example

- Suppose a student measuring the boiling temperature of a certain liquid observes the readings (in degrees Celsius) 102.5, 101.7, 103.1, 100.9, 100.5, and 102.2 on 6 different samples of the liquid. He calculates the sample mean to be 101.82. If he knows that the standard deviation for this procedure is 1.2 degrees, what is the confidence interval for the population mean at a 95% confidence level?

Evanston Township High School

Data Analysis and Statistics

Introduction to Statistics

Jiangtao Gou
March 6, 2013

Lecture 7: Hypothesis Testing

What we have learned

- Normal Distribution
- Z-Scores
- Normal Proportions

$$z = \frac{x - \mu}{\sigma}$$

Outline

- Hypothesis Testing
  - Criminal Trial
  - Playing Bowling
  - Environment contamination
- Small Project (Due March 13th)

Criminal Trials

- The basic concepts in hypothesis testing are actually quite analogous to those in a criminal trial.

Criminal Trials

		Person is: Innocent	Guilty
Jury Says:	Innocent	No Error	Error
	Guilty	Error	No Error

		Person is: Innocent	Guilty
Jury Says:	Innocent	No Error	Error
	Guilty	Error	No Error

- Are both of these errors equally important?
- Is it as bad to decide that a guilty person is innocent and let them go free as it is to decide an innocent person is guilty and punish them for the crime?

Null and Alternative Hypothesis

- In a criminal trial, there actually is a **favored assumption**, an initial bias if you will. The jury is instructed to assume the person is innocent, and only decide that the person is guilty if the evidence convinces them of such.
- Null Hypothesis: The person is innocent
- Alternative Hypothesis: The person is guilty

Example: Bowling

- Hanging out with Tom
  - Let us go bowling.
  - Tom says “My long-term average is 150”
  - Over 4 games, Tom’s average score is 50
  - Do you believe him?

Example: Bowling

- Over 4 games, Tom's average score is 50
  - Don't believe him

What if

- Over 4 games, Tom’s average score is 140

What if

- Over 4 games, Tom's average score is 140
  - More likely to believe him

- At what point, between 50 and 140, do you make the decision to believe Tom or not?

Example: Bowling

- What is your cut-off score for Tom?
- There is a claim, and if a sample outcome falls below a cut-off value (based on the assumption that the claim is true), then we reject the claim.

Hypothesis

$$H_0 : \mu = 150$$

$$H_a : \mu < 150$$

Standard Deviation of a Sample Mean

- A population has mean μ and standard deviation σ
- The average of a Simple Random Sample has mean μ and standard deviation $\sigma/\sqrt{n}$

- Averages are less variable than individual obsrvations.

Z-Scores

$$Z = \frac{x - \mu_0}{\sigma}$$

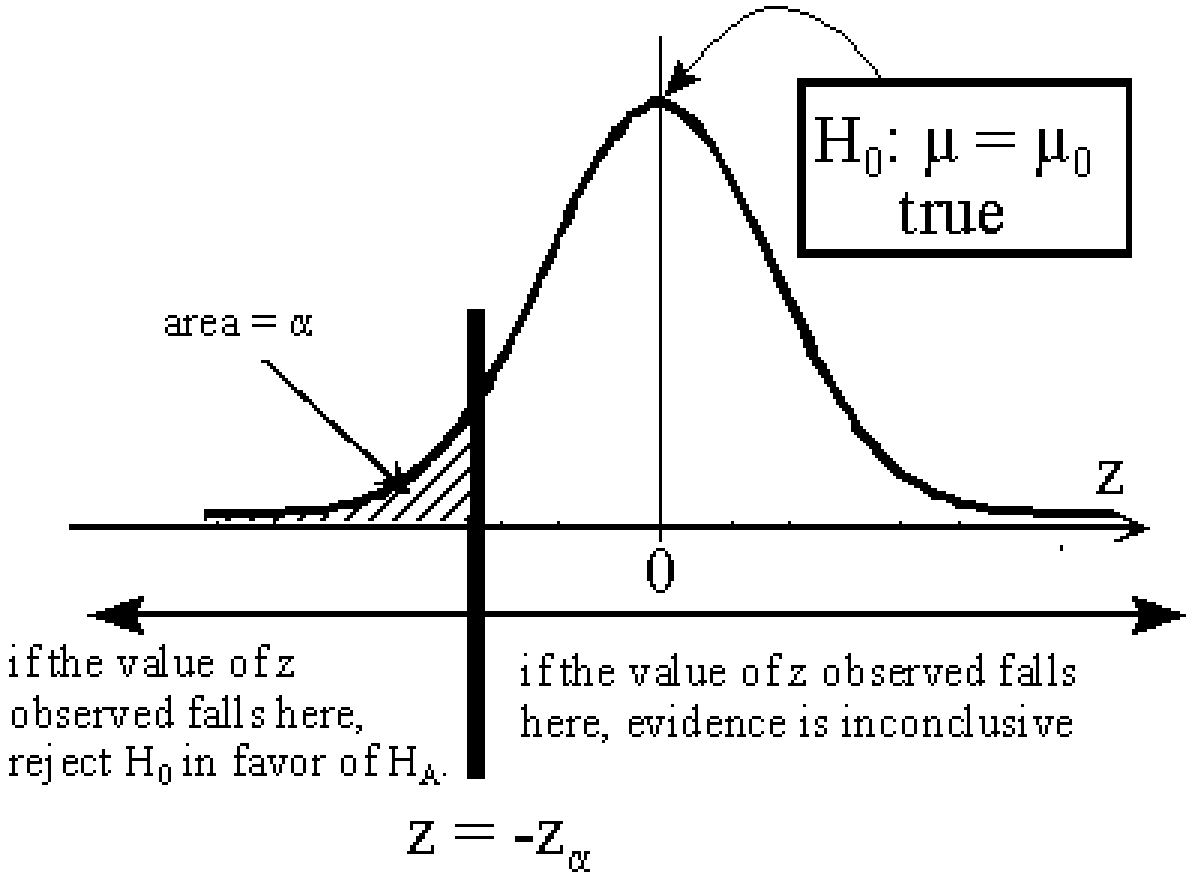
$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

Z-Scores

$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

- Now, by assuming the standard deviation σ is equal to 70, can you draw any conclusion?

When α is 0.05, the critical value is -1.645



$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

- Tom said his average is 150, but his four-game average was only 40.
- Assume that the standard deviation σ is equal to 70
- What's the Z-score? By comparing it with the critical value -1.645, what conclusion do you get?

Practice: Mercury in the river

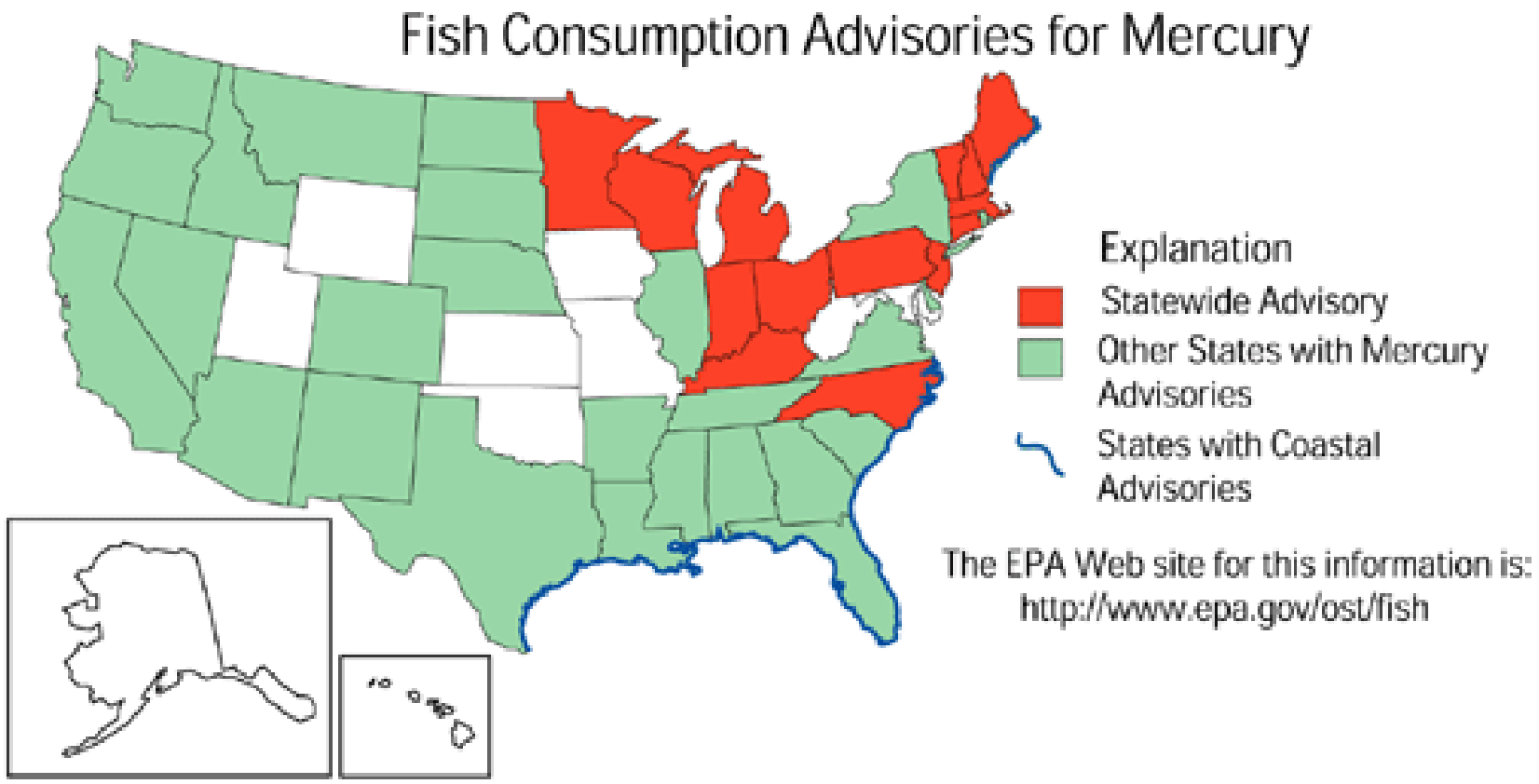
Practice: Mercury in the river

If you regularly eat types of fish that are high in methylmercury, it can accumulate in your blood stream over time

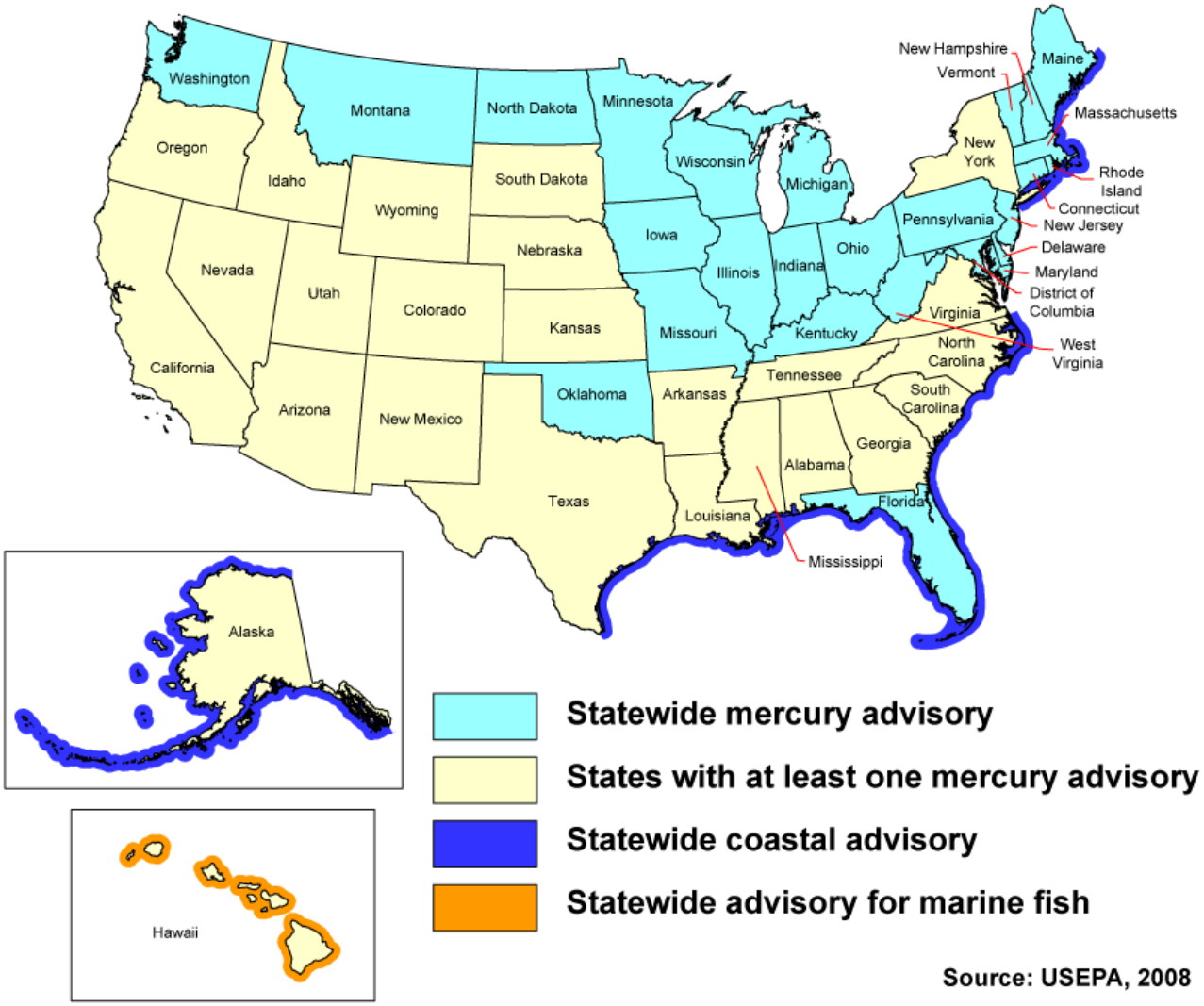
Customer Advisory: Mercury in Seafood

The FDA recommends that pregnant and nursing women, women who may become pregnant and young children should avoid consumption of fish such as swordfish, shark, tilefish and king mackerel. These groups should also limit their intake of fresh, frozen and canned tuna. To learn more, please visit our website at www.wholefoods.com, or ask for our "Methylmercury in Seafood" brochure at the seafood counter.

States with mercury fish consumption advisories (2000)



States with mercury fish consumption advisories (2008)




WARNING
AVISO
Catfish
(Bagre)
Largemouth Bass
(Perca de boca grande)
Elevated levels of PCBs have been found in some catfish in this lake. Largemouth bass may have elevated levels of mercury.
Do not eat any catfish and no more than 2 meals per month of largemouth bass from this lake.
If you are pregnant, planning to get pregnant, are nursing, or are a child under 15 years of age, do not eat any of these fish.
Swimming, boating, and handling fish do not present a known health risk.
For more information call:
N.C. CARELINE at 1-800-662-7030
Se han encontrado niveles altos de BPCs en algunos peces bagre de este lago. La perca de boca grande puede tener niveles altos de mercurio.
No coma ningún pez bagre y no más de dos porciones al mes de perca de boca grande de este lago.
No coma ninguno de estos peces, si está embarazada, planea quedar embarazada, está amamantando, o es un niño/a menor de 15 años.
El nadar, pasear en bote o tocar los peces no presenta un riesgo conocido para la salud.
Para más información llame
N.C. CARE-LINE al 1-800-662-7030
Jeffrey P. Engel, M.D., State Health Director
1/6/2011

Advisory Levels for Mercury in Tennessee

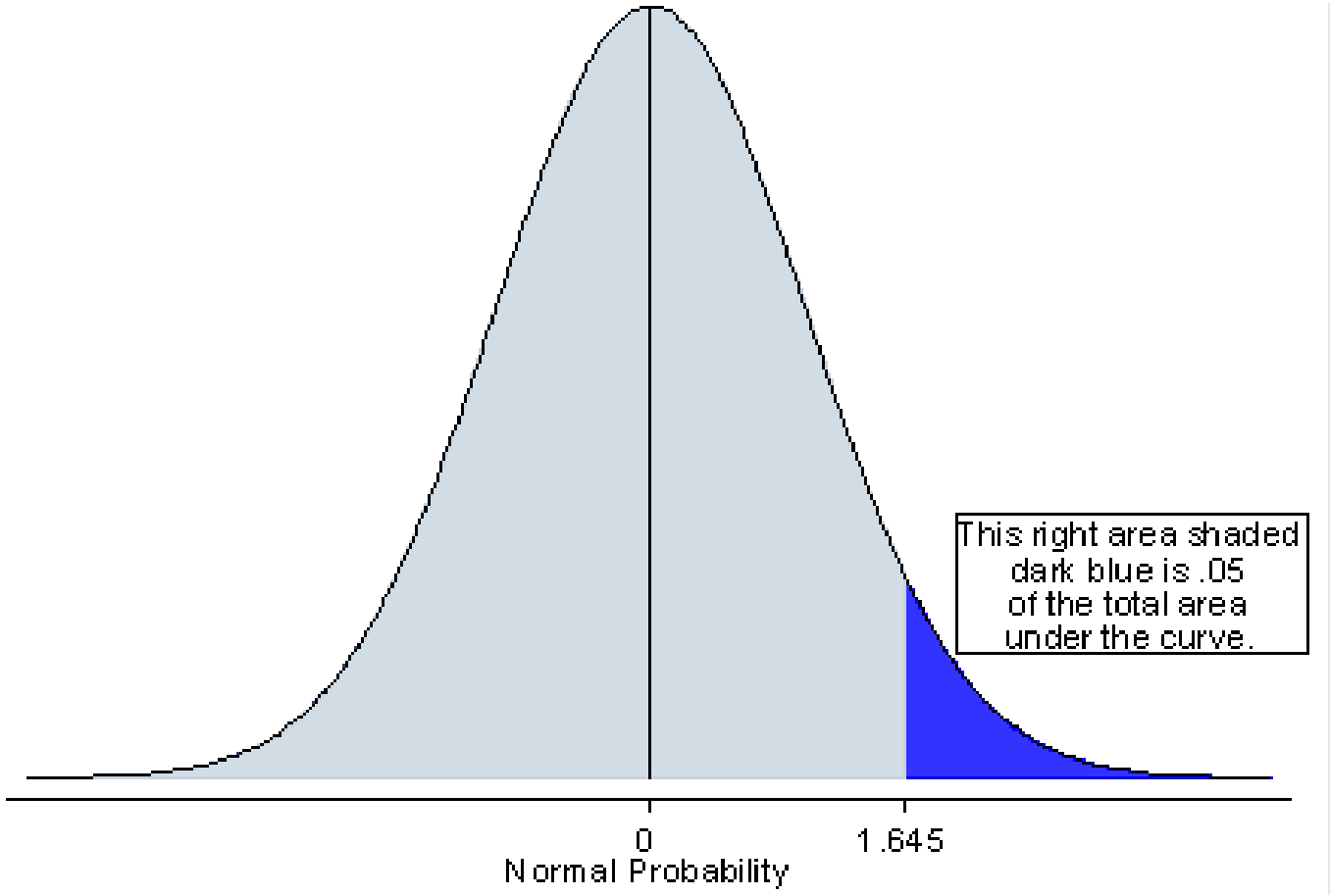
Precautionary Advisory **0.3 ppm average concentration**

Children, pregnant women, or nursing mothers should not eat these fish in any amount. Persons with a previous occupational exposure to the chemical of concern should also not eat these fish. All other persons should limit consumption of these fish to one or two meals per month.

Do Not Consume Advisory **1.0 ppm average concentration**

No persons should eat these fish in any amount.

When α is 0.05, the critical value is 1.645



Bruce D. Spencer

Professor of Statistics

Department Chair

Ph.D., 1979, Yale University

bspencer@northwestern.edu

Research Interests

I study the production and use of public statistics, particularly in policy-laden contexts. For example, should an organization (such as the Federal government) spend more money (or less) on statistical programs such as the decennial census? Such a question requires an interdisciplinary attack, and I have focused on questions such as (i) how are the data used, (ii) what is the quality (including accuracy, relevance, timeliness, etc.) of the data, (iii) how would changes in the quality affect uses, and (iv) how can we assign measures of value to those effects. I work actively with government agencies on major statistical programs and conduct related research in sampling theory and methods, particularly weighting. I am currently assessing the accuracy of population forecasts and trying to get the government to provide stochastic estimates of uncertainty along with their forecasts, such as Social Security forecasts. I am also working with the Census Bureau on how to estimate population and how to evaluate the accuracy of their estimates. With support from The Searle Fund, I am analyzing the accuracy of randomized social experiments, in particular the Head Start Impact Study.

Evanston township High School

Data Analysis and Statistics

Introduction to Statistics

Jiangtao Gou
March 15, 2013

2013 Evanston Township High School Statistics Research Colloquium

Student Presentations

Presentations

- Average Salaries in the Chicago area
  - by Paul Lemon, Marilu Sierra and Enrique Miyasato
- Women's height on the Chicago Sky Women's Basketball team
  - by Erin Boothe and Kelsea Frazier
- Weight of a Wrestler
  - by Anthony Derrick
- and more…

Videos

- The Joy of Stats: 200 Countries, 200 Years, 4 Minutes
  - http://www.youtube.com/watch?v=jbkSRLYSojo&list=PLE63F079456D5E210

- Monty Hall Problem: Two Goats and One Car
  - http://www.youtube.com/watch?v=OBpEFqjkPO8&list=PLE63F079456D5E210